AntiCP3: Prediction of Anticancer Proteins using Evolutionary Information from Protein Language Models

by

Amisha Gupta

Under the Supervision of

Prof. G.P.S Raghava

Indraprastha Institute of Information Technology Delhi

May, 2025

AntiCP3: Prediction of Anticancer Proteins using Evolutionary Information from Protein Language Models

by

Amisha Gupta

Submitted

in partial fulfillment of the requirements for the degree of

Master of Technology

to

Indraprastha Institute of Information Technology Delhi

May, 2025

Certificate

This is to certify that the thesis titled **AntiCP3: Prediction of Anticancer Proteins using Evolutionary Information from Protein Language Models** being submitted by **Amisha Gupta** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

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.

May, 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 **Milind Chauhan** 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.

Amisha Gupta
MT23225

Abstract

Several computational tools have been developed to predict the anticancer nature of peptides, including AntiCP and AntiCP2. While these methods have been widely adopted by the scientific community, they are not suitable for predicting anticancer proteins, as they differ significantly in composition and sequence characteristics. In this study, we introduce AntiCP3, the first dedicated platform for the accurate prediction of anticancer proteins. Our approach begins with an in-depth compositional analysis, which revealed clear differences between anticancer peptides and proteins, reinforcing the need for a distinct predictive framework. To build this, we first implemented similarity-based methods, which provided only moderate performance. We then developed a range of machine learning and deep learning models using conventional protein features such as amino acid composition (AAC), dipeptide composition (DPC), and physicochemical properties (PCP). The Extra Trees classifier achieved the best performance among traditional models, with a maximum AUROC of 0.72. To enhance performance, we integrated evolutionary features by extracting Position-Specific Scoring Matrix (PSSM) profiles, which improved the AUROC to 0.79. We further fine-tuned the pre-trained ESM2-t33 protein language model on our curated dataset, using its ability to capture both structural and contextual information. This led to a significant increase in the performance, achieving an AUROC of 0.90. Finally, we developed a hybrid model that combines BLAST-based sequence similarity scores with the fine-tuned ESM2 model, resulting in the highest performance with an AUROC of 0.91. All models were rigorously trained using manual five-fold cross-validation, and the performance was further validated using an independent test set. To facilitate widespread usage, AntiCP3 has been implemented as both a user-friendly web server and a standalone package. Additionally, the best-performing model has been deployed on Hugging Face as open access enabling direct integration into computational pipelines and promoting reproducible research.

Contents

List of Tables

List of Figures

List of Abbreviations

AAC:	Amino Acid Composition
ACPs:	Anticancer Peptides
AUC:	Area Under the Curve
BERT:	Bidirectional Encoder Representations from Transformers
BiGRU:	Bidirectional Gated Recurrent Unit
BiLSTM:	Bidirectional Long Short-Term Memory
BLAST:	Basic Local Alignment Search Tool
CLI:	Command Line Interface
CNN:	Convolutional Neural Network
DPC:	Di-peptide Composition
DL:	Deep Learning
DSSP:	Dictionary of Secondary Structure in Proteins
DT:	Decision Tree
ESM:	Evolutionary Scale Modelling
ET:	Extra Trees
FDA:	Food and Drug Administration
KNN:	K-nearest Neighbor
mABs:	Mono-clonal Antibodies
MCC:	Matthews Correlation Coefficient
MERCI:	Motif-EmeRging and with Classes-Identification
ML:	Machine Learning
MLP:	Multi Layer Perceptron
NCBI:	National Center for Biotechnology Information
PCP:	Physicochemical Properties
PDB:	Protein Data Bank
PLM:	Protein Language Model
PSSM:	Position-Specific Scoring Matrices
PSI-BLAST:	Position-Specific Iterative BLAST
RF:	Random Forest
ResCNN:	Residual CNN
ROC:	Receiver Operating Characterstic Curve
RNN:	Recurrent Neural Network
RSA:	Relative Solvent Accessibility
SVC:	Support Vector Classifier
WHO:	World Health Organisation
XGB:	Extreme Gradient Boosting

Chapter 1

Introduction

1.1 Background

Cancer is a disease characterized by uncontrolled growth of cells that invade surrounding tissues and may metastasize to healthy organs [1]. They pose serious global threat in healthcare as it can arise in any part of the body and are of different types, each with unique genetic, molecular and pathological features. As per the World Health Organization (WHO) [2], approximately 20 million new cancer cases and 9.7 million cancer-related deaths were reported globally in 2022. Several factors such as environmental conditions and sedentary lifestyle also contribute to increased incidences of different types of cancers. Presently available treatment strategies like chemotherapy, immunotherapy and radiotherapy have developed a lot with time and have been successful in early detection and remission but they come with their own set of challenges. High recurrence rates, lack of specificity to cancer-specific sites and severe side effects are some of the major drawbacks of conventional therapeutic strategies.

To overcome these challenges, protein- and peptide-based therapeutics are becoming effective alternatives in cancer treatment. Anticancer peptides (ACPs) and proteins are two structurally but functionally complementary categories of biomolecules. These molecules not only act as direct anticancer agents but also play important role in development of targeted therapeutics. ACPs are typically short length peptides with length in the range of 5-50 amino acids, mostly rich in cationic and amphipathic residues like lysine, arginine etc. [3]. Cancer cells undergo different types of modifications to resist immune system response of host organism. They are more negatively charged as compared to the membranes of normal cells due to the presence of more negatively charged lipids like phosphatidyl serine on the outer leaflet of the membrane [4]. The positive charge of these peptides forms strong electrostatic interactions with the negatively charged components of plasma membrane resulting in membrane disruption via different mechanisms like pore formation, membrane destabilization or intracellular targeting of organelles like mitochondria and the nucleus. These ACPs also have immunomodulatory and antiangiogenic properties, thus contributing to both direct and indirect anticancer effects. However, their lesser half-life and rapid degradation in-vivo limited their potential. In contrast, Anticancer Proteins are longer length, functionally diverse biomolecules, with promising anticancer activities [5]. These include cytokines, tumor necrosis factors, mABs and enzymes. They act through various biological mechanisms like immune modulation, receptor targeting, and apoptosis induction. Due to their complex structures, they are able to interact specifically with target sites, resulting in target-specific response.

In this study, we focus on anticancer proteins ranging from 50 to 1000 amino

acids, curated from the CancerPPD2 [6] database to capture full-length, biologically active domains. The present study focuses primarily on anticancer proteins longer than 50 amino acids, which are structurally more complex and functionally diverse than short ACPs. Unlike smaller peptides, anticancer proteins may retain stable tertiary or quaternary structures, allowing them to interact with specific cell-surface receptors or signaling molecules involved in tumor progression. Mechanism of action of Anticancer Proteins is depicted in Figure 1.1. The potential of such proteins is reflected in the growing number of FDA-approved biologics for cancer therapy, including mABs like trastuzumab, immunocytokines, and enzyme-based therapies. These approvals highlight the clinical relevance and therapeutic promise of protein-based agents, particularly in accessing targets and pathways inaccessible to small molecules.

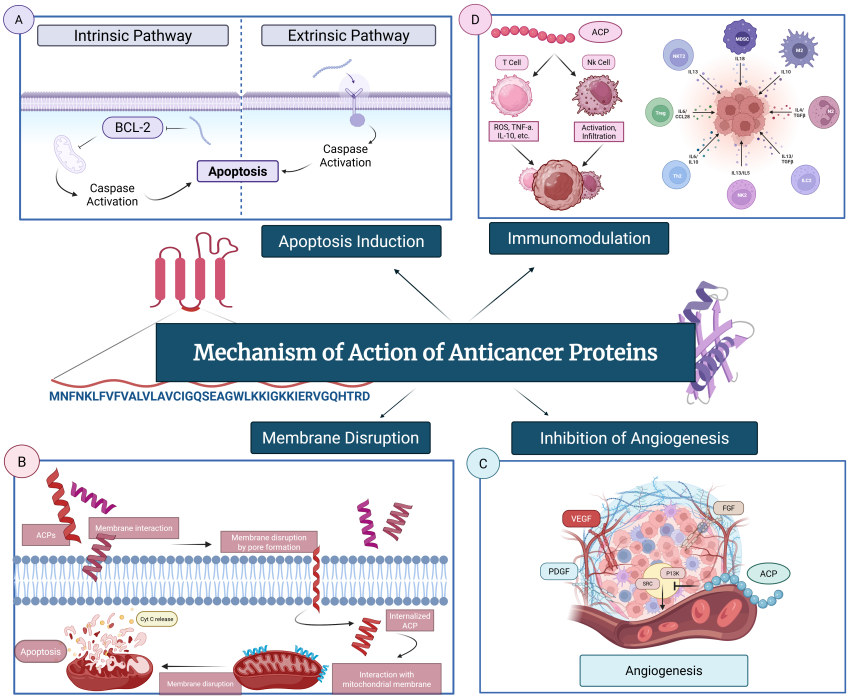


Fig. 1.1. *Mechanism of Anticancer Proteins*

The growing interest in Anticancer peptides (ACPs) and proteins as therapeutic agents has been significantly supported by advances in computational biology, bioinformatics, and high-throughput data mining. The traditional pipeline of drug discovery and validation is very time-consuming and resource intensive. In-silico tools and publicly available databases have accelerated this process of identification and design of peptides and protein-based therapeutics by correctly annotating and predicting potential candidates, thus reducing the burden on laboratory resources [7]. Databases like Uniprot [8], PDB [9] and NCBI [10] provide vast resources on biological data, but they sometimes lack domain-specific annotations. Several attempts have been made to combine data from all these primary databases and literature to provide high quality manually annotated information on specialized proteins and peptides with their domain specific role.

For eg. CancerPPD2 [6] is an updated repository containing information on experimentally validated Anticancer proteins and peptides. Data was mined from these biological databases like Uniprot [8] and PDB [9] as well as from literature sources like Pubmed and Patent Lens. These specialized databases provide annotated information on sequences, structure, chemical modifications and experimental activity of peptides helping in data-driven identification of novel peptides and proteins using in-silico tools like. These computational resources can complement the traditional Drug discovery pipeline and accelerate the development of protein- and peptide- based therapeutics.

1.2 Motivation of Work

In the era of precision medicine, there is growing emphasis on developing targeted therapies with minimal off-target effects. Among emerging biologics, therapeutic proteins have demonstrated considerable potential in treating complex diseases, including various forms of cancer. Despite this, the landscape of computational prediction tools in oncology remains heavily skewed toward short anticancer peptides (ACPs), with relatively little focus on full-length proteins that may show anticancer activity. This lack of dedicated tools for protein-level anticancer prediction restricts our ability to mine large proteomic datasets for potential therapeutic candidates. To address this gap, our study proposes a binary classification framework — **AntiCP3** — specifically designed to predict anticancer proteins by utilizing a curated dataset of experimentally validated anticancer protein sequences. The model incorporates evolutionary information and learned representations from protein language models, which capture the biological context of sequences more effectively than traditional feature engineering approaches.

1.3 Objective

The aim of this study is to develop an efficient computational framework, AntiCP3, for predicting Anticancer proteins using only their amino acid sequences, in order to support the discovery and development of novel protein-based therapeutics. To achieve this, different AI- based strategies were explored and training models was done on a high-quality, manually curated dataset, and the best-performing model after evaluating through rigourous evaluation metrics.

To facilitate accessibility and promote further research within the scientific community, AntiCP3 has been made publicly available as a web server (https://webs.iiitd.edu.in/raghava/anticp3/), as a standalone software package, and is also deployed on Hugging Face (https://huggingface.co/raghavagps-group/anticp3) and GitHub (https://github.com/raghavagps/anticp3) for broader usability and integration. Figure 1.2 shows the overall architecture of this study.

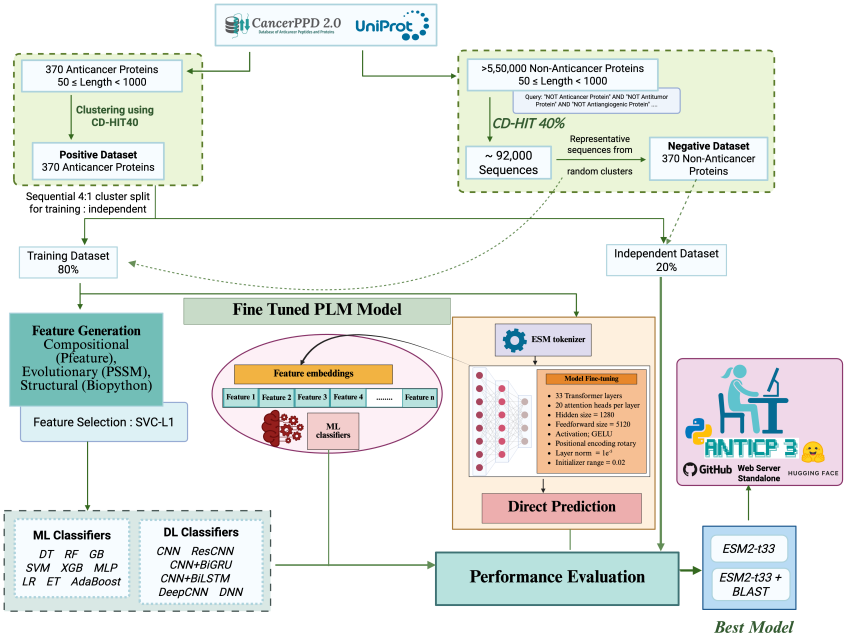


Fig. 1.2. *Overall architecture of AntiCP3*

Chapter 2

Literature Review

With the growing interest in protein- and peptide- based therapeutics, several attempts have been made in development of tools for prediction of Anticancer proteins and peptides, with most early efforts focussing on short peptide sequences due to their relatively simple structure and ease of synthesis. One of the first major contributions in this area was AntiCP [11], developed in 2013 using a dataset of 225 experimentally validated ACPs, all with lengths below 50 amino acids. Subsequently, various methods were developed such as Hajisharifi et al. [12], mACPred2.0 [13], ACPred-BMF [14], AntiCP2.0 [15], ACPML [16], MLACP-2.0 [17]. Table 2.1 summarizes the list of available Anticancer peptide prediction methods. While these tools achieved good performance integrating various sequence-based features such as amino acid composition, physicochemical properties, and motif information, their scope was limited to short peptides and did not extend to full-length protein sequences. Although tools such as CancerGram [18] and mACPred 2.0 [13] appear to support protein inputs via their web interfaces, a closer examination reveals that their underlying models were trained exclusively on peptide datasets. When evaluated on the curated validation dataset used in AntiCP3—comprising full-length, experimentally validated Anticancer proteins—CancerGram [18] exhibited poor generalization, misclassifying the majority of sequences and achieving an AUC close to random (0.52). While mACPred 2.0 [13] showed moderate improvement (AUC 0.66), its performance still remained limited when applied beyond its original peptide-focused training scope. In the present study, we have focussed on the prediction and classification of Anticancer proteins as proteins, unlike peptides, often function through complex domain interactions, structural motifs, and evolutionary conservation that are not fully captured by peptide-based predictors. Till Now, no such method has been developed that specifically focusses on longer length proteins. This is first such attempt where longer protein sequences were used for training the model to differentiate between Anticancer and Non-Anticancer Proteins.

