



Development of AI-based Tools For Designing Subunit Vaccines

BY

NISHANT KUMAR

PhD21206

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

NEW DELHI-110020

March, 2026



Development of AI-based Tools For Designing Subunit Vaccines

A THESIS SUBMITTED

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

NISHANT KUMAR

(PhD21206)

UNDER THE GUIDANCE OF

PROF. GAJENDRA PAL SINGH RAGHAVA

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY

NEW DELHI-110020

March, 2026



Prof. Gajendra P.S. Raghava
Professor, Department of Computational Biology

Certificate

This is to certify that the thesis titled “**Development of AI-based tools for designing subunit vaccines**” being submitted by **Mr. Nishant Kumar** 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 regulation relating to the degree.

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

Place: New Delhi

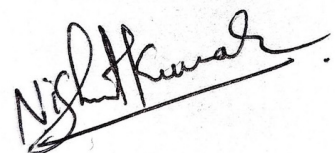
Date: July, 2026.

Prof. Gajendra P.S. Raghava
(Supervisor)

Declaration

I hereby declare that the thesis entitled “**Development of AI-based Tools for Designing Subunit Vaccines**” submitted to the Indraprastha Institute of Information Technology Delhi (IIIT-Delhi) for the award of the degree of Doctor of Philosophy (Ph.D.) is a record of original research work carried out by me under the supervision of **Prof. Gajendra P.S. Raghava**, Department of Computational Biology, IIIT-Delhi, New Delhi.

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



Place: New Delhi

Date: July, 2026

Nishant Kumar

Roll no. PhD21206

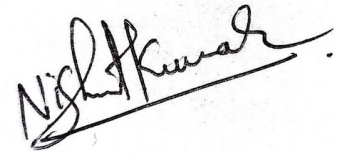
Department of Computational Biology

Indraprastha Institute of Information Technology Delhi

New Delhi – 110020

Declaration Regarding Use of AI

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.

A handwritten signature in black ink, reading "Nishant Kumar", with a horizontal line underneath the name.

Nishant Kumar
Roll No. PhD21206

Acknowledgements

I express my heartfelt gratitude to the Almighty for granting me strength, patience, and perseverance throughout this demanding journey. Completing this doctoral work has been a transformative experience, and I feel fortunate to have received support from many individuals and institutions along the way.

I would like to convey my deepest gratitude to my Ph.D. supervisor, **Prof. Gajendra Pal Singh Raghava**, for his invaluable guidance, constant encouragement, and unwavering support throughout my doctoral research. His scientific insight and enthusiasm for research greatly shaped my understanding of computational biology. I am especially thankful for the academic freedom he provided to explore new ideas, along with his timely advice and constructive feedback. His mentorship helped me grow both as a researcher and as an independent thinker, and I feel privileged to have carried out my doctoral work under his supervision.

I sincerely thank the institute authorities and faculty members of Indraprastha Institute of Information Technology, Delhi (IIIT-Delhi) for providing a stimulating research environment and excellent infrastructure. I am grateful to **Prof. Ranjan Bose**, the former Director of IIIT-Delhi, for granting me the opportunity to pursue my doctoral studies at the institute, and to **Prof. Gajendra P.S. Raghava**, the current Director of IIIT-Delhi, for his continued support and encouragement during the course of my Ph.D. I would also like to thank my doctoral committee members, **Prof. Ganesh Bagler and Dr. Gaurav Ahuja**, for their valuable suggestions, critical comments, and periodic evaluations, which helped improve the quality of my research work.

I acknowledge with appreciation all the faculty members who taught me during my coursework and helped build a strong academic foundation. I am also thankful to the administrative and academic staff of IIIT-Delhi, especially Ms. Priti Patel, Dr. Ravi Bhasin, Mr. Ajay, Ms. Shipra Jain, Ms. Sheetu Ahuja, Ms. Anshu Dureja, Mr. Imran Khan, Mr. Kapil Dev Garg, Mr. Mohit Kumar, Mr. Raju Biswas, Ms. Sonal, Mr. Tinku, Mr. Binod, Mr. Shyam, Mr. Adarsh, for their continuous assistance in academic and administrative matters. Their timely help made many complicated procedures smooth and manageable.

I gratefully acknowledge the financial support provided by the **University Grants Commission (UGC), India**, through the CSIR-UGC Fellowship, which made it possible for me to focus on my research work. My special thanks go to all my lab members Shubham Choudhury, Pratik Shinde

Pankaj Kumar, Nisha Bajiya, Kunal Choudhary, Dr. Ritu Tomar, Anand Singh Rathore, Naman Mehta, Sahil Raina, Dr. Akansha Arora, Saloni, Pushpendra, Shivani, Piyush, Devraj, Manish, Sudhanshu, Sachin, Anshuma, seniors Dr. Sumeet Patiyal, Dr. Anjali Dhall, Dr. Leimarembi Devi Naorem, Dr. Neetesh Pandey Dr. Neelam, Dr. Indra Prakash Jha, colleagues, Saveena Solanki, Alok Anand, Dilip Yadav, Amit Samal, Anshul, Dr. Madhu, Dr. Vishaka, Sanjay, Sonam, Suvendhu, Gaurav, Shruti, Manas, Dr. Mansi Goel, Akansha Gupta, Pradeep, Akshya, Saurav Goswami, Swarnava, Abhisekh, Jasmine. The friendly environment, collaborative discussions, and shared learning experiences made the research journey enjoyable and memorable. I am particularly grateful to my colleagues of Premas Biotech pvt. Ptd., Ms. Swati, Mr. Parvesh, Dr. Ambica Baru, Dr. Manish Rana, and Prof. Dr. Pradeep Panwar, C.C.S. University, for guiding me during the initial stages of my Ph.D. and helping me understand both research methodology and lab culture. I also thank my colleagues and batchmates for their intellectual discussions, encouragement during difficult phases, and for making the lab a supportive and motivating workplace. I would also like to acknowledge the library staff, hostel staff, mess workers, canteen staff, BMS staff Naveen Ji and supporting personnel at IIIT-Delhi for taking care of daily needs and creating a comfortable living environment away from home. Additionally, I want to thank Mr. Rajesh, whose tea became a vital component of my everyday schedule. During long workdays, his constantly stimulating tea and the quick chats that accompanied it provided much-needed breaks.

I owe my deepest gratitude to my family. I dedicate this thesis to my parents, **Smt. Kavita and Mr. Vijay Kumar**, whose unconditional love, sacrifices, and belief in me have been the foundation of my education and life. Their encouragement and patience sustained me during challenging times. I am equally thankful to my brother, **Vibhanshu Kumar** and family members for their constant support, motivation, and understanding throughout this journey. Finally, I thank all my friends, well-wishers, and everyone who directly or indirectly contributed to my Ph.D. journey. This accomplishment belongs not only to me but also to all those who stood beside me and believed in me.

Abstract

Immunization continues to be a cornerstone of global public health, drastically reducing mortality rates worldwide. Over time, the field of vaccinology has progressively shifted from utilizing whole-pathogen preparations toward highly targeted epitope and subunit formulations. However, traditional laboratory-based methods for vaccine discovery remain heavily constrained by high financial costs and extended developmental timelines. To mitigate these bottlenecks, this thesis introduces a comprehensive suite of computational resources for designing epitope or subunit based vaccines. The work presented in this thesis is organized into three major components: identification of epitope-based vaccine candidates, safety evaluation of predicted candidates, and assessment of stability under physiological and environmental conditions. Under the first objective, a tool named CLBtope was developed to predict both linear and conformational B-cell epitopes from antigen sequences, facilitating the identification of potential vaccine candidates capable of eliciting protective B-cell and antibody responses. To ensure candidate safety, two predictive models were established: DMPPred, which identifies peptides associated with type 1 diabetes mellitus, and AlgPred3, which comprehensively evaluates the allergenic risks of proposed vaccine components. Finally, to address peptide and protein stability under diverse physiological and environmental conditions, three distinct platforms were developed: AFProPred for the characterization and design of antifreeze proteins, PPTStab for engineering thermostable proteins, and PlifePred2 for estimating peptide half-lives in body fluids. All tools developed in this thesis are freely available to the scientific community as web servers and standalone software through platforms such as GitHub and PyPI. State-of-the-art AI methodologies, including machine learning, deep learning, and large language models, were deployed to design effective subunit vaccines.

List of Publications, and Poster Presentations

Publications

Thesis related

Published

1. **Kumar N**, Patiyal S, Choudhury S, et al. DMPPred: a tool for identification of antigenic regions responsible for inducing type 1 diabetes mellitus. Brief. Bioinform. 2023; 24.
2. **Kumar N**, Bajiya N, Patiyal S, et al. Multi-perspectives and challenges in identifying B-cell epitopes. Protein Sci. 2023; 32:e4785.
3. **Kumar N**, Tripathi S, Sharma N, et al. A method for predicting linear and conformational B-cell epitopes in an antigen from its primary sequence. Comput. Biol. Med. 2024; 170:108083.
4. **Kumar N**, Choudhury S, Bajiya N, et al. Prediction of anti-freezing proteins from their evolutionary profile. Proteomics 2025; 25:e202400157.
5. Tijare P, **Kumar N**, Raghava GPS. Prediction and design of thermostable proteins with a desired melting temperature. Sci. Rep. 2025; 15:16683

Others

Published

6. Patiyal S, Dhall A, **Kumar N**, et al. HLA-DR4Pred2: An improved method for predicting HLA-DRB1*04:01 binders. Methods 2024; 232:18–28.
7. Bajiya N, **Kumar N**, Raghava GPS. Prediction of inhibitory peptides against E.coli with desired MIC value. Sci. Rep. 2025; 15:4672.
8. Gahlot PS, Choudhury S, Bajiya N, **Kumar N**, et al. Prediction of plant resistance proteins using alignment-based and alignment-free approaches. Proteomics 2025; 25:e202400261.
9. Rathore AS, **Kumar N**, Choudhury S, et al. Prediction of hemolytic peptides and their hemolytic concentration. Commun. Biol. 2025; 8:176.
10. Sahni R, **Kumar N**, Raghava GPS. HAIRpred: Prediction of human antibody interacting residues in an antigen from its primary structure. Protein Sci. 2025; 34:e70212.

Under Review

11. Shendre A, Mehta NK, Rathore AS, **Kumar N**, et al. MAP format for representing chemical modifications, annotations, and mutations in protein sequences: An extension of the FASTA format. arXiv [q-bio.BM] 2025.
12. Mathur K, Jain S, Bajiya N, **Kumar N**, et al. In silico tool for identification of colorectal cancer from cell-free DNA biomarkers. arXiv [q-bio.GN] 2025.
13. Alam U, Chaudhary K, **Kumar N**, et al. PEPlife2: A updated repository of the half-life of peptides. bioRxiv 2025.
14. Pandey A, Megha, **Kumar N**, et al. CBTOPE2: An improved method for predicting of conformational B-cell epitopes in an antigen from its primary sequence. arXiv [q-bio.BM] 2025.

Poster Presentations

1. **Kumar N**, Patiyal S, Choudhury S, et al. (2025) Identification of antigenic regions responsible for inducing type 1 diabetes mellitus. JBI2025 XV Bioinformatics Symposium (JBI2025) at Madrid, Spain, 22 – 24 October.
2. **Kumar N**, Bajiya N, Patiyal S, et al. (2024) Multi-perspectives and challenges in identifying B-cell epitopes at IIIT-Delhi (RIISE), 15-16 March.

Table of Contents

Certificate.....	iii
Declaration.....	iv
Acknowledgements.....	v
Abstract.....	viii
List of Publications, and Poster Presentations	ix
Table of Contents.....	xi
List of Abbreviations	xvii
List of Figures.....	xix
List of Tables	xxii
Chapter 1. Introduction	1
1.1 Motivation and Background	2
1.2 Types of Immunity.....	2
1.3 Vaccination	4
1.4 Subunit Vaccine Design: Challenges and Opportunities	6
1.5 Role of Artificial Intelligence in Vaccine and Therapeutic Design.....	8
1.6 A Unified Computational Framework for Subunit Vaccine Design.....	8
1.7 Thesis Objectives	9
1.8 Organization of the chapter.....	11
Chapter 2. Review of Literature.....	12
2.1 Vaccination and Immunotherapy.....	13
2.1.1 Different Immune System arms	14
2.2 B-cell Epitopes (Humoral Immunity)	15
2.2.1 Experimental Identification of BCE	16
2.2.2 B-cell epitopes Repositories	17
2.2.3 Linear B-Cell Epitope Prediction Methods.....	17
2.2.4 Prediction of conformational B-Cell Epitopes.....	21
2.3 Safety Evaluation.....	23
2.3.1 Autoimmune Responses and Vaccine Safety	23
2.3.2 Allergenicity and Vaccine Safety	25
2.4 Stability Evaluation.....	27

2.4.1	Temperature Stability.....	28
2.4.2	Cold Denaturation and Antifreeze Proteins	30
2.4.3	Half-Life in Body Fluids.....	32
2.5	Summary	34
Chapter 3. Datasets & Methodology.....		36
3.1	Overview.....	37
3.2	Dataset Compilation and Curation.....	37
3.3	Redundancy Removal and Dataset Balancing	38
3.4	Dataset Partitioning.....	41
3.5	Compositional Analysis	41
3.6	Positional Analysis	41
3.7	Feature Extraction.....	42
3.7.1	Sequence composition descriptors.....	42
3.7.2	Binary Profiles Features.....	43
3.7.3	Evolutionary Profile-based Features.....	44
3.7.4	Protein Language Model Embeddings.....	45
3.8	Feature Selection Methods.....	46
3.9	Model Development.....	49
3.9.1	Traditional Machine Learning Algorithms	49
3.9.2	Large Language models.....	50
3.10	Alignment-Based Approaches	51
3.10.1	Similarity search approach.....	51
3.10.2	Motif Identification.....	51
3.11	Model Training and Hyperparameter Optimization	53
3.12	Model Evaluation.....	54
3.13	Performance Evaluation Metrics.....	55
3.13.1	Classification Metrics	55
3.13.2	Regression Metrics.....	56
3.14	Feature Importance Analysis Using SHAP.....	58
3.15	Web server Implementation and software	58
Chapter 4. Prediction of Linear and Conformational Epitopes.....		60
4.1	Introduction.....	61
4.2	Overview of the study	62

4.3	Materials and Methods.....	63
4.3.1	Dataset Acquisition.....	63
4.3.2	Feature Extraction and Representation	63
4.3.3	Feature selection	64
4.3.4	Model Development and Evaluation	64
4.3.5	Sequence Similarity Search using BLAST	64
4.3.6	Motif Derived features.....	65
4.4	Results.....	65
4.4.1	Compositional analysis	65
4.4.2	Positional Specific analysis.....	66
4.4.3	Performance on Individual Descriptors	66
4.4.4	Performance on Combined Descriptors	67
4.4.5	Performance of selected features	68
4.4.6	Sequence Similarity Approach	70
4.4.7	Hybrid methods.....	70
4.4.8	Benchmarking with available methods.....	72
4.5	SHAP-based interpretation	74
4.6	Case study: analysis of BCEs in SARS-CoV2	74
4.7	Web server implementation	75
4.8	Discussion and Conclusion.....	75
Chapter 5. Identification of T1DM-Associated Peptides.....		77
5.1	Introduction.....	78
5.2	Overview of the study.....	79
5.3	Materials and Methods.....	80
5.3.1	Dataset collection.....	80
5.3.2	Feature generation.....	81
5.3.3	Feature selection	81
5.3.4	ML Model Development and Evaluation	81
5.3.5	Similarity Search.....	81
5.4	Results.....	82
5.4.1	Compositional analysis	82
5.4.2	Positional analysis.....	82
5.4.3	HLA Allele Distribution	83
5.4.4	Performance of compositional descriptors.....	84
5.4.5	Performance of Feature Selection Models.....	84
5.4.6	Similarity search approach.....	86
5.4.7	Hybrid model performance analysis	87

5.5	SHAP-based interpretation	88
5.6	Case Study: Coxsackievirus B4 Antigenic Regions	89
5.7	Web server	90
5.8	Discussion and Conclusion	91
Chapter 6. Allergenicity Prediction of Therapeutic Peptides		93
6.1	Introduction.....	94
6.2	Overview of the study	95
6.3	Materials and Methods.....	96
6.3.1	Dataset Compilation and Curation.....	96
6.3.2	Feature generation.....	98
6.3.3	Feature fusion and Feature selection.....	98
6.3.4	Development and Performance Evaluation of Models	99
6.3.5	Large Language Models as Classifiers	99
6.3.6	Protein Language Model Embeddings.....	99
6.3.7	Quantum SVM	100
6.4	Results.....	101
6.4.1	Amino Acid Composition Analysis	101
6.4.2	Positional Analysis	102
6.4.3	Preliminary Machine Learning Evaluation	103
6.4.4	Multi-feature Fusion Results.....	103
6.4.5	Performance evaluation of selected feature descriptors.....	104
6.4.6	Performance of binary profile-based models.....	105
6.4.7	Performance on Large Language Models (LLMs)	106
6.4.8	Quantum SVM	107
6.4.9	Motif Analysis	108
6.4.10	Final Model.....	109
6.5	Benchmarking of available methods.....	110
6.6	Limitations	112
6.7	Web Server Implementation	112
6.8	Discussion and conclusion.....	113
Chapter 7. Prediction and Design of Thermostable Proteins with desired T _m		114
7.1	Introduction.....	115
7.2	Overview of the study	115
7.3	Material and Methods	116

7.3.1	Data collection and Preparation	116
7.3.2	Features engineering	117
7.3.3	Normalization	118
7.3.4	Regression Model Development.....	118
7.3.5	Model evaluation	119
7.4	Results.....	120
7.4.1	Protein Composition Analysis	120
7.4.2	Correlation of Tm with Descriptors.....	120
7.4.3	Development of Prediction Models	122
7.4.4	Comparison with Existing Methods.....	126
7.4.5	Application: Large-scale Screening of Thermostable Proteins	127
7.5	SHAP-based interpretation	129
7.6	Web server and Standalone Implementation	130
7.7	Discussion and Conclusion	130
Chapter 8. Prediction of Anti-Freezing Proteins.....		132
8.1	Introduction.....	133
8.2	Overview of the study	134
8.3	Materials and Methods.....	135
8.3.1	Dataset Compilation and Pre-processing	135
8.3.2	Composition-based Feature Extraction.....	136
8.3.3	Evolutionary Feature Generation (PSSM Profiles).....	136
8.3.4	Feature Selection.....	137
8.3.5	Selection of appropriate ML models.....	137
8.3.6	Alignment-based and Motif-based Methods.....	137
8.4	Results.....	138
8.4.1	Compositional analysis	138
8.4.2	Alignment-Based Method.....	139
8.4.3	Alignment-Free Method.....	141
8.4.4	Hybrid approaches	144
8.5	SHAP-based interpretation	146
8.6	Benchmarking	146
8.7	Implementation of web server	147
8.8	Discussion and Conclusion	147
Chapter 9. Half-life prediction of natural and modified peptides		149
9.1	Introduction.....	150

9.2	Overview of the study	151
9.3	Material and Methods	151
9.3.1	Dataset construction.....	151
9.3.2	Datasets Used for Model Development	153
9.3.3	Feature extraction.....	154
9.3.4	Feature Selection.....	155
9.3.5	ML Regression Development	156
9.3.6	Evaluation of performance.....	156
9.4	Result	156
9.4.1	Prediction of half-life on the combine dataset	157
9.4.2	Prediction of half-life on the natural dataset.....	157
9.4.3	Prediction of half-life on the modified dataset	158
9.4.4	Feature Fusion.....	159
9.4.5	Comparative Analysis Across the Three Datasets	161
9.4.6	Models developed on selected features.....	161
9.4.7	Performance on Protein Language Model Embedding.....	162
9.4.8	Final Model.....	165
9.5	Implementation of web-server	166
9.6	Discussion and Conclusion	166
Chapter 10.	Summary	168
10.1	Limitations and Future Directions	171
Bibliography		175

List of Abbreviations

NK	Natural Killer Cells
Acc	Accuracy
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUROC	Area Under The Receiver Operating Characteristic Curve
BAC	Balanced Accuracy
BCE	B-Cell Epitope
BFD	Big Fantastic Database
BP	Binary Profile
DC	Dendritic Cells
DL	Deep Learning
DT	Decision Tree
DTR	Decision Tree Regressor
EM	Electron Microscopy
EN	Elastic Net
ET	Extra Trees
FAO	Food And Agriculture Organization
FIPA	Fluorescence-Based Ice Plane Affinity
GNB	Gaussian Naïve Bayes
HDX	Hydrogen-Deuterium Exchange
KNN	K-Nearest Neighbors
LGBM	Light Gradient Boosting Machine
LR	Logistic Regression
MAE	Mean Absolute Error
MCC	Matthews Correlation Coefficient
MHC	Major Histocompatibility Complex
ML	Machine Learning
MLM	Masked Language Modeling
MLP	Multilayer Perceptron Classifier
mRMR	Minimum Redundancy Maximum Relevance
MSE	Mean Squared Error
NCI	National Cancer Institute

NLP	Natural Language Processing
NMR	Nuclear Magnetic Resonance
NuSVR	Nu-Support Vector Regression
PAMPs	Pathogen-Associated Molecular Patterns
PCC	Pearson Correlation Coefficient
pI	Isoelectric Point
PLM	Protein Language Models
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool
PSSM	Position-Specific Scoring Matrix
QML	Quantum Machine Learning
QSVMs	Quantum Support Vector Machines
R ²	Coefficient Of Determination
RF	Random Forest
RFE	Recursive Feature Elimination
RMSE	Root Mean Squared Error
Sens	Sensitivity
SER	Shannon Entropy Of Residues
Spec	Specificity
SVC	Support Vector Classifier
SVMs	Support Vector Machines
T1DM	Type 1 Diabetes Mellitus
T _m	Thermostability
TPP	Thermal Proteome Profiling
TSL	Two Sample Logo
WHO	World Health Organization
XGB	Extreme Gradient Boosting
YI	Youden's Index

List of Figures

Figure 1.1 The overall organization of the human immune system	3
Figure 1.2 Schematic illustration of innate and adaptive immunity	3
Figure 1.3 Estimated lives saved through childhood vaccinations between 1974 and 2024	5
Figure 1.4 Scope of the thesis within the adaptive immune system. Of the two arms of adaptive immunity, the study focuses on the B-cell (humoral) arm, analysed through three tasks, B-cell epitope identification, vaccine safety screening, and vaccine stability evaluation	9
Figure 1.5 Illustrates the unified computational framework proposed in this thesis, organised into three modules. The framework first identifies candidate B-cell epitopes, followed by sequential safety assessment and stability evaluation, ultimately enabling the rational design of safe and stable vaccine candidates.	11
Figure 1.6 Organization of the thesis and chapter titles	11
Figure 2.1 Diagram illustrating the overview of innate and adaptive immunity	14
Figure 2.2 Classification of B-cell epitopes (a) linear epitope (b) conformational epitope	16
Figure 2.3 Experimental techniques commonly used for B-cell epitopes mapping	17
Figure 2.4 Major repositories related to B-cell epitope and prediction	18
Figure 2.5 Disease associated with an autoimmune disorder	24
Figure 3.1 General computational workflow employed in this thesis	37
Figure 3.2 Clustering based redundancy removal and representative sequence selection for non-redundant dataset construction	39
Figure 3.3 Binary profile matrix showing the dimension of 21 amino acids	44
Figure 3.4 The diagrammatic representation of 5-fold cross-validation ($k=5$)	55
Figure 4.1 Overview of CLBTope workflow for prediction of B-cell epitopes	62
Figure 4.2 Average compositional analysis of B-cell epitopes, non B-cell epitopes, and general proteome (adopted from [19])	65
Figure 4.3 Sequence logo representation of BCEs. The height of each amino acid indicates its relative frequency at that position, and the overall stack height represents positional conservation (adopted from [19])	66
Figure 4.4 ROC curve and AUROC scores of RF models based on ten composition descriptors (adopted from [19])	67
Figure 4.5. ROC curve and AUROC values showing the performance on hybrid features set (adopted from [19])	67

Figure 4.6 Comparative evaluation of classifier performance before and after feature selection for (A) training and (B) validation datasets (adopted from [19])	68
Figure 4.7 ROC curves and AUROC values demonstrating the performance of machine learning models trained using DPC features (adopted from [19])	69
Figure 4.8 Classifier performance across different DPC feature set sizes (400, 200, 150, and 100) for (A) training and (B) validation datasets (adopted from [19])	69
Figure 4.9 ROC curve and AUROC values showing the model performance on hybrid method (BLAST and DPC-400) (adopted from [19])	71
Figure 4.10 ROC curve and AUROC values comparing the performance of existing methods	73
Figure 4.11 Top 15 features driving B-cell epitope prediction, ranked by mean SHAP value	73
Figure 5.1 Brief on Type 1 Diabetes and immune destruction of β -cells (adopted from [102])	79
Figure 5.2 Overall architecture of the study	80
Figure 5.3 Amino acid composition analysis of T1DM-associated peptides, and random peptides	82
Figure 5.4 Illustration of sequence logo of T1DM associated peptides (adopted from [102])	83
Figure 5.5 Performance comparison of classifiers across feature subsets: (A) training, (B) validation datasets (adopted from [102])	85
Figure 5.6. SHAP-based feature importance for the T1DM-associated peptide prediction model	87
Figure 6.1 The complete architecture of the study	96
Figure 6.2 The length distribution of Allergenic vs non-Allergenic balanced dataset	97
Figure 6.3 Different protein language models used in this study including parameters	100
Figure 6.4 The amino acid composition comparison of allergen, and non-allergen	101
Figure 6.5 The amino acid composition comparison of allergen proteins, allergenic peptides	102
Figure 6.6 TSL plot showing positional amino acid preferences in allergenic and non-allergenic peptides across the N- and C-terminus	102
Figure 6.7 The AUROC of different feature types on different classifiers on the validation dataset	103
Figure 6.8 The top classifier performance on different sets of features (validation dataset) using the mRMR, LassoL1, and SVC-L1 feature selection (A) on hybrid features (B) DPC features.	105
Figure 6.9 The AUROC curve showing the performance of N8-C8 sets of binary profile patterns	106
Figure 6.10 The AUROC curve showing the performance of N-terminal (X) padding of binary profile patterns	106
Figure 6.11 The AUROC curve performance of (A) LLMs, (B) ET classifier on the embeddings extracted using LLM on the validation dataset	107
Figure 6.12 Ten discriminative motifs that were found in allergen (A) with minimum number of sites 10 (B) with minimum number of sites 5	109

Figure 6.13 SHAP-based feature importance for the AlgPred3 allergenicity prediction model	111
Figure 7.1 The overall architecture of the method (adopted from [143])	116
Figure 7.2 . Architectural diagram of ANN combined with MLP model (adopted from [143])	119
Figure 7.3 . Average percent amino acid composition of thermostable ($T_m > 50\text{ }^{\circ}\text{C}$ and $T_m < 50\text{ }^{\circ}\text{C}$) proteins	121
Figure 7.4 The correlation between the percent composition of AAC, SER, PAAC, and APAAC residues and T_m (adopted from [143])	122
Figure 7.5 The PCC of different regressors on different feature type (Validation dataset)	122
Figure 7.6 The R2 of different regressors on different feature type (Validation dataset)	123
Figure 7.7. SHAP-based feature importance for the PPTstab thermostability prediction model.	130
Figure 8.1 The overall workflow of the study (adopted from [158])	136
Figure 8.2 Average amino acid composition of antifreeze proteins (AFPs) and non-antifreeze proteins (non-AFPs).	138
Figure 8.3 The differences in average-composition scores of AFPs compared to general proteome and non-AFPs (adopted from [158])	139
Figure 8.4 ROC curve for models built using combined features on training and validation dataset (adopted from [158])	143
Figure 8.5 ROC curve for models built with mRMR-selected features on training and validation dataset (adopted from [158])	144
Figure 8.6 ROC curve built using PSSM-composition with AAC composition (adopted from [158])	145
Figure 8.7. SHAP-based feature importance for the antifreeze-protein prediction model	147
Figure 9.1 The workflow of the PlifePred2	152
Figure 9.2 The construction of plifepred2 dataset	154
Figure 9.3 Performance of composition-based descriptors on the combined peptide with flags	157
Figure 9.4 Performance of composition-based descriptors on the natural peptide	158
Figure 9.5 Performance of composition-based descriptors on the modified peptide with flags	159
Figure 9.6 illustrates the SHAP-based feature importance for the peptide half-life prediction model (composition-based features)	168

List of Tables

Table 1.1 Different types of vaccine with example	6
Table 1.2 Peptide-based vaccines in different phases of clinical trial, at NIH website [12]	7
Table 2.1 Early computational approaches for Linear B-cell epitope prediction	18
Table 2.2 ML and DL methods for linear B-cell epitope prediction	20
Table 2.3 Sequence based prediction methods of conformational B-cell epitopes	21
Table 2.4 Structure based prediction method for conformational B-cell epitopes	22
Table 2.5 Computational methods for allergenicity prediction of proteins/peptides	26
Table 2.6 Available computational methods for prediction of protein thermal stability	29
Table 2.7 Computational methods for antifreeze protein identification	31
Table 3.1 The majorly used databases with description	38
Table 3.2 Summary of redundancy removal strategies employed across different prediction tools	39
Table 3.3 Sequence-derived descriptor used for protein/peptide characterization	42
Table 3.4 Protein language models and their embeddings dimensions	45
Table 3.5 ML algorithms used in model development with grid search key hyperparameters	52
Table 4.1 Evaluation of BLAST similarity search at different E-values on the validation dataset (adopted from [19]).	70
Table 4.2: Model performance developed using the Hybrid method (BLAST and DPC-400) on training and validation dataset	71
Table 4.3. Comparative analysis of existing methods of B-cell prediction on the Validation Dataset	72
Table 4.4 Potential BCE in the SARS-CoV2 receptor-binding domain predicted using the proposed method	74
Table 5.1 Frequency of HLA-binding alleles observed in T1DM-associated peptides (adopted from [102])	83
Table 5.2 Performance of 12 composition-based descriptors using ET algorithm (adopted from [102])	84
Table 5.3 ML performance using DPC-150 feature subset on training and validation datasets (adopted from [102])	85
Table 5.4 Showing the frequency of top dipeptide composition (adopted from [102])	86
Table 5.5 Evaluation of BLAST similarity search on validation dataset (adopted from [102])	87
Table 5.6 Performance of Hybrid method (BLAST and DPC-150) (adopted from [102])	88
Table 5.7 Prediction of potential T1DM inducing antigenic peptides in the CBV4 polyprotein (adopted from [102])	90
Table 6.1 Various classifiers performance on ALL_COMB features	104

Table 6.2 Performance of comparison of QSVM with classical ML classifiers on composition-based features	108
Table 6.3 The performance of motifs on best performing ML score on validation dataset	109
Table 6.4 Performance of different classifiers on the SVCL1 (C 0.01) features	110
Table 6.5 The comparison of available methods on the validation dataset	111
Table 7.1 Top positively and negatively correlation features across descriptor (adopted from [143])	121
Table 7.2 Performance of different regressor models on SER features	124
Table 7.3 Performance of regression models using different LLM embeddings	125
Table 7.4 Performance of regression models using SER features with different features	126
Table 7.5 Reported performance of existing regression-based predictors (adopted from [143])	127
Table 7.6 Distribution of proteins across Tm ranges in selected microorganisms (adopted from [143])	128
Table 8.1 Top 5 motifs exclusive and inclusive motifs along with the number of sequences in which each motif was observed (adopted from [158])	139
Table 8.2 AUROC of different descriptors models on the validation dataset (adopted from [158])	141
Table 8.3 Performance of PSSM based ML models (adopted from [158])	142
Table 8.4. Performance on the PSSM composition profile combined with AAC composition on main (for training) and validation dataset (for Validation)	145
Table 8.4 Performance comparison of existing predictors on the validation dataset (adopted from [158])	147
Table 9.1 The flags with their description	153
Table 9.2 The performance of different regressors on ALL_COMP on combined peptide with flags	159
Table 9.3 The performance of different regressors on ALL_COMP on natural peptide	160
Table 9.4 The performance of different regressors on ALL_COMP on modified peptide with flags	160
Table 9.5 Performance of RFR model after feature selection on the combined dataset	162
Table 9.6 Performance of regression models using Protein Language Model embeddings on the combined dataset	163
Table 9.7 The complete performance of RFE-SVR 100 features on different regressor	165

CHAPTER

1

Introduction

1.1 Motivation and Background

The immune system is a highly coordinated and complex defense system composed of immune organs, proteins, leukocytes, and several chemicals. Together, these components protect the body (host) from invading pathogens such as parasites, bacteria, and viruses, and also help in eliminating abnormal cells including tumors and transplanted cells. At the same time, the system is able to recognize self-components and maintain tolerance towards host organisms, thereby preventing self-damage. Broadly, immune response can be divided into two types: innate immunity and adaptive immunity. The innate immune system is present from the birth and provides rapid, nonspecific defense that does not adapt to the characteristics of the individual pathogens. In contrast, adaptive immunity is characterized by antigen-specific responses mediated primarily by T and B lymphocytes, which undergo genetic rearrangements and generate immunological memory in response to antigen exposure [1]. Therefore, researchers are increasingly employing bioinformatics based sequence analysis, epitope mapping approaches, microarray platforms, and high-throughput immunological assays to identify the minimal and essential regions of B-cell epitopes [2]. This work focuses on the B-cell component of the adaptive immunity, emphasizing the identification and analysis of the B-cell epitopes and their relevance to vaccine development. The central idea is to integrate epitope identification, safety evaluation, and stability assessment into an integrated computational framework for rational vaccine design. A schematic representation of the human immune system is shown in Figure 1.1.

1.2 Types of Immunity

The immune system functions through two major types of immunity: innate immunity, which generates immediate but non-specific protection, and adaptive (acquired) immunity, which provides specific responses and immunological memory (see Figure 1.2) [3]. Innate immunity is the body's primary line of defence and provides immediate protection against invading pathogens. It is present from birth and acts as a general, non-specific defense mechanism against a wide range of microorganisms. For example, physical and chemical barriers, such as the skin, prevent the entry of pathogens into the body. In addition, the immune cells that contribute to innate immunity include the complement system, epithelial cells, neutrophils, natural killer (NK) cells, dendritic cells, mast cells, basophils, and eosinophils [4]. On the other hand, acquired immunity develops

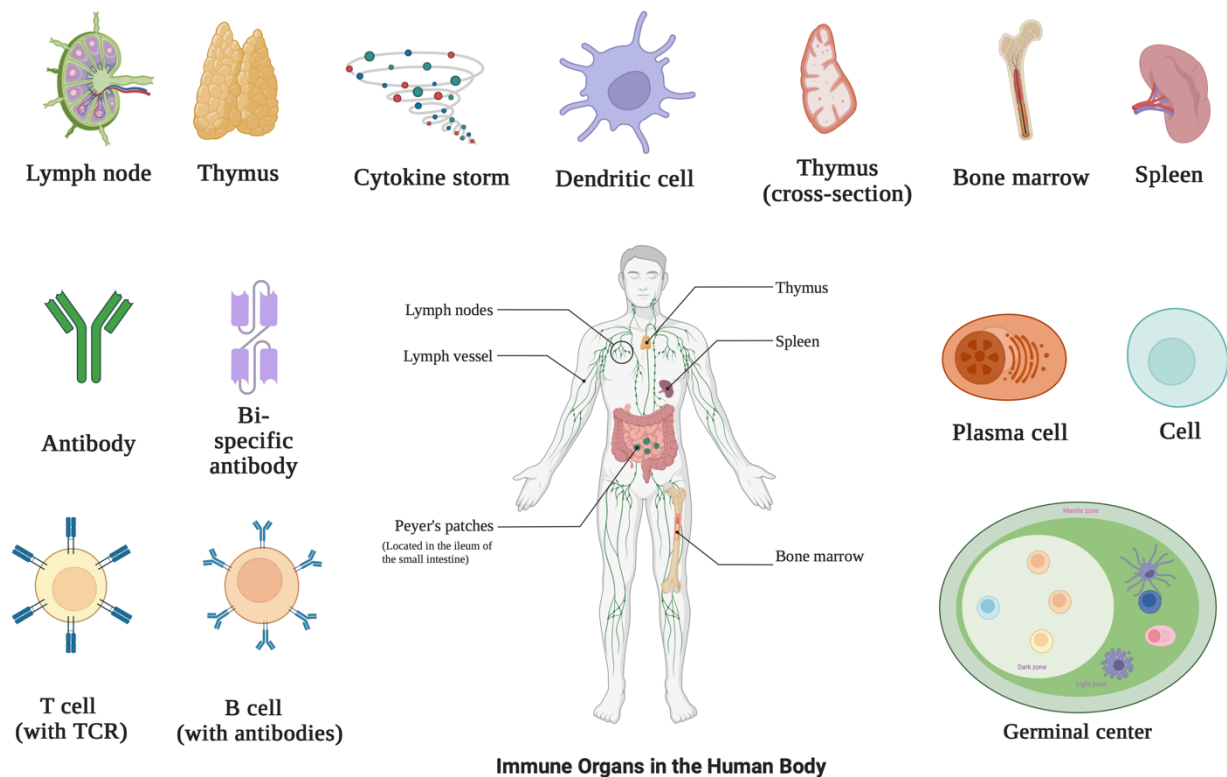


Figure 1.1 The overall organization of the human immune system

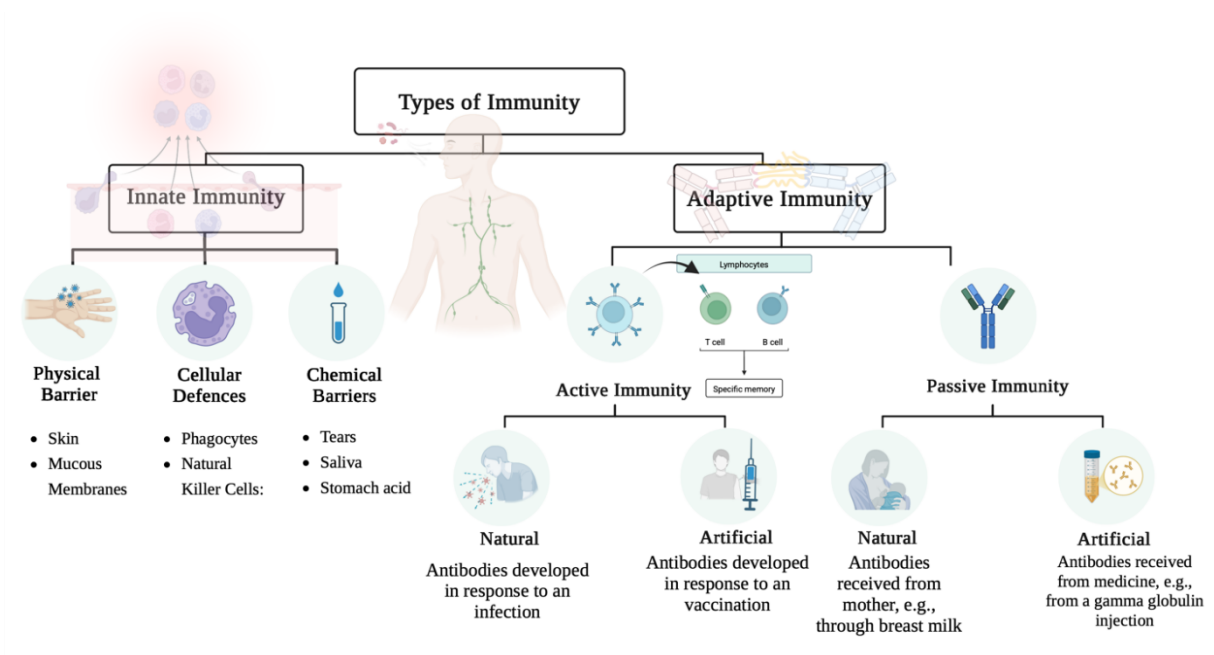


Figure 1.2 Schematic illustration of innate and adaptive immunity

after exposure to specific pathogens. This type of immunity is specific and is mediated by antigen-recognizing T lymphocytes and B lymphocytes. Although the adaptive response is slower to initiate compared with innate immunity, it is highly precise and less damage to host tissues. Adaptive immunity's ability to recall previously encountered pathogens is a fundamental characteristic; as a result, subsequent exposures elicit a more robust and rapid immune response. This characteristic is referred to as immunological memory [1].

1.3 Vaccination

Vaccines remain one of the most effective biomedical interventions in public health. They have contributed to the control and near elimination of several life-threatening infectious diseases, preventing millions of deaths worldwide [5,6]. There are five types of vaccines given in Table 1.1. By safely exposing individuals to a weakened, inactivated, or subunit form of a pathogen, vaccines enable the immune system to develop long-lasting protection without causing the disease itself. When vaccination coverage becomes sufficiently high, community-level protection (herd immunity) reduces disease transmission and protects even those who cannot be vaccinated [7]. Despite these successes, there are some challenges that remain, such as gaps in coverage persist across regions, particularly in resource-limited settings, and vaccine development, manufacturing, and cold-chain storage require substantial financial and logistical investment [8]. Efforts to accelerate and optimize vaccine development increasingly rely on computational approaches for antigen selection, epitope mapping, and immunogenicity assessment [6]. The continued progress, therefore, requires strategies that improve accessibility, reduce development cost, and enable the design of safe and effective vaccines for emerging and re-emerging diseases [9]. A recent global assessment illustrates the transformative role of vaccination; Andrew Shattock and colleagues estimated that routine immunization programs prevented approximately 154 million deaths over the last five decades, with measles vaccination alone saving an estimated 94 million lives [7] (shown in Figure 1.3 A-G).

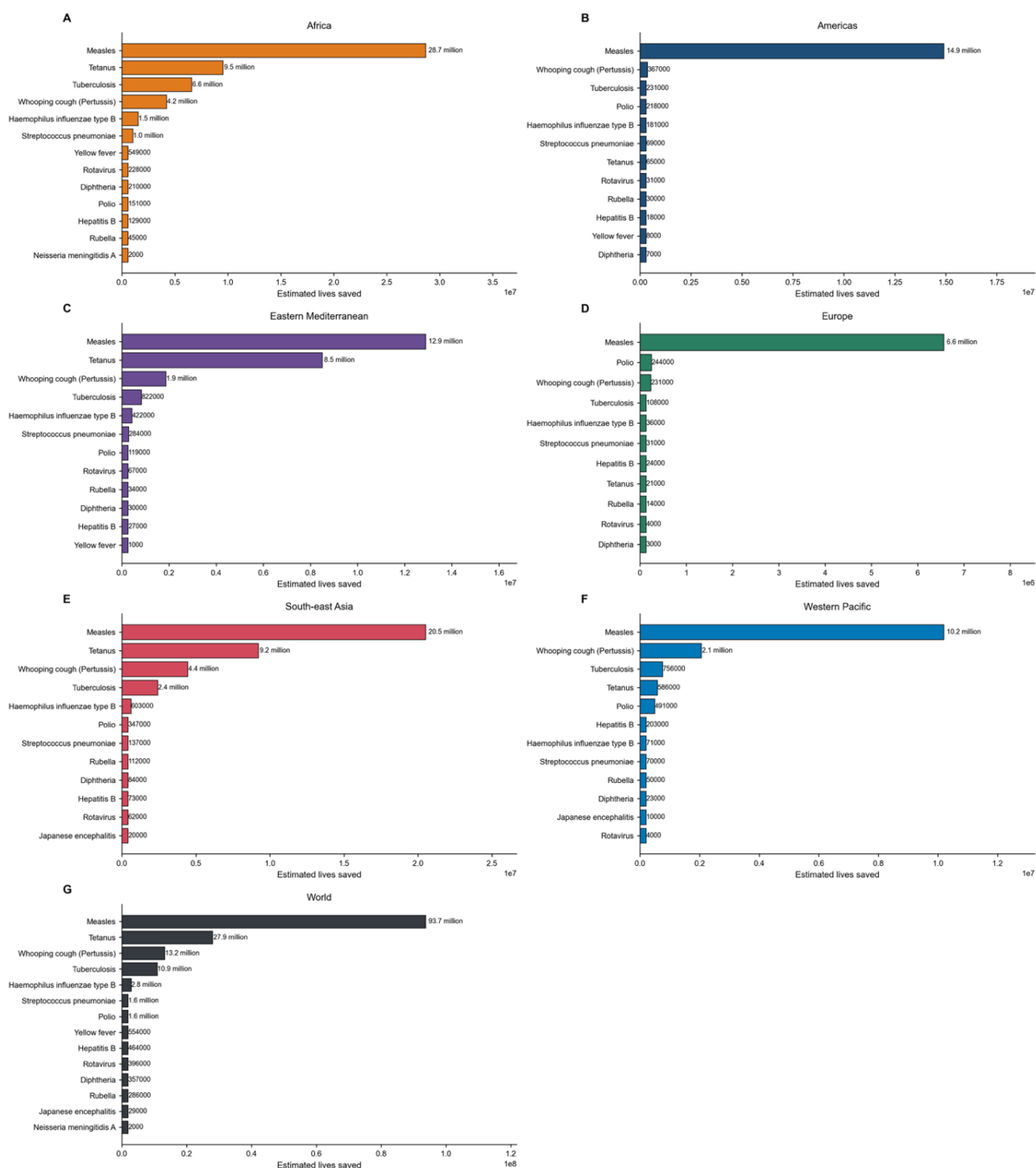


Figure 1.3 *Estimated lives saved through childhood vaccinations between 1974 and 2024*

Table 1.1 Different types of vaccine with example

S.No.	Vaccines	Examples	Description
1.	Inactivated (Killed) Vaccines	Rabies, Flu, Hepatitis A	The infectious agent is killed or inactivated via chemical and physical method so it cannot replicate.
2.	Live Attenuated Vaccines	Rotavirus, Variola (Smallpox), Varicella (Chickenpox), Yellow Fever, MMR	It has a weakened (attenuated) version of the microbe that causes the illness.
3.	mRNA (messenger RNA) vaccines	Comirnaty, Spikevax	It contains mRNA that encodes a specific viral protein. After entering the host cells, the mRNA is translated into protein, which then stimulates an immune response.
4.	Recombinant, Subunit, Polysaccharide, and Conjugate Vaccines	Hib, HPV, whooping cough, hepatitis B, a pneumococcal infection, Shingles and meningococcal disease	It includes fragments of the pathogen, as opposed to the whole microbe (such as sugar, protein, or capsid).
5.	Viral vector vaccines	COVID-19 Vaccine	It utilizes a genetically modified, harmless virus as a vector to deliver antigenic material and induce a protective immune response.
6.	Toxoid Vaccines	Diphtheria, Tetanus	It contains inactivated toxins (toxoids) produced by the pathogen. These non-toxic forms stimulate the immune system to recognize and neutralize the toxin during a natural infection

Hib, *Haemophilus influenzae* type b; HPV, Human papillomavirus; MMR, Measles, mumps, rubella;

1.4 Subunit Vaccine Design: Challenges and Opportunities

Traditionally, vaccines typically use whole pathogens administered either in live-attenuated or inactivated form. Although these approaches have proven effective, they may raise concerns related to the safety, reactogenicity, and production complexity. To address these limitations, subunit vaccines were developed. Instead of using the entire pathogen, subunit vaccines include only selected components, or antigens, that are capable of stimulating the immune system. These components may consist of proteins, peptides, polysaccharides, or their combinations, and they help address many safety and manufacturing challenges associated with whole-pathogen vaccines [10]. Peptide-based vaccines represent a particularly promising category. Instead of administering full-length proteins, these vaccines use short immunogenic regions that specifically target immune responses. Recent advances in genomics and immunoinformatics have enabled systematic in silico

scanning of pathogen genomes to identify antigenic proteins, locate epitope regions, and design peptide vaccine candidates [11]. Table 1.2 summarizes several peptide-based vaccine candidates that are currently under clinical evaluation.

Table 1.2 Peptide-based vaccines in different phases of clinical trial, at NIH website [12]

Peptide Vaccine	Firm/Institution	Trial Stage	Reference
Multi-epitope linear peptide based vaccine against influenza A and B	BiondVax Pharmaceuticals	Phase II	NCT02691130
PIpepTolDC: An immunotherapeutic vaccine for the treatment of T1DM	City of Hope Medical Center	Phase I	NCT04590872
Efficacy and safety Evaluation of Omalizumab with AIT	Qianfoshan Hospital	Active trial	NCT06771934
Idiotypic Antigen Vaccine Immunotherapy for Follicular Lymphoma	NCI	Phase I completed	NCT00001512
Peptide based therapeutic vaccine for HIV	United Biomedical Inc.	Phase I/II completed	NCT00002428
Fluvirin®, Agriflu subunit vaccine	Novartis Vaccines and Diagnostics	US FDA-approved in 2009	[13]
Multimeric-001 (M-001) synthetic peptide vaccine	NIAID	Phase II	NCT03058692, [14]
BOPRIVA [®] peptide vaccine induces antibodies against GnRF to controls aggressive behavior in bulls	Pfizer	launched	[15,16]

NCI, National Cancer Institute; NIAID, National Institute of Allergy and Infectious Diseases; AIT, allergen specific immunotherapy; HIV, human immunodeficiency virus; GnRF, Gonadotropin-releasing factor;

However, several challenges remain with the currently available vaccine approaches. These include incomplete protection against diverse pathogen strains, variability in immune responses, and potential safety and specificity concerns [17]. Subunit vaccines aim to address these issues by focusing on epitope-driven immune activation. Peptide epitopes capable of interacting with immune cell receptors can be identified using experimental and computational approaches and subsequently synthesized as vaccine components [11,17,18]. The length, sequence context, and physicochemical characteristics of peptides influence their ability to bind immune receptors and activate B and T cell responses, ultimately shaping vaccine efficacy [11]. Thus, accurate identification and careful selection of epitope regions are critical steps in rational vaccine design.

1.5 Role of Artificial Intelligence in Vaccine and Therapeutic Design

Compared with conventional experimental screening, computational epitope prediction offers significant advantages in speed, scalability, and cost. Over the last decade, intensive efforts have focused on developing *in silico* approaches for epitope detection, antigen characterization, and vaccine design supported by advances in structure biology, computational biology and immunology [11]. The combination of artificial intelligence (AI) and machine learning (ML) is a particularly active field. These approaches enable the analysis of large-scale biological sequence data to identify complex relationships between protein features and their biological properties, including antigenicity, allergenicity, autoimmune potential, thermostability, and peptide stability. The ability to design subunit vaccines that are immunogenic, safe, and stable while also reducing dependence on cold-chain systems has become an important goal in the next-generation vaccine research [10]. This work builds upon these advancements by developing AI-based computational approaches that:

1. Identify B-cell epitopes from primary sequences,
2. Evaluate their immunological safety, and
3. Predict and enhance the stability for vaccine development.

Together, these components contribute to a unified framework for rational vaccine and therapeutic protein design. In Figure 1.4, we have presented the study's complete objective.

1.6 A Unified Computational Framework for Subunit Vaccine Design

Although each prediction tool developed in this thesis addresses a distinct computational problem, together they constitute an integrated framework for rational subunit vaccine design. The overall workflow begins with the identification of potential B-cell epitopes from pathogen proteins, which represent candidate vaccine components capable of eliciting humoral immune responses. These candidate peptides are subsequently evaluated for safety by screening their potential to induce allergenic or autoimmune responses, thereby minimizing the risk of adverse immunological effects. Finally, the stability of the selected candidates is assessed using computational tools for predicting protein thermostability, peptide half-life, and antifreeze protein characteristics, ensuring that the proposed vaccine candidates possess favourable physicochemical properties for storage, delivery, and biological activity. Thus, the individual prediction tools developed in this thesis are interconnected components of a unified artificial intelligence-driven pipeline that facilitates the

identification, safety evaluation, and stability optimization of peptide- and protein-based vaccine candidates.

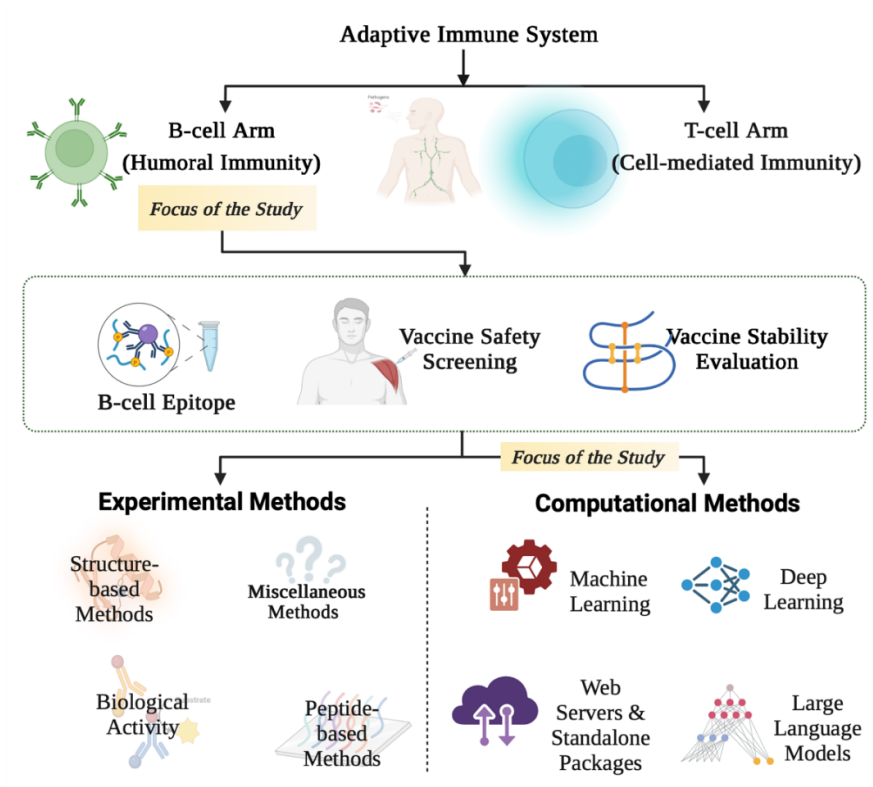


Figure 1.4 Scope of the thesis within the adaptive immune system. Of the two arms of adaptive immunity, the study focuses on the B-cell (humoral) arm, analysed through three tasks, B-cell epitope identification, vaccine safety screening, and vaccine stability evaluation.

1.7 Thesis Objectives

The overall objective of this study is to develop advanced tools for prediction, design, and screening potential candidates for subunit/protein vaccines. Broadly divided into the following three objectives:

Objective 1: Identification of Vaccine Candidates

Firstly, we focus on identifying B-cell epitopes (BCEs) that serve as critical targets for vaccine development. B-cells constitute a central arm of adaptive immunity, providing protection against pathogens through the secretion of specific antibodies following antigen recognition [19]. BCE have significant potential in immunological applications, particularly in the design of epitope-based vaccines, antibody-based therapeutic, and an immune diagnostic tool [20,21]. Therefore, the identification of both linear and conformational BCEs is essential for improving peptide-based

vaccine design and development. This work provides a foundation for computational subunit vaccine design, significantly accelerating vaccine and antibody development pipelines.

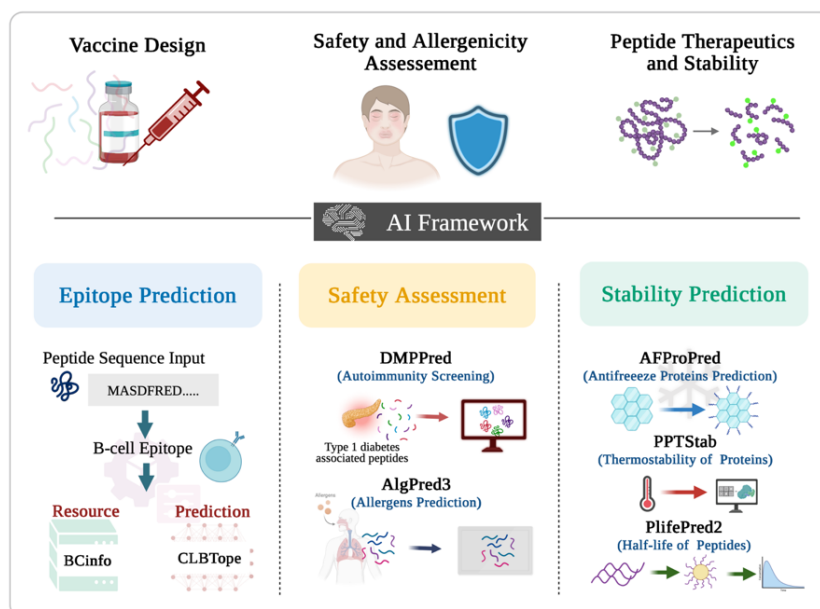
Objective 2: Safety Evaluation of Vaccine Candidates

There are several practical safety challenges when translating antigens into vaccines. First, vaccines can occasionally trigger autoimmunity in genetically predisposed individuals, observed in both animal studies and human cases. Molecular mimicry occurs when vaccine antigens resemble human proteins, causing the immune system to attack self-tissues. Secondly, allergic reactions, particularly immediate IgE-mediated responses, most individuals who are initially labeled as allergic to a vaccine can safely tolerate a subsequent dose without any clinical reaction [22]. Machine learning and large language models based approaches were used to predict allergenic peptide and to identify peptides associated with Type-1 diabetes mellitus. Together, these methods enable the early filtering of potentially unsafe antigenic peptides.

Objective 3: Stability and half-life

Over the last two decades, scientists have become increasingly concerned with maintaining the structures of proteins and peptides. Evaluating their stability under varying temperature conditions, including both heating and freezing/cooling, is essential for understanding temperature-induced damage (thermal injury), as thermal stress can lead to protein denaturation and loss of biological activity. An important aspect of computational protein design is that it will enable the creation of new proteins with valuable applications in medicine, industry, and research [23]. In this research, we develop methods for predicting the design of antifreeze proteins, the thermostability (T_m) of proteins, and the prediction of the half-life of peptides. This work focuses on developing and integrating computational approaches to assess and enhance the thermal stability, cold stability, and half-life of peptide/proteins. Three modules were developed to address this objective: antifreeze protein identification, thermostability (melting temperature) prediction, and half-life prediction of natural and modified peptides in biological systems. The schematic summary of the overall work carried out in this study is shown in Figure 1.5.

Figure 1.5 Illustrates the unified computational framework proposed in this thesis, organised into three modules. The framework first identifies candidate B-cell epitopes, followed by sequential safety assessment and stability evaluation, ultimately enabling the rational design of safe and stable vaccine candidates.



1.8 Organization of the chapter

Chapter 1	Introduction	
Chapter 2	Review of Literature	
Chapter 3	Dataset and Methodology	
Chapter 4	Prediction of B-cell epitopes	
Chapter 5	Prediction of T1DM associated peptides	
Chapter 6	Prediction of allergen epitopes	
Chapter 7	Antifreeze Protein Prediction	
Chapter 8	Thermostability (Tm) Prediction	
Chapter 9	Prediction of half-life	
Chapter 10	Summary	

Figure 1.6 Organization of the thesis and chapter titles

C H A P T E R

2

Review of Literature

In this chapter, we review the biological and computational background relevant to this study. First, the role of vaccination in immunotherapy and the functioning of the immune system are introduced to provide the biological context for peptide-based vaccine design. Then we discuss the growing importance of computational methods in comparison with traditional experimental methods, which helps in reducing cost, time, and in vitro screening efforts. After that, we mainly focus on the three major objectives addressed in this thesis. The first objective covers B-cell epitope identification, including experimental techniques, available biological repositories, and prediction methods based on sequence and structural information. The second objective discusses computational identification of disease-associated antigens and allergenic peptides, which is critical for early safety assessment in subunit vaccine development. Finally, the third objective reviews stability-related properties of proteins and peptides, including thermostability, antifreeze proteins, and peptide half-life, and emphasizes their importance for therapeutic efficacy, storage, and drug delivery.

2.1 Vaccination and Immunotherapy

One of the most advantageous public health strategies for both treating and preventing infectious and non-infectious diseases is vaccination. Traditionally, vaccines relied on whole pathogens in live-attenuated or inactivated forms. However, safety concerns and efficacy concerns associated with vaccines have motivated the development of more targeted approaches [24]. Therapeutic vaccines are a significant area of immunotherapy, especially for the treatment of chronic infections and cancers. The goal of these vaccines is to boost the host immune system's ability to identify and eliminate diseased cells [25,26]. The development of modern vaccines focuses on peptides or epitopes that may induce humoral and cellular immune responses against the disease. Peptide-based vaccines, utilizing defined B-cell and T-cell epitopes, have emerged over the past two decades as promising alternatives to conventional vaccine platforms. Their key advantage includes improved safety, reduced toxicity, and precise immunological targeting. However, successful development requires accurate identification of immunogenic regions and evaluation of safety and stability, motivating the application of computational approaches [27,28].

2.1.1 Different Immune System arms

The immune system serves as the primary defense against invading pathogens, including bacteria and viruses, and also eliminates abnormal cells, such as cancer cells, through a coordinated network of defense mechanisms. It is broadly categorized into two major branches: (i) Innate Immunity, and (ii) Adaptive (Acquired) Immunity (see Figure 2.1).

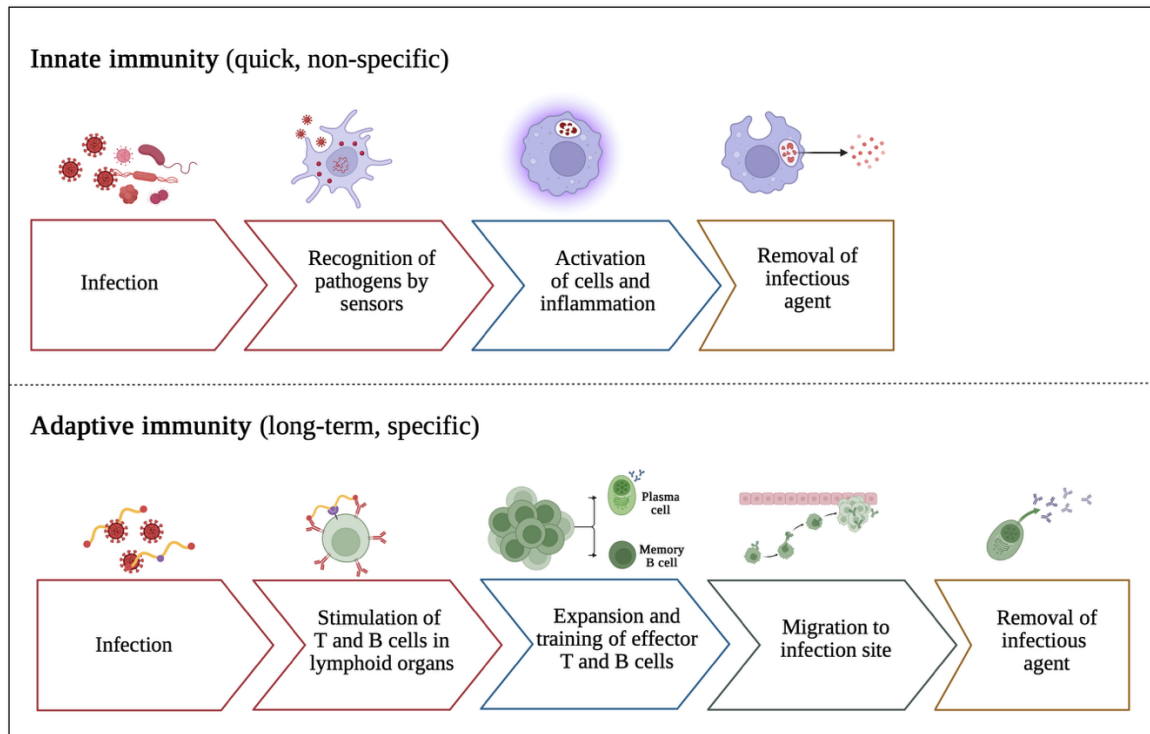


Figure 2.1 Diagram illustrating the overview of innate and adaptive immunity

2.1.1.1 Innate Immunity

Innate immunity serves as the first line of host defense, providing rapid and non-specific protection through physical barriers such as the skin and mucosal surfaces, as well as immune cells, including neutrophils and macrophages. Because this system responds in a similar manner to a broad range of foreign agents, it is often referred to as the non-specific immune response. The major cellular players include neutrophils, NK cells, macrophages, and dendritic cells (DC), which detect conserved pathogen-associated molecular patterns (PAMPs) using pattern recognition and subsequently trigger inflammatory responses. Although innate immunity does not generate long-term immunological memory, it is essential for initiating and directing the subsequent adaptive immune response [29,30].

2.1.1.2 Adaptive Immunity

When the innate immune response is insufficient to eliminate invading pathogens, the adaptive immune system is activated. Adaptive immunity, mediated by B lymphocytes and T lymphocytes, produces a highly specific and long-lasting defense against antigens. B lymphocytes, also known as B cells, are responsible for humoral immunity. They undergo differentiation into plasma cells, which circulate in the blood and other body fluids and release antigen-specific antibodies. T lymphocytes, also known as T cells, on the other hand, mediate cellular immunity by regulating immune responses and eliminating abnormal or infected cells. Following antigen exposure, a subset of activated lymphocytes develops memory cells, enabling faster and more robust responses upon re-encounter with the same antigen. This immunological memory response forms the basis of vaccination and long-term protection against diseases [3,4].

2.2 B-cell Epitopes (Humoral Immunity)

B lymphocytes are central mediators of humoral immunity in mammals and represent a key component of adaptive immunity. A B-cell epitope (BCE), also referred to as an antigenic determinant, is defined as the specific portion of an antigen that is recognized by the B-cell receptor or its soluble form (antibodies) secreted after B-cell activation [31]. In other words, a BCE is the part of an antigen that has direct contact with immunoglobulins. Usually, antibody recognition occurs at the solvent-exposed surface of the antigen [32]. BCEs are classified into two main categories: linear epitopes and conformational epitopes (Figure 2.2). Linear (continuous) epitopes are short peptide regions composed of sequential amino acid residues. In contrast, conformational (discontinuous) epitopes are formed by amino acids that are separated along the primary sequence but come into close proximity once the protein adopts its three-dimensional structure. As most antigens are proteins, they serve as the primary targets for epitope-prediction approaches. Therefore, identifying BCE is important because selected epitope fragments can replace full antigens for diagnostics, antibody production, and vaccine design. Epitope mapping is therefore widely applied in immunological research, disease monitoring, and therapeutic development, such as designing epitope-based vaccines [31,32].

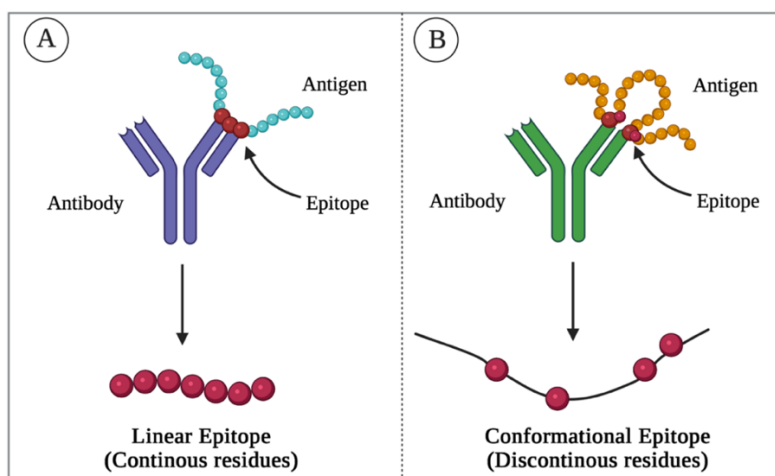


Figure 2.2 *Classification of B-cell epitopes (a) linear epitope (b) conformational epitope*

2.2.1 Experimental Identification of BCE

BCEs can be identified using several experimental approaches, including the determination of antigen-antibody complex structures, screening of peptide libraries for antibody binding, and functional assays in which mutations are introduced into the antigen and the resulting antibody-antigen complex is examined [32]. Broadly, experimental methods for BCE identification are categorized into structural and functional methods. Structural methods involve the analysis of antibody-antigen complexes using nuclear magnetic resonance (NMR), X-ray crystallography, hydrogen-deuterium exchange (HDX), and electron microscopy (EM). Functional approaches, on the other hand, evaluate antibody-antigen interactions through techniques such as surface plasmon resonance [33], mass spectrometry [34], and NMR spectroscopy [35], along with immunological assays including ELISA, ELISPOT, and Western blotting [36] (See Figure 2.3).

Experimental identification of epitopes is often labor-intensive, costly, and time-consuming because a large number of potential epitope candidates must be tested. To address these challenges, computational (*in silico*) prediction methods are now widely used to prioritize potential epitopes prior to laboratory validation, thereby significantly reducing the experimental workload. These methods help prioritize likely epitope regions in advance, thereby reducing the scale of laboratory screening and the overall experimental workload [32]. In the following sections, we first review the available databases and the prediction strategies used for linear BCEs, then discuss the prediction of conformational BCEs, and finally discuss the need for an integrated framework capable of identifying both types of epitopes.

