In Silico Approaches for Biomolecule-Driven Disease Diagnosis and Therapy

By
Ritu Tomer

(PhD20207)

Under the Supervision of Prof. Gajendra P.S. Raghava
Department of Computational Biology
Indraprastha Institute of Information Technology, Delhi
New Delhi – 110020

October, 2025

In Silico Approaches for Biomolecule-Driven Disease Diagnosis and Therapy

By
Ritu Tomer

(PhD20207)

A Thesis
Submitted in Partial Fulfilment of the Requirements for the Degree Of
Doctor of Philosophy

Under the Supervision of Prof. Gajendra P.S. Raghava
Department of Computational Biology
Indraprastha Institute of Information Technology, Delhi
New Delhi – 110020

October, 2025

Certificate

This is to certify that the thesis titled "**In Silico Approaches for Biomolecule-Driven Disease Diagnosis and Therapy**" being submitted by **Ms. Ritu Tomer** to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of Doctor of Philosophy, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

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

December, 2025

Sign

Prof. Gajendra P.S. Raghava

Indraprastha Institute of Information Technology Delhi

New Delhi 110020

Declaration

“I acknowledge that I am fully responsible for the entire content of my thesis, including any sections assisted by online tools, including Artificial Intelligence-based tools. I accept full accountability for any violations of ethical standards in publications arising from the use of such tools.”.

December, 2025

Student Name: Ritu Tomer

Roll Number: PhD20207

Advisor Name: Prof. Gajendra P. S. Raghava

Indraprastha Institute of Information Technology Delhi

New Delhi 110020

Acknowledgement

“I owe every milestone not to my strength alone, but to the hands that lifted me, the words that encouraged me, and the hearts that believed in me.”

As I stand at the juncture between an ending and a new beginning, I find myself reflecting on a journey that has been as much about inner transformation as it has been about scientific discovery. The completion of this thesis marks not only the culmination of years of research but also the evolution of a person shaped by perseverance, faith, and the unwavering support of many remarkable individuals. What began as a pursuit of knowledge gradually became a voyage of patience, humility, and self-discovery - one that taught me that growth often unfolds in silence, and strength often reveals itself through grace.

My deepest and most profound gratitude is reserved for my mentor and supervisor, **Prof. Gajendra P. S. Raghava**, whose brilliance, integrity, and empathy have profoundly influenced my intellectual and personal journey. His mentorship has been far more than guidance in research - it has been a lesson in leadership, patience, and the art of thinking deeply yet simply. His ability to find purpose in precision and meaning in complexity has transformed how I perceive science and life. I remain forever indebted to him for his faith in me, for steering me through uncertainty, and for inspiring me to pursue excellence with humility and sincerity.

I would like to convey my profound thanks to **Prof. Gajendra P. S. Raghava**, and **Prof. Ranjan Bose** (the IIIT-D Director) for allowing me the opportunity to be admitted to the PhD programme and for afterwards allowing me to utilise the institute's resources for productive research. My profound gratitude goes out to **Dr. Gaurav Ahuja** and **Dr. Jaspreet Kaur Dhanjal** for being in my committee member and helping me along the way with their guidance and expertise. Additionally, I'd like to thank, **Prof. Gajendra P. S. Raghava, Dr. Gaurav Ahuja, Dr. Vibhor Kumar, and Dr. Debarka Sengupta** for teaching me during my course work. The administrative personnel at IIIT-D, especially **Ms. Shipra Jain, Ms. Anshu Dureja, and Raju Biswas** deserve a special thank you for always being accessible to answer our queries and immediately resolve our academic issues. I am also appreciative of IIIT-D for offering first-class facilities and infrastructure. An official thanks to Department of Science & Technology (DST), the funding source, for giving me the **"Innovation in Science Pursuit for Inspired**

Research (INSPIRE)" research fellowship to help with my doctoral studies. I would especially like to thank **Mr. Manoj and Dr. Sachin Suriyan,** who taught me at the J.V.C college and **Dr. Bhupesh Taneja,** who guide me during my master's dissertation at CSIR-IGIB, for giving me the information and direction I needed to begin this wonderful scientific journey.

I am deeply thankful to **Dr. Pawan Kumar** and **Dr. Rajan Srivastav**, whose wisdom, guidance, and faith in my abilities have been a continuous source of inspiration. Their encouragement not only strengthened my professional pursuits but also helped me navigate the uncertainties and challenges that accompany a long research journey.

I extend my warmest thanks to my dear friends - **Sneha Agarwal, Pooja Beniwal, Prashasti,** and **Dr. Alok Nikhil Jha** - whose laughter, warmth, and unwavering friendship have brought balance and joy to my life. Their companionship has been a gentle reminder that friendship, like light, can brighten even the darkest moments of struggle.

My sincere appreciation goes to my seniors - their constant guidance **Dr. Anjali Dhall, Dr. Sumeet Patiyal, Dr. Dilraj Kaur, Dr. Harpreet Kaur, Dr. Neetesh Pandey, Dr. Madhu Sharma, Dr. Smriti Chawla**, and **Ms. Shipra Jain** for, generosity, and kindness. Their experience and encouragement have often served as a guiding light during moments of doubt, and their empathy has left an indelible mark on my journey.

To my labmates - **Shubham, Nishant, Pratik, Pushpendra, Dhananjay, Anand, Naman, Shivani, Dikscha, Nisha, Saloni, Pankaj, Anshuma, Kunal, Devraj, Manish, Piyush, Sahil and Sachin** - I extend my sincere appreciation for creating an environment enriched with learning, collaboration, and shared purpose. The numerous discussions, coupled with occasional light-hearted exchanges and collegial banter, added a distinctive character to this journey. Each experience, whether pleasant or challenging, contributed meaningfully to my professional growth and resilience as a researcher.

As I bring these acknowledgments to a close, my heart turns with deepest reverence to the **Almighty** and to my late father, **Mr. BrijPal Singh**, whose blessings, love, and values continue to illuminate my path beyond the realm of sight. His absence has been my greatest sorrow, yet his spirit has been my greatest strength. I owe all that I am today to the values of resilience, integrity, and compassion instilled in me by him and my mother, **Smt. Savita Devi**. My heartfelt gratitude extends to my entire family, including my **in-laws**, for their

constant love, patience, and encouragement - the invisible scaffolding that sustained me through the most challenging times.

A very special acknowledgment goes to my husband, **Mr. Nitin Malik**, whose unwavering faith, patience, and unconditional support have been the foundation upon which my dreams have stood firm. His quiet strength, understanding, and belief in my journey have been both my comfort and my courage. I am also thankful to the youngest members of my family, **Ashika, Advik, Pari, Aayu,** and **Daksh** whose laughter and innocence bring boundless joy to my life and remind me of the beauty of balance between work and love.

As I look back, I realize that this thesis is not merely a reflection of my work but a tapestry woven with the encouragement, compassion, and belief of all those who stood beside me. Every challenge endured and every breakthrough achieved has been a silent lesson in perseverance and grace. This journey has taught me that success is never solitary, it is the sum of countless unseen hands, kind words, and shared moments of hope. To all who have walked beside me in silence and strength, I offer my deepest gratitude. This accomplishment belongs as much to them as it does to me.

Ritu Tomer

Abstract

Developments in computational biology have facilitated the systematic study of biomolecules including peptides, proteins or nucleic acids. Such advancements have provided disease-oriented research and therapeutic discovery opportunities. The thesis deals with two broad disciplines in this field: Disease Diagnosis and Biomolecule-Based Therapeutics, with a particular focus on integrating curated data resources and machine-learning-based predictive systems. The initial part is aimed at enhancing the molecular knowledge and diagnostic studies of mucormycosis, a serious and rapidly spreading fungal infestation also referred to as “Black Fungus”. Here, we have a developed a web-repository titled as “MucormyDB”. The repository is a compilation of genomic, proteomic, virulence and therapeutic information. This repository enables researchers to easily perform comparative analyses and allows the identification of possible genetic biomarkers. With such capabilities MucormyDB is a valuable resource to study the molecular basis of mucormycosis. The second part of the thesis provides computational models for peptide-based therapeutics prediction. This involves IL4pred2, a machine-learning model that predicts peptides that can induce interleukin-4, and the AntiCP4, which predicts anticancer peptides with enhanced predictive capability. In addition to therapeutic prediction, identification of safety of peptide candidates is also taken into account in the thesis. In order to deal with this aspect, we have designed RAIpred to detect peptides that could trigger rheumatoid arthritis. We have also developed CDpred to predict peptides related to the celiac disease. These tests provide a valuable adjunct for immunological risk determination. They help researchers to select safer peptide candidates in the early phase of drug discovery early. Together, the resources developed in this thesis present high-quality molecular data, powerful predictive algorithms, and easily available computing platforms that can be used to study mucormycosis and complementarily improve the systematic evaluation of therapeutic peptides and the potential immunological consequences of peptides.

Table of Contents

Certificate 3

Declaration 4

Acknowledgement 5

Abstract 8

List of Abbreviations 12

List of Figures 14

List of Tables 15

1. Introduction 16

1.1. Background 16
1.1.1. Protein and Peptides in Disease Diagnosis and Therapeutics 17
1.1.2. Transition to Computational Diagnostics and Therapy 17

1.2. Origin of the Proposal 18

1.3. Objective of Thesis 19

1.4. Organization of Chapters 20

2. Review of Literature 22

2.1. Overview 22

2.2. Peptides as Regulatory and Therapeutic Molecules 22

2.3. Computational Biology and AI in Molecular Research 23

2.4. Disease Diagnosis 24
2.4.1. Role of Computational Databases in Disease Identification 24
2.4.2. Fungal Disease Repositories and Evolution 24

2.5. Biomolecule-Based Therapeutics 26
2.5.1. Interleukin-4 (IL-4) Inducing Peptides 26
2.5.2. Anticancer Peptide (ACP) Prediction 27
2.5.3. Limitations in Existing Computational Approaches 32

2.6. Safety assessment of peptides 32
2.6.1. Immunogenic and Autoimmune Reactions Triggered by Peptides 32
2.6.2. Computational Models for Immunological Safety Prediction 33

2.7. Conclusion 33

3. Compilation of resources for Mucormycosis diagnosis 35

3.1. Introduction 35

3.2. Database Design 37

3.3. Data Curation and Annotation 38
3.3.1. Genome and Proteome Data 38
3.3.2. Available therapeutics 38
3.3.3. Vaccine Design 39

3.4. Discussion and Conclusion 42

4. Host Specific Prediction of Interleukin-4 Inducing Peptides 44

4.1. Introduction 44

4.2. Methodology and Algorithm Development 45

4.2.1. Data Preparation and preprocessing 45

4.2.2. Feature extraction using standalone tools 46

4.2.3. Computational Framework 47

4.2.4. Model evaluation 48

4.2.5. Ensemble approach 49

4.3. Result and Analysis 49

4.3.1. Preferential Positional Analysis 49

4.3.2. Compositional Analysis 50

4.3.3. Univariate analysis 52

4.3.4. Alignment-Based Approach 55

4.3.5. Alignment-Free Approaches 56

4.3.6. Benchmarking 64

4.3.7. Webserver Implementation 66

4.4. Discussion and Conclusion 67

5. Identification of anticancer peptides using sequence based features 69

5.1. Introduction 69

5.2. Materials and Methods 70

5.2.1. Overall architecture of the study 70

5.2.2. Creation of dataset and its pre-processing 71

5.2.3. Generation of features 71

5.2.4. Development of model 72

5.2.5. Performance measures for evaluation 72

5.3. Results 72

5.3.1. Sequence based Analysis 72

5.3.1.1. Compositional Analysis 73

5.3.1.2. Mean-Based Univariate Analysis 73

5.3.1.3. Logistic Regression-Based Single Feature Analysis 74

5.3.2. Model Performance and Analysis 75

5.3.3. Comparison with existing approach 77

5.3.4. Web-based services 78

5.4. Discussion and Conclusion 79

6. Prediction of Rheumatoid Arthritis-Inducing Peptides in an Antigen 81

6.1. Introduction 81

6.2. Model Architecture and Performance 82

6.2.1. Data Preparation and Feature Extraction 83

6.2.2. Feature extraction and Selection 84

6.2.3. Model development 84

6.3. Result Analysis 84

6.3.1. Sequence analysis 84

6.3.2. Composition-based feature analysis 85

6.3.3. Motif-based analysis 86

6.3.4. Ensemble Model 86

6.4. Discussion 87

6.5. Conclusion and Limitation 88

7. Determination of Specific Epitopes and Motifs in a protein associated with Celiac Disease *89*

7.1. Introduction **89**

7.2. Algorithm Development **91**

7.2.1. Dataset Preparation 91

7.2.2. Preliminary Analysis 92

7.2.3. Model development using ML & ensemble approach 92

7.3. Result and Analysis **93**

7.3.1. Frequency of HLA-alleles 93

7.3.2. Sequence pattern analysis 93

7.3.3. Motif-based analysis 95

7.3.4. PQ density 96

7.3.5. ML-based approach 97

7.3.6. Ensemble approach 98

7.4. Case Studies of CDpred **98**

7.5. Discussion and Conclusion **99**

8. Summary *101*

8.1. Overview **101**

8.2. MucormyDB **102**

8.3. Host-Specific Modelling of IL-4 Inducing Peptides **103**

8.4. Predicting and Designing Anticancer Peptides **104**

8.5. Prediction of Autoimmune Peptides in Rheumatoid Arthritis **106**

8.6. Identification of Celiac Disease-Inducing Peptides **107**

8.7. Key Contributions **109**

8.8. Limitations and Future Directions **109**

8.9. Conclusion **109**

List of Publications *111*

URL of Computational Resources *113*

References *114*

List of Abbreviations

Acronym	Full Form	Acronym	Full Form
PSA	Prostate-specific antigen	IEDB	Immune epitope database
CRP	C-reactive proteins	MHC	Major Histocompatibility Complex
HGP	Human Genome Project	ANN	Artificial Neural Network
RF	Rheumatoid factor	AAC	Amino acid composition
ACPA	Anti-citrullinated protein antibodies	DPC	Di-peptide composition
GRP-78	Glucose-regulated protein 78	TPC	Tri-peptide composition
FTR1	High-affinity iron permease	BLAST	Basic local alignment search tool
CYP51	cytochrome P450	MERCI	(Motif-EmeRging and with Classes Identification)
GWT1	GPI-anchored wall transfer protein 1	MAST	Motif Alignment and Search Tool
ARP	ADP-ribosylation factor	LLM	Large language model
DHD	Dihydrolipoyl dehydrogenase	DL	Deep learning
CaN	Calcineurin	DT	Decision Tree
SAPs	Serine and Aspartate protease	RF	Random Forest
Cda	Chitin deacetylase	KNN	k-Nearest neighbour
BCGF-1	B cell growth factor-1	GNB	Gaussian Naïve base
BSF-1	B cell stimulatory factor -1	ET	Extra tree
DGP	Deamidated gliadin peptide	SVC	Support vector classifier

EMA	Endomysial antibody	1DCNN	1 dimensional Convolutional Neural Network
tTG	Tissue Transglutaminase	RRI	Repetitive Residue Information
GnRH	Gonadotropin-releasing hormone	AI	Artificial Intelligence
ARF	ADP-ribosylation factor	AABP	Amino acid binary profile
IL-4	Interleukin-4	AD	Autoimmune diseases
ACP	Anticancer peptides	NCBI	National Center for Biotechnology Information
RA	Rheumatoid arthritis	LR	Logistic Regression
CD	Celiac disease	CeTD	Composition enhanced Transition and Distribution

List of Figures

Figure 1.1: The organization of thesis chapters. 21
Figure 2.1: Timeline of fungal databases. 25
Figure 3.1: The disease mechanism of mucorales in healthy human hosts. 36
Figure 3.2: The structural organization of “Mucormydb”. 37
Figure 4.1:Overall computational framework of IL4pred2. 48
Figure 4.2: Preferential amino-acid positional analysis of human and mouse data. 49
Figure 4.3: Preferential positional analysis of residues in human vs mouse. 50
Figure 4.4: Compositional analysis of human and mouse data. 51
Figure 4.5: Compositional analysis between human and mouse. 51
Figure 5.1: Complete pipeline and workflow of the study. 71
Figure 5.2: The average amino acid compositional difference between anticancer, antimicrobial and random peptides. 73
Figure 6.1: The detailed model architecture of RAIpred. 83
Figure 6.2: The average amino acid compositional difference between RA-inducing and RA non-inducing peptides. 85
Figure 7.1: Workflow of the CDpred. 91
Figure 7.2: The Sequence analysis of CD-associated peptides. 94
Figure 7.3: The Compositional analysis of CD-associated peptides. 95
Figure 8.1: The web repository of MucormyDB. 103
Figure 8.2: The webserver of IL4pred2. 104
Figure 8.3: The webserver of Anticp4. 106
Figure 8.4: The web server of RAIpred. 107
Figure 8.5: The CDpred web server. 108

List of Tables

Table 2.1: List of available methods for IL4-inducing peptide prediction. 27
Table 2.2: List of available methods for anticancer peptide prediction. 28
Table 3.1: The highest ranking peptide-based therapeutic candidate. 39
Table 3.2: Top10 nucleotide based adjuvants predictive by VaccineDA. 41
Table 3.3: The highest-ranking RNA-based therapeutic candidates. 42
Table 4.1: Peptide distribution across datasets for human and mouse. 46
Table 4.2: The top five features with highest mean difference in both datasets. 52
Table 4.3:The top five LR-based features with maximum performance. 53
Table 4.4: The BLAST coverage on Human and Mouse datasets. 55
Table 4.5: Performance of ML classifiers developed using different features. 57
Table 4.6: The performance of Scikit based feature selection methods. 58
Table 4.7: The performance of Univariate feature selection methods. 59
Table 4.8: The performance of deep-learning based classifier on both data. 60
Table 4.9: The performance of LLM based classifier on both data. 62
Table 4.10: The performance of different benchmarking models on both data. 65
Table 4.11: The performance of human vs mouse models on both validation data. 66
Table 5.1: Performance of the mean-based univariate analysis. 74
Table 5.2: The performance of LR-based single feature analysis. 74
Table 5.3: The performance of best ML classifiers on both datasets. 75
Table 5.4: The DL model results on different sequence features. 77
Table 5.5: The benchmarking performance of existing ACP methods. 78
Table 6.1: The ML models performance using different features. 85
Table 6.2: The relevant motif list for RA-inducing peptides. 86
Table 6.3: The ensemble model performance using exclusive positive motifs. 86
Table 7.1: The number of specific HLA-associated peptides. 93
Table 7.2: The conserved motif list with their occurrences. 95
Table 7.3: Performance of PQ abundance method on different window sizes. 96
Table 7.4: The performance of ml classifiers on AAC based features. 97
Table 7.5: The cumulative coverage of motifs in positive sequences. 98

1. Introduction

1.1. Background

Biomolecules are chemical compounds that constitute the essential basis of life, ranging from the most primitive unicellular organisms to advanced multicellular mammals[1].These molecules are categorized into small and large molecules. Small molecules have a low molecular weight (< 1000 Da) and are hydrophobic in nature, which helps them to easily pass through the outer plasma membrane of a cell. Large biomolecules such as carbohydrates, proteins, lipids, and nucleic acids, are composed of the polymerization of hundreds or thousands of small molecular units[2]. These molecules play an essential role in metabolism, signal transduction, gene expression, cellular communication, immune response, and cellular homeostasis. These functions provide structural and functional support to cells and tissues by mediating biochemical reactions and pathways essential for life. Disruption or malfunctioning of these molecules leads to disease pathogenesis[3]. For example, mutations or abnormalities in protein post-translational modifications may cause the development of cancer, neurodegenerative disorders, metabolic syndromes, and other diseases[4].

Carbohydrates are the major energy source, and they aid in the performance of fundamental cellular functions. The modification of glycosylation patterns can cause autoimmune diseases, infections, and cancer[3]. Lipids are the major component of the cell membranes. They serve a significant role in cell signalling and as messenger molecule transmitting signals across cell surface receptors to intracellular targets. The malfunctioning of these signaling molecules causes inflammatory and cardiovascular diseases[5]. The hereditary molecules that store, express, and transmit genetic information are known as nucleic acids. These molecules are responsible for the regulation of gene expression through transcriptional and translational processes. Genetic or epigenetic changes during these processes may greatly influence human health and disease[3]. Together, these molecules can be used as disease markers and therapeutic targets and can aid in therapy, prognosis, and precision therapeutics[4]. In order to design novel diagnostic tools and specific therapies, one should understand their intricate interactions and their systemic evolution under disease conditions.

1.1.1. Protein and Peptides in Disease Diagnosis and Therapeutics

Of all biomolecules, the biological activity of proteins and peptides is highly specific and leads to fewer off-target effects and reduced toxicity. They are playing an important role in key pathological mechanisms, such as cancer proliferation, immune escape, angiogenesis, and metastasis[6]. They are used in different diseases such as cancer, infections, inflammation, autoimmune disorders, neurodegenerative disorders, metabolic disorders and cardiovascular diseases as biomarkers. Prostate-specific antigen (PSA) is a protein utilized in the diagnosis of prostate cancer[7]. C-reactive proteins (CRP) is utilized in the diagnosis of bacterial infection[8]. Antimicrobial peptides, such as cathelicidin LL-37 and human neutrophil peptides (HNPs), are found in the diagnosis of conditions such as RA and psoriasis, thus indicating that they can be used as diagnostic markers[9,10].

Besides diagnosis, proteins and peptides are also important in treatment measures. Receptor blockers (e.g., cetuximab) or carriers in antibody-drug conjugates (ADCs) containing cytotoxic agents are administered to tumor cells using monoclonal antibodies (mAbs)[11]. As targeted therapies are applied using peptides, they can selectively identify diseased cells to alter pathological pathways without affecting healthy tissues[12,13]. In clinical research, various therapeutic peptides demonstrate excellent outcomes as tumor-penetrating peptides, immune-modulatory peptides, and anticancer peptides[14]. Therefore, they are important biomolecules in the diagnosis of diseases as well as the development of therapeutics due to their specificity, functional diversity, and clinical relevance.

1.1.2. Transition to Computational Diagnostics and Therapy

The increasing number of diseases worldwide is further supports the need to identify efficient biomolecular markers and therapeutics[15]. Even with the growing need, proper diagnosis and effective therapeutic measures are still difficult to achieve. The classic methods of diagnosis and treatment mainly depend on experimental methods such as imaging, histopathology, culture-based assays, and biochemical tests. Although these approaches are effective in most instances, they are usually tedious, costly, and time consuming[16]. These methods are not practically viable for large scale screening particularly in resource-constrained environments. Moreover, they are not always able to offer early or predictive information regarding the development of a disease, prognosis, or response to treatment.

Computational biology and bioinformatics have been developed to overcome these constraints; as a result, cost effective, high throughput disease diagnosis and therapeutic solutions are being developed. Since the Human Genome Project (HGP) began, there has been an exponential rise in biological data. Computational biology and bioinformatics methods efficiently handle this large-scale data to obtain meaningful biological information[17]. Consequently, contemporary studies are shifting from traditional experimental methods to integrated methods that incorporate experimental validation followed by in silico analysis. This integrated method enhances precision, reduces cost and offers innovative diagnosis and treatment interventions within an effective timeframe.

1.2. Origin of the Proposal

The current work is designed in response to the emerging necessity of in-silico based solutions to the essential problems in disease diagnostics and treatment driven by biomolecules. A sudden outbreak of mucormycosis was witnessed during the COVID-19 pandemic. The information concerning this fatal fungal disease was distributed among various scientific publications and online materials. In response to this gap, it was necessary to compile disease-relevant molecular information and provide in-silico solutions by targeting proteins involved in host invasion and immune modulation.

Peptides also emerged as a significant class of drugs with diverse applications. At present, there are already 80 peptide-based drugs that have been approved by the FDA that testify to their increasing clinical utility. This increasing importance raises the need for effective in-silico methods to determine functionally significant peptides. Interleukin-4 surprisingly has a significant role in the regulation of immunity, particularly those associated with allergic inflammation, antibody production, and defense against parasites. There are few tools available for the identification of IL4-inducing peptides, but they lack host specificity. Anticancer peptides are also promising alternatives to treatment, but current prediction models were developed using limited and outdated experimental data. Such gaps can be filled by developing precise, reliable, host specific approaches using recently generated experimentally validated data.

In spite of the tremendous advances in approaches to the prediction of therapeutic peptides, notable gaps exist in the detection of potentially harmful peptides that can cause detrimental immune responses. Viral peptides appearing as analogues of host proteins, e.g. Epstein-Barr virus heat-shock fragments, may avoid immune surveillance and cause autoimmune-like diseases. Food peptides like gluten also induce inflammatory ailments in those who are

genetically predisposed. The detection of these pathogenic peptide signatures is thus indispensable in disease prevention, therapeutic design, as well as biomedical studies. The collective challenges provide the basis of the current study, that seeks to consolidate disease-specific molecular information and further develop the prediction of peptides.

1.3. Objective of Thesis

This thesis focuses on the development of in-silico tools such as web repositories and prediction methods to enhance biomolecule-based disease diagnosis and therapy. The study is divided into two major sections, each addressing specific biomedical challenges with the help of machine learning and artificial intelligence.

The initial theme is concerned with disease diagnosis and involves developing a comprehensive web repository that consolidates genomic and proteomic information. The aim of this online resource is to facilitate knowledge among clinicians and researchers about disease mechanisms and the identification of potential treatment targets. The necessity of such resources was proved by the sudden increase in the number of mucormycosis cases during the COVID-19 pandemic. The lack of sufficient molecular information and a specific online portal inhibited early diagnosis and management of patients in time. To address this gap, MucormyDB was developed, a web-based repository of information relating to genomic, proteomic, pharmacological, immunotherapy, and antifungal drugs. It helps to thoroughly explore Mucorales organisms and promotes genome-based therapeutic approaches.

The next section “Biomolecule driven therapy” is further divided into two major objectives. The first object is focused on developing in-silico tools to predict the therapeutic potential of peptides based on their sequences and physicochemical characteristics. In this section, we aim to develop the most reliable tools for interleukin-4 and anticancer peptides. The second objective is aimed at addressing an often overlooked aspect of peptide research, "the potential harmful effects of specific peptides". In this study, we aimed to develop in-silico tools to predict peptides or epitopes responsible for causing autoimmune diseases like Rheumatoid arthritis and Celiac disease. The combined study provides integrated computational solutions that support biomolecule-driven approaches for both disease diagnosis and therapeutic development.

1.4. Organization of Chapters

The current research addresses the major challenges of disease diagnostics and biomolecule based therapeutics by developing integrated in-silico tools and web-based solutions that are easy to use. The created tools and resources are openly available via web servers, GitHub, PyPI packages, and Python-based standalone versions so that the scientific community can use them widely. The thesis is separated into eight chapters that describe various aspects of the study.

Chapter 1 is dedicated to the background, importance, requirement, organization, and general structure of the thesis. The chapter provides the scientific and computing foundation of all the other chapters.

Chapter 2 gives detailed literature review of biomolecules, peptide-based diagnostics, and therapeutics. It also covers the current resources available for targeted diseases and peptides of this study. The chapter identifies some important weaknesses of the current tools and serves as a motivation for the further development of the computational models and integrative web-based resources, which can be viewed as a rationale and motivation for the current research.

Chapter 3 explains the first objective of the study titled "Computational Resource for Fungal Disease". This chapter explains mucormycosis in detail along with its epidemiology, pathophysiology, and clinical significance. The chapter also covers the data collection, compilation, and top predicted immunotherapeutic peptides along with development of a web-repository.

Chapter 4 provides the development approach and resulting framework for in-silico prediction of a host-specific IL4-inducing peptide tool. This chapter comprises the importance of IL4 in disease and therapy, the computational approach, sequence based analysis and cross host model comparison.

Chapter 5 presents AntiCP4, a next-generation tool for anticancer peptide prediction. This tool also employs AI/ML-based classifiers, motif-based pattern recognition, and ensemble approaches to achieve high prediction accuracy.

Chapter 6 describes the development of a computational method designed to identify peptides associated with rheumatoid arthritis. This chapter comprises a detailed discussion of the dataset collection, feature extraction, AI model development, and the establishment of online and Python-based platforms.

Chapter 7 discusses the utilisation of the CDpred framework to identify regions in proteins associated with celiac disease.

Chapter 8 summarizes the findings, contributions, and outlines future directions for computational biomolecule-based diagnostics and therapeutics. The overall structure and relationship between the goals and developed resources are schematically depicted in **Figure 1.1.**

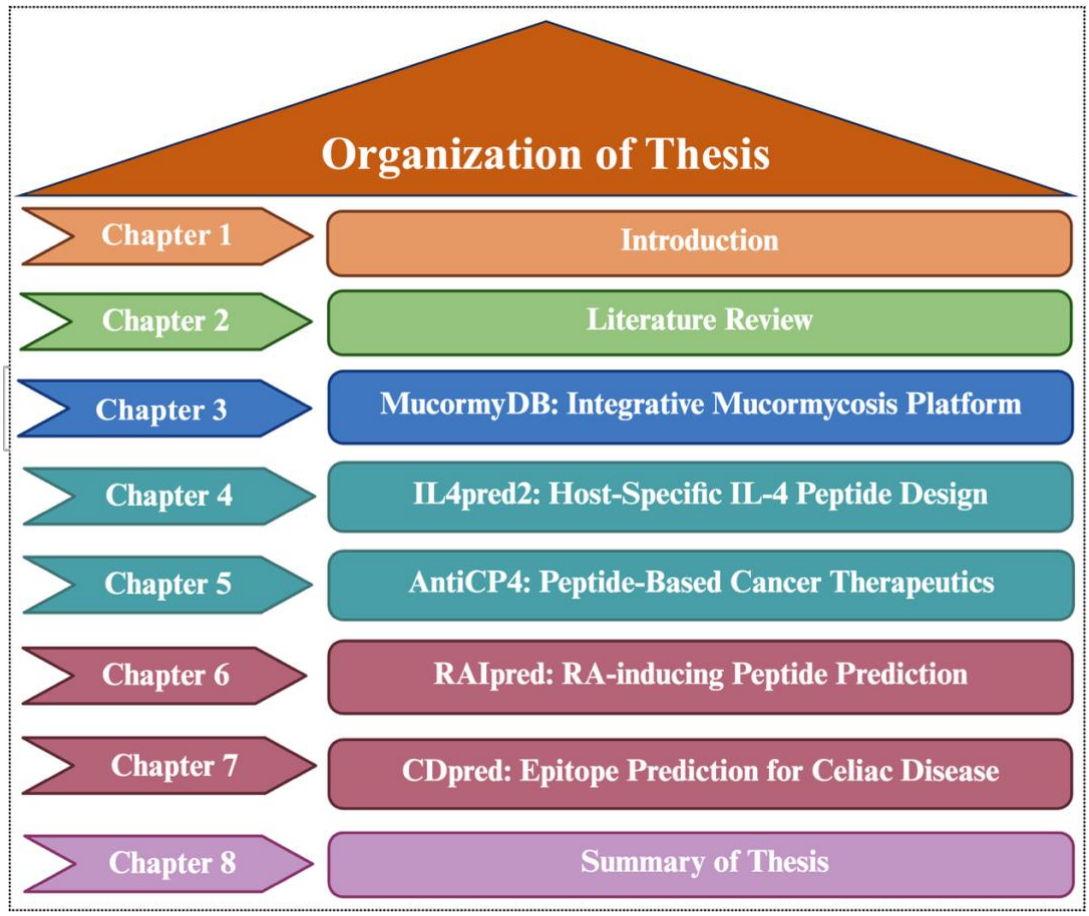

Figure 1.1: The organization of thesis chapters.

2. Review of Literature

2.1. Overview

The cell is the basic structural and functional component of all organisms, both unicellular microorganisms and multicellular ones with great complexity. There is an organized biochemical process within each cellular entity that nourishes life, controls metabolic pathways, and provides maximum functionality of the cellular entity. Biomolecules, proteins, peptides, nucleic acids, lipids, and carbohydrates work to regulate these processes and ensure that cells stay healthy[18]. Of special interest, proteins play a central role due to their variety of structural, catalytic, transport, and regulatory roles[19]. The amino acid sequence of a protein determines its three-dimensional conformation and the resulting functional characteristics. Proteins are used as enzymes in important biochemical reactions, as receptors in cellular communication, as well as in the translocation of molecules across membranes. They also give structural support to cells and tissues, serve as hormones that coordinate physiological activity, and serve as antibodies against pathogenic agents[20,21]. In their complex reactions with other biomolecules, proteins form networks of interactions that support the molecular communication of the cell and its general homeostasis[22].

2.2. Peptides as Regulatory and Therapeutic Molecules

Peptides that are composed of short chains of amino acids have a broad variety of applications as regulatory and signalling molecules. The different subclasses of peptides such as hormones, cytokines, immunomodulators, and antimicrobial peptides have indispensable roles in homeostasis and immune equilibrium of cells[23]. The major immune modulators that regulate inflammation and host defense mechanisms include cytokine peptides, such as interleukin 4 (IL-4) and interleukin 10 (IL-10)[24,25]. Abnormal regulation of these peptides may cause immune homeostasis disorders and contribute to the occurrence of inflammatory or autoimmune diseases. In pathologies like celiac disease and rheumatoid arthritis, the modulation of the interactions between peptide and proteins or peptide and MHC has been involved in the generation of pathogenic immune responses[26,27]. Peptides also have significant therapeutic potential besides being regulators in intercellular communication and immune modulation.

Peptides are a strong class of next-generation therapeutics because of their accuracy, biocompatibility, and ability to regulate complex biochemical functions. The unique

specificity and structural flexibility of these peptides allow them to interact with protein-protein interactions that are inaccessible to small-molecule agents, and hence place them at the interface between chemical and biological therapeutics[14]. Several peptide therapeutics have been approved by the FDA and/or are undergoing clinical trials, such as insulin, glucagon-like peptide-1 analogs, and antimicrobial peptides[28]. The development of novel therapeutic compounds with high efficacy and low adverse reactions requires a thorough knowledge of the sequence-function relationship of these peptides. However, protein and peptide properties are difficult to characterize experimentally, as they are too complex biologically. Thousands of proteins are expressed in every organism with dynamic post-translational modifications, interactions, and conformational states that change over time and context[29]. Genetic mutations and post-translational modifications and aberrant expression patterns are often pathological factors triggering the activation of diseases[4]. Mutations affecting the active site of an enzyme or the disruption of signal proteins can lead to the development of oncogenesis or metabolic disorders[30,31]. Pathogenic proteins can likewise enhance disease, including mucormycosis, by evading host defenses with virulence factors [32]. Therefore, the study of these molecular pathways requires thorough knowledge of how to develop effective therapeutic and diagnostic interventions.

2.3. Computational Biology and AI in Molecular Research

The recent developments in proteomics, structural modeling, and genomic sequencing provide large datasets that clarify biomolecular interactions, activities, and sequences. Such datasets can be easily analyzed in the framework of computational biology and bioinformatics and thus help identify biomarkers, functional residues, and bioactive peptides by using predictive models. In addition, machine learning and artificial intelligence currently allow classification of peptide activity, prediction of binding affinity, and discovery of therapeutically relevant peptide sequences[33–35]. These classifications, in turn, make the accurate identification of disease-related entities and the rational design of therapeutic candidates more feasible by combining biological information with computational approaches. The fine insight into biomolecular interactions does not only help to decipher disease pathophysiology but also forms the foundation of successful disease management. Early and accurate diagnosis methods determine molecular changes that trigger pathological states, thus enabling early intervention. As a result, the initial therapeutic approaches, such as peptide-based drugs, are designed to reverse or repair such changes and reestablish normal physiological activity[36,37].