S.No.	Method	Year of Publication	Peptide Length	Webserver	Github
1	AntiCP [11]	2013	≤50	Yes	NA
2	AntiCP 2.0 [15]	2021	4–50	Yes	Yes
3	ACP-DL [19]	2019	Not mentioned	NA	Yes
4	ACPP [20]	2015	15–100	Yes (Not Working)	Yes
5	ENNAACT [21]	2020	7–40	Yes	NA
6	DeepACP [22]	2020	Not mentioned	NA	Yes
7	CancerGram [18]	2020	≤50	Yes	Yes
8	AACFlow [23]	2024	5–50	NA	Yes
9	AntiMF [24]	2022	10–50	NA	NA
10	ACP-PDAFF [25]	2024	Not mentioned	NA	Yes
11	DeepACPpred [26]	2020	NA	NA	Yes
12	StackACPred [27]	2022	5–50	NA	Yes
13	mACPpred 2.0 [13]	2024	5–50	Yes	NA
14	mACPpred [28]	2019	5–52	Yes	NA
15	ME-ACP [29]	2022	Not mentioned	NA	Yes
16	ACP-MLC [30]	2023	Not mentioned	NA	Yes
17	ACP-BC [31]	2023	Not mentioned	NA	Yes
18	xDEEP-AcPEP [32]	2021	10–38	Yes	Yes
19	ACP-Check [33]	2022	Not mentioned	Yes (Not Working)	NA
20	MLACP 2.0 [17]	2022	50	Yes	NA
21	ACPPFel [34]	2024	Not mentioned	Yes (Not Working)	NA
22	ACP-ML [16]	2024	Not mentioned	NA	Yes
23	MA-PEP [35]	2024	Not mentioned	NA	Yes
24	ACP-DRL [36]	2024	Not mentioned	NA	Yes
25	Li & Wang's Method [37]	2020	Not mentioned	NA	NA
26	Hajishrafi et al. [12]	2014	Not mentioned	NA	NA
27	PEPred-Suite [38]	2019	Not mentioned	Yes (Not Working)	NA
28	ACPred-Fuse [39]	2019	Not mentioned	Yes (Not Working)	NA
29	PTPD [40]	2019	Not mentioned	NA	NA
30	ACPred [41]	2019	Not mentioned	Yes	Yes
31	MLACP [42]	2017	Not mentioned	Yes (Not Working)	NA
32	iACP [43]	2016	Not mentioned	Yes (Not Working)	NA
33	ACPred-FL [44]	2018	10–50	Yes (Not Working)	NA
34	Zhao et al. [45]	2021	5–70	NA	NA
35	ACP-GBDT [46]	2023	Not mentioned	NA	Yes

Table 2.1. *List of available Anticancer Peptide prediction methods*

Chapter 3

Materials and Methodology

The datasets were mined and manually curated from publicly available repositories. All the models were trained, tested, and evaluated on this manually curated non-redundant dataset.

3.1 Dataset Collection & Compilation

Positive Data representing Anticancer Proteins were mined from CancerPPD2 [6] - by applying a length filter to retain sequences between 50 and 1000 amino acids, a total of 370 natural protein sequences were selected. To compile the **Negative dataset**, the UniProt database [8] was queried using the exclusion keywords "NOT - Anticancer protein," AND "NOT - Antitumor protein," AND "NOT - Antiangiogenic protein," along with the filter length: 50 TO 1000. This search yielded 558,717 protein sequences that do not exhibit known anticancer activity.

3.2 Data Preprocessing

Data preprocessing is a preliminary and important step in any machine learning task. Here, we used several preprocessing techniques to clean the datasets and ensure its suitability for model training. Both the positive and negative datasets were subjected to CD-HIT clustering with a 40% sequence identity threshold (CD-HIT40) to remove highly similar sequences, as recommended in earlier studies like Li, W. et al. [47], Khanduja et al. [48] and Sangaraju et al. [13].

For the **positive dataset**, after applying CD-HIT40 we obtained 205 clusters where sequences have high intra-cluster similarity and low inter-cluster similarity. To prevent data leakage during training we ensured that all sequences originating from the same CD-HIT cluster were kept within the same fold during cross-validation. This approach avoids the artificial inflation of model performance and promotes generalizability. Figure 3.1 illustrates the manual five-fold splitting strategy employed for the positive dataset.

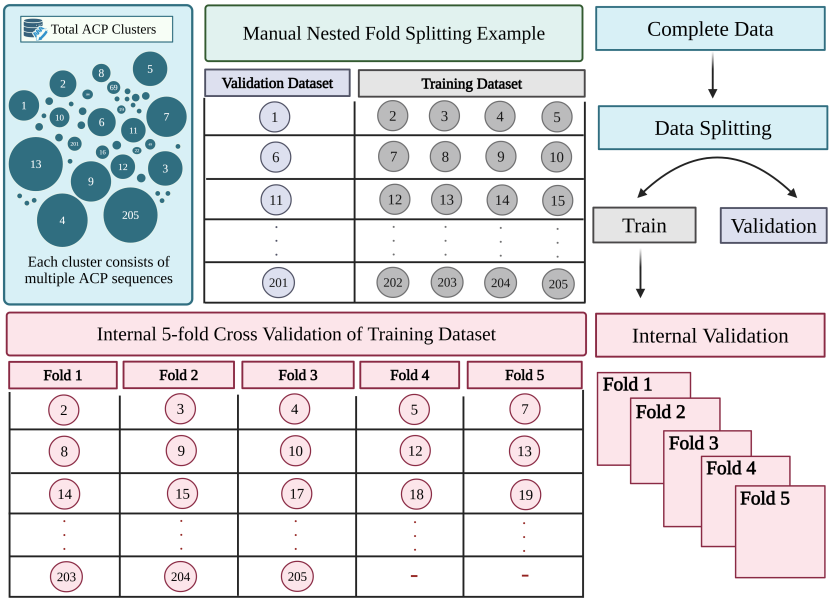


Fig. 3.1. *Manual Five-fold Splitting of Positive Dataset*

For the **negative dataset**, CD-HIT40 clustering reduced the dataset from 558,717 to 92,388 representative sequences. Since, these 92,388 sequences in the negative set were all non-redundant sequences we randomly selected 370 sequences with the same length distribution as positive dataset. Following redundancy removal, both datasets were screened to exclude sequences containing non-natural amino acids.

3.3 Feature Generation

Accurate classification of proteins relies heavily on extracting informative features that reflect key structural, physicochemical, and compositional properties of these biological sequences. In this study, we considered various categories of features to provide a comprehensive representation of the proteins:

- Composition-based Features
- 3D Structure-based Features
- Evolutionary profile-based Features
- Anticancer Peptide Mapping Features
- Sequence Embeddings from PLMs

3.3.1 Composition-based Features

Composition-based features provide a fundamental representation of protein sequences by quantifying the presence and arrangement of amino acids and their properties. To extract these features, the Pfeature tool [49] was used. It

offers a variety of features for peptide and protein analysis. Here, we specifically focussed on three key features – Amino Acid Composition (AAC), Di-Peptide Composition (DPC) and Physicochemical Properties (PCP). To calculate AAC (3.1) , DPC (3.2) , and PCP (3.3) features, the following formulas were used:

$$\text{AAC}_i = \frac{R_i}{L} \tag{3.1}$$

where R_i is the number of residues of amino acid type i, and L is the total length of the sequence.

$$\text{DPC}_i^j = \frac{D_i^j}{L} \tag{3.2}$$

where D_i^j represents the count of dipeptide i-j in the sequence and L is the total length of the sequence.

$$\text{PCP}_i = \frac{P_i}{L} \tag{3.3}$$

where P_i is the cumulative value of a given physicochemical property i across all residues in the sequence and L is the total length of the sequence.

By combining these compositional features, the model gains a better understanding of protein characteristics, allowing better predictions.

3.3.2 3D Structure-based Features

To calculate structure-based features the 3D structures of these proteins were required. However, experimental 3D structures for all the anticancer proteins included in this study were not available. As an alternative, we used the 3D structures predicted by AlphaFold [50]. It is a state-of-the-art AI-based tool used for predicting 3D structures of proteins. These predicted PDB structures were then used to calculate two types of structure-based features. We used biopython's DSSP module for structure-based features extraction (Miller et al. [51], Rost & Sandler [52], Tien at al. [53]). This module is designed to extract Secondary structure state and Relative solvent accessibility (RSA) against each residue in a protein sequence.

1. **Secondary Structure States**

   The DSSP module assigns each residue in the protein sequence one of these secondary states:

   - H: Alpha helix (4-12)
   - B: Isolated beta-bridge residue
   - E: Strand
   - G: 3-10 helix
   - I: Pi helix
   - T: Turn

- S: Bend
- - : None
- P: PolyProline Helix II [54]

We then calculated the percentage of residues in each secondary state in individual protein sequence and used these percentages as secondary state features in N ×9 vector, where N refers to protein sequence. Figure 3.2 illustrates an example matrix of secondary structure state features.

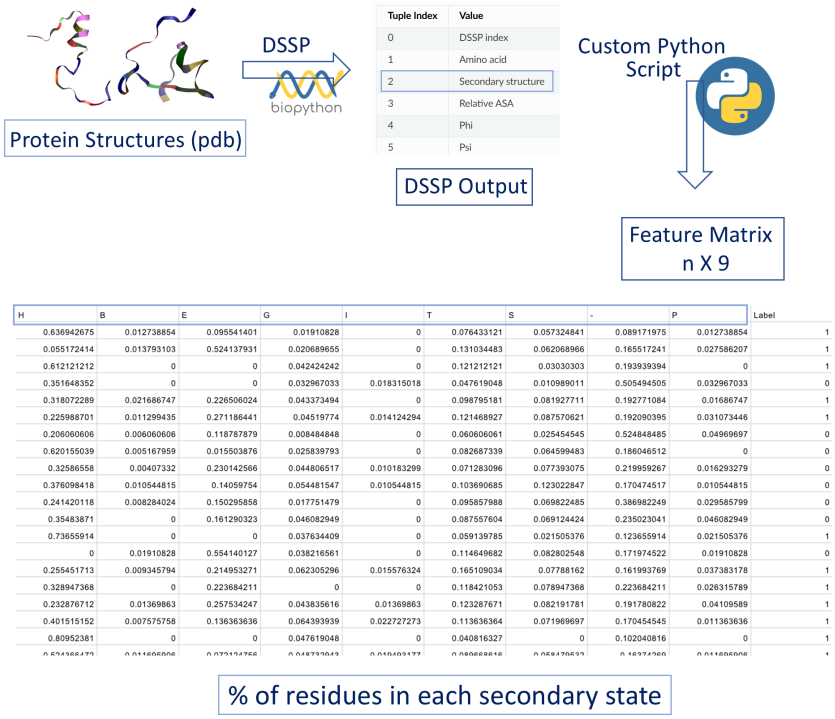


Fig. 3.2. *Generation of Secondary state-based feature matrix from PDB structures*

2. **Relative Surface Area (RSA)**

RSA assigns individual residue a unique score based on their extent of burial or exposure of that residue in 3D structure [53]. These values are then pre-processed and a threshold is set to categorize each residue in the sequence into three categories -

- Buried ($0 \leq \text{RSA} \leq 0.1$)
- Partially Buried ($0.1 < \text{RSA} < 0.3$)
- Exposed ($\text{RSA} > 0.3$)

Using this approach, we finally created a feature vector of size N ×60. Figure 3.3 shows an example RSA feature matrix.

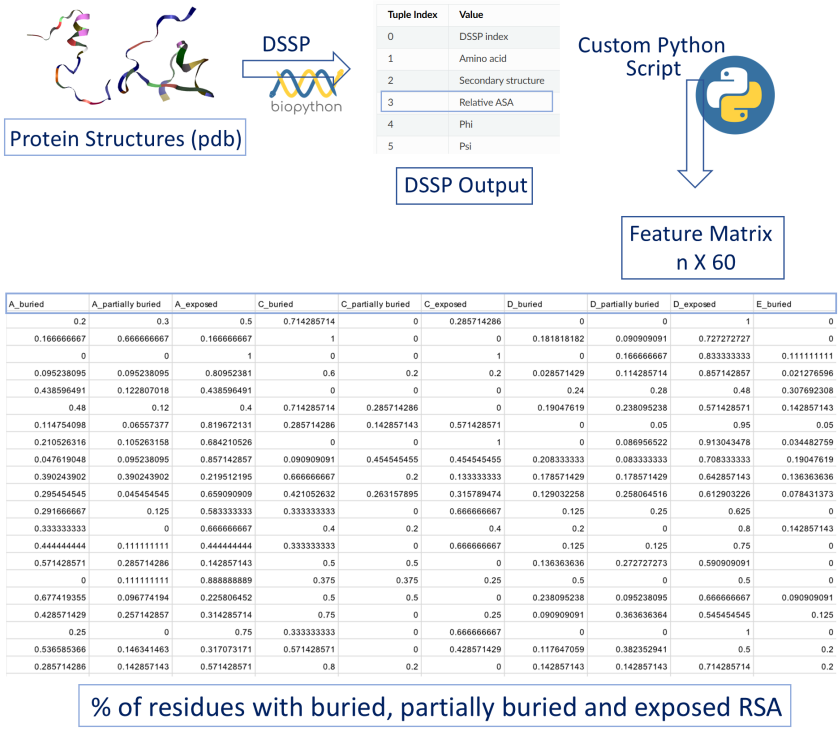


Fig. 3.3. *Generation of RSA-based feature matrix from PDB structures*

3.3.3 Evolutionary Profile-based features

Evolutionary information provides information on how protein sequences have maintained their structure and function across evolutionary timescales. To capture this information, we extracted features from **Position-Specific Scoring Matrices (PSSMs)**, which represent residue-level substitution probabilities derived from multiple sequence alignments. PSSMs are created using **Position-Specific Iterated BLAST (PSI-BLAST)** [55] by aligning a query sequence against comprehensive protein databases like Swissprot. These matrices reflect the likelihood of each amino acid being substituted at every position, thereby highlighting conserved regions that are often biologically important and functionally relevant.

Compared to sequence composition features (e.g., AAC, DPC, PCP) and structure-based features, evolutionary descriptors offer a deeper layer of functional insight. While composition-based features capture global residue distributions and structure-based features encode spatial conformation, PSSMs incorporate evolutionary constraints—enabling the identification of conserved motifs that are essential for biological activity, particularly relevant in the context of Anticancer protein classification.