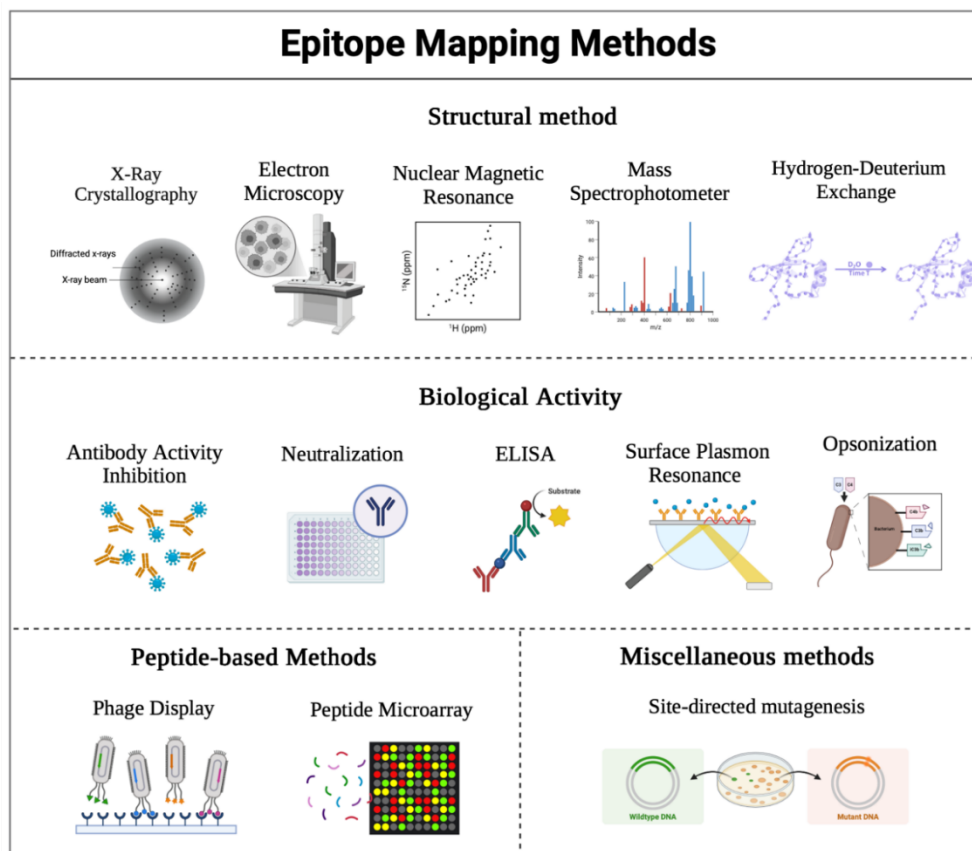


Figure 2.3 Experimental techniques commonly used for B-cell epitopes mapping

2.2.2 B-cell epitopes Repositories

A number of well-known databases related to BCEs are valuable resources for researchers developing computational prediction tools. To increase the accuracy and reliability of BCE prediction methods, access to experimentally validated data is crucial. With advances in biological and immunological techniques, a large volume of BCE-related data is regularly generated and shared through publicly available web databases and scientific publications. Figure 2.4 summarizes some of the most important databases related to BCEs [37].

2.2.3 Linear B-Cell Epitope Prediction Methods

Linear BCEs are short peptide segments that can often act as substitutes for whole antigens in immunization and antibody production studies. Although they represent a smaller fraction of all epitopes, they have attracted considerable attention because of their practical usefulness. Usually, sequence-based computational techniques are used to predict these epitopes directly from the primary amino acid sequence of antigens [32].

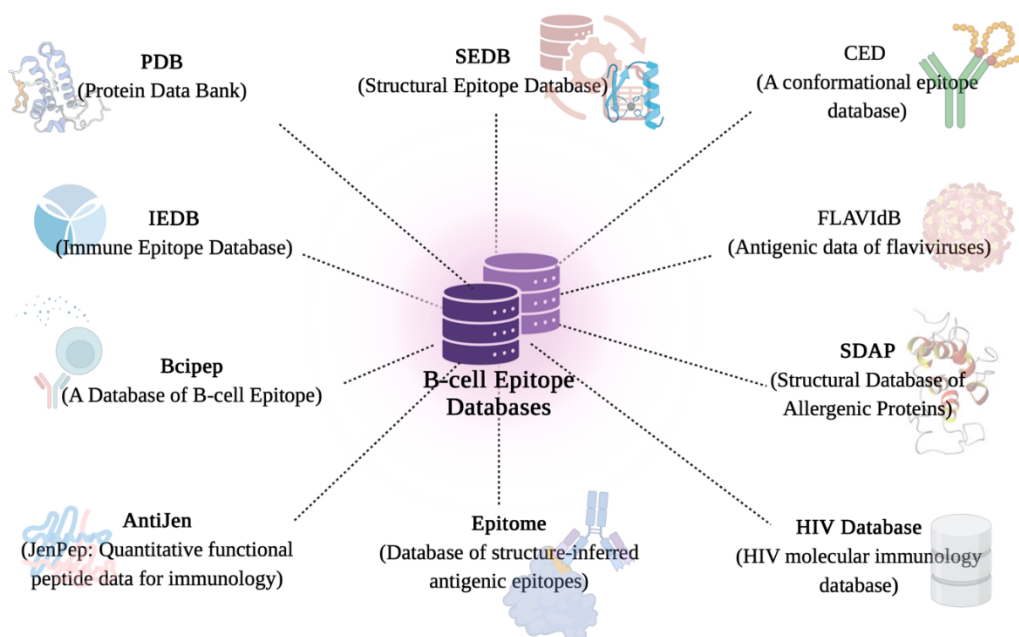


Figure 2.4 Major repositories related to B-cell epitope and prediction

2.2.3.1 Propensity Scales-Based Methods

Early approaches for predicting BCEs relied on amino acid propensity scales, derived from experimental observations. For example, Hopp and Wood used hydrophilic residue values to identify potential epitope regions [37,38]. Subsequently, additional propensity scales were proposed, including those based on solvent-accessible surface area [39], secondary structure [40], polarity [41], flexibility [42], accessibility [43], hydrophilicity [44], antigenic scale [45], and turns [46]. Later methods integrated several physicochemical characteristics, including hydrophilicity, flexibility, and polarity, to improve prediction accuracy. PEOPLE, PREDITOP, BEPITOPE, and BcePred are significant techniques developed under this category (Table 2.1). A prediction technique that integrated amino acid propensity scales with a hidden Markov model was subsequently proposed in 2006 to further enhance the performance [47,48].

Table 2.1 Early computational approaches for Linear B-cell epitope prediction.

Name	Year	Description	References
Janin and Wodak	1978	Protein prediction based on side-chain amino acid structure	[39]
Levitt Method	1978	Protein structure is used to predict the conformational preferences of amino acids.	[40]

Ponnuswamy	1980	Globular proteins' hydrophobic packing and spatial organization of amino acid residues	[41]
Hopp and Woods	1981	Antigenic determinants in proteins are predicted using a hydrophilicity-based scale.	[38]
Karplus & Schulz	1985	A scale for predicting B-cell epitopes based on flexibility.	[42]
Emini Method	1985	A score for identification based on accessibility.	[43]
Parker Method	1986	Hydrophilic, accessible, or mobile regions are used to predict antigenic regions.	[44]
Kolaskar Method	1990	Antigenic determinants in a protein are predicted using the physicochemical characteristics of amino acid residues.	[45]
Pellequer Method	1991	identifying a protein's continuous epitopes based on its primary structure.	[49]
PREDITOP	1993	A tool that employs 22 normalized scores based on physicochemical properties to predict antigenicity.	[50]
PEOPLE	1999	It analyze every residual property that was utilized in earlier techniques for linear B-cell epitope prediction.	[51]
BEPITOPE	2003	It is a newer version of PREDITOP that makes advantage of a variety of features, such as protein turns.	[52]
BcePred	2004	Predicting B-cell epitopes continuously by integrating all physicochemical characteristics.	[53]
BepiPred	2006	This method combines a hidden Markov model with propensity scale techniques to predict linear B-cell epitopes.	[47]
AAPPred	2009	linear prediction of B-cell epitopes based on amino acid pair residue properties.	[54]

2.2.3.2 Machine Learning Approaches

Several studies have shown that propensity scale-based methods possess inherent limitations [55] and generally exhibit poor predictive performance. Even with techniques created using propensity scales for all physicochemical properties, the accuracy is still below satisfactory. To address these limitations, researchers are applying machine learning techniques to develop more reliable prediction methods. In 2006, Saha and Raghava introduced ABCPred [56], a linear BCE prediction method based on neural networks. Similar to this, Söllner and Mayer [57] created a prediction method using the molecular operating environment framework that integrated the nearest neighbor algorithm and a decision tree. Since then, numerous tools employing machine learning strategies for BCE prediction have been reported (listed in Table 2.2). Singh et al. (2013) introduced the

LBtope [58] method, which was among the first to use K-Nearest Neighbor and Support Vector Machine algorithms to predict continuous BCEs using experimentally validated non-B-cell epitopes as negative datasets. Since then, several deep learning-based methods have been developed to improve prediction performance further.

Table 2.2 ML and DL methods for linear B-cell epitope prediction

Name	Year	Description	References
ABCpred	2006	For the first time, continuous B-cell epitopes have been predicted using artificial neural networks.	[56]
Söllner	2006	Propensity scales are used in a machine learning-based algorithm to predict epitopes.	[57]
FBCPred	2008	Linear B-cell epitope prediction using SVM with compositional features; variable length	[59]
BCPREDS	2008	A customized kernel string is used in an SVM-based approach to predict B-cell epitopes.	[60]
COBEpro	2009	Prediction in two stages based on residues and antigenic region propensity.	[61]
BayesB	2010	Bayes-based feature extraction for SVM-based linear B-cell epitope prediction	[62]
LEPS	2011	SVM model for linear epitope prediction based on physicochemical propensities	[63]
BEOracle/ BROracle	2012	Protein sequence-specific immune response determinants of antigenicity and specificity	[64]
SVMTriP	2012	Tripeptide similarity and propensity score were used in the development of this SVM-based approach.	[65]
BEST	2012	A sliding window technique developed utilizing SVM	[66]
LBTope	2013	A technique developed using B-cell and non-B-cell epitopes that have been experimentally verified.	[58]
EPMLR	2014	An epitope prediction approach based on sequencing.	[67]
LBEEP	2015	A model based on machine learning that was created utilizing the dipeptide deviation from the expected mean	[68]
DMN-LBE	2015	A deep maxout network using dropout training methods was used to develop a sequence-based predictor.	[69]
DRREP	2017	A B-cell epitope predictor based on deep learning.	[70]
iBCE-EL	2018	Highly randomized tree and gradient boosting classifiers are combined in an ensemble approach for B-cell epitope prediction.	[71]
DLBEpitope	2020	A DL-based prediction that integrates available techniques	[72]
BCEPS	2021	It has been attempted to predict B-cell epitopes that can cause native antigens to cause cross-reactive antibodies.	[73]

LBCEPred	2022	A RF model using statistical moment-based descriptors was developed to predict the B-cell epitopes.	[74]
Epidope	2021	Using a deep neural network, identify linear B-cell epitope regions on a large number of protein sequences.	[75]
BepiPred-3.0	2022	Using protein language models, it is an enhanced version of the BepiPred series.	[76]

2.2.4 Prediction of conformational B-Cell Epitopes

Nearly 90% of the BCEs occur as conformational epitopes rather than linear sequences [32]; however, predicting conformational epitopes has progressed more slowly than predicting linear epitopes. The limitation is due to two main factors. First, the three-dimensional (3D) structure of the antigen is typically necessary for conformational epitope prediction, and only a few proteins have this information. Second, it can be challenging to separate conformational epitopes from their native protein environment for the purpose of producing specifically targeted antibodies, and this often requires the use of appropriate structural scaffolds for epitope grafting. As a result, conformational epitope prediction is currently underutilized in antibody-based technologies and epitope vaccine development. However, for structure-function analyses of antibody-antigen interactions, determining conformational BCEs is still important [32,77]. The two main categories of conformational BCE prediction tools are sequence-based prediction methods (see Table 2.3) and structure-based prediction methods (see Table 2.4). Recently, several studies have attempted to extract linearized representations of conformational epitopes from antibody-antigen complex structures to improve prediction strategies [73]. In addition, integrated approaches capable of predicting both linear and conformational BCEs have been proposed to support computational vaccine design.

Table 2.3 Sequence based prediction methods of conformational B-cell epitopes

Name	Year	Description	References
Epitopia	2009	Machine-learning prediction of antigenic regions from sequence/structure	[78]
CBTOPE	2010	SVM-based prediction of conformational epitopes from sequence	[79]
CBEP	2014	Ensemble approach using spatial clustering information	[80]
BepiPred-2.0	2017	Random-forest model trained on antigen–antibody structures	[81]
SEPIa	2017	Naïve Bayes and random-forest hybrid predictor	[82]

Table 2.4 Structure based prediction method for conformational B-cell epitopes

Name	Year	Description	References
CEP	2005	Predicts sequential and conformational epitopes from 3D structure	[83]
DiscoTope	2006	Structure-based discontinuous epitope predictor with specificity.	[84]
ElliPro	2008	Structure-based epitope prediction and visualization	[85]
PEPITO	2008	Propensity and half-sphere exposure scoring	[86]
SEPPA	2009	The unit patch of residue triangles at the protein surface is used to predict the spatial epitope of a protein.	[87]
EPCES	2009	Surface conformation feature-based prediction	[88]
EPSVR	2010	SVM model using six scoring terms for epitope prediction	[89]
EPMeta	2010	Meta-server combining five prediction servers	[89]
DiscoTope 2.0	2012	Improved precision version of DiscoTope	[90]
SEPPA 2.0	2014	Incorporates relative surface accessibility	[91]
Seppa 3.0	2019	Extended SEPPA for glycoprotein antigens	[92]
CluSmote	2020	ML model using cluster sampling techniques	[93]
SEMA	2022	Deep-learning prediction from sequence and structure	[94]
Epitope3D	2022	Graph-based machine-learning epitope prediction	[95]

Despite significant progress in computational B-cell epitope prediction, several limitations remain. Early propensity scale-based methods rely on individual physicochemical properties and generally exhibit limited predictive accuracy because B-cell epitopes are determined by multiple sequence and structural characteristics rather than a single residue property. Machine learning-based predictors have substantially improved performance by integrating diverse sequence-derived descriptors; however, many models are trained exclusively on either linear or conformational epitopes, restricting their applicability across different epitope types. More recently, linearized conformational epitopes obtained from antibody-antigen structures, enabling their use in epitope prediction [73]. There is a need to develop a unified computational framework that can predict

both linear and conformational B-cell epitopes using only primary sequence information to improve the reliability and applicability of B-cell epitope prediction in vaccine design.

2.3 Safety Evaluation

Adaptive immunity, driven by antigen-specific T and B lymphocytes, provides a precise yet relatively delayed defense against pathogens through recognition of specific antigenic epitopes. Once the antigen is eliminated, immune homeostasis is restored by inhibitory receptors, regulatory immune cells, and anti-inflammatory mediators that terminate the response and limit tissue damage. In some situations, this regulatory control fails, leading to chronic inflammation or an abnormal immune reaction against self-antigens, known as autoimmunity. Therefore, identifying the initiating antigens and triggering factors is important for developing preventive measures and early therapeutic strategies [96]. Worldwide COVID19 immunization programs have markedly lowered the severity and death rate of SARS-Cov2 related disease. While their benefit to public health is beyond doubt, administering the vaccines to billions of individuals has allowed rare adverse events to come to light, among them autoimmune conditions such as systematic lupus erythematosus, rheumatoid arthritis, and type diabetes [97].

2.3.1 Autoimmune Responses and Vaccine Safety

The diabetes is a autoimmune chronic metabolic syndrome and type 1 diabetes mellitus is mainly characterized by the immune system's destruction of pancreatic β -cells, which produce insulin, a hormone that is necessary for blood glucose regulation [97] (see Figure 2.5). It affects millions of people worldwide; recent estimates suggest that approximately 9.5 million individuals were living with T1DM in 2025, including about 1.85 million individuals below 20 years of age [98]. Several clinical reports by Kanako Sakurai [99], Toshihiro Sato [100], and Toshimitsu Tanaka [101] have suggested a possible association between COVID-19 and vaccination and the onset of T1DM in individuals without a prior history of diabetes. These observations indicate that, in genetically susceptible individuals, vaccination may act as a trigger for autoimmune response, ultimately leading to irreversible damage to pancreatic islet cells and the destruction of β -cells. These cases indicate that T1DM should be considered as a rare but possible side effect of COVID-19 vaccination, even in previously healthy individuals, and emphasize the significance of careful clinical monitoring, especially in patients undergoing immunotherapy [97]. In the context of

vaccine development, this risk underscores the need for careful antigen selection, particularly for subunit vaccines composed of short immunogenic peptides.

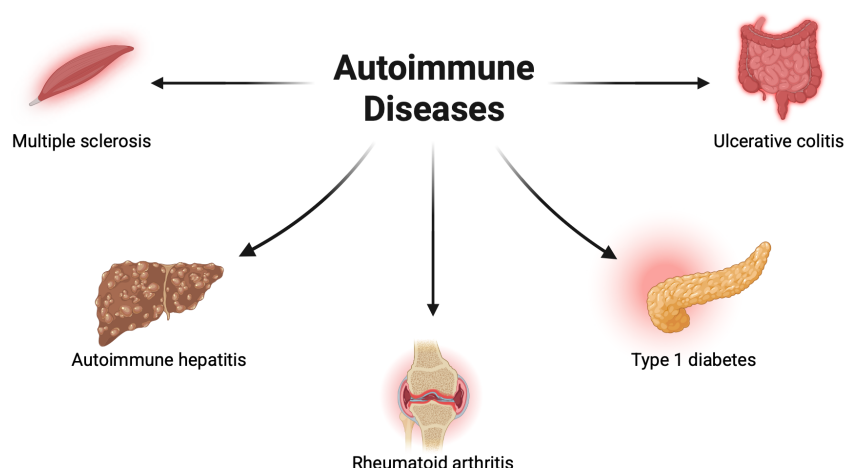


Figure 2.5 Disease associated with an autoimmune disorder

2.3.1.1 Computational Identification of T1DM-associated Regions

Identifying peptides associated with T1DM is essential for understanding disease initiation and progression. Usually, these peptides attach to major histocompatibility complex (MHC) class II molecules and trigger immunological reactions that ultimately destroy pancreatic β -cells [102]. However, only a limited number of studies have explored computational methods for predicting peptides that are associated with T1DM. For example, the GPS-MBA method was created by Cai et al. [103] to predict MHC class II binders associated with type 1 diabetes; however, the study relied on a small dataset restricted to HLA-DQ8 binders, which limited its general applicability. Thus, it is clear that a more reliable and accurate prediction tool must be created using the most recent data [102]. Early identification of T1DM-associated peptides is also crucial for improving safety risks during vaccine development and therapeutic design. Peptide immunotherapy employs short peptide epitopes derived from disease-associated autoantigens rather than the entire antigen, allowing modulation of the autoimmune response while minimizing unwanted adverse effects [104]. To date, only a few *in silico* approaches have attempted to predict T1DM-associated peptides. A major limitation of existing methods is that they were developed on a small dataset comprising only HLA-DQ8 binders [103], which restricts their applicability for the comprehensive safety assessment of vaccine candidates. Therefore there is a need for more accurate prediction that leverages the latest available dataset. These limitations highlight the need for robust

computational models trained on carefully curated datasets with rigorous independent validation for reliable prediction of T1DM-associated peptides.

2.3.2 Allergenicity and Vaccine Safety

A rising global health concern, allergic diseases affect more than 20-30% of the world's population and are becoming more complex and severe, affecting both adults and children [105]. Immune-mediated inflammatory disorders, including food allergies, atopic dermatitis, asthma, and allergic rhinitis, are caused by hypersensitivity of the immune system to normally harmless environmental substances [106]. Allergic responses to vaccines occur when the immune system reacts to one or more components of the vaccine formulation. Similar to other pharmaceutical drugs, vaccines have the potential to induce hypersensitivity reactions, which may be triggered by the antigen itself or by other formulation components [107]. Despite the rarity of serious allergic reactions, vaccination remains highly beneficial, with its advantages far exceeding the associated risks for the vast majority of the population. However, vaccine hesitancy can be influenced by concerns about potential allergic reactions, which can impact efforts to control the spread of infectious diseases and herd immunity. Immediate allergy is driven mainly by allergen specific IgE, which bind the high affinity FcεRI receptor on the mast cells and basophils [108–110], leading to the release of inflammatory mediators and the development of hypersensitivity symptoms. Improved understanding of vaccine-related allergic reactions can help enhance vaccine safety and the manufacturing process. Identifying the specific components responsible for hypersensitivity could enable their removal or substitution with safer alternatives, thereby lowering the risk of vaccine-related allergic reactions. This information is also important for clinical management, as it supports the design of personalized vaccination strategies for individuals at higher risk of allergic reactions [110].

2.3.2.1 Computational Approaches for Allergen Prediction

In the post-genomic era, the rapid growth of protein sequence data poses challenges for traditional experimental techniques, which are laborious and time-consuming [111,112]. In order to solve this, the Food and Agriculture Organization (FAO) and the World Health Organization (WHO) developed guidelines for predicting allergenicity[113]. These include: (i) identifying proteins containing a stretch of 6–8 consecutive amino acids identical to known allergens, and (ii) classifying proteins with >35% sequence similarity over an 80-residue window as allergens. These

two approaches identify allergenic proteins by using sequence alignment techniques and the homology concept. While effective for homologous allergens, these methods struggle with non-homologous proteins, often leading to false identifications [113,114].

To overcome such limitations, computational approaches have been developed to predict allergenicity using machine learning have been widely explored, as shown in Table 2.5. Despite these advances, significant challenges remain. Currently available allergen prediction methods considered for safety evaluations mainly focus on whole-protein analysis and present significant limitations, failing to explicitly evaluate discrete allergen epitopes or peptides. An epitope-focused prediction tool provides a more robust assessment of allergenicity [115]. Therefore, an efficient machine learning method, based on identifying a set of allergenic peptides, is the need of the hour.

Table 2.5 Computational methods for allergenicity prediction of proteins/peptides

Methods	Year	Description	Reference
AlgPred	2006	Hybrid SVM model integrating motifs, BLAST similarity, and physicochemical features	[116]
AllerTool	2007	SVM and sequence-similarity approach for IgE cross-reactivity detection	[117]
AllerHunter	2009	SVM model using pairwise sequence similarity encoding and FAO/WHO allergenicity rules for improved prediction	[118]
PREAL	2013	SVM predictor utilizing biochemical and physicochemical properties of amino acids	[119]
AllerTOP	2013	k-nearest neighbor classifier based on auto- and cross-covariance transformation of protein sequences	[120]
Allerdicator	2014	Text-classification inspired SVM approach representing protein sequences as word-frequency vectors	[121]
AllergenFP	2014	Descriptor fingerprint method using physicochemical property vectors for allergen identification	[122]
AllerTop v.2	2014	Improved k-NN classifier employing amino acid E-descriptors and ACC features	[123]
AllerCatPro	2019	Predicts allergenicity using sequence homology together with structural similarity and epitope evidence	[124]
AlgPred 2.0	2021	Ensemble random-forest model combining motif search, BLAST similarity, and allergen representative peptides	[125]
AllerCatPro 2.0	2022	Updated platform integrating improved structural modeling and epitope similarity analysis	[126]
ProAll-D	2022	LSTM-based deep learning model capturing sequential dependencies in protein sequences	[127]

ALLERDET	2022	Hybrid decision tree and restricted Boltzmann machine deep-learning classifier	[128]
DeepAlgPro	2023	Convolutional neural network with multi-head self-attention to capture contextual sequence features	[129]
iAller	2023	Random forest using multi-feature fusion including composition and physicochemical descriptors	[130]
SEP-AlgPro	2024	Machine- and deep-learning prediction based on protein language model embeddings	[131]
AllergenAI	2024	CNN framework incorporating predicted three-dimensional structural information	[132]
PreAlgPro	2024	CNN classifier utilizing ProtT5 protein language model representations	[133]
AllergyPred	2025	Deep neural network trained on extensive physicochemical descriptors	[134]

Numerous computational methods have been developed for allergen prediction using sequence similarity, motif-based approaches, physicochemical descriptors, and machine learning algorithms. Despite their success, several limitations remain. Currently, available allergens prediction methods that is considered for safety evaluations majorly focuses on the whole protein based analysis. These methods do not explicitly evaluate discrete IgE-binding epitope or allergenic peptides. Therefore, an efficient machine learning method, dedicated allergenic peptide prediction framework specifically optimized for short allergenic peptides is the need of the hour for accurate, scalable, and broadly applicable allergen prediction methods suitable for modern peptide therapeutics and vaccine design.

2.4 Stability Evaluation

Most currently available vaccines require maintenance of a continuous cold chain, meaning they must be stored and transported between 2-8 °C from the time of production until administration. Vaccines that experience temperature excursions outside this range may undergo loss of potency. Although this system is highly effective when properly implemented, it relies on refrigerated packaging, temperature-controlled transportation vehicles, healthcare centres equipped with reliable refrigerators, and trained personnel. This makes the process costly and also susceptible to lapses in temperature control [135,136]. In many developing countries, maintaining uninterrupted refrigeration is often difficult due to infrastructure limitations and irregular electricity supply.

Consequently, substantial quantities of vaccine material are lost each year, and there is also a risk that vaccines with reduced protective efficacy may be administered [137].

Preserving the structural integrity of protein- and peptide-based therapeutics has therefore become a major scientific objective over the past few decades. These biomacromolecules possess only marginal stability and remain stable within a narrow range of temperature, concentration, pH, and ionic strength. Both cold-induced and heat-induced denaturation can significantly affect their stability, making them highly prone to physical and chemical degradation. The WHO has emphasized the importance of addressing these stability challenges [138], as temperature deviations, either above or below the recommended 2-8 °C range, can have serious consequences for vaccine effectiveness. Therefore, there is a clear need for vaccines that can withstand both freezing and elevated temperature conditions, and the development of thermostable vaccines is receiving increasing attention. For this reason, the introduction of thermally stable vaccines that do not rely heavily on cold-chain logistics has become an important global health priority [137].

2.4.1 Temperature Stability

Elevated temperatures can adversely affect different types of vaccines in multiple ways. In subunit vaccines, high temperature may lead to unfolding of protein antigens; in conjugate vaccines, it can cause dissociation of polysaccharides from their protein carriers; and in live attenuated vaccines, it may reduce the viability and infectivity of the organism. Although a few proteins show remarkable resistance to heat, most proteins are unable to maintain their native conformation under thermal stress and tend to convert into aggregated or inactive forms [138,139]. A protein's melting temperature (T_m) is typically used to determine its thermal stability [140], which corresponds to the temperature at which approximately half of the protein molecules become unfolded. This value varies among proteins and typically ranges from 40 to 80 °C in natural systems. During the production, purification, storage, packaging, and transportation of therapeutic proteins, it is therefore essential to maintain temperatures below the protein's melting point to preserve activity and efficacy [138]. Experimental determination of protein T_m relies on advanced methods like Fourier Transform Infrared Spectroscopy, Differential Scanning Calorimetry (DSC), Mass Spectrometry-based Thermal Proteome Profiling (TPP), and Circular Dichroism (CD) [141]. However, due to their cost, time requirements, and experimental complexity, these methods are not suitable for large-scale genome-wide screening of proteins with unknown T_m values. The

Meltome Atlas and ProThermoDB are two specialized databases that have been created recently using experimentally measured T_m values [142]. Therefore, there is a growing need for computational (in silico) approaches that can efficiently identify thermostable proteins directly from sequence databases [140,143]. Although several prediction methods have been proposed previously, many of them are trained on redundant datasets (as shown in Table 2.6), which may limit their performance on proteins that are dissimilar to those in the training set. Consequently, there is a clear requirement to develop prediction models using carefully curated non-redundant protein datasets and standard bioinformatics protocols.

Table 2.6 Available computational methods for prediction of protein thermal stability

Method	Year	Methodology	Reference
(A) Regressor based predictors			
ProtStab	2019	Gradient boosting regression trees using sequence descriptors from PROFEAT, ProtDCal, and ProtParam	[144]
iStable 2.0	2020	Integrated regression/classification framework using Weka models and XGBoost	[145]
ProtStab2	2022	LightGBM regression trained on large-scale feature descriptors (protr, ProtDCal, ProtParam)	[146]
DeepSTABp	2023	Transformer-based protein language model embeddings with MLP predictor	[147]
DeepTM	2023	Graph convolutional networks with self-attention and MLP for melting temperature prediction	[140]
(B) Classifier based predictors			
SCMTPP	2021	SVM classifier using dipeptide propensity scores	[148]
TMPpred	2022	SVM classifier with ANOVA-selected 188-dimensional feature set	[149]
SAPPHIRE	2022	Stacking ensemble model for thermophilic protein identification	[150]
DeepTP	2023	CNN + BiLSTM with self-attention followed by MLP classification	[151]
BertThermo	2023	BERT-based embeddings with feature engineering and ML classifiers (LGBM/LR)	[152]
ProLaTherm	2023	ProtT5 protein language model embeddings for thermophilicity prediction	[153]
TemStaPro	2024	Binary classifier using ESM-2 and ProtT5 language-model embeddings	[154]

Although several computational approaches have been developed for predicting protein thermostability and melting temperature (T_m), important challenges remain. Most existing predictors have been trained on highly redundant protein datasets, which can lead to overly optimistic performance estimates and limit their ability to generalize to evolutionarily distant proteins. Consequently, these models may perform poorly when applied to novel proteins sharing low sequence similarity with those used during training. Furthermore, the majority of available methods are restricted to predicting protein thermostability and do not provide capabilities for designing proteins with user-defined or desired melting temperatures. These limitations highlight the need for robust prediction models developed using rigorously curated non-redundant datasets and standardized bioinformatics protocols, together with computational frameworks that support both prediction and rational protein design.

2.4.2 Cold Denaturation and Antifreeze Proteins

The study of the effects of low temperature on protein structure can be traced back to the 1930s, particularly the work of Hopkins [155]. It was later understood that cold-induced unfolding primarily results from alterations in the interaction pattern between water molecules and the non-polar groups present in proteins [138]. In the 1950s, Scholander and his colleagues reported that fish species could survive freezing temperatures, challenging the prevailing assumptions about the limits of life in cold environments [156]. Subsequently, in 1969, DeVries and co-workers associated this remarkable adaptation with the presence of antifreeze proteins (AFPs) [157]. Since then, AFPs have been identified in a wide variety of organisms, including insects, fungi, bacteria, and mammals [158]. Antifreeze proteins are specialized proteins, glycopeptides, or peptides produced by certain organisms to enable cellular survival under sub-zero conditions. They function by lowering the freezing point of water and inhibiting the growth of ice crystals during freezing. Their ability to modify ice crystal growth stabilizes ice structures within a specific temperature range and prevents ice recrystallization, which also helps reduce drip loss during thawing. In recent years, research on AFPs has focused on their discovery, purification, detection, mechanism of action, and bioinformatics-based characterization, highlighting their potential for diverse applications [159]. Owing to their freeze-resistance properties, AFPs have been explored for multiple practical uses, including food preservation [160], medical applications [161], cryosurgery, and even the production of yoghurt [158,162]. Several experimental approaches have

been developed to identify and characterize AFPs. These include ice-etching methods [163], fluorescence-based ice plane affinity (FIPA) assays [164], ammonium precipitation and ice-affinity chromatography for purification [165], and site-directed mutagenesis to identify ice-binding planes and to understand the mechanism of AFPs [166]. These experimental methods are also expensive and time-consuming. Furthermore, distantly related AFPs are frequently eliminated by traditional similarity-based sequence search techniques like BLAST and PSI-BLAST [167]. Consequently, computational prediction approaches have gained importance for identifying AFPs, particularly due to their significant biomedical and industrial relevance. Although several methods have been proposed to distinguish AFPs from non-AFPs, many of them rely on unreviewed datasets for evaluation [158]. Therefore, the development of prediction methods based on curated, experimentally validated data has become essential; an overview of currently available methods is presented in Table 2.7.

Table 2.7 Computational methods for antifreeze protein identification

Name	Year	Source/Database	Training Dataset	Independent/ Test Dataset	Reference
AFP-Pred*	2011	Pfam database	481 AFPs, 9,493 non-AFPs	-	[167]
iAFP*	2011	PDB, UniProKB	44 AFPs, 3,762 non-AFPs	369 AFPs	[168]
AFP_PSSM*	2012	AFP-Pred	481 AFPs, 9,493 non-AFPs	-	[169]
AFP-PseAAC	2014	AFP-Pred	481 AFPs, 9,493 non-AFPs	-	[170]
TargetFreeze*	2015	AFP-Pred, Pfam database	300 AFPs, 300 non-AFPs	181 AFPs, 8,293 non-AFPs	[171]
AFP-ensemble*	2015	AFP-Pred	371 AFPs, 7,266 non-AFPs	93 AFPs, 1,817 non-AFPs	[172]
CryoProtect	2016	AFP-Pred	478 AFPs, 9,139 non-AFPs	-	[173]
iAFP-Ense	2016	AFP-Pred, Pfam	481 AFPs, 9,193 non-AFPs	-	[174]
RAFP-Pred*	2018	AFP-Pred, iAFP, UniProKB	Dataset1: 481 AFPs, 9,493 non-AFPs; Dataset2: 44 AFPs, 3,762 non-AFPs	Dataset3: 369 AFPs	[175]

afpCOOL*	2018	AFP-Pred, UniProtKB	Dataset1: 481 AFPs, 9,493 non-AFPs	Dataset2: 517 AFPs, 517 non-AFPs	[176]
AFP-CKSAAP	2019	AFP-Pred, Pfam	481 AFPs, 9,493 non-AFPs	-	[177]
W-GDipc-LRMR-Ri	2019	AFP-Pred, MemType-2L	480 AFPs, 374 non-AFPs	7,582 MemType-2L proteins	[178]
AFP-LSE	2020	AFP-Pred	481 AFPs, 9,493 non-AFPs	-	[179]
Sun et al.	2020	AFP-Pred	481 AFPs, 9,493 non-AFPs	-	[180]
AFP-CMBPred	2021	AFP-Pred, Pfam database	300 AFPs, 300 non-AFPs	181 AFPs, 8,293 non-AFPs	[181]
AFP-LXGB*	2022	AFP-Pred, Pfam database	481 AFPs, 9,193 non-AFPs	-	[182]
AFP-SPTS*	2022	AFP-Pred, Pfam database	481 AFPs, 9,193 non-AFPs	-	[183]
AFP-SRC	2022	AFP-Pred	481 AFPs, 9,493 non-AFPs	-	[184]

*Web server is not working properly or not available.

Although numerous computational methods have been developed for the identification of antifreeze proteins (AFPs), several limitations remain. Most existing predictors rely on sequence composition, evolutionary features, or sequence similarity; however, sequence similarity-based approaches often fail to identify distantly related AFPs with low sequence identity. A major limitation of the current methods lies in the datasets used for model development and evaluation. Most predictors have been trained and benchmarked using the AFP-Pred dataset, which comprises 481 AFPs and 9,493 non-AFPs, while a few studies have utilized proteins from the PDB. However, these datasets predominantly contain unreviewed or computationally annotated sequences rather than experimentally validated, manually curated proteins. Consequently, the reliability and generalizability of these models may be compromised. The availability of reviewed AFP sequences in the UniProt/Swiss-Prot database provides an opportunity to develop more robust and biologically reliable prediction models based on experimentally supported data.

2.4.3 Half-Life in Body Fluids

Despite the successful development of a number of vaccines, significant challenges still exist. These include weak immune durability, particularly in subunit vaccines, limited protection against rapidly evolving variants, and insufficient antibody titres, which are frequently caused by their

short-term *in vivo* exposure. Furthermore, because of their rapid systemic clearance, protein-based vaccines and therapeutics typically have a short biological half-life after administration. It has been demonstrated that increasing the half-life can significantly improve drug efficacy [185–187], primarily because of prolonged circulation time and sustained concentrations in the body. However, the influence of *in vivo* persistence on vaccine performance, particularly for subunit vaccines, and its impact on immune response and protection against infectious diseases has been comparatively underexplored [188]. Even though there is a growing need for therapeutic proteins and peptides, their short half-life, primarily from enzymatic degradation and low bioavailability, often limits their clinical efficacy. The estimation of peptide half-life has been investigated using both experimental and *in silico* methods. Thus, extending the peptide's *in vivo* stability is crucial to optimizing therapeutic benefit while reducing the frequency of dosages and administration burden [189].

To address this limitation, several strategies have been proposed. Proteins with a molecular weight of 20-40 kDa have frequently been conjugated with polyethylene glycol (PEG) to increase their half-life. However, PEGylation has gradually become less favourable because of immunogenicity concerns and limited long-term clinical success. One of the more successful approaches is peptide acylation or lipidation. In this method, a peptide is conjugated with a fatty acid (FA) molecule, which binds to human serum albumin (HSA), thereby extending plasma half-life from minutes to several days. Consequently, fatty-acid acylation has emerged as the most effective and successful technique for improving peptide stability and half-life [190,191]. Structural modification strategies have also been applied in therapeutics. For example, in semaglutide, a modification at the second N-terminal amino acid position (Ala8) protects the active GLP-1 peptide from degradation by DPP-4, thereby improving its half-life [192]. As an alternative to PEGylation, genetic fusion with human serum albumin (HSA; 67 kDa) or antibody Fc domains offers another effective approach [193]. Additional stabilization methods include peptide cyclization, substitution of L-amino acids with D-amino acids, and the development of hybrid peptides [194]. Physical methods have also been studied; for example, coating gold nanorods (GNRs) with amphiphilic peptides (AAPs) increased the mice's blood circulation half-life to 37.8 hours, as compared to 26.6 hours for PEG-coated GNRs [195]. Many experimental and computational studies have been utilized to enhance and optimize peptide half-life because of the significance of this field [189].

Previously, many attempts to enhance peptide half-life in blood have included peptide cyclization, terminal modifications, and incorporation of modified amino acid residues. Experimental techniques for determining peptide half-life are well established and highly accurate; however, they are also costly, labor-intensive, and time-consuming. Hence, alternative predictive approaches are required, and in this context, computational prediction methods are becoming increasingly valuable [196]. Several tools have already been developed to estimate protein or peptide half-life. For example, based on the N-end rule, ProtLifePred and ProtParam [197] predict the half-life of proteins in mammalian, *S. cerevisiae*, and *E. coli* systems. HIV-derived peptides in the cytosol of human peripheral blood mononuclear cells can be assessed for stability using the stability prediction tool [198]. Proteins with half-lives less than 30 minutes in human embryonic kidney 293T cells are detected by the SprotP server [199]. HLP predicts the half-life of peptides in intestine-like environments [200], whereas PlifePred determines the half-life of both natural and modified peptides in blood [196]. To further assist researchers and accelerate therapeutic peptide design, there remains a need to develop robust in silico models capable of predicting the half-life of both natural and modified peptide sequences.

Determining peptide half-life experimentally is accurate but costly and time-consuming, motivating computational alternatives. Early tools such as ProtLifePred and ProtParam were built for proteins using the N-end rule and perform poorly on peptides in blood ($R \approx 0.05$). The first peptide-specific predictor, PlifePred, was an important advance but was trained on only 261 peptides, relied on classical chemical descriptors without leveraging modern protein language model representations, and required a three-dimensional structure for its structure-based predictions, which is often unavailable for novel peptides. These constraints highlight the need for a peptide half-life predictor trained on a substantially larger and more diverse dataset, combining an expanded feature space with protein language model embeddings and an explicit representation of the modifications that strongly influence peptide stability.

2.5 Summary

The literature review highlights the substantial progress made in computational immunology and protein engineering for vaccine development. Among the available vaccine strategies, protein subunit vaccines are gaining increasing attention because of their inherent advantages, making them particularly suitable for high-risk populations like infants, elders, and immunocompromised

individuals. In addition, protein-based vaccines generally offer improved safety profiles and can be engineered for better stability, thereby reducing dependence on strict cold-chain infrastructure and facilitating their distribution in resource-limited settings. Early in silico safety screening provides an effective approach to evaluate potential antigenic epitopes prior to experimental validation, significantly saving both time and resources. Through rational antigen design and appropriate adjuvant selection, protein vaccines can also be tailored to modulate both humoral and cellular immune responses [201]. As a result, vaccination remains an essential part of public health while evolving toward safer and more precise immunization strategies. Still, there are a number of significant gaps despite these advancements. The following significant limitations are aimed to address in this work:

1. Lack of a unified framework capable of predicting both Linear and Conformational B-cell epitopes.
2. The need for early computational safety screening of vaccine candidates, enabling removal of peptides associated with (i) peptides associated with autoimmune diseases, and (ii) potential allergens, can be filtered out during the design stage.
3. To establish a reliable method for predicting protein melting temperature (T_m) using a non-redundant protein dataset
4. Development of a prediction model for antifreeze proteins based on annotated or experimentally validated data.
5. To build a computational (in silico) predictor to estimate the half-life of both modified and natural peptide sequences.

Addressing these challenges is expected to improve the design of vaccines and therapeutic proteins and enhance their long-term protective efficacy.

C H A P T E R

3

Datasets
&
Methodology

3.1 Overview

The computational tools described in this work address different biological objectives related to subunit vaccine design and therapeutic protein design, including B-cell epitope prediction, safety evaluation, and protein stability assessment. Although these objectives span distinct biological problems, they share a common standard of bioinformatics methodologies. This chapter outlines the general computational framework adopted across all the studies, beginning with data sources, redundancy removal, sequence-derived feature representation, dimension reduction, supervised machine learning, alignment-based and non-alignment-based methods, parameter optimization, evaluation metrics, and software. The overall methodological workflow employed in this thesis is shown in Figure 3.1.

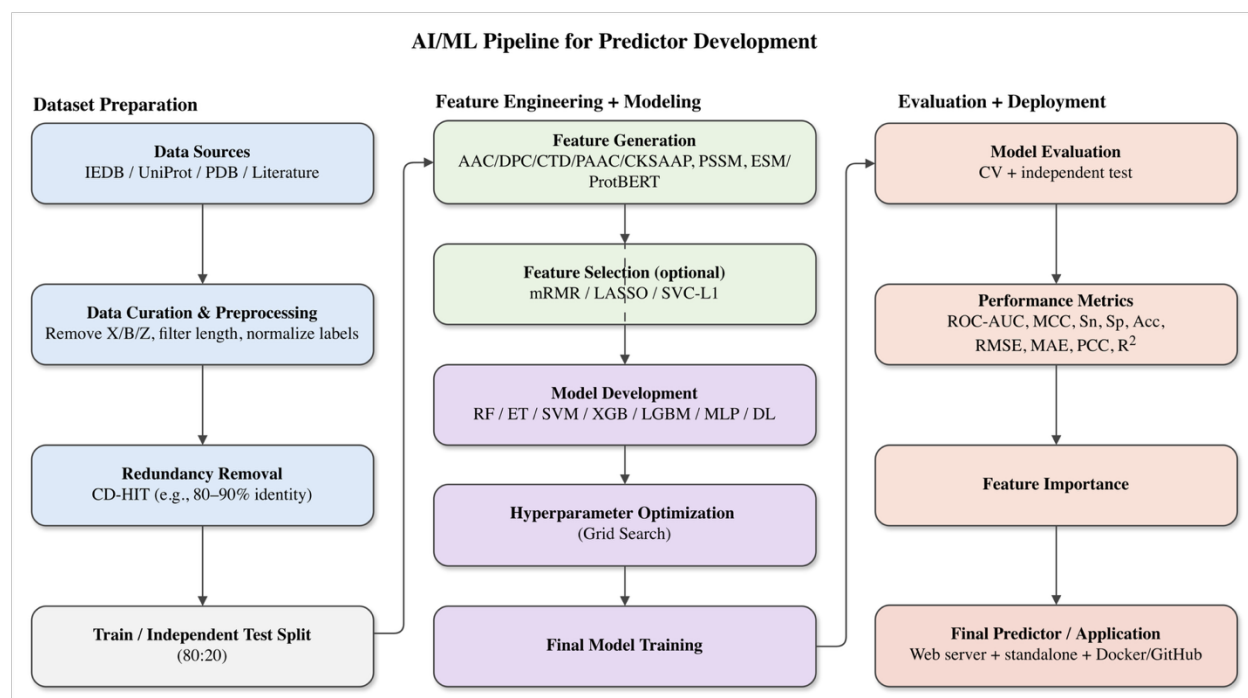


Figure 3.1 General computational workflow employed in this thesis

3.2 Dataset Compilation and Curation

The dataset used in this work was compiled from different, well-established biological databases and curated public resources, selected according to the specific objectives of each work. These datasets encompass experimentally validated B-cell epitopes, full-length protein sequences, disease-associated peptides, allergen and non-allergen peptide collections, as well as protein thermostability datasets. Major publicly available repositories (listed in Table 3.1) such as UniProt (Swiss-Prot), Immune Epitope Database (IEDB), structure-derived annotations from the Protein

Data Bank (PDB), Meltome Atlas, and curated datasets extracted from peer-reviewed literature [73,147] were extensively utilized to ensure broad biological coverage and high data quality.

Table 3.1 *The majorly used databases with description*

Database	Description	Reference
IEDB	Immune Epitope Database is a manually curated data from the literature.	[202]
PDB	A collection of experimentally determined 3D structures of proteins and biological molecules.	[203]
UniProt	A central resource for protein sequences and functional information.	[204]
Meltome	The organismal optimal growth temperatures.	[205]

To maintain the data reliability and minimize noise, preference was given to experimentally validated entries and reviewed entries wherever possible. Standard bioinformatics preprocessing steps were applied uniformly across all datasets, including the removal of duplicate entries, the exclusion of unreviewed, predicted, or ambiguous annotations, and the elimination of sequences containing non-standard or ambiguous amino acids. The length selection criteria were also applied to prediction methods. When experimentally validated negative datasets were limited or unavailable, we carefully generated the negative dataset from curated protein repositories such as Swiss-Prot by following established guidelines to avoid bias and data leakage. Overall, this systematic and standardized dataset compilation strategy ensured the development of robust, unbiased, and reproducible predictive models across diverse applications in subunit vaccine design and therapeutic protein engineering.

3.3 Redundancy Removal and Dataset Balancing

The quantity of available sequencing data is expanding rapidly due to the development of next-generation sequencing technology. In such large-scale datasets, redundancy is common, as highly similar or identical sequences frequently arise from closely related organisms, isoforms, homologs, or repeated experimental sampling. In various analysis methods, removing duplication from such data through clustering may be essential for reducing the computational time, storage space, and noise interference [206]. When using random splitting, this redundancy impacts the evaluation of a machine learning (ML) model's performance, resulting in poor performance on out-of-distribution data and overestimate predictive performance. Redundancy elimination is considered a standard necessity in immunoinformatics and bioinformatics, where the issue is well-known for

protein function prediction. By ensuring that the sequence similarity between samples is greater than a particular threshold, tools such as CD-HIT are utilized to reduce redundancy [207]. The CD-HIT emerged as one of the most widely adopted tools and is regarded as a state-of-the-art method for rapid clustering of large-scale protein and nucleotide datasets. Early versions of CD-HIT were designed for efficient processing of large sequence collections [208], while later generations incorporated an optimized greedy strategy that greatly improves runtime and memory efficiency, enabling redundancy removal even in datasets with hundreds of thousands of sequences [206,207]. CD-HIT determines redundancy based on sequence identity between each candidate sequence and the representative (“centroid”) of an existing cluster: if identity meets or exceeds a user-defined threshold, the sequence is assigned to that cluster; otherwise, it becomes the representative of a new cluster. For example, at a 90% identity threshold, sequences in the same cluster share $\geq 90\%$ identity with their representative, and representatives from different clusters share $< 90\%$ identity with each other. The method proceeds by sorting sequences by descending length, selecting the longest as the first cluster representative, and iteratively comparing subsequent sequences to existing representatives to produce both full clustering assignments and a non-redundant representative set. Threshold selection remains application-dependent; prior evaluations report wide usage spanning 40% to 100% identity, reflecting trade-offs between strict redundancy removal and retention of biological diversity (see Figure 3.2) [209].

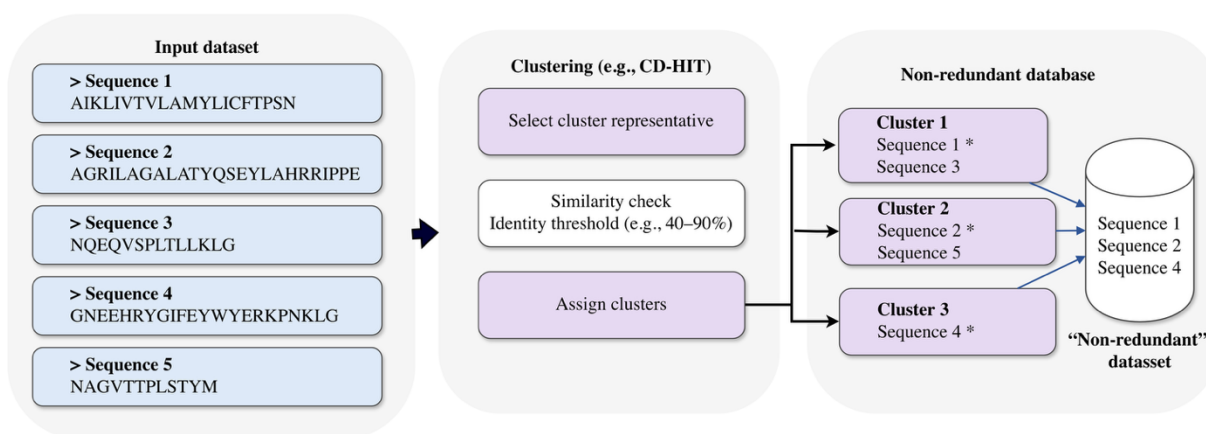


Figure 3.2 Clustering based redundancy removal and representative sequence selection for non-redundant dataset construction

The sequence identity threshold used for redundancy removal varied according to the characteristics of each prediction problem and is summarized in Table 3.2. Following redundancy removal, the resulting non-redundant datasets were used for feature extraction, feature selection, model development, and performance evaluation.

Table 3.2. Summary of redundancy removal employed across different prediction tools

Prediction Tool	Dataset	Redundancy Removal	Rationale
CLBTope	B-cell epitopes	CD-HIT <80%	To remove highly similar epitope sequences while preserving sequence diversity for unbiased model development.
DMPPred	T1DM-associated peptides	Exact duplicate removal	To eliminate duplicate peptide sequences.
AlgPred3	Allergenic and non-allergenic Peptides	CD-HIT 90%	To construct a non-redundant benchmark dataset and improve model generalizability.
PPTStab	Thermostable proteins	CD-HIT 40%	To reduce homologous proteins and develop a robust regression model for melting temperature prediction.
AFProPred	Antifreeze proteins	CD-HIT 40%	To remove redundant protein sequences while maintaining representative AFP diversity.
PlifePred2	Natural and modified peptides	Exact duplicate removal	To eliminate duplicate peptide sequences.

Prior to dataset partitioning, appropriate preprocessing was performed to reduce data redundancy. Depending on the characteristics of the prediction task, either sequence redundancy was removed using CD-HIT at the specified sequence identity threshold or exact duplicate sequences were eliminated. This preprocessing minimized redundancy within the datasets and reduced the likelihood of identical or highly similar sequences influencing model evaluation. The processed datasets were subsequently partitioned into training and testing subsets and used for feature extraction, feature selection, model development, and performance evaluation. The preprocessing strategy adopted for each prediction task was selected according to the characteristics of the respective dataset and the objectives of the study.

3.4 Dataset Partitioning

A rigorous and unbiased evaluation of machine-learning models depends strongly on appropriate dataset partitioning and class balance. Class imbalance is a common issue in biological datasets and may lead to biased learning where models favor the majority class. To address this, wherever classification tasks were performed, datasets were constructed in a balanced manner, ensuring that the number of positive and negative samples was equal (or nearly equal). This strategy provides fair learning conditions and avoids inflated accuracy due to over-representation of one class. Therefore, across all studies presented in this work, we adopted standardized protocols to construct representative training and validation datasets and to minimize evaluation bias. In general, datasets were divided into two components: (i) a training dataset used for feature engineering, model training, and hyperparameter optimization, and (ii) a fully unseen independent validation dataset reserved exclusively for final model assessment. For most prediction tasks, the complete dataset was partitioned into 80% training and 20% independent validation dataset. This splitting strategy is widely used in bioinformatics and provides a practical trade-off between training set size and reliability of independent evaluation. Importantly, the independent dataset was never used during model building or parameter tuning, ensuring an unbiased estimate of real-world predictive performance.

3.5 Compositional Analysis

To examine the distribution of amino acids within each dataset, we calculated the amino acid composition using Equation 1. For this purpose, the composition module of the Pfeature [210] standalone was employed, and the compositions of the positive and negative sets were computed separately for each dataset.

$$CP_x = \frac{NR_x}{TR} \quad [1]$$

Where, CP_x represents the composition of residue x ; NR_x denotes the total number of residues of type x ; and TR indicates the total number of residue in the sequence.

3.6 Positional Analysis

We used web server to create the logos using two-sample logo (TSL) [211] and weblogo [212] in order to understand the position-specific preference of residues. In order to create the logo, it is a prerequisite to fix the peptide length. Therefore, we have taken the minimum length residues from the N-terminal and the minimum length residues from the C-terminal for the purpose of fulfilling

the fixed length criteria. Finally, we have joined both regions to produce a fixed-length peptide [213,214].

General rule

If peptide length = L , and you want terminal length $k=12$:

N-terminus ($N \rightarrow C$) = sequence $[1:k]$

C-terminus ($N \rightarrow C$) = sequence $[L-k+1 : L]$

C-terminus ($C \rightarrow N$) = reverse of C-terminus above

24-mer = N-terminus + reversed-C-terminus

3.7 Feature Extraction

Proteins and other biological sequence data are composed of variable-length amino acid sequences. Therefore, it is important to calculate a certain number of informative features from each sequence so that machine-learning models use them as inputs. In the previous few decades, significant efforts have been made to transform protein amino acid sequences into numerical vectors of consistent length. Such representations have enabled computational methods to address various protein-related problems, including structure prediction [215], function annotation [216], side-chain conformation prediction [217], and enzyme engineering [218]. To facilitate this process, several feature-extraction tools and packages have been developed, including iFeature [219], iFeature Omega [220], iLearnPlus [221], protr [222], Rcpi [223], Pfeature [210], and MathFeature [224,225]. In our work, we employed multiple complementary strategies to represent protein/peptide sequences, including composition-based descriptors, binary profile encodings, evolutionary information, and embeddings generated using protein language models (PLMs).

3.7.1 Sequence composition descriptors

In this study, a number of features were generated from protein sequences using the standalone version of Pfeature [210]. Pfeature is a comprehensive library designed to compute descriptors from protein and peptide sequences. It consists of five main modules for feature calculation: (i) descriptors based on composition, (ii) binary profiles, (iii) evolutionary information, (iv) structure, and (v) pattern. In the present work, we specifically transformed each sequence into a numerical feature vector using the Pfeature software's composition module. We have listed the composition-based features that are commonly used in Table 3.3.

Table 3.3 Sequence-derived descriptor used for protein/peptide characterization

Descriptor	Description	No. of features
SEP	Shannon entropy of protein	1
SOCN	Sequence order coupling number	2
BTC	Bond composition	4
ATC	Atomic composition	5
AAC	Amino acid composition	20
RRI	Residue repeat Information	20
DDOR	Distance distribution of residue	20
SER	Shannon entropy of residues	20
PAAC	Pseudo amino acid composition	21
APAAC	Amphiphilic pseudo amino acid composition	23
SPC	Shannon entropy of physico-chemical properties	25
PCP	Physico-chemical properties composition	30
QSO	Quasi-sequence order	42
CeTD	Composition-enhanced transition distribution	189
CTC	Conjoint triad calculation of the descriptors	343
DPC	Dipeptide composition	400

3.7.2 Binary Profiles Features

An essential step of developing an efficient machine learning model is feature generation, which converts biological sequences into fixed-dimensional numerical vectors that are suitable for model training. Among the widely adopted encoding strategies, binary profile features, sometimes referred to as one-hot encoding, provide a simple yet highly effective representation and have been used successfully for predicting diverse functional properties from multi-omics datasets [226–228]. There are several existing methods that use the binary profile (BP) [79,226,227,229]. In BP encoding, each amino acid residue from the standard alphabet consists of 20 amino acids- Valine (V), Methionine (M), Glycine (G), Alanine (A), Leucine (L), Isoleucine (I), Proline (P), Phenylalanine (F), Tryptophan (W), Tyrosine (Y), Serine (S), Threonine (T), Cysteine (C), Asparagine (N), Glutamine (Q), Aspartic acid (D), Glutamic acid (E), Lysine (K), Arginine (R), and Histidine (H), is represented by a binary vector in which only the position corresponding to that residue is assigned a value of 1, while all other positions are set to 0. In this study, residue was represented using a 21-dimensional vector, where the additional dimension (X) was used for padding/unknown residue handling where applicable. For example, valine (V) is encoded as

(1,0,0,...,0), Methionine (M) as (0,1,0,...,0), whereas all other residues occupy distinct positions in the binary alphabet. Consequently, a peptide segment (pattern) of length n is represented by a $21 \times n$ feature matrix (Figure 3.3), thereby preserving the positional organization of residues [79]. Formally, for a peptide sequence P , a fixed-length segment of length n can be encoded as:

$$BP(n) = [f(p_1), f(p_2) \dots f(p_i)] \quad [2]$$

Where $f(p_i)$ denotes the 21-dimensional one-hot vector for the i -th residue.

The parameter n is treated as a task-specific window-length parameter (tunable during development) and is selected empirically based on the dataset and biological objective. In practice, multiple windowing strategies have been adopted in earlier studies such as N-terminus only, C-terminus only, combined terminal encoding (N + C), Overlapping window patterns [79,226,227,229].

Amino acid	Binary Profile																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	X
Ala	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cys	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Asp	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Glu	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Phe	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gly	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
His	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ile	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Lys	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Leu	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Met	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Asn	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Pro	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Gln	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Arg	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Ser	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Thr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Val	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Trp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Tyr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure 3.3 Binary profile matrix showing the dimension of 21 amino acids

3.7.3 Evolutionary Profile-based Features

An important component of biological sequence analysis is evolutionary conservation, which frequently reflects structural and functional significance. Conserved sequences, which remain similar or identical across evolutionary time, tend to retain essential biological roles. Therefore,

incorporating residue conservation patterns can substantially enhance protein function prediction. In order to preserve rich evolutionary information while converting variable-length protein sequences into fixed-length feature vectors, Position-Specific Scoring Matrix (PSSM) profiles are commonly utilized [230]. We have used the Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) to create PSSM profiles [231]. Each query protein was searched against the Swiss-Prot database [232] to identify the homologous sequences, employing three iterations and an E-value cutoff of 0.001. This approach enabled the extraction of evolutionary profile information of each protein sequence. Following previously reported approaches, evolutionary information was incorporated in the form of a “PSSM-400” composition profile [233], which has been successfully applied in several studies [233–237]. To compute these features, we used the “pssm_composition” module from the POSSUM package [238]. The resulting PSSM-400 feature is a 20×20 -dimensional vector representing the occurrence composition of each of the 20 amino acids corresponding to every amino acid type in the protein sequence [158].

3.7.4 Protein Language Model Embeddings

Recently, protein language models (PLM) have become an effective tool that has shown excellent performance in a variety of tasks, including protein design, interaction prediction, and structure prediction. PLM-derived embeddings have emerged as state-of-the-art for numerous protein sequence-based property-prediction tasks, more often outperforming conventional compositional descriptors [158,239,240]. PLMs are typically based on transformer architectures trained on millions to billions of protein sequences, which enables them to learn contextual representations that capture biochemical properties, structural constraints, and evolutionary signals embedded in primary sequences. A pre-trained PLM generates a representation of a protein sequence of length n in the form of a token-level embedding matrix of size $n \times d$, where d represents the dimensionality of the per-residue (per-token) embedding. Every token-specific embedding records the protein's local and global characteristics in addition to the biochemical data of the amino acid residue. To obtain a single fixed-length protein-level embedding suitable for machine-learning models, these residue-level embeddings are typically aggregated using pooling strategies. The most popular method is “average pooling”, which creates a vector with dimensions of $1 \times d$ by taking the mean of each feature dimension along the protein's length [240,241]. In this work, the embeddings generated by transformer-based protein language models (shown in Table 3.4) were explored for

several prediction tasks, including thermostability prediction, allergens prediction. Overall, PLM embeddings provide a learned and context-aware alternative to classical feature engineering.

Table 3.4 Protein language models and their embeddings dimensions

Protein Language Model	Embeddings vector size
ESM2 (t6)	320
ESM2 (t12)	480
ESM2 (t30)	640
Ankh	768
ProtBert	1024
ProtT5-XL-UniRef50	1024
ProstT5	1024
ProtGPT2	1280
ESM2 (t33)	1280

3.8 Feature Selection Methods

The most important step in machine learning-based prediction models is feature selection, which finds the most informative data points that affect prediction performance [242]. In order to best predict the response variable, it seeks to identify a relevant subset of features from a potentially huge feature space. As an important preprocessing step, feature selection reduces data dimensionality and eliminates redundant or irrelevant information. Consequently, removing non-informative features can improve model accuracy while also enhancing interpretability [243]. Therefore, to reduce dimensionality and choose the more informative features, we used regularization-based embedded methods (Lasso-L1, Elastic Net, SVC-L1), tree-based embedded methods (Boruta), wrapper methods (RFE-SVR), and redundancy-aware ranking (mRMR) as feature selection techniques. Collectively, these complementary strategies enabled a systematic reduction in feature dimensionality, minimized redundancy, and enhanced predictive accuracy and interpretability across the different computational tools developed in this study.

1. **mRMR Method:** The Minimum Redundancy Maximum Relevance (mRMR) method [244] is a widely used feature selection approach that selects features exhibiting high correlation with the target class while maintaining low inter-correlation among themselves.

The method ranks features according to the minimal-redundancy–maximal-relevance criterion, which is computed using mutual information [243].

2. **SVCL1 Method:** The Support Vector Classifier with L1 regularization (SVC-L1) is another feature selection method that identifies relevant features by retaining only those with non-zero coefficients after applying an L1 penalty, thereby reducing data dimensionality [245]. This approach incorporates both a loss function and a regularization term during optimization. The L1 regularization induces sparsity in the model by forcing many feature weights to zero, effectively selecting a subset of informative features while discarding the rest. An important parameter in this method is the regularization parameter “C”, which controls the strength of regularization and consequently influences the number of selected features. A smaller value of “C” enforces stronger regularization and results in fewer selected features. In this study, we used the default value of the “C” parameter (0.01) [246].
3. **LassoL1 Method:** LASSO (L1) regularization has become a popular feature selection technique because it is both simple and efficient, and numerous variations of the original L1-based approach have since been developed [247,248]. L1 regularization, commonly referred to as LASSO [249] is used to remove redundant and irrelevant features by shrinking less informative feature coefficients toward zero. Notably, even a very small regularization parameter (λ_1) value of 0.001 can substantially reduce the feature space, decreasing the number of features by more than a factor of ten [250].
4. **Boruta Method:** Boruta is a wrapper algorithm for all-relevant feature selection that can be used with any classification technique that can produce a variable importance measure (VIM); by default, it makes use of the Random Forest classifier. The technique compares the original attribute’s importance with the importance predicted by random chance, which is calculated using their permuted (shadow) copies, in order to conduct a top-down search for relevant features. Irrelevant features are gradually removed until the statistical test stabilizes [251]. Boruta is one of the feature selection methods that received the most attention due to its capacity to handle high-dimensional datasets and detect statistically significant features. Boruta uses statistical testing with shadow features, a Random Forest-based method, to assess each feature’s relevance. By comparing the significance of each

original feature with that of its corresponding shadow feature, it separates relevant from irrelevant attributes [242].

5. **Elastic Net Method:** Elastic Net (EN) [247] is a generalized regularized regression model that combines both L1(Lasso) and L2 (Ridge) norm penalties within a single objective function [247]. Owing to this hybrid regularization, EN is particularly suitable for high-dimensional feature selection problems. Unlike LASSO, which may select only one feature from a group of highly correlated variables, Elastic Net can retain groups of correlated features. Consequently, it effectively handles multicollinearity and overcomes limitations associated with using LASSO or Ridge regression individually in high-dimensional datasets [247].
6. **RFE SVM Method:** Support Vector Machines have long been recognized as effective tools for identifying informative patterns in data [252]. Support Vector Machine–Recursive Feature Elimination (SVM-RFE), one of the most popular wrapper-based feature selection techniques for SVMs, was proposed by Guyon et al. [253]. This method is essentially a backward elimination procedure in which features are iteratively removed based on their contribution to the classifier [254]. In this method, feature ranking is determined by the magnitude of the SVM weight vector, and features with the smallest weights are eliminated at each iteration. However, the features ranked highest (i.e., eliminated last) are not necessarily individually the most relevant; rather, they are the most informative within the context of the selected feature subset used by the model [253,254].

The selection of feature descriptors was guided by their predictive performance rather than by a priori assumptions regarding their biological relevance. Initially, machine learning models were developed using each descriptor independently (e.g., AAC, DPC, PAAC, APAAC, CTD, and other sequence-derived descriptors), and their performance was evaluated on the validation dataset. The predictive ability of each descriptor was compared using appropriate evaluation metrics, such as AUROC for classification tasks and PCC/ R^2 for regression tasks. Subsequently, all composition-based descriptors were combined into a comprehensive feature set (ALL_COMP), and feature selection algorithms were applied to identify the most informative subset of features while reducing redundancy. The performance of the selected hybrid features was then compared with that of the best-performing individual descriptors. In cases where an individual descriptor achieved comparable or superior predictive performance, it was preferred because of its lower

computational complexity, improved interpretability, and reduced risk of overfitting. This systematic evaluation ensured that the final descriptor set represented the optimal balance between predictive performance, computational efficiency, and biological interpretability. Furthermore, to improve model interpretability, SHAP (SHapley Additive exPlanations) analysis was performed on the final optimized models to quantify the contribution of individual features and to facilitate the biological interpretation of the selected descriptors.

3.9 Model Development

Machine learning (ML) algorithms are strong computational techniques that may transform complicated data into reliable predictions. This study used a variety of machine learning algorithms to create a reliable predictor that can provide accurate and efficient results. We utilized both regression and classification models as part of our supervised learning techniques. For the given problem, each algorithm has a set of hyperparameters that need to be carefully tuned to find the best configuration and provide the best predictive performance [19].

3.9.1 Traditional Machine Learning Algorithms

For the development of predictive models, this study used a number of machine learning algorithms, including:

1. **Classification Algorithms:** This study employed the following algorithms: Multilayer Perceptron Classifier (MLP), Decision Tree (DT), Logistic Regression (LR), Gaussian Naïve Bayes (GNB), Random Forest (RF), k-Nearest Neighbors (KNN), Extra Trees (ET), Extreme Gradient Boosting (XGB), Light Gradient Boosting Machine (LGBM), and Support Vector Classification (SVC). The scikit-learn Python library was used to implement these classification algorithms.
2. **Regression Models:** regression analysis, multiple algorithms were employed, including Decision Tree Regressor (DTR), Ridge Regression (Ridge), Gradient Boosting Regressor (GBR), Random Forest Regressor (RFR), Elastic Net Regressor (EN), Support Vector Regressor (SVR), Linear Regression (LR), Extreme Gradient Boosting (XGBoost), AdaBoost Regressor (AD), LASSO Regression (LASSO), Multilayer Perceptron Regressor (MLP), Extra Trees Regressor (Extremely Randomized Trees), Nu-Support Vector Regression (NuSVR), Light Gradient Boosting Machine (LGBM Regressor), and Categorical Boosting (CatBoost; gradient boosting on decision trees).

3.9.2 Large Language models

Protein language models (PLMs) based on natural language processing (NLP) have recently advanced and shown excellent performance in a variety of protein analysis tasks. We utilized the well-known LLMs ESM-2, ProtBERT, and PeptideBERT. Additionally, we only use ProtGPT2, ProtT5-XL-Uniref50, ProsT5, and Ankh for embedding extraction [255]. We adjusted the hyperparameters of these models to make them more suitable for our task.