2.4. Disease Diagnosis

2.4.1. Role of Computational Databases in Disease Identification

Genomic and proteomic technologies have increased exponentially, generating a large amount of molecular information with respect to viral, microbial, and human diseases[38]. Due to their size and complexity, conventional experimental analysis cannot fully represent the complex interactions between disease manifestations and molecular entities. Computational biology and bioinformatics are therefore now vital tools for identifying biomarkers, predicting molecular interactions, and designing effective diagnostic systems[39,40]. Because web-based resources enable open-access platforms for the study, visualization, and interpretation of biomolecular data relevant to disease, they have revolutionized biomedical research.

A variety of specialized resources such as CoronaVIR, ZikaVR, VirHostNet, CoronaVR, and ViPR have been developed to aid disease-specific research, combining genomic, proteomic, transcriptomic, and metabolomic datasets with pertinent clinical and pharmacological information[41–45]. Such resources provide a centralized interface to explore pathogen genomes, virulence factors, host-pathogen interactions, immune epitopes, and potential drug targets. The development of these resources helps both researchers and clinicians understand the disease mechanism, previous studies, known symptoms, available therapeutics, and further details about the disease. Mucormycosis, among the several infectious diseases, has received a considerable amount of attention because of its rapid progression and high mortality rate, especially in immunocompromised and post-COVID-19 patients[46]. The infection is caused by fungi of the order Mucorales and remains difficult to diagnose and treat because of limited molecular data and resistance to existing antifungal drugs[47].

2.4.2. Fungal Disease Repositories and Evolution

Various fungal databases have been developed in the past, that consolidate genomic, proteomic, and functional information for pathogenic and non-pathogenic fungi. These repositories have significantly advanced fungal genomics, comparative analyses, and molecular biology research. The e-Fungi database developed in 2007, was one of the earliest integrative fungal resources, offering genome data, functional annotations, and comparative analyses for over 30 fungal species, including pathway and clustering information[48]. Subsequently, CandidaDB was established in 2008 to focus specifically on Candida albicans

and related CTG-clade species, supporting comparative genomic analyses across this medically important group[49]. With development support from Stanford and the Broad Institute, the Aspergillus Genome Database (AspGD), which was launched in 2014, offered a consolidated repository for genomic and functional annotations of Aspergillus species[50]. DemaDb (2016) offered full genomic profiles and metadata, including taxonomy, assembly statistics, and morphological information, for dematiaceous (darkly pigmented) fungi associated with human disorders[51]. FungiDB (2018) broadened the field by combining proteomic, transcriptomic, and genomic datasets for a wide range of fungal illnesses and non-pathogenic fungi[52]. It became part of a broader initiative aimed at developing organism-specific databases. More recently, in 2025 PHI-base has emerged as a comprehensive database for pathogen-host interactions, covering a wide spectrum of microbial and fungal pathogens[53]. The Candida Genome Database (CGD) (2025) further strengthens the fungal data landscape by providing NIH-supported, actively curated genomic and functional data for Candida species[54]. The generation of fungal specific databases and web repositories is shown in **Figure 2.1**.

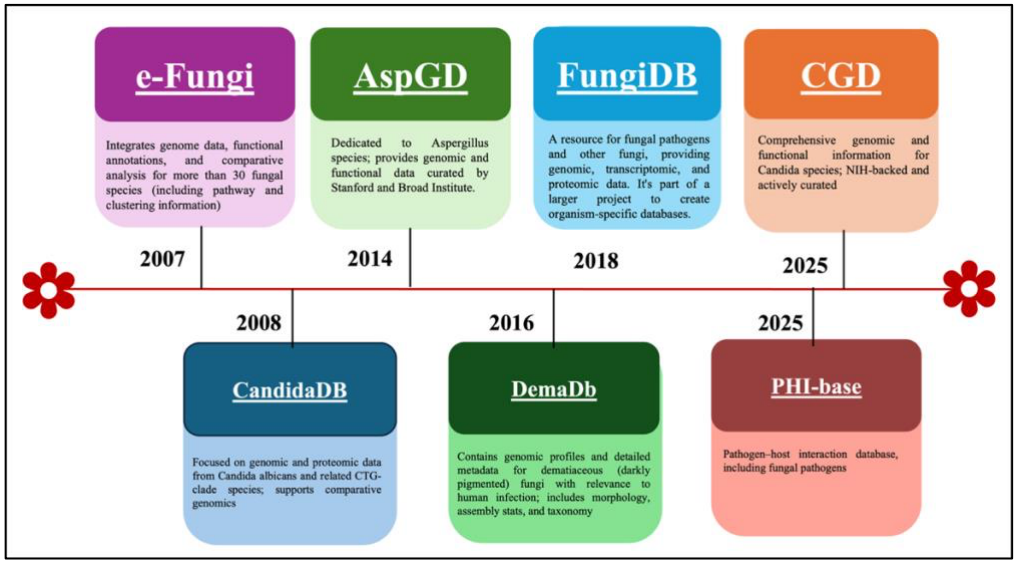

Figure 2.1: Timeline of fungal databases.

Although mucormycosis is clinically significant, no web-based repository has so far been designed that could bring together genomic, proteomic, and therapeutic information in a single platform. The available resources are also incomplete and usually target single datasets or narrow strain data. These limitations inhibit the comprehensive understanding of host-pathogen interactions and drug resistance. To fill this void, we have created a specific web platform that gathers the molecular, clinical, and therapeutic information regarding the

disease in a systematic way. This initiative is intended to expedite scientific studies and enhance translational progress in the field of mucormycosis biology and treatment.

2.5. Biomolecule-Based Therapeutics

Biomolecule-based therapeutics are designed using different biological molecules, including proteins, peptides, and nucleic acids. These are more specific, effective, and safe compared to traditional small-molecule drugs[55]. They mimic or alter normal physiology, are less associated with side effects, and are more specific to treatment. The possibility of using peptides as therapeutic agents is justified by their selective interactions with biological receptors and enzymes, their structural heterogeneity, and biocompatibility[56]. They are also promising new-generation therapeutics because of their various biological activities, including immunomodulatory, antiviral, antibacterial, and anticancer activities[57].

2.5.1. Interleukin-4 (IL-4) Inducing Peptides

IL-4 is a pivotal cytokine that drives T-helper type 2 (Th2) immune responses, stimulates antibody production, and contributes to maintaining immune homeostasis[58]. Peptides capable of inducing IL-4 secretion are therefore of great interest in immunotherapy and vaccine design, where controlled modulation of immune responses is essential[59]. Based upon the essential function of IL-4 in immune regulation and involvement in a variety of disease processes, several computational approaches have been developed to identify IL-4-inducing peptides.

IL4pred was the first web-based tool that used experimentally validated datasets and machine learning models based on dipeptide composition and motif features to distinguish IL-4-inducing from non-inducing peptides[60]. In order to enhance predictive performance, Meta-IL4 (2023) later used ensemble learning by combining several classifiers; nevertheless, it is only available through an open-source repository[61]. The latest approach, PLM-IL4 (2025), achieves higher accuracy and scalability by automatically extracting deep contextual information using transformer-based protein language models[62]. Despite these advancements, there are still certain drawbacks, such as a lack of host-specific prediction accuracy, limited interpretability, and a restricted dataset diversity, because the majority of current models are generic and do not take into consideration differences in IL-4 induction among various host systems. To improve the biological relevance and translational potential of IL-4 peptide prediction, it will be crucial to incorporate host-specific data and

immunological background in addition to experimentally verified datasets. **Table 2.1** provides a list of available methods for predicting IL-4 inducing peptides.

Table 2.1: List of available methods for IL4-inducing peptide prediction.

Tool	Description	Year	Working Status
IL4pred	A machine-learning-based tool designed to predict peptides IL-4-inducing peptides. (https://webs.iiitd.edu.in/raghava/il4pred/)	2013	Yes
Meta-IL4	A meta-prediction tool that integrates multiple models to accurately identify IL-4-inducing peptides. (https://github.com/MirTanveer/Meta-IL4)	2023	No
Plm-Il4	A prediction tool that uses protein language models to accurately identify IL-4-inducing peptides. (http://www.bioai-lab.com/PLM-IL4)	2025	Yes

2.5.2. Anticancer Peptide (ACP) Prediction

Anticancer peptides (ACPs) have been identified as a very potent therapeutic agent that can selectively attack malignant cells and spare normal tissues [63]. They affect the tumor cell membranes, influence oncogenic signaling pathways, prevent angiogenesis, and disrupt tumor cell membranes. These properties make ACPs viable substitutes to traditional chemotherapeutic over drugs that often have low selectivity and high toxicity [64]. A wide range of computational methods has been designed to identify and characterize ACPs using physicochemical properties, sequence-derived properties, and advanced machine-learning models [65–114].

The first computational aids, AntiCP (2013), were based on the maximization of binary feature profiles and amino-acid composition to differentiate between ACPs and non-ACPs [115]. Later models, iACP (2016), MLACP (2017), and ACPred-FL (2018) added more advanced encoding schemes and the use of an ensemble of classifiers to improve predictive performance [116–118]. Even though several of these servers are now outdated, more recent ones, such as ACP-MHCNN (2021), xDEEP-AcPEP (2021), and MLACP 2.0 (2022), use convolutional neural networks and deep-learning frameworks to learn hierarchical sequence motifs, which has led to higher generalization and accuracy [119–121]. Newer models such as AntiCP 2.0 (2021) and ACP-EPC (2025) have also been created based on the combination of ensemble methods, contextual embeddings, and evolutionary information to screen ACP

at large scale [122,123]. However, biases pertaining to old datasets, lack of independent validation, and limited biological interpretability still limit the interpretability of most models. Future ACP prediction models should also combine experimentally validated data, structural, and functional annotations to improve validity and clinical utility. A summary of existing ACP prediction methods is given in **Table 2.2**.

Table 2.2: List of available methods for anticancer peptide prediction.

Tool	Description	Year	Working Status
AntiCP	A machine-learning tool designed to identify anticancer peptides from amino acid-based features. (webs.iiitd.edu.in/raghava/anticp)	2013	Yes
Hajishrafi et al	A predictive model that classifies anticancer peptides using PseAAC and alignment based kernel learning.	2014	No
iACP	A computational method that predicts ACPs by leveraging optimized g-gap dipeptide patterns. (http://lin.uestc.edu.cn/server/iACP)	2016	No
MLACP	An integrated ML framework combining multiple sequence and physicochemical descriptors to detect ACPs. (http://www.thegleelab.org/MLACP/MLACP.html)	2017	No
iACP-GAEnSc	The identification of anticancer peptides using evolutionary intelligent genetic algorithm-based ensemble model.	2017	No
ACPred-FL	A feature learning based approach that identifies ACPs from diverse sequence descriptors. (http://server.malab.cn/ACPred-FL)	2018	No
TargetACP	A model that predicts anticancer peptides using evolutionary profiles and structure-related attributes.	2018	No
ACP-DL	A deep learning classifier that uses binary profiles and reduced alphabet k-mers to detect ACPs.	2019	No
mACPpred	A machine learning model integrating sequence and physicochemical signatures for ACP prediction.	2019	No
PEPred-Suite	PEPred-Suite is a bioinformatics tool for the generic prediction of therapeutic peptides. (http://server.malab.cn/PEPred-Suite)	2019	No
PTPD	A computational pipeline that uses deep learning and word2vec to predict therapeutic peptides, including those with anticancer properties.	2019	No
ACPred	An an interpretable tool for the prediction and characterization of the anticancer.	2019	No

	(https://codes.bio/acpred/)		
ACP-GCN	Graph based deep learning model for predicting anticancer peptides (ACPs).	2020	No
DeepACP	Sequence-based recurrent neural network (RNN) model to predict the likelihood of a given peptide sequence exhibiting anticancer activity.	2020	No
CancerGram	Tool designed to computationally predict anticancer peptides. (http://biongram.biotech.uni.wroc.pl/CancerGram/)	2020	No
ACPred-Fuse	A fused-feature ML predictor incorporating amino acid, dipeptide, and physicochemical encodings. (http://server.malab.cn/ACPred-Fuse)	2020	No
Li & Wang's Method	A classifier built on refined physicochemical property profiles to detect ACPs.	2020	No
ACPP	A composition based tool that predicts anticancer peptides using amino-acid and physicochemical traits. (http://acpp.bicpu.edu.in/predict.php)	2021	No
ACP-MHCNN	A multi-head CNN architecture that integrates sequence and physicochemical maps for ACP prediction. (https://anticancer.pythonanywhere.com/)	2021	Yes
iACP-DRLF	Identify anticancer peptides via deep representation learning features. (http://public.aibiochem.net/iACP-DRLF/)	2021	No
ACP-DA	A ML model uses data augmentation for insufficient samples to improve the prediction performance of anticancer peptides.	2021	No
AntiCP 2.0	An improved method to predict anticancer peptides using amino acid and dipeptide composition with improved algorithmic tuning. (webs.iiitd.edu.in/raghava/anticp2)	2021	Yes
ENNAACT	An Ensemble neural network based tool to predict anticancer properties of peptides. (https://research.timmons.eu/ennaact)	2021	No
xDEEP-AcPEP	Deep neural architecture based tool for quantitative prediction of anticancer peptide activity against specific cancer cell lines. (https://app.cbbio.online/acpep/home)	2021	Yes
Zhao et al	CNN-based model using one-hot and embedding representations to predict anticancer peptides	2021	No
StackACPred	Stacked ensemble model developed to predict anticancer peptides by combining multiple sequence features.	2022	No

ME-ACP	Multi-view neural networks with Ensemble model for identification of anticancer peptides	2022	No
ACP-check	The ACP-check combines the Bi-LSTM network and a fully connected network to predict ACPs. (http://www.cczubio.top/ACP-check/)	2022	No
AntiMF	A deep learning model that utilizes multi-view mechanism to predict anticancer peptides.	2022	No
MLACP 2.0	Enhanced ML model with optimized composition and property-based features to predict anticancer peptides. (https://balalab-skku.org/mlacp2/)	2022	Yes
iACP-RF	Ensemble Random Forest with composition and evolutionary features to predict anticancer peptides.	2023	No
GRDF	Graph-based residue distance features integrated with deep learning models to classify anticancer peptides.	2023	No
TriNet	A tri-fusion neural network to accurately predict anticancer and antimicrobial peptides. (http://liulab.top/TriNet/server)	2023	No
ACPs-ASSF	Improve anticancer peptide (ACP) prediction method by intelligently selecting high-quality data from augmented datasets.	2023	No
ACP-BC	A deep learning model for predicting Anticancer Peptides by combining sequence-based features with handcrafted features in a three-channel system.	2023	No
ACP-MLC	two-level framework for identifying anticancer peptides and classifying their functional roles.	2023	No
ACP-GBDT	A computational method for identifying potential anticancer peptides by using the GBDT algorithm.	2023	No
LGBM-ACp	LightGBM-based model to identify potential anticancer peptides from large datasets	2024	No
Annprob-acps	Artificial neural network using probabilistic feature fusion approach to rapidly and accurately identify potential anticancer peptides. (https://circular-palatable-term.anvil.app/)	2024	No
CAPTURE	An advanced AI-driven platform for predicting anticancer peptides. (https://sds_genetic_analysis.opendfki.de/CAPTURE/Predict/)	2024	No
ACP-PDAFF	A cutting-edge computational model for predicting Anticancer Peptides by integrating powerful Pretrained language models (like ProtBert) with Dual-channel Attentional Feature Fusion (DAFF).	2024	No

ACP-CapsPred	An advanced, explainable AI framework for identifying and predicting the anticancer activity of peptides	2024	No
ACPPfel	An advanced deep ensemble learning model designed to accurately predict anticancer peptides. (http://lmylab.online:5001/)	2024	No
MA-PEP	A cutting-edge deep learning framework for predicting anticancer peptides by fusing sequence and chemical features using multiple attention mechanisms.	2024	No
ACP-LSE	A powerful computational method used for predicting potential anticancer peptides from protein sequences.	2024	No
ACPScanner	An advanced computational tool for predicting anticancer peptides and their specific functions. (http://acpscanner.denglab.org/)	2024	No
AACFlow	A cutting-edge deep learning model designed to accurately predict anticancer peptides by extracting complex sequence features using an Attention Augmented Convolutional Neural Network (AAConv) and a novel flow-attention mechanism.	2024	No
mACPpred 2.0	An advanced machine learning tool for predicting anticancer peptides by using a stacked deep learning approach. (https://balalab-skku.org/mACPpred2/)	2024	No
PLMACpred	ML-based tool designed to predict anticancer peptides directly from their amino acid sequences.	2024	No
ACP-ML	Machine learning-based model designed to classify anticancer peptides.	2024	No
ACP-DRL	A computational method using advanced deep learning, especially protein language models and Bi-LSTM, to predict anticancer peptides from sequences.	2024	No
EnsemPred-ACP	An advanced computational framework that uses a blend of machine learning (ML) and deep learning (DL) to accurately predict anticancer peptides (http://www.thegleelab.org/EnsemPred-ACP/)	2025	No
ACP-CLB	An advanced deep learning model for predicting anticancer peptides.	2025	No
ACP-ESM2	Utilizes pretrained ESM2 protein embeddings for predicting anticancer peptides. (http://www.bioai-lab.com/ACP-ESM2)	2025	No
ACP-DPE	Deep learning framework to predict anticancer peptides.	2025	No
iACP-DPNeT	Deep parallel network used for the computational prediction of anticancer peptides.	2025	No

pACP-HybDeep	An advanced AI model for accurately predicting anticancer peptides by blending deep learning.	2025	No
pACPs-DNN	Deep neural network based model for anticancer peptide prediction.	2025	No
ACP-EPC	A cutting-edge deep learning framework for predicting anticancer peptides from protein sequences. (http://www.bioai-lab.com/ACP-EPC)	2025	Yes

2.5.3. Limitations in Existing Computational Approaches

There are a number of computational tools which can be used to predict IL-4-inducing peptides and anticancer peptides. Nonetheless, most of these tools are limited in such a way that their accuracy and reliability are diminished. One of the biggest problems is the reliance on older or smaller datasets, which are not able to represent the entire range of peptide sequences. Their practical application and reproducibility are further limited because of poor validation, lack of benchmarking, and the absence of user-friendly interfaces. Moreover, there are web servers that become unavailable or are no longer supported, making large-scale analysis challenging. These shortcomings explain why the performance, interpretability, and usability of peptide prediction tools should be improved. Curated datasets, sophisticated feature representations and AI-based algorithms should be introduced as next-generation techniques. In addition to precise prediction, immune compatibility and safety should be assessed, as they are necessary to determine undesirable immune responses and determine therapeutic suitability.

2.6. Safety assessment of peptides

2.6.1. Immunogenic and Autoimmune Reactions Triggered by Peptides

Peptide safety assessment is a vital part of biomedical and biotechnological studies. Even naturally occurring or bioactive peptide can result in unpredictable biological or immunological responses [124]. Activation of immune cells, antigenic mimicry, or interaction with host-proteins may result in hypersensitivity, inflammation, or autoimmune reactions caused by peptides [125,126]. These negative outcomes are commonly caused by improper post-translational modifications, unexpected molecular interactions, or structural homology to pathogenic or host-derived epitopes [127]. Peptides derived from gluten are a thoroughly-reported paradigm of abnormal immune system regulation and increased production of

cytokines in response to particular MHC class II alleles[26]. The effects of these peptides trigger an immune reaction by causing inflammation and damage to the intestinal epithelium. The subsequent pathology is celiac disease, an autoimmune disease with chronic intestinal injury and malabsorption[128]. Another example is heat shock proteins (HSPs) of the Epstein Barr virus that have a strong sequence homology with human proteins. Viral HSPs attach to MHC class II alleles and activate autoreactive $CD4^+$ T lymphocytes[129,130]. This activates B cells to produce autoantibodies such as rheumatoid factor (RF) and anti-citrullinated protein antibodies (ACPAs)[131,132]. Rheumatoid arthritis (RA), a chronic inflammatory disease marked by ongoing joint inflammation and gradual tissue loss, is mostly caused by this molecular mimicry and the immune activation that follows[133].

2.6.2. Computational Models for Immunological Safety Prediction

To avoid unwanted immune or autoimmune reactions, peptides must therefore pass a thorough immunological safety assessment prior to being used in industrial, therapeutic, or diagnostic applications[134]. In this regard, computational screening and predictive modelling have become essential instruments for detecting peptides that may have autoimmune or immunogenic characteristics. This makes it easier to choose and produce safer peptides for a variety of biological and pharmacological uses. Despite significant breakthroughs in peptide-based therapeutics and immunoinformatic, no specialised computational methods have been designed to predict peptides that have potential to cause autoimmune responses in healthy individuals with specific genetic or environmental trigger. This gap emphasises the importance of specialised models to identify and assess peptides associated with autoimmune disorders including rheumatoid arthritis and celiac disease.

2.7. Conclusion

The recent developments in the area of bioinformatics and computational biology have revolutionized the study of biomolecules and increased their potential in therapeutic development. There has been improvement in the process of peptide prediction, integration of disease specific data, and molecular modelling. Nonetheless, there are still some significant challenges. There is a lack of accuracy and generalizability of current computational tools in predicting peptide-based therapies. This limitation can commonly be caused by small or incomplete datasets, use of traditional algorithms, and low consistency in validation practices. Secondly, numerous existing web servers are obsolete, non-operational, or have complicated interfaces. These problems also prevent their application

in biology or clinical research on a large scale. Another major gap is the lack of specific methods to evaluate the immunological safety and autoimmune potential of peptides. Although peptide mimicry has been well characterized to produce immunogenic effects, there is currently no specialized computational technique to identify or assess peptides that can cause autoimmune responses. This limitation highlights the need for precise, interpretable, and experimentally validated models that can detect potentially harmful sequences before they advance to clinical use.

Even though a number of fungal and peptide databases have enhanced our knowledge of pathogen biology and biomolecule-based therapy, the existing data is still scattered. There is no single comprehensive platform that incorporates genomic, proteomic, and therapeutic data regarding diseases like mucormycosis. The development of a single, data-driven repository would increase the identification of novel diagnostic and treatment approaches as well as reinforce research on pathogen biology.

The literature review sheds light on the existing development and defines gaps in the area of biomolecule-based diagnostics and therapies. The development of advanced AI-driven frameworks and web-based services ought to become a critical requirement of future research. Along with enhancing predictive power, such tools should guarantee clinical safety and biological pertinence, thus contributing to the translational applicability of computational results and supporting the creation of the next-generation therapy of complex and emergent diseases.

3. Compilation of resources for Mucormycosis diagnosis

3.1. Introduction

Mucormycosis is an opportunistic fungal infection that emerged during the second wave of the COVID-19 pandemic, leaving a severe impact on the health of affected individuals[135]. Although it is not the first time, cases of mucormycosis have been identified. It was first reported in 1855 and then in 1876 in a cancer patient[136]. During the COVID-19 pandemic, a sudden rise shocked the community with an almost 85% mortality rate [137]. The primary reason behind this unexpected surge in mucormycosis cases was the immunocompromised state of patients due to COVID-19[138]. This unexpected increase was also caused by overcrowded hospitals, a shortage of medical resources, overworked healthcare professionals, and inadequate diagnostics[139]. It is caused by exposure to a variety of saprophytic fungi of the Mucorales order, including Rhizopus, Lichtheimia, Mucor, and Rhizomucor species, with the first three accounting for three-fourths of cases[140]. The mortality and morbidity rates of mucormycosis vary depending on the site of infection, ranging from 46% to 96%[141]. The infection is categorized based on the infected areas, such as Rhino-orbital-cerebral mucormycosis, which infects the sinus and nearby tissues, pulmonary mucormycosis, which infects the lungs, cutaneous mucormycosis, which infects the skin, gastrointestinal mucormycosis, which mainly infects the stomach, followed by colon and ileum, and disseminated mucormycosis, which spreads throughout the circulatory system, affecting multiple organs and resulting in serious complications[142–146].

Mucormycosis pathogenesis involves a complex interplay between a susceptible host environment and the aggressive virulence factors of the causative fungi (order Mucorales) [147]. In a healthy individual with intact mucosal/skin barriers and full innate immunity, spores entering the body are typically handled effectively; however, conditions such as neutropenia, immunosuppression, uncontrolled hyperglycaemia, ketoacidosis and elevated free iron levels significantly compromise host defences[148]. Under such circumstances the fungi exploit the impaired barrier and immune systems to germinate, adhere to and invade endothelial cells, leading to angioinvasion and rapid tissue necrosis[149].

At the molecular level, Mucorales express spore-coat homolog proteins (CotH) that bind to the host endothelial cell receptor GRP78 (glucose-regulated protein 78). GRP78 is a

receptor whose surface expression is up-regulated in conditions of hyperglycaemia, acidosis or iron overload[32]. This binding facilitates fungal endocytosis into and damage of endothelial cells, thereby promoting angioinvasion, thrombosis, infarction and dissemination into adjacent tissues[150]. Concurrently, Mucorales possess potent iron-assimilation mechanisms, for example, high-affinity iron permeases (FTR1), reductases, and siderophore receptors (Fob1/Fob2) that enable the fungus to access iron liberated in the host milieu (especially in diabetic ketoacidosis or deferoxamine therapy) to fuel its rapid growth. The result is fulminant vascular invasion, tissue infarction, and necrosis, which explains the high mortality and destructive nature of the disease[151,152]. The detailed disease mechanism of mucormycosis in the host can be seen in **figure 3.1.**

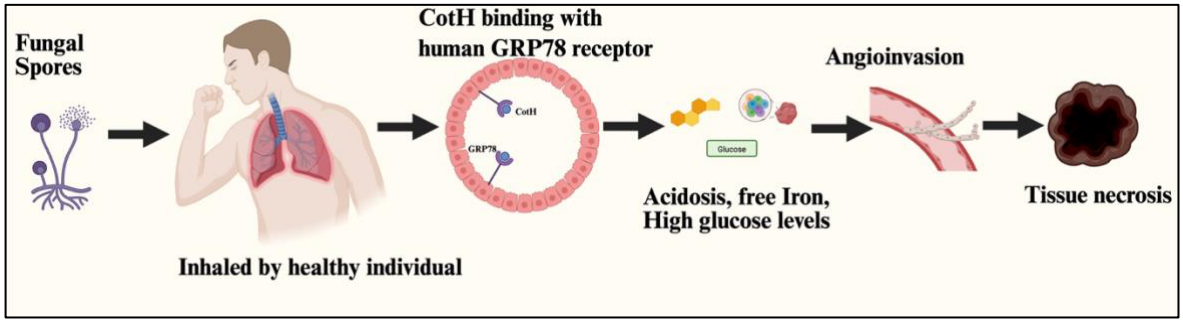

Figure 3.1: The disease mechanism of mucorales in healthy human hosts.

Early detection of mucormycosis is beneficial for improved understanding and management of the condition. Antifungal drug and surgical debridement are widely used therapies for mucormycosis[153]. Acquired resistance to drugs is a problem, though[154]. Therefore, in order to produce safer and more efficient treatment strategies, it is necessary to identify and assess newer molecular targets. The establishment and promotion of open-source materials is necessary to swiftly find potential therapeutics for a life-threatening condition. These materials can help experimental researchers and medical experts find possible solutions for any pandemic or epidemic. To overcome this gap, we have developed a platform to give computer-aided solutions for mucormycosis management in this study. This resource was created with the most up-to-date in-silico available approaches. It has been observed that Rhizopus oryzae is the most common cause of mucormycosis and is found in approximately 70% of cases,[141,155,156] and the remaining ~30% are caused by other members of the order Mucorales. These include less frequently isolated species from the family Mucoraceae, such as Rhizopus microsporus (including R. microsporus var. rhizopodiformis), Lichtheimia (Absidia) corymbifera and L. ramosa, Mucor species (e.g., Mucor circinelloides), Rhizomucor pusillus, and Apophysomyces species (e.g., Apophysomyces elegans and A.

variabilis). In addition, an increasing number of mucormycosis cases have been attributed to Cunninghamella species, particularly Cunninghamella bertholletiae (family Cunninghamellaceae). Rare case reports have also documented infections caused by species belonging to other families within Mucorales, including Saksenaea vasiformis, Basidiobolus ranarum, and Conidiobolus coronatus. Collectively, these organisms account for the remaining proportion of mucormycosis cases reported in the literature[148,155–157]. Thus, in this study, we have used *Rhizopus oryzae* (*R. oryzae*) proteins for the prediction of therapeutic peptides.

3.2. Database Design

The repository contains all important information about the disease, arranged into three main modules. The "Genomic and Proteomic Module" includes data on all known Mucorales species, including whole-genome, protein, and nucleotide sequences. The "Drugs Module" includes information about FDA-approved drugs and therapies, potential drug molecules, and designed 3D structures of targeted proteins across various species. The "Immunotherapy Module" contains computationally designed therapeutic targets using 10 essential *R. oryzae* proteins, which are crucial for the invasion of fungal spores into the host organism[158]. The organization of modules and submodules is represented by **Figure 3.2**.

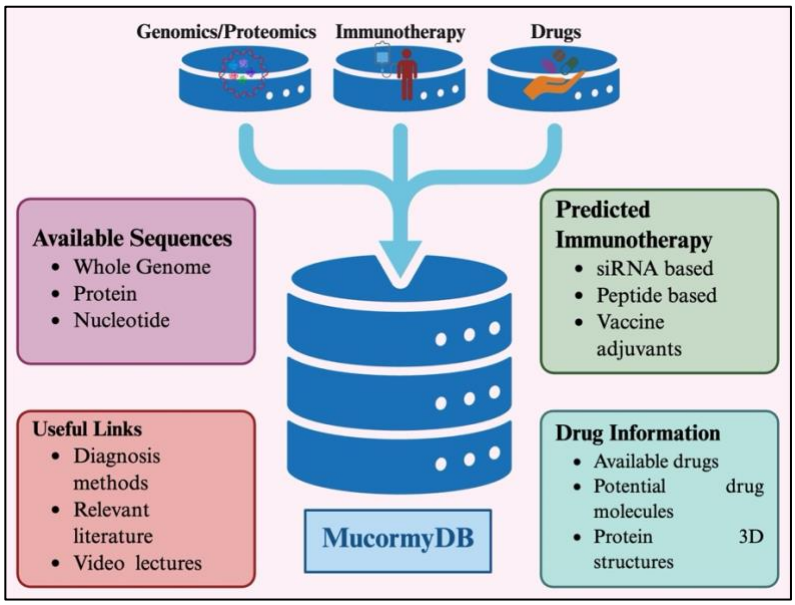

Figure 3.2: The structural organization of "Mucormydb".

3.3. Data Curation and Annotation

3.3.1. Genome and Proteome Data

The whole-genome and proteome sequence data of Mucorales species were obtained from the National Center for Biotechnology Information (NCBI). The data include genomic sequences of three Cunninghamelae, four Apophysomyes, and seven Mucor species. Every genomic record has extensive metadata, such as genus, species, strain name, genome size, GC content, sequencing coverage, and NCBI Taxonomy Identifier, and has direct hyperlinks to the corresponding NCBI records. In total, the module contains 31 sequences of Mucorales including both nucleotide and protein sequences, indicating the number of available sequences by species. The information provided by this module helps to gain a better insight into the diversity of species and enables sequence-level and comparative genomic studies.

3.3.2. Available therapeutics

Information regarding FDA-approved drugs, combination therapies, and experimental antifungal compounds was obtained from extensive literature mining. A total of 15 drugs have been documented, including those that are in different clinical phases or have been the subject of scientific studies. The target, mode of action, and therapeutic use of each drug are described in detail in each entry. Users can access detailed information in the MucormyDB submodule "Available Drugs" (https://webs.iiitd.edu.in/raghava/mucormydb/drug.php). Ten essential proteins that are required for the pathogenicity and survival of *R. oryzae*, the most common etiological agent of mucormycosis, were chosen as potential therapeutic targets. The "Target Sequence" module in DrugBank was used to find possible therapeutic compounds. The search employed optimized BLAST parameters: Cost to open a gap (-1), Penalty for mismatch (-3), Expectation value (0.001), Cost to extend a gap (-1), along with filters for Drug Type: Approved and Protein Type: Target.

Through this analysis, 23 potential drug molecules were identified that act as inhibitors, binders, antagonists, substrates, or agonists against the target proteins. Detailed results are available in the "Potential Drug Targets" submodule of MucormyDB (https://webs.iiitd.edu.in/raghava/mucormydb/pdt.php). The absence of experimentally determined 3D structures for several fungal proteins presents a major limitation in structure-based drug design. To address this, tertiary structures for the ten selected *R. oryzae* proteins were predicted using AlphaFold 2.0 with default parameters. Protein sequences were obtained from UniProt, and predicted structures are available in interactive form and as

downloadable PDB files in the 3D Structure Submodule (https://webs.iiitd.edu.in/raghava/mucormydb/3ds.php).

3.3.3. Vaccine Design

3.3.3.1. Peptide-based Therapeutics

All feasible 9-mer peptides were created from the ten essential *R. oryzae* proteins involved in host invasion in order to find viable peptide-based vaccine candidates. Multiple immunoinformatic tools were employed for the prediction of B-cell epitopes, T-cell epitopes, major histocompatibility complex (MHC) binders, interleukin-4 and interleukin-10 inducing peptides, interferon-gamma inducing epitopes, Nucleotide-based adjuvants, and RNA-based therapeutics. The list of epitopes shown in table 3.1 have B-cell, T-cell, IL-4, IL-10 and IFN- γ inducing predicted properties.

B-cell epitope prediction was carried out using LBtope with a 60% probability threshold, consistent with parameters used in prior studies[41,159]. CTLpred was employed for cytotoxic T-cell epitope prediction using its artificial neural network (ANN) module with a cut-off score of 0.51 and default parameters[160]. The major histocompatibility complex (MHC) class I and II binders were predicted using ProPred1 and ProPred, respectively, with thresholds of 4% and 3%[161,162]. Both tools also enabled the identification of promiscuous binders capable of interacting with multiple alleles. Cytokine-inducing peptides were predicted using IL4Pred, IL10Pred, and IFNepitope, utilizing thresholds of 0.2, −0.3, and default window length 15, respectively[60,163,164]. To identify potential peptide-based adjuvants, VaxinPAD was used with an SVM-based model at a threshold of -0.5[165]. The top 5 predicted B-cell epitopes, T-cell epitopes, IL-4 epitopes, IL-10 epitopes, and IFN-γ epitopes with at least 60% accuracy are displayed on the MucormyDB web server. Similarly, the top five MHC class I and MHC class II binders with the highest affinity scores can be found online.

Table 3.1: The highest ranking peptide-based therapeutic candidate.