1. **PSSM Composition-based Features**

   To transform PSSMs into fixed-length representations suitable for classical machine learning models, we computed the PSSM-400 descriptor using the pssm_composition module from the POSSUM package [56]. This

descriptor results in a 20 ×20 matrix that captures the average substitution scores between all pairs of amino acids across the entire protein sequence. Flattened into a 400-dimensional vector, this representation efficiently summarizes the evolutionary composition of the protein while preserving informative patterns relevant for classification tasks.

2. **Raw PSSM Profiles for Deep Learning Models**

   While composition-based features are concise, they abstract away position-specific signals that are crucial for capturing sequential dependencies. To preserve this positional information, we also utilized the raw PSSM matrices as input for deep learning models, which have been successfully used in other studies [57]. Each raw PSSM is a matrix of size L ×20, where L denotes the length of the protein sequence. This format retains the substitution score of every amino acid at every position, offering a rich, sequentially-aware feature space. These matrices can be used in deep learning methods like CNNs and RNNs as these networks can learn spatial, temporal, and evolutionary motifs associated with the Anticancer activity of proteins. Figure 3.4 shows the workflow for extracting and utilizing raw PSSM profiles in neural networks.

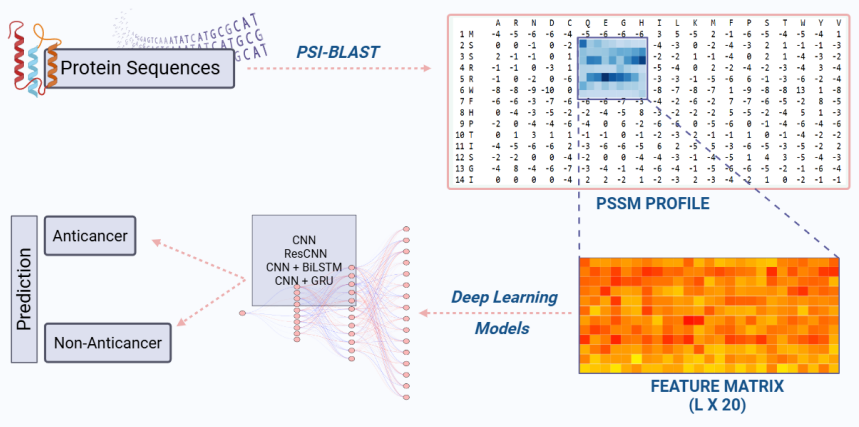


Fig. 3.4. *Work flow of generating PSSM profiles and using them in DL-methods*

3.3.4 Anticancer Peptide Mapping Features

This approach was based on the assumption that certain localized regions within a protein sequence may exhibit anticancer activity. We assumed that if a complete protein sequence has more number of smaller anticancer peptide fragments, the protein as a whole may also exhibit greater anticancer ability. AntiCP2 [15] is an ET-based classifier that predicts anticancer peptides with a very high accuracy. We used the protein scan module from the AntiCP2 method. We extracted overlapping 10-mer peptides from each protein sequence and evaluated their anticancer potential using prediction thresholds of 0.15, 0.30, 0.45, 0.60, 0.75, and 0.90. The number of peptides predicted to be anticancer at each threshold was then used as a quantitative feature. In addition to

these absolute counts, we also used the length of the full protein sequence as a feature such that model learns the correlation between the length and number of peptides and proteins with longer length do not overpower our dataset. We also calculated normalized counts by dividing the number of predicted anti-cancer peptides at each threshold by the total length of the protein. Together, these features capture the distribution and density of localized anticancer-like patterns within a protein sequence. An example matrix illustrating this feature extraction methodology is shown in Figure 3.5

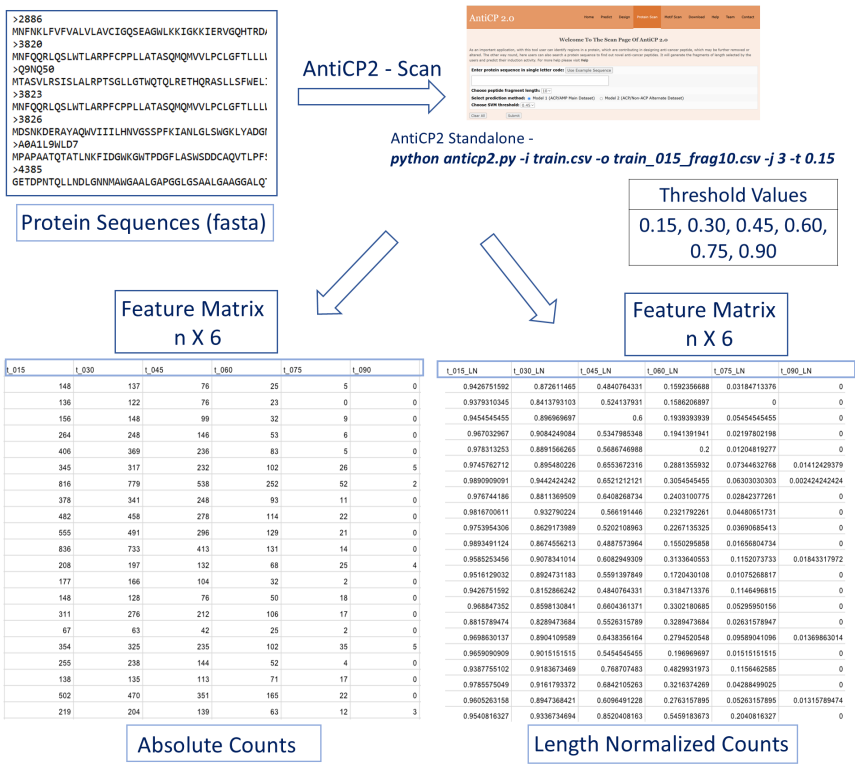


Fig. 3.5. *Generation of Peptide Mapping-based features using AntiCP2*

3.3.5 Sequence Embeddings from PLMs

Recent advances in Protein Language Models (PLMs) have revolutionized the way protein sequences are represented and analysed. Inspired by methodologies from natural language processing, these models treat amino acid sequences as biological "sentences," learning to capture intricate patterns of residue co-occurrence, conserved motifs, and functional relationships—without requiring multiple sequence alignments or structural annotations. Among the state-of-the-art PLMs, **ESM-2 (Evolutionary Scale Modelling 2)** [58], developed by **Meta AI**, stands out due to its strong performance in a variety of protein classification and function prediction tasks. ESM-2 is a transformer-based architecture trained on hundreds of millions of protein sequences, enabling it to implicitly encode both evolutionary and structural characteristics.

Here, we used a **33-layer ESM-2 model** fine-tuned on our dataset to extract

high-dimensional embeddings for each protein sequence. Specifically, the model generates a 1280-dimensional embedding vector that encapsulates the global contextual information inherent to the sequence. These embeddings provide a rich, alignment-free feature representation, offering significant advantages over traditional handcrafted features by capturing latent sequence semantics learned from large-scale datasets. By incorporating ESM-2 embeddings into our feature set, we augment the model's capacity to recognize subtle functional signals and identify Anticancer proteins with improved sensitivity and generalizability.

3.4 Alignment based Approach

3.4.1 Similarity Search Using BLAST

BLAST is a widely used tool in bioinformatics for identifying sequence similarity across protein and nucleotide databases. Based on the premise that homologous protein sequences also share functional similarity [59], BLAST uses local similarity searches to align biological sequences giving information about their evolutionary relationship. In this study, we employed protein–protein BLAST (BLASTp) version 2.2.29+ to classify input sequences based on their similarity to known anticancer and non-anticancer proteins.

The query sequences were aligned against our curated reference database containing both positive (anticancer) and negative (non-anticancer) protein sequences. The query was labeled as Anticancer if the hit is against positive sequences and non-anticancer if the match was found against negative sequences. To increase sensitivity, the hit was done against various E-value thresholds. These methods are easy to use and interpretable but often fail in cases where evolutionary information about protein is not available like that in the case of novel proteins. Therefore, in our framework, BLAST - based predictions were used alongside ML and DL approaches to enhance their predictive powers.

3.4.2 Motif based Approach

To identify conserved sequence patterns characteristic of anticancer proteins, we used the **Motif-EmeRging and with Classes-Identification (MERCI)** tool. It is a pattern discovery algorithm designed to extract class-specific motifs from biological sequences. In this study, we applied it to detect statistically significant and recurring motifs present uniquely in anticancer proteins, which may be indicative of their functional or structural roles.

The analysis was performed using the default parameters of MERCI, following established protocols [60]. The identified motifs provide interpretable biological signatures and serve as valuable complementary features, potentially capturing sequence-level determinants of anticancer activity that are not readily extracted through alignment or embedding-based methods.

3.5 Alignment Free Approach

3.5.1 Machine Learning based classification

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 Anticancer proteins. 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 neighbour (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.

3.5.2 Feature Selection Techniques

To reduce the dimensionality of the features and minimizing the noise, as not all features contribute significantly to the classification task, feature selection is required sometimes in ML related tasks to improve the model performance. We have used the Support Vector Classifier (SVC)-L1-based feature selection from the Scikit-learn library to subset the relevant features. SVC-L1 uses a linear kernel SVC combined with L1 regularization. It prevents the overfitting of the model as it penalizes coefficients with high values and shrinks the coefficients of less important features towards zero. As a result, we were able to retain only the most significant features. Through this method, we explored multiple feature combinations and identified those that contributed most to the model's predictive power.

3.5.3 Deep Learning-Based Classification

To classify Anticancer proteins using raw evolutionary profiles, we implemented the following DL networks -

1. **Multilayer Perceptron (MLP):** A simple DL network which consists of multiple fully connected layers, which are able to learn non-linear mappings between the input and output layers.

2. **Convolutional Neural Network (1D-CNN):** It uses convolutional layers to learn local patterns in the protein sequences.

3. **Deeper CNN:** Building on the 1D-CNN, additional layers were introduced to capture more abstract, hierarchical features in the data.

4. **Residual CNN (ResCNN):** By adding residual connections, we tried to prevent the vanishing gradient problem and allow for deeper networks to be trained more effectively.

5. **CNN + BiGRU (Bidirectional Gated Recurrent Unit):** Hybrid model that combines CNN layer for local feature extraction with BiGRU, which learns dependencies in both forward and backward directions within the sequence.

6. **CNN + BiLSTM (Bidirectional Long Short-Term Memory):** Just like the CNN + BiGRU model, this hybrid uses BiLSTM layers which can capture long-range sequential dependencies, that are important in understanding the functional roles of distant amino acids in a protein sequence.

The models were trained using the Adam optimizer and a binary cross-entropy loss function and fine-tuned by adjusting hyperparameters like dropout rate, number of layers, and applying regularization techniques to prevent overfitting.

3.5.4 Protein Language Models

Protein language models trained on large-scale data with billions of parameters have pushed the boundaries of these predictive tasks. In our study we have used ESM-2 model [58] developed by MetaAI that is a BERT-based transformer trained used masked language modelling objectives. We have specifically used ESM-2_t33 model (esm2_t33_650M_UR50D) which has 33 layers and 650M parameters in order to balance between both accuracy and computational resources. It gives numerical embeddings that capture complex relationships and evolutionary patterns in protein sequences.

For this task, we fine-tuned the pre-trained ESM-2 t33 model using the *EsmForSequenceClassification* framework from Hugging Face's Transformers library, allowing to add additional classification layer on top of the original model for the task of binary classification. Table 3.1 shows the key hyperparameters used to configure the ESM-2 t33-650M_UR50D model.

During fine-tuning, the model adapts the general protein sequence features learned from the pre-trained ESM-2 model to the specific task of predicting Anticancer proteins. By training on a dataset of labeled anticancer and non-anticancer proteins, the model learns to tell the difference between the two types based on the evolutionary and structural features in the protein sequences. Figure 3.6 shows the fine-tuning workflow, where the pre-trained ESM-2 model is adapted using a binary classification head and trained on task-specific data for binary classification of proteins.

Hyperparameter	Description
Number of Transformer Layers	33 layers with 20 attention heads per layer; captures complex relationships across long sequences
Hidden size	1280; determines the size of the model's internal representations
Feedforward Dimension	5120; controls the intermediate layer size for complex sequence transformations
Activation Function	GELU (Gaussian Error Linear Unit); enables smooth gradients and stable training
Position Embeddings	Rotary position embeddings; efficiently encode relative amino acid positions
Layer Norm Epsilon	1e-5 (0.00001); provides numerical stability during training
Weight Initializer Range	0.02; defines the range for initializing model weights for better convergence

Table 3.1. *Key hyperparameters used to configure the ESM-2 t33-650M_UR50D model*

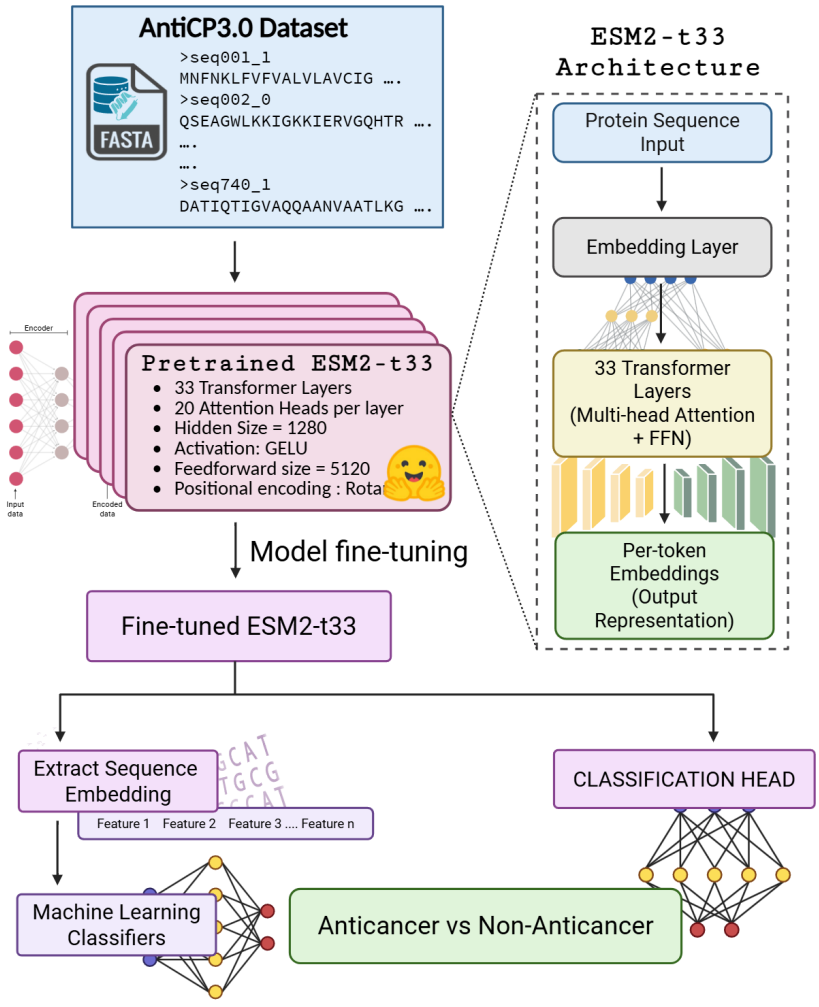


Fig. 3.6. *Workflow for fine-tuning the pre-trained ESM-2 model for Anticancer Protein classification.*

3.6 Cross-validation and performance evaluation

We have performed five-fold cross validation over training data to prevent underfitting and overfitting of the models. In the positive dataset, there were 205 clusters of 370 sequences, so we manually split the clusters into five folds so that each fold can have different set of sequences that are not related. Each cluster were assigned to five folds iteratively. Then the first four folds were placed in the training set and the fifth fold was placed in the validation set. The process was repeated till 306 out of the 370 positive protein sequences were placed in the training dataset, and remaining 64 in the validation dataset or the unseen dataset. This was done to ensure that the model was evaluated on sequences that were not closely related to those in the training set to minimize overfitting of the classifier. The process is clearly shown in Figure 3.1. For negative data, as the sequences retained after CD-HIT were non-redundant, they were randomly split into training and validation and a balanced positive and negative datasets were created as we have randomly selected 306 random sequences in training set and 64 sequences to outer validation set. A balanced distribution of both positive and negative samples was maintained across the training and validation subsets, as well as within each fold during cross-validation. Following evaluation metrices were used to evaluate the performance -

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

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

$$F1\text{-}Score = \frac{2 \times TP}{2 \times TP + FP + FN}$$

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

3.7 Hybrid Approach

To improve the predictive performance of the classifiers we have also developed a hybrid model that combines BLAST results with the predicted labels from the best-performing ML classifiers, so as to take the advantage of both sequence similarity-based BLAST approach and data-driven learning of ML classifiers.