1. **ESM-2:** It was developed by Meta AI, represents one of the most advanced protein language modeling frameworks based on the transformer architecture. This study used several different versions of the ESM-2 model, such as esm2 (t33, t30, t12, and t6). These models have demonstrated excellent performance in tasks like protein structure prediction and variant effect analysis, and they have been pre-trained on large-scale protein sequence datasets. The ESM family's attention-based architecture and extensive pretraining efficiently capture the evolutionary relationships and functional properties of proteins [256].
2. **ProtBert:** ProtBERT [257] is an important protein language pretraining model derived from the BERT architecture. It is trained using the masked language modeling (MLM) objective on large-scale protein sequence datasets such as UniRef100 [258] and the Big Fantastic Database (BFD) [259]. Owing to its large-scale pretraining, ProtBERT has demonstrated strong performance across a wide range of protein analysis tasks [260].
3. **PeptideBERT:** a transformer-based model built on the ProtBERT pretrained architecture, consisting of 12 attention heads and 12 hidden layers [261]. Since ProtBERT itself is a protein-specific adaptation of the BERT framework, pre-trained on a large corpus of protein sequences, PeptideBERT benefits from contextual representations learned from extensive sequence data [260,262].
4. **Ankh:** Google's TPU-v4 infrastructure was used to train the general-purpose protein language model Ankh. Comparing the model to larger protein language models, it is computationally efficient, requiring less than 10% of the parameters for pretraining, less than 7% for inference, and less than 30% of the embedding dimensionality [263].
5. **ProtGPT2:** A protein language model based on the GPT-2 architecture, is a decoder-only transformer. A causal language modeling objective is used to pre-train it on the UniRef50

protein sequence database (version 2021-04). The model, which has 36 transformer layers, can generate representations of protein sequences and learn sequence patterns [264].

6. **ProtTrans**: A text-to-text Transfer Transformer (T5) architecture modified for protein sequences, the T5-3B model, is the source of ProtT5-XL-UniRef50. It is pre-trained on a sizable collection of unprocessed protein sequences in a self-supervised fashion, with labels and inputs produced automatically without the need for human annotation. The BART-like MLM denoising objective used by ProtT5-XL-UniRef50 is specifically designed for protein sequences, in contrast to the span-denoising objective used by the original T5-3B model [255].
7. **ProsT5**: A bilingual protein language model called ProsT5, also known as the protein structure–sequence T5 model, is intended to translate between structural representations and protein sequences. In order to create it, ProtT5-XL-UniRef50 was refined using a dataset of roughly 17 million proteins and superior 3D structural predictions from AlphaFoldDB [265].

3.10 Alignment-Based Approaches

3.10.1 Similarity search approach

Similarity search is a widely adopted strategy for functional annotation of proteins and peptides, as homologous sequences often share related biological roles. BLAST-based methods have therefore been extensively used in computational biology to annotate protein sequences [266]. In order to classify query sequences by searching them against a curated reference database, we implemented a similarity search module in this study using pBLAST (BLAST+ version 2.7.1) [267] with the default parameters (scoring matrix = PAM30; word size = 2; window size = 15; gap open = 9; gap extend = 1; and threshold = 16) [268]. A training dataset was used to build a target BLAST database for performing a blastp search, and the query sequences (typically from the independent/test dataset) were searched against this database at different E-values cutoffs. Each query sequence has been assigned to a class based on the annotation of its aligned hit(s) on the basis of top-hit approach.

3.10.2 Motif Identification

To finding the conserved motifs in biological sequences is a crucial method for identifying shared functions [269]. Motifs are repetitive arrangements of amino acids (in proteins) or nucleotides (in

DNA/RNA) that may correspond to biologically meaningful features (such as functionally conserved sites). In our work, motif-based analysis was employed as an additional feature approach to identify discriminative sequence patterns associated with positive and negative classes.

3.10.2.1 MERCI Motif Analysis

MERCI is a tool that looks for motifs based on their physicochemical properties. It was used to extract motifs enriched in the positive dataset by comparing the sequences from the positive and negative sets in order to find degenerate motifs based on a given classification of amino acids [269]. A number of MERCI parameters were explored to identify both exclusive and inclusive motifs. The negative dataset's maximal frequency (fn) was changed from its default value. With the default setting (fn = 0), MERCI identifies exclusive motifs, i.e., motifs that do not occur in negative sequences. To obtain the inclusive motifs, the value is set to fn = 8. Furthermore, motif discovery was performed across multiple configurations, including no gap motifs, motifs with gap = 1, gap = 2, and using classification Koolman-Rohm [269]. Following motif identification, unique proteins sequences containing the derived motifs were mapped to determine motif coverage and estimate how many proteins/peptides in the dataset could be annotated using these motif signatures.

3.10.2.2 MEME/MAST Motif Analysis

Additionally, we used the MEME suite (version 5.0.5) for systematic motif discovery and motif scanning. The two main modules that make up MEME/MAST are MEME [270], which finds motifs in closely related sequences, and MAST [269], which looks for matches to a set of motifs (from MEME's output). The likelihood of every potential residue at every motif position is represented by position-dependent letter-probability matrices, which are how MEME represents motifs. It takes the training set (FASTA sequences) as input and produces motif models using statistical modeling techniques. Subsequently, MAST uses the discovered motif matrices as input and hitting the query file (test set) on the MEME matrix to search for the match for the motifs. Subsequently, MAST uses the discovered motif matrices as input and searches for motif matches within independent/test sequences, enabling motif-based classification and pattern detection at the dataset scale. To ensure robustness and control noise, multiple motif matrices were generated (corresponding to multiple training datasets where applicable), and motif scanning was performed

using E-value cutoffs during MAST analysis to control statistical stringency. Additionally, e-value cutoffs were considered when performing MAST analysis [271].

3.11 Model Training and Hyperparameter Optimization

We performed hyperparameter tuning over predefined parameter ranges using a grid search with the scikit-learn library in order to maximize the classifier's performance. For each algorithm, important hyperparameters (listed in Table 3.5) were systematically tuned using grid search in order to obtain optimal predictive performance. For regression baselines (e.g., Linear Regression), default model settings were used as no tunable hyperparameters were required. The best models were selected based on performance on the validation set. We evaluated multiple machine learning algorithms and optimized their hyperparameters using the training dataset. For this purpose, GridSearchCV was employed to identify the best-performing configuration for each classifier. The models were optimized by maximizing the area under the receiver operating characteristic curve (AUROC).

Table 3.5 ML algorithms used in model development with grid search key hyperparameters

Algorithm	Library	Hyperparameters
A. Classification Algorithms		
Decision Tree	scikit-learn	criterion, max_depth, max_features, min_samples_leaf, min_samples_split
Random Forest	scikit-learn	bootstrap, criterion, max_depth, max_features, min_samples_leaf, min_samples_split, n_estimators
Extra Trees	scikit-learn	criterion, max_depth, max_features, min_samples_leaf, min_samples_split, n_estimators
Logistic Regression	scikit-learn	C, class_weight, max_iter, penalty, solver
Gaussian Naive Bayes	scikit-learn	Default parameters
K-Nearest Neighbors	scikit-learn	metric, n_neighbors, weights
Support Vector Machine	scikit-learn	C, class_weight, degree, gamma, kernel
Multi-Layer Perceptron	scikit-learn	activation, alpha, hidden_layer_sizes, learning_rate_init, max_iter, solver
Extreme Gradient Boosting	xgboost	colsample_bytree, gamma, learning_rate, max_depth, min_child_weight, n_estimators, reg_lambda, subsample
Light Gradient Boosting	lightgbm	bagging_fraction, feature_fraction, lambda_l1, lambda_l2, learning_rate, max_depth, n_estimators, num_leaves

B. Regression Algorithms		
Decision Tree Regressor	scikit-learn	criterion, max_depth, max_leaf_nodes, min_samples_leaf, min_samples_split
Random Forest Regressor	scikit-learn	bootstrap, max_depth, max_features, min_samples_leaf, min_samples_split, n_estimators
Extra Trees Regressor	scikit-learn	max_depth, max_features, min_samples_leaf, min_samples_split, n_estimators
Gradient Boosting Regressor	scikit-learn	learning_rate, loss, max_depth, n_estimators, subsample
Adaboost Regressor	scikit-learn	learning_rate, loss, n_estimators
Linear Regression	scikit-learn	Default parameters
Ridge Regression	scikit-learn	alpha, max_iter, solver
Lasso Regression	scikit-learn	alpha, max_iter, selection
Elastic Net Regressor	scikit-learn	alpha, l1_ratio, max_iter
Support Vector Regression	scikit-learn	C, epsilon, gamma, kernel
Nu Support Vector Regression	scikit-learn	Default parameters
Xgboost Regressor	xgboost	colsample_bytree, learning_rate, max_depth, n_estimators, objective, reg_lambda, subsample
Lightgbm Regressor	lightgbm	bagging_fraction, feature_fraction, lambda_l1, lambda_l2, learning_rate, max_depth, n_estimators, num_leaves
Catboost Regressor	catboost	depth, iterations, l2_leaf_reg, learning_rate, loss_function
Multi-Layer Perceptron Regressor	scikit-learn	activation, alpha, hidden_layer_sizes, learning_rate_init, max_iter, solver

3.12 Model Evaluation

We used k-fold cross-validation alongside an independent validation dataset to train, test, and evaluate the prediction models in an accurate and unbiased manner, following standard evaluation protocols reported in previous studies [272,273]. Initially, the complete dataset was divided into 80% training data and 20% independent validation dataset. The validation dataset was kept completely aside - it remained unseen during model training and hyperparameter optimization, and it was not included in the cross-validation. The training dataset (80%) was then randomly divided into k subsets (usually $k = 5$ or $k = 10$, depending on dataset size) and subjected to k-fold cross-validation. The model was trained on k-1 folds and tested on the remaining fold in each cycle. Until each fold was used as the test set once, this procedure was repeated (Figure 3.4). The final cross-validation performance was reported as the average score across all folds [266]. Finally, the

best-performing model was evaluated on the independent validation dataset (20%) to estimate its generalization capability.

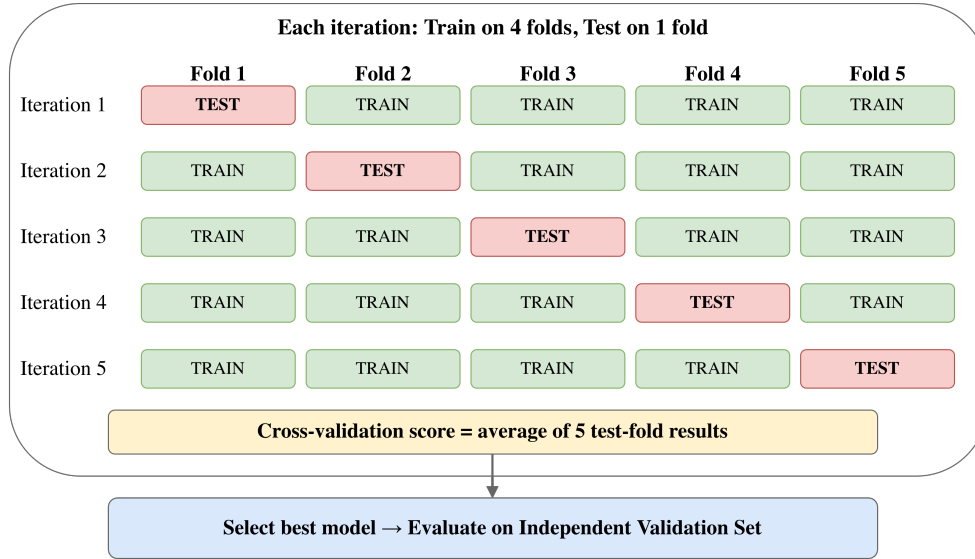


Figure 3.4 The diagrammatic representation of 5-fold cross-validation ($k=5$)

3.13 Performance Evaluation Metrics

3.13.1 Classification Metrics

A combination of threshold-dependent and threshold-independent performance measures were used to assess each classification model created in this study. The confusion matrix summarizes classification outcomes in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), which are used to compute threshold-dependent performance measures. Common evaluation metrics derived from these values include sensitivity (Sens), specificity (Spec), accuracy (Acc), precision, recall, F1-score, and the Matthews correlation coefficient (MCC). These metrics are essential for determining model accuracy as well as understanding the types of errors made during classification [274]. Additionally, we used the area under the receiver operating characteristics curve (AUROC), a threshold-independent metric. The predicted probabilities (or decision scores) produced by the classification models were used to calculate the AUROC. Plotting the true positive rate (sensitivity) against the false positive rate ($1 - \text{specificity}$) across a range of classification thresholds allowed for the creation of the ROC curve [213]. The Following measurements are calculated using Equations (3) - (9):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad [3]$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad [4]$$

$$Specificity = \frac{TN}{TN + FP} \quad [5]$$

$$Sensitivity = \frac{TP}{TP + FN} \quad [6]$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad [7]$$

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} \quad [8]$$

$$Youden's\ Index = Sensitivity + Specificity - 1 \quad [9]$$

Where TP (true positive), TN (true negative), FP (false positive), and FN (false negative)

3.13.2 Regression Metrics

The predictive effectiveness of the regression models was assessed using standard evaluation metrics. By measuring explained variance, prediction error magnitude, and the correlation between observed and predicted values, these metrics offer additional information about model quality. The Pearson Correlation Coefficient (PCC), Coefficient of Determination (R2), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) were specifically utilized. These evaluation metrics were used to choose the top-performing model for further research.

1. **Coefficient of Determination:** The proportion of the dependent variable's variance that can be explained by the independent variables is known as the coefficient of determination. It shows how well that the regression predictions correlate with the actual data. A better fit between the regression model and the data is indicated by an R2 value nearer 1.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i (y_i - x_i)^2}{\sum_i (y_i - \bar{y})^2} \quad [10]$$

SS_{tot} is the total sum of squares, and SS_{res} is the sum of squares of residuals. y_i is the true value and x_i is the prediction.

- 2. Mean squared error:** It is the average of the squared deviations between the actual and predicted values. It measures the regression model's overall prediction error by calculating the average squared deviation between the predicted and observed values. A model that performs better has a lower MSE.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2 \quad [11]$$

Where y_i denotes actual value, x_i is the predicted value, n is the total number of samples

- 3. Mean average error:** The mean absolute error calculates the average size of the discrepancies between the expected and actual values. It does not take into account the direction of the error when calculating the mean of the absolute differences between the expected and actual values. Better predictive accuracy of the regression model is indicated by a lower MAE.

$$MAE = \frac{\sum_{i=1}^N |y_i - x_i|}{N} \quad [12]$$

where y_i is the prediction, and x_i is the true value.

- 4. Root mean squared error:** The root mean squared error calculates the standard deviation of the prediction errors and quantifies the variations between the values that were expected and those that were observed. It gives an error estimate in the same units as the target variable and is the square root of the mean squared error. A model that performs better has a lower RMSE value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_i - x_i)^2}{N}} \quad [13]$$

The y_i is the predicted value, and x_i is the experimental value

- 5. Pearson correlation coefficient (PCC):** The linear relationship between the observed and predicted values is measured by the Pearson Correlation Coefficient, which also indicates its direction. The covariance of the two variables is divided by the product of their standard deviations to produce a normalized covariance. PCC values range from -1 to 1, where 0 denotes no linear correlation, -1 denotes a perfect negative correlation, and 1 denotes a perfect positive correlation.

$$PCC = \frac{cov(X, Y)}{\sigma_X \sigma_Y} = \frac{(E[(X - \mu_X)(Y - \mu_Y)])}{\sigma_X \sigma_Y} \quad [14]$$

Here, σ_X is the standard deviation of X , σ_Y is the standard deviation of Y , μ_X is the mean of X , μ_Y is the mean of Y , and E is the expectation [275].

3.14 Feature Importance Analysis Using SHAP

To improve the interpretability of the developed machine learning models, feature importance analysis was performed using SHAP. It is a game theory-based explainable artificial intelligence (XAI) method that assigns an importance value (Shapley value) to each feature based on its contribution to the model's prediction. Unlike conventional feature importance measures, SHAP provides both global and local explanations by quantifying the contribution of each feature to individual predictions while satisfying desirable theoretical properties such as local accuracy, consistency, and missingness. The method is model-agnostic and can also exploit model-specific algorithms for tree-based ensemble methods, making it particularly suitable for interpreting complex machine learning models. The mathematical foundation of SHAP is based on Shapley values from cooperative game theory, where the contribution of a feature is estimated by averaging its marginal contribution across all possible subsets of features [276,277].

For each prediction model developed in this study, SHAP analysis was performed on the optimized tree-based classifier or regressor using the validation dataset. The mean absolute SHAP value (mean |SHAP|) was calculated for each feature to quantify its overall contribution to model predictions. Features with higher mean absolute SHAP values were considered more influential in the prediction process. Additionally, the top-ranked features were visualized using SHAP summary plots and feature importance bar plots, enabling the identification of the most informative sequence-derived descriptors for each prediction task. These analyses provide a transparent interpretation of the trained models and facilitate the biological interpretation of the sequence features underlying different peptide and protein prediction problems.

3.15 Web server Implementation and software

In order to maximize our research's practical application, we have developed user-friendly web servers for each method: CLBtope, DMPPred, PPTStab, AFProPred, AlgPred3, and PlifePred2 which include the best-performing models from these studies. These web servers are specially designed to facilitate wide accessibility, high-throughput usability, and user-friendliness for users

across the world. The frontend and backend of each are developed using HTML5, CSS, Java, and PHP scripts, ensuring a responsive design that is compatible with almost all devices, including desktops, tablets, and mobile phones. There are three main modules on each web server that can serve the needs of a wide range of researchers:

- (i) **Predict Module:** This module allows users to input a sequence and provide predictions based on the incorporated trained ML models. In this study, this module helps in classifying sequences as B-cell epitope/non-epitope (CLBTope), T1DM-associated/non-associated (DMPPred), AFPs/non-AFPs (AFProPred), T_m of proteins (PPTStab), and half-life of peptides (PlifePred2).
- (ii) **Design Module:** This module mutates one residue at a time for every position in the sequence, assisting in the generation of every potential variation of a protein or peptide along with its predicted value. It assists in designing the best proteins or peptides by determining which mutations in the input sequence are most relevant to the desired interest.
- (iii) **Protein Scan Module:** With the help of the “Protein Scan module”, users can find functional regions in proteins, such as B-cell epitope regions, T1DM-related regions, or thermostable and antifreeze regions of interest, by scanning a region of interest in an amino acid sequence that has been submitted.

Additionally, we provide some other modules such as “BLAST Scan” and “Motif Scan” in CLBTope, DMPPred, and AlgPred3. The query sequence can be compared to a database of known peptides using the BLAST Scan module. On the basis of database hits or matches, the query sequence is predicted. On the other hand, users can map or scan the motif they want in the query sequence using the Motif Scan module. We extracted motifs from peptides using MERCI software. Additionally, we worked to make sure that our tools are both practically deployable and scientifically reliable, which contributes to filling the gaps in protein and peptide therapies. Therefore, with the web interface, we also provide the following services:

- For large-scale usage standalone packages is available;
- Pip packages
- Open source code on GitHub

C H A P T E R

4

Prediction of Linear and Conformational Epitopes

4.1 Introduction

B cells are essential to the immune system because they generate antibodies that defend the body from harmful infections. Hence, B-cell epitopes are crucial for immunological uses such as the development of therapeutic antibodies, vaccines based on epitopes, and diagnosis of diseases [20,21]. B-cell epitopes are generally divided into two groups: linear (continuous) epitopes and conformational (discontinuous) epitopes, based on their structural characteristics and interactions with antibodies. Where conformational epitopes are made up of residues that are far apart in the primary sequence but come into close spatial proximity after protein folding, linear epitopes are made up of a continuous stretch of amino acid residues involved in antibody binding [278]. Historically, both linear and conformational epitopes have been identified through experimental methods like Western blotting, NMR, and X-ray crystallography. Even though these experimental methods are still the best, they are expensive, time-consuming, and labor-intensive (see Figure 2.2). On the other hand, computational methods are simple to create and are cost-effective in terms of both time and money. Propensity scales have previously been used to predict BCEs, which are simple to use and understand. Several studies showed that these methods have limited precision and power. Hence, to overcome these limitations, ML-based approaches have been put forward. Many *in silico* techniques for predicting BCEs have been developed over the past four decades [20,279]. Conformational, linear, and linearized conformational epitope prediction approaches are the general categories into which BCE prediction techniques can be divided. A conformational BCE is a region on a protein structure's surface where residues from different parts of the primary sequence converge in three dimensions and are recognized by B-cell receptors. To predict conformational BCEs, a number of structure-based computational techniques have been developed, such as CEP, DiscoTope, SEPPA, Epitope3D, and SEMA (see Table 2.4). Later, machine learning-based methods like ABCpred, BepiPred, and BCPREDS were developed to enhance prediction performance (refer to Table 2.2). Many of these early predictors were trained on relatively small datasets, primarily from HIVDB [280,281], BCIPEP [282], and AntiJen 2.0 [283]. The majority of computational tools available currently are made to predict BCEs in an antigen either linearly or conformationally. CLBTope, a unified computational framework that can predict both linear and conformational B-cell epitopes using only primary sequence information, was proposed as an approach to this limitation.

4.2 Overview of the study

In this work, we created a single computational framework that accurately predicts both linear and conformational BCEs. The dataset utilized in this study included linearized conformational BCEs and linear BCEs that had been experimentally verified. The proposed approach's overall workflow is depicted in Figure 4.1. The first step involved building machine learning-based models with composition-based features that were obtained from both epitope and non-epitope sequences. After that, BLAST was used to implement an alignment-based strategy in which a query peptide was categorized as an epitope if it demonstrated significant levels of sequence similarity with known BCEs. Although it was noted that each strategy had drawbacks of its own: alignment-based techniques rely significantly on sequence similarity, whereas alignment-free machine learning models could fail in certain difficult circumstances. A hybrid prediction model was created by combining alignment-based and alignment-free techniques in order to get rid of these drawbacks. Additionally, a web server (<https://webs.iitd.edu.in/raghava/clbtope/>) and a standalone software package were created to make use by the scientific community.

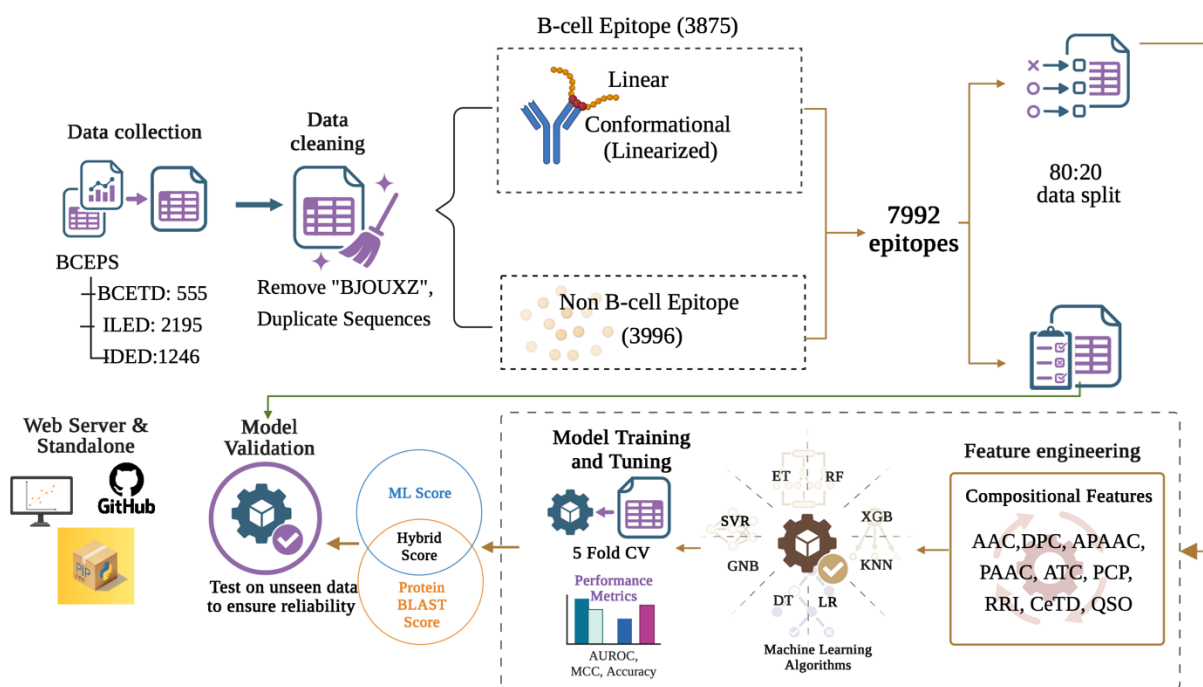


Figure 4.1 Overview of CLBTope workflow for prediction of B-cell epitopes

4.3 Materials and Methods

4.3.1 Dataset Acquisition

The dataset construction followed the standardized protocol described in Chapter 3. Most of the datasets used in this study were obtained from Ras-Carmona et al. [73], including BCETD (linearized conformational B cell epitopes), ILED (IEDB Linear Epitope dataset), and IDED (IEDB Discontinuous Epitope Dataset).

- (i) BCETD is a non-redundant dataset consisting of 555 linearized conformational BCE and 555 non-BCE derived from antibody-antigen structural complexes downloaded from the abYbank/AbDb database. The non-BCEs correspond to antigen regions that are not recognized by antibodies within the same complexes.
- (ii) ILED comprises 2,195 linear BCE and 2,195 non-BCE extracted from the IEDB.
- (iii) IDED is a conformational dataset obtained from IEDB and contains 1,246 discontinuous BCE along with 1,246 corresponding non- BCE.

The three benchmark datasets were merged to create a comprehensive dataset containing 7,992 peptide sequences (3,996 BCE and 3,996 non- BCE). The nearly balanced number of positive and negative samples was retained to avoid class imbalance during model development and to enable unbiased model training and performance evaluation. During dataset preparation, peptide length distribution was also considered to minimize sequence-length bias. To improve dataset quality, sequences containing non-natural amino acids (B, J, O, U, X, and Z) and duplicate entries were removed. After this preprocessing step, the final dataset consisted of 7,871 sequences, including 3,875 BCE and 3,996 non- BCE. Thus, the curated dataset contained both major classes of B-cell epitopes, namely linear epitopes and linearized conformational epitopes. The dataset was subsequently divided into 80% training and 20% independent validation datasets, as described in *section 3.4.1*.

4.3.2 Feature Extraction and Representation

In this study, ten different types of features were computed, including AAC, DPC, ATC, PCP, RRI, SER, CeTD, PAAC, APAAC, and QSO. As summarized in Table 3.3, these descriptors were generated using the composition-based feature module of “Pfeature”, following the protocol described in *Section 3.7.1*, resulting in a 780-dimensional numerical feature vector for each sequence.

4.3.3 Feature selection

Pfeature generates a high-dimensional feature vector, making it challenging to identify meaningful relationships among the extracted descriptors. Therefore, the main goal of feature selection is to find the best subset of informative features from the original feature space. This lowers the chance of overfitting and enhances model performance by making learning easier [232]. The minimum Redundancy Maximum Relevance (mRMR) approach, which is explained in *section 3.8*, was used in this study to choose the features that were most relevant for model development.

4.3.4 Model Development and Evaluation

Machine learning algorithms are effective computational techniques capable of transforming complex data into reliable predictive decisions. In this study, several supervised classifiers-RF, ET, DT, LR, KNN, GNB, and XGB (*explained in Section 3.9.1*) were trained using the training dataset (80% of the total data) with 5-fold cross-validation (*Section 3.12*). Each algorithm contains a distinct set of hyperparameters that were tuned to identify the optimal configuration and achieve the best predictive performance. The final evaluation of the selected models was performed on an independent validation dataset comprising the remaining 20% of the data. Model performance was assessed using AUROC, accuracy, sensitivity, specificity, and MCC, as described in *Section 3.13.1*.

4.3.5 Sequence Similarity Search using BLAST

A commonly utilized similarity search-based method called BLAST was used for protein/peptide annotation. The query is annotated according to the extent to which it resembles the matched sequences after query peptide sequences are aligned against sequences that have known functional annotations. In the present study, blastp (*described in Section 3.10.1*) was used to classify peptides as B-cell epitopes or non-B-cell epitopes [266]. The training dataset was used to create a BLAST-formatted database, which was then used to search the query sequences (test set) using various E-value thresholds between $1e-6$ and $1e+3$. On the basis of top hits, the query sequence was identified as positive if the top hit corresponded to a BCE and as negative if the top hit corresponded to a non-BCE.

4.3.6 Motif Derived features

A motif is a recurring pattern of sequence found in a collection of proteins or peptides. Comparing the motifs in the positive sequences with those in the negative dataset allows for the identification of class-specific motifs that are used to characterize the positive dataset. Based on Koolman and Rohm's proposed amino acid grouping scheme, we used the MERCI motif classification method (described in *Section 3.10.2.1*) in the present study. Aliphatic (A, G, I, L, V), sulfur-containing (C, M), aromatic (F, Y, W), neutral (S, T, N, Q), acidic (D, E), basic (R, H, K), and cyclic (P) are the seven non-overlapping physicochemical groups into which this scheme divides amino acids [269].

4.4 Results

4.4.1 Compositional analysis

To calculate the composition of amino acid we used the *Equation 1* mentioned in *Section 3.5*, which measures the frequency of each residue within a peptide sequence. This analysis facilitates discrimination between BCE and non-BCE. In the present study, the mean compositions of BCE, non-BCE, and the overall proteome were determined (Figure 4.2). The general proteome corresponds to the amino acid distribution obtained from the Swiss-Prot database. The results showed that the positive (BCE) dataset had higher concentrations of Lys (K), Asp (D), Glu (E), Asn (N), Arg (R), Pro (P), and Gln (Q). On the other hand, the negative (non-BCE) dataset had higher levels of Leu (L), Val (V), Ile (I), Phe (F), Cys (C), and Ala (A).

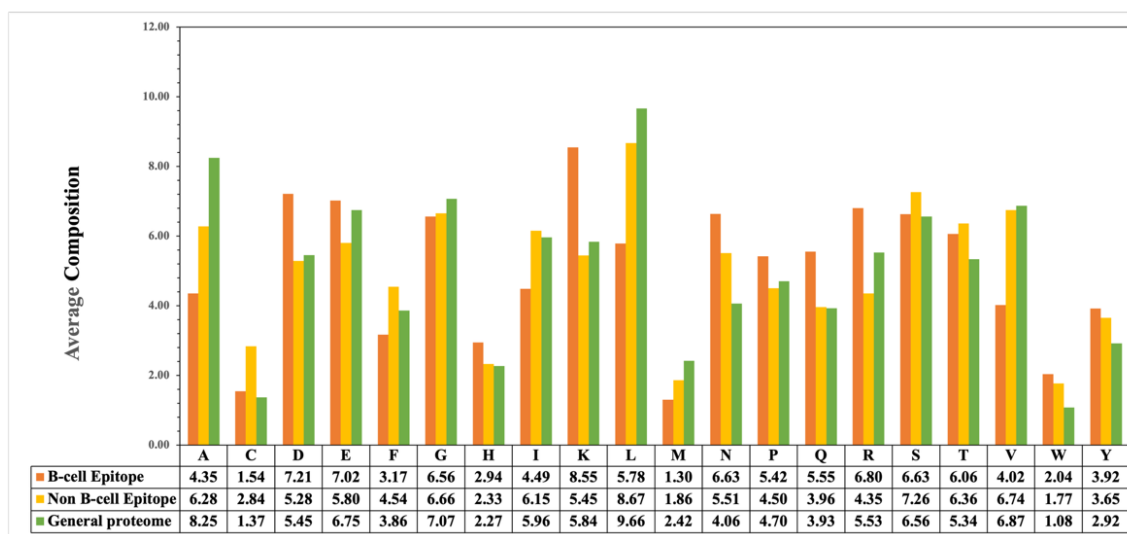


Figure 4.2 Average compositional analysis of B-cell epitopes, non B-cell epitopes, and general proteome (adopted from [19])

4.4.2 Positional Specific analysis

Sequence logos were generated using the web-based application WebLogo (described in *Section 3.6*). In this study, sequence logo analysis was performed to examine residue preferences at specific positions within BCEs. The hydrophilic residue glutamic acid (E) is very abundant and conserved in the majority of positions, as shown in Figure 4.3. This observation is consistent with the fact that epitopes typically reside on the protein surface, where they interact with antibodies or immune receptors. Hydrophilic amino acids are generally enriched on protein surfaces, thereby increasing the likelihood of antibody recognition and binding. Furthermore, it was found that lysine (K) showed a strong preference at position 18, while serine (S) was more conserved at positions 3, 6, 7, and 21.

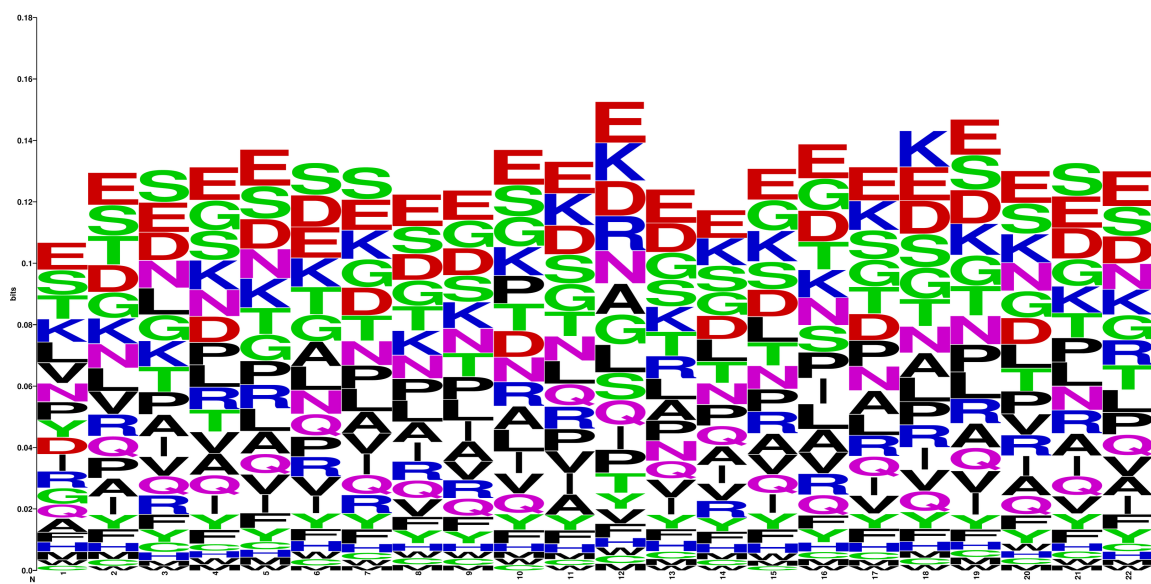


Figure 4.3 Sequence logo representation of BCEs. The height of each amino acid indicates its relative frequency at that position, and the overall stack height represents positional conservation (adopted from [19])

4.4.3 Performance on Individual Descriptors

Various machine learning algorithms, including ET, DT, KNN, LR, XGB, RF, and GNB, were employed to develop the prediction models. One of the most important steps in creating a reliable predictive model is to select the best classifier. Initially, prediction models were constructed using ten composition-based feature sets extracted with the Pfeature tool. Among all classifiers, the RF algorithm consistently demonstrated superior performance across different feature descriptors, achieving AUROC values above 0.72. During 5-fold CV, the best performance was obtained using

the DPC-400 feature set, where the RF-based model achieved the highest AUROC of 0.802 on the validation dataset. This performance was notably better than that obtained with the other descriptors, as illustrated in Figure 4.4.

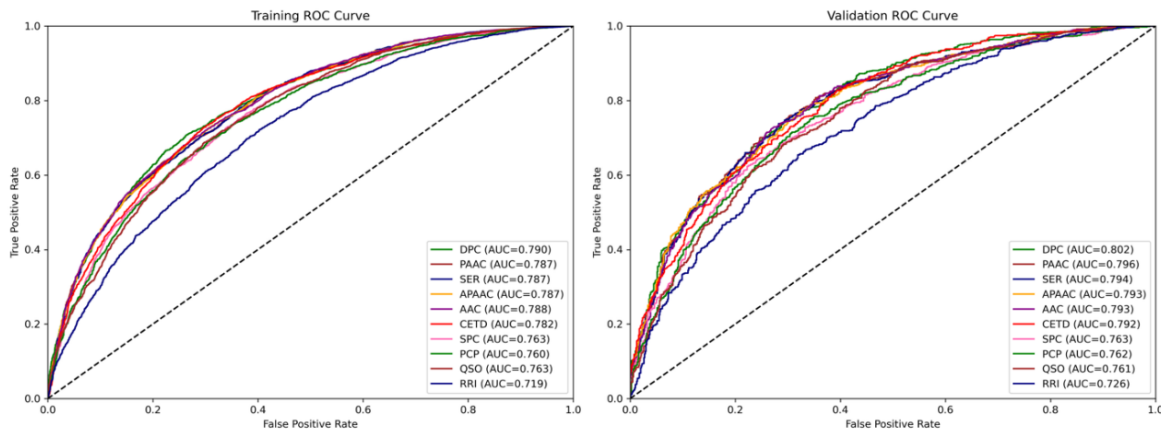


Figure 4.4 ROC curve and AUROC scores of RF models based on ten composition descriptors (adopted from [19])

4.4.4 Performance on Combined Descriptors

The majority of feature descriptors in our analysis performed similarly to the DPC features. Therefore, we combined all descriptors achieving an AUROC greater than 0.760 to construct a hybrid feature vector of 780 dimensions, which also included the DPC-400 features. The resulting hybrid representation comprised AAC, APAAC, CeTD, DPC, PAAC, PCP, SER, SPC, and QSO descriptors. As shown in Figure 4.5, the RF classifier once again performed best using this hybrid feature set, obtaining an AUROC of 0.800 and balanced sensitivity and specificity.

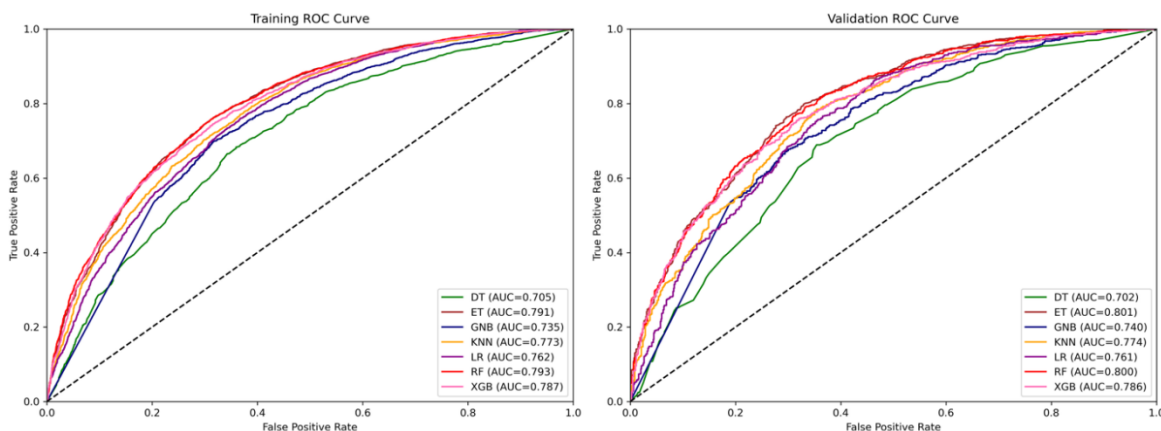


Figure 4.5. ROC curve and AUROC values showing the performance on hybrid features set (adopted from [19])

4.4.5 Performance of selected features

In this study, the mRMR algorithm [243] was applied to eliminate irrelevant and redundant features and to reduce interdependence among descriptors. In order to evaluate the hybrid feature set's performance, the top 400, 200, and 100 features were chosen. The highest AUROC obtained among these subsets was 0.79. Results from a 5-fold cross-validation comparing hybrid features with and without mRMR-based feature selection are shown in Figure 4.6.

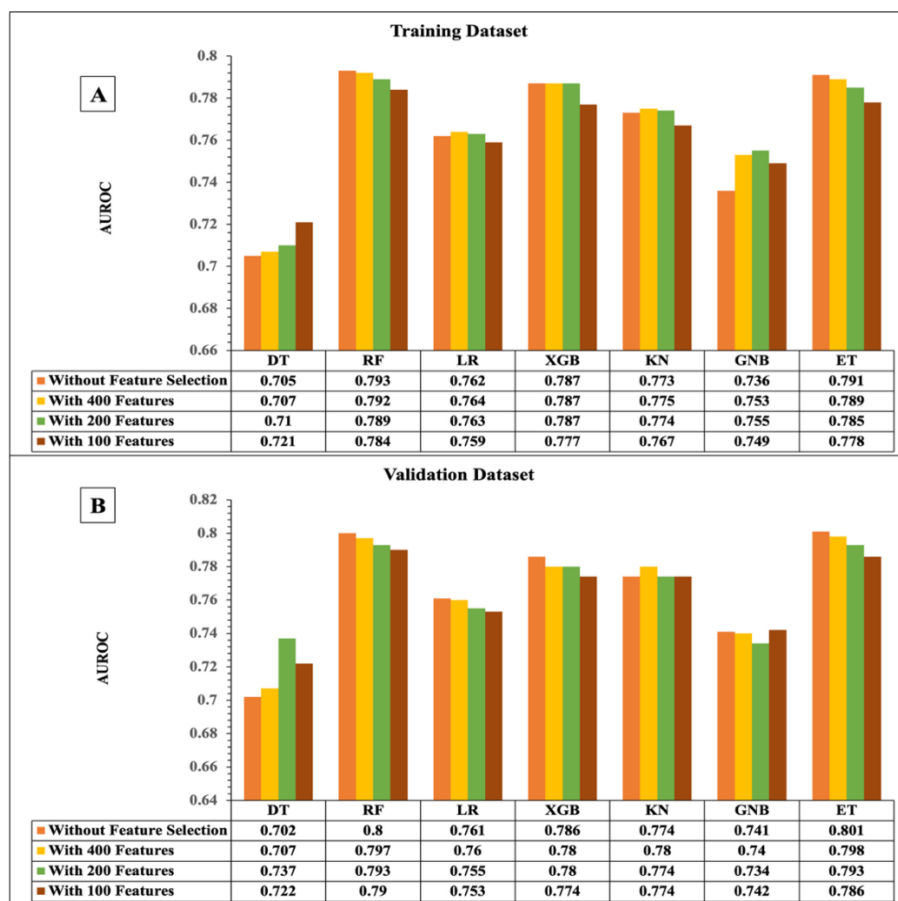


Figure 4.6 Comparative evaluation of classifier performance before and after feature selection for (A) training and (B) validation datasets (adopted from [19])

However, the dipeptide composition (DPC), a 400-dimensional feature representation, showed superior performance compared with the hybrid feature set in terms of AUROC (Figure 4.7). Therefore, in order to improve performance further analysis focused on the DPC descriptor. We have used the mRMR technique on DPC descriptor to extract a smaller set of features, and the top 200, 150, and 100 features were chosen to assess the model's performance. The results indicated that the complete DPC-400 feature set still outperformed all reduced subsets. The performance

(AUROC) of various classifiers trained on the DPC features set for both the training and validation datasets is shown in Figure 4.8.

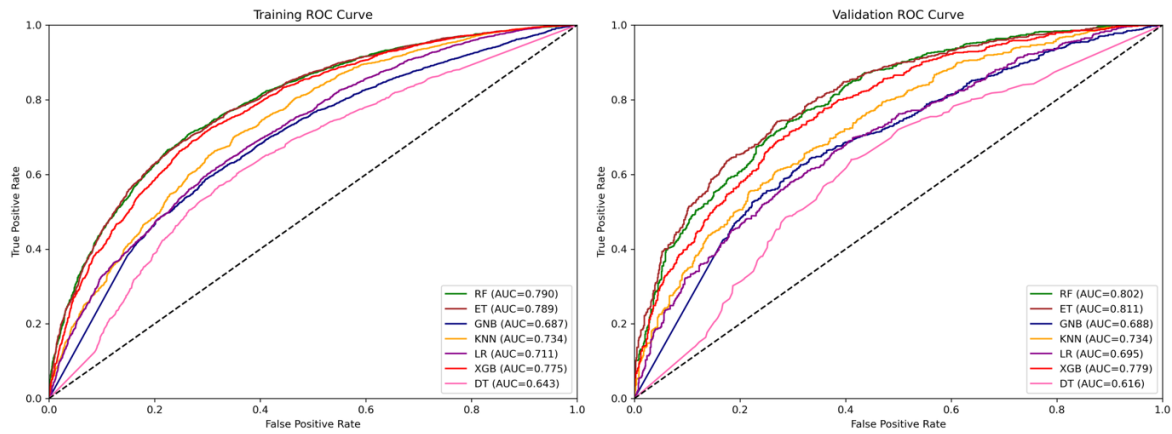


Figure 4.7 ROC curves and AUROC values demonstrating the performance of machine learning models trained using DPC features (adopted from [19])

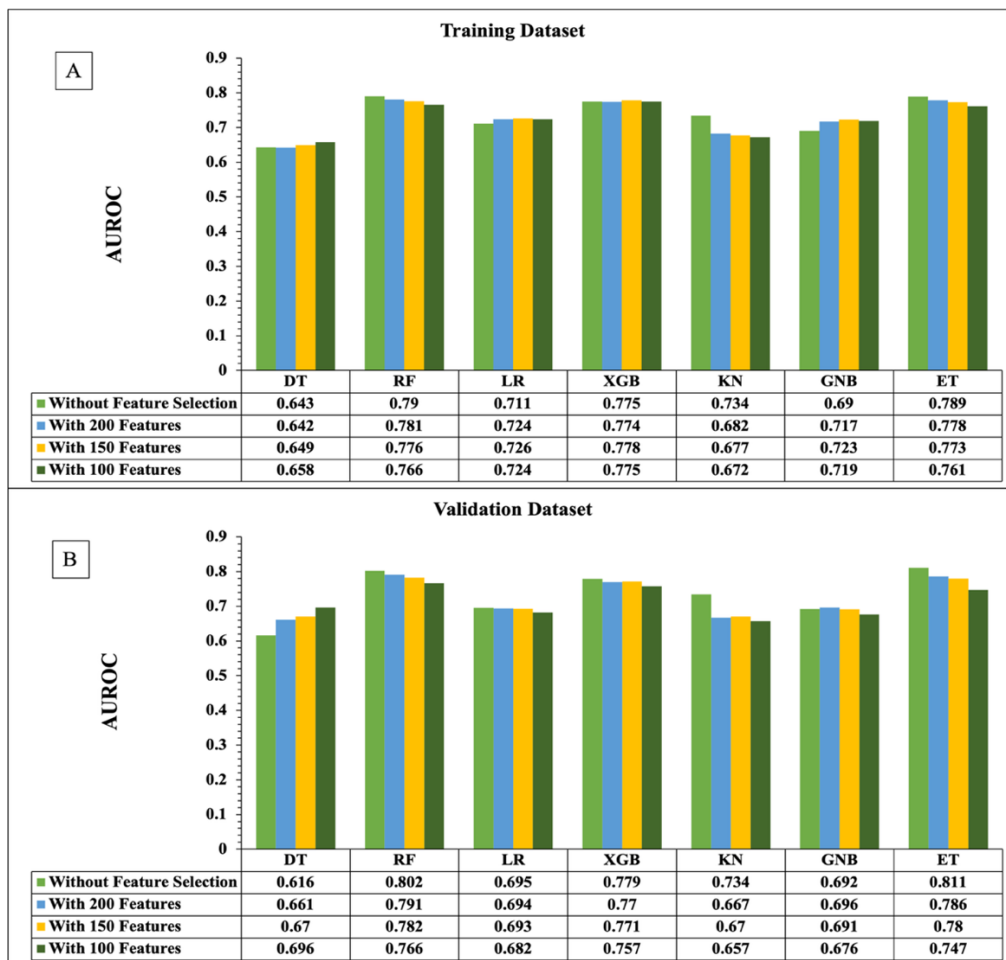


Figure 4.8 Classifier performance across different DPC feature set sizes (400, 200, 150, and 100) for (A) training and (B) validation datasets (adopted from [19])

4.4.6 Sequence Similarity Approach

We created a similarity search-based module to further improve model performance because BLAST has been widely used for protein annotation and functional assignment based on sequence similarity [125,266]. In this module, blastp was employed to classify a query peptide as either a BCE or a non-BCE. A local BLAST database built from the training dataset, and query sequences (test set) were searched against it using E-value thresholds ranging from $1e-6$ to $1e+3$. The top-scoring hit was used to annotate each query sequence. If the top hit matched a known BCE, the query peptide was categorized as a BCE; if not, it was designated as a non-BCE. The probability of correct prediction (C-Hits) varied between 1.94% and 70.32% for the B-cell epitope dataset and between 1.63% and 64.83% for the non-BCE dataset, as summarized in Table 4.1.

Table 4.1 Evaluation of BLAST similarity search at different E-values on the validation dataset (adopted from [19]).

	Positive Dataset						Negative Dataset					
E-value	No Hits		C-Hits		W-Hits		No Hits		C-Hits		W-Hits	
1.00E-06	759	(97.94%)	15	(1.94%)	1	(0.13%)	785	(98.25%)	13	(1.63%)	1	(0.13%)
1.00E-05	754	(97.29%)	18	(2.32%)	3	(0.39%)	779	(97.5%)	19	(2.38%)	1	(0.13%)
1.00E-04	732	(94.45%)	34	(4.39%)	9	(1.16%)	763	(95.49%)	27	(3.38%)	9	(1.13%)
1.00E-03	706	(91.10%)	57	(7.35%)	12	(1.55%)	747	(93.49%)	41	(5.13%)	11	(1.38%)
1.00E-02	673	(86.84%)	85	(10.97%)	17	(2.19%)	721	(90.24%)	59	(7.38%)	19	(2.38%)
1.00E-01	585	(75.48%)	165	(21.29%)	25	(3.23%)	689	(86.23%)	89	(11.14%)	21	(2.63%)
1.00E+00	463	(59.74%)	282	(36.39%)	30	(3.87%)	656	(82.10%)	118	(14.77%)	25	(3.13%)
1.00E+01	325	(41.94%)	386	(49.81%)	64	(8.26%)	519	(64.96%)	205	(25.66%)	75	(9.39%)
1.00E+02	52	(6.71%)	518	(66.84%)	205	(26.45%)	67	(8.39%)	467	(58.45%)	265	(33.17%)
2.00E+02	10	(1.29%)	538	(69.42%)	227	(29.29%)	14	(1.75%)	505	(63.20%)	280	(35.04%)
1.00E+03	1	(0.13%)	545	(70.32%)	229	(29.55%)	0	(0.00%)	518	(64.83%)	281	(35.17%)

*C-Hits: Correct hits; W-Hits: Wrong hits

4.4.7 Hybrid methods

To further improve prediction performance, we combined multiple approaches to develop a hybrid prediction model. Specifically, the machine learning-based prediction scores were integrated with the similarity search scores obtained from BLAST and the motif-based scores derived from the MERCI method.

1. Hybrid1 approach:

In this approach, we combined the machine learning prediction score with the BLAST similarity search score. Among the evaluated models, RF classifier trained on the DPC-400 feature vector. Therefore, using the top-performing RF model to calculate the machine learning prediction score was the first step. Then, each query peptide was classified using BLAST at various E-value thresholds. Following that, a scoring system was implemented, where a correct positive hit (BCE) was given a value of +0.5, a correct negative hit (non-BCE) was given a value of -0.5, and no significant hit was obtained was given a value of 0 [125]. For every query sequence, the final classification was obtained by combining the ML prediction score and the BLAST score. Using this integrated scoring strategy, the hybrid model performance was evaluated at various E-value thresholds (Figure 4.9). At an E-value of 1, the model showed its best performance attained a maximum AUROC of 0.829 and an accuracy of 74.46%. The detailed performance metrics, including sensitivity, specificity, accuracy, MCC, and AUROC, are provided in Table 4.2.

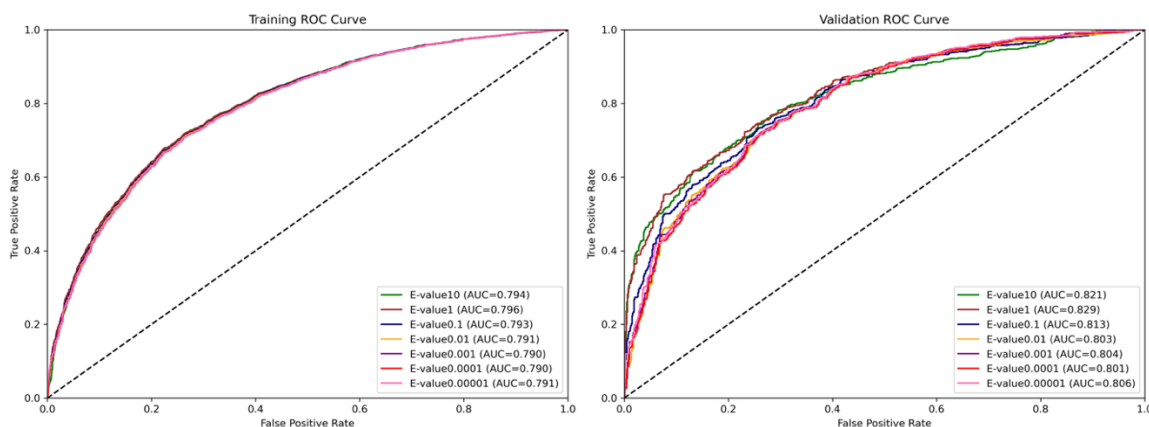


Figure 4.9 ROC curve and AUROC values showing the model performance on hybrid method (BLAST and DPC-400) (adopted from [19])

Table 4.2: Model performance developed using the Hybrid method (BLAST and DPC-400) on training and validation dataset

E-Value	Training dataset					Validation dataset				
	Sens	Spec	Acc	AUROC	MCC	Sens	Spec	Acc	AUROC	MCC
1.00E-05	71.645	72.693	72.177	0.791	0.443	72.903	72.34	72.618	0.806	0.452
1.00E-04	71.677	72.599	72.145	0.79	0.443	72.903	71.965	72.427	0.801	0.449
1.00E-03	71.677	72.631	72.161	0.79	0.443	73.419	71.84	72.618	0.804	0.453
1.00E-02	71.839	72.693	72.273	0.791	0.445	72.387	73.467	72.935	0.803	0.459
1.00E-01	72.097	72.631	72.368	0.793	0.447	73.161	74.218	73.698	0.813	0.474
1.00E+00	72.484	72.693	72.59	0.796	0.452	74.323	74.593	74.46	0.829	0.489
1.00E+01	72.548	72.693	72.622	0.794	0.452	74.065	74.343	74.206	0.821	0.484

2. Hybrid2 approach:

This method combines the ML prediction score obtained from the DPC-400 feature-based model with motif-derived scores generated by the MERCI tool. In order to distinguish between different biological classes, MERCI finds motifs made up of particular patterns of amino acids and physicochemical characteristics. It can also identify gapped motifs and uses two key parameters: FP, the minimum frequency threshold in the positive sequences, and FN, the maximum frequency threshold in the negative sequences [269]. To identify the class-specific motifs, we created distinct positive (BCE) and negative (non-BCE) datasets in order to find class-specific motifs. The top k motifs that were most prevalent in the positive dataset were extracted by MERCI using these datasets. Initially, we assessed the Koolman and Rohm classification without gaps, as well as basic configurations such as 1-gap and 2-gap motifs. Furthermore, both 1-gap and 2-gap motif settings were used to test the Koolman and Rohm grouping scheme. The hybrid model's performance was then assessed by combining the motif score with the DPC-400 machine learning prediction score. The model's maximum AUROC, when using the default MERCI parameters, was 0.67. After optimizing the FN parameter, the AUROC increased to 0.71.

4.4.8 Benchmarking with available methods

We also compared our proposed method with several existing B-cell epitope prediction tools, including iBCEEL [71], ABCPred [56], LBTope [58], CBTope [79], BCEPred [56], BepiPred 3.0 [76], iLBE [284], EpiDope [75], BepiBlast [31], SEMA [94], and BCEPS [73], using the same

validation dataset. The comparative evaluation demonstrated that our method outperformed the existing predictors across multiple performance measures, including sensitivity, specificity, accuracy, AUROC, and Matthews correlation coefficient (MCC). The ROC curves illustrating this comparative analysis are presented in Figure 4.10. The detailed evaluation metrics for all models, including sensitivity, specificity, accuracy, MCC, and AUROC, are provided in Table 4.3.

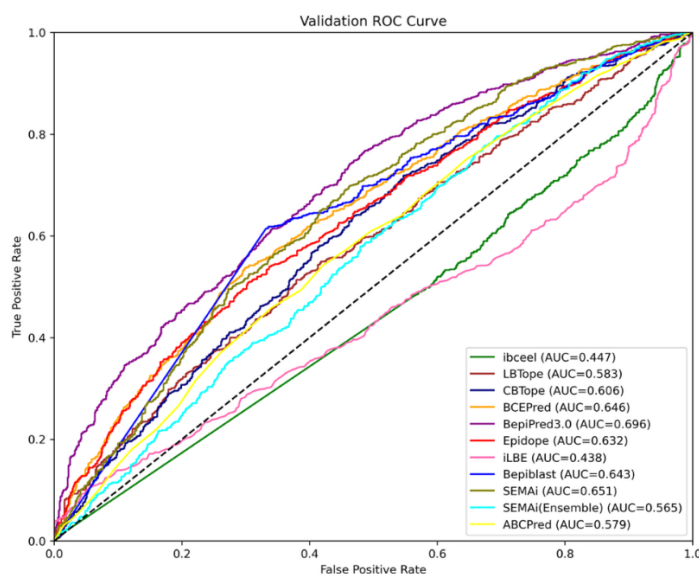


Figure 4.10 ROC curve and AUROC values comparing the performance of existing methods

Table 4.3. Comparative analysis of existing methods of B-cell prediction on the Independent Validation Dataset

Method Name	Year	Sens (%)	Spec (%)	Acc (%)	AUROC	MCC
BCEPred	2004	61.67	60.45	61.05	0.64	0.22
ABCPred	2006	57.43	54.09	55.68	0.57	0.11
CBTope	2010	57.03	59.19	58.13	0.6	0.16
LBTope	2013	55.74	55.94	55.84	0.58	0.11
iBCEEL	2018	48.12	43.8	45.93	0.44	0.08
iLBE	2020	46.96	45.18	46.06	0.43	-0.07
EpiDope	2021	60.51	57.32	58.89	0.63	0.17
BCEPS	2021	7.1	97.87	53.18	NA	0.11
BepiPred3.0	2022	63.48	62.7	63.08	0.69	0.26
SEMA	2022	61.93	59.7	60.8	0.65	0.21
SEMA(Ensemble)	2022	55.22	53.44	54.32	0.56	0.08
BepiBlast	2022	63.35	62.57	62.96	0.64	0.25

4.5 SHAP-based interpretation

To interpret the CLBTope predictions, we applied SHapley Additive exPlanations (SHAP) and ranked features by their mean absolute SHAP value. Figure 4.11 shows the top 15 contributing features. B-cell epitope prediction was driven by dipeptide composition (DPC) descriptors. Among all of the top 15 features the leucine-leucine pair (DPC1_LL) had the highest contribution (mean $|\text{SHAP}| \approx 0.0047$), roughly $1.5\times$ the next feature, followed by the aliphatic pairs DPC1_VL (≈ 0.0032) and DPC1_GL (≈ 0.0030). Below these, the ranking combined acidic/charged pairs such as DPC1_EE, DPC1_EN, DPC1_DD, DPC1_SD, DPC1_KD, with further aliphatic and small-residue pairs such as DPC1_AL, DPC1_LA, DPC1_LV, DPC1_IL, DPC1_AA, DPC1_LG, and DPC1_AI.

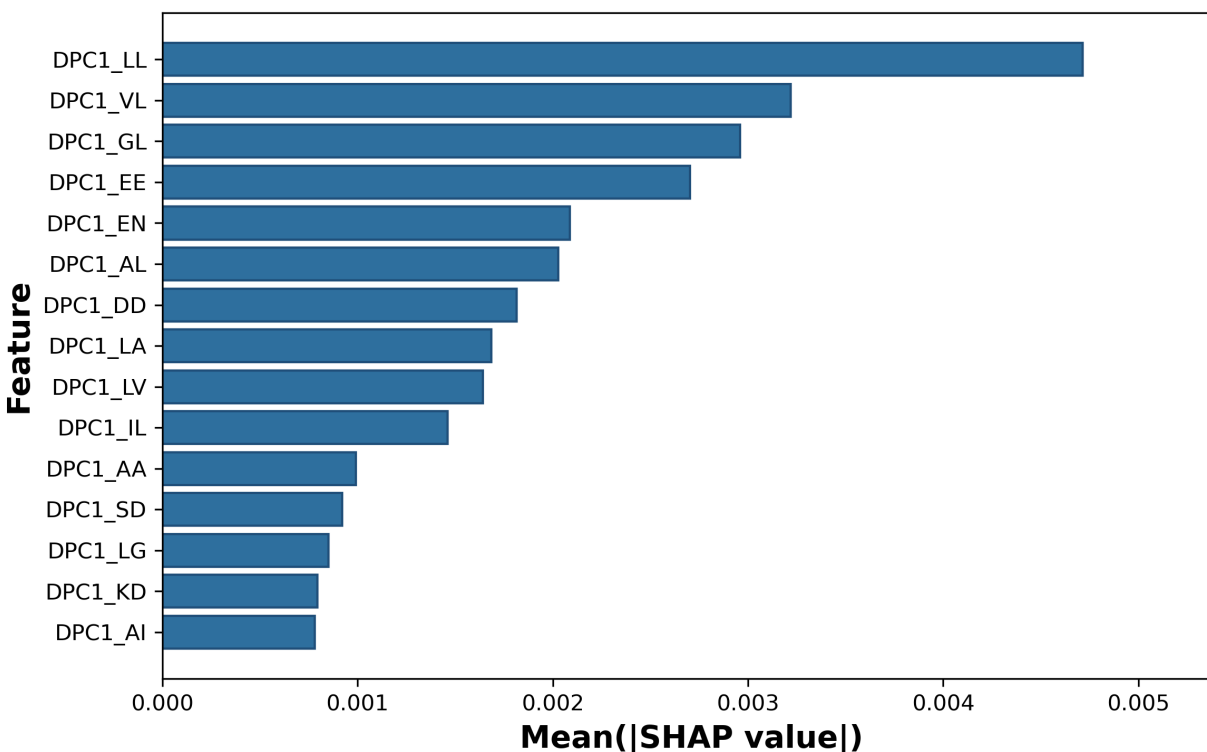


Figure 4.11 Top 15 features driving B-cell epitope prediction, ranked by mean $|\text{SHAP value}|$.

4.6 Case study: analysis of BCEs in SARS-CoV2

The spike protein's interacting surface that binds to the ACE2 receptor is located in its receptor binding domain (RBD), which is also very immunogenic. Previous studies on SARS-CoV2 have shown that most potent neutralizing monoclonal antibodies (mAbs) target regions near the ACE2-binding surface of the RBD, thereby blocking receptor interaction on host cells or disrupting the

prefusion conformation of the spike protein [285]. In this case study, we used the Protein Scan module of the CLBTope method to find BCEs within the 195 amino acid SARS-CoV-2 RBD. We used a default threshold (cut-off = 0.53) and a peptide window length of 11 residues. 103 BCEs and 82 non-BCEs were predicted by the analysis to be present in this region. The immune system may recognize these predicted epitopes as immunogenic regions, which could be potential targets for antibody-mediated responses against SARS-CoV-2. Table 4.3 lists the top 10 high-confidence B-cell epitopes found in the polyprotein.

Table 4.3 Potential BCE in the SARS-CoV2 receptor-binding domain predicted using the proposed method

Start	End	Sequence	Hybrid_Score
37	47	YNSASFSTFKC	1.09
39	49	SASFSTFKCYG	1.02
38	48	NSASFSTFKCY	1.01
40	50	ASFSTFKCYGV	0.97
48	58	YGVSP TKLNDL	0.94
41	51	SFSTFKCYGVS	0.93
44	54	TFKCYGVSP TK	0.93
47	57	CYGVSP TKLND	0.92
42	52	FSTFKCYGVSP	0.91
49	59	GVSP TKLNDLC	0.91

4.7 Web server implementation

For the scientific community's convenience and accessibility, CLBTope (<https://webs.iiitd.edu.in/raghava/clbtope/>) was deployed as a publicly accessible web server. The general web server development methodology is described in *section 3.14*. The best-performing prediction model was implemented within the server and organized into several functional modules, including (i) Predict, (ii) Design, (iii) Scan, (iv) Motif Scan, and (v) Blast Scan.

4.8 Discussion and Conclusion

The ability of epitope-based peptide vaccines to deliver targeted vaccines precisely provides significant advantages over traditional vaccines. The prediction of BCE helps immunologists

create vaccines based on these epitopes, as well as precise diagnostic instruments and successful treatment plans [286]. In this work, we presented a computational approach to predict B-cell and non-B-cell epitopes in an antigen based on its primary sequence. A selective dataset of 3,875 B-cell epitopes (positive set) and 3,996 non-B-cell epitopes (negative set) from BCEPS was used to train the model. Multiple sequence-based descriptors were extracted using the Pfeature package, and ten composition-based feature types were used to evaluate various machine learning algorithms, including DT, RF, LR, XGB, KNN, GNB, and ET. With a maximum AUROC of 0.802 using the DPC descriptor (400-dimensional feature vector), the RF classifier outperformed the rest of the models. An RF model was then trained using a hybrid feature vector (760 dimensions), which produced an AUROC of 0.800. A maximum AUROC of 0.79 was obtained by evaluating further dimensionality reduction using the mRMR feature selection approach on the top 400, 200, and 100 selected features. A hybrid RF-based framework that combines DPC-400 features with BLAST similarity search was created to further improve prediction performance. With a maximum AUROC of 0.829 on the validation dataset, this integrated approach performed the best. A comparison with existing prediction methods (Figure 4.10) showed that the proposed approach performed better than others. A freely available web server available at (<https://webs.iitd.edu.in/raghava/clbtope/>) has been created to make the method accessible to the scientific community. It allows users to predict, scan, and design B-cell epitopes. We hope that researchers in the fields of immunoinformatics, vaccine design, and antibody-based therapies will find this resource useful. Despite the encouraging performance of CLBTope, the present study has certain limitations. First, most of the data were obtained from a single curated source (Ras-Carmona et al.; BCETD, ILED, and IDED). Second, the current predictor relies primarily on sequence-based information and does not explicitly incorporate three-dimensional structural information. Since conformational B-cell epitope prediction depends on the spatial arrangement of residues, integrating structural features in future studies may further improve prediction accuracy. Finally, future studies should validate the model using a completely independent external dataset from different institutions or sources to further establish its generalizability.

C H A P T E R

5

Identification of T1DM-Associated Peptides

5.1 Introduction

Diabetes mellitus is a chronic metabolic condition that can cause blindness, heart attacks, strokes, amputation of limbs, and renal failure, among other serious complications. According to the WHO, the global number of diabetes cases increased from approximately 200 million in 1990 to 830 million in 2022, with about 589 million adults currently affected and projections suggesting a rise to 853 million by 2050. Interestingly, over 81% of those affected are from countries with low or middle incomes [287]. Diabetes mellitus can be divided into two main categories: type 1 diabetes mellitus (T1DM), also known as insulin-dependent diabetes, and type 2 diabetes mellitus (T2DM), also known as adult-onset diabetes [288]. About 5–10% of all cases of diabetes are linked to T1DM, an immune-mediated, chronic autoimmune disease that is most frequently diagnosed in children and adolescents. The disease leads to absolute insulin deficiency due to progressive destruction of insulin-producing pancreatic β -cells. Several autoantigens are known to trigger autoimmune responses against pancreatic β -cells. According to Yoon et al. [289], dendritic cells, β -cell autoantigens, T lymphocytes, B lymphocytes, and macrophages all have important roles in the development of autoimmune diabetes. Environmental factors and genetic susceptibility are linked to influence the development of type 1 diabetes [290] (Figure 5.1).

According to reports, a number of conditions, such as Addison's disease, coronary heart disease, type A gastritis, retinopathy, rheumatoid arthritis, autoimmune thyroid diseases, vitiligo, systemic lupus erythematosus, and celiac disease, are linked to the progression of type 1 diabetes [291]. The development and severity of disease are significantly influenced by genetic susceptibility, especially when it comes to human leukocyte antigen (HLA) class I and class II molecules. According to Primavera et al. [103], the risk of type 1 diabetes is considerably increased when certain allele haplotypes, like DR4–DQ8, are present. Other haplotypes, such as DR3–DQ2, have also been linked to the susceptibility, indicating that it is not just the DR4–DQ8 haplotype. In addition, most T1DM patients exhibit circulating autoantibodies, autoreactive T cells, and anti-islet antigen-specific immune responses, further highlighting the autoimmune nature of the disease [102].