S.No	Protein	Peptide	Vaccine Adjuvants	MHC-I	MHC-II
1	**CYP51**	FMEQKRMIK	✓	HLA-B*5401, HLA-B*2705, HLA-A20 Cattle, HLA-A*3101, HLA-A3, HLA-A*1101, HLA-A1	DRB1_0305, DRB1_1101
2	**GWT1**	i) RAGLMIMTC ii) LYLGLQDWV iii) MTLLGFLRK iv) ARFFLTKSV	i) ✗ ii) ✗ iii) ✗ iv) ✗	i) HLA-B*5801, HLA-B*5103 ii) HLA-A24, MHC-Kd, HLA-Cw*0401, HLA-B*5103	i) ✗ ii) ✗ iii) ✗ iv) ✗

		v) VMTLLGFLR vi) VLYLGLQDW vii) VGSFVFSSG	v) ✕ vi) ✕ vii) ✕	iii) HLA-A68.1, HLA-A*3101, HLA-A3, HLA-A*1101, HLA-A1 iv) HLA-B*3901, HLA-B*2705, HLA-B*2702, HLA-B14 v) HLA-A20 Cattle, HLA-A68.1, HLA-A*3302, HLA-A*3101, HLA-A3, HLA-A*1101 vi) HLA-A3, HLA-B62, HLA-B*2702 vii) ✕	v) DRB5_0101, DRB1_0102, DRB5_0105 vi) ✕ vii) ✕
3	**FTR1**	LEQNAWNQV	✕	MHC-Kk, HLA-B61, HLA-B60, HLA-B*4403, HLA-B40, HLA-B*3701,	✕
4	**CotH3**	i) PQLPWPIEK ii) LPWPIEKDP iii) DRFMRFMVI iv) NLPQLPWPI	i) ✕ ii) ✕ iii) ✓ iv) ✕	i) HLA-A3, HLA-A*1101 ii) ✕ iii) MHC-Kb, MHC-Dd, HLA-Cw*0301, HLA-B*5201, HLA-B*3901, HLA-B*2705, HLA-B*2702, HLA-B14 iv) MHC-Dd, MHC-Db, HLA-B62, HLA-A2.1, HLA-A3, HLA-A*0201, HLA-A2, MHC-Kk	i) ✕ ii) ✕ iii) ✕ iv) ✕
5	**ARP**	i) RKRKAEEEI ii) NPACGFQNF	i) ✕ ii) ✕	i) HLA-B*0702, HLA-A20 Cattle ii) MHC-Ld, HLA-Cw*0702, HLA-Cw*0401, HLA-B*0702, HLA-B62, HLA-B*5801, HLA-B*51, HLA-B*5301, HLA-B*4403, HLA-B*3801, HLA-B*3501	i) ✕ ii) ✕
6	**ARF**	STLKNRQYS	✕	MHC-Db, MHC-Db revised	✕
7	**DHD**	DAVLVSIGR	✕	HLA-A20 Cattle, HLA-A68.1, HLA-A*3302, HLA-A*3101	✕
8	**CaN**	AFSARGNKR	✕	HLA-A*3302, HLA-A*3101	✕
9	**SAPs**	i) FWSYLDRQA ii) PQFTHLLLK iii) FAELLHCSN	i) ✕ ii) ✕ iii) ✕	i) HLA-B*51, HLA-B*5401, HLA-B*5301 ii) HLA-B*2705, HLA-A*3101, HLA-A3, HLA-A*1101 iii) HLA-A1, HLA-B*5401, HLA-B*5103	i) ✕ ii) ✕ iii) ✕

10	**Cda**	i) VNWPYGAQR ii) SYVEMGSNG iii) DSYVEMGSN	i) ✕ ii) ✕ iii) ✕	i) HLA-B*2705, HLA-A20 Cattle, HLA-A*3302 ii) ✕ iii) ✕	i) ✕ ii) ✕ iii) ✕

3.3.3.2. Nucleotide-based adjuvants

VaccineDA was used to perform predictive analyses of nucleotide based vaccine adjuvants in batch mode with default settings[166]. The adjuvant sequences were limited to 30 nucleotides. According to the scoring of support vector machines, the top ten nucleotide-based adjuvants per gene were shortlisted, which are displayed in **Table 3.2**.

TLR9 activation is highly sequence-context dependent, and CpG motifs flanked by thymine-rich regions induce stronger immune responses than GC-dense sequences. High GC content increases structural rigidity and may reduce accessibility to TLR9. Therefore, immunostimulatory sequences typically display moderate GC content (<50%), which is consistent with both the VaccineDA training data and established biological mechanisms of TLR9 recognition.

Table 3.2: Top10 nucleotide based adjuvants predictive by VaccineDA.

S.No.	Protein	Sequence	Start position	Class	SVM score	Length	Mol. Wt.	Tm	GC content (%)
1	**CYP51**	ATGATGCGTCG TGTCGTCGC	1189	Immuno-modulatory	1.631	20	6140.03	55.88	60
2	**GWT1**	TACTTGACAGT CTATCGTGC	118	Immuno-modulatory	1.436	20	6083.02	49.73	45
3	**FTR1**	TACCTTGAACA AAATGCTTG	676	Immuno-modulatory	1.393	20	6100.06	45.63	35
4	**CotH3**	AATATTTCTATC GTTTCCCA	1579	Immuno-modulatory	1.541	20	6017	43.58	30
5	**ARP**	ACGTTGTTTAG GTGATCGAC	1086	Immuno-modulatory	1.355	20	6163.08	49.73	45
6	**ARF**	AATAGACAATA TTCTATTTA	436	Immuno-modulatory	1.411	20	6098.1	37.43	15
7	**DHD**	AGTTTCTATCG GTCGTCGTC	921	Immuno-modulatory	2.024	20	6090.01	51.78	50
8	**CaN**	CTTGAAAGATA ATCAACTTC	345	Immuno-modulatory	0.805	20	6084.06	43.58	30

9	**SAPs**	AGATTCTTTAC AAGCTTCTT	387	Immuno-modulatory	1.274	20	6057.03	43.58	30
10	**Cda**	ACAAGAACGT TGTTGATGTT	1055	Immuno-modulatory	1.728	20	6171.11	45.63	35

3.3.3.3. RNA-based Therapeutics

The messenger RNA sequences of the ten proteins of *R. oryzae* were transcribed into 19-mer oligonucleotides in an attempt to produce siRNA-based treatment approaches. The optimal length of 19 nucleotides can effectively silence a gene[167]. The DesiRm software was then used to screen candidate siRNAs under default parameters and only those with a score of efficacy above 0.80 and a target accessibility of above 0.60 were retained[168]. The resulting siRNAs showed high predicted silencing capacity, and the top ten siRNAs per gene are shown in **Table 3.3**, highlighting their potential as RNA-based therapeutic interventions to treat mucormycosis.

Table 3.3: The highest-ranking RNA-based therapeutic candidates.

S.No.	Protein	Antisense sequences of siRNA	Position on mRNA	mRNA target sequence	Target accessibility
1	**CYP51**	UAAAGUAAACUUGAGAGGG	35	CCCUCUCAAGUUUACUUUA	0.080319
2	**GWT1**	UUCUUUUUGUGGUCUCUUC	84	GAAGAGACCACAAAAAGAA	0.008413
3	**FTR1**	UUAGCCAACUUGACCUUCC	347	GGAAGGUCAAGUUGGCUAA	0.02931
4	**CotH3**	UUGAUAUAGGAGAAUUGAG	773	CUCAAUUCUCCUAUAUCAA	0.06551
5	**ARP**	UAAAUUAUGAGCAUCAAGC	345	GCUUGAUGCUCAUAAUUUA	0.001364
6	**ARF**	UAUACUAGUCUGUCCACCU	204	AGGUGGACAGACUAGUAUA	0.052494
7	**DHD**	UUAGUGGACAUCUUGAACU	806	AGUUCAAGAUGUCCACUAA	0.007881
8	**CaN**	UUGAAGUUGAUUAUCUUUC	348	GAAAGAUAAUCAACUUCAA	0.060585
9	**SAPs**	UUAGGUUGUGAGAAUUUGG	1979	CCAAAUUCUCACAACCUAA	0.021075
10	**Cda**	AAAGAUUUGAAGUUCCUCC	122	GGAGGAACUUCAAAUCUUU	0.005263

3.4. Discussion and Conclusion

The pathogenesis of mucormycosis is investigated to clarify the complicated and dynamic interaction between virulence factors of fungi and the immune responses of the host. Recent studies have demonstrated that the pathogenicity of Mucorales species is not only conditioned by their ability to penetrate the tissues of the host, but also by their extraordinary ability to adapt to adverse physiological factors[157,169]. Fungal persistence even in the strong host defense is made possible by a number of processes, such as metabolic flexibility,

tolerance to oxidative stress, and immunological evasion[170]. Mucorales species have larger gene families involved in iron absorption, carbohydrate metabolism, and stress response pathways, according to comparative genomic research[171]. These genetic features provide a major survival advantage and contribute to the rapid growth of infections once established. Furthermore, the overexpression of adhesion molecules and regulatory proteins facilitates efficient tissue colonization and vascular invasion, making mucormycosis one of the most aggressive forms of fungal infection known to clinical medicine[172]. Managing mucormycosis clinically is still quite difficult. Even with the availability of antifungal drugs like triazoles and amphotericin B, treatment results are frequently unsatisfactory because of delayed diagnosis, low drug efficacy, and the rise of resistant fungal strains[154]. These limitations emphasize the urgent need for early diagnostic markers, improved antifungal drugs, and alternative therapeutic approaches that target molecular determinants of virulence.

Fungal biology and host-pathogen interactions have been greatly improved at the bioinformatics level by a number of fungal databases, including as e-Fungi, CandidaDB, AspGD, DemaDb, FungiDB, PHI-base, and CGD. However, there is little to no emphasis on mucorales in these resources, which mostly concentrate on Aspergillus, Candida, and other clinically important fungi. To bridge this gap, MucormyDB marks a significant step forward. It is a comprehensive platform that combines genetic, proteomic, and therapeutic data relevant to mucorales species. MucormyDB provides researchers a comprehensive tool required to identify targets and develop rational therapies by combining drug discovery resources, 3D protein models, immunoinformatic-based vaccination and siRNA predictions. Unlike earlier fungal databases, it concentrates solely on mucormycosis-related diseases, filling a long-standing need in fungal bioinformatics.

In conclusion, the current knowledge of mucormycosis emphasises how urgently combined computational and experimental research is needed. By providing a centralised and easily navigable platform for examining the molecular foundations of Mucorales pathogenesis, MucormyDB fills a significant knowledge gap. This resource is likely to speed up the search for novel drugs and vaccine candidates, which will ultimately result in the development of more potent mucormycosis diagnostic and therapeutic alternatives.

4. Host Specific Prediction of Interleukin-4 Inducing Peptides

4.1. Introduction

Interleukin 4 (IL-4) was initially discovered in 1982, and is recognized as a key controller of various biological and immunological functions. The modulatory activity of IL-4 has been reported in a range of immune cell types, such as B and T lymphocytes, natural killer (NK) cells, monocytes, dendritic cells, basophils, mast cells, and fibroblasts[173–177]. It also induces Th2 differentiation of naïve T-helper (Th0) cells and Th2 expansion of CD8+ cytotoxic T lymphocytes[61]. It is a multifunctional, pleiotropic cytokine with extensive cellular responses. The receptors of IL-4 are widely expressed, mostly on hematopoietic cells[178,179]. B cell growth factor-1 (BCGF-1) and B cell stimulatory factor-1 (BSF-1) are known as synonyms for this cytokine[177,180]. This cytokine plays a vital role in the regulation of B cell growth and antibody isotype expression; stimulation of T cell growth and the generation of cytotoxic T lymphocytes, and modulation of the growth and differentiation of hematopoietic bone marrow stem cells[173]. With recent advancements, this cytokine has been identified as a potential treatment for several autoimmune diseases, including multiple myeloma, certain tumours, psoriasis, and arthritis[181–183].

Over the last two decades, the application of computational approaches, including the development of machine learning (ML) methods, has led to a better understanding of cytokine potential in various regulatory pathways using experimentally validated data. IL4Pred was developed by the Raghava group in 2013 for predicting IL-4-inducing peptides, regardless of the host specificity. Authors have reported a maximum accuracy of 75.76% and an MCC value of 0.51 for a hybrid ML model that combines amino acid pairs with motif information[60]. Following this, multiple ML-based methods have been developed to predict IL-4-inducing peptides using the same dataset previously employed by IL4pred[61,62]. However, they have a number of fundamental limitations, such as their host-independent nature and dependence on scarce data availability. A next-generation breakthrough in the identification of peptides that induce IL-4 is the development of host-specific models. In contrast to generalised frameworks, these models account for the innate immunological variation between species, especially between humans and mice. Because

cytokine signalling pathways, antigen recognition mechanisms, and MHC allele variations greatly affect IL-4 production, species-specific modelling is essential for precise prediction. Host-specific models make it possible to make more accurate, physiologically relevant predictions by incorporating host-dependent immunological properties. They also offer important insights for developing vaccines and targeted immunotherapeutics that are suited to certain host systems.

In this study, we developed IL4Pred2, developed using separate models for humans and mice to predict IL-4-inducing peptides. The experimentally validated IL-4-inducing and non-inducing peptides with MHC class-II binding specificity were retrieved from the Immune Epitope Database (IEDB). To develop a state-of-the-art classifier tool, we applied various ML-based models, and their performance was evaluated using an independent dataset. A user-friendly web server and a standalone version of 'IL4Pred2' can be accessed using the link https://webs.iiitd.edu.in/raghava/il4pred2/.

4.2. Methodology and Algorithm Development

The IL4pred2 framework was developed to address the complex immunological mechanisms underlying IL-4 induction, which involves both MHC binding affinity and downstream cytokine response. Given this dual biological dependency, the algorithmic design incorporated complementary modeling paradigms to capture sequence motifs, compositional signatures, and contextual residue interactions.

The overall computational strategy was structured around dataset stratification, multi-scale feature representation, diverse predictive modeling approaches, and ensemble integration to enhance biological interpretability and predictive robustness.

4.2.1. Data Preparation and preprocessing

The study was performed using human and mouse as hosts. For this, we individually collected experimentally validated IL-4-inducing, MHC class II binders from IEDB as the positive dataset[184]. To ensure that the model works well with a diverse range of peptides, whether HLA class II binders or non-binders, we used three different experimentally validated IL-4 non-inducing peptide sets as negative datasets. Data statistics are given in Table 4.1.

1. Main dataset includes a mixed set of MHC Class II binders and non-binders with IL-4 non-inducing property.

2. In Alternate1 dataset, our negative set includes MHC Class II binders that do not induce IL-4.
3. In Alternate2 dataset, our dataset has MHC Class II non-binders that do not induce IL-4.

The MHC class II binder/non-binder was categorized based on its IC50 values. As reported in the literature, MHC class II binding affinities were grouped into good, weak, and non-binders categories depending on their IC50 values. The good binders (IC50 ≤ 1000 nM), weak binders (IC50 1,000– 10,000 nM) and non-binders (IC50 > 10000 nM). In this analysis, we identified the peptides whose IC50 exceeds 1000 nM as non-binders of the MHC class II[185,186]. Moreover, we filtered extracted data by eliminating redundant peptide sequences and length filter. The length range was selected on the basis of maximum sequence coverage.

Table 4.1: Peptide distribution across datasets for human and mouse.

Human						
Data Type	**Main Dataset**		**Alternate1 Dataset**		**Alternate2 Dataset**	
	Positive	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
Unique peptide	900	1389	900	534	900	1855
Length (9-22)	845	1333	845	516	845	1817
Final Data	845	845	845	516	845	845
Mouse						
Data Type	**Main Dataset**		**Alternate1 Dataset**		**Alternate2 Dataset**	
	Positive	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
Unique peptide	660	1787	660	905	660	882
Length (9-22)	560	1547	560	698	560	849
Final Data	560	560	560	560	560	560

4.2.2. Feature extraction using standalone tools

Comprehensive compositional descriptors were initially computed to capture global and local residue distributions potentially influencing cytokine induction. However, given the high dimensionality relative to dataset size, subsequent feature selection was essential to mitigate overfitting risk and improve model interpretability. To calculate the sequence-based features, we have used a well-known protein/peptide feature extraction tool “Pfeature” developed in 2019[187]. This tool has the capability to calculate around 9189 composition-

based features. We have used this tool to compute composition features such as amino acid composition (AAC), di-peptide composition (DPC), tri-peptide composition (TPC) etc.

4.2.3. Computational Framework

We have implemented different approaches to predict IL-4-inducing and non-inducing peptides. Initially, we conduct a detailed study of IL-4-inducing peptides using positional, compositional, and univariate analysis (mean-based and LR-based single feature analysis) to gain biological insights. Then, we proceed to the feature selection step, where we utilize multiple feature selection procedures, such as SVC-L1, mRMR, and SFS[188,189]. Multiple feature selection strategies were employed to balance statistical relevance (t-test), redundancy reduction (mRMR), sparsity enforcement (SVC-L1), and sequential optimization (SFS). This multi-stage selection ensured stable and biologically meaningful feature subsets.

Both alignment-based and alignment-free approaches were implemented to evaluate complementary biological signals. Alignment-based tools capture evolutionary similarity and conserved motifs using BLASTshortP (version 2.9.0) and MERCI (Motif EmeRging, and with Classes Identification) and MEME-MAST tools[190–193], while alignment-free machine learning models identify discriminative compositional patterns independent of sequence homology. Their integration allows detection of both conserved and novel IL-4-inducing signatures.

Classical ML models were prioritized for robustness in moderate-sized immunological datasets. Deep learning architectures such as 1D-CNN were implemented to capture spatial motif patterns, while TabNet was utilized for its attention-driven feature prioritization[194]. ProtBERT embeddings were incorporated to leverage contextual representations derived from large-scale protein corpora, enabling modeling of higher-order residue dependencies[195].

The ensemble framework was designed to integrate orthogonal predictive signals derived from alignment-based similarity, motif discovery, and machine learning classification. By combining these complementary paradigms, the ensemble reduces variance, enhances stability, and improves biological reliability in IL-4 peptide prediction. **Figure 4.1** depicts the overall computational framework that summarizes the methodological approaches and forecasting techniques used in this study.

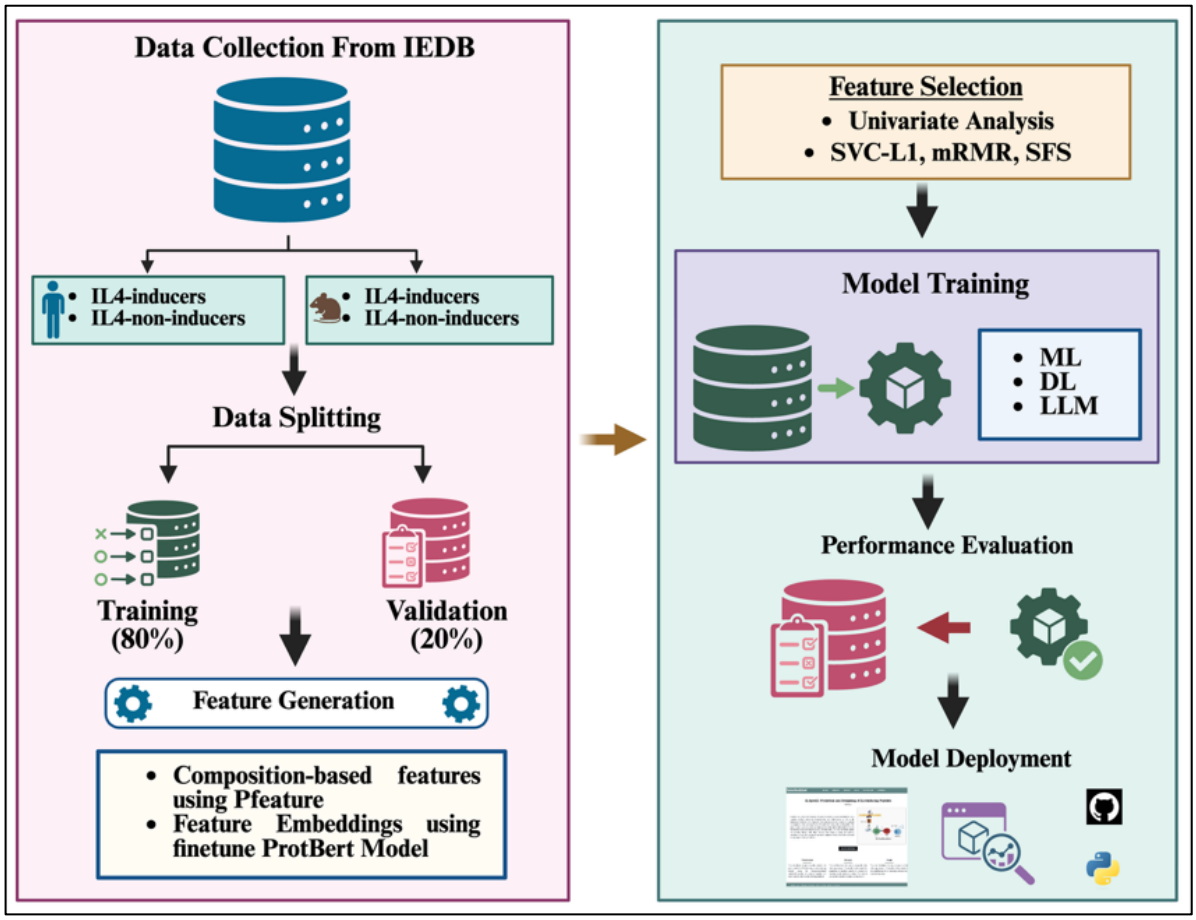

Figure 4.1:Overall computational framework of IL4pred2.

4.2.4. Model evaluation

To evaluate the model performance, the dataset was divided into training and validation set in 80% and 20% ratio. Then a five-fold cross validation approach was used, where sequences in the training sets were first arbitrarily divided into five equivalent folds Thereafter, four of these folds were used for training and remaining fold is used for testing. To test the efficiency of several ML classification models, we utilized well-established evaluation criteria. In this study, we assessed sensitivity (Sens), specificity (Spec), and accuracy (Acc) using both threshold-dependent and independent factors. These parameters were quantified using the formulae below.

$$Sensitivity = TP/(TP + FN) * 100 \quad \text{(i)}$$

$$Specificity = \frac{TN}{TN+FP} * 100 \quad \text{(ii)}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} * 100` \quad \text{(iii)}$$

$$MCC = ((TP * TN) - (FP * FN))/((TP_{FP})(TP + FN)(TN + FP)(TN + FN)) \quad \text{(iv)}$$

4.2.5. Ensemble approach

In order to improve the prediction accuracy of IL-4-inducing peptide identification, we have also tried an ensemble framework in this study that combines machine learning models with conventional sequence-based techniques. Top-performing machine learning classifiers are combined with motif-based techniques like MEME-MAST and MERCI, as well as alignment-based tools like BLAST. The ensemble framework offers a more thorough and biologically significant prediction model by utilizing the complementary strengths of these methods, motif-based analysis to find conserved functional patterns and sequence alignment to capture evolutionary patterns.

4.3. Result and Analysis

4.3.1. Preferential Positional Analysis

We have observed that certain amino acids in the main datasets of Human and mouse hosts are preferred at certain positions. For example, in the human host Lysine (K) is present at the 1st, 4th, 5th, 8th, 9th, 10th,13th and 18th position. Whereas Proline (P) is present at the 1st, 5th, 7th, 8th, 9th, 13th, 15th and 18th position. While in the mouse host, glycine (G) at the 1st and 4th, proline (P) at 3rd, 5th, 6th, 8th, 12th, 13th, 16th and 17th, phenylalanine (F) at the 4th and 5th and tyrosine (Y) are present at the 6th, 8th and 17th position in IL4 inducing peptides. Additionally, leucine (L) is found to be prominent at the 1st, 5th, 6th, 9th, 13th , 14th , 17th and 18th position in IL4 non-inducing peptides (please refer to **figure 4.2**).

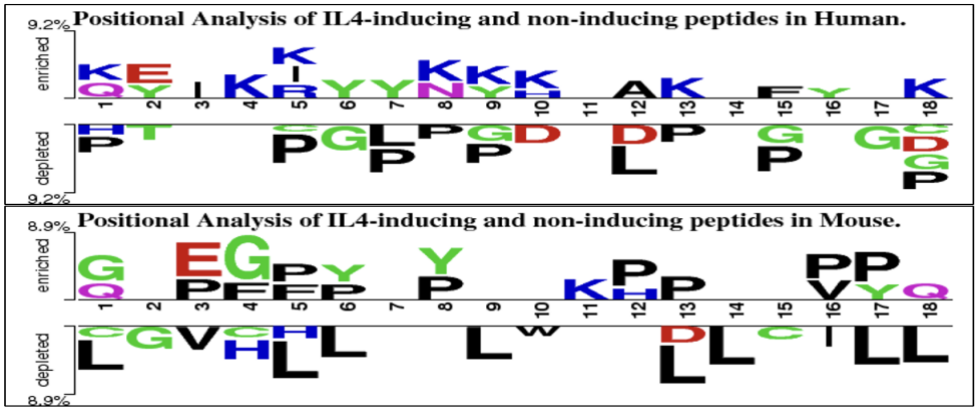

Figure 4.2: Preferential amino-acid positional analysis of human and mouse data.

To identify the amino acid positional difference in Human and mouse hosts' IL-4-inducing peptides, we also made a comparative positional analysis plot between the human and mouse main dataset. Here, we observed that amino acid leucine (L) is abundant at the 1st, 8th, 13th, 14th, 17th and 18th position, methionine (M) is found at the 4th AND 6th position and

cysteine is present at the 11th and 16th position in human IL4-inducers while serine (S), glutamine (Q), proline (P), asparagine (N) and glycine (G) residues are abundant at specific positions in mouse IL4-inducing peptides (please refer to **figure 4.3**).

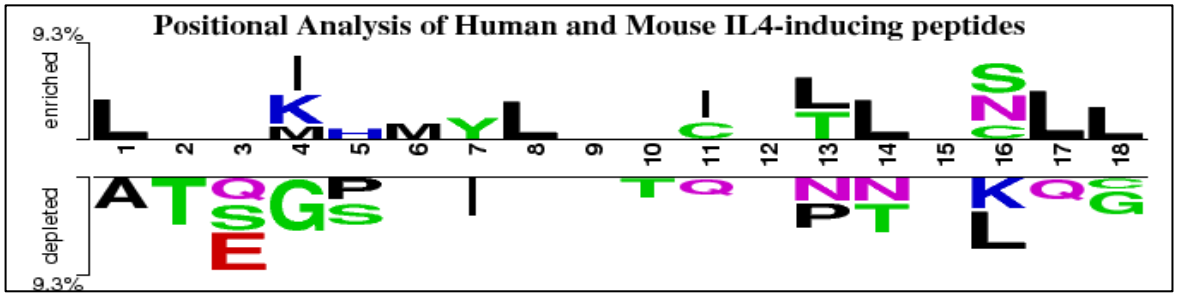

Figure 4.3: Preferential positional analysis of residues in human vs mouse.

4.3.2. Compositional Analysis

The compositional analysis also confirms the average amino acid composition difference between both hosts, as the average amino acid composition of cysteine, phenylalanine, leucine, and methionine residues is significantly higher in human IL-4-inducing peptides while average amino acid composition of glycine and proline residues is abundant in mouse data (see **figure 4.4-4.5**).

Leucine is a hydrophobic amino acid that frequently occupies anchor positions within the MHC class II binding core. Consequently, the main negative dataset, despite being non-inducing, includes a subset of MHC class II-binding peptides that retain hydrophobic anchor residues such as Leu, leading to a higher average Leu composition compared to the positive dataset. In contrast, the alternate2 negative dataset, which is restricted to MHC class II non-binders, is relatively depleted of hydrophobic residues required for stable MHC class II interaction, resulting in the lowest Leu composition. While the Alternate1 negative dataset consists of MHC class II binders that are IL4 non-inducers.

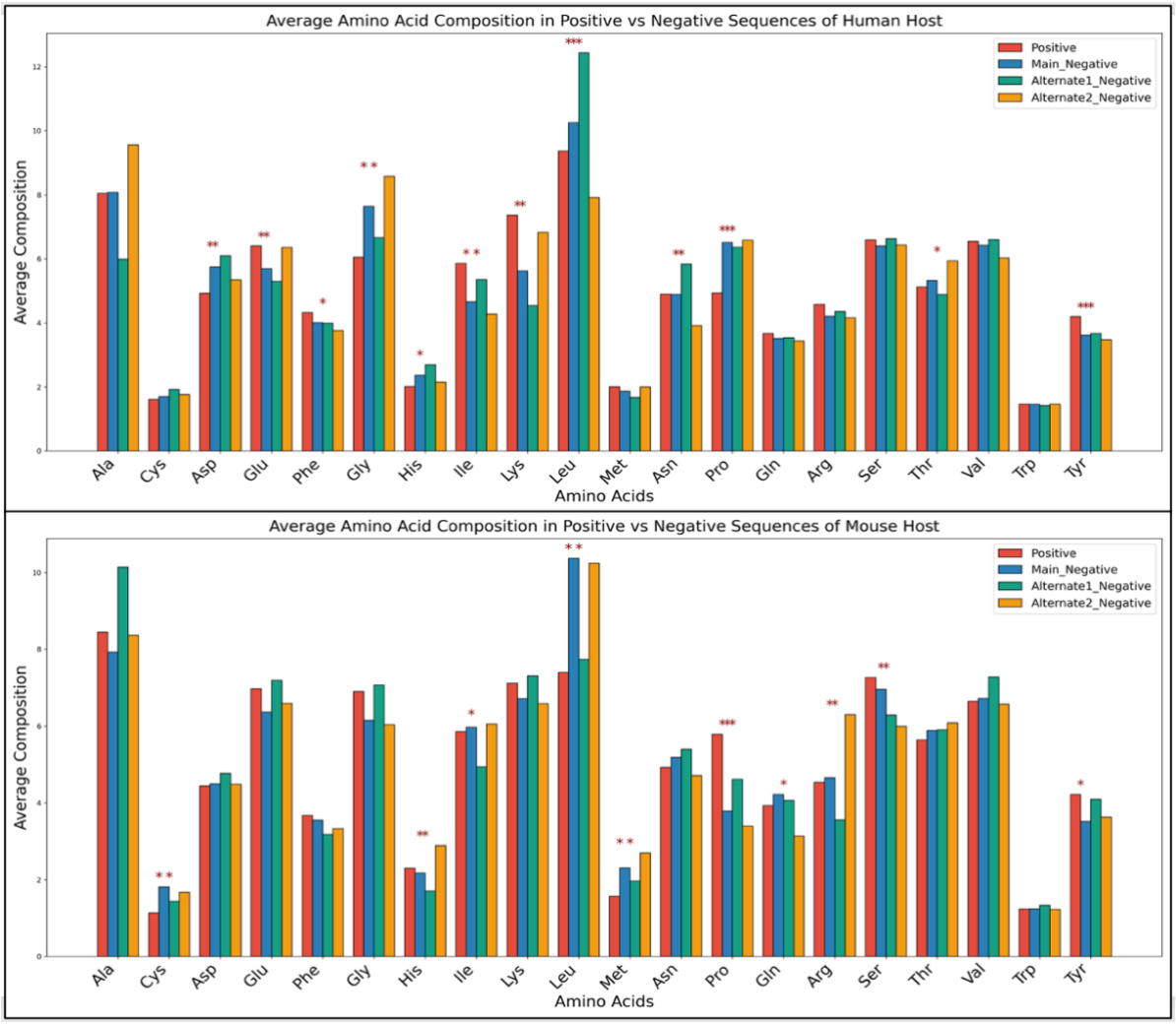

Figure 4.4: Compositional analysis of human and mouse data.

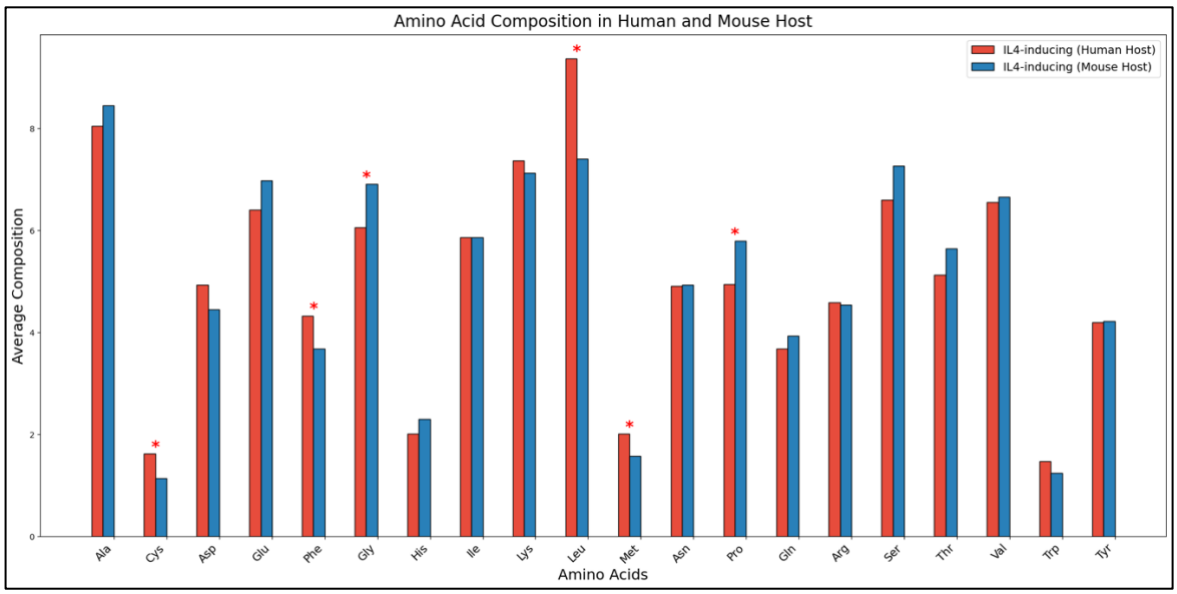

Figure 4.5: Compositional analysis between human and mouse.

4.3.3. Univariate analysis

To identify the most important features in a set of features, we have applied two different approaches.

4.3.3.1. Mean based analysis

In this analysis, we have calculated the mean value of each feature in IL4-inducing and non-inducing sets and calculated the difference between them. We ranked the features based on their higher mean difference. The same process was performed in both Human and Mouse datasets. Finally, **Table 4.2** lists the top five features based on their highest mean difference.

Table 4.2: The top five features with highest mean difference in both datasets.

Human Main Dataset					**Mouse Main Dataset**				
Feature	**Pos mean**	**Neg mean**	**Diff**	**p-value**	**Feature**	**Pos mean**	**Neg mean**	**Diff**	**p-value**
PAAC1_lam1	0.17	-0.17	0.34	1.59E-05	CeTD_100_p_VW1	0.18	-0.18	0.36	1.59E-05
SOC1_SC1	0.16	-0.16	0.32	1.59E-05	CeTD_50_p_VW1	0.18	-0.18	0.36	1.59E-05
CeTD_1-3_SA	0.15	-0.15	0.30	1.59E-05	CeTD_75_p_VW1	0.17	-0.17	0.34	1.59E-05
CeTD_21_PZ	0.15	-0.15	0.30	1.59E-05	SEP_CY	0.17	-0.17	0.34	1.59E-05
CeTD_23_PO	0.15	-0.15	0.30	1.59E-05	CeTD_100_p_PZ1	0.16	-0.16	0.32	1.59E-05
Human Alternate1 Dataset					**Mouse Alternate1 Dataset**				
PAAC1_lam1	0.19	-0.32	0.51	1.59E-05	CeTD_50_p_HB2	0.22	-0.22	0.44	1.59E-05
SOC1_SC1	0.18	-0.3	0.48	1.59E-05	CeTD_100_p_PZ3	0.21	-0.21	0.42	1.59E-05
AAC_K	0.16	-0.26	0.42	1.59E-05	CeTD_21_VW	0.21	-0.21	0.42	1.59E-05
APAAC1_K	0.16	-0.26	0.42	1.59E-05	CeTD_50_p_PZ1	0.21	-0.21	0.42	1.59E-05
PAAC1_K	0.16	-0.26	0.42	1.59E-05	CeTD_50_p_VW1	0.21	-0.21	0.42	1.59E-05
Human Alternate2 Dataset					**Mouse Alternate2 Dataset**				

CeTD_100_p_SA1	0.23	-0.23	0.46	1.59E-05	SER_A	0.12	-0.12	0.24	1.59E-05
CeTD_75_p_SA1	0.23	-0.23	0.46	1.59E-05	DPC1_PS	0.11	-0.11	0.22	1.59E-05
CeTD_50_p_CH1	0.22	-0.22	0.44	1.59E-05	DPC1_GP	0.1	-0.1	0.20	1.59E-05
CeTD_50_p_SA1	0.22	-0.22	0.44	1.59E-05	AAC_P	0.09	-0.09	0.18	1.59E-05
BTC_H	0.21	-0.21	0.42	1.59E-05	AAC_R	0.09	-0.09	0.18	1.59E-05

Note: Pos: Positive; Neg: Negative; Diff: Difference; PAAC1_lam1: Sequence correlation factor for lambda 1; SOC1_SC1: Sequence order coupling number with Schneider matrix for lag 1; AAC_K: Amino acid composition of Lysine; APAAC1_K: Amphiphilic pseudo amino acid composition of Lysine; PAAC1_K : Pseudo amino acid composition of Lysine; BTC_H: Composition of Hydrogen bonds; SER_A: Shannon entropy of Alanine; DPC1_PS: Composition of Proline- Serine; DPC1_GP: Composition of Glycine-Proline; AAC_P: Amino acid composition of Proline; AAC_R: Amino acid composition of Arginine; SEP_CY: Shannon entropy of residues having cyclic side chain; CeTD_100_p_VW1: Number of group 1 residues for normalized Vanderwal's volume present in 100% quartile; CeTD_100_p_PZ1: Composition of group 1 residues for polarizability attribute; CeTD_23_PO: Composition of residues for polarity attribute; CeTD_50_p_SA1: Composition of group 1 residues for solvent accessibility attribute; CeTD_50_p_HB2: Number of group 2 residues for hydrophobicity present in 50% quartile;

4.3.3.2. LR-based single feature analysis

The Logistic regression-based analysis was performed to understand the efficacy of a single feature to discriminate IL-4 inducing and non-inducing peptides in both Human and Mouse. The features were ranked based on their highest AUC value. The top five listed features of all the three datasets are given in **Table 4.3**.