BLAST was used to align the protein sequences against a database of anticancer and non-anticancer proteins and then the test sequences were classified on the basis of hits and non-hits. We used the E-value threshold of 10^{-20}, and

applied the scoring system to change the predicted probabilities, +0.5 was assigned for hits against anticancer proteins, -0.5 for hits against non-anticancer proteins 0 for sequences that had no significant hits. The results of ML classifiers were then used in the next step. The BLAST-based scores were added to the probability values given by ML classifiers and the final scores were obtained. As a result, the hybrid model achieves higher predictive accuracy as compared to using either approaches alone.

3.8 Compositional Analysis

The compositional analysis of the protein sequences was done using an independent t-test to compare the average amino acid compositions between three groups, namely- Anticancer proteins, non-anticancer proteins, and Anticancer peptides from the AntiCP-2 dataset, to see if there were statistically significant differences in the amino acid residue compositions across these groups. A p-value of < 0.05 was considered the threshold for statistical significance. The results showed significant differences in the amino acid compositions between the three groups, indicating that certain amino acid residues are more or less frequent in anticancer proteins compared to non-anticancer proteins and anticancer peptides.

Chapter 4

Results

4.1 Compositional Analysis

To identify the residues, frequent in anticancer proteins, average amino acid compositional analysis was done between Anticancer proteins, peptides and non-anticancer proteins. The independent t-test revealed that cysteine, glycine, leucine, tryptophan and phenylalanine were mostly present in Anticancer proteins and peptides. Methionine, glutamine and tyrosine were significantly abundant in only Anticancer proteins, whereas Lysine was frequent in Anticancer peptides only. On the other hand, residues such as Glutamate, Aspartic acid, Arginine, proline, serine, Threonine and Valine are more abundant in non-anticancer proteins. On the basis of this preliminary analysis, it can be deduced that Aromatic and non-polar aliphatic amino acids are mostly found in Anticancer proteins. Figure 4.1 shows the comparative amino acid composition in the three groups highlighting the residues with statistically significant differences.

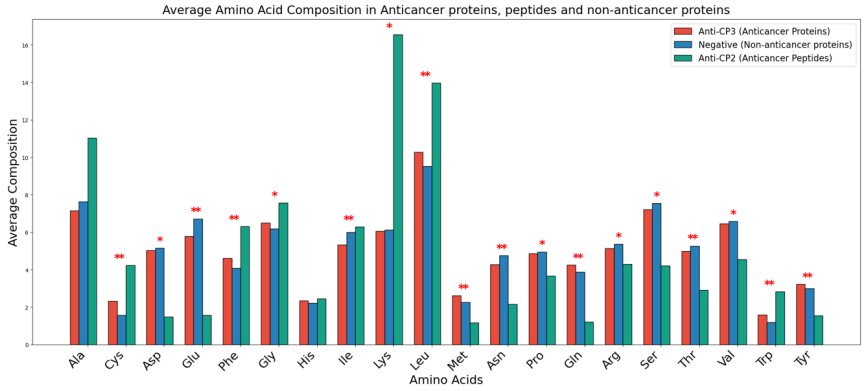


Fig. 4.1. *Average compositional analysis of amino acids in anticancer proteins, peptides and non-anticancer proteins.*

4.2 Alignment-based Analysis

4.2.1 BLAST-based Analysis

For BLAST-based analysis, we used the training dataset to construct a BLAST-formatted database. This database was compared to query sequences from the validation dataset to identify hits across different e-values, ranging from $1e^{-30}$ to $1e^{+10}$. Based on the top hits whether it was against positive or negative sequences, each query sequence was assigned a score as described in section 3.7 and the sequences were classified into positive and negative labels. The detailed BLAST hit results for the validation data are presented in Table 4.1 below.

e-value	Total Hits	ACP Hits	C-ACP Hits	Non-ACP Hits	C-Non-ACP Hits
1.00E+10	128	64	35	64	37
1.00E+05	128	64	35	64	37
1.00E+04	128	64	35	64	38
1.00E+03	128	64	35	64	37
1.00E+02	128	64	36	64	38
1.00E+01	128	64	35	64	37
1.00E+00	64	34	25	30	16
1.00E-01	39	27	21	12	7
1.00E-02	33	23	19	10	6
1.00E-03	31	21	17	10	6
1.00E-04	28	19	16	9	5
1.00E-05	27	18	15	9	5
1.00E-10	21	15	12	6	5
1.00E-20	14	10	9	4	3
1.00E-25	8	4	3	4	3
1.00E-30	5	3	2	2	1

Table 4.1. *The table describes the BLAST coverage over validation data.*

4.2.2 Motif-based Analysis

In this approach, we first identified motifs that were exclusive to anticancer (positive) and non-anticancer (negative) sequences within the training data using the MERCI program and then located in the validation dataset. If a positive motif was found in a sequence, a score of +0.5 was assigned to that sequence, while a negative motif resulted in a score of -0.5. Subsequently, both the motif scores and the predictions from the best-performing machine learning model were combined to compute various evaluation metrics as described in section 3.6. However, the motif-based approach did not give optimal performance in this dataset. Table 4.2 below shows the detailed results.

MERCI Parameters	Sensitivity	Specificity	Accuracy	AUC	Cohen Kappa	MCC
motif_k_10_C_NONE_g0_fp10	78.13	75.00	76.56	0.85	0.53	0.53
motif_k_20_C_NONE_g0_fp10	78.13	75.00	76.56	0.85	0.53	0.53
motif_k_50_C_NONE_g2_fp10	32.81	90.63	61.72	0.85	0.23	0.29
motif_k_50_C_NONE_g2_fp5	32.81	90.63	61.72	0.85	0.23	0.29
motif_k_50_C_NONE_g2_fp2	32.81	90.63	61.72	0.85	0.23	0.29
motif_k_100_C_NONE_g2_fp10	18.75	92.19	55.47	0.85	0.11	0.16
motif_k_100_C_NONE_g2_fp2	18.75	92.19	55.47	0.85	0.11	0.16
motif_k_100_C_NONE_g2_fp5	18.75	92.19	55.47	0.85	0.11	0.16
motif_k_50_C_NONE_g0_fp10	78.13	75.00	76.56	0.84	0.53	0.53
motif_k_100_C_NONE_g0_fp10	78.13	73.44	75.78	0.84	0.52	0.52
motif_k_100_C_NONE_g1_fp2	25.00	90.63	57.81	0.83	0.16	0.21
motif_k_100_C_NONE_g1_fp5	25.00	90.63	57.81	0.83	0.16	0.21
motif_k_50_C_NONE_g1_fp5	23.44	90.63	57.03	0.83	0.14	0.19
motif_k_50_C_NONE_g1_fp2	23.44	90.63	57.03	0.83	0.14	0.19
motif_k_10_C_NONE_g2_fp2	65.63	78.13	71.88	0.83	0.44	0.44
motif_k_10_C_NONE_g2_fp5	65.63	78.13	71.88	0.83	0.44	0.44
motif_k_10_C_NONE_g2_fp10	65.63	78.13	71.88	0.83	0.44	0.44
motif_k_10_C_NONE_g0_fp5	68.75	78.13	73.44	0.82	0.47	0.47
motif_k_10_C_NONE_g0_fp2	68.75	78.13	73.44	0.82	0.47	0.47
motif_k_10_C_NONE_g1_fp10	56.25	81.25	68.75	0.82	0.38	0.39
motif_k_10_C_NONE_g1_fp2	56.25	81.25	68.75	0.82	0.38	0.39
motif_k_10_C_NONE_g1_fp5	56.25	81.25	68.75	0.82	0.38	0.39
motif_k_100_C_NONE_g1_fp10	40.63	84.38	62.50	0.82	0.25	0.28
motif_k_20_C_NONE_g2_fp10	51.56	81.25	66.41	0.82	0.33	0.34
motif_k_20_C_NONE_g2_fp2	51.56	81.25	66.41	0.82	0.33	0.34
motif_k_20_C_NONE_g2_fp5	51.56	81.25	66.41	0.82	0.33	0.34
motif_k_100_C_NONE_g0_fp5	28.13	90.63	59.38	0.82	0.19	0.24
motif_k_100_C_NONE_g0_fp2	28.13	90.63	59.38	0.82	0.19	0.24
motif_k_50_C_NONE_g1_fp10	39.06	84.38	61.72	0.81	0.23	0.26
motif_k_20_C_NONE_g1_fp10	39.06	84.38	61.72	0.81	0.23	0.26
motif_k_20_C_NONE_g1_fp2	39.06	84.38	61.72	0.81	0.23	0.26
motif_k_20_C_NONE_g1_fp5	39.06	84.38	61.72	0.81	0.23	0.26
motif_k_20_C_NONE_g0_fp5	56.25	81.25	68.75	0.80	0.38	0.39
motif_k_20_C_NONE_g0_fp2	56.25	81.25	68.75	0.80	0.38	0.39
motif_k_50_C_NONE_g0_fp2	40.63	85.94	63.28	0.79	0.27	0.30
motif_k_50_C_NONE_g0_fp5	40.63	85.94	63.28	0.79	0.27	0.30

Table 4.2. *Performance of Motif based Ensemble model using default motifs*

4.3 Alignment Free Analysis

4.3.1 Composition-based Features

ML classifiers as described in section 3.5.1 were used for this task of predicting Anticancer proteins. We have used three composition-based features- AAC, DPC and PCP. Detailed performance metrics for each feature type are presented in Table 4.3 (AAC), Table 4.4 (DPC), and Table 4.5 (PCP). The AUROC plot of all classifiers on AAC, DPC and PCP features are given below in Figure 4.2.

Models	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
Decision Tree	52.97 ± 6.85	59.76 ± 9.77	56.36 ± 6.53	0.59 ± 0.08	0.13 ± 0.13	40.62	64.06	52.34	0.55	0.05
Random Forest	70.24 ± 7.49	66.67 ± 3.60	68.46 ± 4.91	0.74 ± 0.06	0.37 ± 0.10	67.19	54.69	60.94	0.68	0.22
Gradient Boosting	80.07±15.90	40.47±21.64	60.27 ± 5.06	0.71 ± 0.07	0.25 ± 0.08	100	6.25	53.12	0.68	0.18
AdaBoost	66.32±12.28	57.50 ± 6.41	61.91 ± 7.10	0.66 ± 0.08	0.24 ± 0.14	65.62	48.44	57.03	0.61	0.14
XGBoost	72.19±10.58	54.91 ± 8.10	63.55 ± 5.46	0.70 ± 0.08	0.28 ± 0.11	73.44	48.44	60.94	0.69	0.23
Extra Trees	71.54±10.09	62.08 ± 4.85	66.81 ± 6.98	0.75 ± 0.08	0.34 ± 0.14	71.88	62.50	67.19	0.69	0.35
Logistic Regression	66.67±11.38	57.82 ± 6.80	62.24 ± 5.97	0.68 ± 0.08	0.25 ± 0.12	54.69	57.81	56.25	0.60	0.13
KNN	73.20±10.09	55.54 ± 2.16	64.37 ± 5.37	0.70 ± 0.08	0.29 ± 0.11	67.19	53.12	60.16	0.64	0.21
SVC	77.76 ± 5.87	59.81 ± 5.49	68.79 ± 2.94	0.75 ± 0.05	0.38 ± 0.06	68.75	50.00	59.38	0.67	0.19
MLP	65.31±10.83	66.03 ± 3.85	65.67 ± 4.49	0.71 ± 0.04	0.32 ± 0.09	64.06	59.38	61.72	0.60	0.23

Table 4.3. *Performance of various ML classifiers on AAC- based features*

Models	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
Decision Tree	48.67 ± 7.23	62.45 ± 5.71	55.56 ± 3.30	0.56 ± 0.03	0.11 ± 0.07	53.12	57.81	55.47	0.55	0.11
Random Forest	70.90 ± 8.80	58.80 ± 3.37	64.85 ± 5.65	0.72 ± 0.08	0.30 ± 0.11	62.5	64.06	63.28	0.64	0.27
Gradient Boosting	70.59 ± 7.83	57.17 ± 6.79	63.88 ± 5.42	0.68 ± 0.08	0.28 ± 0.11	67.19	53.12	60.16	0.64	0.21
AdaBoost	62.40±10.73	61.09 ± 8.32	61.75 ± 8.07	0.65 ± 0.11	0.24 ± 0.16	57.81	50	53.91	0.58	0.08
XGBoost	71.88±10.78	52.92 ± 5.39	62.40 ± 6.59	0.68 ± 0.07	0.26 ± 0.14	59.38	57.81	58.59	0.63	0.17
Extra Trees	58.18 ± 9.95	59.80 ± 6.99	58.99 ± 8.33	0.65 ± 0.12	0.18 ± 0.17	57.81	60.94	59.38	0.63	0.19
Logistic Regression	58.47 ± 4.88	62.06 ± 9.33	60.27 ± 6.28	0.64 ± 0.06	0.21 ± 0.13	60.94	56.25	58.59	0.63	0.17
KNN	78.11±19.54	30.03 ± 8.23	54.07 ± 9.92	0.59 ± 0.13	0.11 ± 0.22	89.06	15.62	52.34	0.65	0.07
SVC	67.97±12.35	62.07±13.90	65.02 ± 8.84	0.71 ± 0.10	0.31 ± 0.18	50	62.5	56.25	0.68	0.13
MLP	64.05 ± 6.34	69.59 ± 6.50	66.82 ± 5.56	0.72 ± 0.07	0.34 ± 0.11	56.25	73.44	64.84	0.72	0.30