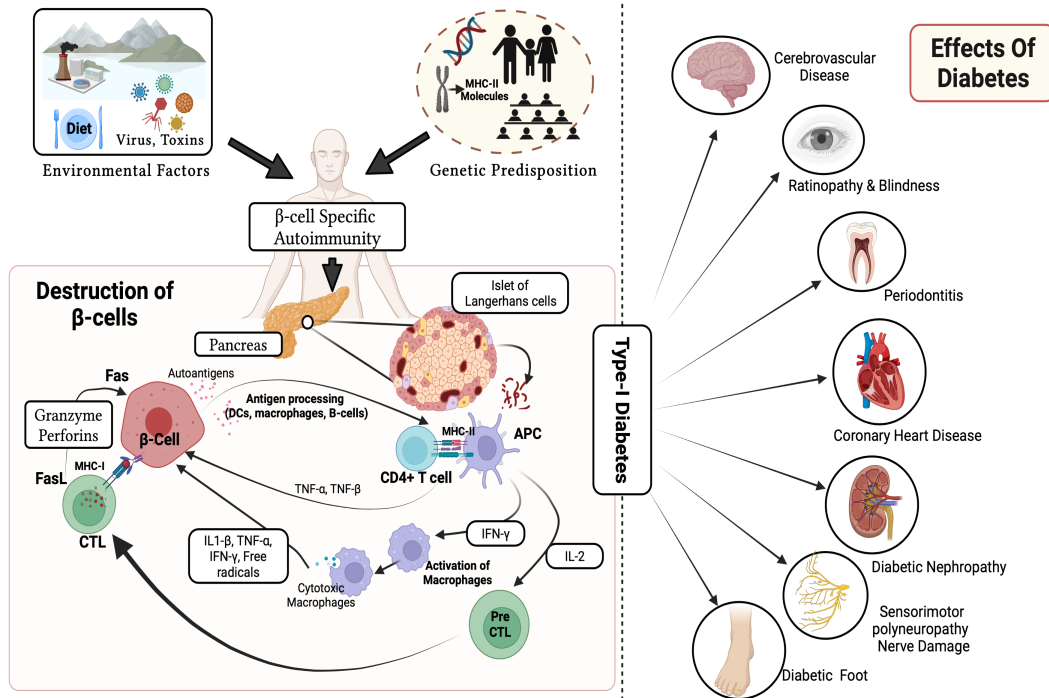


Figure 5.1 Brief on Type 1 Diabetes and immune destruction of β -cells (adopted from [102])

A important step in knowing the progression of the disease and immune-mediated β -cell destruction is the identification of T1DM-associated peptides. MHC class II alleles are usually bind these peptides, which then cause immune system responses that damage pancreatic β -cells. Only a few previous studies have attempted to use in silico methods to predict T1DM-associated peptides. For example, the GPS-MBA method was created by Cai et al. [103] to predict MHC class II binding peptides linked to type 1 diabetes. However, this approach's primary drawback is that it was trained on a comparatively small dataset that only included HLA-DQ8 binders. In order to predict T1DM-associated peptides using larger datasets, a computational method that is both efficient and more accurate needs to be developed.

5.2 Overview of the study

In this work, we provide the first computational framework for precisely predicting T1DM-associated peptide prediction, design, and scanning. Figure 5.2 shows the general architecture of the proposed approach. Both alignment-based (BLAST) and alignment-free (ML-based) methodologies are included in the framework. Additionally, a hybrid model was created to combine the advantages of both approaches, increasing prediction reliability and predictability. A user-friendly web server (<https://webs.iitd.edu.in/raghava/dmpped/>), a standalone package, a

GitHub repository, and a PyPI package for the prediction, scanning, and design of T1DM-associated peptides are among the various implementation options also provided to enable greater utilization by the scientific community.

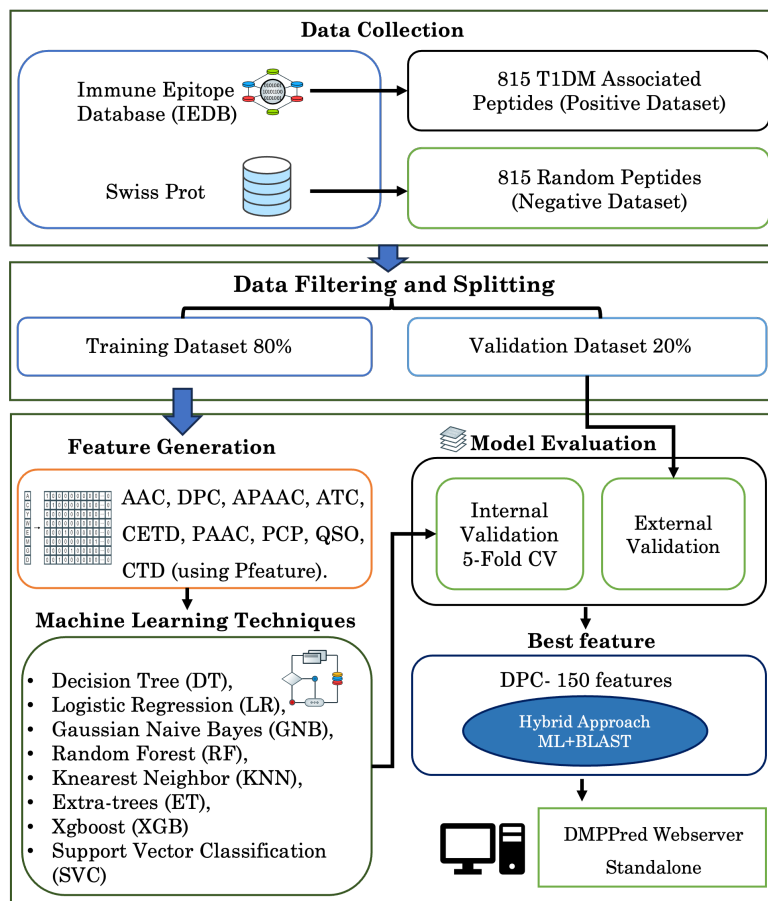


Figure 5.2 Overall architecture of the study

5.3 Materials and Methods

5.3.1 Dataset collection

The dataset was constructed using the standardized strategy described in *Chapter 3*. In this study, we have collected the T1DM-associated peptides from IEDB. After removing redundant entries, peptides were filtered based on length distribution. Since the majority of experimentally validated peptides fall within the 8-30 amino acid range, sequences shorter than 8 or longer than 30 were excluded. This resulted in 815 unique T1DM-associated peptides forming the positive dataset. Due to the limited availability of experimentally validated negative peptides, we generate it by randomly sampling peptides of matching length distribution from Swiss-Prot proteins. An equal

number of negative examples (815 peptides) were generated to avoid class imbalance. The complete dataset (1630 peptides) was then partitioned into training (80%) and independent validation (20%) datasets following the unbiased splitting protocol described in *section 3.4.1*.

5.3.2 Feature generation

To construct an alignment-free prediction model, informative numerical features (descriptors) were generated from peptide sequences. In this study, a diverse set of compositional descriptors was computed, including AAC, DPC, APAAC, ATC, CETD, DDR, PAAC, PCP, QSO, RRI, SPC, and CTD (Table 3.3). Features were calculated using the Pfeature tool as described in *section 3.7.1*. Altogether, a large feature pool (1153 features) was generated, which was further refined using feature selection approaches.

5.3.3 Feature selection

One of the main challenges in this study is choosing an appropriate subset of features from a high-dimensional feature space. As explained in *Section 3.8*, we used the minimum Redundancy Maximum Relevance (mRMR) algorithm [102] among the different feature selection methods. The top 50, 100, and 150 features chosen by the mRMR method were then used to train machine learning models, and their predictive ability was evaluated.

5.3.4 ML Model Development and Evaluation

Multiple supervised classifiers were evaluated, including DT, LR, GNB, RF, KNN, ET, XGB, and SVC, described in *section 3.9.1*. To develop generalizable models and avoid bias or overfitting, training was performed on the training dataset using 5-fold cross-validation, while model performance was evaluated on an independent validation dataset. Evaluation metrics included AUROC and MCC along with sensitivity, specificity and accuracy, described in *section 3.13.1*.

5.3.5 Similarity Search

A similarity-based module was implemented using BLAST (blastp-short task from BLAST+ suite), described in *section 3.10.1*. A local BLAST database was constructed using training peptides. Query peptides from an independent dataset were searched against this database at different e-value thresholds ranging from $1e-6$ to $1e+4$. The class of the query peptide was assigned based on the top hit: if the top hit was T1DM-associated, the query was classified as positive; otherwise, as negative.

5.4 Results

5.4.1 Compositional analysis

In both T1DM-associated and non-T1DM-associated peptides, the amino acid composition (AAC) was determined using *Equation 1, Section 3.5*. The average amino acid composition of randomly generated peptides, T1DM-associated peptides, and the general proteome was calculated as part of the analysis, as shown in Figure 5.3. Leu (L), Met (M), Val (V), and Glu (E) were found to be the most prevalent residues in the positive dataset. In contrast, the negative dataset had greater percentages of preservation for Asp (D), Lys (K), and Arg (R). The enrichment of leucine in T1DM-associated peptides can be explained by the fact that leucine is a bulky, hydrophobic residue. The T1DM risk-associated HLA molecules HLA-A*02:01 (class I) and HLA-DR4 (class II) have hydrophobic anchor pockets that favour such residues, so leucine fits their anchor requirements well [292,293]. This is consistent with the high frequency and positional conservation of leucine in the T1DM-associated peptides (Figure 5.4).

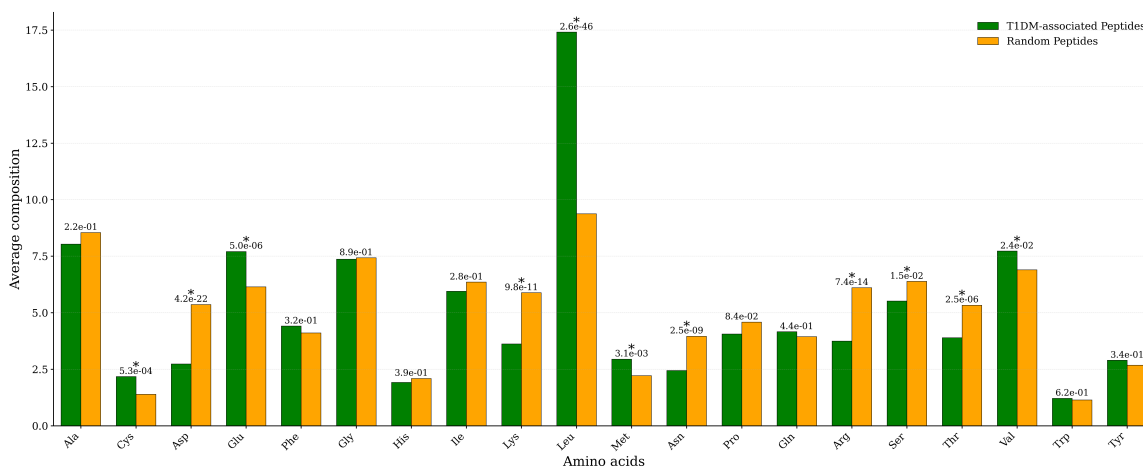


Figure 5.3 Amino acid composition analysis of T1DM-associated peptides, and random peptides

5.4.2 Positional analysis

To investigate the positional preference and frequency of particular amino acid residues within T1DM-associated peptides, a sequence logo (*refer to Section 3.6*) was created. The hydrophobic residue leucine (Leu) was extremely prevalent and conserved across multiple positions, as seen in Figure 5.4. Whereas valine (Val) was primarily found at the 2nd, 3rd, 9th, and 16th positions, alanine (Ala) displayed high conservation at the 2nd, 9th, 11th, and 14th positions. Furthermore, glutamic acid (Glu), a hydrophilic residue, was enriched at positions 2nd, 4th, 11th, and 14th.

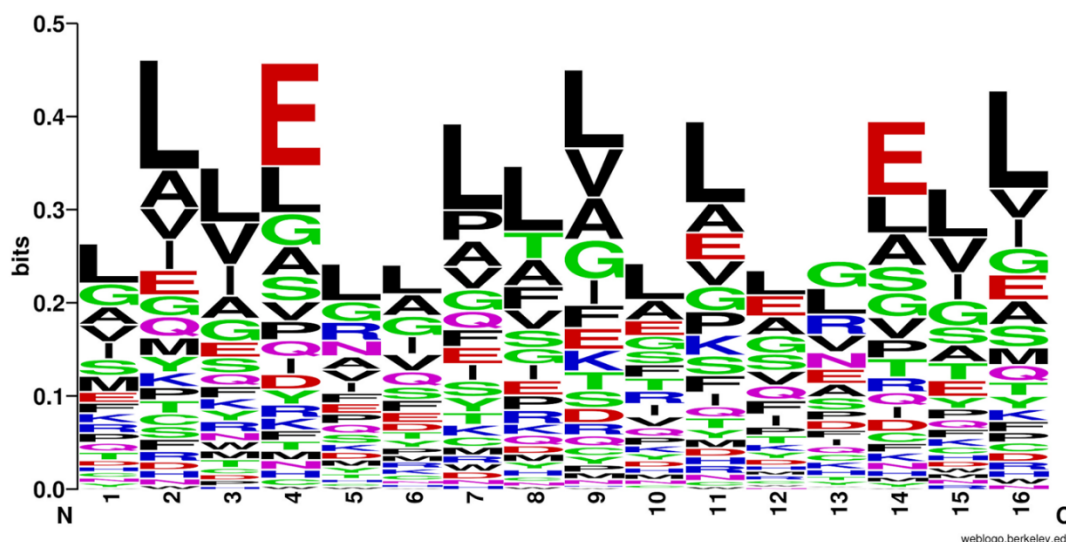


Figure 5.4 Illustration of sequence logo of T1DM associated peptides (adopted from [102])

5.4.3 HLA Allele Distribution

According to previous studies, class II HLA alleles are crucial for the onset of type 1 diabetes [294]. We examined the frequency of HLA allele interactions with T1DM-associated peptides in order to evaluate their binding efficacy. Table 5.1 lists the top 10 class I and class II HLA binders found in the dataset linked to T1DM.

Table 5.1 Frequency of HLA-binding alleles observed in T1DM-associated peptides (adopted from [102])

HLA allele	Frequency
HLA-A*02:01	345
HLA-A2	57
HLA-DQA1*03:01/DQB1*03:02	103
HLA-DQA1*05:01/DQB1*03:02	45
HLA-DR	45
HLA-DR3	39
HLA-DR4	59
HLA-DR53	21
HLA-DRB1*04:01	185
HLA-DRB4*01:01	20

5.4.4 Performance of compositional descriptors

To build prediction models, machine learning algorithms such as DT, RF, LR, GNB, ET, XGB, KNN, and SVC were used. Models were first built using twelve distinct feature types that were calculated using the Pfeature tool. The ET algorithm performed the best overall out of all the classifiers. All 12 feature descriptor's predictive performance is shown in terms of AUROC and MCC in Table 5.2. With an AUROC of 0.893 on the training dataset and 0.891 on the validation dataset, models based on DPC features demonstrated the best performance. On the training and validation datasets, models created with AAC-based features also demonstrated strong performance, obtaining AUROC values of 0.869 and 0.870, respectively.

Table 5.2 Performance of 12 composition-based descriptors using ET algorithm (adopted from [102])

Descriptor	Training Dataset		Validation Dataset	
	AUROC	MCC	AUROC	MCC
AAC	0.869	0.572	0.870	0.571
DPC	0.893	0.621	0.891	0.620
APAAC	0.869	0.551	0.875	0.571
ATC	0.710	0.302	0.749	0.356
CETD	0.849	0.531	0.864	0.546
DDR	0.837	0.500	0.797	0.417
PAAC	0.868	0.569	0.869	0.546
PCP	0.840	0.503	0.847	0.497
QSO	0.858	0.552	0.861	0.565
RRI	0.854	0.540	0.810	0.423
SPC	0.791	0.442	0.812	0.462
CTD	0.856	0.537	0.835	0.473

5.4.5 Performance of Feature Selection Models

Since the highest performance was achieved using DPC features with the ET classifier, these features were further explored to enhance classification performance. The most informative subset of features for differentiating T1DM-associated peptides from non-T1DM-associated peptides was found using the mRMR feature selection algorithm. Based on AUROC, the top 50, 100, and 150

features showed a satisfactory level of discriminatory ability. The performance (AUROC) of several classifiers trained on different feature subsets for both training and validation datasets is shown in Figure 5.5. Table 5.3 shows the detailed performance of the top 150 features chosen using the mRMR method. With balanced sensitivity and specificity and AUROC values of 0.893 and 0.871 on the training and validation datasets, respectively, the ET classifier performed the best. Similar results were achieved by RF models, even with a slight decline in accuracy and AUROC on the training dataset.

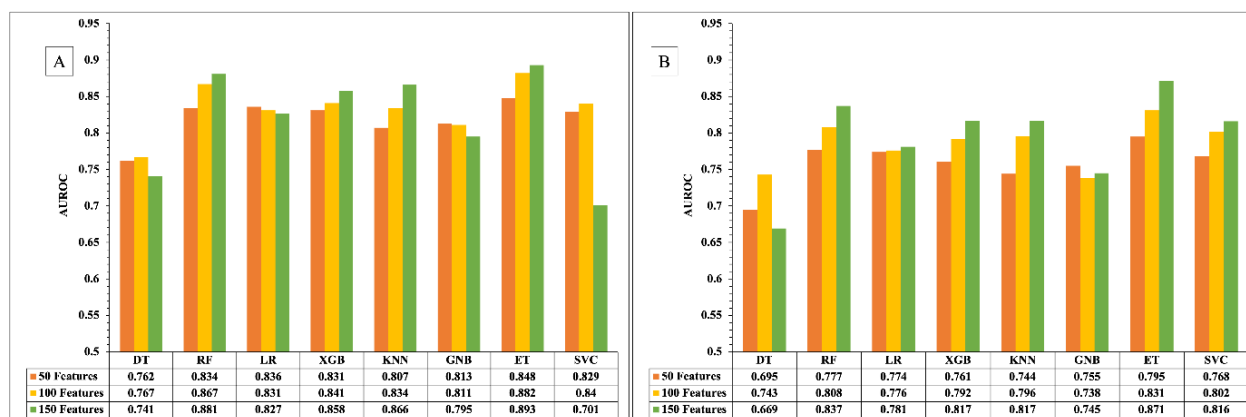


Figure 5.5 Performance comparison of classifiers across feature subsets: (A) training, (B) validation datasets (adopted from [102])

Table 5.3 ML performance using DPC-150 feature subset on training and validation datasets (adopted from [102])

Classifier	Training					Validation				
	Sens (%)	Spec (%)	Acc (%)	AUROC	MCC	Sens (%)	Spec (%)	Acc (%)	AUROC	MCC
DT	69.479	69.939	69.709	0.741	0.394	68.098	64.417	66.258	0.669	0.325
RF	79.908	79.755	79.831	0.881	0.597	76.074	74.233	75.153	0.837	0.503
LR	75.460	75.613	75.537	0.827	0.511	71.779	69.325	70.552	0.781	0.411
XGB	77.607	77.914	77.761	0.858	0.555	74.847	74.233	74.540	0.817	0.491
KNN	77.914	78.374	78.144	0.866	0.563	73.006	76.074	74.540	0.817	0.491
GNB	72.853	74.693	73.773	0.795	0.476	68.712	69.325	69.018	0.745	0.380
ET	81.902	81.135	81.518	0.893	0.630	82.209	77.301	79.755	0.871	0.596
SVC	78.374	77.761	78.067	0.865	0.561	73.620	74.233	73.926	0.831	0.479

Additionally, the 150 features that had been selected were ranked according to how the dipeptide composition of the positive and negative datasets was different. Each dipeptide composition in the

positive and negative sets was analyzed using a two-sample t-test to determine the statistical significance of these differences. It was determined that the observed differences were statistically significant. The top 10 dipeptides LL, VE, LV, SL, AL, LQ, LE, IL, ER, and LA, were found to be significantly more common in T1DM-associated peptides than in non-T1DM-associated peptides based on this analysis (Table 5.4).

Table 5.3 Showing the frequency of top dipeptide composition (adopted from [102])

DPC	Frequency in T1DM	Frequency in Non-T1DM	#Average1	#Average2	Difference	p-value
LL	194	79	3.22	0.92	2.30	1.66E-17
VE	176	48	1.91	0.42	1.49	1.30E-22
LV	145	62	2.04	0.62	1.42	6.39E-14
SL	156	55	1.79	0.49	1.30	2.08E-16
AL	204	94	2.22	0.95	1.27	1.77E-12
LQ	130	38	1.51	0.41	1.10	6.93E-14
LE	148	57	1.57	0.63	0.94	7.83E-10
IL	104	64	1.42	0.63	0.79	3.67E-06
ER	108	39	1.17	0.39	0.78	2.91E-09
LA	148	89	1.62	0.85	0.78	2.40E-06

#Average1: Average values of T1DM associated peptides, #Average2: Average values of non-T1DM associated peptides

5.4.6 Similarity search approach

We created a similarity search module based on BLAST, which has been widely used for protein annotation and functional assignment through sequence similarity [266], in order to further improve model accuracy. To categorize query peptides as T1DM-associated or non-T1DM-associated, the blastp technique was applied. The training dataset was used to build a local BLAST database, and the query sequences (test set) were searched against to this database using E-value thresholds between $1e-6$ and $1e+4$. Based on the top-scoring hit, each query sequence was given a class: if the top hit matched a peptide linked to type 1 diabetes, the query sequence was categorized as T1DM-associated; if not, it was labeled as non-T1DM-associated. The T1DM-associated dataset had a probability of correct prediction (C-Hits) ranging from 17.18% to 87.73%, while the non-T1DM-associated dataset had a probability ranging from 0.00% to 61.96%, as shown in Table 5.5.

Table 5.4 Evaluation of BLAST similarity search on validation dataset (adopted from [102])

E-Value	T1DM Associated			Non-T1DM Associated		
	C-Hits	W-Hits	No Hits	C-Hits	W-Hits	No Hits
1.00E-06	28 (17.18%)	0 (0.00%)	135 (82.82%)	0 (0.00%)	0 (0.00%)	163 (100.00%)
1.00E-05	41 (25.15%)	0 (0.00%)	122 (74.85%)	0 (0.00%)	0 (0.00%)	163 (100.00%)
1.00E-04	54 (33.13%)	0 (0.00%)	109 (66.87%)	0 (0.00%)	0 (0.00%)	163 (100.00%)
1.00E-03	62 (38.04%)	0 (0.00%)	101 (61.96%)	0 (0.00%)	0 (0.00%)	163 (100.00%)
1.00E-02	81 (49.69%)	0 (0.00%)	82 (50.31%)	1 (0.61%)	0 (0.00%)	162 (99.39%)
1.00E-01	91 (55.83%)	0 (0.00%)	72 (44.17%)	1 (0.61%)	0 (0.00%)	162 (99.39%)
1.00E+00	111 (68.10%)	2 (1.23%)	50 (30.67%)	2 (1.23%)	0 (0.00%)	161 (98.77%)
1.00E+01	124 (76.07%)	5 (3.07%)	34 (20.86%)	21 (12.88%)	11 (6.75%)	131 (80.37%)
1.00E+02	143 (87.73%)	19 (11.66%)	1 (0.61%)	90 (55.21%)	46 (28.22%)	27 (16.56%)
2.00E+02	143 (87.73%)	19 (11.66%)	1 (0.61%)	97 (59.51%)	53 (32.52%)	13 (7.98%)
1.00E+03	143 (87.73%)	19 (11.66%)	1 (0.61%)	100 (61.35%)	55 (33.74%)	8 (4.91%)
1.00E+04	143 (87.73%)	19 (11.66%)	1 (0.61%)	101 (61.96%)	56 (34.36%)	6 (3.68%)

*C-Hits: Correct hits; W-Hits: Wrong hits

5.4.7 Hybrid model performance analysis

This study combined BLAST similarity search scores with machine learning predictions to create a hybrid (ensemble) approach. Initially, each query peptide in the independent dataset was classified using BLAST at various E-value thresholds by searching against a local database constructed from the training dataset. A scoring system based on the top hit was employed: 0 represented no significant hit, +0.5 represented a correct positive prediction (T1DM-associated peptide), and -0.5 represented a correct negative prediction (non-T1DM-associated peptide). Previous studies have used a similar scoring system [116]. Next, prediction scores were generated using the machine learning models. The overall prediction score was then calculated by combining the ML prediction score and the BLAST score for every query sequence. The difference between sensitivity and specificity was minimized in order to establish the ideal threshold. As indicated in Tables 5.2 and 5.3, the ET classifier showed the best performance when using DPC and DPC-150

feature sets. Therefore, hybrid models were constructed using the DPC-150 features (Table 5.6). The best results were obtained with an E-value of 0.1, according to performance evaluation at various E-values. The training and validation datasets had accuracies of 88.57% and 90.46%, respectively, and maximum AUROC values of 0.94 and 0.95.

Table 5.5 Performance of Hybrid method (BLAST and DPC-150) (adopted from [102])

Training Dataset					
E-value	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUROC	MCC
1.00E-06	83.55	83.72	83.64	0.91	0.67
1.00E-05	84.05	84.87	84.46	0.91	0.69
1.00E-04	84.87	84.87	84.87	0.92	0.70
1.00E-03	85.53	85.53	85.53	0.92	0.71
1.00E-02	86.84	86.51	86.68	0.93	0.73
1.00E-01	88.98	88.16	88.57	0.94	0.77
1.00E+00	89.80	89.15	89.47	0.94	0.79
Validation Dataset					
1.00E-06	81.58	90.13	85.86	0.93	0.72
1.00E-05	81.58	91.45	86.51	0.93	0.73
1.00E-04	82.24	91.45	86.84	0.93	0.74
1.00E-03	84.87	91.45	88.16	0.94	0.76
1.00E-02	85.53	92.11	88.82	0.94	0.77
1.00E-01	87.50	93.42	90.46	0.95	0.81
1.00E+00	86.84	90.13	88.49	0.94	0.77

5.5 SHAP-based interpretation

To interpret the DMPPred predictions, we applied SHapley Additive exPlanations (SHAP) and ranked features by their mean absolute SHAP value. Figure 5.6 shows the top 15 contributing features. As with the other composition-based models in this thesis, prediction was driven entirely by dipeptide composition (DPC) descriptors — all top 15 features belong to the DPC family. The leucine–valine pair (DPC1_LV) had the largest contribution, followed by DPC1_LL and DPC1_VL. The remaining high-ranking features were also dominated by leucine-, valine-, and glutamate-containing pairs: DPC1_VE, DPC1_FL, DPC1_LE, DPC1_LQ, DPC1_IL, DPC1_QV, DPC1_AL, DPC1_SL, DPC1_DA, DPC1_GK, DPC1_MV, and DPC1_AN.

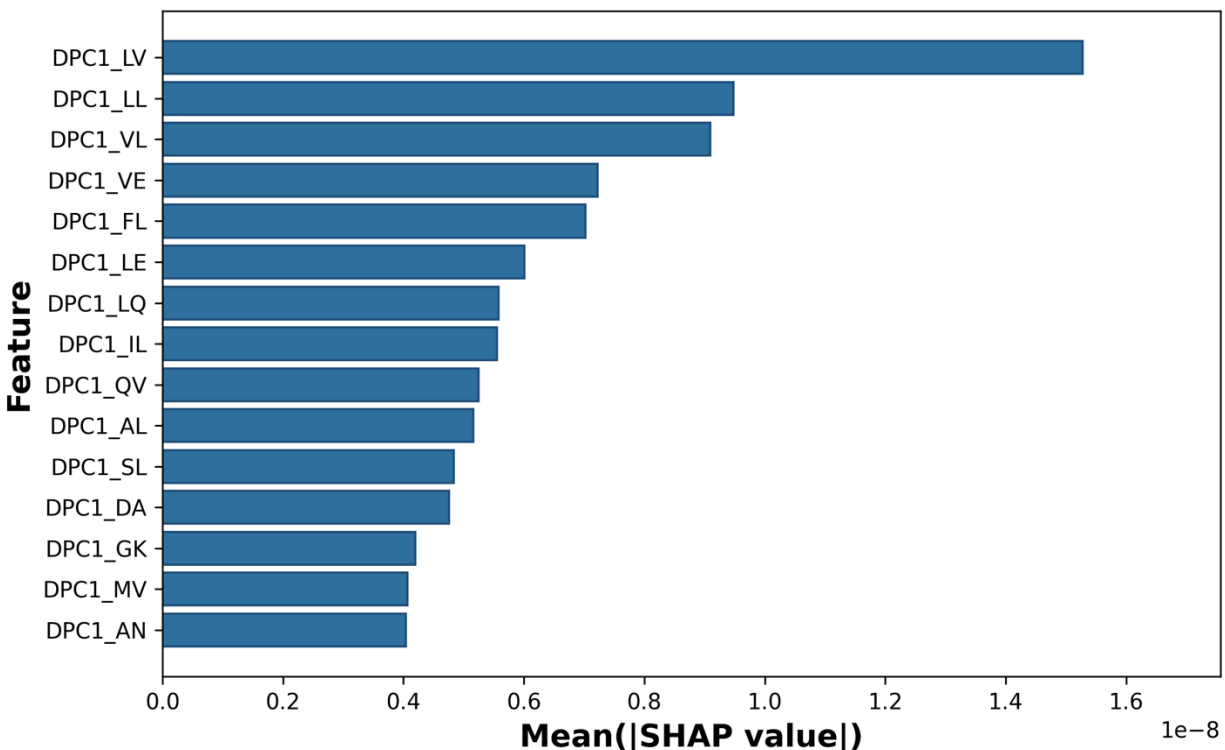


Figure 5.6. SHAP-based feature importance for the T1DM-associated peptide prediction model.

The dominance of hydrophobic pairs (Leu, Val, Met) matches the anchor-residue preferences of the risk alleles HLA-A*02:01 and HLA-DR4, whose anchor pockets favour hydrophobic residues, while the glutamate-containing pairs reflect the acidic anchor preference of the T1DM-associated HLA-DQ8 molecule.

5.6 Case Study: Coxsackievirus B4 Antigenic Regions

Previous studies have suggested that multiple environmental factors, including viral infections, contribute to the progression of T1DM. A number of viruses, including enteroviruses like rotavirus [295], cytomegalovirus [296], mumps virus [297], and coxsackievirus B (CVB) [298], have been linked to T1DM. The most common enteroviral strain found in patients with diabetes and pre-diabetes among these is coxsackievirus B4 (CVB4). The Protein Scan module of DMPPred was used in this study to predict T1DM-associated peptides in two CVB4 proteins, VP1 and the polyprotein, with a peptide window length of 15 residues. The default cut-off value of 0.58 was replaced with a higher threshold score of 0.70 to reduce false-positive predictions. In the polyprotein, we found 46 T1DM-associated peptides with scores ≥ 0.70 . However, none of the

VP1 protein's peptides received a score higher than 0.65. According to DMPPred predictions, these findings imply that the polyprotein is enriched with potential T1DM-associated peptides. This observation aligns with earlier research showing the CVB4 polyprotein plays a role in the development of type 1 diabetes. Table 5.7 lists the top 10 T1DM-associated peptides with high confidence that were found in the polyprotein.

Table 5.6 Prediction of potential T1DM inducing antigenic peptides in the CBV4 polyprotein (adopted from [102])

Start	End	Sequence	Hybrid Score
1128	1142	EWLKVKILPEVREKH	0.97
1129	1143	WLKVKILPEVREKHE	0.97
1130	1144	LKVKILPEVREKHEF	0.97
1122	1136	KIQKFIEWLKVKILP	0.96
1127	1141	IEWLKVKILPEVREK	0.92
1126	1140	FIEWLKVKILPEVRE	0.89
1125	1139	KFIEWLKVKILPEVR	0.86
1131	1145	KVKILPEVREKHEFL	0.86
1123	1137	IQKFIEWLKVKILPE	0.85
1124	1138	QKFIEWLKVKILPEV	0.85

5.7 Web server

To ensure usability for experimental researchers and clinicians, DMPPred was implemented as a public web server and standalone package. The design principles were simplicity, functionality and interpretability. The web interface provides multiple modules:

- (i) **Predict:** classification of peptides as T1DM-associated or random,
- (ii) **Protein Scan:** sliding window identification of T1DM regions within proteins,
- (iii) **Design:** generation of all possible analog peptides and scoring,
- (iv) **BLAST Scan:** similarity-based annotation.

The dataset used for model development is available for download in FASTA format to support reproducibility and future benchmarking.

5.8 Discussion and Conclusion

At present, a definitive cure for type 1 diabetes mellitus (T1DM) remains unavailable, and insulin therapy continues to be the primary treatment for disease management. Recently, antigen-specific immunotherapy has shown promising clinical outcomes, demonstrating both safety and improved efficacy. These antigens can be delivered in a number of forms, such as proteins, peptides, DNA plasmids, and nanoparticles, and they are comparatively easy to synthesize [299]. Identifying T1DM-specific epitopes or peptides is therefore essential for both experimental study and the development of biomedical therapeutics. In this study, we suggested a novel computational technique for predicting peptides that are associated with T1DM and non T1DM-associated peptides. The amino acids L, M, and E are enriched in T1DM-associated peptides when compared to the negative dataset, according to compositional and positional analyses. Among different feature representations, DPC perform the best with AUROC values of 0.893 and 0.891 on the training and validation datasets, respectively. Subsequently, feature selection using the mRMR method was performed on DPC feature subsets (DPC-50, DPC-100, and DPC-150). The DPC-150 feature set provided the best results, achieving AUROC values of 0.893 and 0.871 on the training and validation datasets using the ET classifier. A hybrid model was then developed by integrating BLAST similarity search with DPC-150 features, which further improved performance and achieved AUROC values of 0.945 and 0.951 on the training and validation datasets, respectively. The proposed method was also compared with the existing GPS-MBA predictor. Because GPS-MBA accepts only peptides of length nine, we selected 50 T1DM-associated and 50 non-T1DM-associated peptides of length nine from the independent dataset and evaluated both tools. DMPPred demonstrated superior performance, achieving sensitivity, specificity, and accuracy of 92%, 100%, and 96.0%, respectively, whereas GPS-MBA achieved sensitivity of 6%, specificity of 90%, and accuracy of 48.0%. Furthermore, DMPPred was used to identify antigenic regions within the CVB4 polyprotein. To provide a simpler way for researchers to identify possible disease-associated antigens that could cause T1DM β -cell destruction, we developed DMPPred (<https://webs.iitd.edu.in/raghava/dmppred/>), an intuitive web server that facilitates the design, scanning, and prediction of T1DM-associated peptides, to assist the scientific community. This approach is expected to make a substantial contribution to the study of immunotherapy, autoimmunity, and vaccine development. The major limitation of the present study is that experimentally validated non-T1DM peptides are currently unavailable in public databases.

Therefore, randomly generated peptides were used as surrogate negative samples for model development. Although this strategy has been adopted in several computational peptide prediction studies, these sequences may not fully capture the diversity of biologically relevant non-T1DM peptides. The availability of experimentally validated negative datasets in the future will facilitate the development of more biologically representative prediction models.

Clinical significance: DMPPred is intended as a pre-experimental safety screening tool rather than a diagnostic or standalone decision-making system. Its primary application is in the design of peptide- and subunit-based vaccines, where it can help identify peptides with the potential to induce T1DM-associated autoimmune responses before experimental validation. Early identification of such peptides enables researchers to redesign or exclude potentially unsafe candidates, thereby improving vaccine safety. For example, viral proteins such as the Coxsackievirus B4 (CVB4) polyprotein contain regions exhibiting molecular mimicry with pancreatic β -cell autoantigens and have been associated with T1DM. However, false-positive predictions may lead to the exclusion of peptides that are actually safe and immunogenic. Therefore, the predictions generated by DMPPred should be considered as an initial prioritization step and should always be complemented with appropriate experimental and immunological validation before their application in vaccine or therapeutic development.

Managing false positives: DMPPred allows users to adjust the prediction threshold according to the requirements of the application. By default, the threshold is selected to provide a balanced sensitivity and specificity. However, for safety-critical applications, such as vaccine and biotherapeutic development, users can adopt a higher threshold to reduce false-positive predictions and improve specificity. In addition, the hybrid prediction framework integrates machine learning with BLAST-based similarity search, providing complementary evidence to increase prediction confidence. This strategy was applied in the CVB4 case study, where a stringent cut-off of 0.70 was used in place of the default 0.58 to minimise false positives; this identified 46 high-confidence T1DM-associated peptides in the CVB4 polyprotein, whereas no peptide in VP1 exceeded a score of 0.65. This illustrates how DMPPred can be tuned to achieve more stringent, application-specific screening.

C H A P T E R

6

Allergenicity Prediction of Therapeutic Peptides

6.1 Introduction

Allergic diseases are a growing global health concern, affecting over 20-30% of the world's population and increasing in complexity and severity in both children and adults. These disorders include immune-mediated inflammatory diseases caused by hypersensitivity responses to typically harmless environmental substances, including food allergies, atopic dermatitis, asthma, and allergic rhinitis [105,106,300]. In recent decades, the incidence of allergic diseases has risen considerably, partly due to environmental and lifestyle changes, and patients frequently exhibit heterogeneous clinical manifestations and high recurrence risk [301]. Allergy disorders are especially common in the developed world; for example, approximately 31.8% of adults in the United States suffer from at least one allergic condition [302], such as hay fever (allergic rhinitis), eczema (atopic dermatitis), allergic asthma, and specific food allergies [303]. In 1906, Clemens von Pirquet coined the term “allergy” to refer to increased or decreased reactivity to environmental exposures, but nowadays it is primarily used to refer to immune-mediated hypersensitivity. An exaggerated immune response to foreign substances, known as allergens, is a key feature of allergies. Allergens are most commonly proteins derived from aeroallergens and foods, but they can also originate from personal care products and occupational exposures [131,304]. Type I (IgE-mediated) allergic reactions are frequently triggered by foods (e.g., nuts, milk, fish, eggs), pollen, insect venom, house dust mite excrement, latex, and certain medications. The initial exposure induces sensitization, whereas subsequent encounters activate allergen-specific IgE bound to mast cells and basophils, leading to mediator release (e.g., histamine, leukotrienes). In severe cases, this response can progress to systemic reactions, including anaphylaxis [133,305].

Allergens may be broadly grouped into two categories. The first includes non-infectious environmental substances that trigger IgE production and, upon re-exposure, activate IgE-coated mast cells/basophils, producing the classical Type I hypersensitivity phenotype. This group includes pollens, animal dander, dust-mite particles, common allergenic foods (e.g., peanuts, shellfish, milk, eggs), latex, insect venoms, and several drug classes. The second group involves largely IgE-independent mechanisms that induce adaptive immune responses and local inflammation; allergic contact dermatitis caused by nickel or urushiol (poison ivy) represents a typical example [304,306]. Allergic inflammation typically progresses through an early phase (minutes), a late phase (hours), and chronic inflammation under repeated exposure [304]. Although current therapies can reduce symptoms, they rarely provide durable cure. Therefore, accurate

allergen identification, avoidance of re-exposure, and disruption of key steps in hypersensitivity pathways are crucial to reduce recurrence. This need is intensified by the expanding use of genetically modified or engineered proteins in food products, therapeutic drugs, and vaccines [307,308]. Allergen recognition is thus strongly influenced by amino acid sequence composition and three-dimensional structure, traditionally assessed using immunochemical, biochemical, and immunological assays [130,309,310]. However, in the post-genomic era, the rapid expansion of protein sequence data has made experimental allergen screening increasingly challenging because wet-lab assays remain laborious and time-consuming [111].

To standardize allergenicity assessment, the FAO and the WHO proposed a decision-tree framework for evaluating novel proteins [113]. The widely used criteria include: (i) the presence of short identical peptide segments (typically 6–8 consecutive residues) shared with known allergens, and (ii) >35% sequence similarity over an 80-residue sliding window, indicating potential cross-reactivity. While these sequence-homology rules are useful for detecting allergens closely related to known allergen families, they frequently fail for non-homologous allergens and may generate false-positive or false-negative predictions [114]. Consequently, computational allergenicity prediction has become essential to enable high-throughput screening and to prioritize candidates for experimental validation. A broad range of computational tools has been developed using ML and DL methods (*as shown in Table 2.5*). Traditional ML predictors include AlgPred, AllerTool, AllerHunter, PREAL, AllergenFP, AllerCatPro, AllerCatPro 2.0, AllerTOP v.2, AlgPred 2.0, and iAller. These methods employ classifiers such as RF, KNN, XGBoost, and SVM. Consistent with this, a recent chemometric study on food protein allergenicity reported SVM as an effective classifier [311]. In parallel, DL-based tools have emerged, including DeepAlgPro, AllergenAI, ALLERDET, and ProALL-D. Hybrid and ensemble approaches such as SEP-AlgPro (See Table 2.5) integrate multiple ML/DL components but can be computationally expensive. Together, these computational strategies aim to accelerate allergen identification and support applications in food safety, therapeutic protein development, and vaccine/biopharmaceutical research [312].

6.2 Overview of the study

The proposed method's overall workflow is shown in Figure 6.1. First, allergen and non-allergen datasets are gathered from various carefully selected sources. Next, highly similar sequences are

eliminated through redundancy reduction. After that, the dataset was divided into separate training and validation sets after being balanced. The Pfeature package was then used to perform thorough feature extraction. Then, binary profile representations and protein language model embeddings were generated and evaluated in parallel. Multiple machine learning classifiers were trained on these features, and model performance was optimized through hyperparameter tuning and feature selection techniques, including mRMR, Lasso-L1, and SVC-L1. The best-performing model was finally deployed as a publicly accessible web server, and standalone as well as downloadable software packages were also provided for user convenience.

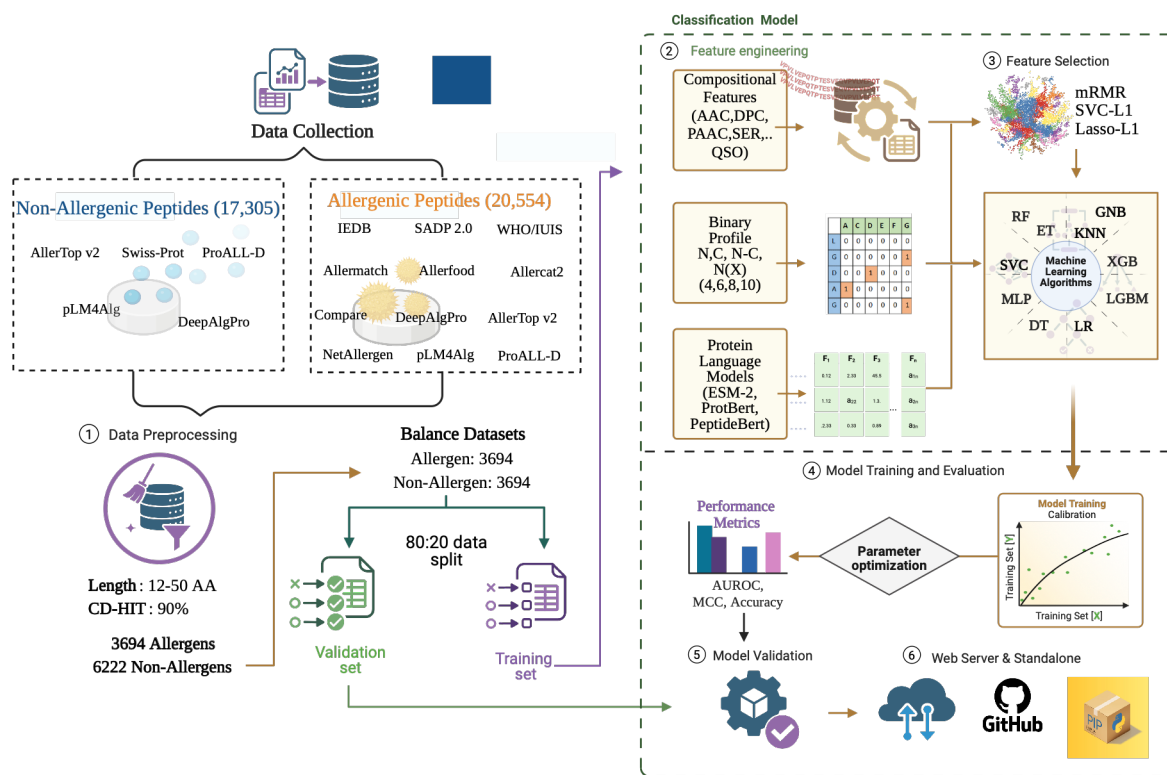


Figure 6.1 The complete architecture of the study

6.3 Materials and Methods

6.3.1 Dataset Compilation and Curation

The dataset used in this study includes 9348 allergens and 9348 non-allergens collected as follows:
(1) Allergenic Peptides, (2) Non- allergenic Peptides, (3) Data Splitting.

- 1. Allergenic Peptides:** The positive dataset consists of allergens collected from databases such as IEDB, SADP 2.0, WHO, Allermatch, and Allerfood, and prediction methods including AlgPred 2.0, pLM4Alg, Compare, DeepAlgo, AllerTop, ProAllD,

Allercat2, and NetAllergen. Initially, we had 20,554 allergenic epitopes/peptides; we then applied standard bioinformatics techniques, such as CD-HIT [206] to remove redundant epitopes. After removing duplicates and applying CD-HIT (90%), we are left with 3,694 allergenic peptides, varying in length from 12 to less than 50.

2. **Non-allergenic Peptides:** The “NOT” keyword “Allergen [KW-020]” was used to search Swiss-Prot for proteins annotated with the label "evidence for the existence of protein at the protein level" in order to gather the negative dataset, which is made up of non-allergens. Additionally, proteins with the terms “allergen”, “allergic”, “allergy”, or “allergens” in their description were not included. The available methods negative dataset includes AllerTop, DeepAlgo, pLM4Alg, and ProAllD. We also collected the negative assays epitope dataset from IEDB. After compiling all the datasets of peptides/epitopes of a length range from 12 to less than 50 and removing redundancy by 90% using CD-hit, we are left with 6,222 peptides. Furthermore, to create a balanced dataset, we have selected the same portion of negative peptides as in the positive dataset. The length distribution analysis was given in Figure 6.2.
3. **Data Splitting:** In this study, we implemented an 80:20 data splitting, as described in *Section 3.4.1*. That resulting in a training dataset of 5,910 peptides and an independent validation dataset of 1,478 peptides. This systematic selection is crucial for developing an unbiased evaluation of our model’s predictive capability.

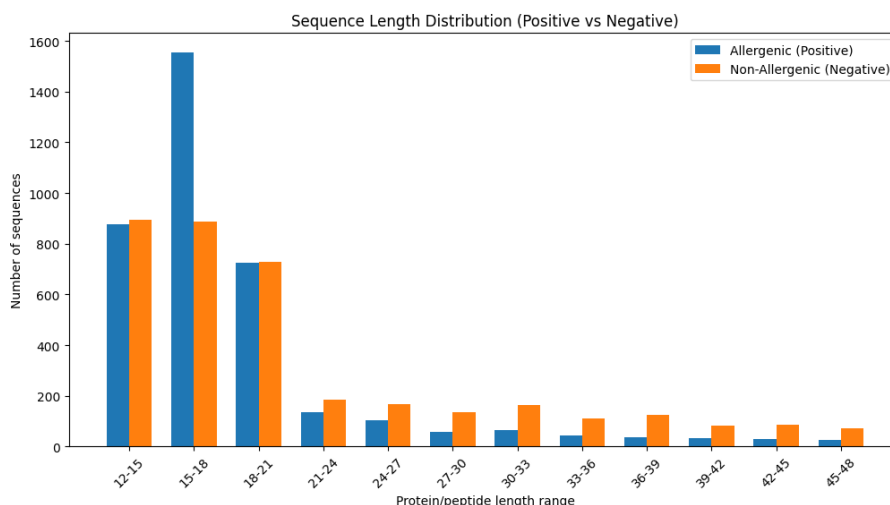


Figure 6.2 The length distribution of Allergenic vs non-Allergenic balanced dataset

6.3.2 Feature generation

6.3.2.1 Composition features

To develop prediction models capable of discriminating between allergen and non-allergen peptides, the sequences were first transformed into numerical feature vectors. For this purpose, the composition module of Pfeature was employed to compute various sequence descriptors, as described in *Section 3.7.1*. The calculated composition-based features included AAC (20), APAAC (23), BTC (4), ATC (5), CeTD (189), DDOR (20), DPC (400), PAAC (21), PCP (30), QSO (42), RRI (20), SOC (2), SER (20), SPC (25), CTC (343), SEP (1), and a combined feature set comprising all descriptors listed in Table 3.3.

6.3.2.2 Binary Profile Patterns

In addition to composition-based descriptors, four types of binary profile feature sets were generated using peptide fragments of different lengths (4, 6, 8, 10, and 12 residues): (i) N-terminal binary profiles (N), (ii) C-terminal binary profiles (C), (iii) combined N- and C-terminal profiles (N-C), and (iv) a fixed-length representation, in which a dummy residue 'X' was appended at the C-terminal end of peptide sequences shorter than 50 residues to standardize their length. All these feature representations were utilized to develop classification models for identifying allergenic peptides. As explained in *Section 3.7.2* (Figure 3.3), each peptide pattern was encoded into a binary profile, with each amino acid represented by a 21-dimensional vector (for example, alanine represented as 1,0]. Binary profile encoding has been widely applied in several existing prediction methods [79,227].

6.3.3 Feature fusion and Feature selection

In biomolecular sequence analysis, feature extraction from a single perspective frequently yields insufficient sequence information, which reduces classification models' predictive performance. To address this limitation, multiple composition-based feature types were combined through feature fusion [130]. However, multi-feature fusion typically produces high-dimensional and redundant feature vectors. Therefore, feature selection is an important preprocessing step in ML, as it identifies relevant set of features from a potentially large feature space. Eliminating irrelevant or redundant features can improve classification accuracy while enhancing model interpretability. Accordingly, to select informative descriptors and reduce dimensionality, we applied feature selection techniques including mRMR, LASSO-L1, and SVC-L1 (described in *Section 3.8*).

6.3.4 Development and Performance Evaluation of Models

A number of ML algorithms, such as XGB, LR, KNN, DT, GNB, RF, ET, LGBM, SVC, and MLP (explained in *Section 3.9.1*), were evaluated. The models were evaluated on an independent validation set after being trained using 5-fold cross-validation on the training dataset (*Section 3.12*). The scikit-learn library was used to perform grid search hyperparameter tuning in order to maximize classifier performance (*Section 3.11*). Along with sensitivity, specificity, and accuracy, the model's performance has been evaluated using AUROC and MCC, as detailed in *Section 3.13.1*.

6.3.5 Large Language Models as Classifiers

Recent developments in protein language models based on NLP have shown excellent performance in a number of protein analysis tasks. Several well-known large language models, such as ESM-2, ProtBERT, and PeptideBERT (explained in *Section 3.9.2*), were used in this study. In particular, several ESM series variants were used, including esm2 (t33, t30, t12, and t6). During the fine-tuning process, these model's hyperparameters were adjusted in order to better perform our prediction task.

6.3.6 Protein Language Model Embeddings

We used embeddings from pretrained protein language models to describe protein sequence data. In this study, ProtBERT and the ESM-2 model were used to obtain protein sequence embeddings (described in *Section 3.7.4*). From input sequences, the pretrained ProtBERT model generates high-dimensional, context-aware representations. The model creates an embedding matrix for a protein sequence of length n , where each row represents an encoded representation of a distinct amino acid residue. Multiple checkpoints of the ESM-2 model, differing in model size, were employed and obtained from the Hugging Face repository [313] (Figure 6.3). Although larger variants such as ESM-2 3B and 15B offer improved representational capacity, they require substantially greater memory and training time, making them less practical for large-scale experiments. The resulting embeddings were subsequently used as input features for developing the prediction models.

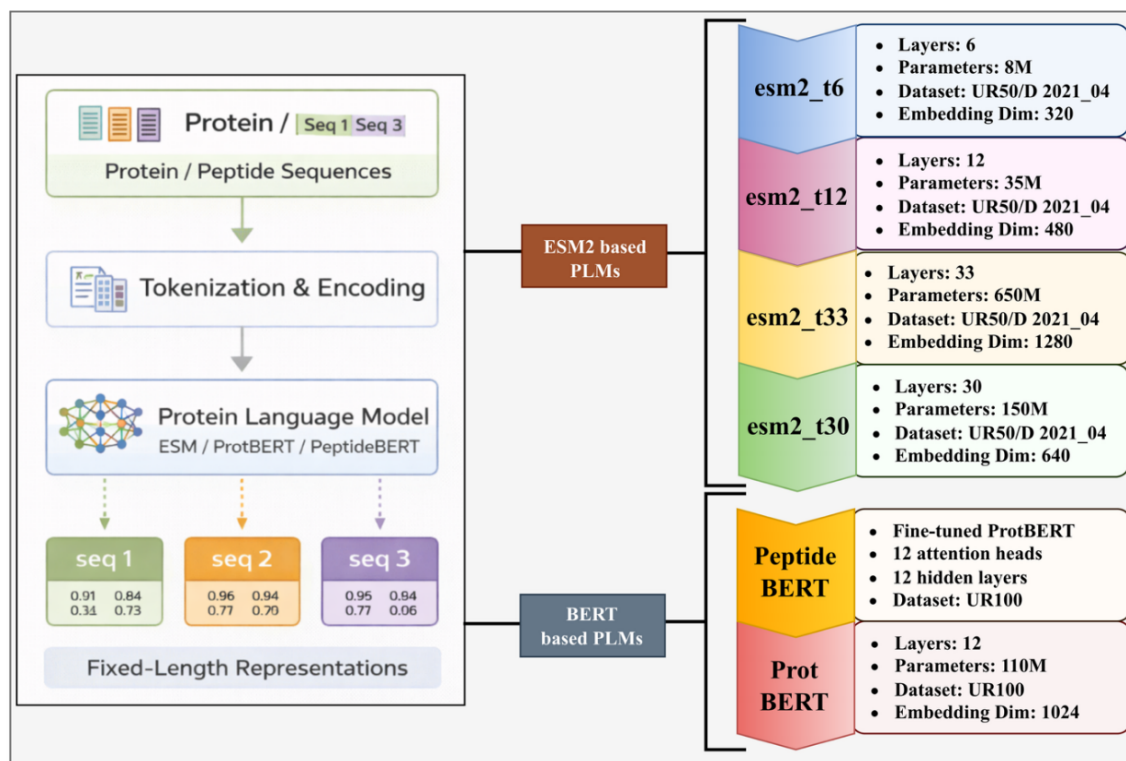


Figure 6.3 Different protein language models used in this study including parameters

6.3.7 Quantum SVM

Over the past ten years, quantum machine learning (QML), and specifically quantum support vector machines (QSVMs), has emerged as a suitable substitute for classical machine learning methods. QML introduces novel computational paradigms for constructing predictive models with the potential to enhance learning performance [314,315]. The PennyLane library was used in this study to implement QSVM in a Python 3.10 environment. By facilitating automatic differentiation and seamless interaction with both quantum simulators and quantum hardware, PennyLane offers a framework for quantum machine learning [316]. We adopted a comparative approach to evaluate quantum and classical machine learning models using the AAC dataset. Specifically, QSVM model performance was compared with classical classifiers, including standard SVC and RF. To investigate the effect of the quantum feature space on classification performance, experiments were conducted with different numbers of qubits (6, 10, and 14).

6.4 Results

6.4.1 Amino Acid Composition Analysis

We analysed the average AAC composition of allergenic and non-allergenic peptides using Equation 1 mentioned in *Section 3.5*. Amino acid composition was compared between allergenic and non-allergenic peptides (Figure 6.4). Thirteen residues differed significantly between the two groups ($p < 0.05$): Cys, Asp, Glu, Gly, Ile, Lys, Leu, Pro, Gln, Ser, Thr, Trp and Tyr. Cys, Gly, Ile and Leu occurred at higher frequency in non-allergenic peptides, whereas Glu, Pro and Gln were enriched in allergenic peptides. The strongest differences were recorded for Gln ($p = 1.6\text{e-}92$), Glu ($p = 3.0\text{e-}28$) and Pro ($p = 3.0\text{e-}20$). These three residues therefore account for much of the compositional signal separating the two classes. In contrast, amino acids such as Phe, His, Asp, Asn, Tyr, and Ser showed no significant differences ($p > 0.05$) across the two groups, indicating that their abundance remains relatively comparable between allergenic and non-allergenic peptides. We also compute the average amino acid composition of allergen proteins and allergenic peptides shown in Figure 6.5.

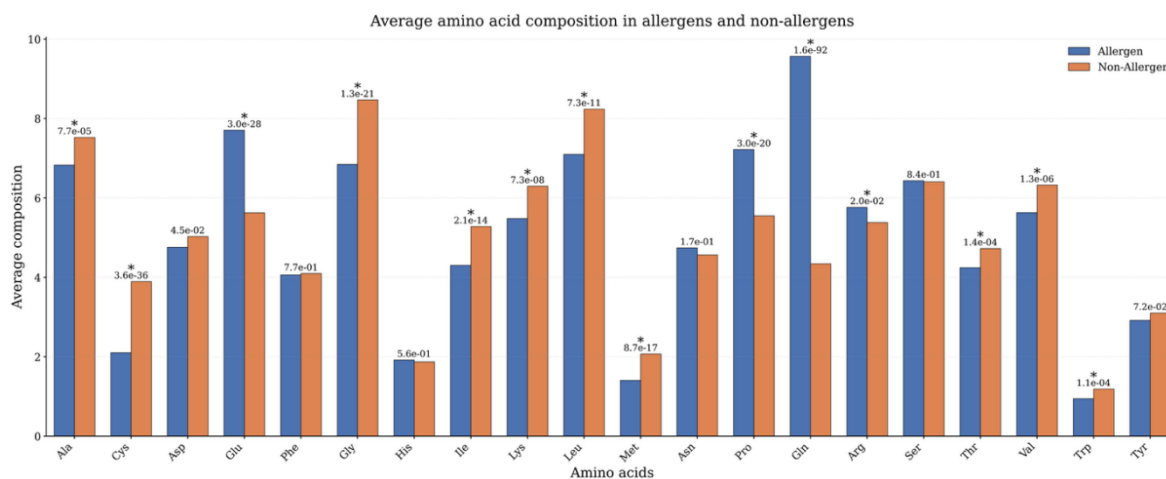


Figure 6.4 The amino acid composition comparison of allergen, and non-allergen

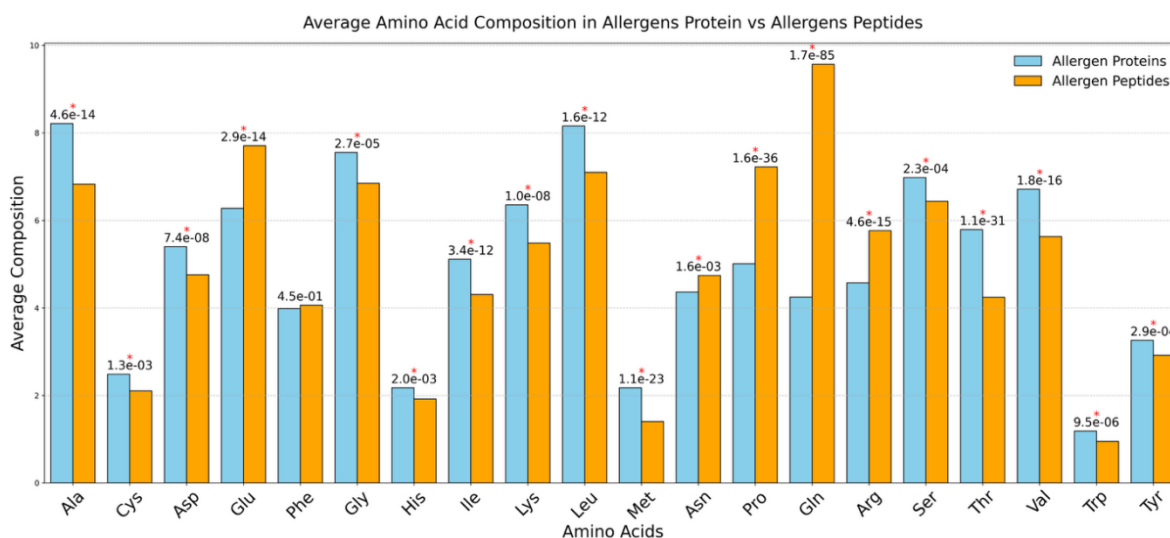


Figure 6.5 The amino acid composition comparison of allergen proteins, allergenic peptides

6.4.2 Positional Analysis

The positional frequency of all 20 natural amino acids at both peptide termini was analyzed. The analysis indicated that certain residues are preferentially enriched in allergenic peptides. To visualize residue preferences at specific positions, sequence logos were generated using the Two Sample Logo (TSL) method (*see Section 3.6 for details*). In this representation, the first eleven positions correspond to the N-terminal residues, whereas the last eleven positions correspond to the C-terminal residues of the peptides. The results revealed that glutamine (Gln) was highly abundant across most positions (Figure 6.6). In addition, proline (Pro), arginine (Arg), and glutamic acid (Glu) were also frequently observed in allergenic peptides.

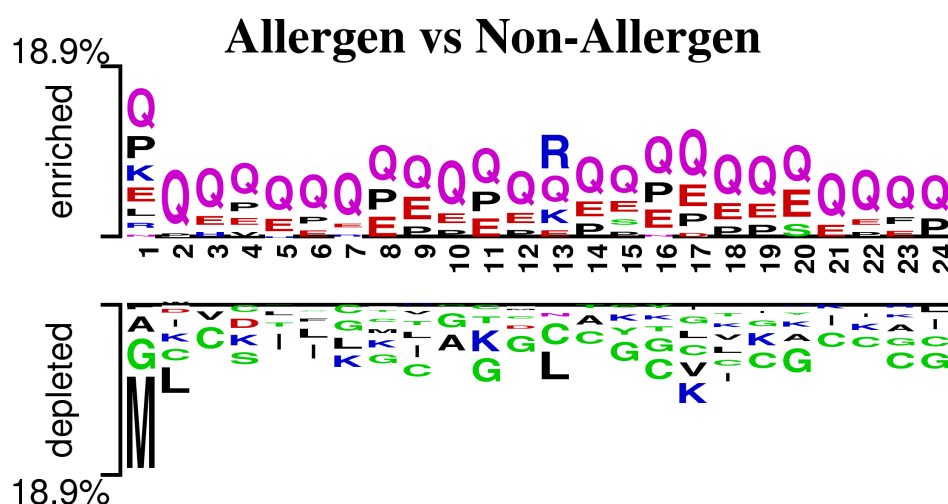


Figure 6.6 TSL plot showing positional amino acid preferences in allergenic and non-allergenic peptides across the N- and C-terminus

6.4.3 Preliminary Machine Learning Evaluation

To select the most suitable feature extraction and model construction methods from the numerous existing algorithms, and to reduce experimental complexity and workload, we performed a preliminary analysis on the dataset [130]. In this study, we first selected 16 types of features to build classification models, as shown in Figure 6.7. Among these 16 types of features, the predictive performance was evaluated, and it was observed that RF performs better when compared to other classifiers. In Figure 6.7, we present the performance of all features obtained using various classifiers. It was shown that DPC-based features achieved the highest AUROC of 0.80 on the validation dataset.



Figure 6.7 The AUROC of different feature types on different classifiers on the validation dataset

6.4.4 Multi-feature Fusion Results

We fused or combined the 16 feature types listed in Figure 5 for ALL_COMB in order to enhance the performance of our models. Multi-feature fusion techniques yield a feature vectors matrix with 1165 features. With an AUROC of 0.79 and an MCC of 0.41 on both training and validation datasets using ET, ALL_COMB features likewise demonstrated strong performance, as indicated in Table 6.1. Our analysis revealed that ET, RF, XGB, and LGBM performed nearly identically.

Table 6.1 Various classifiers performance on ALL_COMP features

Training Dataset						
Model	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUROC	F1	MCC
ET	71.48	70.04	70.76	0.80	0.71	0.42
RF	71.92	69.67	70.80	0.80	0.71	0.42
LGBM	70.30	70.15	70.22	0.78	0.70	0.40
XGB	69.79	69.44	69.61	0.78	0.70	0.39
LR	65.38	66.13	65.75	0.73	0.66	0.32
SVC	66.16	64.91	65.53	0.72	0.66	0.31
KN	63.51	62.07	62.79	0.70	0.63	0.26
MLP	59.61	66.36	63.00	0.70	0.62	0.26
DT	62.29	63.46	62.88	0.68	0.63	0.26
GNB	61.72	63.7	62.71	0.67	0.62	0.25
Validation Dataset						
ET	72.75	71.76	72.26	0.80	0.73	0.45
RF	71.54	68.76	70.16	0.79	0.71	0.40
LGBM	70.87	70.12	70.5	0.79	0.71	0.41
XGB	68.99	70.81	69.89	0.78	0.70	0.40
LR	62.69	69.30	65.97	0.73	0.65	0.32
SVC	66.71	69.03	67.86	0.73	0.68	0.36
KN	66.17	64.39	65.29	0.72	0.66	0.31
MLP	28.19	89.36	58.53	0.71	0.41	0.22
DT	63.36	63.44	63.40	0.68	0.64	0.27
GNB	61.88	64.67	63.26	0.67	0.63	0.27

6.4.5 Performance evaluation of selected feature descriptors

6.4.5.1 Feature selection on hybrid features

To determine the most crucial features from hybrid features for predicting allergen, we initially use the mRMR feature selection technique to extract feature set matrices of sizes 100, 200, 250, 300, 400, and 800, and evaluated using the minimum reliability and maximum significance approach, with model performance assessed using the ET, RF, XGB, and LGBM classifier as in mentioned above section. To further optimize the feature sets, we have performed feature reduction on the ALL_COMP features using two L1-regularized approaches: SVC-L1 and Lasso-L1. Using the SVCL1 with three different regularization strengths ($C = 0.001, 0.01, \text{ and } 0.1$), the feature set is reduced to 31, 273, and 652 features, respectively. Similarly, Lasso L1 regularization with λ values of 0.1, 0.01, and 0.001 selected 19, 143, and 527 features, respectively. Afterwards,

prediction models were developed using different ML algorithms. As shown in Figure 6.8A, the SVCL1 method with $C = 0.01$, resulting in a 273 feature subset, achieved the improved AUROC and MCC (0.80 and 0.45, respectively) while maintaining a balanced sensitivity and specificity on the RF model. Notably, this performance was comparable to that of the best-performing DPC features based models.

6.4.5.2 Feature Selection on DPC Features

As shown in Figure 6.7, the DPC descriptors (400 feature matrix) achieve the highest performance, with a maximum AUROC of 0.80 on the validation dataset. We further applied the above-mentioned feature selection techniques and calculated the performance on the top-performing ML classifiers, including ET, LGBM, RF, and XGB (shown in Figure 6.8B). Overall results showed that the predictive performance improves with an increase in the number of features. While aggressive feature reduction using strong LASSO-L1 regularization ($\lambda = 0.1$) or very low SVC-L1 regularization ($C = 0.001$) led to a notable decline in AUROC and MCC. In particular, feature sets selected using mRMR (200 to 300 features), LassoL1 ($\lambda = 0.001$), and SVC-L1 ($C = 0.01$) achieved the most balanced performance across best classifiers, with validation AUROC values ranging from ~ 0.79 to 0.80 and MCC values of up to ~ 0.41 – 0.43 .

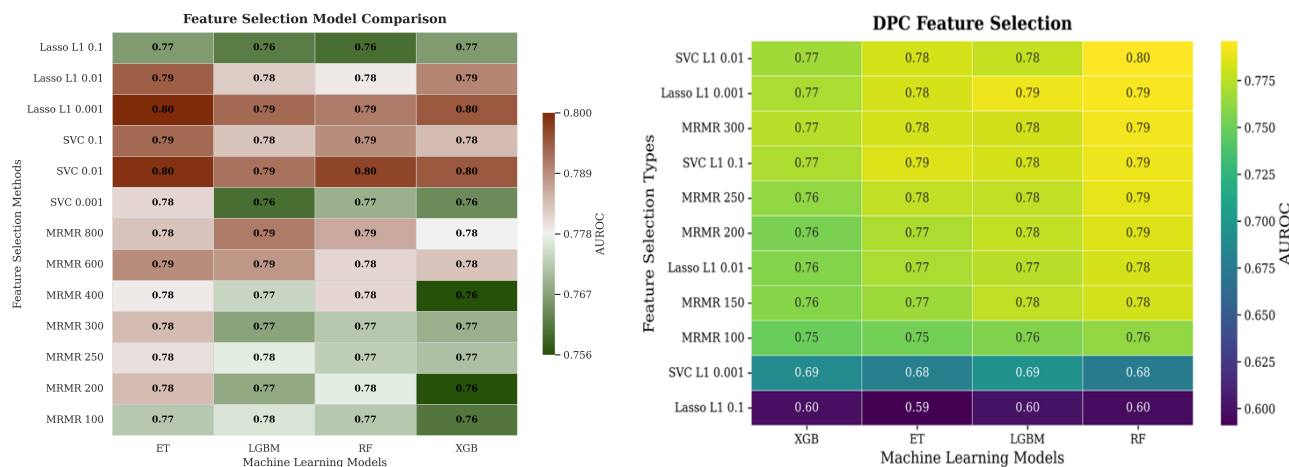


Figure 6.8 The top classifier performance on different sets of features (validation dataset) using the mRMR, LassoL1, and SVC-L1 feature selection (A) on hybrid features (B) DPC features.