Table 4.3:The top five LR-based features with maximum performance.

Human Main Dataset						**Mouse Main Dataset**					
Feature	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Feature**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
PAAC1_lam1	53.85	55.15	54.51	0.59	0.09	CeTD_100_p_VW1	60.71	56.43	58.57	0.61	0.17
SOC1_SC1	57.92	55.88	56.88	0.58	0.14	CeTD_50_p_VW1	60.71	56.43	58.57	0.60	0.17
CeTD_1-3_SA	62.75	50.94	56.73	0.58	0.14	CeTD_75_p_VW1	60.71	56.43	58.57	0.60	0.17

AAC_K	55.35	55.15	55.25	0.57	0.11	AAC_L	58.57	51.43	55.00	0.60	0.10
SER_K	55.35	55.15	55.25	0.57	0.11	SER_L	58.57	51.43	55.00	0.60	0.10
Human Alternate1 Dataset						**Mouse Alternate1 Dataset**					
PAAC1_lam1	57.40	60.27	58.49	0.65	0.17	QSO1_SC_A	59.82	48.04	53.93	0.57	0.08
SOC1_SC1	61.30	59.69	60.69	0.64	0.20	SER_A	56.43	50.36	53.39	0.57	0.07
CeTD_23_SS	57.87	64.73	60.47	0.64	0.22	AAC_A	56.43	50.36	53.39	0.56	0.07
CeTD_100_p_SA2	60.00	57.56	59.07	0.64	0.17	PAAC1_A	56.43	50.36	53.39	0.56	0.07
CeTD_75_p_SA2	60.00	57.56	59.07	0.63	0.17	APAAC1_A	56.43	50.36	53.39	0.56	0.07
Human Alternate2 Dataset						**Mouse Alternate2 Dataset**					
CeTD_75_p_CH1	57.63	61.42	59.53	0.62	0.19	CeTD_100_p_VW1	60.71	57.50	59.11	0.61	0.18
CeTD_100_p_CH1	57.63	61.42	59.53	0.62	0.19	CeTD_50_p_VW1	60.71	57.50	59.11	0.61	0.18
CeTD_50_p_CH1	57.63	61.42	59.53	0.62	0.19	CeTD_75_p_VW1	60.71	57.50	59.11	0.61	0.18
CeTD_100_p_SA1	53.96	62.60	58.28	0.62	0.17	CeTD_50_p_HB2	42.50	76.25	59.38	0.61	0.20
BTC_H	59.53	62.60	61.07	0.62	0.22	CeTD_50_p_PZ1	55.18	62.50	58.84	0.61	0.18

Note: Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUC: Area under curve; MCC: Mathew's correlation coefficient; PAAC1_lam1: Sequence correlation factor for lambda 1; SOC1_SC1: Sequence order coupling number with Schneider matrix for lag 1; AAC_K: Amino acid composition of Lysine; APAAC1_K: Amphiphilic pseudo amino acid composition of Lysine; PAAC1_K: Pseudo amino acid composition of Lysine; BTC_H: Composition of Hydrogen bonds; SER_A: Shannon entropy of Alanine; DPC1_PS: Composition of Proline- Serine; DPC1_GP: Composition of Glycine-Proline; AAC_P: Amino acid composition of Proline; AAC_R: Amino acid composition of Arginine; SEP_CY: Shannon entropy of residues having cyclic side chain; CeTD_100_p_VW1: Number of group 1 residues for normalized vander-waals volume present in 100% quartile; CeTD_100_p_PZ1: Composition of group 1 residues for polarizability attribute; CeTD_23_PO: Composition of residues for polarity attribute; CeTD_50_p_SA1: Composition of group 1 residues for solvent accessibility attribute; CeTD_50_p_HB2: Number of group 2 residues for hydrophobicity present in 50% quartile;

4.3.4. Alignment-Based Approach

4.3.4.1. BLAST

In this approach, a BLAST search was performed for peptides in the independent dataset against those in the training dataset to identify homologous sequences. Each query peptide was classified as an IL-4 inducer or non-inducer based on its top-scoring hit, peptides matching a positive sequence were labelled as inducers, while those matching a negative sequence were labelled as non-inducers. BLAST searches were executed at multiple E-value thresholds (1e-1 to 1e-6) to evaluate prediction performance across varying levels of sequence similarity. The detailed list of BLAST coverage results for both human and mouse hosts on all three validation datasets is provided in **Table 4.4.**

Table 4.4: The BLAST coverage on Human and Mouse datasets.

Human Main Dataset					**Mouse Main Dataset**				
e-value	**TH**	**SC**	**CPH**	**CNH**	**e-value**	**TH**	**SC**	**CPH**	**CNH**
0.1	557	171	89	59	0.1	197	70	30	29
0.01	458	143	75	41	0.01	164	65	28	27
0.001	331	113	60	31	0.001	126	52	22	20
0.0001	207	80	48	21	0.0001	80	35	14	12
1.00E-05	126	54	36	12	1.00E-05	39	24	9	6
1.00E-06	66	32	25	3	1.00E-06	18	11	7	1
Human Alternate1 Dataset					**Mouse Alternate1 Dataset**				
0.1	470	165	89	66	0.1	247	81	30	38
0.01	383	138	76	46	0.01	192	72	28	32
0.001	263	97	60	25	0.001	136	51	22	18
0.0001	174	73	48	15	0.0001	89	38	14	14

1.00E-05	104	48	36	7	1.00E-05	42	23	9	6
1.00E-06	57	29	25	2	1.00E-06	18	10	7	1
Human Alternate2 Dataset					**Mouse Alternate2 Dataset**				
0.1	611	174	90	66	0.1	320	92	31	57
0.01	492	151	76	55	0.01	178	81	28	50
0.001	378	121	61	42	0.001	105	44	22	19
0.0001	276	91	45	32	0.0001	53	22	14	6
1.00E-05	178	66	36	23	1.00E-05	30	14	9	4
1.00E-06	105	37	25	7	1.00E-06	14	9	7	2

Note: TH: Total Hits; SC: Sequence Conserved; CPH: Correct Positive Hit; CNH: Correct Negative Hit.

4.3.4.2. Motif Analysis

To find conserved sequence motifs typical of peptides that induce IL-4 and those that do not, a thorough motif analysis was carried out using two popular tools, MERCI and MEME-MAST. To determine their predictive value, motifs were selected from the training dataset and then searched through separate validation datasets for both human and mouse hosts.

4.3.5. Alignment-Free Approaches

4.3.5.1. *Machine Learning-Based Classifiers*

To build alignment-free predictive models, multiple ML algorithms were implemented using composition-based features such as AAC, DPC**,** and TPC. As summarized in **Table 4.5**, for the human host, the Extra Tree classifier trained on AAC features achieved the highest AUC of 0.76, while for the mouse host, the Random Forest classifier using DPC features yielded the best performance with an AUC of 0.70, demonstrating balanced sensitivity and specificity.

4.3.5.2. *Feature Selection Analysis*

Three well-known feature selection methods were used: SFS, mRMR, and SVC-L1 to improve model interpretability and eliminate redundant features. From the initial 9,189 features generated using Pfeature, 149, 50, and 17 features were selected by SVC-L1, mRMR, and SFS, respectively, for the human host, and 97, 50, and 17 features for the mouse host. The ET classifier with SVC-L1-selected features had the highest AUC of 0.79 for the human host. In contrast, the RF classifier had the highest AUC of 0.72 for the mouse host, as determined by ML classifiers created using these optimized feature sets. Comprehensive evaluation metrics are detailed in **Table 4.6**.

In addition, mean-based univariate analysis using the Student's t-test was employed to identify statistically significant features ($p < 0.05$) with the most considerable mean differences between IL-4-inducing and non-inducing peptides. ML models were developed using the top 100, 150, 200, and 300 ranked features, achieving the maximum AUC of 0.80 (MCC = 0.45) for the human host, and AUC of 0.82 (MCC = 0.50) for the mouse host with the MLP classifier (**Table 4.7**).

Table 4.5: Performance of ML classifiers developed using different features.

Results of Human Main Validation Dataset						
Feature	**Model**	**Sens**	**Spec**	**Accuracy**	**AUC**	**MCC**
AAC	**ET**	71.43	62.18	67.16	0.76	0.34
DPC	**ET**	70.33	66.67	68.64	0.76	0.37
TPC	**RF**	70.33	62.18	66.57	0.73	0.33
Results of Human Alternate1 Validation Dataset						
AAC	**ET**	67.84	73.53	69.96	0.79	0.4
DPC	**ET**	69.59	73.53	71.06	0.82	0.42
TPC	**ET**	72.52	73.53	72.89	0.81	0.45
Results of Human Alternate2 Validation Dataset						
AAC	**RF**	75.82	71.8	73.96	0.83	0.48
DPC	**ET**	74.73	71.8	73.37	0.81	0.47
TPC	**ET**	67.58	74.36	70.71	0.79	0.42
Results of Mouse Main Validation Dataset						

Feature	Model	Sens	Spec	Accuracy	AUC	MCC
AAC	**RF**	64.57	61.86	63.39	0.68	0.26
DPC	**ET**	64.57	62.89	63.84	0.7	0.27
TPC	**ET**	56.69	68.04	61.61	0.67	0.25
Results of Mouse Alternate1 Validation Dataset						
AAC	**RF**	57.48	67.01	61.61	0.65	0.24
DPC	**RF**	61.42	67.01	63.84	0.72	0.28
TPC	**ET**	59.06	62.89	60.71	0.68	0.22
Results of Mouse Alternate2 Validation Dataset						
AAC	**RF**	80.32	79.38	79.91	0.88	0.59
DPC	**ET**	74.8	84.54	79.02	0.87	0.59
TPC	**RF**	61.42	87.63	72.77	0.84	0.5

Note: RF: Random forest; ET: Extra Tree; Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUC: Area under curve; MCC: Mathew's correlation coefficient.

Table 4.6: The performance of Scikit based feature selection methods.

Results of Human Main Validation Dataset							**Results of Mouse Main Validation Dataset**				
Selection Method	**Model**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
SVC-L1	ET	73.08	69.23	71.3	0.79	0.42	69.29	60.83	65.63	0.72	0.3
mRMR	RF	71.43	57.69	65.09	0.69	0.29	65.35	58.76	62.5	0.67	0.24
SFS	LR	61.54	58.33	60.06	0.63	0.2	56.69	64.95	60.27	0.60	0.21
Results of Human Alternate1 Validation Dataset							**Results of Mouse Alternate1 Validation Dataset**				
SVC-L1	RF	70.18	67.65	69.23	0.78	0.37	55.12	68.04	60.71	0.68	0.23
mRMR	RF	70.18	69.61	69.96	0.76	0.39	53.54	74.23	62.5	0.65	0.28
SFS	LR	82.46	40.19	66.66	0.67	0.25	48.82	64.95	55.8	0.60	0.14
Results of Human Alternate2 Validation Dataset							**Results of Mouse Alternate2 Validation Dataset**				
SVC-L1	RF	79.67	69.87	75.15	0.82	0.5	77.95	83.51	80.36	0.87	0.61
mRMR	DT	63.19	63.46	63.31	0.68	0.27	71.65	69.07	70.54	0.76	0.41
SFS	LR	63.74	57.05	60.65	0.66	0.21	60.63	69.07	64.29	0.69	0.29

Note: SVC-L1: Support vector classifier with L1 regularization; mRMR: Minimum redundancy maximum relevance; SFS: Sequential feature selection; Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUC: Area under curve; MCC: Mathew's correlation coefficient.

Table 4.7: The performance of Univariate feature selection methods.

Results of Human Main Validation Dataset						
Feature	**Model**	**Sens**	**Spec**	**Accuracy**	**AUC**	**MCC**
100	**ET**	71	71	71	0.76	0.41
150	**RF**	70	69	70	0.76	0.39
200	**MLP**	74	69	72	0.76	0.43
250	**ET**	74	67	71	0.77	0.41
300	**MLP**	**75**	**70**	**73**	**0.80**	**0.45**
Results of Human Alternate1 Validation Dataset						
100	**ET**	87	42	70	0.78	0.32
150	**ET**	86	44	70	0.80	0.33
200	**ET**	89	49	74	0.81	0.43
250	**ET**	88	46	73	0.81	0.39
300	**XGB**	85	63	77	0.82	0.5
Results of Human Alternate2 Validation Dataset						
100	**ET**	80	70	75	0.82	0.5
150	**ET**	80	71	75	0.83	0.5
200	**ET**	79	74	77	0.84	0.53
250	**ET**	81	72	77	0.84	0.53
300	**ET**	81	70	76	0.85	0.51
Results of Mouse Main Validation Dataset						
Feature	**Model**	**Sens**	**Spec**	**Accuracy**	**AUC**	**MCC**
100	**GradientBoosting**	65	69	67	0.71	0.34
150	**GradientBoosting**	63	72	67	0.72	0.35
200	**ET**	72	65	69	0.73	0.37
250	**ET**	68	69	68	0.73	0.36
300	**MLP**	**74**	**76**	**75**	**0.82**	**0.5**
Results of Mouse Alternate1 Validation Dataset						
100	**LR**	69	67	68	0.74	0.36
150	**LR**	67	74	70	0.80	0.41
200	**LR**	62	75	68	0.79	0.37
250	**LR**	**73**	**72**	**73**	**0.82**	**0.45**
300	**MLP**	67	72	69	0.79	0.39
Results of Mouse Alternate2 Validation Dataset						

100	**ET**	71	79	75	0.82	0.5
150	**ET**	76	81	78	0.86	0.57
200	**ET**	80	84	81	0.87	0.63
250	**ET**	77	84	80	0.87	0.6
300	**ET**	76	86	80	0.88	0.61

Note: MLP: Multilayer Perceptron; ET: Extra Tree; LR: Logistic Regression; Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUC: Area under curve; MCC: Mathew's correlation coefficient.

4.3.5.3. *Deep Learning-Based Classifiers*

To distinguish between peptides that induce IL-4 and those that do not, we employed TabNet and a 1D-CNN architecture. The results shown in **table 4.8**, which employ CNN models, indicate that the TPC feature yields an AUC of 0.64 for the mouse host. In contrast, the RRI feature achieved the best AUC of 0.66 for the human host.

Table 4.8: The performance of deep-learning based classifier on both data.

Results of Human Main Validation Dataset							**Results of Mouse Main Validation Dataset**				
Feature	**Model**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
AAC	CNN	61	55	58	0.61	0.16	55	58	56	0.54	0.13
DPC	CNN	65	53	59	0.62	0.18	51	57	54	0.59	0.08
DPC_LEN	CNN	57	67	62	0.65	0.24	54	68	6	0.63	0.22
RRI	CNN	65	61	63	0.66	0.26	57	45	52	0.52	0.02
TPC	CNN	66	56	62	0.65	0.22	55	62	58	0.64	0.17
ALL_COMP	CNN	66	56	61	0.64	0.22	58	62	6	0.62	0.2
AAC	TabNet	44	6	51	0.51	0.04	5	58	54	0.56	0.08
DPC	TabNet	84	18	54	0.45	0.03	35	68	49	0.49	0.03
DPC_LEN	TabNet	35	74	53	0.58	0.09	1	98	43	0.51	0.05
RRI	TabNet	65	39	53	0.53	0.05	54	52	53	0.53	0.05
TPC	TabNet	3	95	45	0.53	0.06	1	98	43	0.51	0.05
ALL_COMP	TabNet	97	3	53	0.52	0.02	96	1	55	0.46	0.09
Results of Human Alternate1 Validation Dataset							**Results of Mouse Alternate1 Validation Dataset**				
AAC	CNN	81	44	67	0.71	0.27	45	64	53	0.53	0.09
DPC	CNN	84	57	74	0.73	0.42	58	67	62	0.62	0.25
DPC_LEN	CNN	82	52	71	0.74	0.36	51	65	57	0.61	0.16

RRI	CNN	78	49	67	0.71	0.28	48	62	54	0.56	0.1
TPC	CNN	1	0	63	0.5	0	65	58	62	0.65	0.22
ALL_COMP	CNN	1	0	63	0.5	0	63	58	61	0.64	0.21
AAC	TabNet	82	27	62	0.56	0.12	2	7	42	0.47	0.11
DPC	TabNet	99	4	63	0.53	0.09	35	66	48	0.5	0.01
DPC_LEN	TabNet	89	1	59	0.5	0.02	57	43	51	0.48	0
RRI	TabNet	94	19	66	0.64	0.19	36	63	48	0.52	0.01
TPC	TabNet	29	76	47	0.5	0.06	0	1	43	0.58	0
ALL_COMP	TabNet	99	0	62	0.52	0.05	67	28	5	0.48	0.06
Results of Human Alternate2 Validation Dataset							**Results of Mouse Alternate2 Validation Dataset**				
AAC	CNN	69	67	68	0.74	0.36	6	71	65	0.68	0.31
DPC	CNN	65	69	67	0.74	0.34	62	71	66	0.75	0.33
DPC_LEN	CNN	69	66	68	0.73	0.35	64	75	69	0.79	0.39
RRI	CNN	6	63	61	0.66	0.23	52	66	58	0.69	0.18
TPC	CNN	69	68	69	0.75	0.37	69	79	73	0.82	0.48
ALL_COMP	CNN	72	68	7	0.77	0.4	69	8	74	0.82	0.49
AAC	TabNet	72	45	59	0.63	0.18	56	58	57	0.57	0.14
DPC	TabNet	1	94	49	0.5	0.06	3	72	48	0.53	0.02
DPC_LEN	TabNet	59	76	67	0.68	0.35	53	56	54	0.51	0.08
RRI	TabNet	73	37	57	0.57	0.11	59	54	57	0.6	0.13
TPC	TabNet	1	1	46	0.5	0.05	0	1	43	0.55	0
ALL_COMP	TabNet	1	3	55	0.57	0.13	2	94	42	0.48	0.1

Note: AAC: Amino acid composition; DPC: Di-peptide composition; RRI: Repetitive residue information; TPC: Tri-peptide composition; Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUC: Area under curve; MCC: Mathew's correlation coefficient.

4.3.5.4. *Large Language Model*

We utilized the pre-trained ProtBert language model to extract contextual embeddings for peptide sequences and fine-tuned it for the classification of IL-4 peptides. As presented in **table 4.9**, the fine-tuned ProtBert model achieved a maximum AUC of 0.72 and an AUC of 0.75 using extracted embeddings for the human host at epoch 5. In contrast, for the mouse

host, the highest AUC of 0.66 was obtained at epoch 7 for both fine-tuned and embedding-based models.

Table 4.9: The performance of LLM based classifier on both data.

Results of Human Main Dataset						
Epoch	**Model**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
3	Finetune	79	5	65.98	0.72	0.31
	Finetune Embedding (RF)	72	69	71	0.75	0.41
5	Finetune	56	73	64.2	0.69	0.3
	Finetune Embedding (ET)	7	67	69	0.72	0.37
7	Finetune	65	69	67.46	0.7	0.35
	Finetune Embedding (ET)	69	63	66	0.71	0.32
Results of Human Alternate I Dataset						
3	Finetune	76	4	62.64	0.64	0.17
	Finetune Embedding (RF)	74	54	66	0.68	0.28
5	Finetune	69	59	65.93	0.71	0.28
	Finetune Embedding (ET)	74	54	67	0.73	0.28
7	Finetune	76	54	68.5	0.74	0.31
	Finetune Embedding (ET)	8	47	67	0.72	0.28
Results of Human Alternate II Dataset						
3	Finetune	68	64	66.86	0.72	0.33
	Finetune Embedding (RF)	77	65	72	0.76	0.43
5	Finetune	87	52	71.6	0.76	0.43

	Finetune Embedding (ET)	74	65	70	0.77	0.39
7	Finetune	6	85	72.19	0.81	0.47
	Finetune Embedding (ET)	8	72	76	0.82	0.52
Results of Mouse Main Dataset						
3	Finetune	57	63	627	0.6	0.21
	Finetune Embedding (RF)	6	59	59	0.62	0.18
5	Finetune	69	52	62.05	0.63	0.22
	Finetune Embedding (ET)	64	6	62	0.64	0.23
7	Finetune	55	67	671	0.66	0.22
	Finetune Embedding (ET)	61	62	61	0.66	0.22
Results of Mouse Alternate I Dataset						
3	Finetune	58	7	63.39	0.68	0.28
	Finetune Embedding (RF)	62	67	64	0.68	0.29
5	Finetune	55	6	58.04	0.64	0.16
	Finetune Embedding (ET)	58	63	60	0.65	0.21
7	Finetune	44	72	56.25	0.61	0.16
	Finetune Embedding (ET)	51	65	57	0.62	0.16
Results of Mouse Alternate II Dataset						
3	Finetune	69	67	68.3	0.72	0.36
	Finetune Embedding (RF)	72	69	71	0.77	0.77
5	Finetune	52	85	66.96	0.8	0.39

	Finetune Embedding (ET)	7	78	74	0.81	0.48
7	Finetune	55	82	66.96	0.79	0.38
	Finetune Embedding (ET)	65	77	71	0.79	0.42

Note: Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUC: Area under curve; MCC: Mathew's correlation coefficient.

4.3.5.5. *Ensemble Approach*

Using the MERCI and MEME-MAST algorithms, we were able to identify unique motifs and recurrent sequence patterns linked to peptides that induce IL-4 as well as those that do not. To determine their existence and predictive significance, motifs were taken from the training dataset and then looked for in the separate validation peptides. Each query peptide containing a positive motif was classified as an IL-4-inducing peptide and assigned a score of +0.5, while those without a match were assigned 0. The motif-based scores were then combined with the predicted labels from the best-performing machine learning model to form an ensemble prediction. Using the MERCI tool, the ensemble achieved a maximum AUC of 0.80 with an MCC of 0.45 for the human host, and an AUC of 0.80 with an MCC of 0.50 for the mouse host, with a maximum frequency of 10 positive motifs under default parameters. Similarly, using the MEME-MAST tool, the ensemble achieved a maximum AUC of 0.70 with an MCC of 0.31 for the human host, and an AUC of 0.70 with an MCC of 0.28 for the mouse host at a significance threshold of $p \leq 0.05$. However, the ensemble model using BLAST failed to increase performance beyond the ML model. Finally, our ensemble models using motif-based methods or BLAST were not incorporated into the web server, as their predictive performance was equal to or lower than that of the ML classifiers alone.

4.3.6. Benchmarking

In order to assess the effectiveness and suitability of the suggested strategy, we carried out a thorough benchmarking study between IL4pred2 and earlier approaches for IL-4-inducing peptide prediction. Notable advances have been made in this field by earlier research. The first computational method, IL4pred, achieved a maximum accuracy of 75.76% and an MCC of 0.51 using amino acid pair and motif-based features. Hassan et al. (2023) subsequently designed Meta-IL4, demonstrated an improved accuracy of 90.70%. The

maximum accuracy of 93% was attained by Liu et al. (2025) in their more current transformer-based model, PLM-IL4. All of these models were developed using experimentally verified IL-4 data without host differentiation. On the other hand, IL4pred2 captures interspecies variations in immune recognition by introducing a host-specific framework. Leucine had a greater average composition in human IL-4-inducing peptides, while glycine and proline were more common in mouse peptides, according to compositional analysis, which showed distinct differences in the two host using their amino acid sequences (see **Figure 4.4**).

A benchmarking study was performed using both validation and newly added datasets not used in any previous methods to compare performance across models. On the main validation datasets, the hybrid model (SVM + motif) of IL4Pred achieved accuracies of 75% and 79% for human and mouse hosts, respectively. On newly added data, it obtained 50% accuracy for human and 63% for mouse peptides. In comparison, PLM-IL4 achieved 61% (human) and 66% (mouse) accuracy on validation data and 64% (human) and 81% (mouse) on newly added data (see **Table 4.10**). The slightly higher validation performance of older models can be attributed to substantial overlap in training data approximately 60% of positive and 51% of negative sequences were common with IL4Pred2 datasets.

Table 4.10: The performance of different benchmarking models on both data.

Human Newly Added Peptide Data							**Mouse Newly Added Peptide Data**				
IL4pred Model	**IL4pred2 Data Type**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
IL4pred Main	**Main**	60.52	44.51	50.35	0.53	0.05	62.27	42.61	51.19	0.54	0.05
	Alternate1	60.52	58.27	59.82	0.60	0.17	62.27	42.17	52.00	0.53	0.05
	Altermate2	60.52	43.25	47.88	0.53	0.03	62.27	43.93	49.10	0.54	0.06
IL4pred Alternate	**Main**	60.52	64.50	63.43	0.70	0.22	62.27	69.64	67.56	0.74	0.29
	Alternate1	60.52	87.98	77.70	0.88	0.51	62.27	74.64	71.15	0.78	0.35
	Altermate2	60.52	43.31	47.92	0.53	0.03	62.27	43.93	49.10	0.54	0.06
PLM-IL4pred Model	**Main**	71	59	64	NA	0.29	60	97	81	NA	0.62
	Alternate1	71	61	68	NA	0.31	60	61	60	NA	0.21
	Altermate2	71	59	63	NA	0.27	60	63	62	NA	0.20

Note:

i) IL4pred, a previously developed method comprising two models namely Main model and alternate model where negative datasets are different.

ii) IL4pred2 method's newly added data is the data of positive and negative sequences which was previously not present in any method (IL4pred, Meta-IL4, PLM-IL4).

iii) IL4pred2 method's validation dataset is a mixture of previous method's sequences and newly added sequences.

To further confirm the host specificity of IL4Pred2 models, a reverse benchmarking analysis was carried out. The human model showed random discriminating power when applied to the mouse validation dataset, with an AUC of 0.50. Likewise, the mouse model obtained an AUC of 0.58 when tested on the human dataset (refer to **Table 11**). These findings clearly support the hypothesis that distinct host-specific models are necessary due to compositional and immunological variations between human and mouse peptides. Similar host-dependent behaviour was also observed previously in a study predicting interferon-gamma-inducing peptides[196].

Table 4.11: The performance of human vs mouse models on both validation data.

Prediction on IL4pred2 Human data on Mouse model					
Data Type	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
Human_main	56.59	37.82	47.93	0.50	-0.06
Human_alt1	55.56	35.29	47.99	0.49	-0.09
Human_alt2	82.42	30.13	58.28	0.62	0.15
Prediction on IL4pred2 Mouse data on Human model					
Data Type	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
Mouse_main	9.45	90.72	44.64	0.58	0.00
Mouse_alt1	61.42	35.05	50.00	0.52	-0.04
Mouse_alt2	52.76	63.92	57.59	0.60	0.17

Note: Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUC: Area under curve; MCC: Mathew's correlation coefficient.

4.3.7. Webserver Implementation

IL4Pred2, an advanced and user friendly web-based tool, is intended to assist the scientific community in identifying and assessing immunomodulatory peptides. The web interface was developed utilising a responsive HTML framework to ensure compatibility across

multiple devices, such as PCs, tablets, and smartphones, making it accessible and easy to use from anywhere. The three functional modules have been integrated into the webserver to predict, scan, and design IL4-inducing peptides. Researchers can predict interleukin-4 (IL-4)-inducing peptides using this method, which is essential for immune control and the development of focused therapeutics.

4.4. Discussion and Conclusion

Peptide prediction that triggers interleukin-4 (IL-4) is essential both for understanding the mechanism of host immune regulation and for the design of highly specific immunotherapeutic interventions. Proper identification of such peptides requires a careful determination of host-specific immunological and metabolic peculiarities that control the induction of cytokines. In order to fulfil these demands, the present study introduces IL4Pred2, a computational framework that has been designed independently to run on both human and mouse systems. The combination of species-specific differences in MHC class II antigen presentation, and immune signaling pathways, IL4Pred2 increases both predictive and biological relevance. In-depth positional and compositional studies demonstrate that there are specific sequence preferences of IL -4 inducing peptides between the two hosts. Regarding mouse peptides, the percentage of glycine and proline is greater, whereas human peptides have more leucine, cysteine, and phenylalanine. These disparities mirror inherent immunogenetic and evolutionary differences between humans and mice. To gain accurate prediction on heterogeneous data, the research includes alignment-based techniques (BLAST, MERCI, MEME-MAST) and alignment-free methods, such as deep learning, machine learning, and large language model-based methods.

The ET classifier was shown to outperform other ML models, with a maximum AUC of 0.76 on AAC input data for the human host. The best performance was achieved using the RF model with the best AUC of 0.70 on mouse data on DPC. The incorporation of ProtBert embeddings and DL models further improved sequence representation. The hybrid ensemble approach resulted in an overall predictive performance with an AUC of 0.80 in both host organisms. Even though sklearn based feature selection methods showed strong discriminatory power between IL-4 inducers and non-inducers, the implementation of univariate feature selection methods such as mean-based methods improved the interpretability of the models with top 300 highly discriminative features. Finally, the best-performing ML classifier models across all three datasets were incorporated into the web server and standalone tools.

Comparative analysis indicated that IL4Pred2 outperforms existing systems like IL4Pred, Meta-IL4, and PLM-IL4. Cross-host validation indicated the importance of host-specific models, with the human validation data giving only an AUC of 0.50 on the mouse model, whereas the mouse validation dataset gave an AUC of 0.58 on the human model. These results highlight the necessity of host-specific methods.

In conclusion, IL4pred2 is a next-generation, host specific tool that combines both biological understanding and advanced data-driven techniques. It provides a highly specific, convenient, and biologically significant means of identifying IL-4 inducing peptides. The publicly available web server enables its use in vaccine development, immunotherapy studies, and comparative host immunology, thus improving the immunoinformatic domain.

5. Identification of anticancer peptides using sequence based features

5.1. Introduction

Anticancer peptides (ACPs) are short sequences of amino acids, typically ranging from five to fifty residues. These peptides cause tumor cell death through several different pathways, such as membrane-disrupting, apoptosis inducing, angiogenesis inhibitory, and oncogenic signaling pathway modulating pathways[197]. In comparison with the traditional chemotherapeutic agents, ACPs can effectively penetrate cell membranes and bind to negatively charged phospholipids located on the surface of cancer cells, leading to cell death both directly through cytotoxic action and indirectly through the immune system[198]. Their versatile physicochemical characteristics and structural flexibility make them good prospects in next-generation cancer therapeutics. However, experimental screening of potent ACPs is slow, labour intensive, and expensive. Computational methods have thus become dominant due to their ability to handle large volumes of peptide data and predict anticancer potential. ML and artificial-intelligence (AI) methods identify informative features derived from sequences, such as amino-acid composition, frequency of dipeptides, hydrophobicity, charge, and polarity[69,115,117]. Such models are not only very fast in predicting peptides, but they also offer insights into the sequence-function correlation of anticancer activity.

AntiCP, AntiCP 2.0, and MLACP are some of the computational models that have been proposed to predict anticancer peptides based on various representations of sequence and learning algorithms[115,116,122]. However, most of the current models have shortcomings like small or biased training samples, redundant input features, and limited generalisability. The current research thus aims at developing a strong and explainable computational model that can facilitate the accurate detection of anticancer peptides using sequence-based characteristics. This method is intended to improve predictive performance and explain the determinants of sequence in anticancer activity by incorporating various machine-learning algorithms and feature-selection strategies to enable the rational design of peptide-based therapeutics.

The current analysis was conducted based on an experimentally validated dataset of anticancer peptides that were extracted from recently updated database of CancerPPD2[199].

This database contains 6521 entries of anticancer peptides and proteins. To conduct the current analysis, 1568 anticancer peptides (natural) between 5 and 50 residues were assembled as a positive dataset. In order to reduce bias, two negative datasets were created. The initial negative dataset is known as the main dataset which comprises of antimicrobial peptides filtered from DRAMP-v4 database[200]. The second, known as the alternate dataset, contains a random selection of peptides of the same length throughout the proteome, which were extracted from UniProt. AI-based classifiers and alignment-based techniques were used, both independently and in combination, to predict anticancer peptides with high precision. With the AAC-derived features, an ET classifier had the highest AUC of 0.93 in the alternate dataset and 0.75 AUC in the main dataset. A web server and a standalone software are also provided to ensure access by the community to the prediction pipeline.

5.2. Materials and Methods

5.2.1. Overall architecture of the study

The predictive framework was designed to systematically integrate curated anticancer peptide data, multi-scale sequence descriptors, and advanced machine learning paradigms to identify discriminative patterns associated with anticancer activity. The architectural design was guided by three core principles: (i) reliable dataset construction, (ii) comprehensive feature representation, and (iii) robust model generalization.

The workflow, illustrated in Figure 5.1, comprises dataset compilation, feature extraction at global and terminal levels, feature selection to mitigate dimensionality issues, model development using complementary AI paradigms, and deployment through a webserver for practical usability.

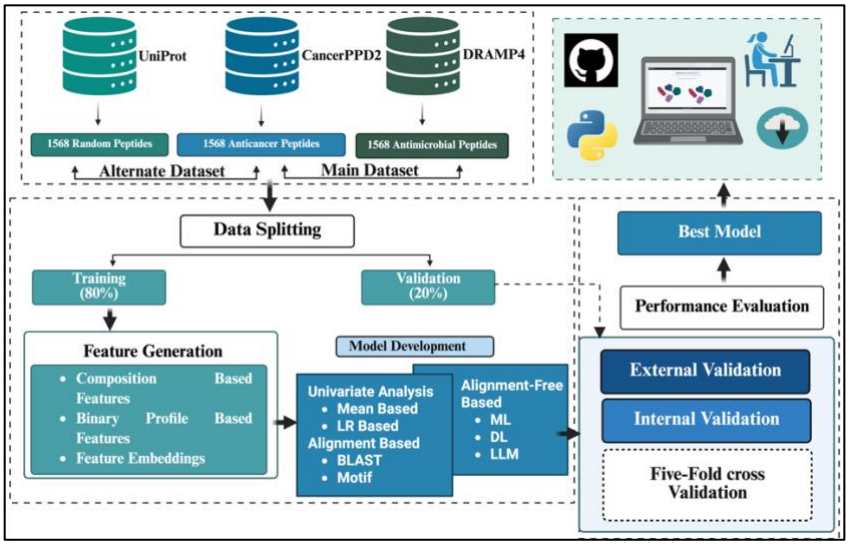

Figure 5.1: Complete pipeline and workflow of the study.

5.2.2. Creation of dataset and its pre-processing

We have extracted the experimentally evaluated natural anticancer peptides from the manually curated CancerPPD2 database. We have gathered a total of 1568 peptides of 5-50 residues as the positive dataset in this study. We have divided the datasets into Main and Alternate based on the negative data. In the main dataset the negative data were compiled from the recently updated the DRAMP database. We ensured that the negative data taken from the DRAMP database were not previously used in any anticancer prediction method. For the alternate dataset, we have collected random peptides from the UniProt database of similar length range. The main dataset utilized experimentally validated antimicrobial peptides from DRAMP as negative samples to ensure biologically relevant discrimination between anticancer and non-anticancer bioactive peptides. The alternate dataset incorporated random peptides from UniProt to evaluate model robustness against generic sequence backgrounds. This dual-dataset strategy allowed assessment of predictive stability under biologically meaningful and generic negative sampling scenarios.