Table 4.4. *Performance of various ML classifiers on DPC- based features*

Models	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
Decision Tree	69.59 ± 9.00	45.70±11.59	57.65 ± 5.61	0.60 ± 0.08	0.16 ± 0.11	62.5	43.75	53.12	0.56	0.06
Random Forest	68.63 ± 6.57	64.70 ± 5.85	66.67 ± 3.93	0.72 ± 0.05	0.34 ± 0.08	65.62	64.06	64.84	0.68	0.30
Gradient Boosting	64.37 ± 7.16	69.92 ± 4.07	67.14 ± 4.66	0.71 ± 0.05	0.34 ± 0.09	43.75	73.44	58.59	0.68	0.18
AdaBoost	76.10±16.44	43.84±15.45	59.97 ± 3.01	0.66 ± 0.03	0.23 ± 0.08	40.62	75.00	57.81	0.66	0.17
XGBoost	69.24±11.67	52.93 ± 6.38	61.09 ± 6.64	0.67 ± 0.08	0.23 ± 0.14	67.19	65.62	66.41	0.70	0.33
Extra Trees	67.32±10.21	61.09 ± 8.59	64.20 ± 3.36	0.72 ± 0.04	0.29 ± 0.07	64.06	64.06	64.06	0.69	0.28
Logistic Regression	66.67±11.43	58.14 ± 5.51	62.41 ± 5.13	0.68 ± 0.07	0.25 ± 0.11	54.69	59.38	57.03	0.60	0.14
KNN	64.36 ± 8.20	55.24 ± 7.36	59.80 ± 6.28	0.65 ± 0.06	0.20 ± 0.13	48.44	70.31	59.38	0.62	0.19
SVC	72.89 ± 7.02	52.96 ± 7.47	62.92 ± 4.12	0.70 ± 0.03	0.27 ± 0.08	68.75	53.12	60.94	0.65	0.22
MLP	63.07 ± 2.91	55.88 ± 4.46	59.47 ± 2.49	0.65 ± 0.04	0.19 ± 0.05	60.94	64.06	62.50	0.66	0.25

Table 4.5. *Performance of various ML classifiers on PCP- based features*

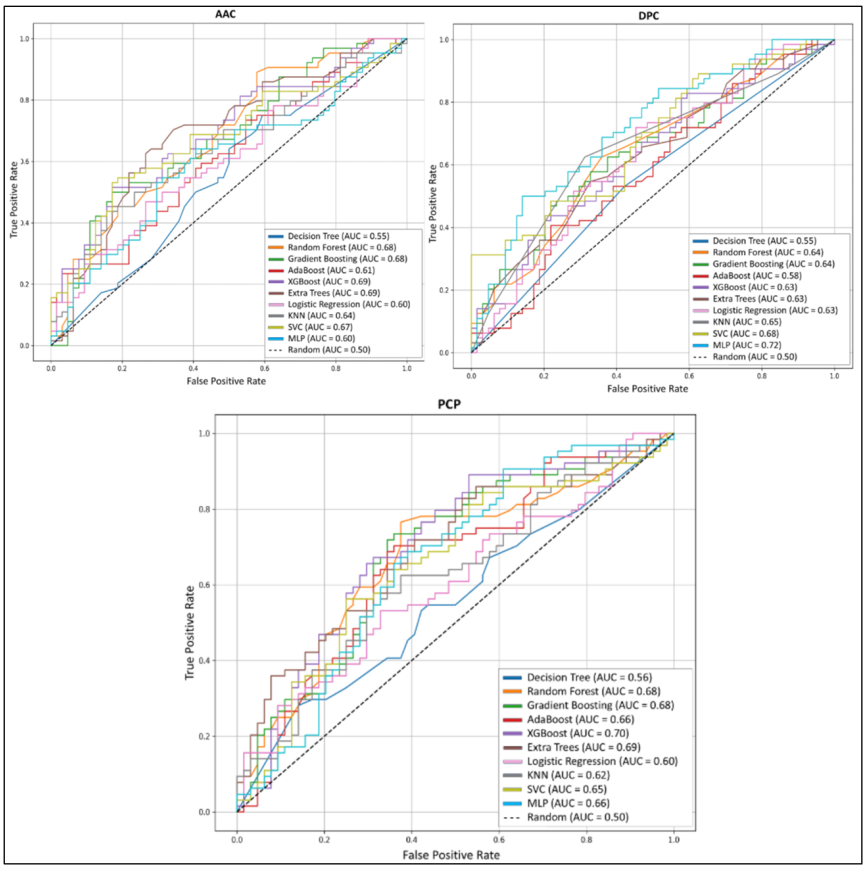


Fig. 4.2. *Receiver Operating Characteristic (ROC) Curve comparing the performance of different ML models for composition-based features on validation dataset*

4.3.2 3D Structure-based Features

To explore the structural characteristics of anticancer proteins, we extracted and pre-processed secondary structure-related features - Relative Solvent Accessibility (RSA) and Secondary Structure States (assigned using DSSP). We evaluated several machine learning classifiers on these features to assess their ability to differentiate anticancer proteins from non-anticancer ones. The Multilayer Perceptron (MLP) classifier achieved the highest performance on RSA features, with an AUROC of 0.67 and an MCC of 0.33 on the validation set. The Extra Trees (ET) classifier performed best on DSSP-based secondary structure features, achieving an AUROC of 0.61 on the validation set. These results suggest that secondary structure elements, particularly solvent accessibility, hold relevant information for classifying anticancer proteins, though their predictive power is moderate compared to other feature types. The AUROC curves for all classifiers trained on RSA and DSSP features are illustrated in Figure 4.3, and a detailed breakdown of performance metrics for RSA is shown in Table 4.6 and DSSP in Table 4.7.

Models	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
Decision Tree	48.96±10.21	60.79 ± 7.35	54.88 ± 5.91	0.55 ± 0.06	0.10 ± 0.12	46.88	68.75	57.81	0.58	0.16
Random Forest	72.49±13.09	53.29 ± 5.16	62.89 ± 6.41	0.67 ± 0.08	0.27 ± 0.14	75.00	54.69	64.84	0.74	0.30
Gradient Boosting	67.91±11.08	55.23 ± 5.53	61.57 ± 6.69	0.67 ± 0.08	0.24 ± 0.14	76.56	67.19	71.88	0.75	0.44
AdaBoost	58.79 ± 8.38	67.31 ± 7.50	63.05 ± 3.19	0.67 ± 0.06	0.26 ± 0.06	51.56	67.19	59.38	0.64	0.19
XGBoost	65.96±12.66	55.59 ± 8.47	60.77 ± 7.94	0.65 ± 0.11	0.22 ± 0.16	65.62	57.81	61.72	0.67	0.24
Extra Trees	67.29 ± 5.52	61.78 ± 6.96	64.53 ± 4.41	0.68 ± 0.06	0.29 ± 0.09	76.56	67.19	71.88	0.78	0.44
Logistic Regression	69.23±11.16	55.24 ± 4.45	62.23 ± 5.54	0.63 ± 0.07	0.25 ± 0.12	75.00	50.00	62.50	0.67	0.26
KNN	54.24 ± 6.84	61.13 ± 4.70	57.69 ± 3.55	0.60 ± 0.03	0.15 ± 0.07	68.75	54.69	61.72	0.62	0.24
SVC	63.71 ± 3.50	68.30 ± 6.00	66.00 ± 4.00	0.69 ± 0.06	0.32 ± 0.08	62.50	78.12	70.31	0.74	0.41
MLP	58.77±12.94	68.30 ± 2.49	63.54 ± 5.73	0.67 ± 0.04	0.27 ± 0.11	67.19	65.62	66.41	0.67	0.33

Table 4.6. *Performance of various ML classifiers on RSA- based features*

Models	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
Decision Tree	58.54 ± 8.47	52.61 ± 4.92	55.57 ± 4.65	0.59 ± 0.02	0.11 ± 0.09	67.19	54.69	60.94	0.6	0.22
Random Forest	58.17 ± 5.63	58.18 ± 4.16	58.17 ± 4.78	0.63 ± 0.05	0.16 ± 0.10	53.12	60.94	57.03	0.63	0.14
Gradient Boosting	51.65 ± 9.15	63.38 ± 4.23	57.51 ± 5.17	0.61 ± 0.05	0.15 ± 0.10	51.56	65.62	58.59	0.63	0.17
AdaBoost	45.89±23.93	69.20±14.22	57.54 ± 6.46	0.60 ± 0.06	0.15 ± 0.14	17.19	76.56	46.88	0.55	−0.08
XGBoost	64.02 ± 9.13	53.59 ± 8.25	58.80 ± 4.86	0.59 ± 0.04	0.18 ± 0.10	51.56	54.69	53.12	0.55	0.06
Extra Trees	58.18 ± 7.41	59.79 ± 1.87	58.98 ± 4.04	0.63 ± 0.05	0.18 ± 0.08	51.56	59.38	55.47	0.61	0.11
Logistic Regression	56.19 ± 6.44	61.11 ± 3.40	58.65 ± 3.72	0.60 ± 0.03	0.17 ± 0.07	45.31	59.38	52.34	0.57	0.05
KNN	58.52 ± 8.39	66.02 ± 2.32	62.27 ± 3.73	0.65 ± 0.03	0.25 ± 0.07	34.38	68.75	51.56	0.56	0.03
SVC	70.91 ± 8.06	59.50 ± 5.24	65.21 ± 4.38	0.68 ± 0.03	0.31 ± 0.09	62.5	54.69	58.59	0.62	0.17
MLP	61.13±15.05	55.18±14.19	58.16 ± 4.14	0.60 ± 0.04	0.17 ± 0.09	42.19	56.25	49.22	0.51	−0.02

Table 4.7. *Performance of various ML classifiers on DSSP- based features*

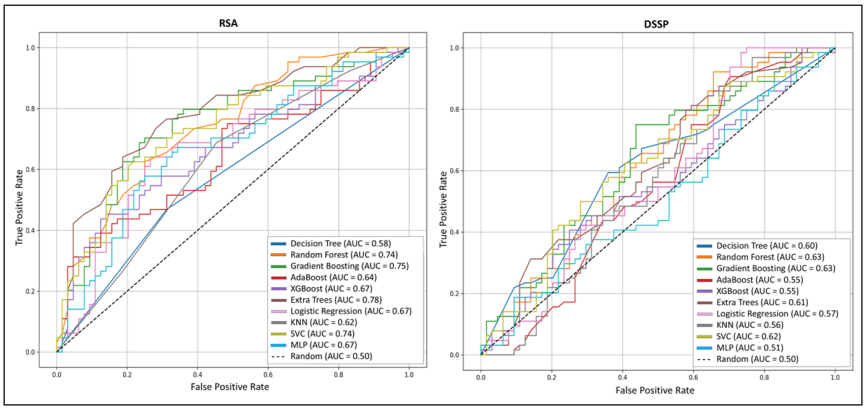


Fig. 4.3. *Receiver Operating Characteristic (ROC) Curve comparing the performance of different ML models for structure-based features on validation dataset*

4.3.3 Evolutionary profile-based Features

1. **PSSM composition-based Features**

   A 20 × 20-dimensional vector of PSSM composition profile was generated for each protein sequence. This vector summarizes the substitution probabilities of amino acids, thereby encoding evolutionary context. We trained multiple machine learning classifiers on these PSSM-based features. Multilayer Perceptron (MLP) classifier achieved the best performance, with an AUROC of 0.79 and a MCC of 0.44 on the validation set. These re-

sults indicate that evolutionary profile-based features are more effective in distinguishing Anticancer proteins compared to sequence- or structure-based features. The AUROC curves for all classifiers trained on PSSM composition features are presented in Figure 4.4. Detailed performance metrics for all classifiers is shown in Table 4.8.

Models	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
Decision Tree	58.47 ± 7.04	54.27 ± 5.81	56.37 ± 2.90	0.58 ± 0.03	0.13 ± 0.06	65.62	56.25	60.94	0.61	0.22
Random Forest	68.62 ± 3.99	67.64 ± 4.14	68.13 ± 3.84	0.76 ± 0.06	0.36 ± 0.08	51.56	78.12	64.84	0.69	0.31
Gradient Boosting	64.04 ± 5.05	72.23 ± 3.98	68.13 ± 3.27	0.76 ± 0.05	0.36 ± 0.06	43.75	81.25	62.50	0.67	0.27
AdaBoost	64.71 ± 4.46	71.25 ± 3.72	67.98 ± 3.27	0.74 ± 0.02	0.36 ± 0.07	43.75	79.69	61.72	0.67	0.25
XGBoost	69.93 ± 7.62	64.02 ± 9.47	66.98 ± 6.23	0.74 ± 0.08	0.34 ± 0.12	59.38	67.19	63.28	0.68	0.27
Extra Trees	73.50 ± 8.51	65.06 ± 4.24	69.28 ± 4.57	0.76 ± 0.05	0.39 ± 0.09	59.38	75.00	67.19	0.71	0.35
Logistic Regression	75.83 ± 3.57	68.97 ± 7.15	72.40 ± 4.29	0.77 ± 0.04	0.45 ± 0.08	62.50	85.94	74.22	0.79	0.50
KNN	62.72 ± 6.29	63.37 ± 4.85	63.05 ± 5.15	0.69 ± 0.07	0.26 ± 0.10	50.00	70.31	60.16	0.64	0.21
SVC	93.11 ± 13.77	14.75 ± 23.51	53.93 ± 5.25	0.63 ± 0.14	0.11 ± 0.13	98.44	7.81	53.12	0.61	0.15
MLP	82.01 ± 4.67	66.37 ± 6.83	74.19 ± 3.19	0.79 ± 0.03	0.49 ± 0.06	70.31	73.44	71.88	0.79	0.44

Table 4.8. *Performance of various ML classifiers on evolutionary profile- based features.*

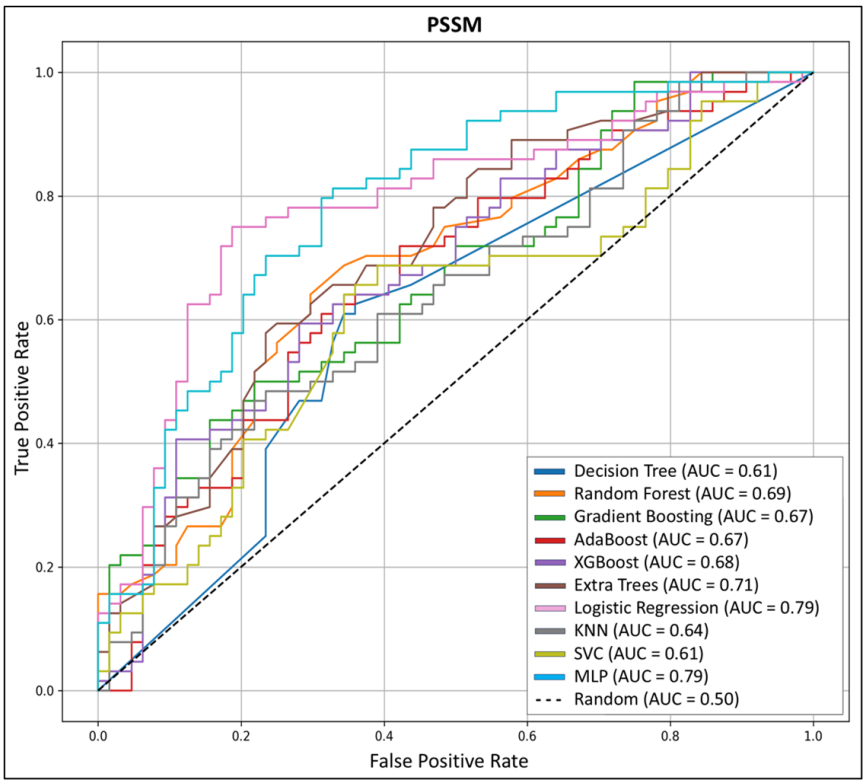


Fig. 4.4. *AUC-ROC plot of evolutionary profile (PSSM) based features on validation dataset.*

2. **Raw PSSM profile for Deep Learning based models** We trained several deep learning models directly on the raw PSSM profiles to capture more detailed local patterns and sequential dependencies present in the evolutionary profiles that are lost in constructing a vector representation

like in the case of composition based PSSM features. We used different neural networks to classify the protein sequences. Detailed results are shown in Table 4.9 below. These models were shown to have slightly decreased performance as compared to PSSM composition. These can be attributed to the limited size of the dataset, as neural networks are known to work better for large amounts of the data.