6.4.6 Performance of binary profile-based models

The classification models have been developed using a binary profile of different patterns such as 4, 6, 8, 10, and 12, where a pattern is represented by a vector of dimensions $N \times 21$ (N is the length of the pattern). We computed the performances, and it was shown that N8C8 achieved the

maximum AUROC of 0.73 (see Figure 6.9). In the case of a binary profile using N(X), the AUROC was 0.75 on the validation dataset (see Figure 6.10).

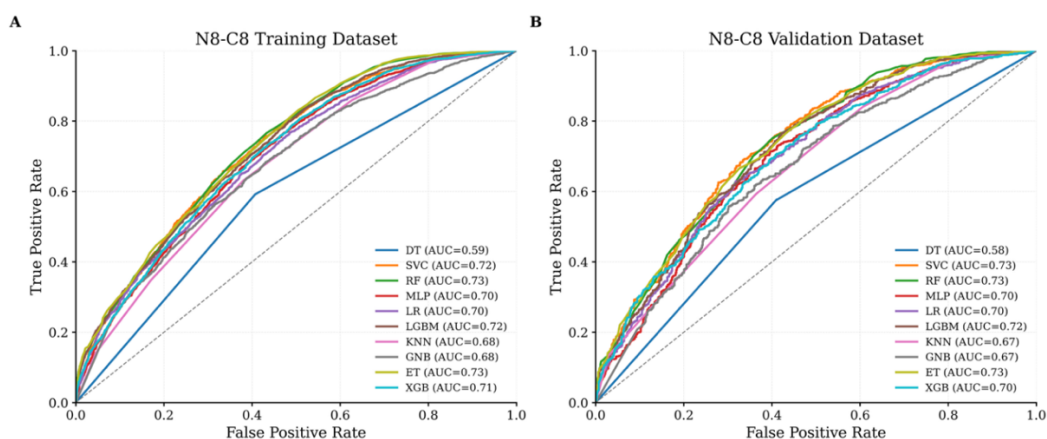


Figure 6.9 The AUROC curve showing the performance of N8-C8 sets of binary profile patterns

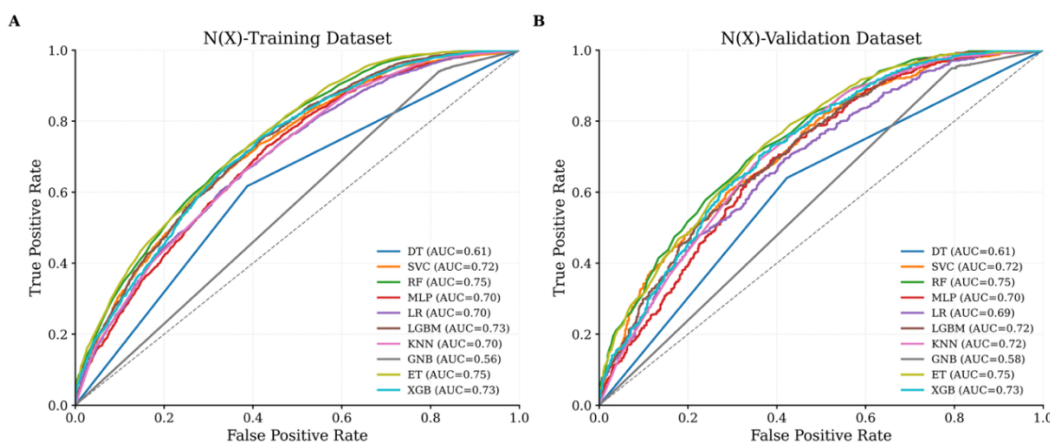


Figure 6.10 The AUROC curve showing the performance of N-terminal (X) padding of binary profile patterns

6.4.7 Performance on Large Language Models (LLMs)

6.4.7.1 Protein language models

Further, we have evaluated the performance of LLMs on the validation dataset by finetuning the model on the training dataset. As shown in Figure 6.11A, among the ESM models, ESM2-t33 achieved the best overall performance with a sensitivity of 68.45%, specificity of 67.25%, accuracy of 67.86%, and an AUROC of 0.76, corresponding to the highest MCC value of 0.36. PeptideBERT demonstrates a balanced behavior with a sensitivity of 67.38% and specificity of

61.53%, achieving an overall accuracy of 64.48%, an AUROC of 0.72, and an MCC of 0.29. ProtBert (epoch 3), on the other hand, showed the highest sensitivity (89.53%) but very low specificity (46.25%), leading to a balanced performance (Accuracy 68.06%, AUROC 0.74, MCC 0.40).

6.4.7.2 Embeddings from fine-tuned PLMs

To investigate whether fine-tuning improves predictive performance, we generated fixed-length protein representations by extracting embeddings from protein language models. Each protein sequence was represented as a vector obtained by averaging the embeddings of its individual residues [317]. We also evaluated embeddings obtained from a fine-tuned protein language model. Machine learning models were then trained using different classifiers, and their performance is presented in Figure 6.11B. Among the evaluated models, the Extra Trees (ET) classifier achieved the best performance, with a maximum AUROC of 0.76 using ESM2-t33 embeddings on the validation dataset. Although larger ESM-2 variants (e.g., 3B and 15B parameters) may provide richer representations, their substantially higher memory requirements and computational cost make them impractical for large-scale analyses.

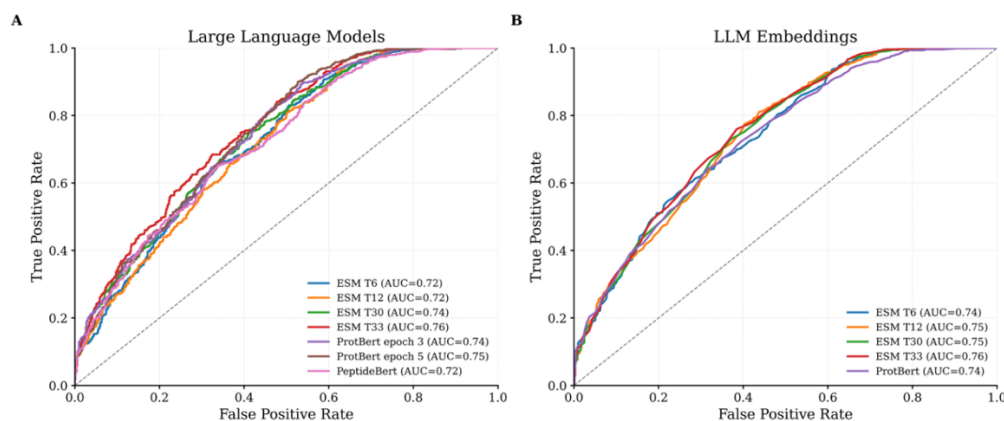


Figure 6.11 The AUROC curve performance of (A) LLMs, (B) ET classifier on the embeddings extracted using LLM on the validation dataset

6.4.8 Quantum SVM

Among quantum ML techniques, QSVM have attracted considerable attention as quantum-enhanced extensions of classical SVMs. By exploiting quantum kernel methods and quantum feature spaces, QSVMs aim to process high-dimensional data more efficiently and potentially improve classification performance. However, the practical benefits of these approaches remain dependent on advances in quantum hardware and continued improvements in algorithmic design

[318]. To identify effective predictive models for our classification task, AAC was used as a standardized input representation for both classical ML models and the QSVM framework. In our experiments, the RF classifier emerged as the most robust model, achieving the best validation performance with an AUROC of 0.77 and an MCC of 0.36. Performance of the QSVM improved with increasing qubit count. The best configuration used 14 qubits and gave an AUROC of 0.66 and an MCC of 0.31 on the validation set (Table 6.2), compared with AUROC 0.77 and MCC 0.36 for the RF classifier on the same AAC features. Classical ensemble methods therefore remain the stronger option for this task. The gap is consistent with the known limitations of present-day quantum hardware, namely restricted qubit availability, noisy qubits and high error rates. Whether quantum kernels can eventually exceed classical methods on problems of this type will depend on progress in kernel design and hardware scalability [319].

Table 6.2 Performance of comparison of QSVM with classical ML classifiers on composition-based features

	Model	Sens (%)	Spec (%)	Accuracy (%)	AUROC	F1	MCC
QSVM	Training	76.47	73.38	74.93	0.75	0.75	0.50
	Validation	68.34	63.06	65.70	0.66	0.67	0.31
RF	Training	69.96	71.53	70.75	0.79	0.71	0.42
	Validation	67.92	68.35	68.13	0.77	0.68	0.36
SVC	Training	66.77	65.69	66.23	0.72	0.66	0.33
	Validation	64.83	63.85	64.34	0.71	0.65	0.29

6.4.9 Motif Analysis

Comprehensive motif analysis was performed using the established MEME–MAST suite to identify signature sequence patterns associated with allergenic and non-allergenic peptides. Motifs were first discovered from the training dataset and subsequently scanned against the independent validation dataset to assess their occurrence and biological relevance. The top 10 motif hits, identified under minimum site thresholds of 10 and 5, are presented in Figures 6.12 A and 6.12 B, respectively. In addition, detailed motif coverage, along with the ensemble results obtained by combining MEME–MAST analysis with the best-performing SVC-L1 ($C = 0.01$) feature-selected machine learning prediction scores, is provided in Table 6.3.

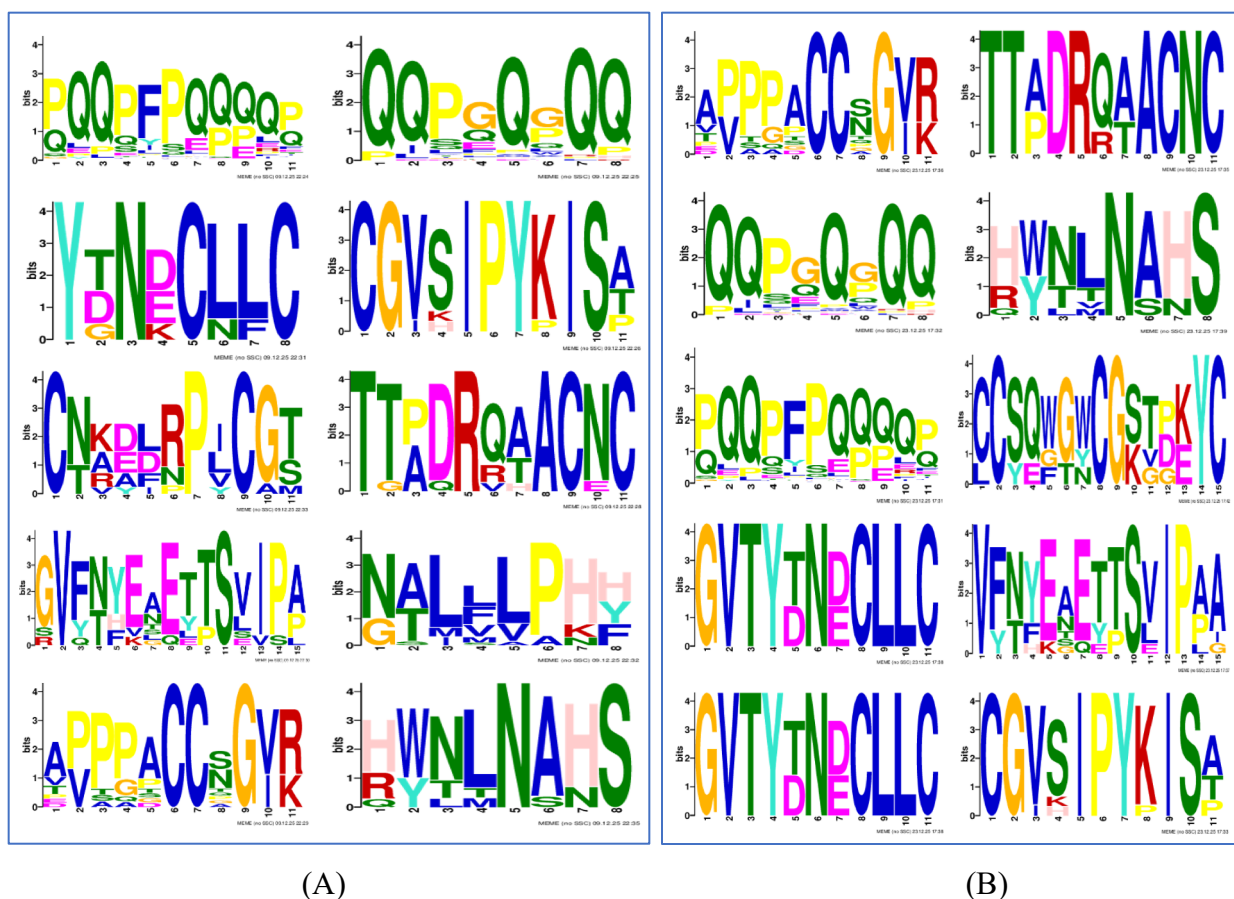


Figure 6.12 Ten discriminative motifs that were found in allergen (A) with minimum number of sites 10; (B) with minimum number of sites 5

Table 6.3. The performance of motifs on best performing ML score on validation dataset

MEME MAST	Number of Motifs	Number of sites	Sens (%)	Spec (%)	Acc (%)	AUROC	F1	MCC
	10	10	71.81	70.94	71.38	0.80	0.43	0.43
	10	5	71.54	70.67	71.11	0.79	0.71	0.42

6.4.10 Final Model

We observed that SVCL1 ($C = 0.01$) selects 273 features from the 1165 features of ALL_COMP, yielding the best performance with a smaller feature subset. Therefore, we have selected these features as a final prediction model. The complete performance was shown in Table 6.4. The ET classifier performs the best among different classifiers, achieving an AUROC of 0.81 and an MCC of 0.44 on the validation dataset with balanced sensitivity and specificity. Other classifiers, such as RF, XGB, and LGBM, also have similar performance. To improve the interpretability of the model, SHAP was performed and features were ranked according to their mean absolute SHAP

value. As shown in Figure 6.13, the most influential feature was a composition-enhanced transition-distribution descriptor (CeTD_100_p_PO1, mean |SHAP| = 0.182), followed by the amino-acid-composition features AAC_Q (glutamine, 0.164) and AAC_C (cysteine, 0.123), a distribution descriptor (DDR_M, 0.155), and additional CeTD descriptors (CeTD_100_p_SA1, CeTD_SA1, CeTD_100_p_HB3, CeTD_VW2).

Table 6.4. Performance of different classifiers on the SVCL1 (C 0.01) features

Model	Training Dataset				
	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUROC	MCC
ET	71.58	72.27	71.93	0.81	0.44
LGBM	71.72	71.23	71.47	0.8	0.43
XGB	70.5	70.72	70.61	0.79	0.41
RF	71.24	71.94	71.59	0.8	0.43
SVC	71.21	70.42	70.81	0.78	0.42
KN	67.82	68.96	68.39	0.76	0.37
MLP	68.53	68.25	68.39	0.76	0.37
LR	67.75	67.17	67.46	0.75	0.35
GNB	64.19	64.17	64.18	0.7	0.28
DT	64.73	62.51	63.62	0.68	0.27
Validation Dataset					
ET	73.56	70.4	71.99	0.81	0.44
LGBM	70.6	73.12	71.85	0.8	0.44
XGB	72.35	70.81	71.58	0.8	0.43
RF	71.68	69.58	70.64	0.8	0.41
SVC	71.68	71.76	71.72	0.79	0.43
KN	67.79	69.44	68.61	0.77	0.37
MLP	75.71	62.62	69.22	0.75	0.39
LR	65.77	70.26	68	0.74	0.36
GNB	66.04	63.17	64.61	0.7	0.29
DT	68.73	58.94	63.87	0.69	0.28

6.5 Benchmarking of available methods

We compared our method with available existing allergen prediction methods to benchmark performance on the validation dataset (Table 6.5). The pLM4Alg (2024) models demonstrated moderate performance compared with existing approaches, achieving accuracies between 66.31% to 67.79% and AUROC values of 0.73 to 0.75. Among these, pLM4Alg_1280 achieved an AUROC of 0.75 but remained below balanced performance. The available methods showed

weaker and less balanced performance. AllerTop v.2 (2014) and ProAll-D (2022) achieved an AUROC of 0.57 and 0.54, respectively with low MCC values (0.14 and 0.07). Meanwhile, AlgPred2 (2021) and AlgPred (2006) achieved a high specificity (83.89% and 88.19%) and low sensitivity (23.46% and 23.87%), with AUROC (0.54 and 0.52) and MCC (0.09 and 0.16). This underscores the need for more robust strategies for peptide prediction that demonstrate more balanced predictive performances.

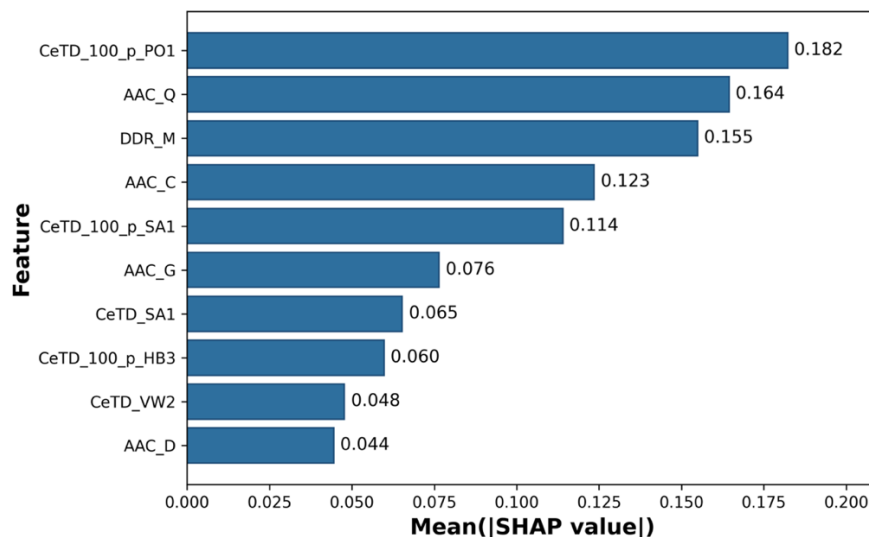


Figure 6.13 SHAP-based feature importance for the allergenicity prediction model.

Table 6.5. The comparison of available methods on the validation dataset

Method	Validation Dataset					
	Year	Sens (%)	Spec (%)	Acc (%)	AUROC	MCC
PLM4Alg_1280	2024	77.72	57.71	67.79	0.75	0.36
PLM4Alg_640	2024	69.93	65.48	67.73	0.74	0.35
PLM4Alg_480	2024	74.63	59.07	66.91	0.73	0.34
pLM4Alg_320	2024	70.20	62.35	66.31	0.73	0.33
AllerTop v.2	2014	60.00	53.47	56.76	0.57	0.14
ProAll-D	2022	63.62	43.66	53.72	0.54	0.07
AlgPred2	2021	83.89	23.46	53.92	0.54	0.09
AlgPred	2006	88.19	23.87	56.29	0.52	0.16

6.6 Limitations

Currently, available allergens prediction methods that is considered for safety evaluations majorly focuses on the whole protein based analysis, which present significant limitations. These methods do not explicitly evaluate discrete IgE-binding epitope or allergenic peptides. Therefore, an efficient machine learning method, based on identifying a set of allergenic peptides is the need of the hour. In this study we have a prediction method for the identification of allergenic and non-allergenic peptides. However, this study has certain limitations. First, the negative dataset was less rigorously curated than the positive dataset. Although, a large portion of non-allergenic peptides were obtained from experimentally validated negative epitopes and previously published datasets, a small portion of the negative set consisted of Swiss-Prot derived peptides that were assumed non-allergenic based on the absence of allergen annotations. Second, potential sequence overlap between the training datasets of previously published allergen prediction methods and our independent validation. We did not perform a rigorous assessment of data overlap for the existing methods, primarily because their performance was already substantially lower than that of our method. Finally, the study lacks of a completely independent external validation dataset derived from separate experimental studies or platforms, primarily due to the limited availability of well-characterized allergenic and non-allergenic peptide data. Future studies incorporating more rigorously curated negative datasets and independent external validation cohorts will further strengthen the robustness and generalizability of the proposed method.

6.7 Web Server Implementation

A web based platform named “AlgPred3” was developed to provide user-friendly access to the prediction models and enable researchers to analyze peptide sequences without local installation (available at <https://webs.iiitd.edu.in/raghava/algpred3/>). The web server provides four modules: (i) Prediction, (ii) Design, (iii) Protein Scan and (iv) Motif Scan. The Prediction module accepts peptide sequences in FASTA format and classifies each as allergenic or non-allergenic peptides. The Design module generates all single-residue mutants of a query peptide and returns a prediction score for each, allowing users to identify substitutions that reduce allergenic potential. The Protein Scan module locates allergenic regions within a full-length protein, and the Motif Scan module detects allergenic-associated motifs. The server uses a responsive HTML front end and runs on

standard desktop and mobile browsers. Standalone and pip-installable versions are distributed through GitHub and PyPI.

6.8 Discussion and conclusion

Traditional experimental approaches for identifying allergenic proteins and mapping their IgE-binding epitopes require substantial time, cost and laboratory effort. These constraints restrict the number of candidates that can be assessed and therefore slow the development of immunological diagnostics and targeted therapeutics [130,320]. The computational tools play a crucial role in achieving these objectives by enabling rapid and cost-effective allergen prediction. Previously, developed computation tools not accessible and not specifically designed for peptides prediction. In the present study, we developed AlgPred3, an advanced computational framework for the prediction of allergenic peptides. The allergens dataset was collected from several databases and prediction methods. For the non-allergens dataset, we have generated random peptides of the same length using the Swiss-Prot database. Most epitopes fall within a narrow range, typically between 8 and 20 amino acids, indicating a strong concentration of data points in this region. Firstly, we analyzed the amino acid composition of allergens and non-allergens. Among all the amino acids Glu, Pro, Arg and Gln were more abundant in allergens. The positional analysis reveals that amino acid 'Pro', 'Arg', and 'Glu', highly conserved in the allergens. We have used composition-based features, binary profile patterns, LLM and LLM embeddings. We further use mRMR, Lasso-L1 and SVC-L1 method of feature selection. After that various prediction models using different ML techniques such as DT, ET, KNN, LR, GNB, XGB, MLP, LGBM, RF, and SVC was developed. It was shown that ET classifier model with SVC-L1 ($C=0.01$) composition consistently outperformed others, achieving the AUROC of 0.81 and MCC of 0.44 on an independent validation dataset. To facilitate accessibility and reproducibility, we have provided AlgPred3 as a freely available web server (<https://webs.iiitd.edu.in/raghava/algpred3/>), along with standalone, PyPI, and GitHub packages. AlgPred3 represents a significant advancement offering improved predictive applicability for peptide allergens. We anticipate that it will serve as a valuable resource for new research opportunities in fields such as therapeutic development, vaccine design, protein engineering, and other biotechnological applications, thereby contributing to safer product development and better management of allergic diseases.

C H A P T E R

7

Prediction and Design of
Thermostable Proteins with
desired T_m

7.1 Introduction

Biotechnological research and development frequently requires the identification of proteins capable of maintaining their structural integrity under high-temperature conditions [154]. Proteins are extremely versatile biomolecules that are necessary for all living things. They control a wide range of biological processes, and their ability to function is primarily dependent on how stable they are at room temperature. The melting temperature (T_m), which is the temperature at which a protein loses its native conformation, is frequently used to describe the thermal stability of proteins. Thermostable proteins are widely utilized in medical research and clinical applications, where they serve as stable scaffolds for drug delivery systems, enzyme engineering, diagnostic assays, and therapeutic interventions. Therefore, rational protein design requires an understanding of the factors governing protein stability [141,321]. Additionally, thermostable proteins are widely used in pharmaceutical development, industrial production, and medical treatments. [23,140]. However, accurate determination of protein T_m in laboratory settings requires complex experimental procedures, including protein expression, purification, and specialized instrumentation. Examples of frequently used experimental methods include Fourier Transform Infrared (FTIR) spectroscopy, Thermal Proteome Profiling (TPP) based on Mass Spectrometry, Differential Scanning Calorimetry (DSC), Circular Dichroism (CD) spectroscopy, and Differential Scanning Fluorimetry (DSF) [140]. Owing to their high cost and technical complexity, these methods are impractical for genome-scale screening of protein thermostability. Consequently, computational approaches that predict protein thermostability directly from sequence information are highly desirable. Several computational tools have previously been proposed to predict protein T_m , including DeepTM, DeepSTABp, ProtStab2, iStable2, TemStaPro, and ProLaTherm. However, most of these models are trained on redundant datasets (Table 2.6) and may fail when applied to proteins with low similarity to training sequences. Thus, it is necessary to use standardized bioinformatics protocols to develop predictive methods based on non-redundant protein datasets.

7.2 Overview of the study

The study's overall workflow is shown in Figure 7.1, and the subsequent subsections offer detailed descriptions. DeepSTABp, a deep learning-based method created to forecast protein thermal stability, provided the dataset used in this study [147]. To reduce sequence redundancy, CD-HIT

clustering was applied at a 40% sequence identity threshold (described in Section 3.3). The resultant dataset was used to train, test, and evaluate models built with deep learning, large language models, and different machine learning techniques. Protein features were generated using the Pfeature software as well as embeddings derived from protein language models. Following model development, 10-fold cross-validation was used to evaluate performance, and an independent validation dataset was used to evaluate the final model that was chosen. A standalone software (<https://webs.iiitd.edu.in/raghava/pptstab/stand.php>) and web-based platform PPTStab (<https://webs.iiitd.edu.in/raghava/pptstab/>) have been developed to make it easier for researchers to predict protein melting temperature (T_m) and design proteins with optimum thermal stability.

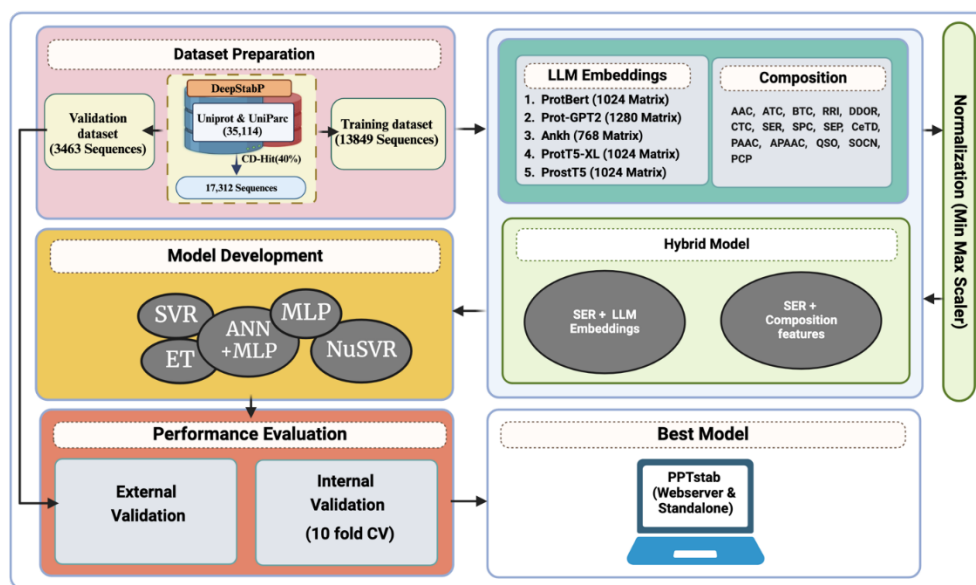


Figure 7.1 The overall architecture of the method (adopted from [143])

7.3 Material and Methods

7.3.1 Data collection and Preparation

The DeepSTABp dataset, which is primarily based on the Meltome Atlas [205], yielded 35,114 protein entries in total. The Meltome Atlas, which collects TPP experiments conducted on a variety of organisms, includes studies involving heat treatment of both intact cells and protein lysates [140]. The corresponding protein sequences were obtained from the UniProt and UniParc databases. The dataset covers a wide range of taxa, including *Bacillus subtilis*, *Oleispira antarctica*, *Thermus thermophilus*, *Pyrococcus torridus*, *Geobacillus stearothermophilus*, *Mus musculus*, *Homo sapiens* (K562 and hepatocytes), *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Danio rerio*, *Arabidopsis thaliana*, *Drosophila melanogaster*, and *Escherichia coli*[140,205]. Protein

sequences with non-standard residues (U, Z, O, B, J, or X) or longer than 2,500 amino acids were not included. With a 40% sequence identity cutoff based on its greedy incremental clustering algorithm, CD-HIT clustering [206] was used to reduce redundancy, guaranteeing that no pair of sequences shared more than 40% similarity. 17,312 non-redundant protein sequences were left over after preprocessing. The dataset was split 80:20 between training and validation sets to build and evaluate the model; 3,463 sequences were used for validation and 13,849 sequences for training.

7.3.2 Features engineering

7.3.2.1 Protein composition descriptors

The composition module of the Pfeature software was utilized to transform protein sequences into numerical feature vectors, as explained in *Section 3.7.1*. Pfeature is a comprehensive library for calculating numerical features from peptide and protein sequences and their structures. It provides five primary modules for feature computation: composition-based descriptors, binary profile-based features, structure-based features, evolutionary information-based descriptors, and pattern-based descriptors. In the present study, only composition-based features were utilized. A total of 15 descriptor types were calculated, including AAC, RRI, DDOR, ATC, BTC, PCP, CTC, CeTD, SEP, SER, SPC, QSO, SOCN, PAAC, and APAAC. The dimensionality of the resulting feature vectors is summarized in Table 3.3.

7.3.2.2 Protein Language Model Embeddings

The embeddings of sequence were generated using ProtBERT, a BERT architecture based pretrained PLM [257]. From ProtBERT's last hidden layers, 1024-dimensional vectors were extracted to represent each protein sequence, commonly referred to as embeddings. In contrast to the initial BERT model, that incorporates next-sentence prediction, ProtBERT considers every protein sequence to be a complete document. The model employs a masked language modeling strategy in which up to 15% of amino acids are randomly masked during training. Protein sequences were tokenized as uppercase amino acid characters separated by spaces. The vocabulary contained 20 tokens that represented the 20 standard amino acids (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, W, Y, and V) and excluded non-natural amino acids. For optimization, the LAMB (Layer-wise Adaptive Moments) optimizer was used with a learning rate of 0.002 and a weight decay of 0.01. This procedure generated a 1024-dimensional embedding vector for each sequence,

placing proteins in a high-dimensional feature space that captures contextual relationships among amino acid residues [257]. Additionally, embeddings were also extracted using other protein language models, including Ankh, ProtGPT2, ProsT5, ProtT5-XL-UniRef50, and ProtBERT (describe in Section 3.9.2).

7.3.3 Normalization

A critical preprocessing step that rescales calculated features to a common numerical range is normalization. This transformation improves model stability, speeds up convergence, and enhances predictive performance by preventing variables with large magnitudes from disproportionately influencing the learning process. In this study, the Min–Max scaling method was used to map feature values to a fixed interval of either $[0, 1]$ or $[-1, 1]$, depending on the requirements of the learning algorithm [322]. The normalization procedure, defined in Equation 15, ensures that all input attributes contribute comparably during model training and aids in identifying meaningful patterns. The Min–Max scaler was used to first normalize the composition-based descriptors in the training dataset. The independent validation dataset was then subjected to the same scaling parameters. By using the same transformation parameters, the suggested method's overall predictive performance is enhanced, consistency across datasets is maintained, and model generalization is supported.

$$z' = \frac{v - v_{min}}{v_{max} - v_{min}} \quad [15]$$

where z is the normalized value, v represents the original feature value, and V_{min} and V_{max} denote the minimum and maximum values of that feature, respectively.

7.3.4 Regression Model Development

To construct a reliable framework, several types of regression algorithms were evaluated. For this purpose, the scalable automated machine-learning platform H2O AutoML (version 3.46.0.4) was employed [323]. Standard performance metrics, such as the PCC and the R^2 , were used to evaluate all available regression algorithms after they had been trained on the training dataset. The six top-performing techniques SVR, ET, NuSVR, LightGBM, CatBoost, and MLP, were chosen for further investigation based on their performance. While the LGBM model [324,325] were implemented in Python using the scikit-learn library. Initially, algorithms that had significant

predictive potential were found using default hyperparameters. Additionally, a custom ANN was made using TensorFlow. Using the ReLU activation function, the network had four fully connected hidden layers with 256, 128, 64, and 32 neurons each, followed by a single output layer [326]. The model was trained using mean squared error (MSE) as the loss function and then optimized using the Adam optimizer to increase convergence and decrease overfitting [327]. A unified predictive framework was then created by combining the ANN with an MLP regressor (Figure 7.2). The outputs of both models were averaged to produce final predictions, which led to an ANN+MLP ensemble intended to enhance overall prediction performance.

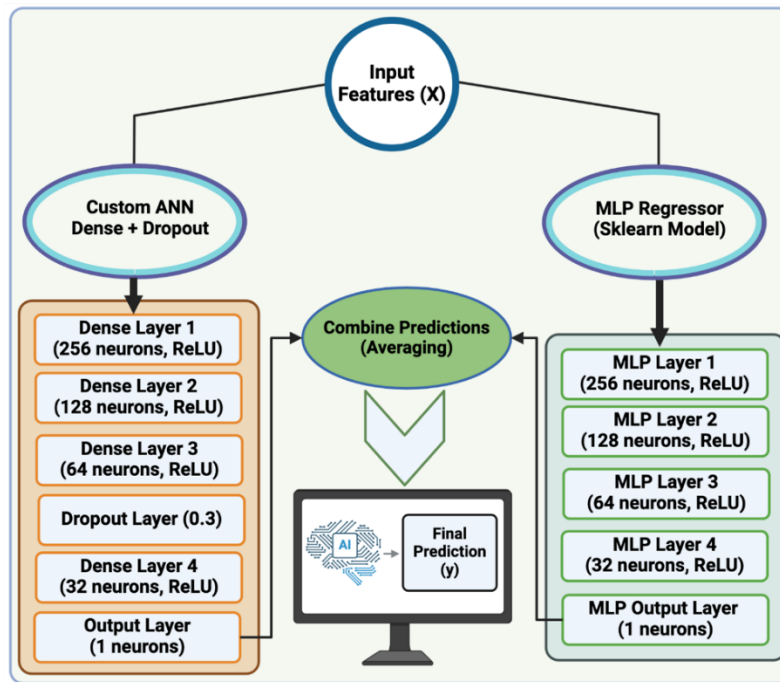


Figure 7.2. Architectural diagram of ANN combined with MLP model (adopted from [143])

7.3.5 Model evaluation

Following the procedure described in *Section 3.12*, ten-fold cross-validation ($k = 10$) and an independent validation dataset were used to assess the model's performance. Using the standard regression metrics described in *Section 3.13.2*, such as the PCC, MSE, MAE, RMSE, and R2, predictive accuracy was measured. The training and validation datasets were used to compute these metrics.

7.4 Results

7.4.1 Protein Composition Analysis

Previous study explored the relationship between the thermal stability of globular proteins and their amino acid composition [328]. We separated proteins into two groups according to their melting temperature (T_m): $T_m > 50\text{ }^{\circ}\text{C}$ and $T_m < 50\text{ }^{\circ}\text{C}$ in order to examine this relationship in our dataset. Figure 7.3 shows the results of applying Equation 1 (*Section 3.5*) to calculate the mean residue composition for each amino acid for both groups. Higher frequencies of leucine (L), alanine (A), glycine (G), and glutamic acid (E) were found in thermostable proteins. However, proteins with lower melting temperatures ($T_m < 50\text{ }^{\circ}\text{C}$) had higher amounts of serine (S), lysine (K), glutamine (Q), and histidine (H). These compositional differences can be explained by the physicochemical roles of the residues. The enrichment of glutamic acid (E) in the more thermostable group is consistent with the well-documented contribution of charged residues to thermal stability: glutamate forms salt bridges and ion-pair networks that stabilise the folded state and raise the melting temperature, an effect shown to be more prevalent in thermophilic proteins [329]. Leucine (L) and alanine (A) act mainly through the hydrophobic effect and tight packing, leucine strengthens the hydrophobic core, while alanine, a strong α -helix former with a small side chain, stabilises secondary structure and allows dense packing. Notably, both glutamate and leucine belong to the IVYWREL set (Ile, Val, Tyr, Trp, Arg, Glu, Leu), whose combined abundance correlates strongly with the optimal growth temperature of organisms, which supports our observation [330].

7.4.2 Correlation of T_m with Descriptors

We initially calculated the percentage composition for selected descriptors-including AAC, SER, PAAC, and APAAC that demonstrated the best predictive performance (Table 7.2). Subsequently, the correlation between residue composition and T_m was determined for each amino acid type, as illustrated in Figure 7.4. The amino acids exhibiting the strongest positive and negative correlations with T_m are summarized in Table 7.1, which presents the top correlations across the different composition-based descriptors.

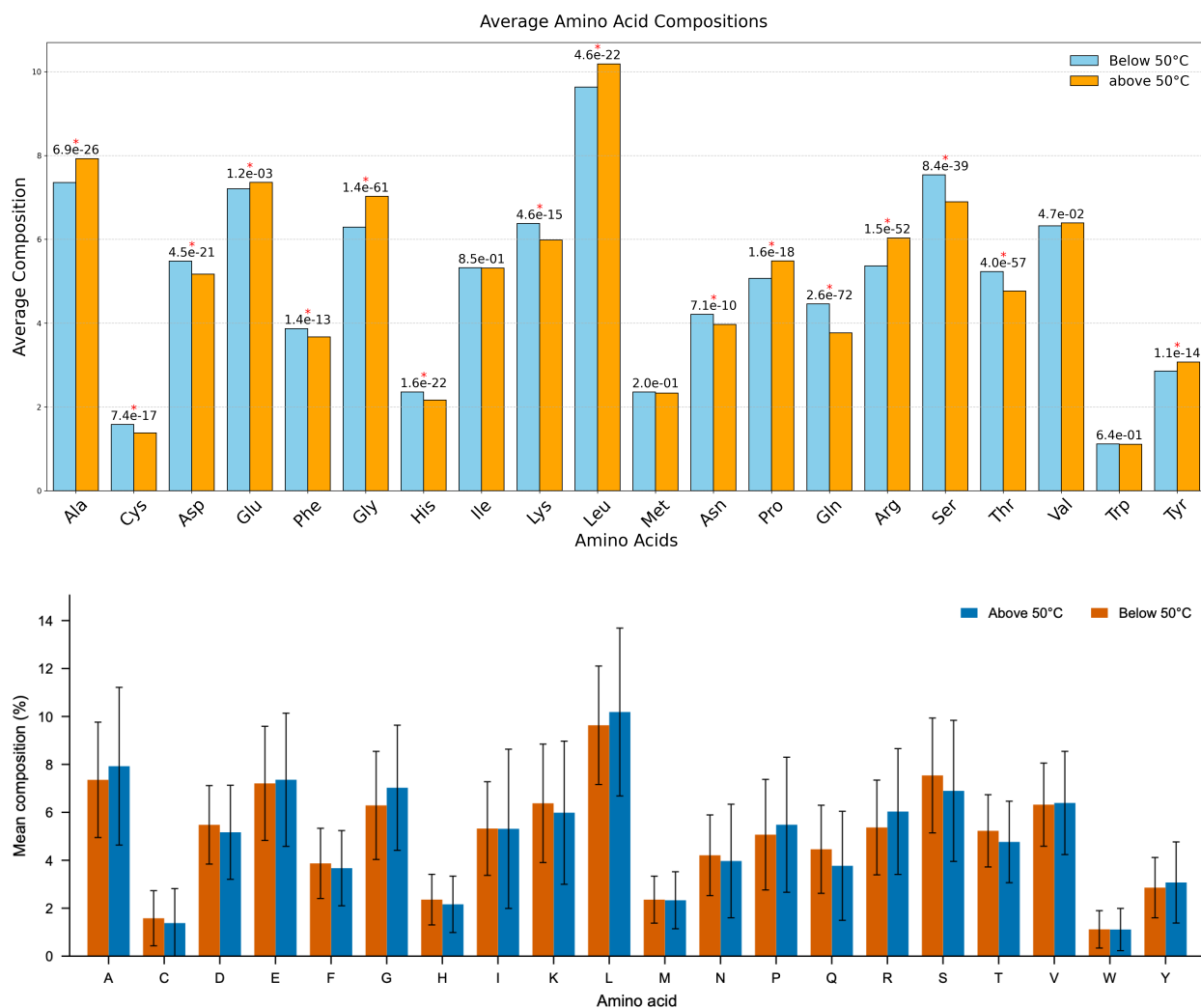


Figure 7.3. Average percent amino acid composition of thermostable ($T_m > 50^\circ\text{C}$ and $T_m < 50^\circ\text{C}$) proteins

Table 7.1 Top positively and negatively correlation features across descriptor (adopted from [143])

Descriptor	Positive Correlation		Negative Correlation	
	Feature	Correlation	Feature	Correlation
AAC	R (Arg)	0.24	Q (Gln)	-0.25
	L (Leu)	0.21	T (Thr)	-0.22
SER	S (Ser)	0.25	R (Arg)	-0.22
	Q (Gln)	0.30	G (Gly)	-0.19
PAAC	lam1	0.26	Q (Gln)	-0.25
	R (Arg)	0.24	T (Thr)	-0.22
APAAC	R (Arg)	0.24	Q (Gln)	-0.25
	L (Leu)	0.21	T (Thr)	-0.22

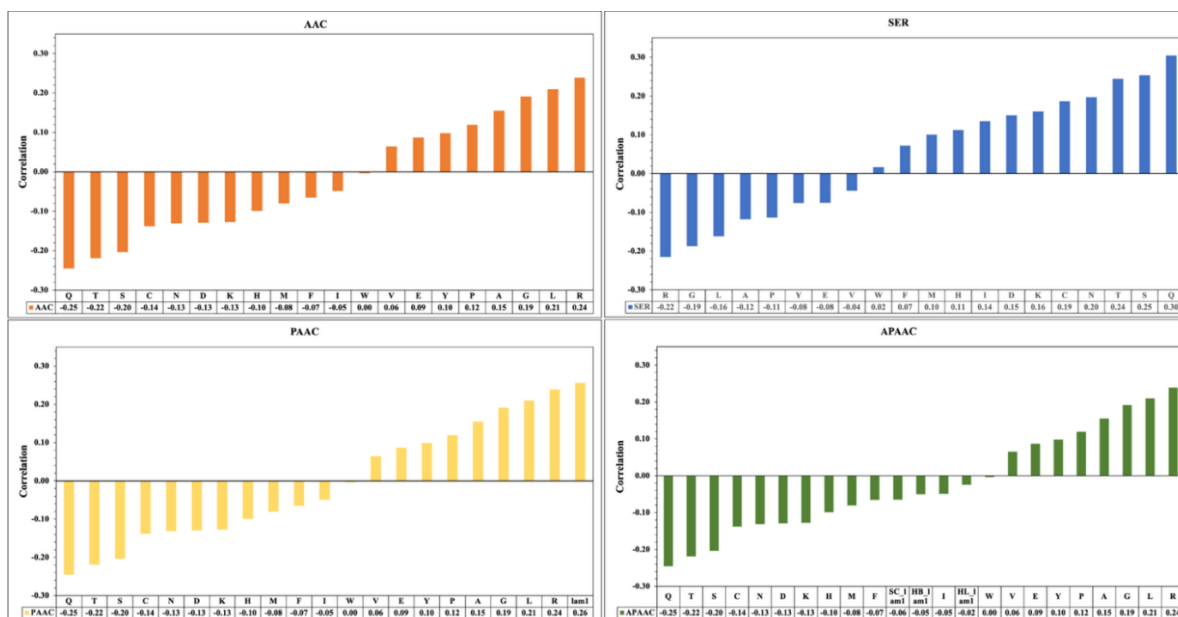


Figure 7.4. The correlation between the percent composition of AAC, SER, PAAC, and APAAC residues and T_m (adopted from [143])

7.4.3 Development of Prediction Models

After developing multiple regression models using various composition-based feature sets, the ANN+MLP ensemble model demonstrated superior performance compared with all other models (Figures 7.5 and 7.6). Among the 15 evaluated descriptors, SER, APAAC, PAAC, and AAC contributed most significantly to predictive performance. Furthermore, on the validation dataset, the SER descriptor achieved the best results, with a PCC of 0.80 and a R^2 value of 0.63.

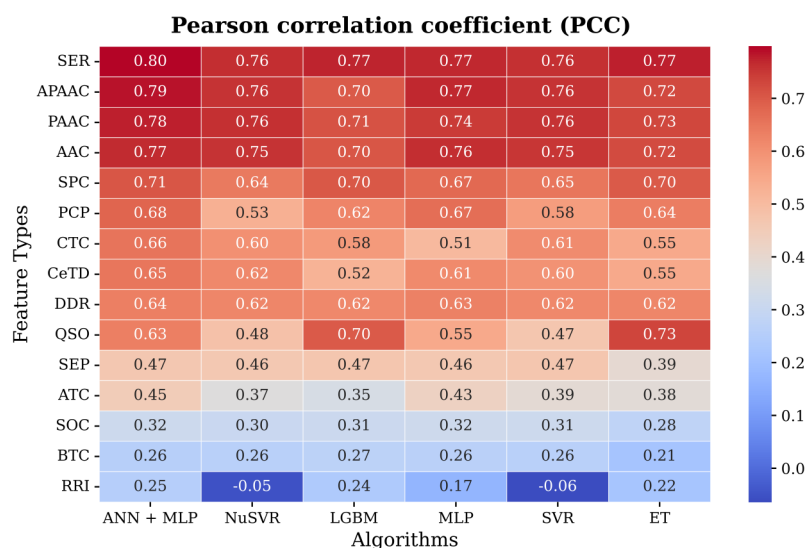


Figure 7.5. The PCC of different regressors on different feature type (Validation dataset)

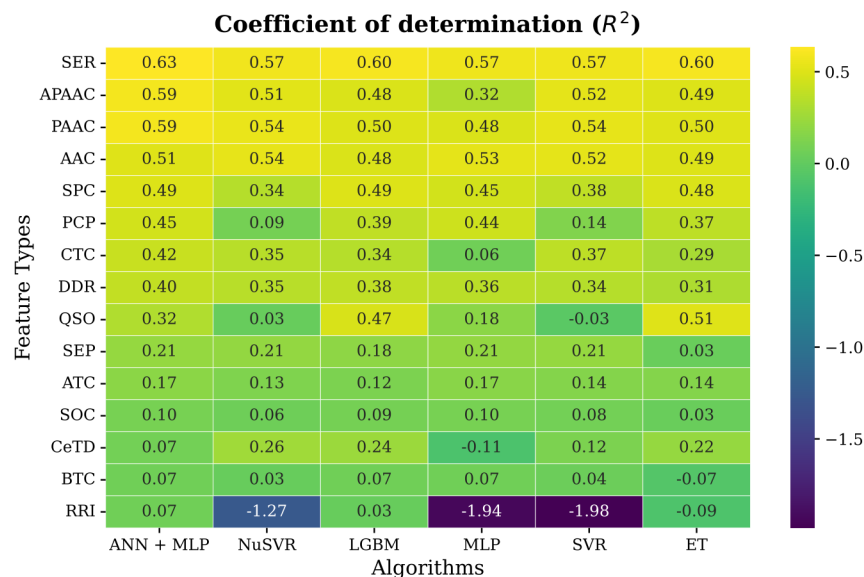


Figure 7.6. The R^2 of different regressors on different feature type (Validation dataset)

7.4.3.1 Composition-Based Model Performance

We developed a prediction framework based on these findings, with the main input feature being the Shannon entropy of residues (SER) composition. A 20-dimensional vector representing the standard amino acids was used to encode each protein sequence. Additionally, a categorical variable was added to indicate whether the measurements were taken under “lysate” or “cell” experimental conditions. These SER descriptors were used to train multiple regression algorithms that estimate protein melting temperature (T_m). The melting temperature was transformed to a 0-1 scale as a preprocessing step, so that all error metrics are expressed as a percentage of the temperature range; ten-fold cross-validation was used to assess model performance. The outcomes are shown in Table 7.2. The ensemble ANN + MLP model performed best, achieving an R^2 of 0.63 and a PCC of 0.80 on the validation dataset, with a mean absolute error of 7.01%, corresponding to a percentage error of approximately 7% (Table 7.2). The remaining regressors gave comparable results (MAE 7.07-7.50%, R^2 0.57-0.60), indicating that performance was limited by the feature representation rather than by the choice of algorithm. Overall, these values show that the SER feature set by itself did not offer optimal predictive accuracy, which motivated the use of protein language model embeddings.

Table 7.2. Performance of different regressor models on SER features

Model	Training Dataset						Validation Dataset					
	RMSE (°C)	MSE (°C ²)	MAE (°C)	PCC	R ²	MAPE	RMSE (°C)	MSE (°C ²)	MAE (°C)	PCC	R ²	MAPE
ANN + MLP	5.50	30.25	4.15	0.80	0.64	8.86	5.62	31.55	4.22	0.80	0.63	8.95
ET	5.50	30.29	4.10	0.80	0.64	8.74	5.79	33.57	4.29	0.77	0.59	9.11
LGBM	5.64	31.81	4.19	0.79	0.62	8.90	5.87	34.46	4.33	0.76	0.58	9.11
MLP	5.81	33.81	4.39	0.77	0.60	9.30	5.91	34.96	4.45	0.77	0.58	9.23
NuSVR	5.67	32.10	4.23	0.79	0.62	8.89	5.96	35.56	4.41	0.76	0.57	9.24
SVR	5.69	32.42	4.33	0.78	0.61	9.17	5.87	34.51	4.41	0.77	0.58	9.27

RMSE, MSE, and MAE are reported in actual temperature units (°C and °C²). MAPE (%) is the mean absolute percentage error, i.e. the average absolute prediction error expressed as a percentage of the observed T_m. PCC and R² are unitless.

7.4.3.2 Protein Language Model Embedding Performance

A prediction framework was developed using pretrained protein language model ProtBERT as a fixed feature encoder without fine-tuning to obtain sequence representations [331]. Regression models for T_m prediction were trained using a complete feature set of 1,026 features that were created by combining these embeddings with categorical indicators that represented “lysate” or “cell” conditions. Embeddings generated from several additional PLMs (Ankh, ProstT5, ProtT5-XL-UniRef50 and ProtGPT2) were also evaluated without fine-tuning, and various regression algorithms were trained on each representation. ProtBERT and ProtT5-XL-UniRef50 embeddings performed comparably and both clearly outperformed Ankh, ProstT5 and ProtGPT2. The best overall performance was obtained with the ProtBERT-based ANN+MLP ensemble model, achieving a PCC of 0.89 and an R² of 0.80 on the validation dataset. This model also showed the lowest prediction error, with an MAE of 5.00%, RMSE of 6.85% and MSE of 46.92 %², corresponding to a percentage error of approximately 5% (Table 7.3). Compared with the SER-based models (Section 7.4.3.1), which achieved a best R² of 0.63 and a percentage error of ~7%, the language model embeddings substantially reduced prediction error and improved explained variance, demonstrating the enhanced robustness and generalizability of the proposed approach.

Table 7.3. Performance of regression models using different LLM embeddings

Model	Training dataset					Validation dataset				
	RMSE (°C)	MSE (°C ²)	MAE (°C)	PCC	R ²	RMSE (°C)	MSE (°C ²)	MAE (°C)	PCC	R ²
ProtBert										
ANN+MLP	3.97	15.78	3.00	0.90	0.81	4.11	16.89	3.00	0.89	0.80
CatBoost	4.46	19.87	3.32	0.87	0.76	4.58	20.96	3.34	0.86	0.75
NuSVR	4.6	21.18	3.38	0.86	0.75	4.67	21.79	3.36	0.86	0.74
Ankh										
ANN+MLP	4.93	24.32	3.61	0.84	0.71	5.03	25.28	3.61	0.83	0.69
CatBoost	5.88	34.57	4.35	0.77	0.59	5.9	34.79	4.3	0.76	0.58
NuSVR	5.88	34.57	4.38	0.77	0.59	5.89	34.72	4.36	0.76	0.58
ProtT5										
ANN+MLP	5.33	28.39	3.88	0.81	0.66	5.24	27.44	3.77	0.82	0.67
CatBoost	6.58	43.32	4.83	0.69	0.48	6.62	43.80	4.79	0.69	0.47
NuSVR	6.1	37.23	4.54	0.74	0.55	6.12	37.45	4.5	0.74	0.55
ProtT5-XL										
ANN+MLP	4.07	16.55	3.07	0.89	0.8	4.12	16.99	3.09	0.89	0.79
CatBoost	4.45	19.82	3.32	0.87	0.76	4.49	20.14	3.35	0.87	0.76
NuSVR	4.39	19.29	3.33	0.88	0.77	4.36	19.03	3.23	0.88	0.77
ProtGPT2										
ANN+MLP	4.94	24.38	3.61	0.84	0.70	5.25	27.56	3.64	0.82	0.67
CatBoost	5.44	29.62	3.99	0.80	0.65	5.54	30.67	4.00	0.79	0.63
NuSVR	5.65	31.95	4.15	0.79	0.62	5.87	34.43	4.23	0.77	0.58

7.4.3.3 Hybrid Feature Approaches

In this approach, models were developed using multiple combinations of input features. The analysis revealed that ProtBERT embeddings outperformed all composition-based feature representations. SER performed the best among the composition-based descriptors, with an R² value of 0.63 and a PCC of 0.80 (Table 7.2). Other feature sets, such as AAC, PAAC, and ProtBERT embeddings, were combined with SER descriptors to improve prediction accuracy further. A combined feature matrix consisting of 1,046 attributes-comprising SER features, ProtBERT embeddings, and the “lysate” or “cell” indicator, was used to train several regression models. However, the ANN+MLP ensemble did not show improvement over previous approaches, yielding a PCC of 0.89 and an R² of 0.79 (Table 7.4). It was found that combining SER features with ProtBERT embeddings did not improve performance when compared to using ProtBERT embeddings alone after evaluating various feature combinations. As a result, ProtBERT embeddings are the sole basis for the final model chosen for this investigation and integrated into the web server.

Table 7.4. Performance of regression models using SER features with different features

Model	Training dataset					Validation dataset				
	RMSE (°C)	MSE (°C ²)	MAE (°C)	PCC	R ²	RMSE (°C)	MSE (°C ²)	MAE (°C)	PCC	R ²
A. SER with ProtBert embeddings										
ANN + MLP	3.99	15.92	3.02	0.90	0.80	4.05	16.4	2.97	0.89	0.79
NuSVR	4.6	21.18	3.38	0.86	0.75	4.84	23.39	3.49	0.86	0.72
LGBM	4.51	20.3	3.32	0.87	0.76	4.6	21.18	3.35	0.86	0.74
MLP	4.34	18.82	3.29	0.88	0.77	4.51	20.3	3.37	0.88	0.76
SVR	4.67	21.79	3.53	0.86	0.74	4.85	23.56	3.6	0.85	0.72
ET	4.48	20.03	3.3	0.87	0.76	4.55	20.74	3.31	0.87	0.75
B. SER with AAC descriptor										
ANN + MLP	5.47	29.88	4.1	0.8	0.64	5.75	33.11	4.32	0.79	0.6
NuSVR	5.73	32.83	4.27	0.78	0.61	6.07	36.8	4.45	0.76	0.56
LGBM	5.6	31.4	4.16	0.79	0.62	6.32	39.92	4.57	0.72	0.52
MLP	5.77	33.25	4.33	0.78	0.6	7.91	62.63	6.52	0.76	0.24
SVR	5.77	33.25	4.37	0.78	0.6	6.01	36.07	4.48	0.76	0.56
ET	5.51	30.4	4.09	0.8	0.63	6.05	36.58	4.51	0.76	0.56
C. SER with PAAC descriptor										
ANN + MLP	5.42	29.35	4.08	0.81	0.65	5.72	32.76	4.31	0.79	0.6
NuSVR	5.72	32.7	4.26	0.78	0.61	6.01	36.07	4.41	0.76	0.56
LGBM	5.59	31.27	4.16	0.79	0.62	6.13	37.6	4.47	0.74	0.55
MLP	5.79	33.52	4.37	0.78	0.6	7.11	50.55	5.75	0.76	0.39
SVR	5.76	33.18	4.36	0.78	0.6	5.98	35.78	4.47	0.76	0.57
ET	5.48	30.07	4.07	0.8	0.64	6.01	36.07	4.49	0.76	0.56

RMSE, MAE and MSE are in degrees Celsius (MSE in °C² = RMSE²). PCC and R² are unitless.

7.4.4 Comparison with Existing Methods

Since several of the previously published methods listed in Table 2.6 are classification-based rather than regression-based predictors, it was not possible to compare our approach with all of them. Therefore, comparisons were limited to existing regression-based methods. Among the available tools, only the DeepSTABp web server was operational, while others were either unavailable or non-functional. The majority of DeepSTABp's dataset is derived from the Meltome Atlas, which compiles TPP assays across various organisms, including studies involving heat treatment of cells and protein lysates. ProtStab2 is an updated version of ProtStab that integrates data from the Meltome Atlas. It seems that many of existing approaches are inconsistent with standard bioinformatics guidelines, especially when it comes to removing redundant sequences. In contrast, our model was developed using a carefully curated non-redundant dataset, making direct comparison challenging, as most existing predictors were trained on redundant data. Our approach performed similarly to the best-performing model, DeepSTABp, in terms of R² and

PCC as documented in the literature, even though we used more rigorous dataset preparation criteria. As shown in Table 7.5, our method performs similarly to DeepSTABp and better than earlier methods. Furthermore, our model was trained using 80% of the dataset and evaluated on 20% external validation data, whereas DeepSTABp used a 90% training and 10% validation split. Additionally, it requires the growth temperature of the organism corresponding to the query protein as input, while PPTStab does not require such information. Overall, our method was developed and validated using rigorous and standardized protocols, demonstrating reliable and robust predictive capability.

Table 7.5. Reported performance of existing regression-based predictors (adopted from [143])

Metrics	PPTStab		DeepSTABp		ProTstab2		ProTstab	
	Train	Test	Train	Test	Train	Test	Train	Test
RMSE	3.97	4.11	2.35	4.30	6.99	9.09	0.16	0.19
MSE	0.26	0.28	5.54	18.46	48.89	82.75	0.02	0.04
MAE	3.00	3.00	1.81	3.22	5.20	6.93	0.12	0.15
PCC	0.90	0.89	0.97	0.90	0.76	0.80	0.79	0.74
R²	0.81	0.80	0.93	0.80	0.57	0.58	47.8	-8.50

7.4.5 Application: Large-scale Screening of Thermostable Proteins

PPTStab is prediction tool designed to predict protein T_m and to facilitate the identification of thermostability of proteins. Such proteins are found across a broad range of organisms, from extremophiles adapted to harsh environmental conditions to well-studied species including bacteria, plants, and mammals, and they have significant utility in various biotechnological and industrial applications [332,333]. Thermostable proteins are very useful in industry, especially for enzymatic processes that require tolerance to temperature changes, such as those in the production of biofuel, food, and pharmaceuticals. As stable components in drug delivery systems, diagnostic tests, and therapeutic formulations, they play a significant role in biomedical research and therapeutics in addition to industrial applications. Additionally, their stability facilitates important clinical uses, such as the creation of heat-stable vaccines that can be distributed worldwide and long-lasting biologics for the treatment of specific diseases [334].

To evaluate model performance and practical applicability, the UniProt database was used to obtain reviewed proteomes. These proteomes focused on organisms that are adapted to different

temperature environments, such as thermophiles, mesophiles, and psychrophiles, which survive at temperatures between approximately 0 °C and above 60 °C [335]. The gram-negative, non-motile, aerobic, and osmotolerant bacterium *Psychrobacter frigidicola* was one of the species chosen for research. It is well-known for its use in biotechnological processes such as restriction endonucleases, uracil-DNA glycosylases, and the synthesis of bioactive metabolites with medical uses [336]. We also examine *Shewanella oneidensis*, a mesophilic and electrochemically active bacterium that has been extensively researched in bioelectrochemistry studies and has an optimal growth temperature of about 35 °C [337]. Lastly, a well-known thermophile that is widely employed in structural genomics, genetic modification, and systems biology research, *Thermus thermophilus*, was examined [143]. Table 7.6 provides a summary of the analysis's overall conclusions.

Table 7.6. Distribution of proteins across Tm ranges in selected microorganisms (adopted from [143])

Category	Micro organism	Total Proteins	30-40 °C	40-50 °C	50-60 °C	60-70 °C	70-80 °C	80-90 °C
Psychrophile	<i>Psychrobacter frigidicola</i>	2298	0.26 %	24.07%	53.68%	16.98%	4.70%	0.30%
Mesophile	<i>Shewanella oneidensis</i>	4068	0.32%	27.24%	52.19%	15.71%	4.40%	0.15%
Thermophile	<i>Thermus thermophilus</i>	2227	0.05%	4.28%	31.25%	32.37%	22.51%	9.55%

In this analysis, the percentages represent the distribution of proteins from each microorganism across different Tm ranges. The results indicate that *Psychrobacter frigidicola* shows the highest proportion of proteins with stability in the 50–60 °C range, suggesting adaptability beyond strictly cold environments. *Shewanella oneidensis* exhibits a broader stability profile spanning approximately 40–70 °C, consistent with its mesophilic nature and preference for moderate temperatures. In contrast, *Thermus thermophilus* demonstrates the greatest stability in the 60–70 °C range, which is characteristic of thermophilic organisms. These observations highlight that although each microorganism is adapted to a specific environmental temperature, its proteome can remain stable across a wider thermal range, reflecting evolutionary flexibility. The analysis also shows how PPTStab can be used to identify thermostable proteomes. Interestingly, a similar distribution of proteins across temperature ranges was observed for psychrophilic and mesophilic organisms. One possible explanation is the composition of the training dataset, which

predominantly contains thermophilic proteins or proteins with relatively high melting temperatures. To develop a more generalized predictor, future models should incorporate balanced datasets representing thermophilic, psychrophilic, and mesophilic proteins. Overall, the proposed method is currently optimized for thermophilic proteins, making it particularly useful for applications requiring high thermal stability, including biomedical and biotechnological processes.

7.5 SHAP-based interpretation

To interpret the PPTStab thermostability model, we applied SHapley Additive exPlanations (SHAP) and ranked features by their mean absolute SHAP value. Figure 7.7 shows the top 15 contributing features. Because PPTStab uses ProtBERT embeddings as input features, the most influential features are individual embedding dimensions (e.g., emb_311, emb_15, emb_910, emb_310), with the single experimental feature (lysate) also ranking among the top contributors. The embedding dimension emb_311 had the largest contribution (mean $|\text{SHAP}| \approx 0.0136$), followed by emb_15 (≈ 0.0090) and emb_910 (≈ 0.0083). Unlike composition-based descriptors, individual embedding dimensions do not have a direct biochemical interpretation. The SHAP analysis therefore identifies which learned dimensions carry the most predictive information for melting temperature, but the underlying biochemical meaning of these dimensions is not directly readable from the feature name.

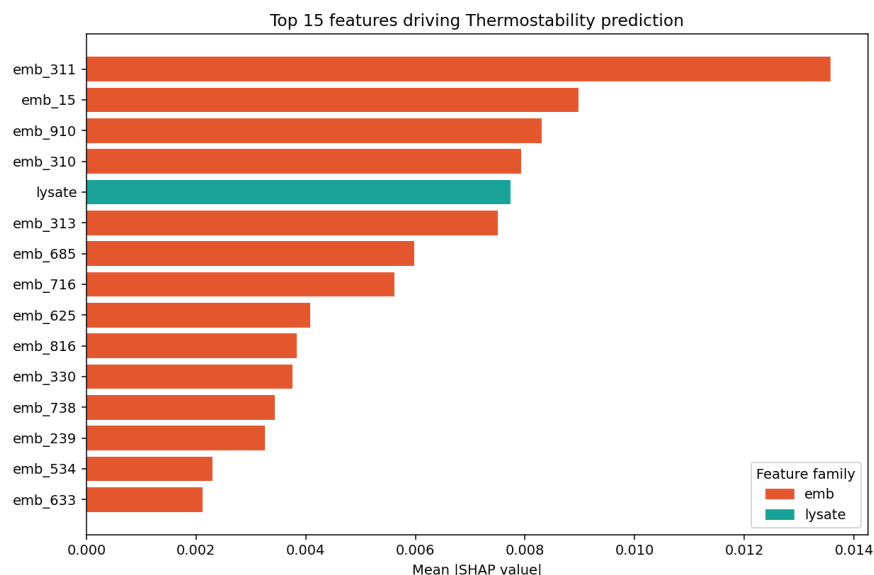


Figure 7.7. SHAP-based feature importance for the thermostability prediction model

7.6 Web server and Standalone Implementation

PPTStab was implemented as an intuitive web server. Both standalone software and `raghava/pptstab` are available at <https://webs.iiitd.edu.in/>. This URL is <https://webs.iiitd.edu.in/raghava/pptstab/stand.php>. The web server offers two main components:

- (i) Predict:** enables users to predict the melting temperature of submitted protein sequences at larger scale.
- (ii) Design:** enables systematic generation of all possible single-point mutant analogs and their predicted T_m values, thereby supporting targeted design of thermostable proteins. The Design module additionally computes multiple physicochemical properties including hydrophobicity, charge, pI, molecular weight, amphipathicity, steric hindrance and others, enabling rational interpretation of variant selection. The web server was developed using responsive HTML templates for compatibility across devices and the standalone versions were made available via GitHub (<https://github.com/raghavagps/pptstab>), and Python package distribution (<https://pypi.org/project/pptstab/>).

7.7 Discussion and Conclusion

Experimental techniques for determining protein thermostability are costly and time-consuming; therefore, computational approaches provide an efficient alternative for predicting protein melting temperature (T_m). Understanding the behavior and functionality of proteins in a variety of biological and industrial applications requires accurate T_m prediction. In contrast to many previously published methods that relied on redundant sequence collections, we developed a new method, using a carefully curated non-redundant dataset. The dataset was obtained from the DeepSTABp resource, which largely derives data from the Meltome Atlas study, and sequence redundancy was reduced using CD-HIT clustering at a 40% identity threshold. A variety of composition-based descriptors and sequence embeddings from protein language models were then used to train multiple regression algorithms, in addition to an extra categorical indicator that showed for “lysate” or “cell” experimental conditions. Performance was evaluated both on individual compositional features and on embeddings generated from different LLMs, as well as combinations of these feature types. Among the traditional descriptors, the SER feature showed the best performance; however, combining SER with other descriptors did not improve prediction accuracy and was therefore excluded from the final model. When using a 1,026-dimensional

feature vector on the validation dataset, the ensemble architecture that combined a custom ANN with an MLP regressor trained on ProtBERT embeddings performed the best, obtaining an R^2 value of 0.80 and PCC of 0.89. Embedding-based models significantly outperformed traditional handcrafted features (e.g., SER, $R^2 = 0.63$). This improvement can be attributed to the contextual nature of protein language model embeddings, which capture higher-order sequence relationships and dataset-specific patterns, whereas traditional descriptors represent static sequence properties. The lack of improvement in the combined SER + embedding model suggests that embeddings already encode most of the informative sequence characteristics, rendering additional handcrafted features redundant. Despite the promise of large language models, their deployment presents challenges such as high computational requirements, scalability limitations, and the substantial resources needed for fine-tuning. These constraints limit widespread accessibility and practical use in many research environments. We created a user-friendly web server, PPTstab (<https://webs.iiitd.edu.in/raghava/pptstab/>), that facilitates the design and prediction of thermostable proteins in order to make the method accessible to researchers.