5.2.3. Generation of features

Given that peptide activity may be influenced by both global residue composition and terminal-specific motifs, features were extracted at whole-sequence and terminal regions (N- and C-termini). This strategy was adopted because terminal residues are often critical for membrane interaction and cytotoxic activity. The amino acid composition-based features such as AAC, DPC, TPC, and ALLCOMP were extracted using the Pfeature tool[187]. We also calculated the composition AAC, DC and AABP of 5, 10, and 15 residues at the N- and C-terminus of peptides. We also integrated the terminal residues such as N5C5, N10C10, and N15C15 and recalculated the composition and binary profile based features. The relevant features can significantly discriminate between the two groups. As the cumulative feature space was high-dimensional, feature selection was employed to reduce redundancy, enhance interpretability, and mitigate overfitting risk. We have used various feature selection methods to reduce the feature size and find more relevant features. The size and relevance of features play a crucial role in determining the accuracy and effectiveness of prediction tasks.

5.2.4. Development of model

Multiple modeling paradigms were systematically evaluated to identify the most suitable predictive architecture for anticancer peptide classification. Classical ML classifiers were initially employed due to their robustness in moderate-sized biological datasets and their interpretability. In the ML-based category, classifiers such as DT, RF, LR, GNB, XGB, KNN, ET, and SVC were employed. To ensure the best predictive model performance, the grid search technique of scikit-learn library was used to tune the hyperparameters for each classifier.

Deep learning architectures such as 1D-CNN were implemented to capture local sequence motifs and spatial residue dependencies, while TabNet was explored for its attention-based feature selection capability in tabular high-dimensional data. Furthermore, evolutionary-scale modeling (ESM2) embeddings were incorporated to capture contextualized residue representations derived from large-scale protein pretraining. Layers t6 and t12 were specifically examined to evaluate intermediate versus deeper contextual embedding performance. Hyperparameter tuning was performed using grid search to ensure optimized classifier performance while maintaining reproducibility. Five-fold cross-validation was adopted to provide statistically reliable performance estimation while preserving sufficient training samples within each fold, particularly important given dataset size constraints.

5.2.5. Performance measures for evaluation

To determine and compare the performance of developed models, a set of evaluation metrics, divided into the threshold-dependent and threshold-independent groups, was computed. The models were tested using the MCC, sensitivity, specificity, accuracy and other threshold-depending indicators that provide information on the ability of the models to appropriately classify the anticancer and non-anticancer peptides. The general discriminative ability of the models was evaluated in terms of the threshold independent measure of AUC. These measures are described in detail in Section 4.2.4 of Chapter 4.

5.3. Results

5.3.1. Sequence based Analysis

In order to clarify the sequence level features that differentiate anticancer peptides and non-anticancer peptides, a stringent sequence-based study was conducted. This included compositional analysis, mean-based univariate analysis, and a logistic regression based

single-feature analysis. The result provides important biological and statistical insights regarding the unique residue motifs and physicochemical features that regulates the peptide activity.

5.3.1.1. Compositional Analysis

A comparison between the mean amino-acid composition of the three datasets (anticancer, antimicrobial, and random peptides) revealed strong differences in the distributions of the residues (see **Figure 5.2**). As with prior research that linked these residues with enhanced membrane affinity and selective cytotoxicity against cancer cells, alanine, phenylalanine, lysine, and leucine had the highest mean composition in the anticancer peptide cohort. In contrast, antimicrobial peptides showed enrichment of cysteine, arginine, and tryptophan, which are commonly associated with disulfide bond formation and electrostatic interactions facilitating antimicrobial activity. However, since they lacked certain structural or functional patterns, random peptides showed higher average compositions of aspartic acid, glutamic acid, methionine, asparagine, proline, glutamine, serine, threonine, valine, and tyrosine. The unique biochemical fingerprints that distinguish functional peptide classes are highlighted by these compositional variations.

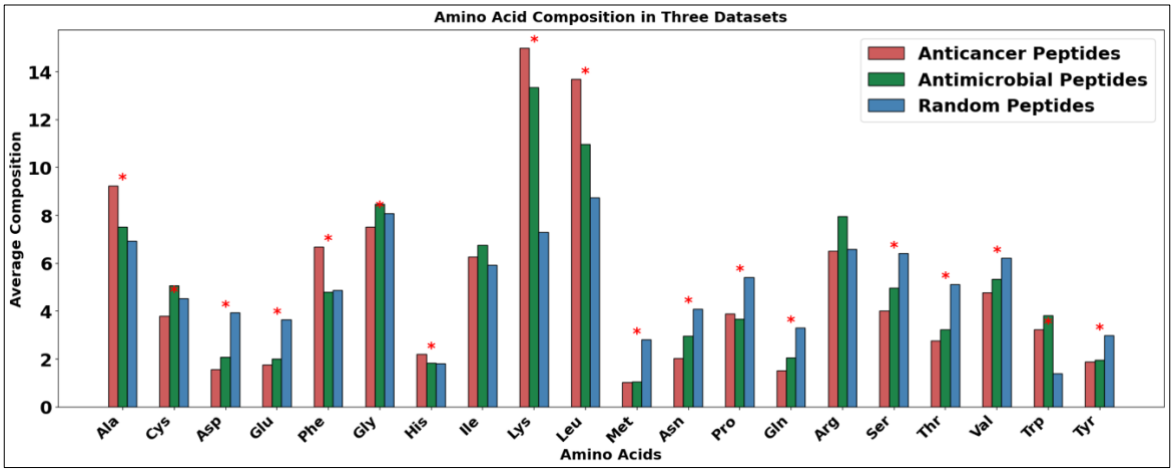

Figure 5.2: The average amino acid compositional difference between anticancer, antimicrobial and random peptides.

5.3.1.2. Mean-Based Univariate Analysis

To evaluate the discriminatory ability of single descriptors, a mean based univariate analysis was performed to compare the means of features in anticancer and non-anticancer data. Some of the descriptors showed statistically significant difference in means implying that there is a significant possibility of differentiating between the two categories of peptides. In the main dataset, such descriptors like TPC_LAK, PRI_SS_HE and DPC1-FA showed the

most significant differences in means, and this highlights their role in the discrimination of anticancer peptides. However, in comparison, the alternate dataset showed that the distinguishing characteristics like AAC_K, PAAC1_K and APAAC1_K had significantly positive differences in means which supports the importance of lysine-containing residues in anticancer activity. All these observations indicate that physicochemical descriptors are essential factors that determine anticancer potential. The comprehensive findings of the mean-based univariate analysis are provided in **Table 5.1**.

Table 5.1: Performance of the mean-based univariate analysis.

Main Data				Alternate Data			
Feature	**Pos mean**	**Neg mean**	**Diff**	**Feature**	**Pos mean**	**Neg mean**	**Diff**
TPC_LAK	0.17	-0.17	0.34	AAC_K	0.35	-0.35	0.7
PRI_SS_HE	0.17	-0.17	0.33	PAAC1_K	0.35	-0.35	0.7
DPC1_FA	0.16	-0.16	0.32	APAAC1_K	0.35	-0.35	0.7
TPC_AKK	0.16	-0.16	0.32	SER_M	0.35	-0.35	0.69
PCP_HB	0.16	-0.16	0.31	PCP_Z4	0.34	-0.34	0.68

Note: Pos: Positive; Neg: Negative; Diff: Difference.

5.3.1.3. Logistic Regression-Based Single Feature Analysis

A univariate analysis based on a logistic regression was conducted to evaluate the discriminative power of individual sequence-derived features further. A logistic regression classifier was used to assess each feature separately. Features like PRI_SS_HE, PCP_Z1, and CeTD_75_p_PZ3 showed moderate predictive power in the main dataset, suggesting that single features can sufficiently distinguish between anticancer and non-anticancer peptides. However, CeTD_100_p_SA3, CeTD_75_p_SA3, and CeTD_50_p_SA3 features in the alternative dataset obtained much higher AUC values, indicating their great predictive usefulness in peptide class differentiation. These findings suggest that sequence-derived descriptors capturing secondary structure tendencies and physicochemical properties play a significant role in anticancer peptide recognition. The detailed results of the LR-based single feature analysis are presented in **Table 5.2**.

Table 5.2: The performance of LR-based single feature analysis.

Main Data						Alternate Data					
Feature	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Feature**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**

PRI_SS_HE	58.42	55.29	56.86	0.6	0.14	CeTD_100_p_SA3	72.7	71.49	72.1	0.8	0.44
PCP_Z1	60.33	55.74	58.04	0.6	0.16	CeTD_75_p_SA3	80.29	63.52	71.91	0.79	0.44
CeTD_75_p_PZ3	53.7	60.91	57.3	0.6	0.15	CeTD_50_p_SA3	80.29	63.52	71.91	0.79	0.44
CeTD_100_p_PZ3	53.7	60.91	57.3	0.6	0.15	SEP	71.94	72.19	72.07	0.78	0.44
CeTD_33_PZ	56.38	60.46	58.42	0.6	0.17	CeTD_100_p_PZ3	73.21	70.73	71.97	0.77	0.44

Note: Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUC: Area under curve; MCC: Mathew's correlation coefficient.

5.3.2. Model Performance and Analysis

We have utilized various composition-based features and even the n- and/or c-term composition of AAC and DPC features for the classification of anticancer peptides. Various ML based classifiers such as (DT, RF, LR, KNN, XGB, GNB, SVC and ET), DL based classifiers such as 1DCNN and Tabnet and esm2 finetuned models were used to perform the classification task. Surprisingly, in both Main and Alternate datasets, ET and RF classifier outperformed other classifiers on different feature.

In the main dataset, the amino-acid composition (AAC) feature achieved highest AUC 0.75 with MCC 0.37 while c15AAC (amino acid composition of 15 residues from c-terminus) achieved the highest AUC 0.78 with MCC 0.44 using ET classifier. While in alternate dataset, amino-acid composition feature achieved highest AUC 0.93 with MCC 0.69. The n10c10AABP amino acid binary profile of main data achieved highest AUC of 0.79 with MCC 0.47 while the same residues binary profile only reaches to AUC 0.93 with MCC 0.70 at alternate dataset (please refer **table 5.3**).

Table 5.3: The performance of best ML classifiers on both datasets.

Main Validation Data							**Alternate Validation Data**					
Feature	**Model**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Model**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
AAC	ET	67.83	69.01	68.42	0.75	0.37	ET	83.44	85.67	84.55	0.93	0.69
DPC	ET	63.69	69.01	66.35	0.74	0.33	ET	84.4	82.48	83.44	0.92	0.67
TPC	ET	56.37	73.48	64.91	0.72	0.3	ET	84.71	80.57	82.64	0.9	0.65
n5 AABP	ET	64.97	73.48	69.22	0.76	0.39	ET	81.85	80.57	81.21	0.90	0.62
n10 AABP	ET	65.61	75.08	70.34	0.77	0.41	ET	82.48	85.03	83.76	0.91	0.68
n15 AABP	ET	61.78	77.64	69.70	0.77	0.40	ET	83.44	85.67	84.55	0.91	0.69

c5 AABP	ET	70.38	64.54	67.46	0.74	0.35	ET	73.89	76.12	75.00	0.83	0.50
c10 AABP	RF	67.83	67.73	67.78	0.75	0.36	ET	79.62	80.57	80.10	0.89	0.60
c15 AABP	RF	64.97	74.12	69.54	0.77	0.39	ET	82.80	79.62	81.21	0.90	0.63
n5c5 AABP	ET	64.33	73.80	69.06	0.76	0.38	ET	80.57	83.12	81.85	0.91	0.64
n10c10 AABP	ET	70.38	76.36	73.37	0.79	0.47	RF	86.31	83.44	84.87	0.93	0.70
n15c15 AABP	ET	66.88	75.72	71.29	0.77	0.43	ET	86.94	83.44	85.19	0.92	0.70
N5 AAC	ET	57.33	70.29	63.80	0.68	0.28	RF	79.62	78.66	79.14	0.89	0.58
N10 AAC	RF	60.83	67.09	63.96	0.73	0.28	ET	84.08	82.80	83.44	0.92	0.67
N15 AAC	ET	62.10	76.04	69.06	0.75	0.39	ET	84.71	86.62	85.67	0.92	0.71
C5 AAC	ET	62.10	64.86	63.48	0.69	0.27	ET	73.57	75.80	74.68	0.82	0.49
C10 AAC	ET	67.20	71.25	69.22	0.75	0.39	ET	79.62	79.94	79.78	0.89	0.60
C15 AAC	ET	72.61	70.93	71.77	0.78	0.44	ET	80.89	84.71	82.80	0.90	0.66
N5C5 AAC	ET	60.51	67.09	63.80	0.70	0.28	ET	82.48	82.17	82.33	0.91	0.65
N10C10 AAC	ET	63.69	74.44	69.06	0.77	0.38	ET	84.08	85.67	84.87	0.93	0.70
N15c15 AAC	ET	67.20	75.72	71.45	0.77	0.43	RF	85.35	86.62	85.99	0.93	0.72

Note: Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUC: Area under curve; MCC: Mathew's correlation coefficient.

We have also implemented ensemble approach using BLAST and Motif. The ensemble approach combining c15AAC model's prediction score with BLAST hits score on validation dataset of main data slightly improved AUC 0.79 with MCC 0.44, while the ensemble approach did not improve performance on the alternate dataset.

The 1DCNN model was only able to achieve the highest AUC 0.70 with MCC 0.28 using TPC features in the main dataset while it achieved an AUC of 0.89 with MCC 0.63 on AAC feature in the alternate dataset. We have also implemented a fine-tuned esm2 t6-layered model that achieve only 0.77 AUC with 0.41 MCC on main dataset and 0.93 AUC with 0.70 MCC on alternate dataset. While embeddings extracted using the finetuned esm2 model

achieved an AUC 0.76 with MCC 0.40 on main dataset and 0.92 AUC with 0.70 MCC on alternate dataset. The best performing DL model results are described below in **table 5.4**.

Table 5.4: The DL model results on different sequence features.

	Main Validation Data					Alternate Validation Data				
Feature	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
AAC_1DCNN	64.97	58.15	61.56	0.66	0.23	78.98	83.76	81.37	0.89	0.63
DPC_1DCNN	52.87	65.81	59.33	0.64	0.19	78.34	84.4	81.37	0.88	0.63
TPC_1DCNN	53.82	73.48	63.64	0.7	0.28	80.57	82.8	81.69	0.87	0.63
AAC_Tabnet	62.1	55.91	59.01	0.64	0.18	72.93	83.44	78.18	0.86	0.57
DPC_Tabnet	73.89	42.81	58.37	0.59	0.18	78.03	58.28	68.15	0.75	0.37
TPC_Tabnet	64.33	62.3	63.32	0.67	0.27	69.11	72.93	71.02	0.76	0.42

Note: Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUC: Area under curve; MCC: Mathew's correlation coefficient.

5.3.3. Comparison with existing approach

In order to predict anticancer peptides (ACPs), a variety of computational models have been developed over the last ten years. These models use deep learning-based representations, physicochemical descriptors, and various sequence-derived features. These approaches differ greatly in terms of their implementation frameworks, dataset composition, and modelling strategies. But a critical analysis shows that many of the current methods have enduring problems, such as a lack of reproducibility, a limited diversity of datasets, and inconsistent availability of web servers or code, all of which limit their wider applicability in large-scale peptide screening and validation.

Benchmarking of these models on main and alternate validation datasets revealed considerable variability in performance across methods and datasets. The classical AntiCP (2013) achieved an accuracy of 48.75% (AUC = 0.51) on the main dataset and 44.21% (AUC = 0.55) on the alternate dataset. The CNN-based ACP-MHCNN (2021) showed a small improvement in accuracy, with 45.53% on the alternative dataset and 49.08% on the main dataset. The updated AntiCP 2.0 (2021) showed a notable improvement in performance, achieving 58.47% accuracy (AUC = 0.59) on the main dataset and 68.21% accuracy (AUC = 0.68). The consistency of the deep learning methods was comparatively greater for xDEEP-AcPEP (2021) and MLACP 2.0 (2022), with MLACP 2.0 achieving the highest results on the alternate dataset with 75% accuracy and AUC = 0.85. The ACP-EPC (2025) model, which was the most current, demonstrated the efficacy of evolutionary-scale

embedding models in capturing contextual sequence dependencies by achieving 47% accuracy (AUC = 0.46) on the main dataset and 73% accuracy (AUC = 0.77). The detailed benchmarking performance of existing ACP prediction methods is presented in **tables 5.5**.

Table 5.5: The benchmarking performance of existing ACP methods.

Main Validation Data						Alternate Validation Data				
Method	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
AntiCP	54	49	52	0.51	0.03	54	64	59	0.55	0.18
ACP-MHCNN	22	74	49	NA	-0.04	22	67	46	NA	-0.12
AntiCP 2.0	59	58	58	0.59	0.17	58	79	68	0.68	0.37
xDEEP-AcPEP	80	12	46	0.46	-0.10	80	60	71	0.78	0.42
MLACP 2.0	85	22	53	0.58	0.09	85	66	75	0.86	0.52
ACP-EPC	81	15	47	0.46	-0.06	81	66	73	0.77	0.47

Note: Sens: Sensitivity; Spec: Specificity; Acc: Accuracy; AUC: Area under curve; MCC: Mathew's correlation coefficient.

5.3.4. Web-based services

To enable effective and convenient research, a web server has been built that is user-friendly for predicting and analyzing anticancer peptides (ACPs). The server uses the most efficient models identified in this study and provides an interactive interface where one can enter sequences, predict and visualise results. The platform is responsive, built on HTML and PHP frameworks, and can be used on a wide range of devices, such as computers, tablets, and smartphones. The server has three main modules namely Protein Scan, Design and Predict. The Predict module allows users to predict the anticancer potential of peptide sequences (in FASTA format) by using an improved ensemble model. The Design module aids the production of single point mutant analogs of provided peptides and the evaluation of the impact that such replacements have on anticancer effects. The Protein Scan module helps users to recognize crucial regions in a protein sequence that may have anticancer properties.

With the use of complete score of each prediction result, probability value, and classification (anticancer or non-anticancer), users may classify sequences for experimental validation. Additionally, a Python program and a standalone version can be downloaded, enabling

interaction with local computational pipelines and large-scale screening. Scikit-learn is used to construct optimal machine learning frameworks in the server's backend. Reproducibility, scalability, and fast response times are provided by effective web server administration. The tool is available for free at: http://webs.iiitd.edu.in/raghava/anticp4/.

5.4. Discussion and Conclusion

Anticancer peptide-based therapeutics have demonstrated encouraging results in the treatment of cancer. Few known anticancer drugs GnRH-targeting peptides LY2510924 and ATSP-7041 are also being used for cancer treatment[201–203]. Because the process of identifying and screening prospective ACPs in a wet lab is labour-intensive, expensive, and time-consuming, the identification of anticancer peptides through computational means has emerged as a crucial step in the rational design of peptide-based therapeutics. Despite the substantial advancements in ML and DL approaches, many of the currently available tools have limitations, including model overfitting, reliance on out-of-date datasets, and limited usefulness because of unavailability or non-functional web servers. By developing a reliable and understandable AI-driven ensemble framework for precise ACP prediction, the current study overcomes these drawbacks. The sequence-based study carried out in this work highlighted the major importance of residue-level characteristics on anticancer activity by revealing distinct compositional preferences between anticancer, antimicrobial, and random peptide groups. Furthermore, the mean-based and logistic regression-based univariate analyses identified statistically significant sequence-derived features that could discriminate anticancer peptides with high precision. The AntiCP4 framework primarily employs an amino acid binary profile for the main model and amino acid composition for the alternate model. To further improve prediction accuracy and interpretability, deep learning architectures, evolutionary-scale language models (ESM2), and several machine learning classifiers were combined through thorough feature selection and model optimization.

AntiCP (2013), AntiCP 2.0 (2021), xDEEP-AcPEP (2021), MLACP 2.0 (2022), and ACP-EPC (2025) are among the previously published methods that have been benchmarked to demonstrate that the developed framework performs competitively while maintaining complete reproducibility and accessibility. The current approach utilizes new, heterogeneous peptide data, which distinctly stands in contrast to the previous models, which relied on old and limited data, and, thus, allows better generalization to new sequences that have been identified recently. Taken together, this study provides a computational platform to predict anticancer peptides that combines biological relevance,

interpretability, and accessibility. The developed model does not only complement the existing state of the art in ACP prediction but also can be a pivotal tool for experimental scientists working in the field of peptide design and therapeutic discovery. In order to strengthen the aspect of biological validation, future plans will involve expansion of the framework to include multifunctional peptide prediction, the enhancement of structure based representations and the incorporation of molecular docking and dynamics simulations.

6. Prediction of Rheumatoid Arthritis-Inducing Peptides in an Antigen

6.1. Introduction

Rheumatoid arthritis (RA) is a chronic systemic autoimmune disease that mainly attacks the synovial joints leading to inflammation of the synovium, degradation of cartilages, and erosion of bones[204]. Having a prevalence of around 1% globally, it is a serious disease in terms of public health because of its progressive course, with the associated risk of long-term disability[205,206]. It is clinically characterized by chronic synovitis, joint deformity, effusion of the joint, and stiffness. There is also reported systemic involvement of neurological, respiratory, and cardiovascular systems[207–209]. The onset of the disease occurs between the ages of 30 and 60 years and is predominant in women, implying a relationship between hormonal and genetic factors and the risk of developing the disease[210].

The pathogenesis of RA is multifactorial and is connected with the interplay of genetic, environmental and immunological factors[211]. Human leukocyte antigen (HLA) class II alleles, specifically HLA-DRB104:01, HLA-DRB101:01, and other shared epitope carriers demonstrate the most significant level of association with RA[212]. These molecules expose extracellular peptides to CD4 + T cells, which trigger pathogenic autoimmune responses. The immunogenic process contributes to the production of autoantibodies, such as rheumatoid factor (RF) and anti-citrullinated protein antibody (ACPA) and encourages the activity of autoreactive B cells, macrophages, and dendritic cells[213]. In RA, of specific clinical interest, ACPAs are the key diagnostic and prognostic biomarkers. One of the main aspects of RA pathophysiology is dysregulation of cytokine production. The pro-inflammatory cytokines, including interleukin-6 (IL-6), interferon-gamma (IFN- γ), tumor necrosis factor-alpha (TNF- α), and interleukin-1 beta (IL-1β), perpetuate inflammation in the synovial environment[214]. These cytokines stimulate osteoclasts and fibroblast-like synoviocytes, thus facilitating the formation of pannus, cartilage destruction, and bone resorption[215]. In addition, inflammatory signaling is amplified by aberrant intracellular signaling pathways, such as NF-κB and JAK/STAT, that contribute to disease progression[216,217].

The recent treatment options in rheumatoid arthritis are aimed at suppressing inflammatory events and avoiding structural destruction of joints. The main category of drugs used to

manage RA is disease-modifying antirheumatic drugs (DMARDs), which fall into the following categories: biologic agents (e.g., TNF-α and IL-6 inhibitors), targeted synthetic DMARDS (e.g., JAK inhibitors), and conventional synthetic medications (e.g., methotrexate, hydroxychloroquine, and sulfasalazine)[215,218–220]. Methotrexate is commonly recognized as the gold standard treatment and it is commonly used as a monotherapy or combination with biologic agents to improve clinical outcome[221,222]. However, treatment advances have failed to eradicate the issue of suboptimal response where 30-40 percent of patients do not respond well. Long term immunosuppression also increases the risk of infection, drug toxicity and the cost of treatment thus necessitating alternative forms of treatment[223].

New developments in immunoinformatic and computational biology have improved the identification of autoimmune drivers and future therapeutic targets[224]. The prediction of epitopes is currently indispensable to the mapping of antigenic peptide sequences that interact with major histocompatibility complex (MHC) molecules and trigger T-cell-responses[225]. The RA-related peptides prediction provides a more profound understanding of the immune recognition system and the process of autoimmunity development[226]. The discovery of immunodominant epitopes also forms the basis of peptide based immunodulators, diagnostic biomarkers, and vaccines designed to regulate adverse immune responses[227].

The present study introduces a computational model used to predict peptides in antigenic protein sequences that induce RA. It uses ML models that are trained on experimentally verified RA-associated epitopes to distinguish between those that cause RA and those that do not. The combination of sequence features with conserved motif patterns, improved the predictive accuracy and biological interpretability of the model. The framework supports the recognition of the peptide candidates that are involved in autoimmune activation and the design of the therapeutic peptides that are unique to the host and disease settings. Finally, the proposed computational strategy will enhance precision immunology in the context of RA by combining molecular processes with translational therapeutic interventions.

6.2. Model Architecture and Performance

The RAIpred framework was designed to integrate classical sequence-derived descriptors with contextual protein language model embeddings in order to comprehensively characterize RA-inducing peptides. The architectural design was guided by three considerations: (i) limited dataset size, (ii) high-dimensional sequence feature space, and

(iii) the potential importance of contextual residue interactions in autoimmune peptide recognition.

The overall architecture is illustrated in **Figure 6.1** and consists of data preprocessing, high-dimensional feature extraction, multi-stage feature selection, model training using ML/DL paradigms, and performance evaluation.

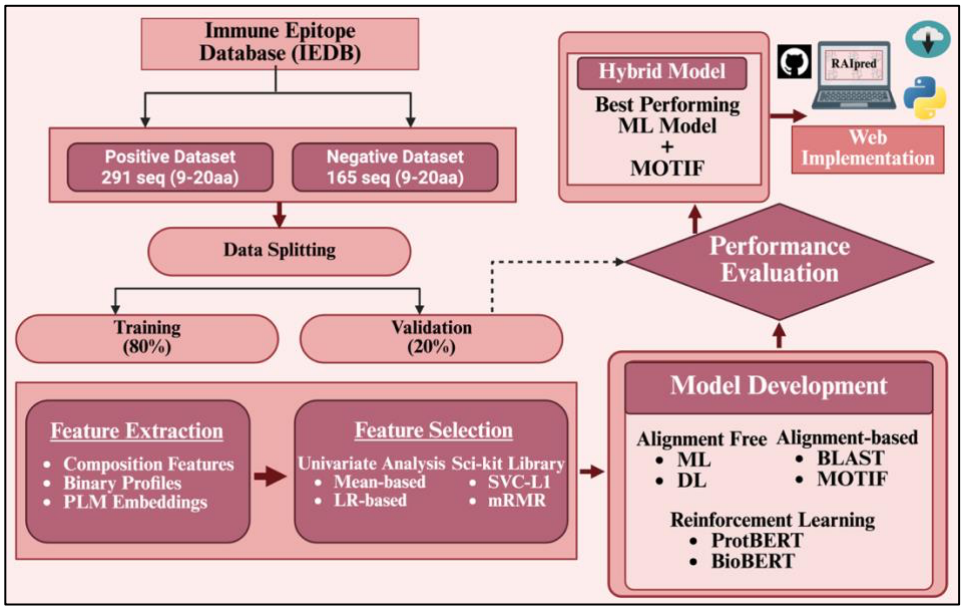

Figure 6.1: The detailed model architecture of RAIpred.

6.2.1. Data Preparation and Feature Extraction

The experimentally validated RA-inducing and non-inducing peptide dataset was retrieved from IEDB. A total of 298 HLA-class II binding RA-inducing peptides were collected as the positive dataset, while 176 HLA-class II binding RA non-inducing peptides were used as the negative dataset. As only 46 HLA-class I RA-inducing peptides were available, this class was excluded to maintain statistical reliability and avoid class imbalance. After applying redundancy removal and length filtering, 291 positive and 165 negative peptides within the length range of 9-20 residues were retained for further analysis. Restricting the dataset to HLA-class II peptides ensured biological consistency and improved modeling robustness. For feature representation, composition-based descriptors and binary profiles were computed using the Pfeature tool. In addition to handcrafted sequence descriptors, ProtBERT embeddings were generated to capture contextual and long-range residue dependencies that may not be adequately represented through conventional frequency-based features. This combined representation strategy enabled modeling of both global compositional characteristics and higher-order sequence context.

6.2.2. Feature extraction and Selection

The high-dimensional feature space (~9189 descriptors) relative to the dataset size necessitated feature selection to mitigate overfitting and improve model interpretability. The sequence-based features were extracted using the Pfeature tool including amino acid, dipeptide, tripeptide, and other types of features. The selection of relevant features facilitates better discrimination between the two categories. To select the relevant features, we emploed two different feature selection techniques from the scikit-learn Python library: SVC-L1 and mRMR. We have also implemented mean-based univariate analysis and LR-based single-feature analysis to identify significant features. SVC-L1 was employed due to its embedded sparsity-inducing regularization, which performs feature selection by shrinking irrelevant coefficients to zero. mRMR was utilized to select features that maximize relevance to the target class while minimizing redundancy among features. Univariate and LR-based analyses were additionally conducted to identify statistically significant discriminative features.

6.2.3. Model development

Given the moderate sample size and nonlinear nature of peptide immunogenicity, multiple modeling paradigms were evaluated, including classical machine learning, deep learning architectures, and ProtBERT-based classifiers. Machine learning models were prioritized for their robustness in small-to-medium biological datasets, while deep learning models were explored to capture higher-order feature interactions. ProtBERT-based models were implemented to assess whether contextual embeddings derived from large-scale protein corpora improve predictive performance over handcrafted features. An 80-20 data split was adopted to maintain an independent validation set for unbiased performance estimation. Five-fold cross-validation was applied on the training subset to ensure statistical robustness and reduce variance in performance estimation.

6.3. Result Analysis

6.3.1. Sequence analysis

Sequence based analysis includes amino acid positional as well as compositional analysis. The positional analysis revealed that glycine (G), glutamine (Q), and phenylalanine (F), isoleucine (I), tyrosine (Y), and proline (P) are predominantly present in RA-inducing (positive) peptides. A similar pattern is observed in the compositional analysis of RA-

inducing peptides, where glycine (G), proline (P), and tyrosine (Y) exhibit the highest average composition, with statistically significant p-values compared to the negative dataset (please refer figure 6.3).

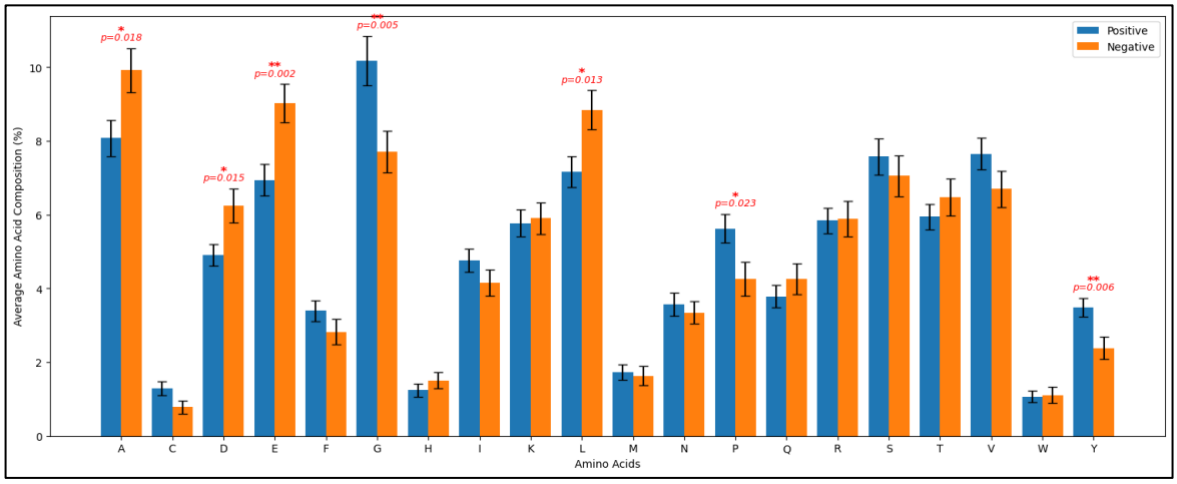

Figure 6.2: The average amino acid compositional difference between RA-inducing and RA non-inducing peptides.

6.3.2. Composition-based feature analysis

Various composition-based features such as AAC, DPC, TPC, RRI etc; and amino-acid binary profile feature were used to classify RA-inducing peptides. Our findings demonstrate that among the various composition-based features, the composition enhanced transition distribution based (CeTD) features performed exceptionally well (please refer **table 6.1**). Using CeTD based features, we achieved a maximum accuracy and AUC of 71% and 0.75 on the training dataset, and 66.30% and 0.75 on the validation dataset, with balanced sensitivity and specificity using the XGB classifier.

Table 6.1: The ML performance on validation data of different features.

Feature Name	ML Model	Sensitivity	Specificity	Accuracy	AUC	Kappa	MCC
CeTD	**XGB**	61.02	75.76	66.3	0.75	0.33	0.35
TPC	**ET**	62.71	69.7	65.22	0.74	0.3	0.31
ALLCOMP	**LR**	57.63	75.76	64.13	0.73	0.3	0.32
APAAC	**ET**	61.02	63.64	61.96	0.72	0.23	0.24
DPC	**KNN**	61.02	69.7	64.13	0.71	0.28	0.3
AAC	**ET**	59.32	60.61	59.78	0.7	0.19	0.19
BTC	**GNB**	50.85	75.76	59.78	0.69	0.23	0.26
DDR	**ET**	52.54	75.76	60.87	0.67	0.25	0.28
CTC	**SVC**	62.71	66.67	64.13	0.65	0.27	0.28
PRI	**LR**	55.93	69.7	60.87	0.63	0.23	0.25

AABP	**KNN**	55.93	51.52	54.35	0.59	0.07	0.07

Note: XGB: extreme gradient boosting, ET: extra tree, LR: logistic regression, KNN: k-Nearest Neighbors, GNB: Gaussian Naïve Baise, SVC: support vector classifier, AUC: area under curve, kappa: Cohen's kappa coefficient, MCC: Mathew's correlation coefficient.

6.3.3. Motif-based analysis

We have also performed motif-based analysis using the MERCI tool. This tool provides various classification methods to identify more relevant motifs from the query sequences. Here, we found that among other classification methods, the BETTS-RUSSELL classification method finds more relevant motifs which are able to discriminate between RA-inducing and non-inducing peptides (please refer **table 6.2**).

Table 6.2: The relevant motif list for RA-inducing peptides.

Motif	Coverage in RA-inducers
tiny polar hydrophobic G hydrophobic	22
hydrophobic L hydrophobic aliphatic small	20
hydrophobic polar hydrophobic tiny polar hydrophobic hydrophobic polar	19
aliphatic aliphatic hydrophobic polar polar aliphatic	19
small S hydrophobic G	18
hydrophobic polar A G hydrophobic	17
G small small G small	17
P hydrophobic polar polar hydrophobic	17
tiny hydrophobic S hydrophobic hydrophobic hydrophobic	16
tiny S hydrophobic G	15
tiny S tiny tiny	15

6.3.4. Ensemble Model

The top-performing ML model (based on CeTD features) and the motif-based method were combined to create an ensemble model. On the validation dataset, this ensemble model achieved the highest AUC of 0.80 and MCC of 0.45 (please refer **table 6.3**). These unique patterns aid in locating particular protein areas that may be responsible for RA.

Table 6.3: The ensemble model performance using exclusive positive motifs.

Classification Method	Requested Number of Motif	Sensitivity	Specificity	Accuracy	AUC	Kappa	MCC

None	**K20**	67.8	66.67	67.39	0.7	0.33	0.33
KOOLMAN-ROHM	**K20**	71.19	66.67	69.57	0.75	0.36	0.37
BETTS-RUSSELL	**K20**	71.19	75.76	72.83	0.8	0.44	0.45

Note: K: number of requested motifs, AUC: area under curve, kappa: Cohen's kappa coefficient, MCC: Mathew's correlation coefficient.