Model	Sens	Spec	Acc	AUC	MCC
MLP (2 Dense Layers, Dropout)	0.46	0.70	0.58	0.61	0.17
CNN (Conv1D + MaxPool)	0.59	0.75	0.67	0.70	0.34
Deeper CNN (3 Conv1D + MaxPool)	0.62	0.61	0.61	0.67	0.23
ResCNN (CNN with Residual Blocks)	0.59	0.75	0.67	0.74	0.34
CNN + BiGRU (1 Conv1D + BiGRU)	0.46	0.78	0.62	0.70	0.26
CNN + BiLSTM (1 Conv1D + BiLSTM)	0.54	0.75	0.65	0.72	0.30

Table 4.9. *Performance metrics of DL models*

4.3.4 SVC-L1-based Feature Selection

To reduce the dimensionality of the features and improve classification performance, we used the SVCL-1 method. From the PSSM composition-based features, SVC-L1 selected a total of 256 optimal features. We then trained various machine learning classifiers using this reduced set of 256 features. Among all classifiers, the Multilayer Perceptron (MLP) achieved the best performance, with an AUROC of 0.78 and MCC of 0.42 on the validation dataset. The performance of different classifiers on the selected PSSM features is presented in Table 4.10

Model	Sensitivity	Specificity	Accuracy	AUC	MCC
Decision Tree	32.81	73.44	53.12	0.58	0.07
Random Forest	57.81	67.19	62.50	0.70	0.25
Gradient Boosting	50.00	78.12	64.06	0.69	0.29
AdaBoost	37.50	75.00	56.25	0.63	0.13
XGBoost	48.44	81.25	64.84	0.70	0.31
Extra Trees	54.69	79.69	67.19	0.73	0.36
Logistic Regression	56.25	82.81	69.53	0.78	0.41
KNN	53.12	64.06	58.59	0.64	0.17
SVC	100.00	0.00	50.00	0.39	0.00
MLP	57.81	82.81	70.31	0.78	0.42

Table 4.10. *Performance of various ML classifiers trained on 256 PSSM features selected using the SVC-L1 method on Validation dataset*

Although the SVC-L1 feature selection method was used to reduce redundancy and highlight the most discriminative features, there was decrease in

performance compared to the model trained on the complete PSSM composition features. One of the reason could be that feature selection techniques like SVC-L1 prioritize individual feature importance and may overlook subtle inter-feature dependencies that models can learn from in high-dimensional spaces. Additionally, some fewer dominant features might still carry complementary information that improves the generalizability. Hence, while feature selection improves efficiency and reduces overfitting risk, it can sometimes lead to the exclusion of useful information, resulting in reduced classification performance.

4.3.5 Protein Language Model

In this study, we used different checkpoints of ESM-2 model with different number of layers for the classification of anticancer proteins. Specifically, we evaluated the performance of layers t6, t12, t30, and t33 by fine-tuning each variant on our labeled dataset. Among these, the ESM2-t33 model, which includes 33 transformer layers, produced the best results—achieving an AUC of 0.90 and a MCC of 0.63 on the validation dataset. The detailed performance results across different layers are summarized in Table 4.11. In addition to

Model	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**	**MCC**
ESM2-t6-8M-UR50D	0.70	0.72	0.71	0.79	0.42
ESM2-t12_35M_UR50D	0.72	0.73	0.73	0.82	0.45
ESM2-t30_150M_UR50D	0.84	0.86	0.85	0.87	0.70
ESM2-t33_650M_UR50D	0.83	0.80	0.81	0.90	0.63

Table 4.11. *Performance of fine-tuned ESM2 model checkpoints*

using the ESM-2 model directly for classification, we also extracted the sequence embeddings generated by the fine-tuned ESM2-t33 model. These embeddings represent rich contextual information from the entire protein sequence and were used as input features for various traditional machine learning classifiers. The models trained on these embeddings performed comparably to the fine-tuned transformer itself, also achieving an AUC of 0.90 (Table 4.12).

Model	Sensitivity	Specificity	Accuracy	AUC	MCC
Decision Tree	53.12	76.56	64.84	0.65	0.31
Random Forest	85.94	64.06	75.00	0.86	0.51
Gradient Boosting	79.69	75.00	77.34	0.87	0.55
AdaBoost	67.19	76.56	71.88	0.80	0.44
XGBoost	84.38	71.88	78.12	0.87	0.57
Extra Trees	73.44	82.81	78.12	0.88	0.56
Logistic Regression	78.12	82.81	80.47	0.88	0.61
KNN	70.31	62.50	66.41	0.78	0.33
SVC	76.56	85.94	81.25	0.88	0.63
MLP	81.25	89.06	85.16	0.90	0.71

Table 4.12. *Performance of embeddings extracted using fine-tuned ESM2-t33 model*

The best-performing ESM2-t33 based model has been integrated into both the webserver and standalone versions of our tool for user-friendly accessibility.

4.3.6 Combined Feature Evaluation

To improve the classification of anticancer proteins, we also evaluated different combinations of features to identify the most informative ones. Among all the combinations tested, the combination of PSSM composition features and DSSP-based secondary structure features gave the best results. This combined feature set achieved an AUROC of 0.81 and an MCC of 0.49 on the validation dataset. The performance of different feature combinations using machine learning classifiers is summarized in Table 4.13.

Feature Type	Model Name	Sens	Spec	Acc	AUC	MCC
AAC + DPC	Extra Trees	65.62	67.19	66.41	0.71	0.33
AAC + PCP	Gradient Boosting	68.75	60.94	64.84	0.70	0.30
AAC + DSSP	Extra Trees	76.56	68.75	72.66	0.77	0.45
AAC + RSA	Gradient Boosting	64.06	75.00	69.53	0.71	0.39
AAC + PSSM	Logistic Regression	62.50	87.50	75.00	0.78	0.52
AAC + DSSP + RSA	Extra Trees	73.44	70.31	71.88	0.79	0.44
DSSP + RSA	Random Forest	56.25	82.81	69.53	0.77	0.41
DPC + PCP	Extra Trees	71.88	71.88	71.88	0.75	0.44
DPC + RSA	Random Forest	68.75	81.25	75.00	0.79	0.50
DPC + DSSP	Random Forest	64.06	65.62	64.84	0.73	0.30
DPC + DSSP + RSA	Extra Trees	60.94	78.12	69.53	0.76	0.40
DPC + PSSM	Random Forest	62.50	75.00	68.75	0.71	0.38
PCP + PSSM	Logistic Regression	62.50	84.38	73.44	0.78	0.48
PCP + DSSP	Gradient Boosting	68.75	71.88	70.31	0.74	0.41
PCP + RSA	AdaBoost	60.94	78.12	69.53	0.72	0.40
PCP + DSSP + RSA	Extra Trees	82.81	62.50	72.66	0.79	0.46
PSSM + DSSP	Logistic Regression	67.19	81.25	74.22	0.81	0.49
PSSM + RSA	Logistic Regression	64.06	82.81	73.44	0.78	0.48
PSSM + DSSP + RSA	Logistic Regression	64.06	82.81	73.44	0.80	0.48

Table 4.13. *Performance of different feature combinations using ML classifiers*

4.3.7 Anticancer Peptide Mapping

As described in section 3.3.4 where we have experimented with different Anticancer peptide mapping strategies, we found that AntiCP2 – Direct Count + Length which is a set of 7 features including direct count (6 features) and peptide length (1 feature) and AntiCP2- Direct Count + Length + Length Normalized counts which is a set of 13 features from the combined set of direct count (6 features), peptide length (1 feature) and normalized count (6 features) obtained 0.58 AUC and 0.56 AUC respectively on the validation set by applying extreme gradient boosting (XGB) and gradient boosting (GB) classifier (see Table 4.14). Based on the performance achieved on these 10-mers, we concluded that this hypothesis did not work here. One of the reasons that we can think of could be the limitations of the 10-mer-based approach itself. Anticancer activity might not be dependent only on the properties of individual 10-mers but instead by complex structural or sequence-level interactions that extend beyond short stretches. Also, AntiCP2 was trained on the dataset that constituted shorter peptides length ≤ 40, so the tool may not be able accurately correlate local peptide properties and overall protein activity.

Feature	Model	Sensitivity	Specificity	Accuracy	AUC	MCC
AntiCP2 - DC	ET	54.69	54.69	54.69	0.55	0.09
AntiCP2 - LN	DT	65.62	46.88	56.25	0.58	0.13
AntiCP2 – DC + Length	XGBoost	53.12	65.62	59.38	0.58	0.19
AntiCP2 – DC + Length + LN	GB	51.56	56.25	53.91	0.56	0.08

Table 4.14. *Performance of Peptide Mapping-based features over independent dataset*

4.4 Hybrid Approach

As discussed in section 3.7 to develop a more accurate model, we combined alignment-based approach (BLAST) with alignment free approach (ML). We combined the best e-value (10e-20) hit BLAST scores with ML scores of our best performing LLM model. The model was able to achieve highest AUC of 0.91 and MCC of 0.63 on validation dataset using the combined scores. The detailed results can be seen below in Table 4.15. The best-performing ESM2-t33-based model, when combined with BLAST, has also been integrated into both the web server and standalone versions of our tool as a hybrid model.

e-value	Sensitivity	Specificity	Accuracy	AUC	Kappa	MCC
1.00×10^{1}	54.69	57.81	56.25	0.76	0.12	0.13
1.00×10^{2}	56.25	59.38	57.81	0.77	0.16	0.16
1.00×10^{3}	54.69	57.81	56.25	0.76	0.12	0.13
1.00×10^{4}	54.69	59.38	57.03	0.77	0.14	0.14
1.00×10^{5}	54.69	57.81	56.25	0.76	0.12	0.13
1.00×10^{10}	54.69	57.81	56.25	0.76	0.12	0.13
0	78.12	65.62	71.88	0.85	0.44	0.44
1.00×10^{-1}	79.69	75.00	77.34	0.89	0.55	0.55
1.00×10^{-2}	82.81	75.00	78.91	0.90	0.58	0.58
1.00×10^{-3}	82.81	75.00	78.91	0.90	0.58	0.58
1.00×10^{-4}	82.81	75.00	78.91	0.90	0.58	0.58
1.00×10^{-5}	82.81	75.00	78.91	0.90	0.58	0.58
1.00×10^{-10}	82.81	79.69	81.25	0.90	0.62	0.63
1.00×10^{-20}	84.38	78.12	81.25	0.91	0.62	0.63

Table 4.15. *Performance variation with different e-value thresholds in hybrid model.*

Chapter 5

Deployment

To facilitate widespread usage and accessibility, the finetuned ESM2-t33 and Hybrid – ESM2-t33 + BLAST model has been deployed across multiple platforms. This multi-platform deployment ensures flexibility for researchers and developers with varying technical preferences. Below are the details of each deployment modality:

5.1 Web Interface

A user-friendly web interface has been developed, allowing users to input protein sequences directly into the browser-based interface without the need for local installations. The interface is intuitive and provides real-time predictions, making it suitable for experimental biologist and non-programmers. The frontend of the interface is developed using HTML, CSS and JavaScript along with Bootstrap framework to ensure consistency and responsiveness across a variety of devices and screen sizes. The model runs on a backend server developed using PHP and Python. Figure 5.1 below shows the homepage of the deployed webserver — AntiCP3 — which offers various utilities and a comprehensive user guide for public use.

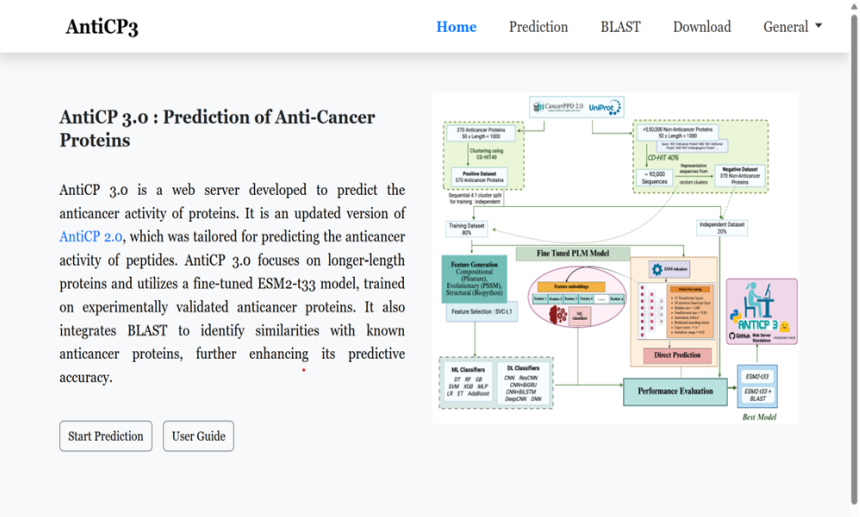


Fig. 5.1. *Homepage of the AntiCP3 webserver*

The website offers two modules –

1. **Prediction** - This module allows users to input protein sequences directly or upload a FASTA file containing multiple sequences. Each sequence is

then classified based on the selected prediction model. Figure 5.2 illustrates the usage of the Predict module within the AntiCP3 web interface, demonstrating how users can input sequences and obtain prediction results.