Limitations: PPTStab was trained on proteins drawn from bacteria, archaea, fungi, plants and animals, so the reported performance represents average accuracy across taxa rather than accuracy within any one lineage. Because thermal adaptation proceeds through different molecular mechanisms in different groups, a single general model may fit some lineages better than others, and taxon-specific models would be needed to resolve lineage-dependent determinants of thermostability. A second constraint is the composition of the training data, which is weighted towards proteins with high melting temperatures. This is the likely reason the predicted T_m distributions of the psychrophilic and mesophilic proteomes in Section 7.6 were almost indistinguishable. Finally, the model was evaluated on a held-out 20 percent split rather than on an independently collected dataset, so its generalisability to data from other laboratories and assay platforms remains untested. Furthermore, external validation using independently collected datasets would further establish the robustness and generalizability of the proposed regression model.

C H A P T E R

8

Prediction of Anti-Freezing Proteins

8.1 Introduction

Antifreeze proteins (AFPs) are specialized polypeptides produced by a wide range of organisms, including fish, plants, and insects, that allow their cells to survive in sub-zero environments. In 1950s, when Scholander and his colleagues discovered that some fish species could survive in freezing seawater, which challenged the prevailing understanding of life in extremely cold climates. AFPs have since been discovered in a wide range of organisms, including bacteria, fungi, insects, and even mammals. By either preventing ice formation or regulating ice growth, these proteins assist organisms in surviving extreme cold temperatures. AFPs have drawn a lot of interest for use in food preservation, medicine, human cryosurgery, and even dairy processing like yogurt production because of their special freeze-resistance property. Furthermore, because current vaccines require a strict cold chain to avoid loss of potency caused by thermal denaturation or degradation, the incorporation of AFP-like stabilization strategies could reduce refrigeration costs and minimize vaccine wastage due to improper temperature exposure [160–162,338].

The global distribution of vaccines depends on an uninterrupted cold chain, and exposure to freezing can be as detrimental as exposure to elevated temperatures, leading to protein denaturation, aggregation, and loss of vaccine potency, particularly in resource-limited settings [339]. Beyond identifying safe and immunogenic candidates, protein- and peptide-based vaccines must remain stable throughout manufacturing, storage, and distribution. Antifreeze proteins (AFPs), naturally produced by cold-adapted organisms, inhibit ice crystal growth and recrystallization, thereby protecting proteins from freeze-induced damage. Consequently, AFPs and AFP-inspired molecules are being explored as cryoprotective agents for the preservation of biologics, including protein therapeutics and vaccine formulations [340,341]. In this context, AFProPred complements the thermostability predictor PPTstab (Chapter 7) by addressing an additional aspect of protein stability. While PPTstab focuses on identifying proteins with enhanced resistance to heat-induced denaturation, AFProPred facilitates the identification of proteins with freeze-protective properties. Together, these complementary tools support the broader objective of developing stable protein therapeutics and vaccine candidates with improved storage and distribution characteristics [342]. AFPs can be identified and characterized using a variety of experimental methods. These include ice-etching [163], ice-affinity chromatography and ammonium precipitation for AFP purification [343], fluorescence-based ice plane affinity (FIPA) [164], and site-directed mutagenesis to identify ice-binding regions and mechanisms

[163,166,343]. However, such experimental approaches are expensive, labour-intensive, and time-consuming. Additionally, traditional similarity-based computational methods such as BLAST or PSI-BLAST are often unable to detect distantly related AFPs due to their high sequence diversity [167,176]. Various in-silico methods have been developed to differentiate AFPs from non-AFPs in order to overcome these limitations (summarized in Table 2.7). One of the initial machine learning-based methods utilizing a Random Forest classifier was presented in 2011 by Kandaswamy and Chou [167]. In the same year, Yu and Lu presented iAFP, a predictor based on SVM that uses a genetic algorithm to select features [168]. Later, several prediction tools were released, such as AFP_PSSM, AFP-PseAAC, AFP-ensemble, TargetFreeze, iAFP-Ense, CryoProtect, afpCOOL, and RAFP-Pred. Miyata et al., AFP-LXGB, and PoGB-pred [344] also explored boosting-based methods. More recently, sparse-representation-based classifiers such as AFP-LSE and AFP-SRC have been proposed for AFP prediction [158]. Despite these developments, a major limitation of many existing predictors lies in the datasets used for training and evaluation, which often rely on unreviewed or redundant sequences. In the present study, we made a systematic effort to construct a high-quality validation dataset of AFPs derived from the reviewed Swiss-Prot database. We further developed a robust prediction framework by integrating multiple strategies, including machine learning-based classification, sequence similarity analysis, and motif/pattern detection. In addition, ensemble approaches were designed by combining two or more predictive strategies to improve the reliability and accuracy of AFP identification.

8.2 Overview of the study

The entire workflow of the study is presented in Figure 8.1. The pipeline begins with compilation of training AFP and non-AFP datasets, followed by redundancy reduction using CD-HIT at 40% identity (*described in Section 3.3*). Two datasets were used: (i) a main dataset for training and optimization, and (ii) an independent validation dataset comprising reviewed Swiss-Prot proteins for unbiased evaluation. The approach evaluates alignment-based (motif discovery, BLAST similarity) and alignment-free (machine learning) approaches. Additionally, it uses evolutionary features from position-specific scoring matrices (PSSM) produced by PSI-BLAST. The best model was ultimately made available as a standalone package and an easy-to-use web server at <https://webs.iitd.edu.in/raghava/afppred>.

8.3 Materials and Methods

8.3.1 Dataset Compilation and Pre-processing

AFProPred utilized two datasets: (1) a validation dataset for external evaluation, and (2) a main dataset for model training.

1. **Validation Dataset:** We derived an independent/validation dataset by acquiring AFPs from the Swiss-Prot's reviewed sequences. We retrieved 297 AFPs from UniProtKB using relevant keyword in reviewed entries, complete list of keywords such as ["Thermal Hysteresis", "sea raven AFP", "insect antifreeze prot motif", "ice structuring proteins", "Hysteresis Protein", "Fish Type II AFP", "antifreeze type 3", "antifreeze type 2", "antifreeze type 1", "antifreeze sf", "antifreeze glycoprotein", "antifreeze"]. After eliminating redundant sequences and fragments, we got 80 non-redundant sequences out of the 297 sequences. These 80 AFPs having ranging from 16 to 2439 amino acids used to validate our models. In order to create set of non-AFP, we searched UniProtKB for the term "NOT antifreeze protein" and obtained 4,29,837. We got 81,339 non-redundant non-AFPs, after removing redundant sequence using CD-hit at cut-off 40%. We randomly picked an equal number of negative sequences and then removed sequences which are not in the range of more than 16 and less than 2439 amino acids. Finally, we have 73 non-AFPs and 80 AFPs in validation dataset for evaluating our models.
2. **Main dataset:** We first extracted 49,352 unreviewed AFP sequences from the UniProt database in order to create the AFP (positive) dataset for training our prediction model. To follow standard bioinformatics practices and avoid redundancy bias, we applied the CD-HIT program (*see Section 3.3*) and reduced sequence identity to 40%. A total of 8,134 positive sequences with lengths ranging from 37 to 12,385 residues were kept after duplicate entries and sequences containing non-natural amino acids were eliminated. For the negative dataset, 9,493 non-AFP sequences were obtained from the existing prediction tool AFP-Pred. After non-natural amino acid-containing sequences were eliminated, 9,439 non-AFP sequences were left. Altogether, the final training dataset comprised 17,573 sequences, which were used for model training and optimization.

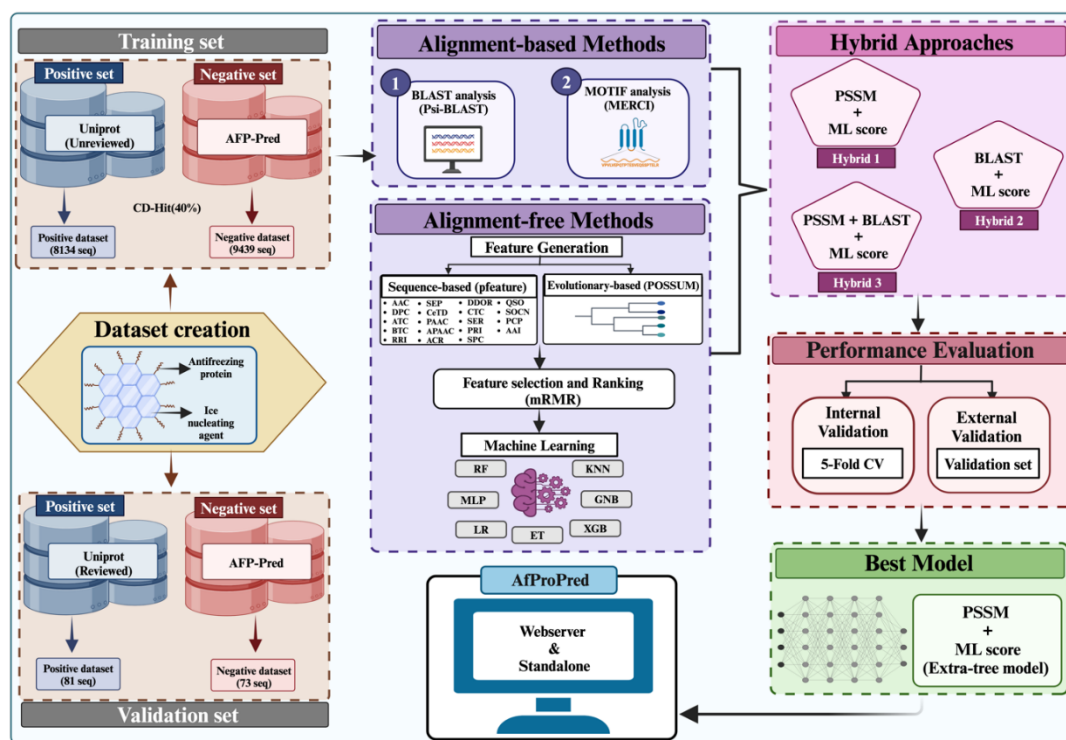


Figure 8.1 The overall workflow of the study (adopted from [158])

8.3.2 Composition-based Feature Extraction

To represent proteins numerically, we computed diverse composition-based descriptors using the Pfeature framework, following the protocol described in *Section 3.7.1*. A total of 1,165 features were computed including AAC (20), DPC (400), ATC (5), BTC (4), PCP (30), RRI (20), DDOR (20), SER (20), SEP (1), CTC (343), PAAC (21), APAAC (23), QSO (42), SOCN (2), CeTD (189), and SPC (25) detailed in Table 3.3.

8.3.3 Evolutionary Feature Generation (PSSM Profiles)

Protein evolutionary features are known to offer significant supplementary information beyond the primary sequence composition. PSI-BLAST [231] was used to create PSSM profiles that included evolutionary information (for more information, see *Section 3.7.3*). The PSSM-400 feature vector was generated in this study using the “psm_composition” module of the POSSUM package [238]. Within a protein sequence, this representation corresponds to a 20×20 dimensional composition matrix that describes the occurrence and substitution propensity of the 20 standard amino acids. For each sequence, PSI-BLAST searches were performed against the Swiss-Prot database to generate the PSSM profile whenever homologous sequences were detected. For sequences where

no homologs were identified and therefore no PSSM could be generated, a zero-filled PSSM-400 vector was assigned to maintain a consistent feature dimension across the dataset.

8.3.4 Feature Selection

A large number of redundant and irrelevant features may negatively impact model performance, whereas a small number of features greatly help identify the target endpoint in most biological datasets. This issue was resolved by applying feature selection techniques to identify an informative subset of features from the original feature space. In a variety of fields, numerous feature selection techniques have been developed and successfully applied [345]. The ensemble minimum redundancy–maximum relevance (mRMR) method was employed in this study to rank features according to mutual information (refer to *Section 3.8*). The top-ranked features selected were extremely relevant in order to differentiate AFPs from non-AFPs while preserving the least amount of redundancy between them. This improved the predictive model's robustness and interpretability.

8.3.5 Selection of appropriate ML models

Several machine learning algorithms, such as KNN, RF, LR, GNB, ET, MLP, and XGB, were evaluated in order to find an appropriate classifier for AFP prediction (*see Section 3.9.1*). Five-fold cross-validation was used to train and evaluate the models, and an independent validation dataset was used for additional validation (details in *Section 3.12*). A number of threshold-dependent and threshold-independent metrics, such as sensitivity, specificity, accuracy, MCC, Youden's Index, balanced accuracy, F1-score, and AUROC, were used to evaluate the model's performance (*see Section 3.13.1*).

8.3.6 Alignment-based and Motif-based Methods

Similarity-based classification was implemented using BLAST+ (blastp). Query proteins were searched against a BLAST database constructed from the training set. Proteins were classified based on top-hit label (*described in Section 3.10.1*). In parallel, motif analysis was performed using MERCI to identify exclusive and inclusive degenerate motifs enriched in AFPs. Several motif configurations were explored, including no-gap, gap=1, gap=2 and Koolman–Rohm classes (*described in Section 3.10.2.1*).

8.4 Results

The Results section of this work is divided into numerous subsections, such as (i) AFP compositional analysis, (ii) alignment-based and alignment-free methods, (iv) hybrid approaches, (v) benchmarking, and (vi) web server implementation.

8.4.1 Compositional analysis

For each dataset, we determined the amino acid composition to analyze the frequency of specific residues (refer to *Section 3.5*). It is possible to find unique patterns linked to potential AFPs by analyzing the average amino acid composition of protein sequences. As shown in Figure 8.2, the average residue composition for AFPs, non-AFPs, and the entire proteome was calculated in this investigation. While leucine (L) and glutamic acid (E) are relatively enriched in the non-AFP dataset, the results show that alanine (A), isoleucine (I), valine (V), and threonine (T) are more abundant in AFPs (positive dataset). The differences between the average composition scores of AFPs and the non-AFP dataset and the general proteome are further shown in Figure 8.3.

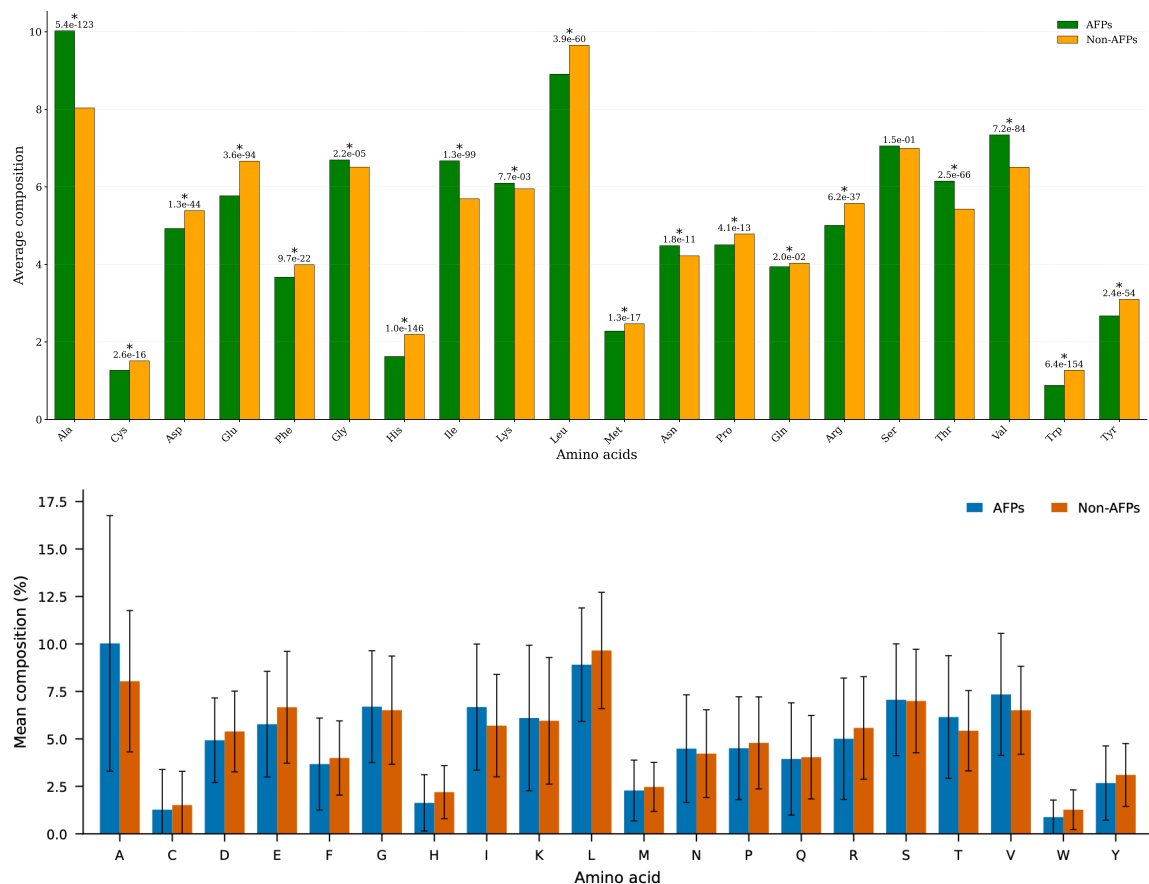


Figure 8.2 Average amino acid composition of antifreeze proteins (AFPs) and non-antifreeze proteins (non-AFPs)

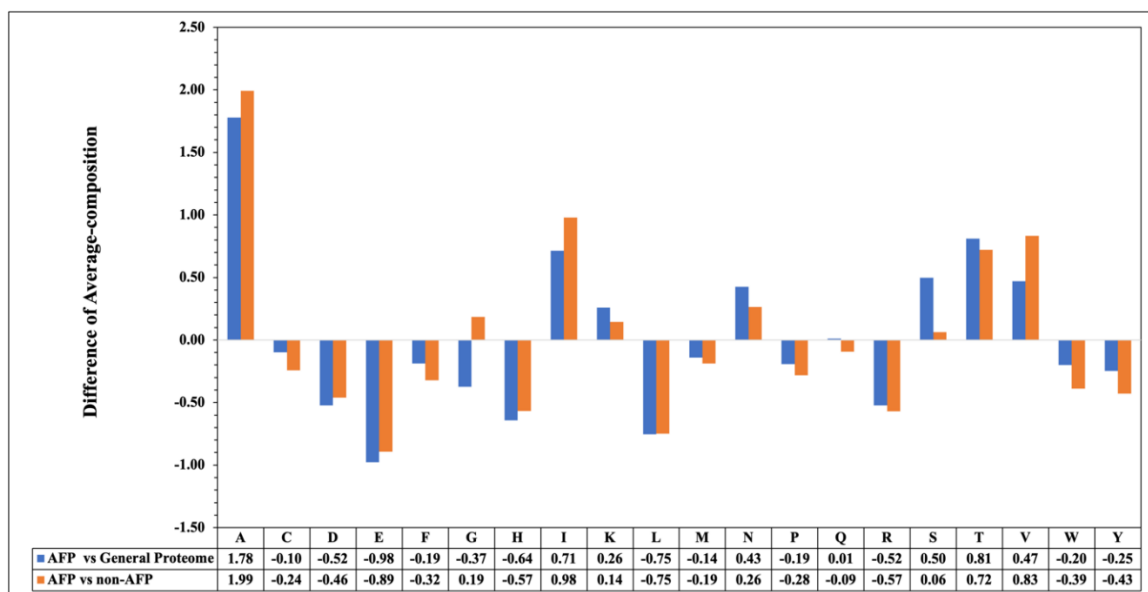


Figure 8.3 The differences in average-composition scores of AFPs compared to general proteome and non-AFPs (adopted from [158])

8.4.2 Alignment-Based Method

8.4.2.1 Motifs Analysis

To identify characteristic sequence patterns associated with AFPs, we searched for exclusive and inclusive motifs that could be used for AFP detection. Motif discovery was performed using the open-source software MERCI. Since MERCI provides multiple search configurations, we evaluated different parameter settings, including no gap (default), one-gap, two-gap, and the Koolman–Rohm classification scheme. By default, MERCI uses an FN parameter (maximum frequency in negative sequences) set to zero, which identifies exclusive motifs specific to the positive dataset. To broaden the analysis and obtain inclusive motifs present in both positive and negative datasets, we increased the FN value to 8. The top five motifs identified in each category, along with the number of sequences in which they occurred, are summarized in Table 8.1. However, only a limited number of sequences contained these motifs, indicating that motif-based methods alone are insufficient for predicting the function of all proteins.

Table 8.1 Top 5 motifs exclusive and inclusive motifs along with the number of sequences in which each motif was observed (adopted from [158])

Exclusive motifs (fn = 0)			Inclusive motifs (fn = 8)		
Motifs	Pos	Neg	Motifs	Pos	Neg

No gap					
A A A A A A A A A A A A A A	234	0	A A A A A A A A A A	367	5
T A A A A A A A A A	216	0	A A A A A A A A A A	328	2
A A A A A A A A A A A A A A	212	0	T A A A A A A	326	7
A A A A A A A A A A A A A A	200	0	A A A A A A T	303	7
A A A A A A A A A A A A A A A	181	0	A A A A A A A A A A A A	292	1
Gap = 1					
A A A A A A gap A A A A A A A A	277	0	A A A A gap A A A A A A	407	5
A A A A A gap A A A A A A A A	272	0	A A A A A gap A A A A A	404	6
A A A A gap A A A A A A A A A A	265	0	A A A gap A A A A A A A A	403	5
A A A A A A A gap A A A A A A A A	257	0	A A A A A A gap A A A A A	399	6
A A A gap A A A A A A A A A A	254	0	A A gap A A A A A A A A A	399	6
Gap = 2					
A A A A gap A A gap A A A A A A A A	296	0	A A A gap A A A A A gap A A	431	6
A A A gap A A A gap A A A A A A A A	291	0	A A gap A A A A gap A A A A A	431	8
A A A gap A A gap A A A A A A A A A	291	0	A A gap A A A gap A A A A A A	431	7
A A gap A A A A gap A A A A A A A A	291	0	A A A A A gap A A A gap A A	430	6
A A gap A A A gap A A A A A A A A A	289	0	A A A A gap A A A A gap A A	430	5
Class = Koolman–Rohm					
aliphatic aliphatic A aliphatic aliphatic	295	0	aliphatic aliphatic A A A aliphatic	420	8
aliphatic aliphatic aliphatic aliphatic A A A			A A alipha		
aliphatic aliphatic					
aliphatic A A aliphatic aliphatic aliphatic A	290	0	A aliphatic aliphatic aliphatic	413	8
aliphatic aliphatic A A aliphatic aliphatic			aliphatic A al		
aliphatic aliphatic A A aliphatic aliphatic	290	0	aliphatic A aliphatic A aliphatic	412	8
aliphatic A aliphatic aliphatic A A aliphatic			aliphatic A		
aliphatic aliphatic A aliphatic aliphatic A A	290	0	aliphatic aliphatic A A A aliphatic	411	8
aliphatic aliphatic aliphatic A A aliphatic			aliphatic		
aliphatic aliphatic A aliphatic aliphatic A	290	0	aliphatic aliphatic A A aliphatic A	411	8
aliphatic aliphatic aliphatic A A A aliphatic			A aliphatic		
aliphatic					

8.4.2.2 Similarity search approach

A widely used sequence similarity search tool, pBLAST (BLAST+ 2.7.1), was employed to perform alignment-based prediction. In this approach, a database was constructed using the training dataset, and query sequences (from the test set) were searched against this database using blastp at various E-value thresholds. Each query sequence was classified as AFP or non-AFP based on the class of its top-scoring hit. Specifically, a query sequence was assigned as an AFP if its best

match corresponded to an AFP; otherwise, it was classified as a non-AFP. A major limitation of this method arises when a query sequence does not produce a significant hit against either AFPs or non-AFPs in the database. Consequently, similarity-based approaches alone are insufficient for predicting the function of all proteins.

8.4.3 Alignment-Free Method

As demonstrated above, alignment-based methods could not offer complete coverage. We built models for antifreeze protein prediction using a variety of machine-learning techniques in order to overcome this difficulty. Among these methods are RF, MLP, LR, XGB, KNN, ET, and GNB. These models were created using a variety of compositional features, combined features, evolutionary-based features, PSSM profiles, and selected features.

8.4.3.1 Compositional features

As discussed in *Section 3.7.1*, we use various composition-based feature types to build machine learning-based prediction models. These features have been generated using the standalone software “Pfeature”; different types of features, including the number of descriptors, are shown in Table 3.3. The performance in terms of AUROC for these models on different types of protein compositions on our validation dataset, which contains 80 AFPs and 73 non-AFPs, is shown in Table 8.2.

Table 8.2 AUROC of different descriptors models on the validation dataset (adopted from [158])

Features	XGB	RF	MLP	LR	KNN	GNB	ET
AAC	0.89	0.87	0.86	0.89	0.89	0.79	0.87
DPC	0.84	0.83	0.90	0.91	0.83	0.70	0.82
ATC	0.77	0.76	0.79	0.86	0.75	0.73	0.78
BTC	0.67	0.70	0.56	0.78	0.65	0.58	0.68
PCP	0.89	0.87	0.87	0.89	0.77	0.75	0.86
RRI	0.74	0.71	0.71	0.69	0.63	0.55	0.70
DDR	0.83	0.84	0.81	0.62	0.74	0.66	0.86
SER	0.88	0.87	0.87	0.87	0.89	0.77	0.86
SEP	0.67	0.65	0.68	0.68	0.62	0.69	0.64

CTC	0.75	0.73	0.79	0.77	0.70	0.65	0.74
PAAC	0.88	0.88	0.87	0.89	0.89	0.78	0.88
APAAC	0.87	0.89	0.87	0.90	0.89	0.80	0.89
QSO	0.87	0.86	0.87	0.91	0.89	0.77	0.83
SOCN	0.66	0.67	0.66	0.65	0.63	0.72	0.68
CeTD	0.80	0.72	0.77	0.82	0.75	0.66	0.74
SPC	0.85	0.87	0.83	0.85	0.74	0.71	0.85

8.4.3.2 PSSM profiles

Previous studies have demonstrated that evolutionary information-based models generally outperform models relying solely on primary sequence features [236,266,346]. Therefore, in this study, we investigated the contribution of evolutionary information to AFP prediction. Evolutionary profiles for each protein sequence were generated using the POSSUM package, which derives features from PSSM profiles. We successfully generated PSSM profiles for 146 out of 153 proteins in the validation dataset; however, profiles could not be produced for seven AFP sequences due to the absence of detectable homologs. The performance of the machine learning models developed using PSSM-based features on the validation dataset (146 proteins) is summarized in Table 8.3. As shown, models based on ET, XGB, and RF classifiers achieved a maximum AUROC of 0.92 on the validation dataset, which contained 73 AFPs and 73 non-AFPs.

Table 8.3 Performance of PSSM based ML models (adopted from [158])

Classifier	Training (8134 AFPs and 9439 non-AFPs)							YI
	Sens (%)	Spec (%)	Accuracy (%)	AUROC	MCC	F1-Score	BAC (%)	
ET	82.18	83.08	82.73	0.91	0.65	0.79	82.63	0.65
GNB	53.89	81.90	70.91	0.77	0.37	0.59	67.90	0.36
KNN	48.20	51.86	50.42	0.51	0.00	0.43	50.03	0.00
LR	79.88	79.39	79.58	0.87	0.58	0.75	79.63	0.59
MLP	62.27	62.25	62.26	0.66	0.24	0.56	62.26	0.25
RF	81.58	82.72	82.28	0.91	0.64	0.78	82.15	0.64
XGB	82.71	82.94	82.85	0.91	0.65	0.79	82.82	0.66
Validation (73 AFPs and 73 non-AFPs)								

ET	83.37	84.88	84.28	0.92	0.68	0.81	84.12	0.66
GNB	49.05	84.26	70.30	0.78	0.36	0.57	66.65	0.31
KNN	97.15	4.75	41.39	0.51	0.05	0.57	50.95	0.49
LR	82.13	81.89	81.98	0.89	0.63	0.78	82.01	0.62
MLP	68.44	81.89	76.56	0.81	0.51	0.70	75.16	0.31
RF	82.41	84.95	83.94	0.92	0.67	0.80	83.68	0.65
XGB	86.22	81.07	83.11	0.92	0.66	0.80	83.65	0.69

8.4.3.3 Combined features

A feature vector of size 1,165 for every sequence in the datasets was then produced by combining all compositional features. The combined feature set was then used to create prediction models using various classifiers, with hyperparameters thoroughly adjusted to maximize AUROC. Our analysis revealed that the ET and XGB classifiers performed comparably, achieving AUROC values of 0.86 and 0.87, respectively, on the validation dataset, as illustrated in Figure 8.4.

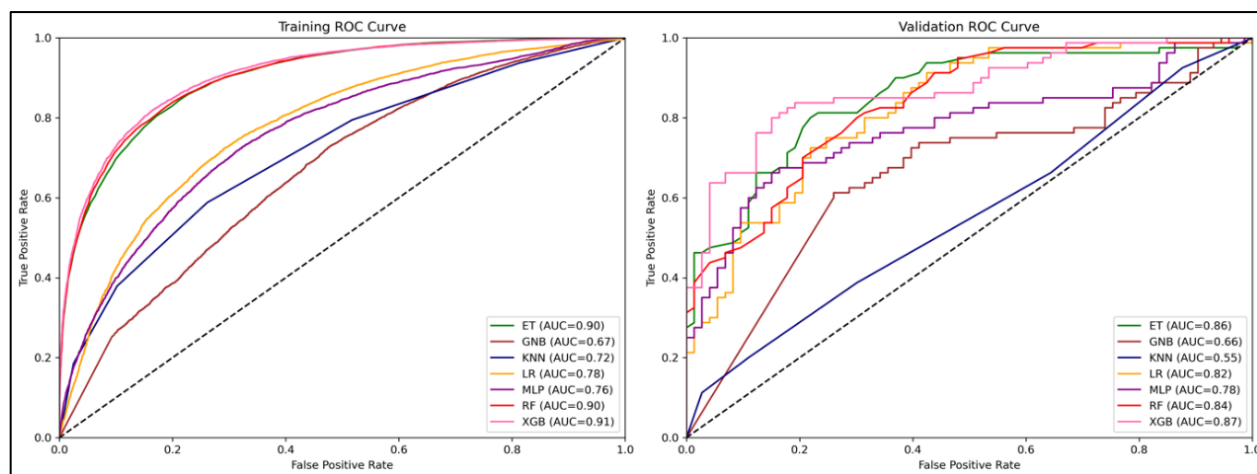


Figure 8.4 ROC curve for models built using combined features on training and validation dataset (adopted from [158])

8.4.3.4 Selected features

In high-dimensional feature sets, many descriptors may be irrelevant or redundant, making model training computationally expensive and potentially reducing predictive performance. Therefore, selecting the most informative features is essential to simplify the model and improve its effectiveness. In this study, a total of 1,165 features were generated using the Pfeature software.

We applied the mRMR feature selection technique to identify the most relevant subset while removing redundant features. We got best set of 150 features after applying mRMR. These features were used to develop machine learning based models for discriminating AFPs and non-AFPs (Figure 8.5).

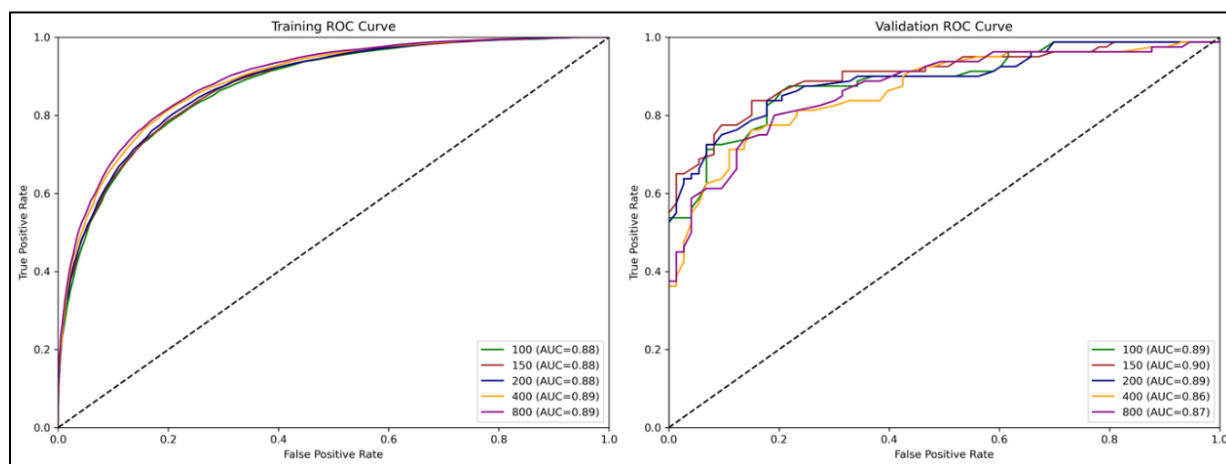


Figure 8.5 ROC curve for models built with mRMR-selected features on training and validation dataset (adopted from [158])

8.4.4 Hybrid approaches

We used both alignment-free (ML techniques) and alignment-based (motif search and BLAST) methods, as explained in the previous sections. Each approach has unique benefits and drawbacks. Since the effectiveness of alignment-based techniques depends on sequence similarity or the presence of conserved motifs, they typically offer high specificity but relatively low sensitivity. Sequence similarity is not a requirement for machine learning-based (alignment-free) models, which are more broadly applicable. Using alignment-based and alignment-free features, we created ensemble (hybrid) models to capitalize on the advantages of both strategies. Machine learning models were first trained using features based on AAC and PSSM. As illustrated in Figure 8.6, the hybrid AAC-PSSM based ET classifier outperformed the others, achieving an AUROC of 0.93 and an MCC of 0.77. The performance on other matrices such as sensitivity, specificity, Accuracy, Matthews correlation coefficient (MCC), F1-Score, Balanced Accuracy (BAC), and Youden's Index (YI) mentioned in Table 8.4. We further constructed models using the selected top 150 features combined with PSSM profiles, where RF and ET classifiers demonstrated nearly identical performance, achieving an AUROC of 0.92. Additionally, integrating the BLAST-based approach with the ET model developed on the AAC+PSSM feature set resulted in a maximum AUROC of

0.90. Several feature combinations were explored to develop optimal hybrid predictors. The highest performance (AUROC = 0.93 on the validation dataset) was obtained using the ET model based on PSSM+AAC features. This final model was selected for implementation in the standalone software and web server.

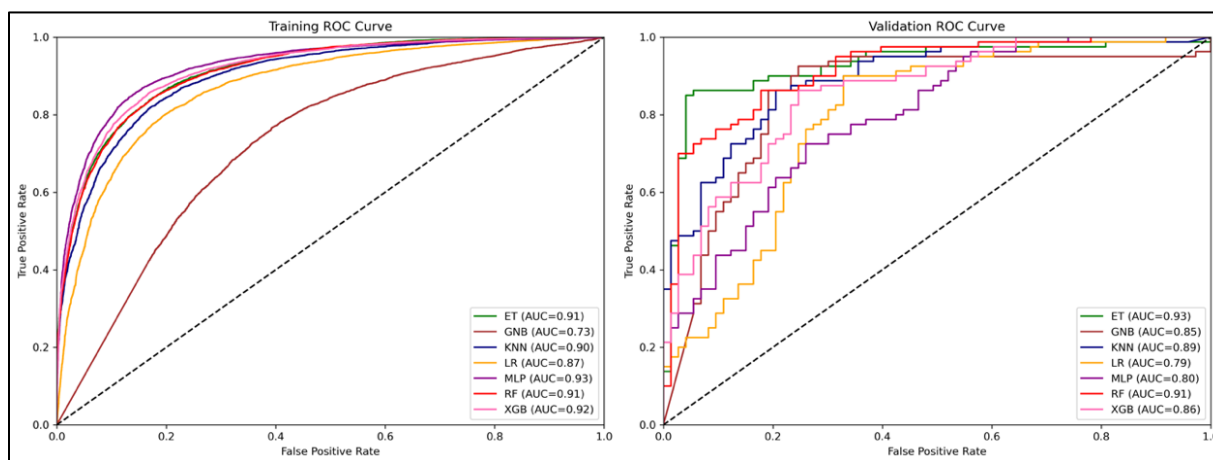


Figure 8.6 ROC curve built using PSSM-composition with AAC composition (adopted from [158])

Table 8.4. Performance on the PSSM composition profile combined with AAC composition on main (for training) and validation dataset (for Validation)

Training (8134 AFPs and 9439 non-AFPs)								
Classifier	Sens (%)	Spec (%)	Acc (%)	AUROC	MCC	F1-Score	BAC (%)	YI
ET	82.91	83.66	83.32	0.92	0.67	0.82	83.29	0.67
GNB	51.48	78.79	66.15	0.73	0.32	0.59	65.14	0.30
KNN	82.35	82.07	82.20	0.90	0.64	0.81	82.21	0.64
LR	80.16	80.06	80.11	0.87	0.60	0.79	80.11	0.60
MLP	85.52	85.68	85.60	0.93	0.71	0.85	85.60	0.71
RF	83.38	83.19	83.28	0.91	0.67	0.82	83.29	0.67
XGB	84.13	84.37	84.26	0.92	0.68	0.83	84.25	0.69
Validation (80 AFPs and 73 non-AFPs)								
ET	85.00	91.78	88.24	0.93	0.77	0.89	88.39	0.77
GNB	46.25	91.78	67.97	0.85	0.42	0.60	69.02	0.38
KNN	68.75	87.67	77.78	0.89	0.57	0.76	78.21	0.56
LR	47.50	79.45	62.75	0.79	0.28	0.57	63.48	0.27
MLP	70.00	73.97	71.90	0.80	0.44	0.72	71.99	0.44
RF	76.25	87.67	81.70	0.91	0.64	0.81	81.96	0.64
XGB	73.75	78.08	75.82	0.86	0.52	0.76	75.92	0.52

8.5 SHAP-based interpretation

To interpret the AFProPred predictions, we applied SHapley Additive exPlanations (SHAP) and ranked features by their mean absolute SHAP value. Figure 8.7 shows the top 15 contributing features. Prediction was driven by a combination of amino-acid-composition (AAC) descriptors and evolutionary (PSSM) features, consistent with the best-performing model built on PSSM and AAC features. The most influential feature was the composition of tyrosine (AAC_Y), followed by two PSSM columns (pssm226, pssm117) and the composition of glutamine (AAC_Q). The remaining top features were dominated by AAC terms, including alanine (AAC_A), threonine (AAC_T), tryptophan (AAC_W), phenylalanine (AAC_F), methionine (AAC_M), glycine (AAC_G), and several charged and polar residues (Arg, Lys, Asn, Pro) together with isoleucine (AAC_I).

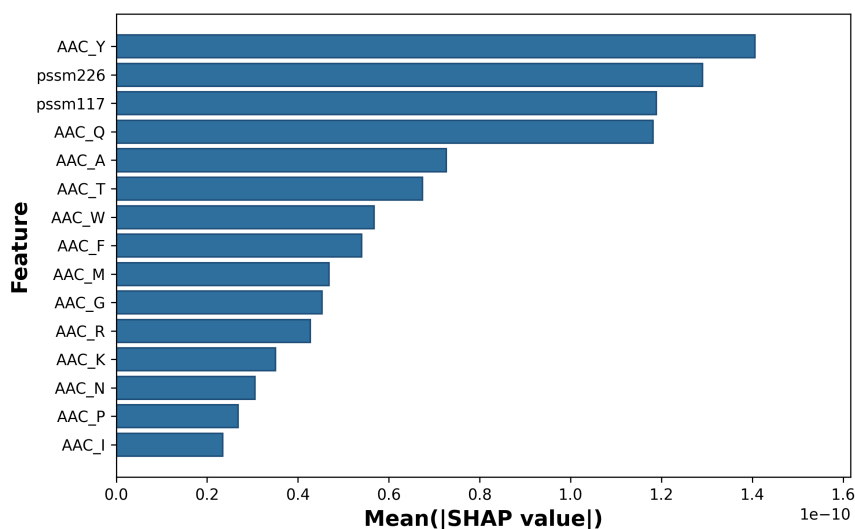


Figure 8.7. SHAP-based feature importance for the antifreeze-protein prediction model. Bars show the top 15 features ranked by mean absolute SHAP value (mean |SHAP value|).

8.6 Benchmarking

To determine the effectiveness of the proposed method, it is necessary to compare its performance with previously developed approaches using the same dataset. Accordingly, we evaluated several existing AFP prediction tools on our validation dataset. However, not all reported methods are currently accessible or functional; therefore, comparisons were limited to publicly available and operational predictors. Our results indicate that the existing methods CryoProtect [173], AFP-CKSAAP [177], AFP-LSE [179], and AFP-SRC [184] exhibited lower performance on the validation dataset compared with AFProPred, as summarized in Table 8.5.

Table 8.5. Performance comparison of existing predictors on the validation dataset (adopted from [158])

Name	Year	Sens (%)	Spec (%)	Acc (%)	Y I	MCC	AUROC	F1	BAC (%)
Cryoprotect	2017	0.79	0.43	60.13	0.22	0.23	0.61	NA	0.61
AFP-CKSAAP	2019	0.78	0.88	82.00	0.65	0.65	0.89	NA	0.83
AFP-LSE	2020	0.73	0.75	74.00	0.48	0.48	NA	0.74	0.74
AFP-SRC	2022	0.56	0.58	57.00	0.14	0.14	NA	0.58	0.57
AFProPred	2024	0.85	0.92	88.24	0.77	0.77	0.93	0.89	0.90

8.7 Implementation of web server

To provide the scientific community with an easily accessible resource, we created AFProPred, a web server accessible at <https://webs.iitd.edu.in/raghava/afpropred/>. This platform is used to implement our top-performing antifreeze protein (AFP) prediction model. Protein Scan, Design, and Prediction are AFProPred's three functional modules. To differentiate AFPs from non-AFPs, the Prediction module allows users submit protein sequences in FASTA format for analysis. Users can use the design module to evaluate potential AFP and create all possible analogs (mutants) of an input sequence. AFP-associated areas in a given protein sequence can be identified using the Protein Scan module. A responsive interface is used in the implementation of the web server to ensure compatibility with a number of devices, including desktop computers, laptops, and smartphones. Furthermore, a standalone Python package called AFProPred has been created to make large-scale genome-wide AFP prediction easier. This package can be downloaded from the web server at <https://webs.iitd.edu.in/raghava/afpropred/standalone.html>.

8.8 Discussion and Conclusion

In the past, a number of computational techniques have been put forth to increase AFP prediction accuracy. Based on the AFP-Pred dataset, which includes 4,81 AFPs and 9,493 non-AFPs, the majority of current methods are used. Some studies used proteins derived from PDB structures, while others (like iAFP) used AFPs that were taken from the Pfam database. A major limitation of these earlier studies is the nature of their datasets, as none employed experimentally validated or well-annotated proteins from the reviewed Swiss-Prot database. Therefore, developing a prediction method using annotated or experimentally supported data is critically important.

In this work, we made a systematic effort to construct a dataset of well-annotated AFPs. Querying Swiss-Prot yielded only 80 reviewed AFPs, which was insufficient for training and optimization. Hence, these 80 AFPs were used exclusively for independent validation. Similarly, 73 non-AFP proteins were retrieved from Swiss-Prot using the keyword “NOT_antifreeze_protein.” Consequently, an independent evaluation dataset comprising 80 AFPs and 73 non-AFPs was constructed and was not used during model training or hyperparameter optimization. Our compositional analysis indicated shared evolutionary patterns among AFPs, where certain residues particularly A, I, V, and T, were conserved. Notably, threonine enhances AFP activity by forming hydrogen bonds on the protein surface that interact with ice crystals [45]. Initially, we applied the similarity-based approach BLAST to identify AFPs. However, due to poor sequence similarity among AFPs, BLAST failed to reliably discriminate AFPs from non-AFPs.

To address this limitation, we employed generalized alignment-free methods. Using these features, multiple ML models were developed, achieving a maximum AUC of 0.90 on the independent dataset. Since evolutionary information is known to be critical for functional prediction, we further incorporated PSSM profiles generated by PSI-BLAST. The inclusion of evolutionary features improved model performance, increasing the AUC from 0.90 to 0.93. We then compared the performance of our method with available predictors using the same independent dataset (80 AFPs and 73 non-AFPs). The results demonstrated that our method outperformed existing tools. Nevertheless, one limitation of the present study is the relatively small evaluation dataset, which restricts robust statistical assessment. Despite extensive efforts, a larger experimentally validated dataset could not be constructed. Future studies should focus on generating larger curated AFP and non-AFP datasets for more comprehensive benchmarking.

Limitations: The present study has several limitations. The independent validation dataset was relatively small, comprising 80 reviewed AFPs and 73 non-AFPs, although all sequences were obtained from high-quality Swiss-Prot entries. In addition, the non-AFP dataset consisted of proteins not annotated as antifreeze proteins rather than experimentally validated non-AFPs. Furthermore, the positive training dataset was primarily derived from unreviewed UniProt entries due to the limited availability of experimentally validated AFPs. Future studies should focus on expanding experimentally validated AFP and non-AFP datasets, performing external validation on independent datasets, and integrating advanced protein language models and evolutionary representations to further improve the robustness and generalizability of AFP prediction models.

C H A P T E R

9

Half-life Prediction of Natural and Modified Peptides

9.1 Introduction

Peptide therapeutics are increasingly used for the treatment of a wide range of diseases. They safely and effectively supplement or replace biologically important peptides in the human body and are applied in the management of neurological disorders, diabetes, cancer, cardiovascular diseases, and several other medical conditions. The global market for peptide therapeutics was estimated to be worth USD 52.59 billion in 2025. It is expected to increase at a compound annual growth rate (CAGR) of 5.19% from 2026 to 2035, from USD 56.06 billion in 2026 to nearly USD 87.21 billion. This growth is largely driven by the increasing prevalence of cancer and metabolic disorders [347]. Despite their potential for therapeutic use, most peptides have a short half-life in vivo ($t_{1/2} \approx 2\text{--}30$ minutes) because they rapidly break down by proteases and are easily eliminated by the kidneys because molecules smaller than ~ 30 kDa are easily filtered and eliminated. Consequently, improving peptide half-life is essential to enhance therapeutic efficacy, reduce dosing frequency, and minimize high-dose administration [194]. The short half-life remains one of the major barriers to the clinical development of peptide drugs, primarily because enzymatic degradation significantly reduces their bioavailability. A number of techniques, such as cyclization, the addition of modified amino acids, and terminal modifications, have been investigated to improve peptide stability. These methods enhance bioavailability in addition to extending in vivo half-life. Determining the half-life of recently created peptides in blood, however, remains to be an important challenge to determining their stability. Despite being accurate and well-established, experimental techniques for determining peptide half-life are expensive, time-consuming, and labor-intensive. Therefore, for effective peptide half-life estimation, alternative computational methods are needed.

An in silico method capable of predicting and designing peptide half-life in blood would serve as a valuable resource for researchers working in therapeutic peptide development. In this study, we aim to develop such a computational predictor using various machine learning techniques, large language models (LLMs), and diverse feature representations, including amino acid composition, dipeptide composition, binary profiles, and atom composition. A wide range of biological activities, including cell-penetrating, tumor-homing, anticancer, antihypertensive, antiviral, and antimicrobial qualities, have been documented for many peptides over the years. The FDA has already approved a number of peptide-based medications due to their high specificity, low toxicity,

chemical diversity, and natural origin; numerous others are presently being developed. However, their susceptibility to proteolytic degradation is the main reason why their clinical use is still restricted. In order to overcome this limitation, we have created an in silico technique to predict the half-life of peptides in an intestine-like environment and develop more stable peptides with enhanced stability and physicochemical characteristics [200].

9.2 Overview of the study

The overall architecture of the method is illustrated in Figure 9.1. Firstly we have collected the half-life dataset from the Peplife2, followed by dataset preprocessing using the parameters mentioned in PlifePred. We construct three datasets: (i) Combined peptide dataset with flags (mentioned in Table 9.1) (ii) Natural peptide dataset (iii) Modified peptide dataset with flags. Pfeature, LLM embeddings, and properties like pI, molecular weight, peptide length, and the percentage composition of individual amino acids were used to extract a large number of features after the dataset was balanced and divided into training and independent validation sets. An independent validation dataset was used to evaluate the final optimized model after the regression models were trained and assessed using five-fold cross-validation. To make the approach accessible, we created the web-based platform PlifePred2 (<https://webs.iiitd.edu.in/raghava/plifepred2>), which makes it possible to predict the half-life of peptides and makes it easier to design proteins with the desired stability.

9.3 Material and Methods

9.3.1 Dataset construction

Peptide sequences and associated half-life information were obtained from PEPLife2, a curated database containing experimentally determined half-life values for approximately 4,500 peptides evaluated under diverse biological environments, including blood, urine, intestinal fluids, kidney and brain homogenates, various cell lines, and standard media such as PBS. From this resource, we first selected peptides with experimentally validated half-life data, resulting in 4,090 entries. To ensure biological relevance and consistency, several filtering criteria were applied. Peptides with lengths greater than 50 residues or fewer than 5 residues were excluded. In addition, peptides with a half-life of less than 20 seconds were removed to avoid extremely unstable sequences that may introduce noise in modeling. After applying these filters and removing duplicates, a final

dataset of 2,679 unique peptides was obtained, with sequence lengths ranging from 5 to 50 residues and half-life values exceeding 20 seconds.

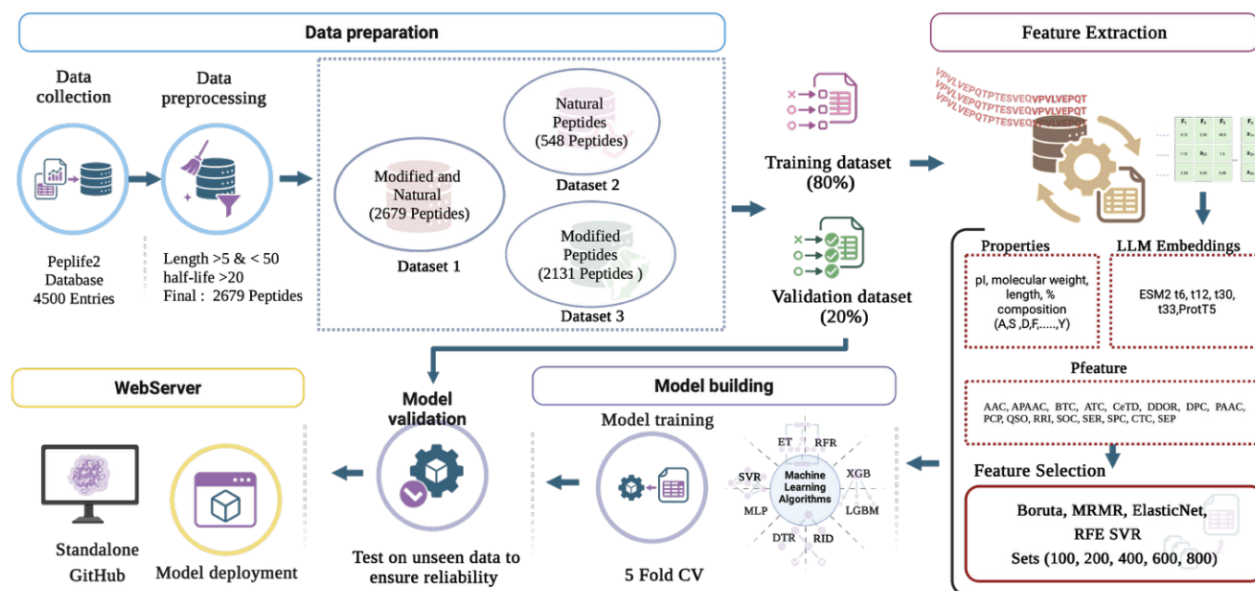


Figure 9.1 The workflow of the PlifePred2

The dataset included both natural peptides (548 sequences) and modified peptides (2,131 sequences). Such modified peptides were retained because prior studies have demonstrated that even a single amino acid substitution or chemical modification can significantly alter peptide half-life [348,349]. To normalize the distribution and represent half-life values on a linear modeling scale, the $\log_{10}$ transformation of half-life (in seconds) was used for ML. In this study, we implemented an 80:20 data splitting, as described in Section 3.4.1.

In peptide engineering, when full structural data are unavailable, binary flags are often used to represent key biochemical features that influence peptide stability and degradation. Many previous studies use flags to represent structural or chemical properties that influence peptide stability, especially when full three-dimensional structures are unavailable. These flags indicate whether a peptide contains specific features known to affect biological half-life. In this study, we developed a model that integrates sequence information with five biological flags - d, ct, nt, cyc, ptm (e.g. MKTL...0,1,0,1,1; 0/1 indicators) to predict a biological property (half-life) from peptide sequences. These flags denote the presence or absence of the following attributes (see Table 1).

Table 9.1 The flags with their description

Flag	Description (in peptide/protein context)
d	D-amino acids present
ct	C-terminal modification
nt	N-terminal modification
cyc	Cyclization
ptm	Post-translational modification

9.3.2 Datasets Used for Model Development

To systematically evaluate the impact of sequence modifications and biological flags on peptide half-life prediction, three datasets were constructed (see Figure 9.2).

(i) Dataset 1: Combined dataset with flags

This dataset included both natural peptides and modified peptides, along with their associated biological flags (d, nt, ct, cyc, ptm). After preprocessing and filtering, a total of 2,679 peptide sequences were retained and finally getting 2,138 sequences for training and 541 sequences for independent validation.

(ii) Dataset 2: Natural peptides

This dataset consisted exclusively of peptides without any chemical or structural modifications. After separation from the combined dataset, the natural peptide subset was split into 438 training sequences and 110 validation sequences

(iii) Dataset 3: Modified peptides with flags

This dataset contained only modified peptides, i.e., sequences possessing at least one of the defined biological flags (D-amino acid, N-terminal modification, C-terminal modification, cyclization, or PTM). Natural peptides were excluded from this set. The final dataset comprised 1,700 training sequences and 431 validation sequences.

These three datasets enabled comparative evaluation of model performance on natural peptides, modified peptides, and the combined dataset incorporating biologically relevant structural flags.

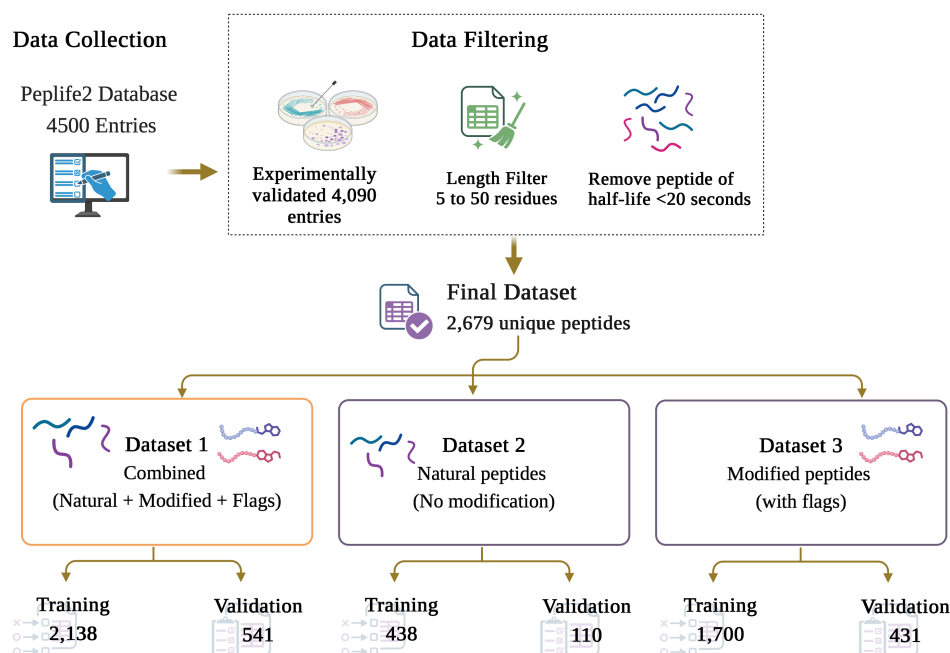


Figure 9.2 The construction of *plifepred2* dataset

9.3.3 Feature extraction

9.3.3.1 Compositional features

To develop predictive models capable of predicting half-life of peptides, it is essential to transform peptide sequences into numerical representations suitable for machine learning algorithms. Therefore, we employed the composition module of the Pfeature tool to compute a diverse set of sequence-derived descriptors, as detailed in *Section 3.7.1*. Specifically, we generated multiple composition-based features, including AAC (20), APAAC (23), BTC (4), ATC (5), CeTD (189), DDOR (20), DPC(400), PAAC (21), PCP (30), QSO (42), RRI (20), SOC (2), SER (20), SPC (25), CTC (343), and SEP (1). In addition to individual descriptors, a hybrid feature set combining all the above composition-based features (listed in Table 3.3, *Section 3.7.1*) was also constructed to capture complementary sequence information and improve model prediction capability.

9.3.3.2 Properties of amino acids

From the protein sequences, several basic properties were calculated to provide informative descriptors for model development. These included molecular weight, sequence length, isoelectric point (pI), and the percentage composition of each amino acid in the sequence. Incorporating these features alongside other descriptors may therefore enhance the model's ability to predict half-life of peptides.

9.3.3.3 Large Language model embeddings

We used pretrained protein language model (PLM)-derived embeddings to describe sequence data. ESM-2 and ProtT5 were used in this study to create protein sequence embeddings (explained in *Sections 3.7.4 and 3.9.2*). These transformer-based models are trained on large-scale protein sequence corpora and learn contextual representations that capture structural, functional, and evolutionary information encoded within amino acid sequences. In this work, we create high-dimensional, context-aware embeddings from input protein sequences using the ESM2 and ProtT5 models. ProtT5 generates an embedding matrix with a size of 1024 for a sequence of length n . Similarly, embeddings from ESM-2 were obtained using different model checkpoints that vary in size and capacity e.g. t6 (320), t12 (480), t30 (640), t33 (1280) available through Hugging Face [313]. Although larger ESM-2 variants such as t36 (2560), and t48 (5120) models may offer richer representations, they require substantial computational memory and training time, which can be impractical for large-scale experiments. Therefore, model size was selected based on a trade-off between representational power and computational feasibility. The resulting embeddings from ProtT5 and ESM-2 were used as input features for prediction models.

9.3.4 Feature Selection

Previous studies have demonstrated that not all input features contribute equally to the performance of an *in silico* predictive model [350]. Therefore, feature selection is an essential step to identify the most informative and biologically relevant descriptors for model construction. In biological analysis, extracting features from a single representation often provides incomplete information, leading to poor predictive performance [130]. To address this limitation, multiple composition-based feature types are often combined to capture complementary sequence characteristics [130]. However, such multi-feature fusion strategies typically produce high-dimensional and redundant feature spaces, which may introduce noise, increase computational burden, and reduce model generalizability [351,352]. In order to increase prediction accuracy, lower overfitting, and improve interpretability, feature selection aims at selecting a subset of appropriate features from a large pool of descriptors [243]. In order to find the most discriminative features and lower dimensionality before model training, we used a number of feature selection techniques in this study. We applied mRMR, RFE-SVR, Boruta, ElasticNet based feature

selection, as described in *Section 3.8*. These methods help eliminate irrelevant and redundant features while preserving the most informative features for prediction method.

9.3.5 ML Regression Development

A number of ML algorithms, such as RFR, XGB, GBR, DTR, SVR, MLP, RID, LGBM, and ET, were used in this study to develop regression models (See *Section 3.9.1*). The training dataset was used to test the effectiveness of these regression algorithms. The Python scikit-learn library was used to implement the models, and the LightGBM framework was used to implement LGBM. During the initial phase, all models were trained with default parameter settings to establish baseline performance and to identify the algorithm with the strongest predictive capability. The best-performing model was subsequently selected for further analysis and hyperparameter optimization.

9.3.6 Evaluation of performance

To evaluate the regression models' performance, five-fold cross-validation was used. This method divided the dataset into five equal subsets, with four of the subsets being used for training and the remaining subset being used as the test set in each iteration. To ensure that each subset was assessed once as an independent test set, this process was carried out five times. The overall cross-validation performance was calculated as the average across all folds [200]. After the model was developed, standard regression metrics such as RMSE, MAE, PCC, and R² were used to further evaluate predictive performance (See *Section 3.13.2*). These metrics were calculated for the independent validation dataset as well as the training dataset.

9.4 Result

A preliminary analysis of the dataset was conducted in order to determine which of these algorithms was the most appropriate, as well as to minimize computational workload and experimental complexity. This initial evaluation allowed us to screen different approaches and focus on the most potential framework. In this study, model performance was calculated across three distinct datasets to understand the impact of sequence modifications on prediction: (i) the combined dataset (natural and modified peptides with flags), (ii) the natural peptide dataset, and (iii) the modified peptide dataset with flags. This comparative framework enabled systematic evaluation of model generalizability across different peptide categories.

9.4.1 Prediction of half-life on the combine dataset

To identify the most informative feature representation for half-life prediction, the performance of individual composition-based descriptors was compared using validation PCC and R^2 , as these metrics reflect both correlation strength and explained variance. First, we have calculated the performance of individual descriptors on different regressors as shown in Figure 9.3 A and 9.3B. It was shown that RFR (Random Forest Regressor) performs the best among all the regressors. Among individual feature sets, CeTD, AAC, APAAC, PAAC, SER, PCP, and DPC showed the strongest predictive performance, achieving validation PCC ~ 0.73 - 0.74 and $R^2 \sim 0.53$ - 0.55 , indicating good correlation between predicted and experimental half-life values.

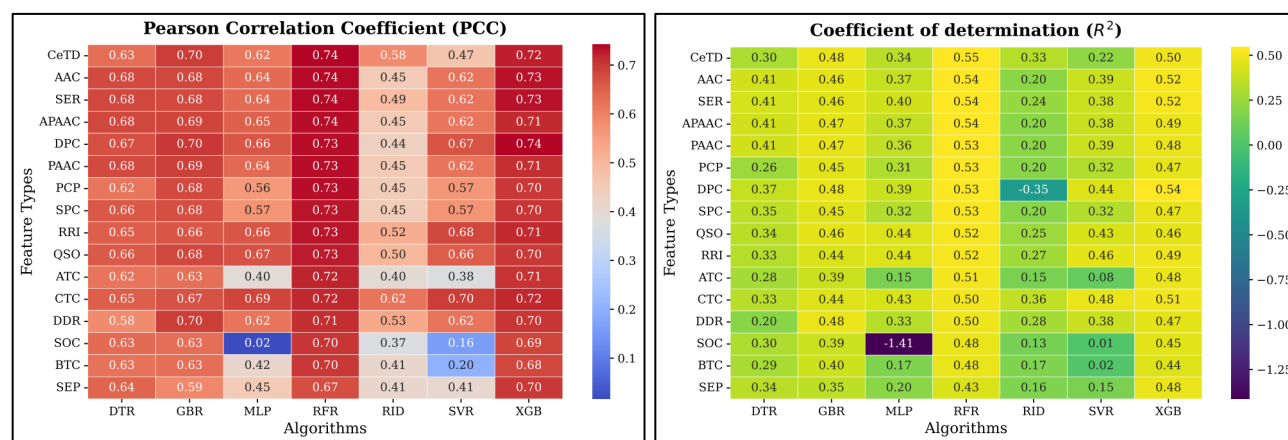


Figure 9.3 Performance of composition-based descriptors on the combined peptide with flags

9.4.2 Prediction of half-life on the natural dataset

To evaluate model performance specifically on peptides without modifications, half-life prediction models were developed using the natural peptide dataset (Dataset 2). This dataset comprised peptides composed solely of standard amino acids, without D-amino acid substitutions, terminal modifications, cyclization, or other post-translational modifications. The same computational pipeline used for the combined datasets was applied here, including sequence-based feature extraction, and regression model training. Performance was evaluated using five-fold cross-validation and an independent validation dataset. The regression models were assessed using PCC, R^2 , RMSE, and MAE. For natural peptides, on validation dataset QSO and CeTD achieved the highest correlation (PCC = 0.71) with $R^2 = 0.49$, closely followed by AAC, and SER, shown in Figure 9.4A and 9.4B. Interestingly, validation performance is higher than training performance for several descriptors, suggesting that the natural dataset is relatively small and may contain

variability that affects training-phase stability. Nevertheless, consistent validation PCC values around 0.69 - 0.71 indicate that sequence-based descriptors alone can reasonably capture determinants of half-life in natural peptides.

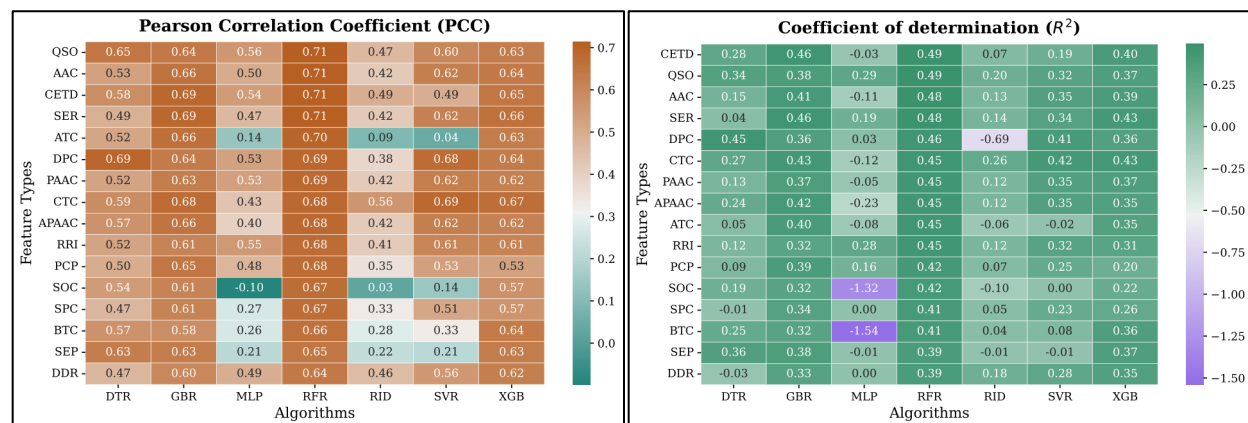


Figure 9.4 Performance of composition-based descriptors on the natural peptide

9.4.3 Prediction of half-life on the modified dataset

In this dataset, model performance was evaluated on peptides annotated with modification flags, as described in Table 9.1. To evaluate the effect of modifications on model performance, half-life models were developed using a modified peptide dataset comprising sequences containing at least one modification (Dataset 3). These included D-amino acid substitutions, N- or C-terminal modifications, cyclization, and other post-translational modifications, represented through predefined flags. The same computational workflow used for the combined dataset was applied to this subset, including feature extraction and regression-based machine learning models. Regression models were trained using composition-based descriptors and flags, and their performance was evaluated using five-fold cross-validation and independent validation. Evaluation metrics included PCC, R^2 , RMSE, and MAE. For peptides containing modifications, PCP (Physicochemical Property Composition) demonstrated the best validation performance (PCC = 0.72, R^2 = 0.51) shown in Figure 9.5(A) and 9.5(B), indicating that physicochemical property distributions play a major role in determining half-life of modified peptides. This suggests that modifications primarily influence peptide stability through changes in hydrophobicity, charge distribution, and other physicochemical traits. Descriptors such as SPC, CeTD, DPC, AAC, APAAC, and RRI also performed strongly (validation PCC ~ 0.70–0.71), showing that both composition and sequence distribution patterns remain important even after chemical

modifications. Compared to the combined dataset, performance values are slightly lower, which is expected since modified peptides introduce additional biochemical variability not fully captured by sequence-only descriptors.

9.4.4 Feature Fusion

To capture complementary sequence information and improve predictive performance, feature fusion was performed for each dataset individually (combined, natural, and modified). In this approach, all composition-based descriptors extracted from the sequences were merged into a unified feature space of 1165, referred to as the ALL_COMP feature set, that includes AAC (20), DPC (400), ATC (5), BTC (4), PAAC (21), APPAC (23), QSO (42), DDR (20), SOCN (2), SEP (1), SER (20), RRI (20), CETD (189), SPC (25), CTC (343), and PCP (30). Regression models developed using the ALL_COMP feature set were evaluated separately on the combined (dataset 1) shown in Table 9.2, natural (dataset 2) shown in Table 9.3, and modified (dataset 3) shown in Table 9.4.