6.4. Discussion

This study was designed to identify RA-inducing peptides from non-inducers. These peptides were retrieved from the IEDB database which provides their detailed experimental information along with source and host. We have observed that most of the peptides are a subset of heat shock proteins, sourced from Epstein-Barr virus. These peptides highly resemble with human proteins, due to which their representation on MHC alleles induces an immune system response against self-peptides as well. In order to classify these peptides, we have implemented various classification approaches for the prediction purpose which includes univariate analysis (Mean-based analysis and LR-based analysis), alignment-based analysis (BLAST and motif), alignment free analysis (ML, DL & LLM) and Ensemble analysis.

A few significant findings from the thorough examination of RA-inducing and non-inducing peptides were noted by performing positional and average amino acid compositional analysis. It has been observed that amino acid glycine and proline are highly abundant in RA-inducers which may be responsible for peptide binding to MHC. Additionally, positional analysis shows that different amino acids have different preferences for the N-terminal and C-terminal regions of positive and negative datasets, showing the diverse roles of residues such as glycine, threonine, and alanine. The mean-based analysis identified 305 features with higher mean differences between the positive and negative datasets, along with statistically significant p-values. These features exhibit the highest mean difference between positive and negative peptides. Further, LR-based single feature analysis shows that Bond Composition (BTC) and CeTD feature achieved a maximum AUC of 0.69 and AUC of 0.68. Further, we have implemented both alignment-free and alignment-based (BLAST and motif) approaches. The BLAST-based method showed somewhat better results at larger e-values, which represent the random odds of receiving hits. However, for the validation dataset, the motif-based method produced the most accurate hits. We found that CeTD composition-based features performed better than all other composition-based features. By

using the XGB classifier, we were able to achieve balanced sensitivity and specificity with maximum accuracy and AUC over training datasets of 71% & 0.75 and 66.30% & 0.75 over validation datasets. This result underscores that CeTD features capture physicochemical peptide properties for our dataset, which is critical for accurate predictions. Finally, our ensemble model achieved a maximum AUC of 0.80 and an MCC of 0.45 on independent validation data. The development of the reliable tool "RAIPred" for predicting HLA class II binding RA-inducing peptides was made possible by the combination of compositional insights and machine learning techniques. We have provided the tool in the form of a webserver, GitHub package, standalone python package and pypi package.

6.5. Conclusion and Limitation

RAIpred is a powerful computational tool for analysing new peptides and creating peptide analogs that may find use in food biotechnology and therapeutic research. Additionally, it has potential for determining the immunogenic risk of peptides found in food goods, especially those that could trigger autoimmune reactions. Despite its efficacy, RAIpred has certain disadvantages. One notable flaw is that it was built on a small dataset that did not adequately represent sequence diversity or discover unusual immunogenic events, resulting in prediction biases.

Overall, RAIpred (https://webs.iiitd.edu.in/raghava/raipred/) is a big step forward in the computational prediction of rheumatoid arthritis-inducing peptides, giving useful insights for medication discovery and food safety evaluation. To make this method more accurate, reliable, and effective in real-life biological and nutritional conditions, it will need larger, experimentally tested, and more varied datasets in the future.

7. Determination of Specific Epitopes and Motifs in a protein associated with Celiac Disease

7.1. Introduction

Celiac disease (CD) is a chronic, immune-mediated enteropathy, which is triggered by the consumption of gluten-containing cereals (wheat, barley, rye) in predisposed people. The condition was first recognized in 1888 as a rare paediatric condition, but in 1953, a breakthrough study discovered gluten to be the main environmental cause of the condition[228,229]. Recent literature reveals that CD is one of the most common autoimmune gastrointestinal diseases with an estimated prevalence of 1.4% of the global population. However, its prevalence varies depending on the genetic background, geographic location, age, and sex[230]. CD, although common is often underdiagnosed, mostly because of its heterogeneous clinical features, which include classical gastrointestinal manifestations like diarrhoea, malabsorption and weight loss, and extra-intestinal manifestations such as anaemia, osteoporosis, neurological complications, infertility, and dermatitis herpetiformis[231]. It has a complex pathophysiology that involves immunological dysregulation, environmental exposure, and genetic predisposition. Over 95% of patients with celiac disease have particular human leukocyte antigen (HLA) class II alleles, which are HLA-DQ2 and HLA-DQ8, the major genetic determinants. HLA-DQ2 is implicated in about 94.5 percent of such cases, and HLA -DQ8 is involved in about 2.7 percent[232]. They encode heterodimeric molecules that bind peptides of gluten and present them to intestinal mucosal CD4+ T cells, which induces an adaptive immunological response[26]. These alleles are necessary but not sufficient to cause disease development and environmental factors and other genetic influences are necessary to develop into overt pathology. It is interesting to note that HLA-DQ8 is also linked to Type 1 diabetes and other autoimmune diseases, which implies common immunogenetic mechanisms[233].

Gluten, a complex of storage proteins, gliadin and glutenin, has proline and glutamine-rich motifs that cannot be fully degraded by gastrointestinal enzymes, is the environmental precipitant of CD[234,235]. These undigested peptides can be subjected to catalysis by tissue transglutaminase (tTG) and thus translocated across the intestinal epithelial barrier to form negatively charged residues which increase their affinity for HLA DQ2/DQ8 molecules[236]. As a result, the intestinal lining becomes inflamed and villous atrophic, which drives the

activation of gluten-specific CD4+ T cells producing pro-inflammatory cytokines, especially interferon-γ (IFN -G)[237]. The 33-mer fragment of α-gliadin is one of the most immunogenic gluten-derived peptides because it contains several overlapping T-cell stimulatory domains and is resistant to proteolytic degradation[238].

The diagnosis of CD is based on an amalgamation of serologic, histologic, and genetic assessments. The common tests in serologic assessment are anti-tissue transglutaminase (anti-tTG), anti-endomysial (EMA), and anti-deamidated gliadin peptide (DGP) antibodies, accepted to be indicators of disease activity[239]. Intestinal biopsies that prove the presence of intraepithelial lymphocytosis, crypt hyperplasia, and villous atrophy are usually used to provide histologic confirmation. In vitro gluten challenge tests can sometimes be used to assess immunologic reactivity, and HLA typing can also be used to further support genetic susceptibility[240]. Strict lifelong gluten-free diet is the only remaining, beyond reasonable doubt, therapeutic intervention provided in the treatment of CD, and it effectively reduces clinical symptoms and promotes mucosal healing[241]. However, the compliance with such a diet is facing significant difficulties, which can be explained by high rates of gluten additives in processed foods, the possibility of cross-contamination, and the lack of sufficient gluten-free alternatives in some geographic areas.

Recent developments in immunoinformatic and molecular immunology have produced substantial information regarding the epitope-level pathogenesis of celiac disease. A better alternative to the traditional experimental screening method is the computational identification of the immunogenic epitopes and motifs found in gluten and related proteins. Such methodologies based on bioinformatics allow quantitative prediction of potential HLA-binding sites, determination of peptide immunogenicity, and exact mapping of T-cell epitopes which induce autoimmune reactions. The current study uses a systematic computational approach in order to identify particular epitopes and conserved sequence motifs in proteins associated with CD. It provides a framework that identifies essential peptide regions that trigger immunological responses in genetically predisposed individuals by combining immunogenicity prediction methods, motif discovery software, and sequence-based analyses. The discovery of such immunodominant epitopes can have many applications, such as, designing antigen-specific immunotherapy to improve immune tolerance, screening gluten-free food products to be safe, and designing therapeutic enzymes or peptide analogs with lower immunogenicity. In the end, computational elucidation of CD-related epitopes provides a platform toward the development of accurate therapeutic strategies in autoimmune and food-related immunopathology.

7.2. Algorithm Development

The CDpred framework was developed through a structured, hypothesis-driven computational pipeline integrating biological sequence analysis and machine learning-based classification. The overall workflow is illustrated in Figure 7.2.

The algorithm development strategy was guided by three primary considerations:

- Biological relevance - CD-inducing peptides are known to exhibit conserved sequence motifs and residue preferences linked to HLA binding and immune activation.
- Dataset characteristics - The dataset comprises short peptides (9-20 residues) with moderate sample size, requiring algorithms robust to high-dimensional feature spaces.
- Model generalization - Given the biomedical application, emphasis was placed on reducing overfitting and ensuring predictive stability.

Accordingly, the framework was divided into sequence-based exploratory analysis, feature-driven machine learning modeling, and ensemble integration to improve predictive robustness.

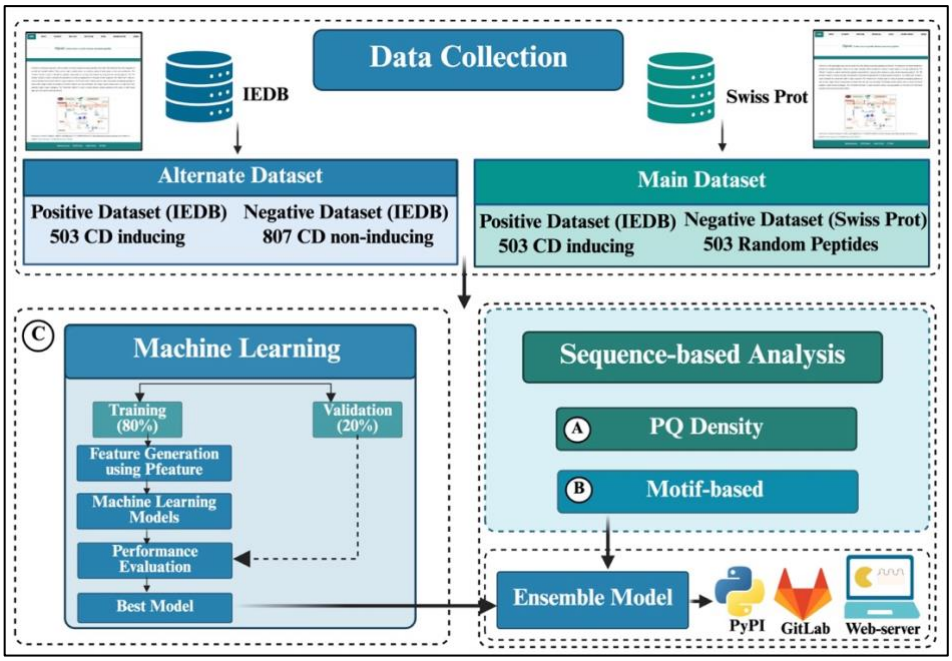

Figure 7.1: Workflow of the CDpred.

7.2.1. Dataset Preparation

To design a computational strategy to locate peptides that are associated with celiac disease, we obtained 521 experimentally validated CD-inducing peptides from the IEDB as the

positive set. Further preprocessing resulted in 503 peptides, with 9-20 amino acids. To construct a negative set we curated experimentally verified CD non-associated peptides from the IEDB and further added random peptides sampled from Swiss-Prot. The main dataset is then composed of 503 peptides related to CD along with the same number of random peptides which are used as negative samples. The alternate dataset includes 503 CD-associated peptides and 807 non-associated peptides that can be involved in other autoimmune diseases. This led to the creation of two datasets, with the first dataset being balanced having equal numbers of positive and negative samples, and the second dataset having 503 positive and 807 negative samples. A balanced dataset was constructed to prevent classifier bias toward the majority class and to enable fair learning of CD-specific sequence patterns. The alternate imbalanced dataset was used to assess model robustness under realistic class distribution scenarios, reflecting biological heterogeneity.

7.2.2. Preliminary Analysis

In order to understand the CD-inducing peptides, we first study these peptides using some preliminary studies. The first step is to determine the amino acid residues within the CD-inducing peptides through web logo and amino acid composition analysis. Subsequently we determine the frequency of HLA-alleles with and without CD-inducing and non-inducing peptides. Lastly, we determine the motif patterns that are conserved in CD-inducing, non-inducing, and random peptides through the MERCI motif tool. The insights obtained from amino acid composition and motif enrichment analysis informed subsequent feature engineering strategies, ensuring that biologically discriminative sequence characteristics were captured during model development.

7.2.3. Model development using ML & ensemble approach

Multiple machine learning paradigms were evaluated to identify an optimal classifier for CD peptide prediction. Considering the high-dimensional feature representation and moderate dataset size, ensemble tree-based methods were prioritized due to their ability to model complex nonlinear feature interactions while reducing overfitting through randomized feature partitioning and implicit feature selection. Model selection was performed based on cross-validated performance using predefined evaluation metrics. The classifier demonstrating the most consistent and superior performance across internal cross-validation and independent validation datasets was selected as the final predictive model.

To ensure reliable performance estimation, the dataset was divided into training (80%) and independent validation (20%) subsets. Five-fold cross-validation was applied to the training data to balance bias and variance in performance estimation while maintaining sufficient training samples per fold. Model evaluation was conducted using both threshold-dependent (Sensitivity, Specificity, Accuracy, MCC) and threshold-independent (AUC) metrics, as described in Section 4.2.4. to ensure comprehensive evaluation in a biomedical prediction setting.

7.3. Result and Analysis

7.3.1. Frequency of HLA-alleles

Previous studies have demonstrated a robust correlation between some HLA molecules, including HLA-DQ2/HLA-DQ8, and CD. Additionally, Sallese et al. proposed that CD susceptibility is also influenced by non-HLA genetic variations. In this study, we found that most of the CD-inducing peptides bind to HLA-DQ2/DQ8 while few also interact with other HLA alleles (refer **table 7.1**). This suggests that CD pathogenesis may involve both innate (non-HLA-DQ mediated) and adaptive (HLA-DQ mediated) immune responses.

Table 7.1: The number of specific HLA-associated peptides.

HLA		CD causing (Positive)	CD non-causing (Negative)
HLA-class I	HLA-A	13	0
HLA-class II	HLA-DQ2	263	148
	HLA-DQ8	18	110
	HLA-DQ2/ DQ8	24	9
	HLA-DR	3	402
	Other	182	138
	Total	503	807

7.3.2. Sequence pattern analysis

The specific position of an amino acid residue plays a critical role in determining the function and structural arrangement of a peptide or protein. To identify key positional preferences associated with CD, we performed both positional and compositional analyses (please refer **figure 7.2** & **7.3**). Both analyses found that amino acid proline (P) and

glutamine (Q) are highly prominent and have the highest average amino acid composition in CD-inducing peptides.

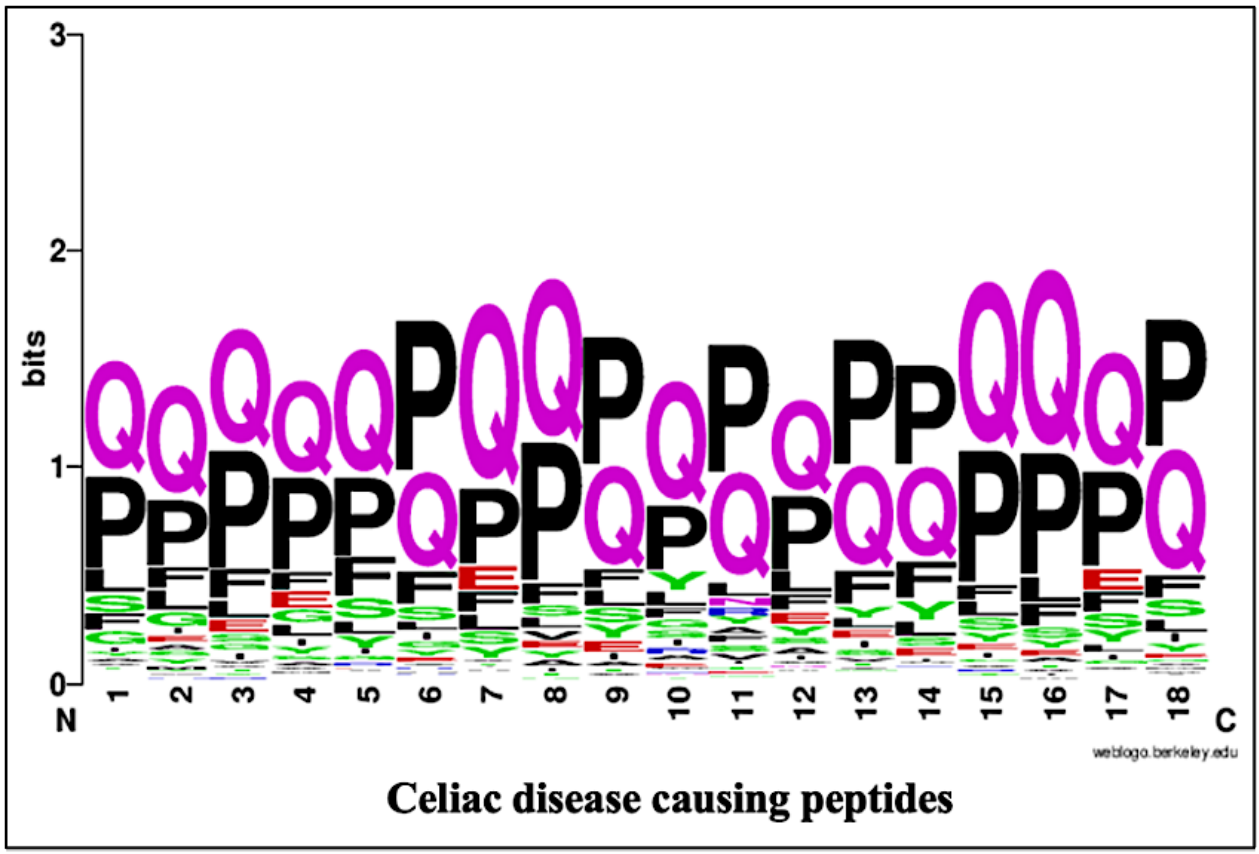

Figure 7.2: The Sequence analysis of CD-associated peptides.

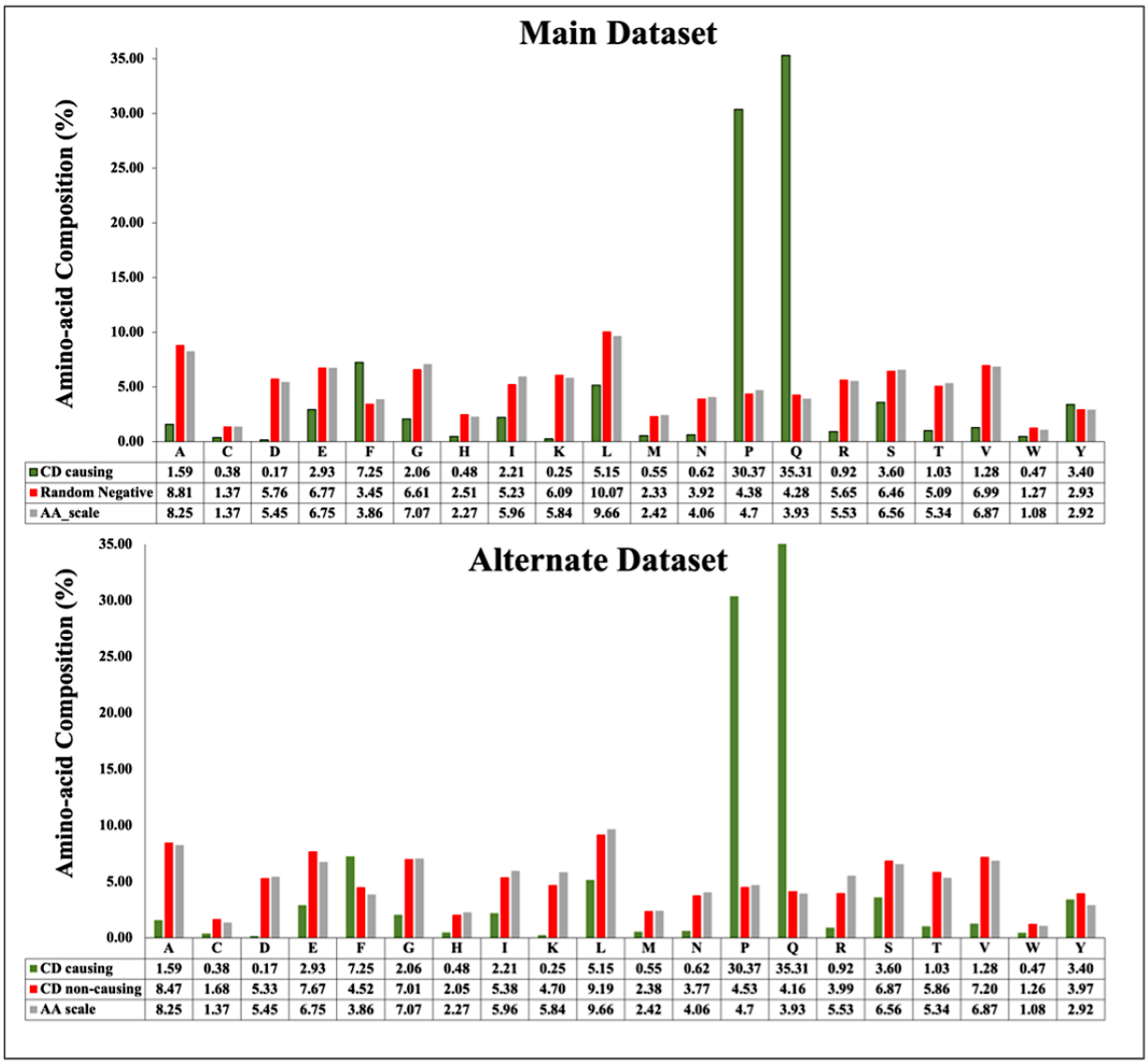

Figure 7.3: The Compositional analysis of CD-associated peptides.

7.3.3. Motif-based analysis

By applying MERCI tool, we identified 50 conserved patterns in CD-inducing peptides of varying lengths where P and Q residues are frequently identified (see **table 7.2**). We also looked for common patterns in disease causing/non-causing and random negative peptides.

Table 7.2: The conserved motif list with their occurrences.

Motifs	Positive	Random Negative	Non-CD-causing
QPF	276	0	4
QQPF	170	0	1
PYP	120	0	3
PEQ	56	0	4
QPQ	350	1	0
PQPQ	189	1	0

QQPQ	131	1	0
PQL	84	1	0

7.3.4. PQ density

Analysis of sequence patterns and motifs showed that proline (P) and glutamine (Q) residues were highly abundant in peptides that induce CD. The study makes the assumption that these residues are capable of efficiently differentiating between peptides that induce CD and those that do not. The study uses the PQ algorithm in conjunction with a density-based classification approach to verify this idea. Using sliding windows that range from three to nine residues, this method creates overlapping sequence patterns for every peptide. It determines the makeup of the P and Q residues inside each window. The study establishes the optimal composition threshold, where each peptide represents the maximum composition value for each pattern size, in order to balance sensitivity and specificity. Patterns with lengths of 5 and 6 exhibit the best discriminative performance for the primary and alternate datasets, respectively, out of all the window sizes that were investigated. **Table 7.3** displays the comprehensive findings.

Table 7.3: Performance of PQ abundance method on different window sizes.

Main Data					
Window size	**Threshold**	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**
3	0.67	85.686	98.807	92.247	0.971
4	0.51	91.65	96.62	94.135	0.977
5	0.41	93.837	94.235	94.036	0.978
6	0.34	95.427	92.445	93.936	0.978
7	0.29	96.421	91.65	94.036	0.979
8	0.38	93.241	97.018	95.129	0.981
9	0.34	94.235	96.62	95.427	0.981
Alternate Data					
Window size	**Threshold**	**Sensitivity**	**Specificity**	**Accuracy**	**AUC**
3	0.67	85.686	98.761	93.74	0.97
4	0.51	91.65	97.77	95.42	0.977
5	0.41	93.837	96.159	95.267	0.979
6	0.34	95.427	94.796	95.038	0.98
7	0.29	96.421	94.3	95.115	0.981
8	0.26	97.018	92.937	94.504	0.983
9	0.34	94.235	98.017	96.565	0.982

7.3.5. ML-based approach

Here, we first calculated the amino-acid composition of each peptide using Equation (v).

$$AACi = \frac{Ri}{L} * 100 \quad \text{(v)}$$

Where AACi is the amino-acid composition of residue type i, Ri is the number of residues of type i, and L is the length of the peptide sequence.

Using the peptide composition, we applied various ML-based classifiers for the prediction purpose. Here, we observed that the ET classifier outperformed other classifiers, achieving an AUC of 0.995 and 0.999, and an accuracy of 97.03% and 98.09% on the validation dataset in the primary and alternate dataset, with a good balance of sensitivity and specificity (please refer to **Table 7.4**).

Table 7.4: The performance of ml classifiers on AAC based features.

Main dataset								
Classifier	**Training**				**Validation**			
	Sens	**Spec**	**Acc**	**AUC**	**Sens**	**Spec**	**Acc**	**AUC**
DT	92.269	92.556	92.413	0.962	97.059	91	94.059	0.982
RF	95.262	95.533	95.398	0.989	98.039	97	97.525	0.994
LR	96.01	96.03	96.02	0.988	98.039	96	97.03	0.99
XGB	95.761	95.782	95.771	0.987	98.039	93	95.545	0.995
KNN	95.262	95.285	95.274	0.986	97.059	96	96.535	0.991
GNB	93.017	98.263	95.647	0.976	93.137	98	95.545	0.99
ET	96.01	96.03	96.02	0.991	98.039	96	97.03	0.995
SVC	95.761	95.782	95.771	0.987	97.059	96	96.535	0.991
Classifier	**Alternate dataset**							
	Training				**Validation**			
DT	92.537	92.57	92.557	0.968	94.059	99.379	97.328	0.99
RF	97.015	97.368	97.233	0.995	98.02	97.516	97.71	0.998
LR	96.269	96.285	96.279	0.99	97.03	96.273	96.565	0.987
XGB	97.015	97.059	97.042	0.992	99.01	93.168	95.42	0.998
KNN	95.771	95.975	95.897	0.992	98.02	95.652	96.565	0.995
GNB	92.537	97.059	95.324	0.977	96.04	96.273	96.183	0.983
ET	97.512	97.523	97.519	0.995	98.02	98.137	98.092	0.999
SVC	97.015	96.904	96.947	0.993	98.02	96.894	97.328	0.996

Note: DT: Decision tress; RF: Random Forest; LR: Logistic regression; XGB: extreme Gradient Boosting; KNN: k-nearest neighbour; GNB: Gaussian naïve base; ET: Extra tree classifier; SVC: support vector classifier

7.3.6. Ensemble approach

In the ensemble approach, we combine the motif-based approach with the ML-based approach. The motif-based approach achieves 81.71% accuracy in the independent dataset. A few peptides that were not identified by the motif-based approach were then covered using the machine learning method. By combining both approaches, we achieve 100% accuracy on an independent dataset (see **Table 7.5**). Our ensemble method is the most effective approach for predicting CD-associated peptides.

Table 7.5: The cumulative coverage of motifs in positive sequences.

Motif	Occurrence	Percentage	Cumulative
QPF	276	54.87	54.87
PQQP	41	8.15	63.02
PYP	33	6.56	69.58
QPQQ	28	5.57	75.15
PFP	14	2.78	77.93
PEQ	12	2.39	80.32
FPQP	4	0.8	81.11
FPQQ	2	0.4	81.51
PQLP	1	0.2	81.71
ML Prediction	92	18.29	100

7.4. Case Studies of CDpred

To assess our model's performance, we conduct a case study using a new dataset (i.e., CD-associated peptide sequences) from the AllergenOnline database (http://www.allergenonline.org), specifically the celiac disease section. This database contains 1040 unique CD-associated peptides with experimental validation. Firstly, we confirm that these peptides are not present in our training or validation data; if found, we removed the standard sequences. After this, we found that a total of 775 unique CD-associated novel sequences remained. Then, using the "Ensemble module's" default parameters, we test performance of CDpred on this unique dataset of 775 peptides and obtain 100% accuracy. The method successfully predicts all 775 CD-associated peptides, with 661 identified through the motif-based approach and 114 through the machine learning-based approach.

7.5. Discussion and Conclusion

Celiac disease (CD) is a complicated autoimmune disease that leads to intestinal inflammation and tissue damage. It is an immune system response to gluten peptide fragments. The mechanism of disease requires the identification of immunogenic epitopes of gluten and associated proteins. This knowledge is also used in the development of safer diets and treatment strategies. The given research elaborates on a computational model to precisely determine CD-associated peptides and conserved motifs that promote immune activation.

Among the main results, there is a strong discriminatory impact of proline (P) and glutamine (Q) residues in CD-inducing peptides. Proline and glutamine-rich gluten peptides are central to celiac disease pathogenesis because their high proline content confers resistance to gastrointestinal proteolysis, resulting in the persistence of long immunogenic peptides. Glutamine residues are selectively deamidated by tissue transglutaminase, increasing peptide negative charge and enhancing binding affinity to HLA-DQ2/DQ8 molecules, which promotes $CD4^+$ T-cell activation and pro-inflammatory cytokine release. Consequently, the enrichment of Pro and Gln contributes to peptide stability and heightened immunogenicity, making them key determinants of celiac disease-triggering epitopes[242,243]. The predictive performance with window size 5 and 6 yielded the best results in all datasets. The density-based classification approach using PQ was successfully used to demonstrate that P- and Q-rich patterns differentiate between inducing and non-inducing peptides in CD. This observation is consistent with previous biochemical research that has reported that proline and glutamine respectively add structural rigidity and hydrophilic properties respectively. These features increase HLA binding and immunogenicity. Frequent appearance of recurrent patterns among CD-associated peptides was also proved by motif-based analysis. These motifs were also P- and Q-enriched regions that indicated their biological utilization on antigen presentation. The motifs were combined into a machine learning (ML) system to form an ensemble model. The model was able to identify all the peptides inducing CD in the training dataset as well as external datasets. The hybrid method revealed the usefulness of motif identification coupled with ML classification. The motif analysis alone identified 661 peptides, on the other hand, the ML model identified 114 peptides that were not detected by the motif analysis. The complementary performance indicates that integrative modelling enhances sensitivity and specificity. The approach was able to predict all the 775 novel CD-associated peptides in the external dataset which

indicated its reliability. CdPred, a web server (https://webs.iiitd.edu.in/raghava/cdpred/), has been built to offer an easy to use interface to make predictions on CD-inducing peptide. It enables users to test the safety of peptides in food and medical applications, determine possible immunogenic regions and analyze protein sequences. Similar principles may be used with more general immunological applications such as predicting peptide-induced autoimmune reactions in other HLA-mediated diseases.

To summarize, CDPred is an important contribution to the computational discovery of CD-associated peptides. It is a dependable, interpretable and biologically significant predictive platform. This system consists of residue composition analysis, machine learning classification, and motif discovery. It has been extensively tested in external validation and is therefore useful as a screening and design tool. The design could be enhanced in the future with structural and physicochemical characteristics, modelling T-cell receptor interactions, and using larger datasets to enhance predictive value and translational applicability.

8. Summary

8.1. Overview

Biological macromolecules and small biomolecules are the foundations of life. Proteins and peptides are the most important of these as they are immunological mediators, signalling molecules, enzymes and receptors. The general health and cellular condition of an organism are affected by the way these molecules interact with lipids, nucleic acids, small ligands, and other molecules. These sensitive molecular interactions are usually perturbed by mutations, post-translational modifications or environmental stimuli, which often lead to disease. The increased interaction between biology and computational science has opened up unexplored possibilities to investigate, analyze, and manage such molecular processes. With the aid of bioinformatics, machine learning, and artificial intelligence (AI), it has become possible to predict immunological signalling efficiently, detect patterns associated with the disease, and design peptide therapeutics with high accuracy. This study describes the recent development of holistic computational tools to tackle immunological knowledge, disease diagnosis and therapeutic peptide prediction.

The study uses in-silico modelling, machine learning, deep learning, and protein language modelling to develop advanced tools and databases that relate diagnostics to the design of therapeutic. The study is separated into two significant sections. The first section is about developing a specialized and thorough web repository on mucormycosis, which is a severe form of fungal infection that is gaining critical clinical relevance. The second is devoted to the identification and design of bioactive peptides, such as interleukin-inducing peptides, anticancer peptides, rheumatoid arthritis-inducing peptides or celiac disease-associated peptides. The combined efforts of every study, such as data integration to algorithmic modelling, translational bioinformatics, create a coherent framework in molecular-level health research.

Surprisingly, we have identified 24 common peptides across all studies, from which 22 peptides are commonly present in IL-4-inducing peptides of the human dataset of IL4pred2. Notably, 10 peptides were shared between the human and mouse IL4pred2 datasets. Two peptides identified as mouse IL-4 inducers were also found individually in the CDpred and RAIpred datasets. Furthermore, one peptide, ENPVVHFFKNIVTPR, was common to the IL4pred2 human, IL4pred2 mouse, and RAIpred datasets. These overlapping peptides are likely to indicate similar immunogenicity and binding properties of MHC class II, that

indicates adaptive activation of immune responses. Peptides that can trigger cytokine responses (e.g. IL-4) tend to have extended immunomodulatory effects and can be used in a variety of immune conditions (e.g. autoimmune and inflammatory). Hence, the commonality observed likely represents biologically important immunogenic core peptides rather than methodological overlap, suggesting that immune-regulatory peptide landscapes are interrelated.

8.2. MucormyDB

MucormyDB, is an extensive and well-curated online database of mucormycosis which is a severe fungal infection that gained significant interest during the COVID-19 epidemic. Despite this, there was no computational resource specifically designed to combine genomic, proteomic, and therapeutic data on the specific Mucorales species causing the disease. For the first time, MucormyDB was designed to address this gap. It assembles the genome and proteome sequences of 31 fungal species, along with strain information, genome size, GC content and taxonomy identifiers. The repository has advanced analytical layers, which provide predicted protein structures, potential drug molecules, and predicted immunotherapeutic peptides using immunoinformatics based tools. The protein structures were predicted using AlphaFold2 and the DrugBank database was used to identify potential druggable targets. The multidimensionality of this database is based on the inclusion of cytokine-inducing peptides, siRNA candidates, B-cell and T-cell epitopes and nucleotide-based therapeutic adjuvants. It is an online platform that is easy to use and is hoped to motivate researchers to employ it and contribute to the growing molecular knowledge of mucormycosis. MucormyDB not only enable the research of disease genes and therapies, but also serves as an example of how the integration of computational data can generate outcomes that have therapeutic relevance. The screenshot image of MucormyDB repository is given in **Figure 8.1**. The repository is found online at https://webs.iiitd.edu.in/raghava/mucormydb/.

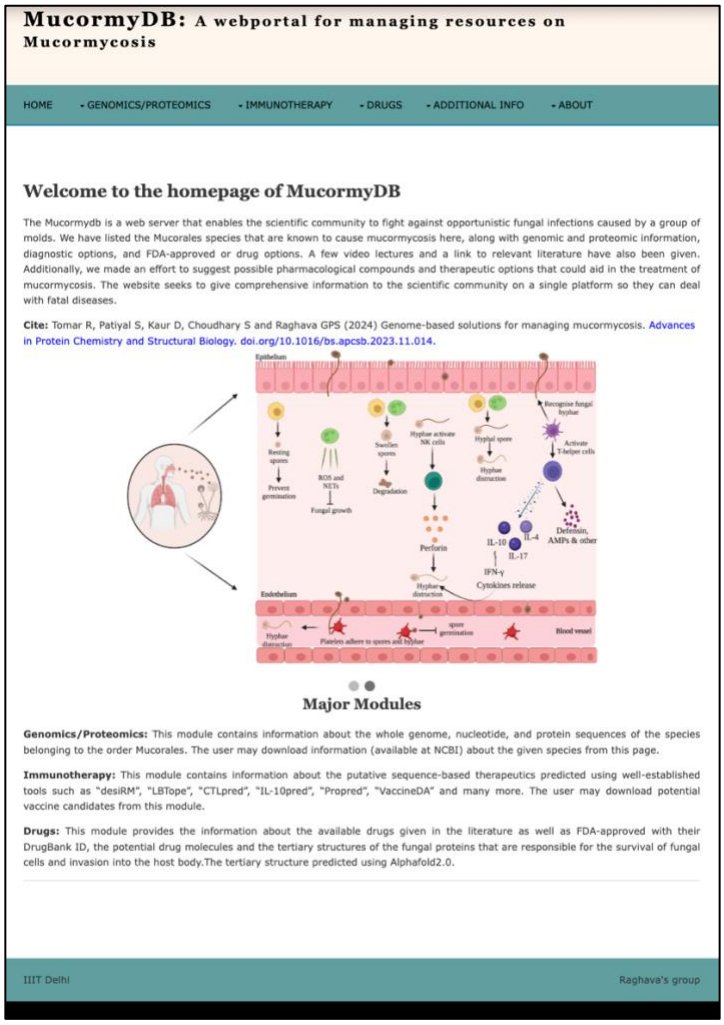

Figure 8.1: The web repository of MucormyDB.