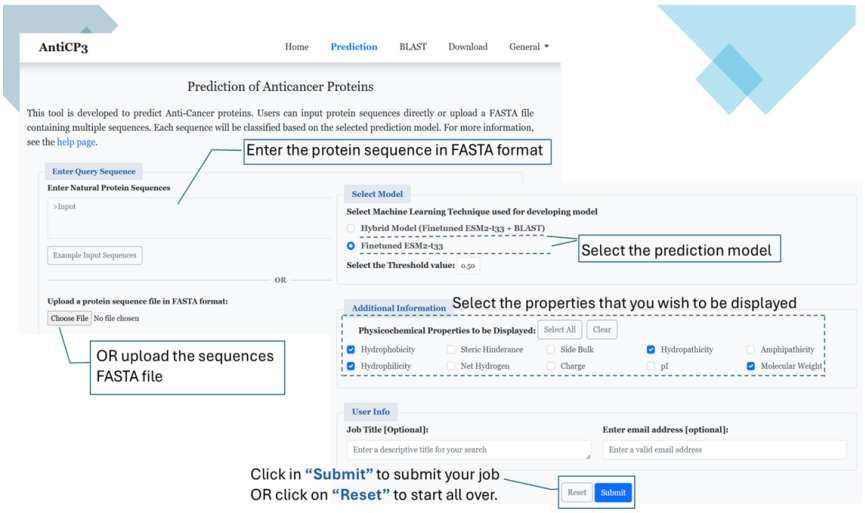
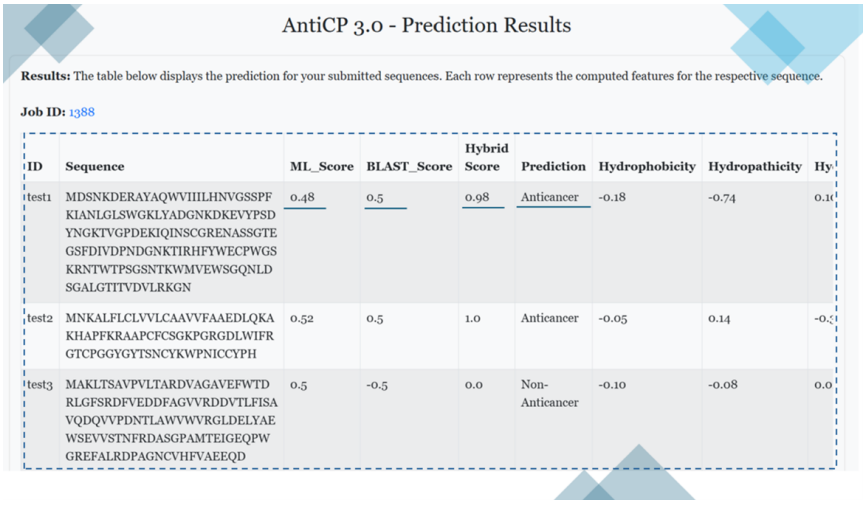


Fig. 5.2. *Predict module of AntiCP3*

The content of the table varies depending on the selected prediction model. When the Hybrid Model is used, the table displays the following columns: ML_Score, BLAST_Score, Hybrid_Score, Prediction, along with all selected physicochemical properties such as hydrophobicity, molecular weight, and others and if Model 1 is selected, the table includes ML_Score, Prediction, and the selected physicochemical properties as shown in Figure 5.3.



AntiCP 3.0 - Prediction Results

Results: The table below displays the prediction for your submitted sequences. Each row represents the computed features for the respective sequence.

Job ID: 1388

ID	Sequence	ML_Score	BLAST_Score	Hybrid Score	Prediction	Hydrophobicity	Hydropathicity	Hy
test1	MDSNKDERAYAQWVIIILHNVGSSPF KIANLGLSWGKLYADGNKDKEVYPSD YNGKTVGPDEKIQINSCGRENASSGTE GSFDIVDPNDGNKTIRHFYWECPWGS KRNTWTPSGSNTKWMVEWSGQNLD SGALGTITVDVLRKGN	0.48	0.5	0.98	Anticancer	-0.18	-0.74	0.1
test2	MNKALFLCLVVLCAAVVFAAEDLQKA KHAPFKRAAPCFCSGKPGRGDLWIFR GTCPGGYGYTSNCYKWPNICCYPH	0.52	0.5	1.0	Anticancer	-0.05	0.14	-0.
test3	MAKLTSAVPVLTARDVAGAVEFWTD RLGFSRDFVEDDFAGVVRDDVTLFISA VQDQVVPDNTLAWVWVRGLDELYAE WSEVVSTNFRDASGPAMTEIGEQPW GREFALRDPAGNCVHFVAEEQD	0.5	-0.5	0.0	Non-Anticancer	-0.10	-0.08	0.0

Fig. 5.3. *Tabular display of prediction results*

2. **BLAST** - This module enables users to perform a BLAST search, comparing their query protein sequence against a database of known anticancer

and non-anticancer proteins. The module classifies the query sequence as anticancer if it matches any entries in the database. Users can input multiple protein sequences at once in the text area or upload them in FASTA format. Figure 5.4 illustrates the usage of the BLAST module.

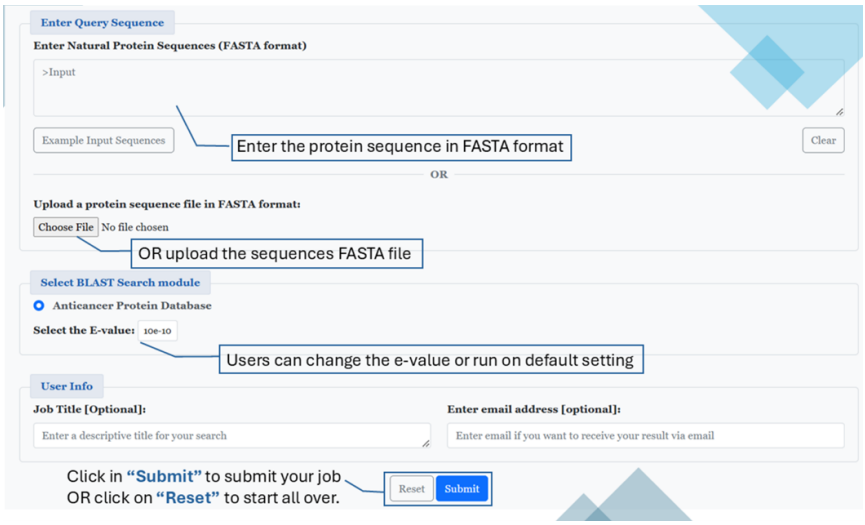


Fig. 5.4. *Usage of the BLAST module of AntiCP3*

Based on the hit, each submitted sequence is given a prediction. Users can also view the detailed BLAST alignment results. Figure 5.5 shows detailed BLAST results.

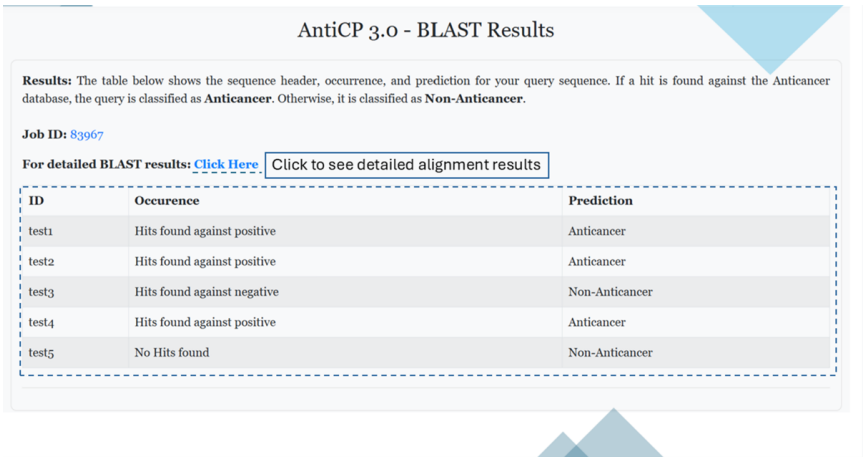


Fig. 5.5. *Tabular display of BLAST results on AntiCP3 webserver*

5.2 Standalone Application

For users requiring a more autonomous solution, a standalone application in the form of a command-line interface (CLI) tool has been developed. This version allows for the local execution of the model, eliminating the need for continuous internet connectivity. The CLI tool is designed to work across all

major operating systems—Windows, Linux, and macOS—ensuring broad accessibility. Users can download the standalone version from (https://webs.iiitd.edu.in/raghava/anticp3/down.html). The application operates based on command-line arguments, with the input file name being the only mandatory argument. Users can customize the execution further using optional arguments, providing flexibility to the users. The available arguments are as follows:

- `-i` or `--input` : The input file name (required). This is the file containing the protein sequences that will be classified.
- `-o` or `--output` : The output file name (optional). If not specified, the results file is saved as `output.csv`.
- `-m` or `--model` : Specifies the model to be used (optional). By default, model – Hybrid (ESM2 + BLAST) will be used.
- `-d` or `--device` : Specifies the device for inference (optional).
- `-t` or `--threshold` : Sets the threshold for classification (optional, default: 0.5).

This CLI tool allow users to use our model locally, especially for batch-processing and/or integrating this model into larger workflows. Figure 5.6 shows the interface of the standalone version of AntiCP3, which provides all core functionalities locally.

Fig. 5.6. *Standalone version of the AntiCP3 tool*

5.3 Pip Package

For users who intend to use the model into their existing Python workflows, a pip-installable package is available. The package is hosted on the Python Package Index (PyPI). The package can be installed simply by running :

```
pip install anticp3
```

Link to PyPi - https://pypi.org/project/anticp3/

5.4 GitHub Repository

The complete source code and model files are made available on GitHub (https://github.com/raghavagps/anticp3). This repository enables developers to access the project directly through GitHub. It serves as an open-source platform, allowing the scientific and development communities to use the model, submit issues, and share improvements. Documentation and installation instructions are provided to ensure ease of use for those who wish to deploy the model in their own environments. Figure 5.7 shows the GitHub repository interface.

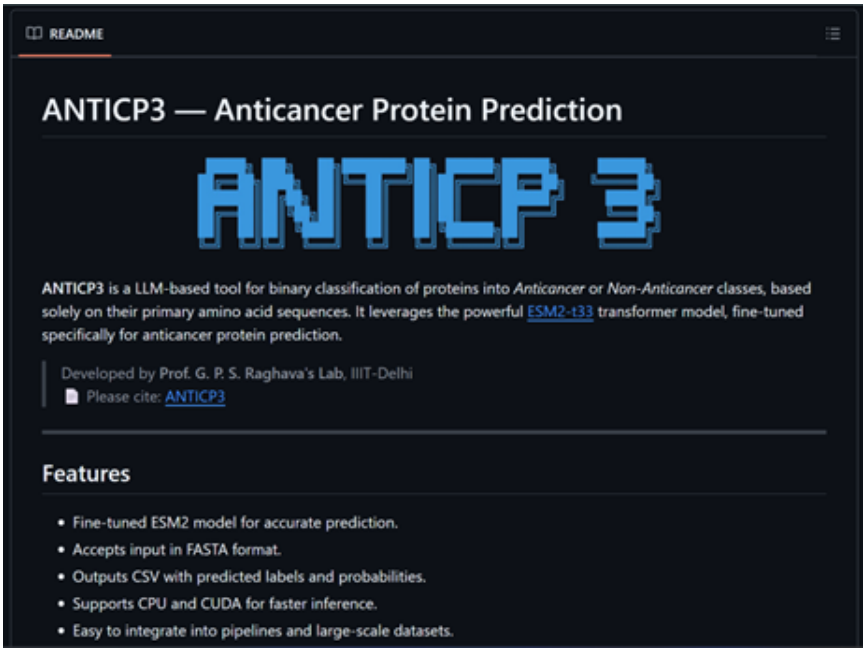


Fig. 5.7. *GitHub repository of AntiCP3*

5.5 Hugging Face Model Hub

To enhance accessibility and integration into modern machine learning workflows, the fine-tuned model has also been deployed on the Hugging Face Model Hub, a widely used platform for sharing and utilizing state-of-the-art machine learning models. Users can deploy the fine-tuned checkpoint for their prediction purposes and can experiment with the architecture. They can directly load the model and tokenizer into their existing bioinformatics pipeline. Below is a sample code snippet demonstrating how to load and use the model for inference:

```
from transformers import AutoTokenizer,
    AutoModelForSequenceClassification
import torch

tokenizer = AutoTokenizer.from_pretrained("raghavagps-group/anticp3")
model = AutoModelForSequenceClassification.from_pretrained("raghavagps
    -group/anticp3")

sequence = "MANCVVGYIGERCQYRDLKWWELRGGGGSGGGGSAPAFSVSPASGLSDGQSVSV"

# Tokenize and run inference
inputs = tokenizer(sequence, return_tensors="pt", truncation=True)

with torch.no_grad():
    logits = model(**inputs).logits
    probs = torch.nn.functional.softmax(logits, dim=-1)
    prediction = torch.argmax(probs, dim=1).item()

labels = {0: "Non-Anticancer", 1: "Anticancer"}
print("Prediction:", labels[prediction])
```

Chapter 6

Discussion

Protein- and peptide- based therapeutics are gaining attention in the field of oncology. These proteins are used in the development of novel treatment strategies, like monoclonal antibodies (mAbs), small-molecule drugs, and tumour growth inhibitors [61]. Some Proteins exhibit multifunctional properties; for eg., several Antimicrobial peptides also show Anticancer activity. Bleomycin, a compound derived from Streptomyces verticillus, is a well-known example, with proven efficacy against various cancers such as head and neck squamous cell carcinomas, Hodgkin's disease, and ovarian cancer [62]. Despite their promising therapeutic potential, the identification and experimental validation of anticancer proteins remain time-consuming, costly, and labour-intensive. This highlights the growing need for reliable in silico prediction tools that can streamline the discovery process. While most previous studies have focused on anticancer peptides, our study specifically targeted anticancer proteins due to their broader and potentially more impactful biological functions.

In this work, we analysed several feature types for modelling anticancer proteins, including composition-based, evolutionary profile-based, secondary structure, and sequence-derived features. Each feature set offers unique insights—for example, composition-based features reveal amino acid arrangement, while evolutionary profiles highlight conserved regions critical to biological function [63]. The integration of diverse features contributed to improved predictive performance. Among the features tested, the Anticancer peptide mapping features were least effective. This could be due to two main reasons:

1. Complex Interactions: Anticancer activity of proteins is likely not determined by short peptides alone. Instead, it might involve more intricate interactions between the entire protein's structure or sequence, which cannot be captured by small peptide segments.