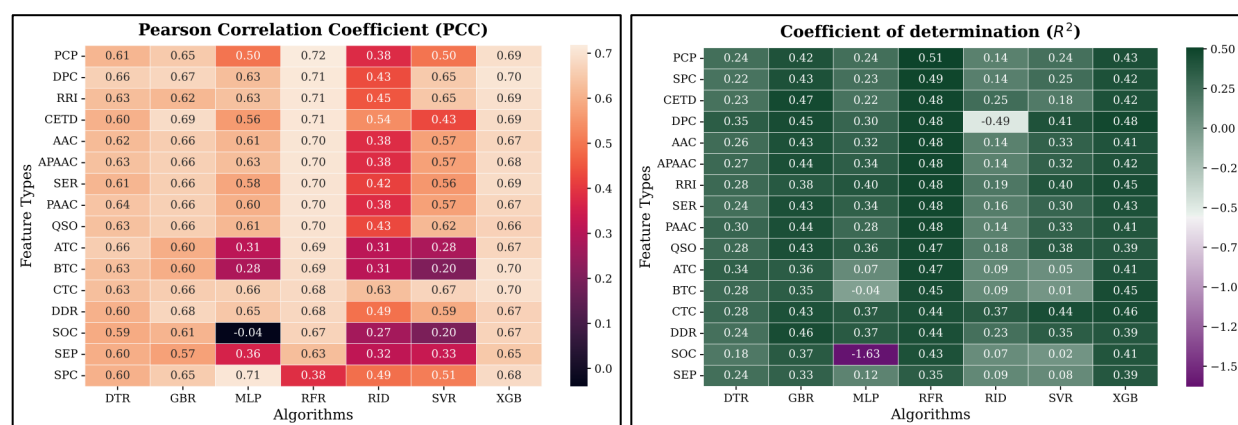


Figure 9.5 Performance of composition-based descriptors on the modified peptide with flags

After evaluating model performance across all datasets (Table 9.2, 9.3, and 9.4), the RFR consistently demonstrated the strong prediction ability. On Dataset 1 (combined dataset with flags), RFR achieved the highest validation performance with $PCC = 0.73$ and $R^2 = 0.53$. For Dataset 2 (natural peptides), the model achieved $PCC = 0.67$ and $R^2 = 0.44$, while for Dataset 3 (modified peptides with flags), performance reached $PCC = 0.71$ and $R^2 = 0.48$. These results clearly indicate that incorporating modification flags improves predictive performance. The combined dataset, which integrates both sequence features and modification-related information, produced the best overall results.

Table 9.2 The performance of different regressors on ALL_COMP on combined peptide with flags

Model	Training Dataset					Validation Dataset				
	MAE	RMSE	PCC	R2	MAPE	MAE	RMSE	PCC	R2	MAPE
RFR	0.52	0.72	0.73	0.52	15.70	0.51	0.70	0.73	0.53	15.97
GBR	0.56	0.74	0.71	0.50	17.01	0.55	0.72	0.72	0.51	17.44
XGB	0.53	0.74	0.72	0.50	15.86	0.52	0.73	0.72	0.50	16.11
DTR	0.58	0.84	0.65	0.34	17.41	0.58	0.81	0.67	0.38	17.34
MLP	0.74	1.01	0.51	0.05	22.83	0.71	0.96	0.54	0.12	20.58
LAS	0.80	0.98	0.36	0.12	24.31	0.79	0.97	0.34	0.10	25.14
SVR	0.82	1.00	0.26	0.07	25.12	0.83	1.01	0.20	0.02	26.39

The half-life label was log10-transformed, so all error metrics are reported and PCC and R² are unitless.

Table 9.3 The performance of different regressors on ALL_COMP on natural peptide

Model	Training Dataset					Validation Dataset				
	MAE	RMSE	PCC	R2	MAPE	MAE	RMSE	PCC	R2	MAPE
RFR	0.56	0.75	0.60	0.35	18.77	0.52	0.69	0.67	0.44	19.40
XGB	0.57	0.78	0.58	0.30	18.79	0.53	0.70	0.66	0.43	19.22
GBR	0.57	0.77	0.58	0.32	18.92	0.53	0.71	0.65	0.41	19.49
DTR	0.60	0.84	0.54	0.19	20.13	0.56	0.81	0.57	0.21	20.83
LAS	0.71	0.87	0.37	0.14	23.56	0.72	0.87	0.35	0.09	26.10
SVR	0.74	0.92	0.20	0.03	24.00	0.75	0.91	0.20	0.01	27.51
MLP	1.07	1.94	0.15	-3.35	35.03	1.15	1.79	0.23	-2.81	39.79

The half-life label was log10-transformed, so all error metrics are reported and PCC and R² are unitless.

Table 9.4 The performance of different regressors on ALL_COMP on modified peptide with flags

Model	Training Dataset					Validation Dataset				
	MAE	RMSE	PCC	R2	MAPE	MAE	RMSE	PCC	R2	MAPE
RFR	0.49	0.69	0.74	0.55	14.38	0.51	0.71	0.71	0.48	15.13
GBR	0.54	0.71	0.72	0.52	15.77	0.55	0.71	0.69	0.47	16.23
XGB	0.50	0.71	0.73	0.52	14.59	0.52	0.74	0.69	0.43	15.51
DTR	0.55	0.79	0.68	0.40	15.89	0.58	0.82	0.64	0.31	16.36
MLP	0.71	1.06	0.47	-0.07	19.93	0.69	0.91	0.56	0.14	18.64
SVR	0.79	0.98	0.31	0.09	23.47	0.77	0.98	0.20	0.01	23.66
RID	0.69	1.25	0.47	-0.48	19.79	0.69	1.14	0.44	-0.35	20.42

The half-life label was log10-transformed, so all error metrics are reported and PCC and R² are unitless.

9.4.5 Comparative Analysis Across the Three Datasets

A comparison of model performance between the natural, modified, and combined peptide datasets provides valuable information about the factors that influence peptide half-life and how modifications affect stability. Top descriptors including CeTD, AAC, SER, and APAAC achieved validation PCC values of approximately 0.74 and $R^2 \approx 0.55$, indicating that the combined dataset (dataset 1) had the best overall predictive performance among the three datasets. By adding flags, the model's capacity to represent modification-driven stability effects is further improved. For the natural peptide (dataset 2), the best performance was observed with sequence-order-based descriptors such as QSO and CeTD (validation PCC ≈ 0.71 , $R^2 \approx 0.49$). Since natural peptides lack flags of modifications, their half-life is more directly governed by compositional features. In contrast, for the modified peptide (dataset 3), descriptors based on physicochemical property composition (PCP) performed best (validation PCC ≈ 0.72 , $R^2 \approx 0.51$). This indicates that stability in modified peptides is strongly influenced by physicochemical properties introduced through terminal modifications, D-amino acids, cyclization, or other PTMs. The combined framework therefore provides the most robust and generalizable approach for peptide half-life prediction.

9.4.6 Models developed on selected features

The combined dataset showed the best predictive performance when the three datasets were compared. Thus, this dataset was used for further model development and optimization. The ALL_COMP feature set obtained from the combined dataset (dataset 3) was subjected to feature selection in order to further increase model efficiency and decrease redundancy. With all composition-based descriptors included, this extensive feature pool offered a rich depiction of sequences. However, noise may be introduced and model generalizability may be diminished due to the high dimensionality of this feature set. Therefore, prior to building the model, feature selection techniques were used to keep only the most discriminative and informative features.

Across all feature subset sizes, validation PCC values remained within a narrow and stable range (0.72–0.74). The best validation performance was observed for 100 features using RFE-SVR (PCC = 0.74, $R^2 = 0.54$) and 800 features using mRMR (PCC = 0.74, $R^2 = 0.54$) shown in Table 9.5. Notably, the 100-feature RFE-SVR model achieved nearly identical performance to larger feature sets, demonstrating that a small, well-selected feature subset can effectively predict peptide half-life. This reduction in dimensionality lowers computational cost and improves model performance.

Overall, RFE-SVR emerged as the most stable feature selection strategy, consistently providing strong performance across different feature sizes.

Table 9.5 Performance of RFR model after feature selection on the combined dataset

No. of Features	Model	Training Dataset				Validation Dataset			
		MAE	RMSE	PCC	R2	MAE	RMSE	PCC	R2
100	RFE SVR	0.52	0.71	0.73	0.53	0.51	0.70	0.74	0.54
	Boruta	0.52	0.72	0.73	0.52	0.51	0.71	0.73	0.53
	MRMR	0.52	0.72	0.72	0.52	0.51	0.71	0.73	0.52
	ElasticNet	0.52	0.71	0.73	0.53	0.53	0.72	0.72	0.50
200	ElasticNet	0.52	0.72	0.73	0.52	0.51	0.70	0.73	0.53
	Boruta	0.52	0.72	0.73	0.52	0.51	0.71	0.73	0.53
	RFE SVR	0.52	0.72	0.73	0.52	0.51	0.71	0.73	0.52
	MRMR	0.52	0.72	0.73	0.52	0.51	0.72	0.72	0.51
400	RFE SVR	0.52	0.72	0.73	0.52	0.51	0.71	0.73	0.53
	ElasticNet	0.52	0.71	0.73	0.53	0.51	0.71	0.73	0.53
	RF	0.52	0.71	0.73	0.53	0.51	0.71	0.73	0.53
	Boruta	0.52	0.72	0.73	0.52	0.51	0.71	0.73	0.53
	MRMR	0.52	0.72	0.73	0.52	0.52	0.71	0.73	0.52
600	RFE SVR	0.52	0.72	0.73	0.52	0.51	0.70	0.73	0.53
	Boruta	0.52	0.72	0.73	0.52	0.51	0.71	0.73	0.53
	MRMR	0.52	0.72	0.73	0.52	0.51	0.71	0.73	0.53
	ElasticNet	0.52	0.72	0.73	0.52	0.51	0.71	0.73	0.52
800	MRMR	0.52	0.72	0.73	0.52	0.51	0.70	0.74	0.54
	Boruta	0.52	0.72	0.72	0.52	0.51	0.71	0.73	0.53
	ElasticNet	0.52	0.72	0.73	0.52	0.51	0.71	0.73	0.53
	RFE SVR	0.52	0.72	0.73	0.52	0.51	0.71	0.73	0.53

9.4.7 Performance on Protein Language Model Embedding

On the combined peptide dataset, we assessed the predictive performance of PLM embeddings in addition to composition sequence descriptors. In order to produce contextual embeddings that directly capture evolutionary, structural, and functional patterns from amino acid sequences, pretrained transformer-based models, such as ESM-2 and ProtT5, were employed. Using pooling techniques (such as mean pooling), residue-level embeddings were produced for every peptide

sequence and combined into fixed-length sequence representations. These embeddings were then used as input features for regression models, following the same training and evaluation protocol applied to composition-based descriptors. The performance of different sets of features using different regression models was shown in Table 9.6.

Table 9.6 Performance of regression models using Protein Language Model embeddings on the combined dataset

Model	Training Dataset				Validation Dataset			
	MAE	RMSE	PCC	R2	MAE	RMSE	PCC	R2
ESM T6								
LGBM	0.52	0.73	0.72	0.51	0.51	0.7	0.74	0.53
RFR	0.53	0.73	0.72	0.51	0.53	0.72	0.72	0.51
GBR	0.56	0.74	0.7	0.50	0.57	0.73	0.7	0.49
ET	0.53	0.74	0.71	0.49	0.53	0.74	0.71	0.49
XGB	0.54	0.76	0.7	0.47	0.54	0.76	0.70	0.46
MLP	0.56	0.76	0.69	0.46	0.56	0.75	0.69	0.46
RID	0.64	0.82	0.61	0.38	0.6	0.79	0.64	0.41
DTR	0.59	0.86	0.64	0.32	0.57	0.81	0.67	0.38
SVR	0.65	0.83	0.6	0.36	0.62	0.81	0.61	0.37
ESM T12								
LGBM	0.53	0.73	0.71	0.50	0.51	0.70	0.73	0.53
RFR	0.53	0.73	0.72	0.51	0.52	0.71	0.72	0.52
GBR	0.56	0.74	0.7	0.49	0.56	0.73	0.71	0.50
ET	0.53	0.74	0.71	0.50	0.53	0.74	0.71	0.49
MLP	0.57	0.77	0.68	0.45	0.54	0.73	0.71	0.49
XGB	0.54	0.75	0.71	0.48	0.55	0.74	0.71	0.47
RID	0.63	0.8	0.63	0.40	0.6	0.78	0.65	0.42
SVR	0.65	0.83	0.6	0.36	0.63	0.82	0.61	0.37
DTR	0.6	0.86	0.63	0.31	0.58	0.83	0.66	0.34
ESM T30								
GBR	0.55	0.73	0.71	0.51	0.55	0.71	0.72	0.52
RFR	0.53	0.72	0.72	0.51	0.53	0.72	0.72	0.51
LGBM	0.52	0.73	0.72	0.51	0.53	0.72	0.72	0.50
ET	0.53	0.74	0.71	0.5	0.53	0.74	0.71	0.48

XGB	0.54	0.75	0.71	0.48	0.54	0.76	0.70	0.46
MLP	0.57	0.78	0.68	0.44	0.55	0.75	0.70	0.46
RID	0.62	0.80	0.64	0.41	0.59	0.78	0.65	0.43
DTR	0.61	0.87	0.63	0.3	0.58	0.81	0.66	0.37
SVR	0.67	0.85	0.57	0.33	0.64	0.83	0.60	0.35
ESM T33								
RFR	0.53	0.73	0.72	0.51	0.53	0.72	0.72	0.51
LGBM	0.53	0.73	0.72	0.51	0.52	0.72	0.72	0.5
GBR	0.56	0.74	0.71	0.5	0.56	0.73	0.7	0.49
ET	0.53	0.74	0.71	0.49	0.54	0.75	0.7	0.47
XGB	0.55	0.76	0.7	0.46	0.54	0.75	0.71	0.47
RID	0.6	0.78	0.66	0.44	0.6	0.77	0.66	0.44
MLP	0.57	0.76	0.69	0.46	0.58	0.77	0.68	0.44
DTR	0.59	0.85	0.64	0.33	0.61	0.85	0.63	0.32
SVR	0.68	0.86	0.57	0.32	0.66	0.85	0.57	0.32
ProtT5								
RFR	0.53	0.72	0.72	0.52	0.52	0.70	0.73	0.53
LGBM	0.53	0.73	0.72	0.50	0.52	0.72	0.72	0.51
GBR	0.55	0.73	0.71	0.50	0.55	0.72	0.71	0.50
ET	0.53	0.74	0.71	0.49	0.53	0.74	0.71	0.49
XGB	0.54	0.75	0.71	0.48	0.54	0.74	0.71	0.48
SVR	0.57	0.76	0.68	0.47	0.55	0.76	0.68	0.46
RID	0.60	0.77	0.67	0.44	0.59	0.77	0.67	0.44
MLP	0.57	0.77	0.69	0.45	0.56	0.79	0.66	0.41
DTR	0.59	0.84	0.65	0.35	0.58	0.83	0.66	0.34

Regression models trained on PLM embeddings demonstrated strong and consistent predictive performance across all ESM-2 layers (T6, T12, T30, T33) and ProtT5. Across all PLM variants, tree-based ensemble methods performed best, such as ESM-T6 + LGBM achieved validation PCC = 0.74 and $R^2 = 0.53$, while ESM-T12 + LGBM and ProtT5 + RFR both validation PCC = 0.73 with $R^2 = 0.53$. These results are comparable to the best composition-based models (PCC $\approx$ 0.74). The evaluation on the combined dataset showed that PLM-based features achieved competitive performance in terms of PCC and R^2 , while maintaining stable error metrics (MAE, RMSE). This

suggests that embeddings derived from large pretrained models can effectively capture determinants of peptide half-life and complement classical feature representations.

9.4.8 Final Model

Among all, the model developed using RFE-SVR with 100 selected features achieved the highest predictive performance, with a validation PCC of 0.74 and R^2 of 0.54 (as shown in Table 9.7). This model demonstrated an optimal use of feature dimensionality, indicating that a compact subset of informative descriptors is sufficient to capture the determinants of peptide half-life. Therefore, this was selected as the final model for implementation in our framework. Based on validation results, the 100-feature RFE-SVR model was selected as the optimal for subsequent model development. Figure 9.6 shows the top 15 contributing features. Prediction of peptide half-life was driven by a diverse set of composition and physicochemical descriptors rather than a single family. The most influential feature by a wide margin was a distribution descriptor for tryptophan (DDR_W, mean $|\text{SHAP}| \approx 0.21$), more than twice the next feature. It was followed by composition-enhanced transition–distribution terms (CeTD_13_SS, CeTD_75_p_SA3, CeTD_HB2, CeTD_12_PZ), conjoint-triad features (CTC_461, CTC_143), a Shannon-entropy residue descriptor (SER_G), quasi-sequence-order terms (QSO1_SC_A, QSO1_SC_P), dipeptide-composition pairs (DPC1_SD, DPC1_LS, DPC1_TV), and shift/entropy-based descriptors (SEP, SEP_SS_CO).

Table 9.7 The complete performance of RFE-SVR 100 features on different regressor

Model	Training Dataset				Validation Dataset			
	MAE	RMSE	PCC	R2	MAE	RMSE	PCC	R2
RFR	0.52	0.71	0.73	0.53	0.51	0.70	0.74	0.54
XGB	0.52	0.73	0.72	0.51	0.53	0.74	0.71	0.49
GBR	0.56	0.74	0.71	0.50	0.57	0.74	0.70	0.48
MLP	0.58	0.78	0.67	0.43	0.61	0.83	0.66	0.35
RID	0.61	0.79	0.65	0.43	0.63	0.83	0.60	0.35
SVR	0.69	0.88	0.54	0.29	0.68	0.88	0.52	0.27
DTR	0.57	0.81	0.67	0.39	0.62	0.90	0.60	0.24

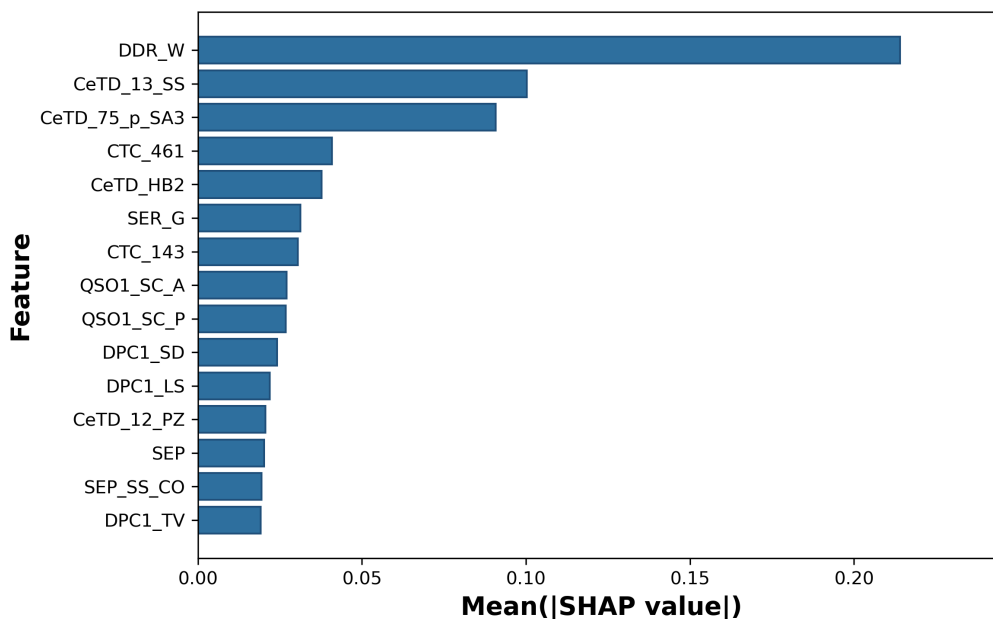


Figure 9.6 illustrates the SHAP-based feature importance for the peptide half-life prediction model (composition-based features).

9.5 Implementation of web-server

To facilitate access for the scientific community, a user-friendly web server was implemented for the developed prediction models (available at <https://webs.iiitd.edu.in/raghava/plifepred2>). The server provides an platform for predicting peptide half-life directly from sequence input, eliminating the need for users to perform complex feature extraction or machine learning computations. The web server backend was developed using a combination of HTML5, CSS, Java, PHP. Upon submission of peptide sequences in FASTA format, the server performs the following steps: (i) Prediction, (ii) Design, using the optimized regression model. The predicted output is provided in a tabular format, along with relevant computed properties. This tool provides a convenient resource for researchers working in peptide therapeutics, vaccine design, and protein engineering by enabling rapid in silico evaluation of peptide stability.

9.6 Discussion and Conclusion

Despite the growing interest in peptide-based drugs, their therapeutic potential is frequently hindered by rapid proteolytic degradation resulting in a short half-life that remains a major obstacle to their development [200]. The present study addresses this major challenge by developing PlifePred2, a web server for predicting peptide half-life using sequence-derived features, modification flags, and protein language model (PLM) embeddings. A major focus of this work is

the significant role of modifications in determining peptide half-life. Comparative evaluation across the three datasets revealed that the combined dataset (natural + modified peptides with flags) consistently showed the highest predictive performance. Models trained on this dataset achieved validation $PCC \approx 0.73-0.74$ and $R^2 \approx 0.53-0.55$, outperforming models trained solely on natural peptides. This demonstrates that flags encoding D-amino acids, terminal modifications, cyclization, and PTMs carry information predictive of half-life that sequence composition alone does not capture.

The composition-based descriptors CeTD, AAC, APAAC, PAAC, SER, and PCP were the most informative. PCP was the most effective descriptor for modified peptides, whereas QSO and CeTD were the best for natural peptides. By incorporating complementary sequence information, feature fusion with the ALL_COMP representation further increased the robustness of the model. However, feature selection analysis showed that predictive performance remained stable across feature subset sizes. The RFE-SVR 100-feature model achieved the best validation performance ($PCC = 0.74$, $R^2 = 0.54$) that helps in reducing the computational complexity. PLM embeddings (ESM-2 and ProtT5) produced performance comparable to compositional descriptors. Tree-based regressors (LGBM, RFR) performed best on these embeddings. This study presents PlifePred2 (<https://webs.iiitd.edu.in/raghava/plifepred2>), an machine learning method for predicting the half-life of natural and modified peptides using the modification flags. The publicly accessible web server further enhances the utility of our method that helps in prediction of half-life and designing of peptide with desired half-life.

C H A P T E R

10

Summary

This work provides a foundation for computational subunit vaccine design, significantly accelerating the development of peptide-based vaccines and protein therapeutics. Identification of peptides that are responsible for inducing type 1 diabetes mellitus. Prediction of allergens to optimize therapeutic design. Designing and predicting antifreeze proteins, thermostable proteins, and the half-life of proteins and peptides to enable stability, safe storage, and distribution without cold-chain dependence. Together, these modules form an combined computational framework for screening, designing, and optimizing therapeutic proteins for biomedical applications. Modern vaccine research increasingly focuses on subunit and peptide vaccines because they are safer and more specific than whole-pathogen vaccines. We use standard bioinformatics methodologies across all modules, including curated datasets, redundancy removal, sequence-based feature extraction, protein language model embeddings, feature selection, supervised learning, and independent validation.

In chapter 1 we introduces the biological background of immunity and vaccination. It explains innate and adaptive immune responses and highlights the role of B-cell epitopes in antibody production and vaccine development. Traditional vaccines based on live or inactivated pathogens are effective but raise safety and manufacturing concerns. Subunit and peptide vaccines offer safer alternatives but require precise epitope selection and stability optimization. The chapter defines the central aim of the thesis: to develop computational tools for identifying antigenic peptides, evaluating safety, and improving stability for therapeutic applications.

In chapter 2 we review experimental and computational approaches for epitope prediction, allergen prediction, T1DM associated peptide identification, protein stability, antifreeze proteins, and peptide half-life prediction. It describes databases such as IEDB, UniProt, and PDB and discusses limitations of existing methods, including reliance on small datasets, homology-based predictions. Identifies major research gaps and limited tools for stability and prediction.

Chapter 3 outlines the general computational workflow. Datasets were collected from curated biological databases and literature and processed using redundancy removal (e.g., CD-HIT clustering). Sequences feature calculated using composition-based descriptors, binary profiles, and protein language model embeddings. Feature selection techniques and machine learning algorithms were applied for model construction, and evaluation was performed using cross-validation and independent validation. The methodology provides a standardized framework used across all subsequent prediction modules.

Chapter 4 presents CLBTope, a method for predicting both linear and conformational B-cell epitopes from primary sequence. Because B-cell epitopes are the targets of antibody responses, their identification is the first step in subunit vaccine design. Models were trained on 7,871 curated peptides, and a Random Forest classifier using dipeptide composition combined with BLAST similarity scores achieved an AUROC of 0.829 on the independent validation set, exceeding published predictors evaluated on the same data.

Chapter 5 presents DMPPred for the identification of peptides associated with type 1 diabetes mellitus. Antigenic peptides that mimic human proteins can trigger autoimmune responses in predisposed individuals, making this a safety concern in vaccine design. An Extra Trees classifier on 150 mRMR-selected dipeptide features, combined with BLAST, reached an AUROC of 0.951. Chapter 6 presents AlgPred3 for predicting allergenic peptides rather than whole-proteins. Composition-based descriptors, binary profiles, protein language models and an exploratory quantum SVM were compared on a balanced set of 7,388 peptides. An Extra Trees classifier on 273 selected features gave the best validation performance, with an AUROC of 0.81 and an MCC of 0.44.

Chapter 7 presents PPTStab for prediction of protein melting temperature (T_m). Protein stability is crucial for pharmaceutical storage and transportation, especially in regions lacking reliable cold-chain infrastructure. After redundancy reduction at 40 percent identity, 17,312 non-redundant proteins were used to train regression models. An ensemble of a custom ANN and an MLP regressor on ProtBERT embeddings achieved an R-squared of 0.80 and a Pearson correlation of 0.89, comparable to the best existing tool despite stricter dataset curation.

Chapter 8 presents AFProPred for the identification of antifreeze proteins, which inhibit ice crystal growth and are relevant to cryopreservation and cold storage of biologics. Because antifreeze proteins share little sequence identity, similarity search alone performed poorly. Adding PSSM evolutionary profiles to amino acid composition raised validation AUROC from 0.90 to 0.93, with an accuracy of 88.24 percent and an MCC of 0.77.

Chapter 9 presents PlifePred2 for prediction of peptide half-life, a determinant of therapeutic efficacy. Models were trained on 2,679 peptides with experimentally measured half-lives, using sequence descriptors together with binary flags encoding D-amino acids, terminal modifications, cyclisation and post-translational modification. A recursive feature elimination model retaining 100 features gave a validation Pearson correlation of 0.74. Including modification flags improved

prediction over sequence composition alone, confirming that chemical modification carries information not recoverable from sequence.

Taken together these methods establishes an computational framework for designing safe, stable, and effective peptide-based therapeutics. The developed modules sequentially perform epitope identification, autoimmune and allergen filtering, and stability optimization (thermal stability, cold resistance, and half-life). The framework significantly reduces experimental screening effort by prioritizing high-quality therapeutic candidates computationally. It supports next-generation subunit vaccine development, therapeutic peptide design, and protein engineering. By integrating artificial intelligence with biological knowledge, the thesis contributes toward precision vaccinology and personalized therapeutic design. Ultimately, the proposed approach provides a practical strategy for accelerating drug discovery while improving safety, efficacy, and storage feasibility of protein-based therapeutics.

10.1 Limitations and Future Directions

Despite the encouraging performance of the developed prediction tools, the present study has several limitations that provide opportunities for future research.

Common limitations

- 1. Completely external validation:** All models were evaluated on a 20% hold-out independent validation set. Protein datasets were split 80:20 into training and validation sets after redundancy reduction, and peptide datasets were de-duplicated before splitting. These hold-out sets were not used during training, so they give an unbiased model performance. Although, this does not completely remove the possibility of similarity between the training and validation sets. Therefore, future work should use completely external validation datasets drawn from independent sources, or family-wise partitioning, to more strictly assess model generalizability.
- 2. Negative datasets:** For several tasks, experimentally validated true negatives are not available in large numbers. One of the major reason for this is that most of the biologist record only positive results and ignoring the record for negative dataset. Therefore, negative datasets were built by curated sampling, such as length-matched random sampling from Swiss-Prot after redundancy reduction. This is a standard approach, but it may not fully represent the diversity of true negatives. Future work should use experimentally validated and more carefully curated negative sets to improve reliability.

Tool-specific limitations:

- 1. CLBTope (B-cell epitopes, Chapter 4):** Most datasets came from a single curated source (Ras-Carmona et al.; BCETD, ILED, IDED). This may carry over any annotation bias from that source. Also, conformational epitope prediction is complex than linear-epitope prediction, because it depends on three-dimensional structure that sequence features capture only indirectly.
- 2. DMPPred (T1DM peptides, Chapter 5):** The experimentally validated positives from IEDB were limited (815 non-redundant peptides), which restricts the training data. As true experimental negatives do not exist in large numbers, an equal number of length-matched negatives was sampled at random from Swiss-Prot. This balances the classes but shares the common negative-dataset limitation.
- 3. AlgPred3 (allergenicity, Chapter 6):** The negative dataset was less rigorously curated than the positive dataset. Also, sequence- and composition-based prediction captures linear determinants well but does not fully model conformational IgE epitopes that depend on structure.
- 4. PPTStab (thermostability, Chapter 7):** The model was trained on proteins from many taxonomic groups to make it general. This improves applicability across organisms, but organism-specific analysis was beyond the present scope; future work should build taxon-specific predictors to find lineage-dependent factors.
- 5. AFProPred (antifreeze proteins, Chapter 8):** The main limitation is the small well-annotated evaluation set: the reviewed Swiss-Prot database gave only 80 non-redundant AFPs, which with 73 non-AFPs is too small for robust model development. BLAST-based prediction worked poorly due to low sequence similarity among AFPs, motif-based detection covered only a few sequences, and PSSM profiles could not be generated for 7 AFPs that had no detectable homologs. Future work should build larger, experimentally validated AFP datasets.
- 6. PLifePred2 (peptide half-life, Chapter 9):** The protein-language-model embeddings are high-dimensional and not directly interpretable, which limits biological insight into the predictions. Interpretability methods such as SHAP or attention analysis were not applied here and remain an important extension.

Future Directions:

The predictive modules developed in this thesis for epitope identification, autoimmune (T1DM) and allergenicity filtering, and stability optimisation (thermostability, cold/freezing resistance, and half-life) constitute a discriminative pipeline that prioritises candidates from existing sequence space. Several complementary directions would extend this work from candidate filtering toward end-to-end candidate generation and validation.

Generative AI for vaccine design: The tools presented here are discriminative, they predict the properties of existing sequences. A natural next step is to couple them with generative models (such as diffusion models, variational autoencoders, or protein-language-model-based generators) that design novel epitopes and peptides *de novo*. In such a pipeline, the predictors developed in this thesis: immunogenicity, allergenicity, autoimmune risk, thermostability, and half-life, could act as scoring functions that steer generation toward candidates that are simultaneously immunogenic, safe, and stable, moving from prediction to rational design.

Multimodal learning: Most of the present tools rely on sequence-derived features. Future models could integrate several data types such as sequence, predicted structure, physicochemical properties, and omics data such as expression or HLA information, within a single multimodal framework. Combining complementary modalities is likely to improve both accuracy and biological interpretability. Extending PLifePred2 and AlgPred3 to fuse structural and assay-derived modalities is a promising route to higher accuracy and transferability.

Structure-based AI: The recent availability of accurate structure predictors (e.g. AlphaFold2/3 and ESMFold) makes it feasible to add structural information to these tools. Structure-aware models would be particularly valuable for conformational B-cell epitope prediction and for stability tasks, where surface accessibility, solvent exposure, and three-dimensional packing are directly relevant.

Wet-lab validation pipelines: Finally, an important direction is to close the loop between computation and experiment. Predicted epitopes and peptides could be synthesised and tested experimentally for immunogenicity and allergenicity, and for stability properties such as melting temperature and half-life, giving prospective validation of the models and strengthening their translational value. Establishing such computational-experimental pipelines would be a key step toward applying these tools in real vaccine and therapeutic development.

Large biological foundation models: Protein language models such as ESM-2 and ProtT5, already used for embeddings in this thesis (Chapters 6, 7, 9), are early instances of a broader shift toward large biological foundation models trained across proteins, genomes, and functional data. Fine-tuning such models on task-specific labels rather than using frozen embeddings and exploiting their emergent capabilities is a natural extension that could improve performance on the small, curated datasets that limit several chapters of this work [353,354].

Bibliography

- [1] R. Wang, C. Lan, K. Benlagha, N.O.S. Camara, H. Miller, M. Kubo, S. Heegaard, P. Lee, L. Yang, H. Forsman, X. Li, Z. Zhai, C. Liu, The interaction of innate immune and adaptive immune system, *MedComm* 5 (2024) e714. <https://doi.org/10.1002/mco2.714>.
- [2] A.S. De Groot, L. Moise, J.A. McMurry, W. Martin, Epitope-based immunome-derived vaccines: A strategy for improved design and safety, in: *Clinical Applications of Immunomics*, Springer US, New York, NY, 2009: pp. 39–69. https://doi.org/10.1007/978-0-387-79208-8_3.
- [3] In brief: The innate and adaptive immune systems, Institute for Quality and Efficiency in Health Care (IQWiG), 2023. <https://www.ncbi.nlm.nih.gov/books/NBK279396/> (accessed February 24, 2026).
- [4] B. Aristizábal, Á. González, *Innate immune system*, El Rosario University Press, 2013. <https://www.ncbi.nlm.nih.gov/books/NBK459455/> (accessed February 24, 2026).
- [5] F. Spooner, S. Dattani, S. Vanderslott, M. Roser, Vaccination, *Our World in Data* (2022). <https://ourworldindata.org/vaccination> (accessed February 24, 2026).
- [6] R.-A.R. María, C.-V.J. Arturo, J.-A. Alicia, M.-L.G. Paulina, A.-O. Gerardo, The impact of bioinformatics on vaccine design and development, in: *Vaccines*, InTech, 2017. <https://doi.org/10.5772/intechopen.69273>.
- [7] A.J. Shattock, H.C. Johnson, S.Y. Sim, A. Carter, P. Lambach, R.C.W. Hutubessy, K.M. Thompson, K. Badizadegan, B. Lambert, M.J. Ferrari, M. Jit, H. Fu, S.P. Silal, R.A. Hounsell, R.G. White, J.F. Mosser, K.A.M. Gaythorpe, C.L. Trotter, A. Lindstrand, K.L. O'Brien, N. Bar-Zeev, Contribution of vaccination to improved survival and health: modelling 50 years of the Expanded Programme on Immunization, *Lancet* 403 (2024) 2307–2316. [https://doi.org/10.1016/S0140-6736\(24\)00850-X](https://doi.org/10.1016/S0140-6736(24)00850-X).
- [8] G.Y. Scott, A.T. Aborode, R.O. Adesola, M. Khan, E.E. Elebesunu, A.A. Oluwaseun, M. Ahmad, E.O. Nibokun, S.D. Kwaku, A.A. Bakre, P.N. Chukwu, Z. Olapade, A.I. Osu, I.A. Onifade, O.A. Ogundijo, J. Olaoye, T. Bakare-Abidola, S. Fakorede, S. Olapade, Enhancing vaccine quality and accessibility: Strategies for efficient storage and distribution in resource-constrained environment, *Outbreak Management and Response* 1 (2025). <https://doi.org/10.1080/29947677.2024.2443439>.
- [9] M.R. Hilleman, Vaccines in historic evolution and perspective: a narrative of vaccine discoveries, *Vaccine* 18 (2000) 1436–1447. [https://doi.org/10.1016/s0264-410x\(99\)00434-x](https://doi.org/10.1016/s0264-410x(99)00434-x).
- [10] Y.L. Vishweshwaraiah, N.V. Dokholyan, Toward rational vaccine engineering, *Adv. Drug Deliv. Rev.* 183 (2022) 114142. <https://doi.org/10.1016/j.addr.2022.114142>.
- [11] A. Nandy, S.C. Basak, A brief review of computer-assisted approaches to rational design of peptide vaccines, *Int. J. Mol. Sci.* 17 (2016) 666. <https://doi.org/10.3390/ijms17050666>.
- [12] Clinicaltrials.gov, (n.d.). <https://clinicaltrials.gov/search?term=peptide> (accessed February 24, 2026).
- [13] Z. Vajo, The seasonal influenza vaccine Agriflu(®), *Expert Rev. Vaccines* 10 (2011) 1513–1517. <https://doi.org/10.1586/erv.11.133>.
- [14] S. Romeli, S.S. Hassan, W.B. Yap, Multi-Epitope peptide-based and Vaccinia-based universal influenza vaccine candidates subjected to clinical trials, *Malays. J. Med. Sci.* 27 (2020) 10–20. <https://doi.org/10.21315/mjms2020.27.2.2>.

- [15] S.-S. Wong, R.J. Webby, Traditional and new influenza vaccines, *Clin. Microbiol. Rev.* 26 (2013) 476–492. <https://doi.org/10.1128/CMR.00097-12>.
- [16] F. Janett, T. Gerig, A.C. Tschuor, S. Amatayakul-Chantler, J. Walker, R. Howard, M. Piechotta, H. Bollwein, S. Hartnack, R. Thun, Effect of vaccination against gonadotropin-releasing factor (GnRF) with Bopriva® in the prepubertal bull calf, *Anim. Reprod. Sci.* 131 (2012) 72–80. <https://doi.org/10.1016/j.anireprosci.2012.02.012>.
- [17] A.N. Tsoras, K.M. Wong, A.K. Paravastu, J.A. Champion, Rational design of antigen incorporation into subunit vaccine biomaterials can enhance antigen-specific immune responses, *Front. Immunol.* 11 (2020) 1547. <https://doi.org/10.3389/fimmu.2020.01547>.
- [18] I. Mayrose, O. Penn, E. Erez, N.D. Rubinstein, T. Shlomi, N.T. Freund, E.M. Bublil, E. Ruppín, R. Sharan, J.M. Gershoni, E. Martz, T. Pupko, Pepitope: epitope mapping from affinity-selected peptides, *Bioinformatics* 23 (2007) 3244–3246. <https://doi.org/10.1093/bioinformatics/btm493>.
- [19] N. Kumar, S. Tripathi, N. Sharma, S. Patiya, N.L. Devi, G.P.S. Raghava, A method for predicting linear and conformational B-cell epitopes in an antigen from its primary sequence, *Comput. Biol. Med.* 170 (2024) 108083. <https://doi.org/10.1016/j.compbiomed.2024.108083>.
- [20] Y.-T. Lo, T.-C. Shih, T.-W. Pai, L.-P. Ho, J.-L. Wu, H.-Y. Chou, Conformational epitope matching and prediction based on protein surface spiral features, *BMC Genomics* 22 (2021) 116. <https://doi.org/10.1186/s12864-020-07303-5>.
- [21] K.A. Galanis, K.C. Nastou, N.C. Papandreou, G.N. Petichakis, D.G. Pigis, V.A. Iconomidou, Linear B-cell Epitope prediction for in silico vaccine design: A performance review of methods available via command-line interface, *Int. J. Mol. Sci.* 22 (2021) 3210. <https://doi.org/10.3390/ijms22063210>.
- [22] Y. Segal, Y. Shoenfeld, Vaccine-induced autoimmunity: the role of molecular mimicry and immune crossreaction, *Cell. Mol. Immunol.* 15 (2018) 586–594. <https://doi.org/10.1038/cmi.2017.151>.
- [23] J.C. Bischof, X. He, Thermal stability of proteins, *Ann. N. Y. Acad. Sci.* 1066 (2005) 12–33. <https://doi.org/10.1196/annals.1363.003>.
- [24] M. Verdecia, J.F. Kokai-Kun, M. Kibbey, S. Acharya, J. Venema, F. Atouf, COVID-19 vaccine platforms: Delivering on a promise?, *Hum. Vaccin. Immunother.* 17 (2021) 2873–2893. <https://doi.org/10.1080/21645515.2021.1911204>.
- [25] B. Autran, G. Carcelain, B. Combadiere, P. Debre, Therapeutic vaccines for chronic infections, *Science* 305 (2004) 205–208. <https://doi.org/10.1126/science.1100600>.
- [26] L.H. Butterfield, Cancer vaccines, *BMJ* 350 (2015) h988. <https://doi.org/10.1136/bmj.h988>.
- [27] M. Skwarczynski, I. Toth, Peptide-based synthetic vaccines, *Chem. Sci.* 7 (2016) 842–854. <https://doi.org/10.1039/c5sc03892h>.
- [28] C. Verma, V.A. Pawar, S. Srivastava, A. Tyagi, G. Kaushik, S.K. Shukla, V. Kumar, Cancer vaccines in the immunotherapy era: Promise and potential, *Vaccines (Basel)* 11 (2023) 1783. <https://doi.org/10.3390/vaccines11121783>.
- [29] A. Pal, A.K. Chakravarty, Basic concepts of immunogenetics, in: *Genetics and Breeding for Disease Resistance of Livestock*, Elsevier, 2020: pp. 95–100. <https://doi.org/10.1016/b978-0-12-816406-8.00004-8>.
- [30] H. Goodarzi, J. Trowbridge, R.L. Gallo, Innate immunity: a cutaneous perspective, *Clin. Rev. Allergy Immunol.* 33 (2007) 15–26. <https://doi.org/10.1007/s12016-007-0037-4>.

- [31] A. Ras-Carmona, A.A. Lehmann, P.V. Lehmann, P.A. Reche, Prediction of B cell epitopes in proteins using a novel sequence similarity-based method, *Sci. Rep.* 12 (2022) 13739. <https://doi.org/10.1038/s41598-022-18021-1>.
- [32] J.L. Sanchez-Trincado, M. Gomez-Perosanz, P.A. Reche, Fundamentals and methods for T- and B-cell Epitope prediction, *J. Immunol. Res.* 2017 (2017) 2680160. <https://doi.org/10.1155/2017/2680160>.
- [33] P. Pattnaik, Surface plasmon resonance: applications in understanding receptor-ligand interaction, *Appl. Biochem. Biotechnol.* 126 (2005) 79–92. <https://doi.org/10.1385/abab:126:2:079>.
- [34] C. Hager-Braun, K.B. Tomer, Determination of protein-derived epitopes by mass spectrometry, *Expert Rev. Proteomics* 2 (2005) 745–756. <https://doi.org/10.1586/14789450.2.5.745>.
- [35] M.A. Johnson, B.M. Pinto, NMR spectroscopic and molecular modeling studies of protein-carbohydrate and protein-peptide interactions, *Carbohydr. Res.* 339 (2004) 907–928. <https://doi.org/10.1016/j.carres.2003.12.028>.
- [36] J. Ponomarenko, M. Van Regenmortel, B cell epitope prediction, *Structural Bioinformatics* (2009).
- [37] N. Kumar, N. Bajiya, S. Patiyal, G.P.S. Raghava, Multi-perspectives and challenges in identifying B-cell epitopes, *Protein Sci.* 32 (2023) e4785. <https://doi.org/10.1002/pro.4785>.
- [38] T.P. Hopp, K.R. Woods, Prediction of protein antigenic determinants from amino acid sequences, *Proc. Natl. Acad. Sci. U. S. A.* 78 (1981) 3824–3828. <https://doi.org/10.1073/pnas.78.6.3824>.
- [39] J. Janin, S. Wodak, Conformation of amino acid side-chains in proteins, *J. Mol. Biol.* 125 (1978) 357–386. [https://doi.org/10.1016/0022-2836\(78\)90408-4](https://doi.org/10.1016/0022-2836(78)90408-4).
- [40] M. Levitt, Conformational preferences of amino acids in globular proteins, *Biochemistry* 17 (1978) 4277–4285. <https://doi.org/10.1021/bi00613a026>.
- [41] P.K. Ponnuswamy, M. Prabhakaran, P. Manavalan, Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins, *Biochim. Biophys. Acta* 623 (1980) 301–316. [https://doi.org/10.1016/0005-2795\(80\)90258-5](https://doi.org/10.1016/0005-2795(80)90258-5).
- [42] P.A. Karplus, G.E. Schulz, Prediction of chain flexibility in proteins, *Sci. Nat.* 72 (1985) 212–213. <https://doi.org/10.1007/bf01195768>.
- [43] E.A. Emini, J.V. Hughes, D.S. Perlow, J. Boger, Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide, *J. Virol.* 55 (1985) 836–839. <https://doi.org/10.1128/JVI.55.3.836-839.1985>.
- [44] J.M. Parker, D. Guo, R.S. Hodges, New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites, *Biochemistry* 25 (1986) 5425–5432. <https://doi.org/10.1021/bi00367a013>.
- [45] A.S. Kolaskar, P.C. Tongaonkar, A semi-empirical method for prediction of antigenic determinants on protein antigens, *FEBS Lett.* 276 (1990) 172–174. [https://doi.org/10.1016/0014-5793\(90\)80535-q](https://doi.org/10.1016/0014-5793(90)80535-q).
- [46] J.L. Pellequer, E. Westhof, M.H. Van Regenmortel, Correlation between the location of antigenic sites and the prediction of turns in proteins, *Immunol. Lett.* 36 (1993) 83–99. [https://doi.org/10.1016/0165-2478\(93\)90072-a](https://doi.org/10.1016/0165-2478(93)90072-a).
- [47] J.E.P. Larsen, O. Lund, M. Nielsen, Improved method for predicting linear B-cell epitopes, *Immunome Res.* 2 (2006) 2. <https://doi.org/10.1186/1745-7580-2-2>.

- [48] Y. El-Manzalawy, V. Honavar, Recent advances in B-cell epitope prediction methods, *Immunome Res.* 6 Suppl 2 (2010) S2. <https://doi.org/10.1186/1745-7580-6-S2-S2>.
- [49] J.L. Pellequer, E. Westhof, M.H. Van Regenmortel, Predicting location of continuous epitopes in proteins from their primary structures, *Methods Enzymol.* 203 (1991) 176–201. [https://doi.org/10.1016/0076-6879\(91\)03010-e](https://doi.org/10.1016/0076-6879(91)03010-e).
- [50] J.L. Pellequer, E. Westhof, PREDITOP: a program for antigenicity prediction, *J. Mol. Graph.* 11 (1993) 204–10, 191–2. [https://doi.org/10.1016/0263-7855\(93\)80074-2](https://doi.org/10.1016/0263-7855(93)80074-2).
- [51] A.J. Alix, Predictive estimation of protein linear epitopes by using the program PEOPLE, *Vaccine* 18 (1999) 311–314. [https://doi.org/10.1016/s0264-410x\(99\)00329-1](https://doi.org/10.1016/s0264-410x(99)00329-1).
- [52] M. Odorico, J.-L. Pellequer, BEPITOPE: predicting the location of continuous epitopes and patterns in proteins, *J. Mol. Recognit.* 16 (2003) 20–22. <https://doi.org/10.1002/jmr.602>.
- [53] S. Saha, G.P.S. Raghava, BcePred: Prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties, in: *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004: pp. 197–204. https://doi.org/10.1007/978-3-540-30220-9_16.
- [54] I.I. Davydov, A.G. Tonevitskiĭ, Linear B-cell epitope prediction, *Mol. Biol. (Mosk.)* 43 (2009) 166–174. <https://doi.org/10.1134/s0026893309010221>.
- [55] M.J. Blythe, D.R. Flower, Benchmarking B cell epitope prediction: underperformance of existing methods, *Protein Sci.* 14 (2005) 246–248. <https://doi.org/10.1110/ps.041059505>.
- [56] S. Saha, G.P.S. Raghava, Prediction of continuous B-cell epitopes in an antigen using recurrent neural network, *Proteins* 65 (2006) 40–48. <https://doi.org/10.1002/prot.21078>.
- [57] J. Söllner, B. Mayer, Machine learning approaches for prediction of linear B-cell epitopes on proteins, *J. Mol. Recognit.* 19 (2006) 200–208. <https://doi.org/10.1002/jmr.771>.
- [58] H. Singh, H.R. Ansari, G.P.S. Raghava, Improved method for linear B-cell epitope prediction using antigen's primary sequence, *PLoS One* 8 (2013) e62216. <https://doi.org/10.1371/journal.pone.0062216>.
- [59] Y. El-Manzalawy, D. Dobbs, V. Honavar, Predicting flexible length linear B-cell epitopes, *Comput. Syst. Bioinformatics Conf.* 7 (2008) 121–132. <https://www.ncbi.nlm.nih.gov/pubmed/19642274>.
- [60] Y. El-Manzalawy, D. Dobbs, V. Honavar, Predicting linear B-cell epitopes using string kernels, *J. Mol. Recognit.* 21 (2008) 243–255. <https://doi.org/10.1002/jmr.893>.
- [61] M.J. Sweredoski, P. Baldi, COBepro: a novel system for predicting continuous B-cell epitopes, *Protein Eng. Des. Sel.* 22 (2009) 113–120. <https://doi.org/10.1093/protein/gzn075>.
- [62] L.J.K. Wee, D. Simarmata, Y.-W. Kam, L.F.P. Ng, J.C. Tong, SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction, *BMC Genomics* 11 Suppl 4 (2010) S21. <https://doi.org/10.1186/1471-2164-11-S4-S21>.
- [63] H.-W. Wang, Y.-C. Lin, T.-W. Pai, H.-T. Chang, Prediction of B-cell linear epitopes with a combination of support vector machine classification and amino acid propensity identification, *J. Biomed. Biotechnol.* 2011 (2011) 432830. <https://doi.org/10.1155/2011/432830>.
- [64] Y. Wang, W. Wu, N.N. Negre, K.P. White, C. Li, P.K. Shah, Determinants of antigenicity and specificity in immune response for protein sequences, *BMC Bioinformatics* 12 (2011) 251. <https://doi.org/10.1186/1471-2105-12-251>.
- [65] B. Yao, L. Zhang, S. Liang, C. Zhang, SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity, *PLoS One* 7 (2012) e45152. <https://doi.org/10.1371/journal.pone.0045152>.

- [66] J. Gao, E. Faraggi, Y. Zhou, J. Ruan, L. Kurgan, BEST: improved prediction of B-cell epitopes from antigen sequences, *PLoS One* 7 (2012) e40104. <https://doi.org/10.1371/journal.pone.0040104>.
- [67] Y. Lian, M. Ge, X.-M. Pan, EPMLR: sequence-based linear B-cell epitope prediction method using multiple linear regression, *BMC Bioinformatics* 15 (2014) 414. <https://doi.org/10.1186/s12859-014-0414-y>.
- [68] S. Vijayakumar, G. Namasivayam, Harnessing computational biology for exact linear B-cell Epitope prediction: A novel amino acid composition-based feature descriptor, *OMICS* 19 (2015) 648–658. <https://doi.org/10.1089/omi.2015.0095>.
- [69] Y. Lian, Z.C. Huang, M. Ge, X.M. Pan, An improved method for predicting linear B-cell Epitope using deep maxout networks, *Biomed. Environ. Sci.* 28 (2015) 460–463. <https://doi.org/10.3967/bes2015.065>.
- [70] G. Sher, D. Zhi, S. Zhang, DRREP: deep ridge regressed epitope predictor, *BMC Genomics* 18 (2017) 676. <https://doi.org/10.1186/s12864-017-4024-8>.
- [71] B. Manavalan, R.G. Govindaraj, T.H. Shin, M.O. Kim, G. Lee, IBCE-EL: A new ensemble learning framework for improved linear B-cell Epitope prediction, *Front. Immunol.* 9 (2018) 1695. <https://doi.org/10.3389/fimmu.2018.01695>.
- [72] T. Liu, K. Shi, W. Li, Deep learning methods improve linear B-cell epitope prediction, *BioData Min.* 13 (2020) 1. <https://doi.org/10.1186/s13040-020-00211-0>.
- [73] A. Ras-Carmona, H.F. Pelaez-Prestel, E.M. Lafuente, P.A. Reche, BCEPS: A web server to predict linear B cell epitopes with enhanced immunogenicity and cross-reactivity, *Cells* 10 (2021) 2744. <https://doi.org/10.3390/cells10102744>.
- [74] W. Alghamdi, M. Attique, E. Alzahrani, M.Z. Ullah, Y.D. Khan, LBCEPred: a machine learning model to predict linear B-cell epitopes, *Brief. Bioinform.* 23 (2022). <https://doi.org/10.1093/bib/bbac035>.
- [75] M. Collatz, F. Mock, E. Barth, M. Hölzer, K. Sachse, M. Marz, EpiDope: a deep neural network for linear B-cell epitope prediction, *Bioinformatics* 37 (2021) 448–455. <https://doi.org/10.1093/bioinformatics/btaa773>.
- [76] J.N. Clifford, M.H. Høie, S. Deleuran, B. Peters, M. Nielsen, P. Marcatili, BepiPred-3.0: Improved B-cell epitope prediction using protein language models, *Protein Sci.* 31 (2022) e4497. <https://doi.org/10.1002/pro.4497>.
- [77] M. Levitt, Nature of the protein universe, *Proc. Natl. Acad. Sci. U. S. A.* 106 (2009) 11079–11084. <https://doi.org/10.1073/pnas.0905029106>.
- [78] N.D. Rubinstein, I. Mayrose, E. Martz, T. Pupko, Epitopia: a web-server for predicting B-cell epitopes, *BMC Bioinformatics* 10 (2009) 287. <https://doi.org/10.1186/1471-2105-10-287>.
- [79] H.R. Ansari, G.P. Raghava, Identification of conformational B-cell Epitopes in an antigen from its primary sequence, *Immunome Res.* 6 (2010) 6. <https://doi.org/10.1186/1745-7580-6-6>.
- [80] J. Zhang, X. Zhao, P. Sun, B. Gao, Z. Ma, Conformational B-cell epitopes prediction from sequences using cost-sensitive ensemble classifiers and spatial clustering, *Biomed Res. Int.* 2014 (2014) 689219. <https://doi.org/10.1155/2014/689219>.
- [81] M.C. Jespersen, B. Peters, M. Nielsen, P. Marcatili, BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes, *Nucleic Acids Res.* 45 (2017) W24–W29. <https://doi.org/10.1093/nar/gkx346>.

- [82] G.A. Dalkas, M. Rومان, SEPIa, a knowledge-driven algorithm for predicting conformational B-cell epitopes from the amino acid sequence, *BMC Bioinformatics* 18 (2017) 95. <https://doi.org/10.1186/s12859-017-1528-9>.
- [83] U. Kulkarni-Kale, S. Bhosle, A.S. Kolaskar, CEP: a conformational epitope prediction server, *Nucleic Acids Res.* 33 (2005) W168-71. <https://doi.org/10.1093/nar/gki460>.
- [84] P. Haste Andersen, M. Nielsen, O. Lund, Prediction of residues in discontinuous B-cell epitopes using protein 3D structures, *Protein Sci.* 15 (2006) 2558–2567. <https://doi.org/10.1110/ps.062405906>.
- [85] J. Ponomarenko, H.-H. Bui, W. Li, N. Fusseder, P.E. Bourne, A. Sette, B. Peters, ElliPro: a new structure-based tool for the prediction of antibody epitopes, *BMC Bioinformatics* 9 (2008) 514. <https://doi.org/10.1186/1471-2105-9-514>.
- [86] M.J. Sweredoski, P. Baldi, PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure, *Bioinformatics* 24 (2008) 1459–1460. <https://doi.org/10.1093/bioinformatics/btn199>.
- [87] J. Sun, D. Wu, T. Xu, X. Wang, X. Xu, L. Tao, Y.X. Li, Z.W. Cao, SEPPA: a computational server for spatial epitope prediction of protein antigens, *Nucleic Acids Res.* 37 (2009) W612-6. <https://doi.org/10.1093/nar/gkp417>.
- [88] S. Liang, D. Zheng, B. Yao, C. Zhang, EPCES and EPSVR: Prediction of B-cell antigenic epitopes on protein surfaces with conformational information, *Methods Mol. Biol.* 2131 (2020) 289–297. https://doi.org/10.1007/978-1-0716-0389-5_16.
- [89] S. Liang, D. Zheng, D.M. Standley, B. Yao, M. Zacharias, C. Zhang, EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results, *BMC Bioinformatics* 11 (2010) 381. <https://doi.org/10.1186/1471-2105-11-381>.
- [90] J.V. Kringelum, C. Lundegaard, O. Lund, M. Nielsen, Reliable B cell epitope predictions: impacts of method development and improved benchmarking, *PLoS Comput. Biol.* 8 (2012) e1002829. <https://doi.org/10.1371/journal.pcbi.1002829>.
- [91] T. Qi, T. Qiu, Q. Zhang, K. Tang, Y. Fan, J. Qiu, D. Wu, W. Zhang, Y. Chen, J. Gao, R. Zhu, Z. Cao, SEPPA 2.0--more refined server to predict spatial epitope considering species of immune host and subcellular localization of protein antigen, *Nucleic Acids Res.* 42 (2014) W59-63. <https://doi.org/10.1093/nar/gku395>.
- [92] C. Zhou, Z. Chen, L. Zhang, D. Yan, T. Mao, K. Tang, T. Qiu, Z. Cao, SEPPA 3.0-enhanced spatial epitope prediction enabling glycoprotein antigens, *Nucleic Acids Res.* 47 (2019) W388–W394. <https://doi.org/10.1093/nar/gkz413>.
- [93] B. Solihah, A. Azhari, A. Musdholifah, Enhancement of conformational B-cell epitope prediction using CluSMOTE, *PeerJ Comput. Sci.* 6 (2020) e275. <https://doi.org/10.7717/peerj-cs.275>.
- [94] T.I. Shashkova, D. Umerenkov, M. Salnikov, P.V. Strashnov, A.V. Konstantinova, I. Lebed, D.N. Shcherbinin, M.N. Asatryan, O.L. Kardymon, N.V. Ivanisenko, SEMA: Antigen B-cell conformational epitope prediction using deep transfer learning, *Front. Immunol.* 13 (2022) 960985. <https://doi.org/10.3389/fimmu.2022.960985>.
- [95] B.M. da Silva, Y. Myung, D.B. Ascher, D.E.V. Pires, epitope3D: a machine learning method for conformational B-cell epitope prediction, *Brief. Bioinform.* 23 (2022). <https://doi.org/10.1093/bib/bbab423>.
- [96] B.A. Suliman, Potential clinical implications of molecular mimicry-induced autoimmunity, *Immun. Inflamm. Dis.* 12 (2024) e1178. <https://doi.org/10.1002/iid3.1178>.

- [97] M. Peng, Z. Wang, Vaccine-associated autoimmunity: From clinical signals to immune pathways, *Vaccines (Basel)* 13 (2025) 1112. <https://doi.org/10.3390/vaccines13111112>.
- [98] G.D. Ogle, F. Wang, A. Haynes, G.A. Gregory, T.W. King, K. Deng, D. Dabelea, S. James, A.J. Jenkins, X. Li, R.C.W. Ma, D.M. Maahs, R.A. Oram, C. Pihoker, J. Svensson, Z. Zhou, D.J. Magliano, J. Maniam, Global type 1 diabetes prevalence, incidence, and mortality estimates 2025: Results from the International diabetes Federation Atlas, 11th Edition, and the T1D Index Version 3.0, *Diabetes Res. Clin. Pract.* 225 (2025) 112277. <https://doi.org/10.1016/j.diabres.2025.112277>.
- [99] K. Sakurai, D. Narita, N. Saito, T. Ueno, R. Sato, S. Niitsuma, K. Takahashi, Z. Arihara, Type 1 diabetes mellitus following COVID-19 RNA-based vaccine, *J. Diabetes Investig.* 13 (2022) 1290–1292. <https://doi.org/10.1111/jdi.13781>.
- [100] T. Sato, S. Kodama, K. Kaneko, J. Imai, H. Katagiri, Type 1 diabetes mellitus associated with nivolumab after second SARS-CoV-2 vaccination, Japan, *Emerg. Infect. Dis.* 28 (2022) 1518–1520. <https://doi.org/10.3201/eid2807.220127>.
- [101] T. Tanaka, S. Nagasu, T. Furuta, M. Gobaru, H. Suzuki, Y. Shimotsuura, J. Akiba, M. Nomura, F. Fujita, T. Kawaguchi, K. Miwa, Case report: A case of fulminant type 1 diabetes mellitus after COVID-19 vaccination during treatment of advanced gastric cancer: pitfall in managing immune-related adverse events, *Front. Oncol.* 13 (2023) 1264281. <https://doi.org/10.3389/fonc.2023.1264281>.
- [102] 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>.
- [103] R. Cai, Z. Liu, J. Ren, C. Ma, T. Gao, Y. Zhou, Q. Yang, Y. Xue, GPS-MBA: computational analysis of MHC class II epitopes in type 1 diabetes, *PLoS One* 7 (2012) e33884. <https://doi.org/10.1371/journal.pone.0033884>.
- [104] M. Alhadj Ali, Y.-F. Liu, S. Arif, D. Tatovic, H. Shariff, V.B. Gibson, N. Yusuf, R. Baptista, M. Eichmann, N. Petrov, S. Heck, J.H.M. Yang, T.I.M. Tree, I. Pujol-Autonell, L. Yeo, L. Baumard, R. Stenson, A. Howell, A. Clark, Z. Boulton, J. Powrie, L. Adams, F.S. Wong, S. Luzio, G. Dunseath, K. Green, A. O’Keefe, G. Bayly, N. Thorogood, R. Andrews, N. Leech, F. Joseph, S. Nair, S. Seal, H. Cheung, C. Beam, R. Hills, M. Peakman, C.M. Dayan, Metabolic and immune effects of immunotherapy with proinsulin peptide in human new-onset type 1 diabetes, *Sci. Transl. Med.* 9 (2017). <https://doi.org/10.1126/scitranslmed.aaf7779>.
- [105] E. Alska, A. Doligalska, K. Napiórkowska-Baran, M. Dolina, K. Osińska, A. Pilichowicz, A. Wojtkiewicz, J.J. Kaczor, B. Szymczak, Z. Bartuzi, Global burden of allergies: Mechanisms of development, challenges in diagnosis, and treatment, *Life (Basel)* 15 (2025) 878. <https://doi.org/10.3390/life15060878>.
- [106] S.M. Miescher, M. Vogel, Molecular aspects of allergy, *Mol. Aspects Med.* 23 (2002) 413–462. [https://doi.org/10.1016/s0098-2997\(02\)00009-2](https://doi.org/10.1016/s0098-2997(02)00009-2).
- [107] R.A. Wood, Allergic reactions to vaccines, *Pediatr. Allergy Immunol.* 24 (2013) 521–526. <https://doi.org/10.1111/pai.12102>.
- [108] V. Sampath, G. Rabinowitz, M. Shah, S. Jain, Z. Diamant, M. Jesenak, R. Rabin, S. Vieths, I. Agache, M. Akdis, D. Barber, H. Breiteneder, S. Chinthrajah, T. Chivato, W. Collins, T. Eiwegger, K. Fast, W. Fokkens, R.E. O’Hehir, M. Ollert, L. O’Mahony, O. Palomares, O. Pfaar, C. Riggioni, M.H. Shamji, M. Sokolowska, M. Jose Torres, C. Traidl-Hoffmann, M. van Zelm, D.Y. Wang, L. Zhang, C.A. Akdis, K.C. Nadeau, Vaccines and allergic reactions:

- The past, the current COVID-19 pandemic, and future perspectives, *Allergy* 76 (2021) 1640–1660. <https://doi.org/10.1111/all.14840>.
- [109] C.A. Stone Jr, C.R.F. Rukasin, T.M. Beachkofsky, E.J. Phillips, Immune-mediated adverse reactions to vaccines, *Br. J. Clin. Pharmacol.* 85 (2019) 2694–2706. <https://doi.org/10.1111/bcp.14112>.
- [110] M.-H. Tsai, C.-Y. Chiu, Allergic reactions to vaccines in children: From constituents to specific vaccines, *Biomedicines* 11 (2023) 620. <https://doi.org/10.3390/biomedicines11020620>.
- [111] F.E. Agamah, G.K. Mazandu, R. Hassan, C.D. Bope, N.E. Thomford, A. Ghansah, E.R. Chimusa, Computational/in silico methods in drug target and lead prediction, *Brief. Bioinform.* 21 (2020) 1663–1675. <https://doi.org/10.1093/bib/bbz103>.
- [112] V. Marzano, B. Tilocca, A.G. Fiocchi, P. Vernocchi, S. Levi Mortera, A. Urbani, P. Roncada, L. Putignani, Perusal of food allergens analysis by mass spectrometry-based proteomics, *J. Proteomics* 215 (2020) 103636. <https://doi.org/10.1016/j.jprot.2020.103636>.
- [113] Evaluation of allergenicity of genetically modified foods: report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology, (2001) 22–25.
- [114] M.B. Stadler, B.M. Stadler, Allergenicity prediction by protein sequence, *FASEB J.* 17 (2003) 1141–1143. <https://doi.org/10.1096/fj.02-1052fje>.
- [115] A. Silvanovich, M.A. Nemeth, P. Song, R. Herman, L. Tagliani, G.A. Bannon, The value of short amino acid sequence matches for prediction of protein allergenicity, *Toxicol. Sci.* 90 (2006) 252–258. <https://doi.org/10.1093/toxsci/kfj068>.
- [116] S. Saha, G.P.S. Raghava, AlgPred: prediction of allergenic proteins and mapping of IgE epitopes, *Nucleic Acids Res.* 34 (2006) W202–9. <https://doi.org/10.1093/nar/gkl343>.
- [117] Z.H. Zhang, J.L.Y. Koh, G.L. Zhang, K.H. Choo, M.T. Tammi, J.C. Tong, AllerTool: a web server for predicting allergenicity and allergic cross-reactivity in proteins, *Bioinformatics* 23 (2007) 504–506. <https://doi.org/10.1093/bioinformatics/btl621>.
- [118] H.C. Muh, J.C. Tong, M.T. Tammi, AllerHunter: a SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins, *PLoS One* 4 (2009) e5861. <https://doi.org/10.1371/journal.pone.0005861>.
- [119] J. Wang, D. Zhang, J. Li, PREAL: prediction of allergenic protein by maximum Relevance Minimum Redundancy (mRMR) feature selection, *BMC Syst. Biol.* 7 Suppl 5 (2013) S9. <https://doi.org/10.1186/1752-0509-7-S5-S9>.
- [120] I. Dimitrov, D.R. Flower, I. Doytchinova, AllerTOP--a server for in silico prediction of allergens, *BMC Bioinformatics* 14 Suppl 6 (2013) S4. <https://doi.org/10.1186/1471-2105-14-S6-S4>.
- [121] H.X. Dang, C.B. Lawrence, Allerdicator: fast allergen prediction using text classification techniques, *Bioinformatics* 30 (2014) 1120–1128. <https://doi.org/10.1093/bioinformatics/btu004>.
- [122] I. Dimitrov, L. Naneva, I. Doytchinova, I. Bangov, AllergenFP: allergenicity prediction by descriptor fingerprints, *Bioinformatics* 30 (2014) 846–851. <https://doi.org/10.1093/bioinformatics/btt619>.
- [123] I. Dimitrov, I. Bangov, D.R. Flower, I. Doytchinova, AllerTOP v.2--a server for in silico prediction of allergens, *J. Mol. Model.* 20 (2014) 2278. <https://doi.org/10.1007/s00894-014-2278-5>.
- [124] S. Maurer-Stroh, N.L. Krutz, P.S. Kern, V. Gunalan, M.N. Nguyen, V. Limviphuvadh, F. Eisenhaber, G.F. Gerberick, AllerCatPro-prediction of protein allergenicity potential from the

- p
-
- protein sequence, Bioinformatics 35 (2019) 3020–3027.
-
- <https://doi.org/10.1093/bioinformatics/btz029>
- .
-
- [125] N. Sharma, S. Patiyal, A. Dhall, A. Pande, C. Arora, G.P.S. Raghava, AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes, Brief. Bioinform. 22 (2021).
- <https://doi.org/10.1093/bib/bbaa294>
- .
-
- [126] M.N. Nguyen, N.L. Krutz, V. Limviphuvadh, A.L. Lopata, G.F. Gerberick, S. Maurer-Stroh, AllerCatPro 2.0: a web server for predicting protein allergenicity potential, Nucleic Acids Res. 50 (2022) W36–W43.
- <https://doi.org/10.1093/nar/gkac446>
- .
-
- [127] P.M. Shanthappa, R. Kumar, ProAll-D: protein allergen detection using long short term memory - a deep learning approach, ADMET DMPK 10 (2022) 231–240.
- <https://doi.org/10.5599/admet.1335>
- .
-
- [128] F.M. Garcia-Moreno, M.A. Gutiérrez-Naranjo, ALLERDET: A novel web app for prediction of protein allergenicity, J. Biomed. Inform. 135 (2022) 104217.
- <https://doi.org/10.1016/j.jbi.2022.104217>
- .
-
- [129] C. He, X. Ye, Y. Yang, L. Hu, Y. Si, X. Zhao, L. Chen, Q. Fang, Y. Wei, F. Wu, G. Ye, DeepAlgPro: an interpretable deep neural network model for predicting allergenic proteins, Brief. Bioinform. 24 (2023).
- <https://doi.org/10.1093/bib/bbad246>
- .
-
- [130] B. Liu, Z. Yang, Q. Liu, Y. Zhang, H. Ding, H. Lai, Q. Li, Computational prediction of allergenic proteins based on multi-feature fusion, Front. Genet. 14 (2023) 1294159.
- <https://doi.org/10.3389/fgene.2023.1294159>
- .
-
- [131] S. Basith, N.T. Pham, B. Manavalan, G. Lee, SEP-AlgPro: An efficient allergen prediction tool utilizing traditional machine learning and deep learning techniques with protein language model features, Int. J. Biol. Macromol. 273 (2024) 133085.
- <https://doi.org/10.1016/j.ijbiomac.2024.133085>
- .
-
- [132] C. Yang, S.S. Negi, C.H. Schein, W. Braun, P. Kim, AllergenAI: a deep learning model predicting allergenicity based on protein sequence, bioRxivorg (2024).
- <https://doi.org/10.1101/2024.06.22.600179>
- .
-
- [133] L. Zhang, T. Liu, PreAlgPro: Prediction of allergenic proteins with pre-trained protein language model and efficient neutral network, Int. J. Biol. Macromol. 280 (2024) 135762.
- <https://doi.org/10.1016/j.ijbiomac.2024.135762>
- .
-
- [134] E. Kemmler, E.K. Fath, R. Preissner, P. Banerjee, AllergyPred: a web server for allergen prediction, Nucleic Acids Res. 53 (2025) W4–W10.
- <https://doi.org/10.1093/nar/gkaf383>
- .
-
- [135] X. Chen, G.J.P. Fernando, M.L. Crichton, C. Flaim, S.R. Yukiko, E.J. Fairmaid, H.J. Corbett, C.A. Primiero, A.B. Ansaldo, I.H. Frazer, L.E. Brown, M.A.F. Kendall, Improving the reach of vaccines to low-resource regions, with a needle-free vaccine delivery device and long-term thermostabilization, J. Control. Release 152 (2011) 349–355.
- <https://doi.org/10.1016/j.jconrel.2011.02.026>
- .
-
- [136] P. Das, Revolutionary vaccine technology breaks the cold chain, Lancet Infect. Dis. 4 (2004) 719.
- [https://doi.org/10.1016/s1473-3099\(04\)01222-8](https://doi.org/10.1016/s1473-3099(04)01222-8)
- .
-
- [137] T. Sun, H. Han, G.A. Hudalla, Y. Wen, R.R. Pompano, J.H. Collier, Thermal stability of self-assembled peptide vaccine materials, Acta Biomater. 30 (2016) 62–71.
- <https://doi.org/10.1016/j.actbio.2015.11.019>
- .
-
- [138] M. Akbarian, S.-H. Chen, Instability challenges and stabilization strategies of pharmaceutical proteins, Pharmaceutics 14 (2022) 2533.
- <https://doi.org/10.3390/pharmaceutics14112533>
- .