8.3. Host-Specific Modelling of IL-4 Inducing Peptides

The following study section reinforces the disease-centered approach with the help of immunomodulatory peptide prediction and the use of biomolecules as a form of treatment. The complexity of the immune system arises because of the molecular interactions, and the cytokines like interleukin-4 (IL-4) that play a key role in the regulation of the immunological processes. The precise determination of peptides capable of triggering IL-4 activity will have a significant effect on immune modulation, allergy, and in vaccine development. This work led to the design of a host-specific predictive model of IL-4-inducing peptides. Unlike previous general models that ignored interspecies variation, IL4Pred2 identifies the unique immunogenetic properties of human and mouse hosts. A comprehensive feature extraction was done on experimentally confirmed IL-4-inducing and IL-4-non-inducing peptides available at the Immune Epitope Database (IEDB) using pfeature. Several composition-based features were computed along with the higher order physicochemical descriptors. The

best discriminative features were identified with the help of sophisticated feature selection techniques, including SVC-L1, mRMR, and sequential feature selection. Finally, the study used a multiple learning methods, both conventional machine learning classifiers, including Random Forest, XGBoost, and Support Vector Machines, as well as deep learning architectures like 1D-CNN and TabNet. Embeddings of ProtBert were applied to give contextual sequence representations, as in natural language processing. The MLP classifier was found to be the best performer in both mouse datasets (AUC 0.82), and in human datasets (AUC 0.80) using top 300 features which were selected in mean-based univariate analysis. More importantly, cross-host benchmarking reveals that models trained on one species incurred poor predictions on the other, which shows the biological necessity of the host-specific prediction. The IL4Pred2 web server integrates prediction, design and scanning modules to allow users to not only predict peptides triggering IL-4, but also to design analog mutations, and identify immunogenic regions on proteins. The screenshot image of IL4pred2 server is given in **Figure 8.2**. This comprehensive model is an important step in immunoinformatics modelling that is host specific.

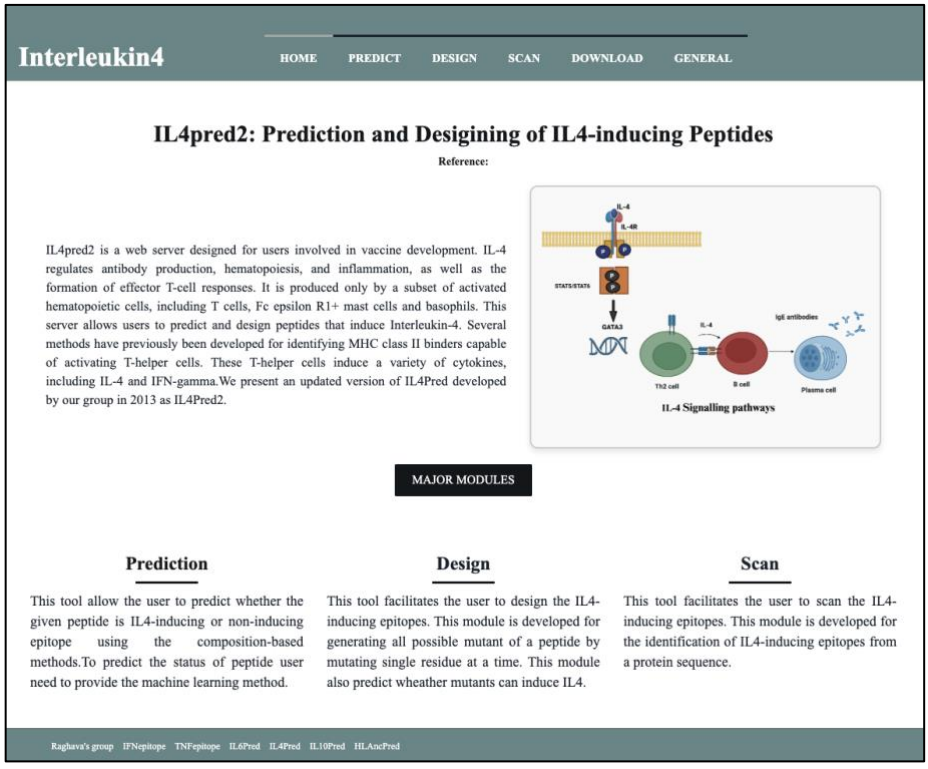

Figure 8.2: The webserver of IL4pred2.

8.4. Predicting and Designing Anticancer Peptides

Adhering to the immunological paradigm, the research then shifts into anticancer therapeutics by developing AntiCP4, a strong machine learning paradigm to detect

anticancer peptides (ACPs). Cancer remains among the major causes of mortality in the world, and though personalized medicines, chemotherapy, and radiation therapy form the backbone of treatment regimes, their limitations including the lack of specificity, systemic toxicity, and resistance to the drug is a factor that precipitates the adoption of other measures. The ACPs provide a possible treatment option as they are selectively cytotoxic to tumour cells and are less associated with adverse effects. AntiCP4 was developed using experimentally verified data retrieved from the CancerPPD2 database of 1,568 naturally occurring anticancer peptides. In order to develop a generalized model, two negative datasets were created, which included antimicrobial peptides (DRAMP-v4) and random peptides from UniProt. The sequence-derived features like amino acid composition, dipeptide composition, tripeptide composition, and terminal-specific features were calculated with Pfeature. A set of machine learning algorithms (Extra Trees, Random Forest, XGBoost, SVC, and deep learning architectures 1D-CNN, TabNet) was used to build the models.

Moreover, highly optimized ESM2 transformer models were used to calculate evolutionary embeddings. The extra Tree classifier, generated the highest performance achieving AUC of 0.93 and MCC of 0.69 on the Main dataset. Single-feature analyses using mean-based and logistic regression methods revealed the presence of important discriminative residues and physicochemical properties that are related to anticancer activity. AntiCP4 proved to be more predictive and even more transparent, reliable, and to use updated datasets compared to other current methods, including AntiCP 2.0, MLACP, and xDEEP-AcPEP. Three key modules of the user-friendly AntiCP4 web platform can be used to predict, design, and scan proteins that enable sequence-based prediction, logical design with mutation analysis, and comprehensive protein scanning to locate anticancer regions. The platform is easily accessible to both computational and experimental researchers for computational peptide therapies. The screenshot image of the webserver is given below in Figure 8.3, and the webserver is available at https://webs.iiitd.edu.in/raghava/anticp4/.

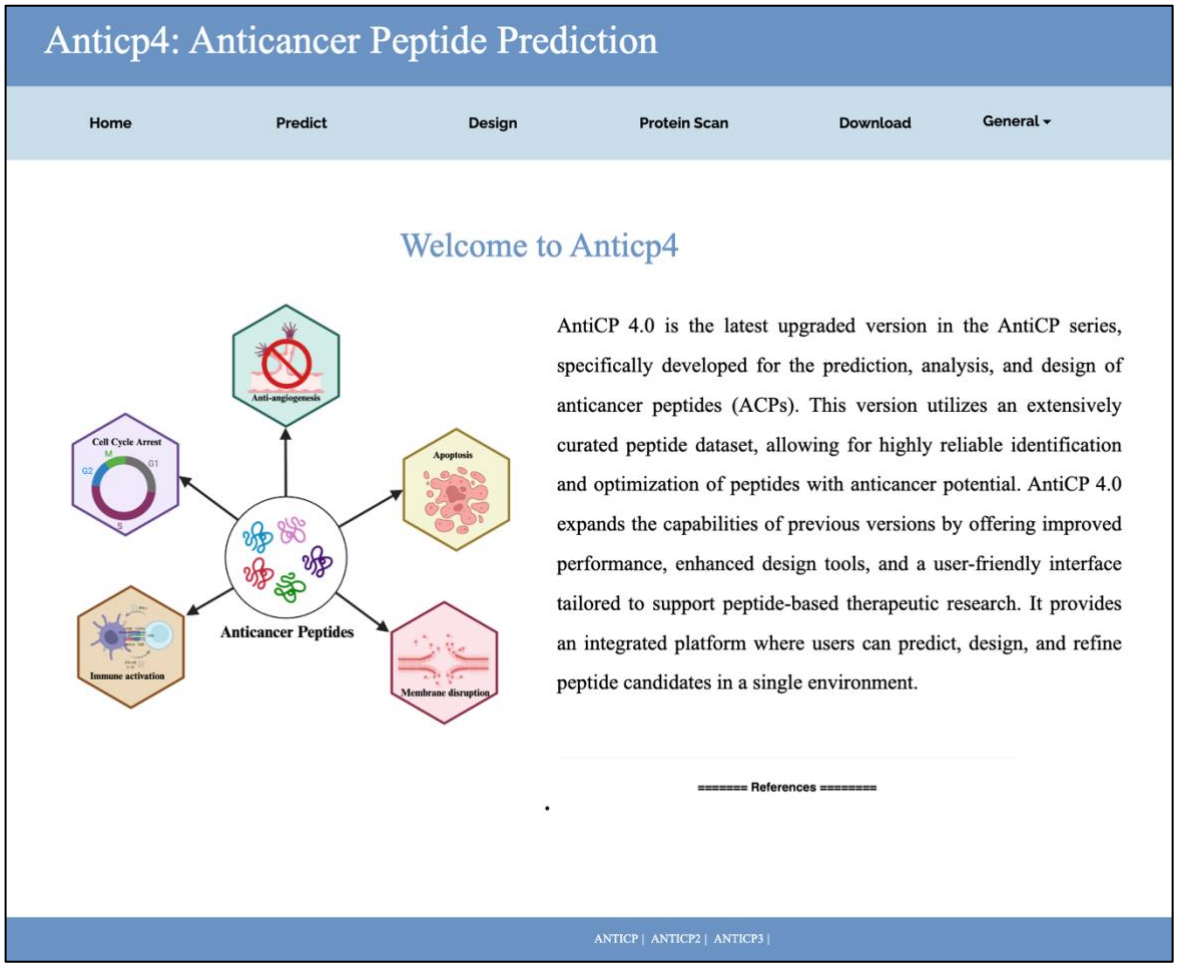

Figure 8.3: The webserver of Anticp4.

8.5. Prediction of Autoimmune Peptides in Rheumatoid Arthritis

The next step in the study is the identification of autoimmune peptides through the development of a model (RAIpred) to predict peptides that trigger rheumatoid arthritis. RA is a chronic autoimmune disorder, which is associated with chronic inflammation of the synovia, autoantibodies, and damage to the joints. The combination of the HLA class II alleles including HLA-DRB1 variations and pathogen-derived peptides results in abnormal immune activation. RAIpred was constructed based on experimentally validated RA-associated peptides retrieved from IEDB including both RA-inducing and non-inducing HLA class II binders. After calculating composition-based features and ProtBert embeddings, dimensionality reduction was performed using SVC-L1 and mRMR feature selection methods. Some of the machine learning classifiers used included Support Vector Machines, Logistic Regression and XGBoost. The best accuracy and AUC on training data (71% and 0.75) and validation data (66% and 0.75), were obtained with XGBoost model

using CeTD features. Additional motif-based insight as implemented by MERCI, specifically the BETTS-RUSSELL classification helped further to enhance the AUC to 0.80 with an MCC of 0.45. This method revealed that the number of glycine and proline residues in the RA-inducing peptides is high and could contribute to the binding of MHC and immunological cross-reactivity to self-proteins. The predictive tool that makes predictions about RA-associated peptides and calculates the risk of immunogenicity of a peptide due to its source (food or pharmaceutical), RAIpred was published as an independent application and online server. The screenshot image of the tool is presented in **Figure 8.4**. Although the model was found to be highly discriminative, its small size and the possibility of class imbalance are limitations that further studies should consider incorporating larger and experimentally verified datasets.

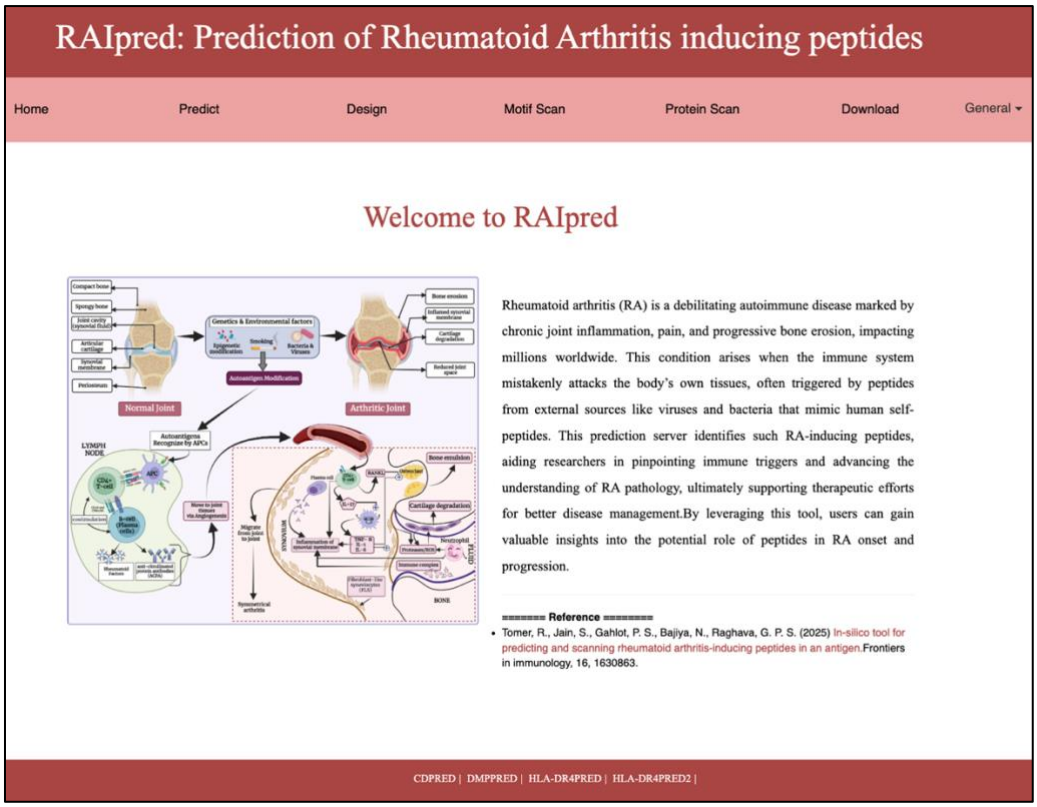

Figure 8.4: The web server of RAIpred.

8.6. Identification of Celiac Disease-Inducing Peptides

The final predictive tool developed in this study, CDpred, focuses on identifying celiac disease (CD)- inducing peptides. CD is a gluten-sensitive autoimmune disorder primarily mediated by HLA-DQ2 and HLA-DQ8 molecules, which present proline- and glutamine-rich gluten peptides to T cells, triggering intestinal inflammation and villous atrophy. CDpred was developed using 503 experimentally verified CD-inducing peptides and

multiple sets of non-inducing controls. According to preliminary sequence analysis, proline and glutamine residues predominate in CD-associated peptides, which is consistent with their high HLA binding and resistance to proteolytic digestion. To measure proline-glutamine abundance over sliding windows of different lengths and determine the best discriminatory thresholds for peptide categorization, the PQ-density algorithm was presented. Traditional ML classifiers were also evaluated, with the ExtraTree algorithm achieving near-perfect accuracy (97–98%) and AUC values approaching 0.99 on the validation datasets. The ensemble model, a combination of ML and motif-based predictions obtained 100% accuracy when tested on a separate dataset of 775 experimentally confirmed CD peptides from AllergenOnline, with motif-based analysis identifying conserved PQ-rich motifs. The solid biological basis of the PQ-density characteristic and the resilience of the ensemble approach are demonstrated by the high success rate. The CDpred web platform enables researchers to screen food proteins for immunogenic peptides, design safer peptide analogs, and explore disease mechanisms related to HLA-mediated recognition. The webserver is available at https://webs.iiitd.edu.in/raghava/cdpred, and the screenshot image is shown below in **figure 8.5**.

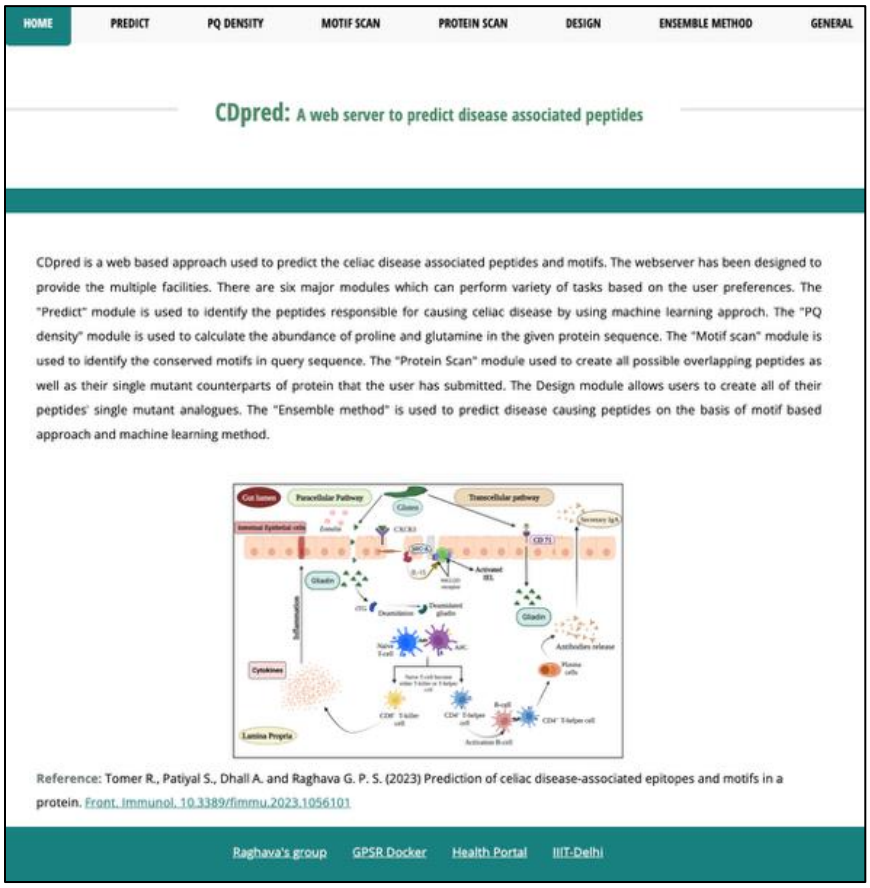

Figure 8.5: The CDpred web server.

8.7. Key Contributions

Together, these studies enhance the development of computational immunology and peptide therapeutics in several aspects. For each model, performance and interpretability are equally important in order to ensure that predictions may be associated with biologically significant characteristics like residue composition, position-specific preferences, and structural implications. Advances in methodology, like the incorporation of host-specific frameworks, ensemble learning, and large language models, show how computational pipelines can help in bridging the gap between molecular biology and clinical applications. Reproducibility and usability are also guaranteed by the creation of independent packages and publicly accessible web servers. Researchers can utilize these tools in various fields for peptide screening, hypothesis formulation, and experimental prioritization.

8.8. Limitations and Future Directions

The studies admit several limitations in spite of their success. Model generalization is limited in many fields (especially RA) by the lack of experimentally validated datasets. Since it is challenging to prove authentic non-immunogenic peptides experimentally, negative dataset production remains a difficulty. Furthermore, structural and dynamic elements of peptide-receptor interactions are still mostly unexplored in current frameworks, despite the fact that composition and sequence-based properties capture crucial information. Molecular dynamics simulations, docking, and structure-based modelling might all be used to enhance mechanistic understanding and improve predictions. Higher-order contextual and structural information can be captured in a promising way by combining multimodal learning architectures with newly developed protein language models.

8.9. Conclusion

To summarize the current work, a logical computational ecosystem has been designed that combines feature engineering, data curation, AI-based modelling, and translational application. It demonstrates how clever algorithms may speed up biomedical research by tackling both the diagnostic and therapeutic fronts, from host-specific cytokine production and immunotherapeutic peptide identification to the integration of fungal pathogen data. In addition to being scientifically proven, the developed methods namely MucormyDB, IL4Pred2, AntiCP4, RAIpred, and CDpred, are also easily accessible, reliable, and flexible enough to face new biological and clinical problems in the future.

The predictions generated by these models can be experimentally validated using established immunological and biochemical assays tailored to the specific peptide function. Predicted active peptides can first be synthesized and screened in vitro using functional assays, such as cytokine secretion assays (e.g., ELISA or ELISpot for IL-4), T-cell activation or proliferation assays, and peptide-HLA binding assays where relevant. For disease-associated or immunogenic peptides, cellular assays using primary human immune cells or relevant cell lines can be employed to assess functional responses relative to predicted activity. In the case of anticancer peptides, cytotoxicity assays (e.g., MTT/XTT or live-dead assays) against cancer and normal cell lines can evaluate selective anticancer activity. Promising candidates may subsequently be validated in vivo using appropriate disease or tumor models to confirm biological relevance and efficacy.

These tools can serve as early-stage in silico screening platforms within clinical and drug-discovery pipelines. By prioritizing candidate peptides based on predicted immunological or therapeutic relevance, they reduce the experimental search space and associated time and cost. High-confidence predictions can be advanced to targeted synthesis and functional validation, enabling faster identification of disease-associated epitopes, immunomodulatory peptides, or anticancer candidates. Clinically, such tools may support hypothesis generation, biomarker discovery, and patient stratification by highlighting peptides linked to specific immune responses or disease states. In drug-discovery workflows, they can complement experimental screening by guiding lead selection and optimization, particularly in settings where data are limited and interpretability is critical. When combined, they highlight the expanding contribution of computational biology to understanding immunological systems, directing the generation of peptides, and eventually enhancing precision medicine.

List of Publications

Thesis Related Publications

Peer Reviewed Papers

- **R. Tomer**, S. Patiyal, A. Dhall, G.P.S. Raghava, Prediction of celiac disease associated epitopes and motifs in a protein , Front. Immunol. . 14 (2023). https://www.frontiersin.org/articles/10.3389/fimmu.2023.1056101.
- **Tomer, R.,** Jain, S., Gahlot, P. S., Bajiya, N., & Raghava, G. P. S. (2025). In-silico tool for predicting and scanning rheumatoid arthritis-inducing peptides in an antigen. Frontiers in immunology, 16, 1630863. https://doi.org/10.3389/fimmu.2025.1630863
- **R. Tomer**, S. Patiyal, D. Kaur, S. Choudhury, G.P.S. Raghava, Genome-based solutions for managing mucormycosis, Adv. Protein Chem. Struct. Biol. 139 (2024) 383—403. https://doi.org/10.1016/bs.apcsb.2023.11.014.

Preprints and Paper under review/preparation

- **R. Tomer**, N.K. Mehta, S. Malik, S. Jain, G.P.S. Raghava, IL4Pred2: Prediction of Interleukin-4 Inducing Peptides in Human and Mouse, BioRxiv. (2025). https://doi.org/10.1101/2025.04.23.650150. **(Preprint)**
- **R. Tomer,** M. Chauhan, G.P.S. Raghava, Anticp4: An updated model fro predicting anticancer peptides using sequence derived features. **(Under Preparation)**

Other Publications

- N. Kumar, S. Patiyal, S. Choudhury, **R. Tomer**, A. Dhall, G.P.S. Raghava, DMPPred: a tool for identification of antigenic regions responsible for inducing type 1 diabetes mellitus., Brief. Bioinform. 24 (2023). https://doi.org/10.1093/bib/bbac525.
- A. Arora, D. Kaur, S. Patiyal, D. Kaur, **R. Tomer**, G.P.S. Raghava, SalivaDB-a comprehensive database for salivary biomarkers in humans., Database (Oxford). 2023 (2023). https://doi.org/10.1093/database/baad002.
- M. Chauhan[#], A. Gupta[#], **R. Tomer[#]**, G.P.S. Raghava, CancerPPD2: an updated repository of anticancer peptides and proteins, Database. 2025 (2025) baaf030. https://doi.org/10.1093/database/baaf030.

- Jain S, **Tomer R**, Patiyal S, Raghava GPS. NfκBin: a machine learning based method for screening TNF-α induced NF-κB inhibitors. Front Bioinforma 2025;Volume 5-. https://doi.org/10.3389/fbinf.2025.1573744.
- S. Malik, **R. Tomer**, A. Arora, G.P.S. Raghava, Identification of Multiple Prognostic Biomarker Sets for Risk Stratification in SKCM, Front Bioinforma 2025;Volume 5- https://doi.org/10.3389/fbinf.2025.1624329 .
- Bajiya N, Najrin S, Kumar P, Choudhury S, **Tomer R**, Raghava GPS. CPPsite3: An updated large repository of experimentally validated cell-penetrating peptides. Drug Discov Today 2025:104421. https://doi.org/10.1016/j.drudis.2025.104421.

URL of Computational Resources

The list of web-services developed during this study.

Name	Webserver	Dataset	Standalone
MucormyDB	*mucormydb	-	-
IL4pred2	*il4pred2	*il4pred2/download.html	#il4pred2
AntiCP4	*anticp4	*anticp4/download.html	#anticp4
CDpred	*cdpred	* cdpred/dataset.php	@cdpred
RAIpred	*raipred	*raipred/download.html	#raipred

*: https://webs.iiitd.edu.in/raghava/; #: https://github.com/raghavagps/ ;

@: https://gitlab.com/raghavalab/

References

1. Datta, L. P., Manchineella, S. & Govindaraju, T. Biomolecules-derived biomaterials. *Biomaterials* **230**, 119633 (2020).
2. Kumar, J., Narnoliya, L. K. & Alok, A. A CRISPR Technology and Biomolecule Production by Synthetic Biology Approach. in *Current Developments in Biotechnology and Bioengineering* 143–161 (Elsevier, 2019).
3. Ko, H.-J. & Lee, C. H. Emerging and Promising Keywords in Biomolecules and Therapeutics for 21st Century Diseases. *Biomol Ther (Seoul)* **33**, 1–4 (2025).
4. Zhong, Q. *et al.* Protein posttranslational modifications in health and diseases: Functions, regulatory mechanisms, and therapeutic implications. *MedComm (2020)* **4**, e261 (2023).
5. Thitame, S. N. & Aher, A. A. Algal Biomolecules in Cardiovascular Disease: A Review of Current Evidence and Emerging Therapeutic Avenues. *J Pharm Bioallied Sci* **17**, S12–S15 (2025).
6. Seo, J. H. *et al.* Protein and Peptide in Cancer Research: From Biomarker to Biotherapeutics. *Cancers (Basel)* **17**, 3031 (2025).
7. Sundaresan, V. M. *et al.* Prostate-specific antigen screening for prostate cancer: Diagnostic performance, clinical thresholds, and strategies for refinement. *Urol Oncol* **43**, 41–48 (2025).
8. Escadafal, C., Incardona, S., Fernandez-Carballo, B. L. & Dittrich, S. The good and the bad: using C reactive protein to distinguish bacterial from non-bacterial infection among febrile patients in low-resource settings. *BMJ Glob Health* **5**, e002396 (2020).
9. Hoffmann, M. H. *et al.* The cathelicidins LL-37 and rCRAMP are associated with pathogenic events of arthritis in humans and rats. *Annals of the Rheumatic Diseases* **72**, 1239–1248 (2013).

10. Mleczko, M., Kowalska-Kępczyńska, A., Gerkowicz, A., Kowal, M. & Krasowska, D. Serum Levels of Human Neutrophil Peptides 1–3 (HNP1–3) as Potential Biomarkers in Psoriasis and Associated Comorbidities. *Biomedicines* **13**, 1635 (2025).
11. Gogia, P., Ashraf, H., Bhasin, S. & Xu, Y. Antibody–Drug Conjugates: A Review of Approved Drugs and Their Clinical Level of Evidence. *Cancers* **15**, 3886 (2023).
12. Li, C. M. *et al.* Novel Peptide Therapeutic Approaches for Cancer Treatment. *Cells* **10**, 2908 (2021).
13. Vadevoo, S. M. P. *et al.* Peptide-based targeted therapeutics and apoptosis imaging probes for cancer therapy. *Arch. Pharm. Res.* **42**, 150–158 (2019).
14. Wang, L. *et al.* Therapeutic peptides: current applications and future directions. *Sig Transduct Target Ther* **7**, 48 (2022).
15. De Gaetano, S. *et al.* Global Trends and Action Items for the Prevention and Control of Emerging and Re-Emerging Infectious Diseases. *Hygiene* **5**, 18 (2025).
16. Alsharksi, A. N., Sirekbasan, S., Gürkök-Tan, T. & Mustapha, A. From Tradition to Innovation: Diverse Molecular Techniques in the Fight Against Infectious Diseases. *Diagnostics* **14**, 2876 (2024).
17. Berger, B., Daniels, N. M. & Yu, Y. W. Computational biology in the 21st century: scaling with compressive algorithms. *Commun. ACM* **59**, 72–80 (2016).
18. Khan, Y. S. & Farhana, A. Histology, Cell. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).
19. LaPelusa, A. & Kaushik, R. Physiology, Proteins. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).
20. Chu, W.-T. & Zheng, Q.-C. Conformational Changes of Enzymes and DNA in Molecular Dynamics. in *Advances in Protein Chemistry and Structural Biology* vol. 92 179–217 (Elsevier, 2013).

21. Rehman, I. & Botelho, S. Biochemistry, Tertiary Protein Structure. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).

22. Morris, R., Black, K. A. & Stollar, E. J. Uncovering protein function: from classification to complexes. *Essays in Biochemistry* **66**, 255–285 (2022).

23. Md Fadilah, N. I. *et al.* Discovery of bioactive peptides as therapeutic agents for skin wound repair. *J Tissue Eng* **15**, 20417314241280360 (2024).

24. Chatterjee, P., Chiasson, V. L., Bounds, K. R. & Mitchell, B. M. Regulation of the Anti-Inflammatory Cytokines Interleukin-4 and Interleukin-10 during Pregnancy. *Front. Immunol.* **5**, (2014).

25. Yu, Y. *et al.* Cytokines Interleukin 4 (IL-4) and Interleukin 10 (IL-10) Gene Polymorphisms as Potential Host Susceptibility Factors in Virus-Induced Encephalitis. *Med Sci Monit* **23**, 4541–4548 (2017).

26. Ciccocioppo, R., Di Sabatino, A. & Corazza, G. R. The immune recognition of gluten in coeliac disease. *Clinical and Experimental Immunology* **140**, 408–416 (2005).

27. Ishina, I. A. *et al.* MHC Class II Presentation in Autoimmunity. *Cells* **12**, 314 (2023).

28. Jain, S., Gupta, S., Patiyal, S. & Raghava, G. P. S. THPdb2: compilation of FDA approved therapeutic peptides and proteins. *Drug Discovery Today* **29**, 104047 (2024).

29. Orellana, L. Large-Scale Conformational Changes and Protein Function: Breaking the in silico Barrier. *Front. Mol. Biosci.* **6**, 117 (2019).

30. Chitluri, K. K. & Emerson, I. A. The importance of protein domain mutations in cancer therapy. *Heliyon* **10**, e27655 (2024).

31. Mahé, M., Rios-Fuller, T. J., Karolin, A. & Schneider, R. J. Genetics of enzymatic dysfunctions in metabolic disorders and cancer. *Front Oncol* **13**, 1230934 (2023).

32. Ibrahim, A. S. Host-iron assimilation: pathogenesis and novel therapies of mucormycosis. *Mycoses* **57**, 13–17 (2014).

33. Hashemi, S., Vosough, P., Taghizadeh, S. & Savardashtaki, A. Therapeutic peptide development revolutionized: Harnessing the power of artificial intelligence for drug discovery. *Heliyon* **10**, e40265 (2024).

34. Mamoudou, H. & Mune, M. A. M. AI-driven bioactive peptide discovery of next-generation metabolic biotherapeutics. *Appl. Food Res.* **5**, 101291 (2025).

35. Wan, F., Wong, F., Collins, J. J. & de la Fuente-Nunez, C. Machine learning for antimicrobial peptide identification and design. *Nat. Rev. Bioeng.* **2**, 392–407 (2024).

36. Bustin, S. A. & Jellinger, K. A. Advances in molecular medicine: Unravelling disease complexity and pioneering precision healthcare. *Int. J. Mol. Sci.* **24**, 14168 (2023).

37. Xie, Y., Chen, X., Xu, M. & Zheng, X. Application of the human proteome in disease, diagnosis, and translation into precision medicine: Current status and future prospects. *Biomedicines* **13**, 681 (2025).

38. Tanca, A., Deligios, M., Addis, M. F. & Uzzau, S. High throughput genomic and proteomic technologies in the fight against infectious diseases. *J. Infect. Dev. Ctries.* **7**, 182–190 (2013).

39. Oulas, A. *et al.* Systems Bioinformatics: increasing precision of computational diagnostics and therapeutics through network-based approaches. *Brief. Bioinform.* **20**, 806–824 (2019).

40. You, Y. *et al.* Artificial intelligence in cancer target identification and drug discovery. *Signal Transduct. Target. Ther.* **7**, 156 (2022).

41. Patiyal, S. *et al.* A web-based platform on Coronavirus disease-19 to maintain predicted diagnostic, drug, and vaccine candidates. *Monoclon. Antib. Immunodiagn. Immunother.* **39**, 204–216 (2020).

42. Gupta, A. K. *et al.* ZikaVR: An integrated Zika virus resource for genomics, proteomics, phylogenetic and therapeutic analysis. *Sci. Rep.* **6**, 32713 (2016).

43. Navratil, V. *et al.* VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. *Nucleic Acids Res.* **37**, D661-8 (2009).

44. Gupta, A. K. *et al.* CoronaVR: A computational resource and analysis of epitopes and therapeutics for severe acute respiratory syndrome Coronavirus-2. *Front. Microbiol.* **11**, 1858 (2020).

45. Brister, J. R. *et al.* Towards viral genome annotation standards, report from the 2010 NCBI Annotation Workshop. *Viruses* **2**, 2258–2268 (2010).

46. Balushi, A. A. *et al.* COVID-19-associated mucormycosis: An opportunistic fungal infection. A case series and review. *Int. J. Infect. Dis.* **121**, 203–210 (2022).

47. Bouza, E., Muñoz, P. & Guinea, J. Mucormycosis: an emerging disease? *Clin. Microbiol. Infect.* **12**, 7–23 (2006).

48. Hedeler, C. *et al.* e-Fungi: a data resource for comparative analysis of fungal genomes. *BMC Genomics* **8**, 426 (2007).

49. Rossignol, T. *et al.* CandidaDB: a multi-genome database for Candida species and related Saccharomycotina. *Nucleic Acids Res.* **36**, D557-61 (2008).

50. Cerqueira, G. C. *et al.* The Aspergillus Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. *Nucleic Acids Res.* **42**, D705-10 (2014).

51. Kuan, C. S. *et al.* DemaDb: an integrated dematiaceous fungal genomes database. *Database (Oxford)* **2016**, baw008 (2016).

52. Basenko, E. Y. *et al.* FungiDB: An integrated bioinformatic resource for fungi and Oomycetes. *J. Fungi (Basel)* **4**, (2018).

53. Urban, M. *et al.* PHI-base - the multi-species pathogen-host interaction database in 2025. *Nucleic Acids Res.* **53**, D826–D838 (2025).

54. Lew-Smith, J., Binkley, J. & Sherlock, G. The Candida Genome Database: annotation and visualization updates. *Genetics* **229**, (2025).