2. Limitations of AntiCP2: AntiCP2, which was trained on shorter peptides (length ≤ 40) , may not be capable of identifying the complex relationships between local peptide properties and the overall anticancer function of proteins.

We also explored different combinations of features and applied feature selection methods to improve our model. Our best-performing model, based on the ESM-2 fine-tuned model, achieved an AUC of 0.90 and an MCC of 0.71 on the validation set, outperforming other classifiers. Building on this, we created a hybrid model by combining ESM-2 with BLAST, which improved performance even further, reaching an AUC of 0.91 and an MCC of 0.63. This hybrid model demonstrates how combining sequence similarity searches with machine learning can boost the prediction accuracy of anticancer proteins. To serve the scientific

community and developers' community, we developed a web server and provided standalone versions, GitHub repositories, and a pip package, making it easier for the users to use our model as per their convenience.

In conclusion, the development of AntiCP3 tool is an important step in the direction of predicting and selecting candidates for protein-based therapeutics. By developing models that combine various types of features and using hybrid approaches, we have created a tool that can predict Anticancer proteins with high accuracy using high quality manually annotated dataset. We believe that AntiCP3 will help researchers in their search for new anticancer therapies, and will prove to be a valuable resource in the direction of advancing cancer treatment as cancer still remains one of the biggest threats in healthcare.

Chapter 7

Limitation and Future Scope

While this study represents a novel step toward the in-silico prediction of Anticancer proteins, there are several limitations must be acknowledged. A major limitation is the size of the dataset used for model training and finetuning. Because of the limited number of experimentally validated anticancer proteins, the current model was developed using a relatively small dataset, comprising 370 positive anticancer protein sequences and equal number of negative sequences. The results were good for the preliminary analysis, but for more generalise results, larger dataset is required, as it can capture the broad spectrum of sequence variability found in Anticancer proteins and improve predictive performance.

To address this limitation, we plan for continuous retraining of the model using an expanded dataset as the databases grow and more annotated sequences become available. Ultimately, such advancements may support the identification of novel therapeutic candidates and contribute to the ongoing efforts in cancer drug discovery and development.

References

[1] E. Krieghoff-Henning, J. Folkerts, A. Penzkofer, and S. Weg-Remers, "Cancer – an overview," *Medizinische Monatsschrift für Pharmazeuten*, vol. 40, pp. 48–54, Feb. 2017. PMID: 29952494.

[2] W. H. Organization, "Global cancer burden growing, amidst mounting need for services," 2024.

[3] Y. Zhang, C. Wang, W. Zhang, and X. Li, "Bioactive peptides for anticancer therapies," *Biomater Transl*, vol. 4, no. 1, pp. 5–17, 2023.

[4] W. Szlasa, I. Zendran, A. Zalesińska, M. Tarek, and J. Kulbacka, "Lipid composition of the cancer cell membrane," *J Bioenerg Biomembr*, vol. 52, pp. 321–342, Oct 2020. Epub 2020 Jul 26.

[5] M. Xie, D. Liu, and Y. Yang, "Anti-cancer peptides: classification, mechanism of action, reconstruction and modification," *Open Biol*, vol. 10, p. 200004, Jul 2020. Epub 2020 Jul 22.

[6] M. Chauhan, A. Gupta, R. Tomer, and G. P. S. Raghava, "Cancerppd2: an updated repository of anticancer peptides and proteins," *Database*, vol. 2025, p. baaf030, 05 2025.

[7] A. Lee, J. Harris, K. Khanna, and J. Hong, "A comprehensive review on current advances in peptide drug development and design," *Int J Mol Sci*, vol. 20, p. 2383, May 14 2019.

[8] T. U. Consortium, "Uniprot: the universal protein knowledgebase in 2025," *Nucleic Acids Research*, vol. 53, pp. D609–D617, 11 2024.

[9] 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.

[10] National Center for Biotechnology Information (NCBI), "NCBI - National Center for Biotechnology Information." https://www.ncbi.nlm.nih.gov/, 1988. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2025 May 18].

[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] Z. Hajisharifi, M. Piryaiee, M. M. Beigi, M. Behbahani, and H. Mohabatkar, "Predicting anticancer peptides with Chous pseudo amino acid composition and investigating their mutagenicity via Ames test," *Journal of Theoretical Biology*, vol. 341, pp. 34–40, 2014.

[13] V. K. Sangaraju, N. T. Pham, L. Wei, X. Yu, and B. Manavalan, "macppred 2.0: Stacked deep learning for anticancer peptide prediction with integrated spatial and probabilistic feature representations," *Journal of Molecular Biology*, vol. 436, no. 17, p. 168687, 2024. Computation Resources for Molecular Biology.

[14] B. Han, N. Zhao, C. Zeng, Z. Mu, and X. Gong, "ACPred-BMF: bidirectional LSTM with multiple feature representations for explainable anticancer peptide prediction," *Scientific Reports*, vol. 12, p. 21915, Dec. 2022.

[15] 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, Aug. 2020. _eprint: https://academic.oup.com/bib/article-pdf/22/3/bbaa153/37962923/bbaa153.pdf.

[16] 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, 2024.

[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, 2022.

[18] M. Burdukiewicz, K. Sidorczuk, D. Rafacz, F. Pietluch, M. Bakała, J. Słowik, and P. Gagat, "Cancergram: An effective classifier for differentiating anticancer from antimicrobial peptides," *Pharmaceutics*, vol. 12, no. 11, 2020.

[19] H.-C. Yi, Z.-H. You, X. Zhou, L. Cheng, X. Li, T.-H. Jiang, and Z.-H. Chen, "Acp-dl: A deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation," *Molecular Therapy - Nucleic Acids*, vol. 17, pp. 1–9, 2019.

[20] 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, Mar. 2015.

[21] P. B. Timmons and C. M. Hewage, "Ennaact is a novel tool which employs neural networks for anticancer activity classification for therapeutic peptides," *Biomedicine Pharmacotherapy*, vol. 133, p. 111051, 2021.

[22] L. Yu, R. Jing, F. Liu, J. Luo, and Y. Li, "Deepacp: A novel computational approach for accurate identification of anticancer peptides by deep learning algorithm," *Molecular Therapy - Nucleic Acids*, vol. 22, pp. 862–870, 2020.

[23] S. Zhang, Y. Zhao, and Y. Liang, "Aacflow: an end-to-end model based on attention augmented convolutional neural network and flow-attention mechanism for identification of anticancer peptides," *Bioinformatics*, vol. 40, p. btae142, 03 2024.

[24] J. Liu, M. Li, and X. Chen, "Antimf: A deep learning framework for predicting anticancer peptides based on multi-view feature extraction," *Methods*, vol. 207, pp. 38–43, 2022.

[25] X. Wang and S. Wang, "Acp-pdaff: Pretrained model and dual-channel attentional feature fusion for anticancer peptides prediction," *Computational Biology and Chemistry*, vol. 112, p. 108141, 2024.

[26] N. Lane and I. Kahanda, "Deepacppred: A novel hybrid cnn-rnn architecture for predicting anti-cancer peptides," in *Practical Applications of Computational Biology & Bioinformatics, 14th International Conference (PACBB 2020)* (G. Panuccio, M. Rocha, F. Fdez-Riverola, M. S. Mohamad, and R. Casado-Vara, eds.), (Cham), pp. 60–69, Springer International Publishing, 2021.

[27] M. Arif, S. Ahmed, F. Ge, M. Kabir, Y. D. Khan, D.-J. Yu, and M. Thafar, "Stackacpred: Prediction of anticancer peptides by integrating optimized multiple feature descriptors with stacked ensemble approach," *Chemometrics and Intelligent Laboratory Systems*, vol. 220, p. 104458, 2022.

[28] V. Boopathi, S. Subramaniyam, A. Malik, G. Lee, B. Manavalan, and D.-C. Yang, "macppred: A support vector machine-based meta-predictor for identification of anticancer peptides," *International Journal of Molecular Sciences*, vol. 20, no. 8, 2019.

[29] G. Feng, H. Yao, C. Li, R. Liu, R. Huang, X. Fan, R. Ge, and Q. Miao, "Me-acp: Multi-view neural networks with ensemble model for identification of anticancer peptides," *Computers in Biology and Medicine*, vol. 145, p. 105459, 2022.

[30] H. Deng, M. Ding, Y. Wang, W. Li, G. Liu, and Y. Tang, "Acp-mlc: A two-level prediction engine for identification of anticancer peptides and multi-label classification of their functional types," *Computers in Biology and Medicine*, vol. 158, p. 106844, 2023.

[31] M. Sun, H. Hu, W. Pang, and Y. Zhou, "Acp-bc: A model for accurate identification of anticancer peptides based on fusion features of bidirectional long short-term memory and chemically derived information," *International Journal of Molecular Sciences*, vol. 24, no. 20, 2023.

[32] J. Chen, H. H. Cheong, and S. W. I. Siu, "xDeep-AcPEP: Deep Learning Method for Anticancer Peptide Activity Prediction Based on Convolutional Neural Network and Multitask Learning," *Journal of Chemical Information and Modeling*, vol. 61, pp. 3789–3803, Aug. 2021. Publisher: American Chemical Society.

[33] L. Zhu, C. Ye, X. Hu, S. Yang, and C. Zhu, "Acp-check: An anticancer peptide prediction model based on bidirectional long short-term memory and multi-features fusion strategy," *Computers in Biology and Medicine*, vol. 148, p. 105868, 2022.

[34] 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, p. 1352504, 2024.

[35] 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.

[36] 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, p. 1376486, 2024.

[37] Q. Li, W. Zhou, D. Wang, S. Wang, and Q. Li, "Prediction of anticancer peptides using a low-dimensional feature model," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 892, 2020.

[38] L. Wei, C. Zhou, R. Su, and Q. Zou, "Pepred-suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning," *Bioinformatics*, vol. 35, pp. 4272–4280, 04 2019.

[39] B. Rao, C. Zhou, G. Zhang, R. Su, and L. Wei, "Acpred-fuse: fusing multi-view information improves the prediction of anticancer peptides," *Briefings in Bioinformatics*, vol. 21, pp. 1846–1855, 11 2019.

[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] N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, and W. Shoombuatong, "Acpred: A computational tool for the prediction and analysis of anticancer peptides," *Molecules*, vol. 24, no. 10, 2019.

[42] B. Manavalan, S. Basith, T. Hwan Shin, S. Choi, M. Ok Kim, and G. Lee, "Mlacp: machine-learning-based prediction of anticancer peptides," *Oncotarget*, vol. 8, no. 44, pp. 77121–77136, 2017.

[43] 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, pp. 16895–16909, 2016.

[44] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, "Acpred-fl: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides," *Bioinformatics*, vol. 34, pp. 4007–4016, 06 2018.

[45] Y. Zhao, S. Wang, W. Fei, Y. Feng, L. Shen, X. Yang, M. Wang, and M. Wu, "Prediction of anticancer peptides with high efficacy and low toxicity by hybrid model based on 3d structure of peptides," *International Journal of Molecular Sciences*, vol. 22, no. 11, 2021.

[46] Y. Li, D. Ma, D. Chen, and Y. Chen, "Acp-gbdt: An improved anticancer peptide identification method with gradient boosting decision tree," *Frontiers in Genetics*, vol. 14, p. 1165765, 2023.

[47] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, pp. 1658–1659, 05 2006.

[48] A. Khanduja and D. Mohanty, "Sprotfp: a machine learning-based method for functional classification of small orfs in prokaryotes," *NAR Genomics and Bioinformatics*, vol. 7, p. lqae186, 01 2025.

[49] A. Pande, S. Patiyal, A. Lathwal, C. Arora, D. Kaur, A. Dhall, G. Mishra, H. Kaur, N. Sharma, S. Jain, S. S. Usmani, P. Agrawal, R. Kumar, V. Kumar, and G. P. Raghava, "Pfeature: A tool for computing wide range of protein features and building prediction models," *Journal of Computational Biology*, vol. 30, no. 2, pp. 204–222, 2023. PMID: 36251780.

[50] 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.

[51] S. Miller, J. Janin, A. M. Lesk, and C. Chothia, "Interior and surface of monomeric proteins," *Journal of Molecular Biology*, vol. 196, no. 3, pp. 641–656, 1987.

[52] B. Rost and C. Sander, "Conservation and prediction of solvent accessibility in protein families," *Proteins*, vol. 20, pp. 216–226, Nov 1994.

[53] M. Z. Tien, A. G. Meyer, D. K. Sydykova, S. J. Spielman, and C. O. Wilke, "Maximum allowed solvent accessibilites of residues in proteins," *PLoS One*, vol. 8, p. e80635, Nov 2013.

[54] P. Chakrabarti, "On the pathway of the formation of secondary structures in proteins," *Proteins*, vol. 93, pp. 396–399, Jan 2025. Epub 2023 Sep 23.

[55] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, pp. 3389–3402, 09 1997.

[56] J. Wang, B. Yang, J. Revote, A. Leier, T. T. Marquez-Lago, G. Webb, J. Song, K.-C. Chou, and T. Lithgow, "Possum: a bioinformatics toolkit for generating numerical sequence feature descriptors based on pssm profiles," *Bioinformatics*, vol. 33, pp. 2756–2758, 05 2017.

[57] N. Q. Khanh Le, Q. H. Nguyen, X. Chen, S. Rahardja, and B. P. Nguyen, "Classification of adaptor proteins using recurrent neural networks and PSSM profiles," *BMC Genomics*, vol. 20, p. 966, Dec. 2019.

[58] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *PNAS*, 2019.

[59] W. R. Pearson, "An introduction to sequence similarity ("homology") searching," *Current Protocols in Bioinformatics*, vol. Chapter 3, pp. 3.1.1–3.1.8, Jun 2013.

[60] C. Vens, M.-N. Rosso, and E. G. J. Danchin, "Identifying discriminative classification-based motifs in biological sequences," *Bioinformatics*, vol. 27, pp. 1231–1238, 03 2011.

[61] A. G. Mukherjee, U. R. Wanjari, A. V. Gopalakrishnan, P. Bradu, A. Biswas, R. Ganesan, K. Renu, A. Dey, B. Vellingiri, A. El Allali, A. M. Alsamman, H. Zayed, and C. George Priya Doss, "Evolving strategies and application of proteins and peptide therapeutics in cancer treatment," *Biomedicine Pharmacotherapy*, vol. 163, p. 114832, 2023.

[62] T. M. Karpiński and A. Adamczak, "Anticancer activity of bacterial proteins and peptides," *Pharmaceutics*, vol. 10, no. 2, 2018.

[63] D. Harding-Larsen, J. Funk, N. G. Madsen, H. Gharabli, C. G. Acevedo-Rocha, S. Mazurenko, and D. H. Welner, "Protein representations: Encoding biological information for machine learning in biocatalysis," *Biotechnology Advances*, vol. 77, p. 108459, 2024.