- [139] R.J.B.M. Delahaije, P.A. Wierenga, M.L.F. Giuseppin, H. Gruppen, Comparison of heat-induced aggregation of globular proteins, *J. Agric. Food Chem.* 63 (2015) 5257–5265. <https://doi.org/10.1021/acs.jafc.5b00927>.
- [140] M. Li, H. Wang, Z. Yang, L. Zhang, Y. Zhu, DeepTM: A deep learning algorithm for prediction of melting temperature of thermophilic proteins directly from sequences, *Comput. Struct. Biotechnol. J.* 21 (2023) 5544–5560. <https://doi.org/10.1016/j.csbj.2023.11.006>.
- [141] M. Gorania, H. Seker, P.I. Haris, Predicting a protein's melting temperature from its amino acid sequence, *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2010 (2010) 1820–1823. <https://doi.org/10.1109/IEMBS.2010.5626421>.
- [142] H. Jiang, G.-B. Zhang, Y.-X. Wang, F.-Y. Jiang, H.-Y. Zhang, Z.-W. Nie, L. Yuan, J. Chen, W.-B. Zhang, TmPred: Enhancing thermophilic protein melting point prediction with protein language models and deep learning, *Chin. J. Polym. Sci.* 43 (2025) 2191–2200. <https://doi.org/10.1007/s10118-025-3447-y>.
- [143] P. Tijare, N. Kumar, G.P.S. Raghava, Prediction and design of thermostable proteins with a desired melting temperature, *Sci. Rep.* 15 (2025) 16683. <https://doi.org/10.1038/s41598-025-98667-9>.
- [144] Y. Yang, X. Ding, G. Zhu, A. Niroula, Q. Lv, M. Vihinen, ProTstab - predictor for cellular protein stability, *BMC Genomics* 20 (2019) 804. <https://doi.org/10.1186/s12864-019-6138-7>.
- [145] C.-W. Chen, M.-H. Lin, C.-C. Liao, H.-P. Chang, Y.-W. Chu, iStable 2.0: Predicting protein thermal stability changes by integrating various characteristic modules, *Comput. Struct. Biotechnol. J.* 18 (2020) 622–630. <https://doi.org/10.1016/j.csbj.2020.02.021>.
- [146] Y. Yang, J. Zhao, L. Zeng, M. Vihinen, ProTstab2 for prediction of protein thermal stabilities, *Int. J. Mol. Sci.* 23 (2022) 10798. <https://doi.org/10.3390/ijms231810798>.
- [147] F. Jung, K. Frey, D. Zimmer, T. Mühlhaus, DeepSTABp: A deep learning approach for the prediction of thermal protein stability, *Int. J. Mol. Sci.* 24 (2023) 7444. <https://doi.org/10.3390/ijms24087444>.
- [148] P. Charoenkwan, W. Chotpatiwettkul, V.S. Lee, C. Nantasenamat, W. Shoombuatong, A novel sequence-based predictor for identifying and characterizing thermophilic proteins using estimated propensity scores of dipeptides, *Sci. Rep.* 11 (2021) 23782. <https://doi.org/10.1038/s41598-021-03293-w>.
- [149] C. Meng, Y. Ju, H. Shi, TMPpred: A support vector machine-based thermophilic protein identifier, *Anal. Biochem.* 645 (2022) 114625. <https://doi.org/10.1016/j.ab.2022.114625>.
- [150] P. Charoenkwan, N. Schaduangrat, M.A. Moni, P. Lio', B. Manavalan, W. Shoombuatong, SAPPHERE: A stacking-based ensemble learning framework for accurate prediction of thermophilic proteins, *Comput. Biol. Med.* 146 (2022) 105704. <https://doi.org/10.1016/j.combiomed.2022.105704>.
- [151] J. Zhao, W. Yan, Y. Yang, DeepTP: A deep learning model for thermophilic protein prediction, *Int. J. Mol. Sci.* 24 (2023) 2217. <https://doi.org/10.3390/ijms24032217>.
- [152] H. Pei, J. Li, S. Ma, J. Jiang, M. Li, Q. Zou, Z. Lv, Identification of thermophilic proteins based on sequence-based bidirectional representations from transformer-embedding features, *Appl. Sci. (Basel)* 13 (2023) 2858. <https://doi.org/10.3390/app13052858>.
- [153] F. Haselbeck, M. John, Y. Zhang, J. Pirnay, J.P. Fuenzalida-Werner, R.D. Costa, D.G. Grimm, Superior protein thermophilicity prediction with protein language model embeddings, *NAR Genom. Bioinform.* 5 (2023) lqad087. <https://doi.org/10.1093/nargab/lqad087>.

- [154] I. Pudžiuvėlytė, K. Olechnovič, E. Godliauskaite, K. Sermokas, T. Urbaitis, G. Gasiunas, D. Kazlauskas, TemStaPro: protein thermostability prediction using sequence representations from protein language models, *Bioinformatics* 40 (2024). <https://doi.org/10.1093/bioinformatics/btae157>.
- [155] F.G. Hopkins, Denaturation of proteins by urea and related substances, *Nature* 126 (1930) 328–330. <https://doi.org/10.1038/126328a0>.
- [156] P.F. Scholander, L. van Dam, J.W. Kanwisher, H.T. Hammel, M.S. Gordon, Supercooling and osmoregulation in arctic fish, *J. Cell. Comp. Physiol.* 49 (1957) 5–24. <https://doi.org/10.1002/jcp.1030490103>.
- [157] A.L. DeVries, D.E. Wohlschlag, Freezing resistance in some Antarctic fishes, *Science* 163 (1969) 1073–1075. <https://doi.org/10.1126/science.163.3871.1073>.
- [158] N. Kumar, S. Choudhury, N. Bajiya, S. Patiyal, G.P.S. Raghava, Prediction of anti-freezing proteins from their evolutionary profile, *Proteomics* 25 (2025) e202400157. <https://doi.org/10.1002/pmic.202400157>.
- [159] A. Eskandari, T.C. Leow, M.B.A. Rahman, S.N. Oslan, Antifreeze proteins and their practical utilization in industry, medicine, and agriculture, *Biomolecules* 10 (2020) 1649. <https://doi.org/10.3390/biom10121649>.
- [160] X. Zhan, D.-W. Sun, Z. Zhu, Q.-J. Wang, Improving the quality and safety of frozen muscle foods by emerging freezing technologies: A review, *Crit. Rev. Food Sci. Nutr.* 58 (2018) 2925–2938. <https://doi.org/10.1080/10408398.2017.1345854>.
- [161] S.G. Lee, H.Y. Koh, J.H. Lee, S.-H. Kang, H.J. Kim, Cryopreservative effects of the recombinant ice-binding protein from the arctic yeast *Leucosporidium* sp. on red blood cells, *Appl. Biochem. Biotechnol.* 167 (2012) 824–834. <https://doi.org/10.1007/s12010-012-9739-z>.
- [162] Y.-H. Zhang, Z. Li, L. Lu, T. Zeng, L. Chen, H. Li, T. Huang, Y.-D. Cai, Analysis of the sequence characteristics of antifreeze protein, *Life (Basel)* 11 (2021) 520. <https://doi.org/10.3390/life11060520>.
- [163] J. Cheng, Y. Hanada, A. Miura, S. Tsuda, H. Kondo, Hydrophobic ice-binding sites confer hyperactivity of an antifreeze protein from a snow mold fungus, *Biochem. J.* 473 (2016) 4011–4026. <https://doi.org/10.1042/BCJ20160543>.
- [164] K. Basu, C.P. Garnham, Y. Nishimiya, S. Tsuda, I. Braslavsky, P. Davies, Determining the ice-binding planes of antifreeze proteins by fluorescence-based ice plane affinity, *J. Vis. Exp.* (2014) e51185. <https://doi.org/10.3791/51185>.
- [165] M.J. Kuiper, C. Lankin, S.Y. Gauthier, V.K. Walker, P.L. Davies, Purification of antifreeze proteins by adsorption to ice, *Biochem. Biophys. Res. Commun.* 300 (2003) 645–648. [https://doi.org/10.1016/s0006-291x\(02\)02900-5](https://doi.org/10.1016/s0006-291x(02)02900-5).
- [166] M.C. Loewen, W. Gronwald, F.D. Sönnichsen, B.D. Sykes, P.L. Davies, The ice-binding site of sea raven antifreeze protein is distinct from the carbohydrate-binding site of the homologous C-type lectin, *Biochemistry* 37 (1998) 17745–17753. <https://doi.org/10.1021/bi9820513>.
- [167] K.K. Kandaswamy, K.-C. Chou, T. Martinetz, S. Möller, P.N. Suganthan, S. Sridharan, G. Pugalenth, AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties, *J. Theor. Biol.* 270 (2011) 56–62. <https://doi.org/10.1016/j.jtbi.2010.10.037>.

- [168] C.-S. Yu, C.-H. Lu, Identification of antifreeze proteins and their functional residues by support vector machine and genetic algorithms based on n-peptide compositions, *PLoS One* 6 (2011) e20445. <https://doi.org/10.1371/journal.pone.0020445>.
- [169] X. Zhao, Z. Ma, M. Yin, Using support vector machine and evolutionary profiles to predict antifreeze protein sequences, *Int. J. Mol. Sci.* 13 (2012) 2196–2207. <https://doi.org/10.3390/ijms13022196>.
- [170] S. Mondal, P.P. Pai, Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction, *J. Theor. Biol.* 356 (2014) 30–35. <https://doi.org/10.1016/j.jtbi.2014.04.006>.
- [171] X. He, K. Han, J. Hu, H. Yan, J.-Y. Yang, H.-B. Shen, D.-J. Yu, TargetFreeze: Identifying antifreeze proteins via a combination of weights using sequence evolutionary information and pseudo amino acid composition, *J. Membr. Biol.* 248 (2015) 1005–1014. <https://doi.org/10.1007/s00232-015-9811-z>.
- [172] R. Yang, C. Zhang, R. Gao, L. Zhang, An effective antifreeze protein predictor with ensemble classifiers and comprehensive sequence descriptors, *Int. J. Mol. Sci.* 16 (2015) 21191–21214. <https://doi.org/10.3390/ijms160921191>.
- [173] R. Pratiwi, A.A. Malik, N. Schaduangrat, V. Prachayasittikul, J.E.S. Wikberg, C. Nantasenamat, W. Shoombuatong, CryoProtect: A web server for classifying antifreeze proteins from nonantifreeze proteins, *J. Chem.* 2017 (2017) 1–15. <https://doi.org/10.1155/2017/9861752>.
- [174] X. Xiao, M. Hui, Z. Liu, IAFP-Ense: An ensemble classifier for identifying antifreeze protein by incorporating grey model and PSSM into PseAAC, *J. Membr. Biol.* 249 (2016) 845–854. <https://doi.org/10.1007/s00232-016-9935-9>.
- [175] S. Khan, I. Naseem, R. Togneri, M. Bennamoun, RAFP-Pred: Robust prediction of Antifreeze Proteins using localized analysis of n-peptide compositions, *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15 (2018) 244–250. <https://doi.org/10.1109/TCBB.2016.2617337>.
- [176] M. Eslami, R. Shirali Hossein Zade, Z. Takaloo, G. Mahdevar, A. Emamjomeh, R.H. Sajedi, J. Zahiri, afpCOOL: A tool for antifreeze protein prediction, *Heliyon* 4 (2018) e00705. <https://doi.org/10.1016/j.heliyon.2018.e00705>.
- [177] M. Usman, J.A. Lee, AFP-CKSAAP: Prediction of antifreeze proteins using composition of k-spaced amino acid pairs with deep neural network, in: 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), IEEE, 2019. <https://doi.org/10.1109/bibe.2019.00016>.
- [178] S. Wang, L. Deng, X. Xia, Z. Cao, Y. Fei, Predicting antifreeze proteins with weighted generalized dipeptide composition and multi-regression feature selection ensemble, *BMC Bioinformatics* 22 (2021) 340. <https://doi.org/10.1186/s12859-021-04251-z>.
- [179] M. Usman, S. Khan, J.-A. Lee, AFP-LSE: Antifreeze proteins prediction using latent space encoding of composition of k-spaced amino acid pairs, *Sci. Rep.* 10 (2020) 7197. <https://doi.org/10.1038/s41598-020-63259-2>.
- [180] S. Sun, H. Ding, D. Wang, S. Han, Identifying antifreeze proteins based on key evolutionary information, *Front. Bioeng. Biotechnol.* 8 (2020) 244. <https://doi.org/10.3389/fbioe.2020.00244>.
- [181] F. Ali, S. Akbar, A. Ghulam, Z.A. Maher, A. Unar, D.B. Talpur, AFP-CMBPred: Computational identification of antifreeze proteins by extending consensus sequences into multi-blocks evolutionary information, *Comput. Biol. Med.* 139 (2021) 105006. <https://doi.org/10.1016/j.combiomed.2021.105006>.

- [182] A. Khan, J. Uddin, F. Ali, A. Ahmad, O. Alghushairy, A. Banjar, A. Daud, Prediction of antifreeze proteins using machine learning, *Sci. Rep.* 12 (2022) 20672. <https://doi.org/10.1038/s41598-022-24501-1>.
- [183] A. Khan, J. Uddin, F. Ali, H. Kumar, W. Alghamdi, A. Ahmad, AFP-SPTS: An accurate prediction of antifreeze proteins using sequential and pseudo-Tri-slicing evolutionary features with an extremely randomized tree, *J. Chem. Inf. Model.* 63 (2023) 826–834. <https://doi.org/10.1021/acs.jcim.2c01417>.
- [184] M. Usman, S. Khan, S. Park, A. Wahab, AFP-SRC: identification of antifreeze proteins using sparse representation classifier, *Neural Comput. Appl.* 34 (2022) 2275–2285. <https://doi.org/10.1007/s00521-021-06558-7>.
- [185] S. Kontos, J.A. Hubbell, Drug development: longer-lived proteins, *Chem. Soc. Rev.* 41 (2012) 2686–2695. <https://doi.org/10.1039/c2cs15289d>.
- [186] J. Zalevsky, A.K. Chamberlain, H.M. Horton, S. Karki, I.W.L. Leung, T.J. Sproule, G.A. Lazar, D.C. Roopenian, J.R. Desjarlais, Enhanced antibody half-life improves in vivo activity, *Nat. Biotechnol.* 28 (2010) 157–159. <https://doi.org/10.1038/nbt.1601>.
- [187] J. Abraham, Monoclonal antibodies with extended half-life to prevent covid-19, *N. Engl. J. Med.* 386 (2022) 2236–2238. <https://doi.org/10.1056/NEJMe2205563>.
- [188] Z. Shen, C. Li, W. Song, L. Liu, Y. Kong, A. Huang, Q. Bao, T. Ying, Y. Wu, Enhancing vaccine half-life as a novel strategy for improving immune response durability of subunit vaccines, *PLoS Pathog.* 21 (2025) e1012845. <https://doi.org/10.1371/journal.ppat.1012845>.
- [189] U. Alam, K. Chaudhary, N. Kumar, R. Tomer, S. Patiyal, G.P.S. Raghava, PEPLife2: A updated repository of the half-life of peptides, *bioRxiv* (2025). <https://doi.org/10.1101/2025.05.13.653654>.
- [190] B. Sun, F.S. Willard, D. Feng, J. Alsina-Fernandez, Q. Chen, M. Vieth, J.D. Ho, A.D. Showalter, C. Stutsman, L. Ding, T.M. Suter, J.D. Dunbar, J.W. Carpenter, F.A. Mohammed, E. Aihara, R.A. Brown, A.B. Bueno, P.J. Emmerson, J.S. Moyers, T.S. Kobilka, M.P. Coghlan, B.K. Kobilka, K.W. Sloop, Structural determinants of dual incretin receptor agonism by tirzepatide, *Proc. Natl. Acad. Sci. U. S. A.* 119 (2022) e2116506119. <https://doi.org/10.1073/pnas.2116506119>.
- [191] H. Hu, A.S. Ransdell, H. Qu, J.D. Durbin, F.A. Valenzuela, S. Hernandez-Buquer, M.D. Gonciarz, Probing the binding mechanism of acylated peptides to human serum albumin, *ACS Chem. Biol.* 18 (2023) 1158–1167. <https://doi.org/10.1021/acschembio.3c00018>.
- [192] Q. Tan, S.E. Akindehin, C.E. Orsso, R.C. Waldner, R.D. DiMarchi, T.D. Müller, A.M. Haqq, Recent advances in incretin-based pharmacotherapies for the treatment of obesity and diabetes, *Front. Endocrinol. (Lausanne)* 13 (2022) 838410. <https://doi.org/10.3389/fendo.2022.838410>.
- [193] R. Kontermann, *Therapeutic proteins*, Wiley-VCH Verlag, Weinheim, Germany, 2012. <https://doi.org/10.1002/9783527644827>.
- [194] S.C. Penchala, M.R. Miller, A. Pal, J. Dong, N.R. Madadi, J. Xie, H. Joo, J. Tsai, P. Batoon, V. Samoshin, A. Franz, T. Cox, J. Miles, W.K. Chan, M.S. Park, M.M. Alhamadsheh, A biomimetic approach for enhancing the in vivo half-life of peptides, *Nat. Chem. Biol.* 11 (2015) 793–798. <https://doi.org/10.1038/nchembio.1907>.
- [195] Y.-J. Li, L. Zhang, P.-P. Yang, K. Zhang, X.-F. Gong, D.-Y. Hou, H. Cao, X.-C. Wu, R. Liu, K.S. Lam, L. Wang, Bioinspired screening of anti-adhesion peptides against blood proteins for intravenous delivery of nanomaterials, *Nano Lett.* 22 (2022) 8076–8085. <https://doi.org/10.1021/acs.nanolett.2c02243>.

- [196] D. Mathur, S. Singh, A. Mehta, P. Agrawal, G.P.S. Raghava, In silico approaches for predicting the half-life of natural and modified peptides in blood, *PLoS One* 13 (2018) e0196829. <https://doi.org/10.1371/journal.pone.0196829>.
- [197] M.R. Wilkins, E. Gasteiger, A. Bairoch, J.C. Sanchez, K.L. Williams, R.D. Appel, D.F. Hochstrasser, Protein identification and analysis tools in the ExPASy server, *Methods Mol. Biol.* 112 (1999) 531–552. <https://www.ncbi.nlm.nih.gov/pubmed/10027275>.
- [198] E. Lazaro, C. Kadie, P. Stamegna, S.C. Zhang, P. Gourdain, N.Y. Lai, M. Zhang, S.A. Martinez, D. Heckerman, S. Le Gall, Variable HIV peptide stability in human cytosol is critical to epitope presentation and immune escape, *J. Clin. Invest.* 121 (2011) 2480–2492. <https://doi.org/10.1172/JCI44932>.
- [199] X. Song, T. Zhou, H. Jia, X. Guo, X. Zhang, P. Han, J. Sha, SProtP: a web server to recognize those short-lived proteins based on sequence-derived features in human cells, *PLoS One* 6 (2011) e27836. <https://doi.org/10.1371/journal.pone.0027836>.
- [200] A. Sharma, D. Singla, M. Rashid, G.P.S. Raghava, Designing of peptides with desired half-life in intestine-like environment, *BMC Bioinformatics* 15 (2014) 282. <https://doi.org/10.1186/1471-2105-15-282>.
- [201] X. Tie, J. Li, H. Wu, R. Weng, Y. Zhang, J. Li, K. Chen, Advances in the structure-based design of protein vaccines for respiratory infectious diseases, *Comput. Struct. Biotechnol. J.* 29 (2025) 259–270. <https://doi.org/10.1016/j.csbj.2025.10.008>.
- [202] R. Vita, N. Blazeska, D. Marrama, IEDB Curation Team Members, S. Duesing, J. Bennett, J. Greenbaum, M. De Almeida Mendes, J. Mahita, D.K. Wheeler, J.R. Cantrell, J.A. Overton, D.A. Natale, A. Sette, B. Peters, The Immune Epitope Database (IEDB): 2024 update, *Nucleic Acids Res.* 53 (2025) D436–D443. <https://doi.org/10.1093/nar/gkae1092>.
- [203] S.K. Burley, H.M. Berman, G.J. Kleywegt, J.L. Markley, H. Nakamura, S. Velankar, Protein Data Bank (PDB): The single global macromolecular structure archive, *Methods Mol. Biol.* 1607 (2017) 627–641. https://doi.org/10.1007/978-1-4939-7000-1_26.
- [204] E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, A. Bairoch, UniProtKB/Swiss-Prot, *Methods Mol. Biol.* 406 (2007) 89–112. https://doi.org/10.1007/978-1-59745-535-0_4.
- [205] A. Jarzab, N. Kurzawa, T. Hopf, M. Moerch, J. Zecha, N. Leijten, Y. Bian, E. Musiol, M. Maschberger, G. Stoeck, I. Becher, C. Daly, P. Samaras, J. Mergner, B. Spanier, A. Angelov, T. Werner, M. Bantscheff, M. Wilhelm, M. Klingenspor, S. Lemeer, W. Liebl, H. Hahne, M.M. Savitski, B. Kuster, Meltome atlas-thermal proteome stability across the tree of life, *Nat. Methods* 17 (2020) 495–503. <https://doi.org/10.1038/s41592-020-0801-4>.
- [206] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, *Bioinformatics* 28 (2012) 3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>.
- [207] Q. Li, N. Fu, S.S. Omee, J. Hu, MD-HIT: Machine learning for material property prediction with dataset redundancy control, *Npj Comput. Mater.* 10 (2024). <https://doi.org/10.1038/s41524-024-01426-z>.
- [208] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics* 22 (2006) 1658–1659. <https://doi.org/10.1093/bioinformatics/btl158>.
- [209] Q. Chen, Y. Wan, Y. Lei, J. Zobel, K. Verspoor, Evaluation of CD-HIT for constructing non-redundant databases, in: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2016. <https://doi.org/10.1109/bibm.2016.7822604>.

- [210] A. Pande, S. Patiyal, A. Lathwal, C. Arora, D. Kaur, A. Dhall, G. Mishra, H. Kaur, N. Sharma, S. Jain, S.S. Usmani, P. Agrawal, R. Kumar, V. Kumar, G.P.S. Raghava, Pfeature: A tool for computing wide range of protein features and building prediction models, *J. Comput. Biol.* 30 (2023) 204–222. <https://doi.org/10.1089/cmb.2022.0241>.
- [211] V. Vacic, L.M. Iakoucheva, P. Radivojac, Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments, *Bioinformatics* 22 (2006) 1536–1537. <https://doi.org/10.1093/bioinformatics/btl151>.
- [212] G.E. Crooks, G. Hon, J.-M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, *Genome Res.* 14 (2004) 1188–1190. <https://doi.org/10.1101/gr.849004>.
- [213] A. Dhall, S. Patiyal, N. Sharma, S.S. Usmani, G.P.S. Raghava, Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19, *Brief. Bioinform.* 22 (2021) 936–945. <https://doi.org/10.1093/bib/bbaa259>.
- [214] A. Dhall, S. Patiyal, G.P.S. Raghava, HLAnCPred: a method for predicting promiscuous non-classical HLA binding sites, *Brief. Bioinform.* 23 (2022). <https://doi.org/10.1093/bib/bbac192>.
- [215] M. AlQuraishi, Machine learning in protein structure prediction, *Curr. Opin. Chem. Biol.* 65 (2021) 1–8. <https://doi.org/10.1016/j.cbpa.2021.04.005>.
- [216] R. Bonetta, G. Valentino, Machine learning techniques for protein function prediction, *Proteins* 88 (2020) 397–413. <https://doi.org/10.1002/prot.25832>.
- [217] K. Nagata, A. Randall, P. Baldi, SIDEpro: a novel machine learning approach for the fast and accurate prediction of side-chain conformations, *Proteins* 80 (2012) 142–153. <https://doi.org/10.1002/prot.23170>.
- [218] S. Mazurenko, Z. Prokop, J. Damborsky, Machine learning in enzyme engineering, *ACS Catal.* 10 (2020) 1210–1223. <https://doi.org/10.1021/acscatal.9b04321>.
- [219] Z. Chen, P. Zhao, F. Li, A. Leier, T.T. Marquez-Lago, Y. Wang, G.I. Webb, A.I. Smith, R.J. Daly, K.-C. Chou, J. Song, iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences, *Bioinformatics* 34 (2018) 2499–2502. <https://doi.org/10.1093/bioinformatics/bty140>.
- [220] Z. Chen, X. Liu, P. Zhao, C. Li, Y. Wang, F. Li, T. Akutsu, C. Bain, R.B. Gasser, J. Li, Z. Yang, X. Gao, L. Kurgan, J. Song, iFeatureOmega: an integrative platform for engineering, visualization and analysis of features from molecular sequences, structural and ligand data sets, *Nucleic Acids Res.* 50 (2022) W434–W447. <https://doi.org/10.1093/nar/gkac351>.
- [221] Z. Chen, P. Zhao, C. Li, F. Li, D. Xiang, Y.-Z. Chen, T. Akutsu, R.J. Daly, G.I. Webb, Q. Zhao, L. Kurgan, J. Song, iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization, *Nucleic Acids Res.* 49 (2021) e60. <https://doi.org/10.1093/nar/gkab122>.
- [222] N. Xiao, D.-S. Cao, M.-F. Zhu, Q.-S. Xu, Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences, *Bioinformatics* 31 (2015) 1857–1859. <https://doi.org/10.1093/bioinformatics/btv042>.
- [223] D.-S. Cao, N. Xiao, Q.-S. Xu, A.F. Chen, Rcp: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions, *Bioinformatics* 31 (2015) 279–281. <https://doi.org/10.1093/bioinformatics/btu624>.
- [224] R.P. Bonidia, D.S. Domingues, D.S. Sanches, A.C.P.L.F. de Carvalho, MathFeature: feature extraction package for DNA, RNA and protein sequences based on mathematical descriptors, *Brief. Bioinform.* 23 (2022). <https://doi.org/10.1093/bib/bbab434>.

- [225] D. Guevara-Barrientos, R. Kaundal, ProFeatX: A parallelized protein feature extraction suite for machine learning, *Comput. Struct. Biotechnol. J.* 21 (2023) 796–801. <https://doi.org/10.1016/j.csbj.2022.12.044>.
- [226] H.-C. Yi, Z.-H. You, X. Zhou, L. Cheng, X. Li, T.-H. Jiang, Z.-H. Chen, ACP-DL: A deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation, *Mol. Ther. Nucleic Acids* 17 (2019) 1–9. <https://doi.org/10.1016/j.omtn.2019.04.025>.
- [227] X. Xiao, S. Shao, Y. Ding, Z. Huang, K.-C. Chou, Using cellular automata images and pseudo amino acid composition to predict protein subcellular location, *Amino Acids* 30 (2006) 49–54. <https://doi.org/10.1007/s00726-005-0225-6>.
- [228] S.M. Azim, A. Sharma, I. Noshadi, S. Shatabda, I. Dehzangi, A convolutional neural network based tool for predicting protein AMPylation sites from binary profile representation, *Sci. Rep.* 12 (2022) 11451. <https://doi.org/10.1038/s41598-022-15403-3>.
- [229] X. Xiao, P. Wang, K.-C. Chou, GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes, *J. Comput. Chem.* 30 (2009) 1414–1423. <https://doi.org/10.1002/jcc.21163>.
- [230] A.A. Schäffer, L. Aravind, T.L. Madden, S. Shavirin, J.L. Spouge, Y.I. Wolf, E.V. Koonin, S.F. Altschul, Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, *Nucleic Acids Res.* 29 (2001) 2994–3005. <https://doi.org/10.1093/nar/29.14.2994>.
- [231] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* 25 (1997) 3389–3402. <https://doi.org/10.1093/nar/25.17.3389>.
- [232] R. Yang, C. Zhang, R. Gao, L. Zhang, A novel feature extraction method with feature selection to identify Golgi-resident protein types from imbalanced data, *Int. J. Mol. Sci.* 17 (2016) 218. <https://doi.org/10.3390/ijms17020218>.
- [233] L. Zou, C. Nan, F. Hu, Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles, *Bioinformatics* 29 (2013) 3135–3142. <https://doi.org/10.1093/bioinformatics/btt554>.
- [234] R. Kaundal, G.P.S. Raghava, RSLpred: an integrative system for predicting subcellular localization of rice proteins combining compositional and evolutionary information, *Proteomics* 9 (2009) 2324–2342. <https://doi.org/10.1002/pmic.200700597>.
- [235] X. Zhang, S. Liu, RBPPred: predicting RNA-binding proteins from sequence using SVM, *Bioinformatics* 33 (2017) 854–862. <https://doi.org/10.1093/bioinformatics/btw730>.
- [236] R. Verma, G.C. Varshney, G.P.S. Raghava, Prediction of mitochondrial proteins of malaria parasite using split amino acid composition and PSSM profile, *Amino Acids* 39 (2010) 101–110. <https://doi.org/10.1007/s00726-009-0381-1>.
- [237] M. Kumar, M.M. Gromiha, G.P.S. Raghava, SVM based prediction of RNA-binding proteins using binding residues and evolutionary information, *J. Mol. Recognit.* 24 (2011) 303–313. <https://doi.org/10.1002/jmr.1061>.
- [238] J. Wang, B. Yang, J. Revote, A. Leier, T.T. Marquez-Lago, G. Webb, J. Song, K.-C. Chou, T. Lithgow, POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles, *Bioinformatics* 33 (2017) 2756–2758. <https://doi.org/10.1093/bioinformatics/btx302>.
- [239] K. Kaminski, J. Ludwiczak, K. Pawlicki, V. Alva, S. Dunin-Horkawicz, pLM-BLAST: distant homology detection based on direct comparison of sequence representations from

- p
-
- protein language models, Bioinformatics 39 (2023).
-
- https://doi.org/10.1093/bioinformatics/btad579.
-
- [240] R. Singh, S. Sledzieski, B. Bryson, L. Cowen, B. Berger, Contrastive learning in protein language space predicts interactions between drugs and protein targets,
- Proc. Natl. Acad. Sci. U. S. A.*
- 120 (2023) e2220778120. https://doi.org/10.1073/pnas.2220778120.
-
- [241] N. NaderiAlizadeh, R. Singh, Aggregating residue-level protein language model embeddings with optimal transport,
- Bioinform. Adv.*
- 5 (2025) vbaf060. https://doi.org/10.1093/bioadv/vbaf060.
-
- [242] G. Manikandan, B. Pragadeesh, V. Manojkumar, A.L. Karthikeyan, R. Manikandan, A.H. Gandomi, Classification models combined with Boruta feature selection for heart disease prediction,
- Inform. Med. Unlocked*
- 44 (2024) 101442. https://doi.org/10.1016/j.imu.2023.101442.
-
- [243] M. Radovic, M. Ghalwash, N. Filipovic, Z. Obradovic, Minimum redundancy maximum relevance feature selection approach for temporal gene expression data,
- BMC Bioinformatics*
- 18 (2017) 9. https://doi.org/10.1186/s12859-016-1423-9.
-
- [244] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, in:
- Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003, IEEE Comput. Soc., 2004.*
- https://doi.org/10.1109/csb.2003.1227396.
-
- [245] S. Bhalla, H. Kaur, A. Dhall, G.P.S. Raghava, Prediction and analysis of skin cancer progression using genomics profiles of patients,
- Sci. Rep.*
- 9 (2019) 15790. https://doi.org/10.1038/s41598-019-52134-4.
-
- [246] V. Kumar, S. Patiyal, A. Dhall, N. Sharma, G.P.S. Raghava, B3Pred: A random-forest-based method for predicting and designing blood-brain barrier penetrating peptides,
- Pharmaceutics*
- 13 (2021) 1237. https://doi.org/10.3390/pharmaceutics13081237.
-
- [247] H. Zou, T. Hastie, Regularization and variable selection via the elastic net,
- J. R. Stat. Soc. Series B Stat. Methodol.*
- 67 (2005) 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.
-
- [248] Z. Xu, H. Zhang, Y. Wang, X. Chang, Y. Liang, L 1/2 regularization,
- Sci. China Inf. Sci.*
- 53 (2010) 1159–1169. https://doi.org/10.1007/s11432-010-0090-0.
-
- [249] R. Tibshirani, Regression shrinkage and selection via the lasso,
- J. R. Stat. Soc. Series B Stat. Methodol.*
- 58 (1996) 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
-
- [250] O. Demir-Kavuk, M. Kamada, T. Akutsu, E.-W. Knapp, Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features,
- BMC Bioinformatics*
- 12 (2011) 412. https://doi.org/10.1186/1471-2105-12-412.
-
- [251] M.B. Kursu, W.R. Rudnicki, Feature Selection with the Boruta Package,
- J. Stat. Softw.*
- 36 (2010). https://doi.org/10.18637/jss.v036.i11.
-
- [252] I.M. Guyon, N. Matic, V. Vapnik, Discovering Informative Patterns and Data Cleaning, (1996) 181–203. https://cdn.aaai.org/Workshops/1994/WS-94-03/WS94-03-013.pdf (accessed February 24, 2026).
-
- [253] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines,
- Mach. Learn.*
- 46 (2002) 389–422. https://doi.org/10.1023/a:1012487302797.
-
- [254] H. Sanz, C. Valim, E. Vegas, J.M. Oller, F. Reverter, SVM-RFE: selection and visualization of the most relevant features through non-linear kernels,
- BMC Bioinformatics*
- 19 (2018) 432. https://doi.org/10.1186/s12859-018-2451-4.

- [255] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, B. Rost, ProtTrans: Toward understanding the language of life through self-supervised learning, *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (2022) 7112–7127. <https://doi.org/10.1109/TPAMI.2021.3095381>.
- [256] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic level protein structure with a language model, *bioRxiv* (2022). <https://doi.org/10.1101/2022.07.20.500902>.
- [257] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, B. Rost, ProtTrans: Towards cracking the language of life's code through self-supervised learning, *bioRxiv* (2020). <https://doi.org/10.1101/2020.07.12.199554>.
- [258] B.E. Suzek, Y. Wang, H. Huang, P.B. McGarvey, C.H. Wu, UniProt Consortium, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches, *Bioinformatics* 31 (2015) 926–932. <https://doi.org/10.1093/bioinformatics/btu739>.
- [259] M. Steinegger, M. Mirdita, J. Söding, Protein-level assembly increases protein sequence recovery from metagenomic samples manifold, *Nat. Methods* 16 (2019) 603–606. <https://doi.org/10.1038/s41592-019-0437-4>.
- [260] F. Zhang, J. Li, Z. Wen, C. Fang, FusPB-ESM2: Fusion model of ProtBERT and ESM-2 for cell-penetrating peptide prediction, *Comput. Biol. Chem.* 111 (2024) 108098. <https://doi.org/10.1016/j.compbiolchem.2024.108098>.
- [261] C. Guntuboina, A. Das, P. Mollaei, S. Kim, A. Barati Farimani, PeptideBERT: A language model based on transformers for peptide property prediction, *J. Phys. Chem. Lett.* 14 (2023) 10427–10434. <https://doi.org/10.1021/acs.jpcclett.3c02398>.
- [262] Y. Liu, B. Tian, Protein-DNA binding sites prediction based on pre-trained protein language model and contrastive learning, *Brief. Bioinform.* 25 (2023). <https://doi.org/10.1093/bib/bbad488>.
- [263] A. Elnaggar, H. Essam, W. Salah-Eldin, W. Moustafa, M. Elkerdawy, C. Rochereau, B. Rost, Ankh: Optimized protein language model unlocks general-purpose modelling, *arXiv [Cs.LG]* (2023). <https://doi.org/10.48550/ARXIV.2301.06568>.
- [264] N. Ferruz, S. Schmidt, B. Höcker, ProtGPT2 is a deep unsupervised language model for protein design, *Nat. Commun.* 13 (2022) 4348. <https://doi.org/10.1038/s41467-022-32007-7>.
- [265] M. Heinzinger, K. Weissenow, J.G. Sanchez, A. Henkel, M. Mirdita, M. Steinegger, B. Rost, Bilingual language model for protein sequence and structure, *bioRxiv* (2023). <https://doi.org/10.1101/2023.07.23.550085>.
- [266] D. Kaur, C. Arora, G.P.S. Raghava, A hybrid model for predicting pattern recognition receptors using evolutionary information, *Front. Immunol.* 11 (2020) 71. <https://doi.org/10.3389/fimmu.2020.00071>.
- [267] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T.L. Madden, BLAST+: architecture and applications, *BMC Bioinformatics* 10 (2009) 421. <https://doi.org/10.1186/1471-2105-10-421>.
- [268] W.R. Pearson, Selecting the right similarity-scoring matrix, *Curr. Protoc. Bioinformatics* 43 (2013) 3.5.1–3.5.9. <https://doi.org/10.1002/0471250953.bi0305s43>.
- [269] C. Vens, M.-N. Rosso, E.G.J. Danchin, Identifying discriminative classification-based motifs in biological sequences, *Bioinformatics* 27 (2011) 1231–1238. <https://doi.org/10.1093/bioinformatics/btr110>.

- [270] T.L. Bailey, J. Johnson, C.E. Grant, W.S. Noble, The MEME suite, *Nucleic Acids Res.* 43 (2015) W39-49. <https://doi.org/10.1093/nar/gkv416>.
- [271] T.L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2 (1994) 28–36. <https://www.ncbi.nlm.nih.gov/pubmed/7584402>.
- [272] F. Pucci, R. Bourgeas, M. Rooman, Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC, *Sci. Rep.* 6 (2016) 23257. <https://doi.org/10.1038/srep23257>.
- [273] Y. Li, C.R. Middaugh, J. Fang, A novel scoring function for discriminating hyperthermophilic and mesophilic proteins with application to predicting relative thermostability of protein mutants, *BMC Bioinformatics* 11 (2010) 62. <https://doi.org/10.1186/1471-2105-11-62>.
- [274] X. Liang, F. Li, J. Chen, J. Li, H. Wu, S. Li, J. Song, Q. Liu, Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification, *Brief. Bioinform.* 22 (2021). <https://doi.org/10.1093/bib/bbaa312>.
- [275] Pearson's correlation coefficient, in: *Encyclopedia of Public Health*, Springer Netherlands, Dordrecht, 2008: pp. 1090–1091. https://doi.org/10.1007/978-1-4020-5614-7_2569.
- [276] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, *arXiv [Cs.AI]* (2017). <https://doi.org/10.48550/arXiv.1705.07874>.
- [277] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, *Nat. Mach. Intell.* 2 (2020) 56–67. <https://doi.org/10.1038/s42256-019-0138-9>.
- [278] T.A. Najar, S. Khare, R. Pandey, S.K. Gupta, R. Varadarajan, Mapping protein binding sites and conformational epitopes using cysteine labeling and yeast surface display, *Structure* 25 (2017) 395–406. <https://doi.org/10.1016/j.str.2016.12.016>.
- [279] A.S. De Groot, A. Bosma, N. Chinai, J. Frost, B.M. Jesdale, M.A. Gonzalez, W. Martin, C. Saint-Aubin, From genome to vaccine: in silico predictions, ex vivo verification, *Vaccine* 19 (2001) 4385–4395. [https://doi.org/10.1016/s0264-410x\(01\)00145-1](https://doi.org/10.1016/s0264-410x(01)00145-1).
- [280] HIV Immunology Database, (n.d.). <https://www.hiv.lanl.gov/content/immunology> (accessed February 24, 2026).
- [281] J. Liu, W. Zhang, Databases for B-cell epitopes, *Methods Mol. Biol.* 1184 (2014) 135–148. https://doi.org/10.1007/978-1-4939-1115-8_7.
- [282] S. Saha, M. Bhasin, G.P.S. Raghava, Bcipep: a database of B-cell epitopes, *BMC Genomics* 6 (2005) 79. <https://doi.org/10.1186/1471-2164-6-79>.
- [283] H. McSparron, M.J. Blythe, C. Zygouri, I.A. Doytchinova, D.R. Flower, JenPep: a novel computational information resource for immunobiology and vaccinology, *J. Chem. Inf. Comput. Sci.* 43 (2003) 1276–1287. <https://doi.org/10.1021/ci030461e>.
- [284] M.M. Hasan, M.S. Khatun, H. Kurata, ILBE for computational identification of linear B-cell epitopes by integrating sequence and evolutionary features, *Genomics Proteomics Bioinformatics* 18 (2020) 593–600. <https://doi.org/10.1016/j.gpb.2019.04.004>.
- [285] W. Dejnirattisai, D. Zhou, H.M. Ginn, H.M.E. Duyvesteyn, P. Supasa, J.B. Case, Y. Zhao, T.S. Walter, A.J. Mentzer, C. Liu, B. Wang, G.C. Paesen, J. Slon-Campos, C. López-Camacho, N.M. Kafai, A.L. Bailey, R.E. Chen, B. Ying, C. Thompson, J. Bolton, A. Fyfe, S. Gupta, T.K. Tan, J. Gilbert-Jaramillo, W. James, M. Knight, M.W. Carroll, D. Skelly, C. Dold, Y. Peng, R. Levin, T. Dong, A.J. Pollard, J.C. Knight, P. Klenerman, N. Temperton,

- D.R. Hall, M.A. Williams, N.G. Paterson, F.K.R. Bertram, C.A. Siebert, D.K. Clare, A. Howe, J. Radecke, Y. Song, A.R. Townsend, K.-Y.A. Huang, E.E. Fry, J. Mongkolsapaya, M.S. Diamond, J. Ren, D.I. Stuart, G.R. Screaton, The antigenic anatomy of SARS-CoV-2 receptor binding domain, *Cell* 184 (2021) 2183–2200.e22. <https://doi.org/10.1016/j.cell.2021.02.032>.
- [286] A.W. Purcell, J. McCluskey, J. Rossjohn, More than one reason to rethink the use of peptides in vaccine design, *Nat. Rev. Drug Discov.* 6 (2007) 404–414. <https://doi.org/10.1038/nrd2224>.
- [287] Diabetes, (n.d.). <https://www.who.int/news-room/fact-sheets/detail/diabetes> (accessed February 24, 2026).
- [288] A. Sapra, P. Bhandari, Diabetes, in: StatPearls, StatPearls Publishing, Treasure Island (FL), 2025. <https://www.ncbi.nlm.nih.gov/pubmed/31855345>.
- [289] J.-W. Yoon, H.-S. Jun, Autoimmune destruction of pancreatic beta cells, *Am. J. Ther.* 12 (2005) 580–591. <https://doi.org/10.1097/01.mjt.0000178767.67857.63>.
- [290] J.M. Forbes, M.E. Cooper, Mechanisms of diabetic complications, *Physiol. Rev.* 93 (2013) 137–188. <https://doi.org/10.1152/physrev.00045.2011>.
- [291] G.J. Kahaly, M.P. Hansen, Type 1 diabetes associated autoimmunity, *Autoimmun. Rev.* 15 (2016) 644–648. <https://doi.org/10.1016/j.autrev.2016.02.017>.
- [292] A.M. Anderson, L.G. Landry, A.A. Alkanani, L. Pyle, A.C. Powers, M.A. Atkinson, C.E. Mathews, B.O. Roep, A.W. Michels, M. Nakayama, Human islet T cells are highly reactive to preproinsulin in type 1 diabetes, *Proc. Natl. Acad. Sci. U. S. A.* 118 (2021) e2107208118. <https://doi.org/10.1073/pnas.2107208118>.
- [293] R.J. Ellis, R. Varela-Calvino, T.I.M. Tree, M. Peakman, HLA Class II molecules on haplotypes associated with type 1 diabetes exhibit similar patterns of binding affinities for coxsackievirus P2C peptides, *Immunology* 116 (2005) 337–346. <https://doi.org/10.1111/j.1365-2567.2005.02233.x>.
- [294] N. Principi, M.G. Berioli, S. Bianchini, S. Esposito, Type 1 diabetes and viral infections: What is the relationship?, *J. Clin. Virol.* 96 (2017) 26–31. <https://doi.org/10.1016/j.jcv.2017.09.003>.
- [295] M.C. Honeyman, B.S. Coulson, N.L. Stone, S.A. Gellert, P.N. Goldwater, C.E. Steele, J.J. Couper, B.D. Tait, P.G. Colman, L.C. Harrison, Association between rotavirus infection and pancreatic islet autoimmunity in children at risk of developing type 1 diabetes, *Diabetes* 49 (2000) 1319–1324. <https://doi.org/10.2337/diabetes.49.8.1319>.
- [296] C.Y. Pak, H.M. Eun, R.G. McArthur, J.W. Yoon, Association of cytomegalovirus infection with autoimmune type 1 diabetes, *Lancet* 2 (1988) 1–4. [https://doi.org/10.1016/s0140-6736\(88\)92941-8](https://doi.org/10.1016/s0140-6736(88)92941-8).
- [297] H. Hyöty, P. Leinikki, A. Reunanen, J. Ilonen, H.M. Surcel, A. Rilva, M.L. Käär, T. Huupponen, A. Hakulinen, A.L. Mäkelä, Mumps infections in the etiology of type 1 (insulin-dependent) diabetes, *Diabetes Res.* 9 (1988) 111–116. <https://www.ncbi.nlm.nih.gov/pubmed/3243043>.
- [298] H. Hyöty, K.W. Taylor, The role of viruses in human diabetes, *Diabetologia* 45 (2002) 1353–1361. <https://doi.org/10.1007/s00125-002-0852-3>.
- [299] R. Mallone, S. You, The SAgA of antigen-specific immunotherapy for type 1 diabetes, *Diabetes* 70 (2021) 1247–1249. <https://doi.org/10.2337/dbi21-0011>.
- [300] A. Hjern, Chapter 5.8: major public health problems - allergic disorders, *Scand. J. Public Health Suppl.* 67 (2006) 125–131. <https://doi.org/10.1080/14034950600677139>.

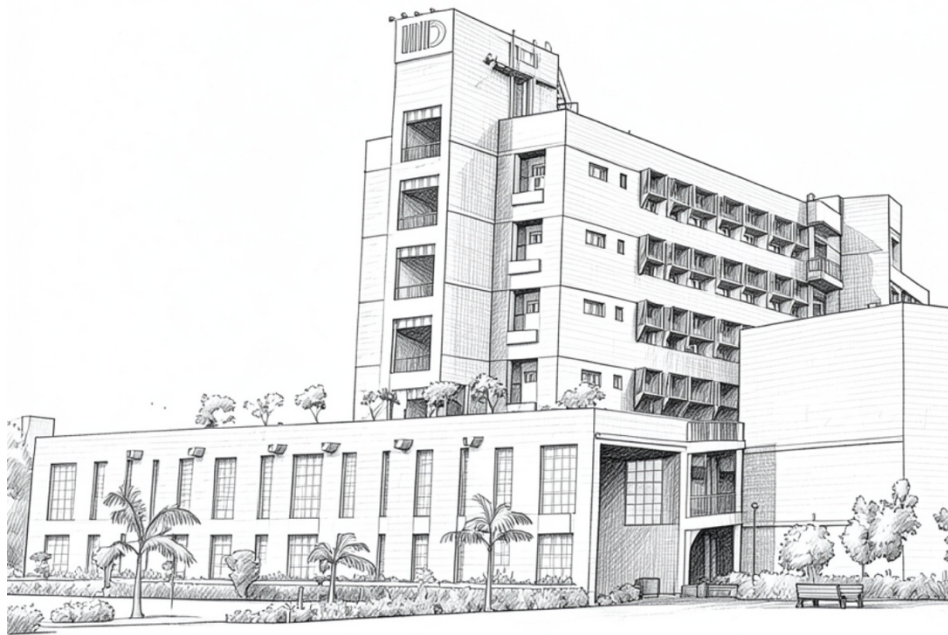
- [301] J. Wang, Y. Zhou, H. Zhang, L. Hu, J. Liu, L. Wang, T. Wang, H. Zhang, L. Cong, Q. Wang, Pathogenesis of allergic diseases and implications for therapeutic interventions, *Signal Transduct. Target. Ther.* 8 (2023) 138. <https://doi.org/10.1038/s41392-023-01344-4>.
- [302] A.E. Ng, P. Boersma, Diagnosed allergic conditions in adults: United States, 2021, *NCHS Data Brief* (2023) 1–8. <https://www.ncbi.nlm.nih.gov/pubmed/36700875>.
- [303] C. Traidl-Hoffmann, J. Afghani, C.A. Akdis, M. Akdis, H. Aydin, K. Bärenfaller, H. Behrendt, T. Bieber, P. Bigliardi, M. Bigliardi-Qi, C.M. Bonefeld, S. Bösch, M.C. Brüggen, S. Diemert, H.-W. Duchna, M. Fährndrich, D. Fehr, M. Fellmann, R. Frei, L.H. Garvey, R. Gharbo, M. Gökkaya, K. Grando, C. Guillet, E. Guler, J. Gutermuth, N. Herrmann, D.J. Hijnen, C. Hülpmisch, A.D. Irvine, E. Jensen-Jarolim, H.H. Kong, H. Koren, C.C.V. Lang, R. Lauener, L. Maintz, P.-Y. Mantel, E. Maverakis, M. Möhrenschrager, S. Müller, K. Nadeau, A.U. Neumann, L. O'Mahony, F.R. Rabenja, H. Renz, C. Rhyner, E. Rietschel, J. Ring, C. Roduit, M. Sasaki, M. Schenk, J. Schröder, D. Simon, H.-U. Simon, M. Sokolowska, S. Ständer, M. Steinhoff, D.S. Piccirillo, A. Taïeb, R. Takaoka, M. Tapparo, H. Teixeira, J.P. Thyssen, S. Traidl, M. Uhlmann, W. van de Veen, M. van Hage, C. Virchow, A. Wollenberg, M. Yasutaka, A. Zink, P. Schmid-Grendelmeier, Navigating the evolving landscape of atopic dermatitis: Challenges and future opportunities: The 4th Davos declaration, *Allergy* 79 (2024) 2605–2624. <https://doi.org/10.1111/all.16247>.
- [304] S.J. Galli, M. Tsai, A.M. Piliponsky, The development of allergic inflammation, *Nature* 454 (2008) 445–454. <https://doi.org/10.1038/nature07204>.
- [305] M. Jutel, I. Agache, M. Zemelka-Wiacek, M. Akdis, T. Chivato, S. Del Giacco, P. Gajdanowicz, I.E. Gracia, L. Klimek, A. Lauerma, M. Ollert, L. O'Mahony, J. Schwarze, M.H. Shamji, I. Skypala, O. Palomares, O. Pfaar, M.J. Torres, J.A. Bernstein, A.A. Cruz, S.R. Durham, S.J. Galli, R.M. Gómez, E. Guttman-Yassky, T. Haahtela, S.T. Holgate, K. Izuhara, K. Kabashima, D.E. Larenas-Linnemann, E. von Mutius, K.C. Nadeau, R. Pawankar, T.A.E. Platts-Mills, S.H. Sicherer, H.-S. Park, S. Vieths, G. Wong, L. Zhang, M.B. Bilò, C.A. Akdis, Nomenclature of allergic diseases and hypersensitivity reactions: Adapted to modern needs: An EAACI position paper, *Allergy* 78 (2023) 2851–2874. <https://doi.org/10.1111/all.15889>.
- [306] C. Guo, Y. Tang, Q. Li, Z. Yang, Y. Guo, C. Chen, Y. Zhang, Deciphering the immune heterogeneity dominated by natural killer cells with prognostic and therapeutic implications in hepatocellular carcinoma, *Comput. Biol. Med.* 158 (2023) 106872. <https://doi.org/10.1016/j.compbiomed.2023.106872>.
- [307] A. Syed, A. Kohli, K.C. Nadeau, Food allergy diagnosis and therapy: where are we now?, *Immunotherapy* 5 (2013) 931–944. <https://doi.org/10.2217/imt.13.93>.
- [308] K.C.M. Verhoeckx, Y.M. Vissers, J.L. Baumert, R. Faludi, M. Feys, S. Flanagan, C. Herouet-Guicheney, T. Holzhauser, R. Shimojo, N. van der Bolt, H. Wichers, I. Kimber, Food processing and allergenicity, *Food Chem. Toxicol.* 80 (2015) 223–240. <https://doi.org/10.1016/j.fct.2015.03.005>.
- [309] I. Kimber, R.J. Dearman, A.H. Penninks, L.M.J. Knippels, R.B. Buchanan, B. Hammerberg, H.A. Jackson, R.M. Helm, Assessment of protein allergenicity on the basis of immune reactivity: animal models, *Environ. Health Perspect.* 111 (2003) 1125–1130. <https://doi.org/10.1289/ehp.5813>.
- [310] G.S. Ladics, M.K. Selgrade, Identifying food proteins with allergenic potential: evolution of approaches to safety assessment and research to provide additional tools, *Regul. Toxicol. Pharmacol.* 54 (2009) S2-6. <https://doi.org/10.1016/j.yrtph.2008.10.010>.

- [311] M. Nedyalkova, M. Vasighi, A. Azmoon, L. Naneva, V. Simeonov, Sequence-based prediction of plant allergenic proteins: Machine learning classification approach, *ACS Omega* 8 (2023) 3698–3704. <https://doi.org/10.1021/acsomega.2c02842>.
- [312] J. Dey, S.R. Mahapatra, S. Patnaik, S. Lata, G.S. Kushwaha, R.K. Panda, N. Misra, M. Suar, Molecular characterization and designing of a novel multiepitope vaccine construct against *Pseudomonas aeruginosa*, *Int. J. Pept. Res. Ther.* 28 (2022) 49. <https://doi.org/10.1007/s10989-021-10356-z>.
- [313] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T.L. Scao, S. Gugger, M. Drame, Q. Lhoest, A.M. Rush, HuggingFace’s transformers: State-of-the-art natural language processing, *arXiv [Cs.CL]* (2019). <https://doi.org/10.48550/ARXIV.1910.03771>.
- [314] J. Qi, C.-H. Yang, S.Y.-C. Chen, P.-Y. Chen, Quantum machine learning: An interplay between quantum computing and machine learning, *arXiv [Quant-Ph]* (2024). <http://arxiv.org/abs/2411.09403>.
- [315] Y. Du, X. Wang, N. Guo, Z. Yu, Y. Qian, K. Zhang, M.-H. Hsieh, P. Rebentrost, D. Tao, Quantum machine learning: A hands-on tutorial for machine learning practitioners and researchers, *arXiv [Quant-Ph]* (2025). <http://arxiv.org/abs/2502.01146>.
- [316] V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, S. Ahmed, V. Ajith, M.S. Alam, G. Alonso-Linaje, B. AkashNarayanan, A. Asadi, J.M. Arrazola, U. Azad, S. Banning, C. Blank, T.R. Bromley, B.A. Cordier, J. Ceroni, A. Delgado, O. Di Matteo, A. Dusko, T. Garg, D. Guala, A. Hayes, R. Hill, A. Ijaz, T. Isaacson, D. Ittah, S. Jahangiri, P. Jain, E. Jiang, A. Khandelwal, K. Kottmann, R.A. Lang, C. Lee, T. Loke, A. Lowe, K. McKiernan, J.J. Meyer, J.A. Montañez-Barrera, R. Moyard, Z. Niu, L.J. O’Riordan, S. Oud, A. Panigrahi, C.-Y. Park, D. Polatajko, N. Quesada, C. Roberts, N. Sá, I. Schoch, B. Shi, S. Shu, S. Sim, A. Singh, I. Strandberg, J. Soni, A. Száva, S. Thabet, R.A. Vargas-Hernández, T. Vincent, N. Vitucci, M. Weber, D. Wierichs, R. Wiersema, M. Willmann, V. Wong, S. Zhang, N. Killoran, PennyLane: Automatic differentiation of hybrid quantum-classical computations, *arXiv [Quant-Ph]* (2018). <http://arxiv.org/abs/1811.04968>.
- [317] B.G. Iovino, H. Tang, Y. Ye, Protein domain embeddings for fast and accurate similarity search, *bioRxiv* (2023). <https://doi.org/10.1101/2023.11.27.567555>.
- [318] S. Tomar, R. Tripathi, S. Kumar, Comprehensive survey of QML: From data analysis to algorithmic advancements, *arXiv [Quant-Ph]* (2025). <http://arxiv.org/abs/2501.09528>.
- [319] E.H. Houssein, Z. Abohashima, M. Elhoseny, W.M. Mohamed, Machine learning in the quantum realm: The state-of-the-art, challenges, and future vision, *Expert Syst. Appl.* 194 (2022) 116512. <https://doi.org/10.1016/j.eswa.2022.116512>.
- [320] L. Yu, Y. Luo, S. Wu, S. Chen, L. Xue, R. Jing, J. Luo, Multimodal deep learning for allergenic proteins prediction, *BMC Biol.* 23 (2025) 232. <https://doi.org/10.1186/s12915-025-02347-z>.
- [321] S. Timr, D. Madern, F. Sterpone, Protein thermal stability, *Prog. Mol. Biol. Transl. Sci.* 170 (2020) 239–272. <https://doi.org/10.1016/bs.pmbts.2019.12.007>.
- [322] K. Cabello-Solorzano, I. Ortigosa de Araujo, M. Peña, L. Correia, A. J. Tallón-Ballesteros, The impact of data normalization on the accuracy of machine learning algorithms: A comparative analysis, in: *Lecture Notes in Networks and Systems*, Springer Nature Switzerland, Cham, 2023: pp. 344–353. https://doi.org/10.1007/978-3-031-42536-3_33.

- [323] E.L.A. Poirier, H2O AutoML: Scalable Automatic Machine Learning, ICML 2020 AutoML Workshop (2020). <https://api.semanticscholar.org/CorpusID:221338558>.
- [324] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E., scikit-learn, *Journal of Machine Learning Research* 12 (2011) 2825–2830. <http://jmlr.org/papers/v12/pedregosa11a.html>.
- [325] M. Hajihosseini, A. Maghsoudi, R. Ghezelbash, A novel scheme for mapping of MVT-type Pb–Zn prospectivity: LightGBM, a highly efficient gradient boosting decision tree machine learning algorithm, *Nat. Resour. Res.* 32 (2023) 2417–2438. <https://doi.org/10.1007/s11053-023-10249-6>.
- [326] P. Koprinkova-Hristova, V. Mladenov, N.K. Kasabov, eds., *Artificial neural networks*, 2015th ed., Springer International Publishing, Cham, Switzerland, 2014. <https://doi.org/10.1007/978-3-319-09903-3>.
- [327] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, *arXiv [Cs.LG]* (2014). <https://doi.org/10.48550/ARXIV.1412.6980>.
- [328] P.K. Ponnuswamy, R. Muthusamy, P. Manavalan, Amino acid composition and thermal stability of proteins, *Int. J. Biol. Macromol.* 4 (1982) 186–190. [https://doi.org/10.1016/0141-8130\(82\)90049-6](https://doi.org/10.1016/0141-8130(82)90049-6).
- [329] S. Kumar, C.J. Tsai, R. Nussinov, Factors enhancing protein thermostability, *Protein Eng.* 13 (2000) 179–191. <https://doi.org/10.1093/protein/13.3.179>.
- [330] K.B. Zeldovich, I.N. Berezovsky, E.I. Shakhnovich, Protein and DNA sequence determinants of thermophilic adaptation, *PLoS Comput. Biol.* 3 (2007) e5. <https://doi.org/10.1371/journal.pcbi.0030005>.
- [331] C. Marquet, M. Heinzinger, T. Olenyi, C. Dallago, K. Erckert, M. Bernhofer, D. Nechaev, B. Rost, Embeddings from protein language models predict conservation and variant effects, *Hum. Genet.* 141 (2022) 1629–1647. <https://doi.org/10.1007/s00439-021-02411-y>.
- [332] Y. Dehouck, B. Folch, M. Rومان, Revisiting the correlation between proteins' thermoresistance and organisms' thermophilicity, *Protein Eng. Des. Sel.* 21 (2008) 275–278. <https://doi.org/10.1093/protein/gzn001>.
- [333] R. Cavicchioli, K.S. Siddiqui, D. Andrews, K.R. Sowers, Low-temperature extremophiles and their applications, *Curr. Opin. Biotechnol.* 13 (2002) 253–261. [https://doi.org/10.1016/s0958-1669\(02\)00317-8](https://doi.org/10.1016/s0958-1669(02)00317-8).
- [334] L. Zhao, J. Zhao, K. Zhong, A. Tong, D. Jia, Targeted protein degradation: mechanisms, strategies and application, *Signal Transduct. Target. Ther.* 7 (2022) 113. <https://doi.org/10.1038/s41392-022-00966-4>.
- [335] L. García-Descalzo, E. García-López, C. Cid, Comparative proteomic analysis of psychrophilic vs. Mesophilic bacterial species reveals different strategies to achieve temperature adaptation, *Front. Microbiol.* 13 (2022) 841359. <https://doi.org/10.3389/fmicb.2022.841359>.
- [336] M.-L. García-López, J.A. Santos, A. Otero, J.M. Rodríguez-Calleja, Psychrobacter, in: *Encyclopedia of Food Microbiology*, Elsevier, 2014: pp. 261–268. <https://doi.org/10.1016/b978-0-12-384730-0.00285-8>.
- [337] S. Ikeda, Y. Takamatsu, M. Tsuchiya, K. Suga, Y. Tanaka, A. Kouzuma, K. Watanabe, *Shewanella oneidensis* MR-1 as a bacterial platform for electro-biotechnology, *Essays Biochem.* 65 (2021) 355–364. <https://doi.org/10.1042/EBC20200178>.

- [338] C.L. Karp, D. Lans, J. Esparza, E.B. Edson, K.E. Owen, C.B. Wilson, P.M. Heaton, O.S. Levine, R. Rao, Evaluating the value proposition for improving vaccine thermostability to increase vaccine impact in low and middle-income countries, *Vaccine* 33 (2015) 3471–3479. <https://doi.org/10.1016/j.vaccine.2015.05.071>.
- [339] L. Opsomer, S. Jana, R. Hoogenboom, N. Sanders, Overcoming cold chain challenges: Cryoprotection of self-amplifying mRNA polyplexes for long-term storage at ambient temperature, *J. Control. Release* 387 (2025) 114259. <https://doi.org/10.1016/j.jconrel.2025.114259>.
- [340] Y. Sun, D. Maltseva, J. Liu, T. Hooker 2nd, V. Mailänder, H. Ramløv, A.L. DeVries, M. Bonn, K. Meister, Ice recrystallization inhibition is insufficient to explain cryopreservation abilities of antifreeze proteins, *Biomacromolecules* 23 (2022) 1214–1220. <https://doi.org/10.1021/acs.biomac.1c01477>.
- [341] M.T. Warren, C.I. Biggs, A. Bissoyi, M.I. Gibson, G.C. Sosso, Data-driven discovery of potent small molecule ice recrystallisation inhibitors, *Nat. Commun.* 15 (2024) 8082. <https://doi.org/10.1038/s41467-024-52266-w>.
- [342] A. Leyba, A. Francian, M. Razjmoo, A. Bierle, R. Janardhana, N. Jackson, B. Chackerian, P. Muttil, Formulation, characterization, and in vivo immunogenicity of heat-stabilized dissolvable microneedles containing a novel VLP vaccine, *Pharm. Res.* 42 (2025) 1587–1601. <https://doi.org/10.1007/s11095-025-03927-y>.
- [343] X. Ding, H. Zhang, H. Chen, L. Wang, H. Qian, X. Qi, Extraction, purification and identification of antifreeze proteins from cold acclimated malting barley (*Hordeum vulgare* L.), *Food Chem.* 175 (2015) 74–81. <https://doi.org/10.1016/j.foodchem.2014.11.027>.
- [344] A. Alim, A. Rafay, I. Naseem, PoGB-pred: Prediction of antifreeze proteins sequences using amino acid composition with feature selection followed by a sequential-based ensemble approach, *Curr. Bioinform.* 16 (2021) 446–456. <https://doi.org/10.2174/1574893615999200707141926>.
- [345] A. Benkessirat, N. Benblidia, Fundamentals of feature selection: An overview and comparison, in: 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), IEEE, 2019. <https://doi.org/10.1109/aiccsa47632.2019.9035281>.
- [346] R. Kumar, B. Panwar, J.S. Chauhan, G.P. Raghava, Analysis and prediction of cancerlectins using evolutionary and domain information, *BMC Res. Notes* 4 (2011) 237. <https://doi.org/10.1186/1756-0500-4-237>.
- [347] Peptide therapeutics market size to hit USD 87.21 billion by 2035, (2025). <https://www.precedenceresearch.com/peptide-therapeutics-market> (accessed February 25, 2026).
- [348] J.J. Dwyer, K.L. Wilson, D.K. Davison, S.A. Freel, J.E. Seedorff, S.A. Wring, N.A. Tvermoes, T.J. Matthews, M.L. Greenberg, M.K. Delmedico, Design of helical, oligomeric HIV-1 fusion inhibitor peptides with potent activity against enfuvirtide-resistant virus, *Proc. Natl. Acad. Sci. U. S. A.* 104 (2007) 12772–12777. <https://doi.org/10.1073/pnas.0701478104>.
- [349] L.T. Nguyen, J.K. Chau, N.A. Perry, L. de Boer, S.A.J. Zaat, H.J. Vogel, Serum stabilities of short tryptophan- and arginine-rich antimicrobial peptide analogs, *PLoS One* 5 (2010) e12684. <https://doi.org/10.1371/journal.pone.0012684>.
- [350] P. Wang, L. Hu, G. Liu, N. Jiang, X. Chen, J. Xu, W. Zheng, L. Li, M. Tan, Z. Chen, H. Song, Y.-D. Cai, K.-C. Chou, Prediction of antimicrobial peptides based on sequence alignment and feature selection methods, *PLoS One* 6 (2011) e18476. <https://doi.org/10.1371/journal.pone.0018476>.

- [351] C. Yan, J. Ma, H. Luo, G. Zhang, J. Luo, A novel feature selection method for high-dimensional biomedical data based on an improved binary clonal flower pollination algorithm, *Hum. Hered.* 84 (2019) 34–46. <https://doi.org/10.1159/000501652>.
- [352] Z.-Y. Zhang, Z.-J. Sun, Y.-H. Yang, H. Lin, Towards a better prediction of subcellular location of long non-coding RNA, *Front. Comput. Sci.* 16 (2022). <https://doi.org/10.1007/s11704-021-1015-3>.
- [353] H. Yu, AI foundation models in plant biology, *New Phytol.* (2026). <https://doi.org/10.1111/nph.71395>.
- [354] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. Dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model, *Science* 379 (2023) 1123–1130. <https://doi.org/10.1126/science.ade2574>.



INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**