55. Yin, L., Yuvienco, C. & Montclare, J. K. Protein based therapeutic delivery agents: Contemporary developments and challenges. *Biomaterials* **134**, 91–116 (2017).

56. Zheng, B., Wang, X., Guo, M. & Tzeng, C.-M. Therapeutic peptides: Recent advances in discovery, synthesis, and clinical translation. *Int. J. Mol. Sci.* **26**, 5131 (2025).

57. Lombardi, L., Genio, V. D., Albericio, F. & Williams, D. R. Advances in peptidomimetics for next-generation therapeutics: Strategies, modifications, and applications. *Chem. Rev.* **125**, 7099–7166 (2025).

58. Keegan, A. D., Leonard, W. J. & Zhu, J. Recent advances in understanding the role of IL-4 signaling. *Fac. Rev.* **10**, 71 (2021).

59. Zhu, J. T helper 2 (Th2) cell differentiation, type 2 innate lymphoid cell (ILC2) development and regulation of interleukin-4 (IL-4) and IL-13 production. *Cytokine* **75**, 14–24 (2015).

60. Dhanda, S. K., Gupta, S., Vir, P. & Raghava, G. P. S. Prediction of IL4 inducing peptides. *Clin. Dev. Immunol.* **2013**, 263952 (2013).

61. Hassan, M. T., Tayara, H. & Chong, K. T. Meta-IL4: An ensemble learning approach for IL-4-inducing peptide prediction. *Methods* **217**, 49–56 (2023).

62. Liu, R. *et al.* PLM-IL4: Enhancing IL-4-inducing peptide prediction with protein language model. *Comput. Biol. Chem.* **118**, 108448 (2025).

63. Ghavimi, R., Mahmoudi, S., Mohammadi, M., Khodamoradi, E. & Jahanian-Najafabadi, A. Exploring the potential of anticancer peptides as therapeutic agents for cancer treatment. *Res. Pharm. Sci.* **20**, 165–187 (2025).

64. Chiangjong, W., Chutipongtanate, S. & Hongeng, S. Anticancer peptide: Physicochemical property, functional aspect and trend in clinical application (Review).

Int. J. Oncol. **57**, 678–696 (2020).

65. Hajisharifi, Z., Piryaiee, M., Mohammad Beigi, M., Behbahani, M. & Mohabatkar, H. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol.* **341**, 34–40 (2014).
66. Akbar, S., Hayat, M., Iqbal, M. & Jan, M. A. iACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space. *Artif. Intell. Med.* **79**, 62–70 (2017).
67. Kabir, M. *et al.* Intelligent computational method for discrimination of anticancer peptides by incorporating sequential and evolutionary profiles information. *Chemometr. Intell. Lab. Syst.* **182**, 158–165 (2018).
68. Yi, H.-C. *et al.* ACP-DL: A deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. *Mol. Ther. Nucleic Acids* **17**, 1–9 (2019).
69. Boopathi, V. *et al.* MACPpred: A support vector machine-based meta-predictor for identification of anticancer peptides. *Int. J. Mol. Sci.* **20**, 1964 (2019).
70. Wei, L., Zhou, C., Su, R. & Zou, Q. PEPred-Suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning. *Bioinformatics* **35**, 4272–4280 (2019).
71. Wu, C., Gao, R., Zhang, Y. & De Marinis, Y. PTPD: predicting therapeutic peptides by deep learning and word2vec. *BMC Bioinformatics* **20**, 456 (2019).
72. Schaduangrat, N., Nantasenamat, C., Prachayasittikul, V. & Shoombuatong, W. ACPred: A computational tool for the prediction and analysis of anticancer peptides. *Molecules* **24**, 1973 (2019).
73. Rao, B., Zhang, L. & Zhang, G. ACP-GCN: The identification of anticancer peptides based on graph convolution networks. *IEEE Access* **8**, 176005–176011 (2020).

74. Yu, L., Jing, R., Liu, F., Luo, J. & Li, Y. DeepACP: A novel computational approach for accurate identification of anticancer peptides by deep learning algorithm. *Mol. Ther. Nucleic Acids* **22**, 862–870 (2020).

75. Burdukiewicz, M. *et al.* CancerGram: An effective classifier for differentiating anticancer from antimicrobial peptides. *Pharmaceutics* **12**, 1045 (2020).

76. Rao, B., Zhou, C., Zhang, G., Su, R. & Wei, L. ACPred-Fuse: fusing multi-view information improves the prediction of anticancer peptides. *Brief. Bioinform.* **21**, 1846–1855 (2020).

77. Li, Q., Zhou, W., Wang, D., Wang, S. & Li, Q. Prediction of anticancer peptides using a low-dimensional feature model. *Front. Bioeng. Biotechnol.* **8**, 892 (2020).

78. Vijayakumar, S. & Ptv, L. ACPP: A web server for prediction and design of anti-cancer peptides. *Int. J. Pept. Res. Ther.* **21**, 99–106 (2015).

79. Lv, Z., Cui, F., Zou, Q., Zhang, L. & Xu, L. Anticancer peptides prediction with deep representation learning features. *Brief. Bioinform.* **22**, (2021).

80. Chen, X.-G., Zhang, W., Yang, X., Li, C. & Chen, H. ACP-DA: Improving the prediction of anticancer peptides using Data Augmentation. *Front. Genet.* **12**, 698477 (2021).

81. Timmons, P. B. & Hewage, C. M. ENNAACT is a novel tool which employs neural networks for anticancer activity classification for therapeutic peptides. *Biomed. Pharmacother.* **133**, 111051 (2021).

82. Zhao, Y. *et al.* Prediction of anticancer peptides with high efficacy and low toxicity by hybrid model based on 3D structure of peptides. *Int. J. Mol. Sci.* **22**, 5630 (2021).

83. Arif, M. *et al.* StackACPred: Prediction of anticancer peptides by integrating optimized multiple feature descriptors with stacked ensemble approach. *Chemometr. Intell. Lab. Syst.* **220**, 104458 (2022).

84. Feng, G. *et al.* ME-ACP: Multi-view neural networks with ensemble model for identification of anticancer peptides. *Comput. Biol. Med.* **145**, 105459 (2022).

85. Zhu, L., Ye, C., Hu, X., Yang, S. & Zhu, C. ACP-check: An anticancer peptide prediction model based on bidirectional long short-term memory and multi-features fusion strategy. *Comput. Biol. Med.* **148**, 105868 (2022).

86. Liu, J., Li, M. & Chen, X. AntiMF: A deep learning framework for predicting anticancer peptides based on multi-view feature extraction. *Methods* **207**, 38–43 (2022).

87. Azim, S. M. *et al.* Accurately predicting anticancer peptide using an ensemble of heterogeneously trained classifiers. *Inform. Med. Unlocked* **42**, 101348 (2023).

88. Yao, L. *et al.* Accelerating the discovery of anticancer peptides through deep forest architecture with deep graphical representation. *Int. J. Mol. Sci.* **24**, 4328 (2023).

89. Zhou, W. *et al.* TriNet: A tri-fusion neural network for the prediction of anticancer and antimicrobial peptides. *Patterns (N. Y.)* **4**, 100702 (2023).

90. Tao, H., Shan, S., Fu, H., Zhu, C. & Liu, B. An augmented sample selection framework for prediction of anticancer peptides. *Molecules* **28**, 6680 (2023).

91. Sun, M., Hu, H., Pang, W. & Zhou, Y. ACP-BC: A model for accurate identification of anticancer peptides based on fusion features of bidirectional long short-term memory and chemically derived information. *Int. J. Mol. Sci.* **24**, 15447 (2023).

92. Deng, H. *et al.* ACP-MLC: A two-level prediction engine for identification of anticancer peptides and multi-label classification of their functional types. *Comput. Biol. Med.* **158**, 106844 (2023).

93. Li, Y., Ma, D., Chen, D. & Chen, Y. ACP-GBDT: An improved anticancer peptide identification method with gradient boosting decision tree. *Front. Genet.* **14**, 1165765 (2023).

94. Garai, S., Thomas, J., Dey, P. & Das, D. LGBM-ACp: an ensemble model for anticancer peptide prediction and in silico screening with potential drug targets. *Mol. Divers.* **28**, 1965–1981 (2024).

95. Karim, T., Shaon, M. S. H., Sultan, M. F., Hasan, M. Z. & Kafy, A.-A. ANNprob-ACPs: A novel anticancer peptide identifier based on probabilistic feature fusion approach. *Comput. Biol. Med.* **169**, 107915 (2024).

96. Ghafoor, H., Asim, M. N., Ibrahim, M. A., Ahmed, S. & Dengel, A. CAPTURE: Comprehensive anti-cancer peptide predictor with a unique amino acid sequence encoder. *Comput. Biol. Med.* **176**, 108538 (2024).

97. Wang, X. & Wang, S. ACP-PDAFF: Pretrained model and dual-channel attentional feature fusion for anticancer peptides prediction. *Comput. Biol. Chem.* **112**, 108141 (2024).

98. Yao, L. *et al.* ACP-CapsPred: an explainable computational framework for identification and functional prediction of anticancer peptides based on capsule network. *Brief. Bioinform.* **25**, (2024).

99. Liu, M. *et al.* ACPPfel: Explainable deep ensemble learning for anticancer peptides prediction based on feature optimization. *Front. Genet.* **15**, 1352504 (2024).

100. Liang, X., Zhao, H. & Wang, J. MA-PEP: A novel anticancer peptide prediction framework with multimodal feature fusion based on attention mechanism. *Protein Sci.* **33**, e4966 (2024).

101. Khan, S. Deep-representation-learning-based classification strategy for anticancer peptides. *Mathematics* **12**, 1330 (2024).

102. Zhong, G. & Deng, L. ACPScanner: Prediction of anticancer peptides by integrated machine learning methodologies. *J. Chem. Inf. Model.* **64**, 1092–1104 (2024).

103. Zhang, S., Zhao, Y. & Liang, Y. AACFlow: an end-to-end model based on attention

augmented convolutional neural network and flow-attention mechanism for identification of anticancer peptides. *Bioinformatics* **40**, (2024).

104. Sangaraju, V. K., Pham, N. T., Wei, L., Yu, X. & Manavalan, B. MACPpred 2.0: Stacked deep learning for anticancer peptide prediction with integrated spatial and probabilistic feature representations. *J. Mol. Biol.* **436**, 168687 (2024).
105. Arif, M., Musleh, S., Fida, H. & Alam, T. PLMACPred prediction of anticancer peptides based on protein language model and wavelet denoising transformation. *Sci. Rep.* **14**, 16992 (2024).
106. Bian, J. *et al.* ACP-ML: A sequence-based method for anticancer peptide prediction. *Comput. Biol. Med.* **170**, 108063 (2024).
107. Xu, X. *et al.* ACP-DRL: an anticancer peptides recognition method based on deep representation learning. *Front. Genet.* **15**, 1376486 (2024).
108. Kwon, M. *et al.* EnsemPred-ACP: Combining machine and deep learning to improve anticancer peptide prediction. *Comput. Biol. Med.* **196**, 110668 (2025).
109. Geng, A. *et al.* ACP-CLB: An anticancer peptide prediction model based on multichannel discriminative processing and integration of Large Pretrained Protein Language Models. *J. Chem. Inf. Model.* **65**, 2336–2349 (2025).
110. Gao, S. *et al.* ACP-ESM2: Enhancing anticancer peptide prediction with pre-trained protein language models. *IEEE Trans Comput Biol Bioinform* **22**, 1041–1051 (2025).
111. Huang, G., Cao, Y., Dai, Q. & Chen, W. ACP-DPE: A dual-channel deep learning model for anticancer peptide prediction. *IET Syst. Biol.* **19**, e70010 (2025).
112. Zhang, Z., Wang, X. & Shang, W. iACP-DPNet: a dual-pooling causal dilated convolutional network for interpretable anticancer peptide identification. *Funct. Integr. Genomics* **25**, 147 (2025).
113. Shahid *et al.* pACP-HybDeep: predicting anticancer peptides using binary tree growth

based transformer and structural feature encoding with deep-hybrid learning. *Sci. Rep.* **15**, 565 (2025).

114. Shahid *et al.* pACPs-DNN: Predicting anticancer peptides using novel peptide transformation into evolutionary and structure matrix-based images with self-attention deep learning model. *Comput. Biol. Chem.* **117**, 108441 (2025).
115. Tyagi, A. *et al.* In silico models for designing and discovering novel anticancer peptides. *Sci. Rep.* **3**, 2984 (2013).
116. Manavalan, B. *et al.* MLACP: machine-learning-based prediction of anticancer peptides. *Oncotarget* **8**, 77121–77136 (2017).
117. Chen, W., Ding, H., Feng, P., Lin, H. & Chou, K.-C. iACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget* **7**, 16895–16909 (2016).
118. Wei, L., Zhou, C., Chen, H., Song, J. & Su, R. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* **34**, 4007–4016 (2018).
119. Ahmed, S. *et al.* ACP-MHCNN: an accurate multi-headed deep-convolutional neural network to predict anticancer peptides. *Sci. Rep.* **11**, 23676 (2021).
120. Chen, J., Cheong, H. H. & Siu, S. W. I. XDeep-AcPEP: Deep learning method for anticancer peptide activity prediction based on convolutional neural network and multitask learning. *J. Chem. Inf. Model.* **61**, 3789–3803 (2021).
121. Thi Phan, L. *et al.* MLACP 2.0: An updated machine learning tool for anticancer peptide prediction. *Comput. Struct. Biotechnol. J.* **20**, 4473–4480 (2022).
122. Agrawal, P., Bhagat, D., Mahalwal, M., Sharma, N. & Raghava, G. P. S. AntiCP 2.0: an updated model for predicting anticancer peptides. *Brief. Bioinform.* **22**, (2021).
123. Lv, J. *et al.* ACP-EPC: an interpretable deep learning framework for anticancer peptide prediction utilizing pre-trained protein language model and multi-view feature

extracting strategy. *Mol. Divers.* (2025) doi:10.1007/s11030-025-11352-x.

124. Achilleos, K., Petrou, C., Nicolaidou, V. & Sarigiannis, Y. Beyond efficacy: Ensuring safety in peptide therapeutics through immunogenicity assessment. *J. Pept. Sci.* **31**, e70016 (2025).
125. Mahadik, R., Kiptoo, P., Tolbert, T. & Siahaan, T. J. Immune modulation by antigenic peptides and antigenic peptide conjugates for treatment of multiple sclerosis. *Med. Res. Arch* **10**, (2022).
126. Fehringer, M. & Vogl, T. Molecular mimicry in the pathogenesis of autoimmune rheumatic diseases. *J. Transl. Autoimmun.* **10**, 100269 (2025).
127. Cusick, M. F., Libbey, J. E. & Fujinami, R. S. Molecular mimicry as a mechanism of autoimmune disease. *Clin. Rev. Allergy Immunol.* **42**, 102–111 (2012).
128. Bethune, M. T. & Khosla, C. Parallels between pathogens and gluten peptides in celiac sprue. *PLoS Pathog.* **4**, e34 (2008).
129. Tan, L. C. *et al.* Specificity of T cells in synovial fluid: high frequencies of CD8(+) T cells that are specific for certain viral epitopes. *Arthritis Res.* **2**, 154–164 (2000).
130. Fazou, C., Yang, H., McMichael, A. J. & Callan, M. F. Epitope specificity of clonally expanded populations of CD8+ T cells found within the joints of patients with inflammatory arthritis. *Arthritis Rheum.* **44**, 2038–2045 (2001).
131. Tishler, M. & Shoenfeld, Y. Anti-heat-shock protein antibodies in rheumatic and autoimmune diseases. *Semin. Arthritis Rheum.* **26**, 558–563 (1996).
132. Mahmud, S. A. & Binstadt, B. A. Autoantibodies in the pathogenesis, diagnosis, and prognosis of juvenile idiopathic arthritis. *Front. Immunol.* **9**, 3168 (2018).
133. Gao, Y., Zhang, Y. & Liu, X. Rheumatoid arthritis: pathogenesis and therapeutic advances. *MedComm* **5**, e509 (2024).
134. Puig, M. & Shubow, S. Immunogenicity of therapeutic peptide products: bridging the

gaps regarding the role of product-related risk factors. *Front. Immunol.* **16**, 1608401 (2025).

135. García-Carnero, L. C. & Mora-Montes, H. M. Mucormycosis and COVID-19-associated mucormycosis: Insights of a deadly but neglected mycosis. *J. Fungi (Basel)* **8**, 445 (2022).

136. Mahalaxmi, I. *et al.* Mucormycosis: An opportunistic pathogen during COVID-19. *Environ. Res.* **201**, 111643 (2021).

137. Suvvari, T. K., Arigapudi, N., Kandi, V. R. & Kutikuppala, L. S. Mucormycosis: A killer in the shadow of COVID-19. *J. Mycol. Med.* **31**, 101161 (2021).

138. Aranjani, J. M., Manuel, A., Abdul Razack, H. I. & Mathew, S. T. COVID-19-associated mucormycosis: Evidence-based critical review of an emerging infection burden during the pandemic's second wave in India. *PLoS Negl. Trop. Dis.* **15**, e0009921 (2021).

139. Rammaert, B. *et al.* Healthcare-associated mucormycosis. *Clin. Infect. Dis.* **54 Suppl 1**, S44-54 (2012).

140. Gomes, M. Z. R., Lewis, R. E. & Kontoyiannis, D. P. Mucormycosis caused by unusual mucormycetes, non-Rhizopus, -Mucor, and -Lichtheimia species. *Clin. Microbiol. Rev.* **24**, 411–445 (2011).

141. Roden, M. M. *et al.* Epidemiology and outcome of zygomycosis: a review of 929 reported cases. *Clin. Infect. Dis.* **41**, 634–653 (2005).

142. Bhansali, A. *et al.* Presentation and outcome of rhino-orbital-cerebral mucormycosis in patients with diabetes. *Postgrad. Med. J.* **80**, 670–674 (2004).

143. Luo, Z. & Zhang, L. Diagnosis and treatment of pulmonary mucormycosis: A case report. *Exp. Ther. Med.* **14**, 3788–3791 (2017).

144. Castrejón-Pérez, A. D., Welsh, E. C., Miranda, I., Ocampo-Candiani, J. & Welsh, O.

Cutaneous mucormycosis. *An. Bras. Dermatol.* **92**, 304–311 (2017).

145. Spellberg, B. Gastrointestinal mucormycosis: an evolving disease. *Gastroenterol. Hepatol. (N. Y.)* **8**, 140–142 (2012).

146. Elitzur, S. *et al.* Disseminated mucormycosis in immunocompromised children: Are new antifungal agents making a difference? A multicenter retrospective study. *J. Fungi (Basel)* **7**, 165 (2021).

147. Brunke, S., Mogavero, S., Kasper, L. & Hube, B. Virulence factors in fungal pathogens of man. *Curr. Opin. Microbiol.* **32**, 89–95 (2016).

148. Binder, U., Maurer, E. & Lass-Flörl, C. Mucormycosis--from the pathogens to the disease. *Clin. Microbiol. Infect.* **20 Suppl 6**, 60–66 (2014).

149. Challa, S. Mucormycosis: Pathogenesis and pathology. *Curr. Fungal Infect. Rep.* **13**, 11–20 (2019).

150. Chibucos, M. C. *et al.* An integrated genomic and transcriptomic survey of mucormycosis-causing fungi. *Nat. Commun.* **7**, 12218 (2016).

151. Alqarihi, A., Kontoyiannis, D. P. & Ibrahim, A. S. Mucormycosis in 2023: an update on pathogenesis and management. *Front. Cell. Infect. Microbiol.* **13**, 1254919 (2023).

152. Moheb-Alian, A. *et al.* Mucormycosis and COVID-19: Unraveling the interplay of fungal infection in a global health crisis: An overview. *Infect. Disord. Drug Targets* **25**, e18715265310191 (2025).

153. Pappas, P. G. *et al.* Invasive fungal infections among organ transplant recipients: results of the Transplant-Associated Infection Surveillance Network (TRANSNET). *Clin. Infect. Dis.* **50**, 1101–1111 (2010).

154. Pemán, J., Cantón, E. & Espinel-Ingroff, A. Antifungal drug resistance mechanisms. *Expert Rev. Anti. Infect. Ther.* **7**, 453–460 (2009).

155. Spellberg, B., Edwards, J., Jr & Ibrahim, A. Novel perspectives on mucormycosis:

pathophysiology, presentation, and management. *Clin. Microbiol. Rev.* **18**, 556–569 (2005).

156. Ribes, J. A., Vanover-Sams, C. L. & Baker, D. J. Zygomycetes in human disease. *Clin. Microbiol. Rev.* **13**, 236–301 (2000).

157. Morales-Franco, B. *et al.* Host-pathogen molecular factors contribute to the pathogenesis of Rhizopus spp. In diabetes mellitus. *Curr. Trop. Med. Rep.* **8**, 6–17 (2021).

158. Tahiri, G. *et al.* Mucorales and mucormycosis: Recent insights and future prospects. *J. Fungi (Basel)* **9**, 335 (2023).

159. Singh, H., Ansari, H. R. & Raghava, G. P. S. Improved method for linear B-cell epitope prediction using antigen's primary sequence. *PLoS One* **8**, e62216 (2013).

160. Bhasin, M. & Raghava, G. P. S. Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine* **22**, 3195–3204 (2004).

161. Singh, H. & Raghava, G. P. S. ProPred1: prediction of promiscuous MHC Class-I binding sites. *Bioinformatics* **19**, 1009–1014 (2003).

162. Singh, H. & Raghava, G. P. ProPred: prediction of HLA-DR binding sites. *Bioinformatics* **17**, 1236–1237 (2001).

163. Singh, O., Hsu, W.-L. & Su, E. C.-Y. ILeukin10Pred: A computational approach for predicting IL-10-inducing immunosuppressive peptides using combinations of amino acid global features. *Biology (Basel)* **11**, 5 (2021).

164. Dhanda, S. K., Vir, P. & Raghava, G. P. S. Designing of interferon-gamma inducing MHC class-II binders. *Biol. Direct* **8**, 30 (2013).

165. Nagpal, G., Chaudhary, K., Agrawal, P. & Raghava, G. P. S. Computer-aided prediction of antigen presenting cell modulators for designing peptide-based vaccine adjuvants. *J. Transl. Med.* **16**, 181 (2018).

166. Nagpal, G. *et al.* VaccineDA: Prediction, design and genome-wide screening of oligodeoxynucleotide-based vaccine adjuvants. *Sci. Rep.* **5**, 12478 (2015).

167. McIntyre, G. J. *et al.* 96 shRNAs designed for maximal coverage of HIV-1 variants. *Retrovirology* **6**, 55 (2009).

168. Ahmed, F. & Raghava, G. P. S. Designing of highly effective complementary and mismatch siRNAs for silencing a gene. *PLoS One* **6**, e23443 (2011).

169. Hussain, M. K. *et al.* Mucormycosis: A hidden mystery of fungal infection, possible diagnosis, treatment and development of new therapeutic agents. *Eur. J. Med. Chem.* **246**, 115010 (2023).

170. Hernández-Chávez, M. J., Pérez-García, L. A., Niño-Vega, G. A. & Mora-Montes, H. M. Fungal strategies to evade the host immune recognition. *J. Fungi (Basel)* **3**, (2017).

171. Lax, C. *et al.* Genes, pathways, and mechanisms involved in the virulence of Mucorales. *Genes (Basel)* **11**, 317 (2020).

172. Baldin, C. & Ibrahim, A. S. Molecular mechanisms of mucormycosis—The bitter and the sweet. *PLoS Pathog.* **13**, e1006408 (2017).

173. Jansen, J. H., Fibbe, W. E., Willemze, R. & Kluin-Nelemans, J. C. Interleukin-4. A regulatory protein. *Blut* **60**, 269–274 (1990).

174. Chomarat, P. & Banchereau, J. An update on interleukin-4 and its receptor. *Eur. Cytokine Netw.* **8**, 333–344 (1997).

175. Yoshimoto, T., Bendelac, A., Watson, C., Hu-Li, J. & Paul, W. E. Role of NK1.1+ T cells in a TH2 response and in immunoglobulin E production. *Science* **270**, 1845–1847 (1995).

176. Seder, R. A. *et al.* Mouse splenic and bone marrow cell populations that express high-affinity Fc epsilon receptors and produce interleukin 4 are highly enriched in basophils. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 2835–2839 (1991).

177. Brown, M. A. *et al.* B cell stimulatory factor-1/interleukin-4 mRNA is expressed by normal and transformed mast cells. *Cell* **50**, 809–818 (1987).

178. Luzina, I. G. *et al.* Regulation of inflammation by interleukin-4: a review of "alternatives." *J. Leukoc. Biol.* **92**, 753–764 (2012).

179. Paul, W. E. History of interleukin-4. *Cytokine* **75**, 3–7 (2015).

180. Howard, M. *et al.* Identification of a T cell-derived b cell growth factor distinct from interleukin 2. *J. Exp. Med.* **155**, 914–923 (1982).

181. Röcken, M., Racke, M. & Shevach, E. M. IL-4-induced immune deviation as antigen-specific therapy for inflammatory autoimmune disease. *Immunol. Today* **17**, 225–231 (1996).

182. Ghoreschi, K. *et al.* Interleukin-4 therapy of psoriasis induces Th2 responses and improves human autoimmune disease. *Nat. Med.* **9**, 40–46 (2003).

183. Lubberts, E. *et al.* IL-4 gene therapy for collagen arthritis suppresses synovial IL-17 and osteoprotegerin ligand and prevents bone erosion. *J. Clin. Invest.* **105**, 1697–1710 (2000).

184. Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2019).

185. Harrison, L. C. *et al.* A peptide-binding motif for I-A(g7), the class II major histocompatibility complex (MHC) molecule of NOD and Biozzi AB/H mice. *J. Exp. Med.* **185**, 1013–1021 (1997).

186. Rajapakse, M., Schmidt, B., Feng, L. & Brusic, V. Predicting peptides binding to MHC class II molecules using multi-objective evolutionary algorithms. *BMC Bioinformatics* **8**, 459 (2007).

187. Pande, A. *et al.* Pfeature: A tool for computing wide range of protein features and building prediction models. *J. Comput. Biol.* **30**, 204–222 (2023).

188. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1226–1238 (2005).

189. Whitney, A. W. A direct method of nonparametric measurement selection. *IEEE Trans. Comput.* **C–20**, 1100–1103 (1971).

190. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

191. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202-8 (2009).

192. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME suite. *Nucleic Acids Res.* **43**, W39-49 (2015).

193. Vens, C., Rosso, M.-N. & Danchin, E. G. J. Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics* **27**, 1231–1238 (2011).

194. Li, Z., Liu, F., Yang, W., Peng, S. & Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **33**, 6999–7019 (2022).

195. Elnaggar, A. *et al.* ProtTrans: Towards cracking the language of life's code through self-supervised learning. *bioRxiv* (2020) doi:10.1101/2020.07.12.199554.

196. Dhall, A., Patiyal, S. & Raghava, G. P. S. A hybrid method for discovering interferon-gamma inducing peptides in human and mouse. *Sci. Rep.* **14**, 26859 (2024).

197. Xie, M., Liu, D. & Yang, Y. Anti-cancer peptides: classification, mechanism of action, reconstruction and modification. *Open Biol.* **10**, 200004 (2020).

198. Sood, A., Jothiswaran, V. V., Singh, A. & Sharma, A. Anticancer peptides as novel immunomodulatory therapeutic candidates for cancer treatment. *Explor. Target. Antitumor Ther.* **5**, 1074–1099 (2024).

199. Chauhan, M., Gupta, A., Tomer, R. & Raghava, G. P. S. CancerPPD2: an updated repository of anticancer peptides and proteins. *Database (Oxford)* **2025**, (2025).

200. Ma, T. *et al.* DRAMP 4.0: an open-access data repository dedicated to the clinical translation of antimicrobial peptides. *Nucleic Acids Res.* **53**, D403–D410 (2025).

201. Huerta-Reyes, M. *et al.* Treatment of breast cancer with gonadotropin-releasing hormone analogs. *Front. Oncol.* **9**, 943 (2019).

202. Salgia, R. *et al.* A randomized phase II study of LY2510924 and carboplatin/etoposide versus carboplatin/etoposide in extensive-disease small cell lung cancer. *Lung Cancer* **105**, 7–13 (2017).

203. Ishikawa, R., Saito, K., Misawa, T., Demizu, Y. & Saito, Y. Identification of the stapled α-helical peptide ATSP-7041 as a substrate and strong inhibitor of OATP1B1 in vitro. *Biomolecules* **13**, 1002 (2023).

204. Zhao, J., Guo, S., Schrodi, S. J. & He, D. Molecular and cellular heterogeneity in rheumatoid arthritis: Mechanisms and clinical implications. *Front. Immunol.* **12**, 790122 (2021).

205. Huang, J. *et al.* Promising therapeutic targets for treatment of rheumatoid arthritis. *Front. Immunol.* **12**, 686155 (2021).

206. Ngian, G.-S. Rheumatoid arthritis. *Aust. Fam. Physician* **39**, 626–628 (2010).

207. Smolen, J. S., Aletaha, D. & McInnes, I. B. Rheumatoid arthritis. *Lancet* **388**, 2023–2038 (2016).

208. Gibofsky, A. Epidemiology, pathophysiology, and diagnosis of rheumatoid arthritis: A Synopsis. *Am. J. Manag. Care* **20**, S128-35 (2014).

209. Wu, D. *et al.* Systemic complications of rheumatoid arthritis: Focus on pathogenesis and treatment. *Front. Immunol.* **13**, 1051082 (2022).

210. Zhu, J. *et al.* Age at menarche, age at natural menopause, and risk of rheumatoid

arthritis - a Mendelian randomization study. *Arthritis Res. Ther.* **23**, 108 (2021).

211. Alivernini, S., Firestein, G. S. & McInnes, I. B. The pathogenesis of rheumatoid arthritis. *Immunity* **55**, 2255–2270 (2022).
212. Scholz, E. *et al.* A comparative analysis of the peptide repertoires of HLA-DR molecules differentially associated with rheumatoid arthritis. *Arthritis Rheumatol.* **68**, 2412–2421 (2016).
213. Wu, C.-Y., Yang, H.-Y., Luo, S.-F. & Lai, J.-H. From rheumatoid factor to anti-citrullinated protein antibodies and anti-carbamylated protein antibodies for diagnosis and prognosis prediction in patients with rheumatoid arthritis. *Int. J. Mol. Sci.* **22**, 686 (2021).
214. McInnes, I. B. & Schett, G. Pathogenetic insights from the treatment of rheumatoid arthritis. *Lancet* **389**, 2328–2337 (2017).
215. Radu, A.-F. & Bungau, S. G. Management of rheumatoid arthritis: An overview. *Cells* **10**, 2857 (2021).
216. Simon, L. S. *et al.* The Jak/STAT pathway: A focus on pain in rheumatoid arthritis. *Semin. Arthritis Rheum.* **51**, 278–284 (2021).
217. Ding, Q. *et al.* Signaling pathways in rheumatoid arthritis: implications for targeted therapy. *Signal Transduct. Target. Ther.* **8**, 68 (2023).
218. Cronstein, B. N. The mechanism of action of methotrexate. *Rheum. Dis. Clin. North Am.* **23**, 739–755 (1997).
219. Bedoui, Y. *et al.* Methotrexate an old drug with new tricks. *Int. J. Mol. Sci.* **20**, 5023 (2019).
220. Zhao, Z. *et al.* Application and pharmacological mechanism of methotrexate in rheumatoid arthritis. *Biomed. Pharmacother.* **150**, 113074 (2022).
221. Conley, B. *et al.* What are the core recommendations for rheumatoid arthritis care?

Systematic review of clinical practice guidelines. *Clin. Rheumatol.* **42**, 2267–2278 (2023).

222. Moreira, P. M., Correia, A. M., Cerqueira, M. & Gil, M. F. Perioperative management of disease-modifying antirheumatic drugs and other immunomodulators. *ARP Rheumatol.* **1**, 218–224 (2022).
223. Angelini, J. *et al.* JAK-inhibitors for the treatment of rheumatoid arthritis: A focus on the present and an outlook on the future. *Biomolecules* **10**, 1002 (2020).
224. Song, Y., Li, J. & Wu, Y. Evolving understanding of autoimmune mechanisms and new therapeutic strategies of autoimmune disorders. *Signal Transduct. Target. Ther.* **9**, 263 (2024).
225. Sun, R., Qian, M. G. & Zhang, X. T and B cell epitope analysis for the immunogenicity evaluation and mitigation of antibody-based therapeutics. *MAbs* **16**, 2324836 (2024).
226. Sokolova, M. V., Schett, G. & Steffen, U. Autoantibodies in rheumatoid arthritis: Historical background and novel findings. *Clin. Rev. Allergy Immunol.* **63**, 138–151 (2022).
227. Prawiningrum, A. F., Paramita, R. I. & Panigoro, S. S. Immunoinformatics approach for Epitope-based vaccine design: Key steps for breast cancer vaccine. *Diagnostics (Basel)* **12**, 2981 (2022).
228. Losowsky, M. S. A history of coeliac disease. *Dig. Dis.* **26**, 112–120 (2008).
229. Dicke, W. K., Weijers, H. A. & Van de Kamer, J. H. Coeliac disease. II. The presence in wheat of a factor having a deleterious effect in cases of coeliac disease. *Acta Paediatr.* **42**, 34–42 (1953).
230. Singh, P. *et al.* Global prevalence of celiac disease: Systematic review and meta-analysis. *Clin. Gastroenterol. Hepatol.* **16**, 823-836.e2 (2018).
231. Parzanese, I. *et al.* Celiac disease: From pathophysiology to treatment. *World J.*

Gastrointest. Pathophysiol. **8**, 27–38 (2017).

232. Stanković, B. *et al.* HLA genotyping in pediatric celiac disease patients. *Bosn. J. Basic Med. Sci.* **14**, 171–176 (2014).
233. van Lummel, M. *et al.* Type 1 diabetes-associated HLA-DQ8 transdimer accommodates a unique peptide repertoire. *J. Biol. Chem.* **287**, 9514–9524 (2012).
234. Holding, D. R. Recent advances in the study of prolamin storage protein organization and function. *Front. Plant Sci.* **5**, 276 (2014).
235. Shewry, P. R. & Halford, N. G. Cereal seed storage proteins: structures, properties and role in grain utilization. *J. Exp. Bot.* **53**, 947–958 (2002).
236. Aboulaghras, S. *et al.* Pathophysiology and immunogenetics of celiac disease. *Clin. Chim. Acta* **528**, 74–83 (2022).
237. Garrote, J. A., Gómez-González, E., Bernardo, D., Arranz, E. & Chirdo, F. Celiac disease pathogenesis: the proinflammatory cytokine network. *J. Pediatr. Gastroenterol. Nutr.* **47 Suppl 1**, S27-32 (2008).
238. Zhou, L. *et al.* Abrogation of immunogenic properties of gliadin peptides through transamidation by microbial transglutaminase is acyl-acceptor dependent. *J. Agric. Food Chem.* **65**, 7542–7552 (2017).
239. Gujral, N., Freeman, H. J. & Thomson, A. B. R. Celiac disease: prevalence, diagnosis, pathogenesis and treatment. *World J. Gastroenterol.* **18**, 6036–6059 (2012).
240. Lindfors, K. *et al.* Coeliac disease. *Nat. Rev. Dis. Primers* **5**, 3 (2019).
241. Lebwohl, B., Sanders, D. S. & Green, P. H. R. Coeliac disease. *Lancet* **391**, 70–81 (2018).
242. Paolella, G., Sposito, S., Romanelli, A. M. & Caputo, I. Type 2 transglutaminase in coeliac disease: A key player in pathogenesis, diagnosis and therapy. *Int. J. Mol. Sci.* **23**, 7513 (2022).

243. Kumar, J., Kumar, M., Pandey, R. & Chauhan, N. S. Physiopathology and management of gluten-induced celiac disease. *J. Food Sci.* **82**, 270–277 (2017).