# Development of in silico tools for designing cancer immunotherapy or subunit vaccine

**By**

**Anjali Dhall**

**(PhD17207)**

**Under the Supervision of Prof. Gajendra P.S. Raghava**

Department of Computational Biology

Indraprastha Institute of Information Technology

New Delhi – 110020

October, 2022

# Development of in silico tools for designing cancer immunotherapy or subunit vaccine

**By**

**Anjali Dhall**

**(PhD17207)**

A Thesis

Submitted in Partial Fulfilment of the Requirements for the Degree Of

**Doctor of Philosophy**

**Under the Supervision of Prof. Gajendra P.S. Raghava**

Department of Computational Biology

Indraprastha Institute of Information Technology

New Delhi – 110020

October, 2022

# Certificate

This is to certify that the thesis entitled **"Development of in silico tools for designing cancer immunotherapy or subunit vaccine"** being submitted by **Miss. Anjali Dhall** 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 her under my supervision. In my opinion, the thesis has reached the standards, fulfilling the requirements of the regulations relating to the degree.

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

October, 2022

**Prof. Gajendra P.S. Raghava**

Supervisor Name

Indraprastha Institute of Information Technology Delhi

New Delhi - 11002

# Acknowledgements

First and foremost, I would like to express my gratitude to **GOD Almighty**, for giving me the courage, strength, and resources to start a Ph.D. project, persevere through it, and successfully complete it.

I would like to express my sincere gratitude and thank to my Ph.D. supervisor, **Prof. Gajendra P.S. Raghava,** who accepted me as a student back in 2018 and gave me the chance to work in his lab in order to complete my research work. He has always been there to offer his sincere support and counsel, and he has provided me with essential direction, inspiration, and suggestions in my search for knowledge. He has granted me complete flexibility to conduct my study and to engage in productive discussions about novel concepts, all the while ensuring that I keep on track and don't lose sight of my objectives. Without his tremendous assistance and support, this thesis would not have been possible, and I will always be grateful to him for that. He has served as the ideal example for me of how to maintain positivity and kindness in both my work and personal life. For my Ph.D. research, I could not have asked for a better supervisor than him.

I would like to convey my profound thanks to **Prof. Gajendra P. S. Raghava**, **Prof. Pankaj Jalote** and **Prof. Ranjan Bose** (the IIIT-D Director) for allowing me the opportunity to be admitted to the PhD programme and for afterwards allowing me to utilise the institute's resources for productive research. My profound gratitude goes out to **Dr. Debarka Sengupta** and **Dr. Vibhor Kumar** for being in my committee member and helping me along the way with their guidance and expertise. Additionally, I'd like to thank, **Prof. Gajendra P. S. Raghava, Dr. Debarka Sengupta, Dr. Subhadip Raychaudhuri, Dr. Vibhor Kumar, and Dr. Angshul Majumdar** for teaching me during my course work. The administrative personnel at IIIT-D, especially **Mrs. Priti Patel, Ms. Shipra Jain, Ms. Sheetu Ahuja,** and **Ms. Anshu Dureja,** deserve a special thank you for always being accessible to answer our queries and immediately resolve our academic issues. My sincere gratitude goes out to **Mr. Imran Khan, Nidhi Mam,** and **Mr. Kapil Dev Garg** for timely delivering my stipend. I am also appreciative of IIIT-D for offering first-rate facilities and infrastructure. An official thanks to Department of Science & Technology (DST), the funding source, for giving me the **"Innovation in Science Pursuit for Inspired Research (INSPIRE)"** research fellowship to help with my doctoral studies. I would especially like to thank **Dr. Harpreet Singh, Mrs. Purnima Sharma,** and **Dr. Shallu Kalia,** who taught me at the H.M.V. college, for giving me the information and direction I needed to begin this wonderful scientific journey.

My deepest thanks to my lab-mates and seniors, who are always encouraging and ensured a safe and fun working environment. I am incredibly appreciative of everyone in the lab, both those who were there before me and those who came after me. My seniors, **Dr. Kumardeep Chaudhary, Dr. Sandeep Kumar Dhandha, Dr. Gandharva nagpal, Dr. Sherry Bhalla, Dr. Piyush Agrawal, Dr. Salman Sadullah Usmani, Dr. Pawan Kumar Raghav, Dr. Akshara Pande, Dr. Lubna Maryam, Dr. Harpreet Kaur, Dr. Vinod Poriya, Dr. Rajesh Kumar, Dr. Anjali Lathwal, Dr. Chakit Arora, Dr. Dilraj Kaur, Dr. Neelam Sharma and Dr. Leimarembi Devi Naorem** taught me a lot. It's impossible for me to think of better, more helpful, and cooperative lab mates. With my seniors **Dr. Sherry Bhalla, Dr. Piyush Agrawal, Dr. Salman Sadullah Usmani, Dr. Harpreet Kaur, and Dr. Vinod Poriya,** I had most delightful experience. My special and deepest thank to **Dr. Sherry Bhalla and Dr. Harpreet Kaur** for initiating my first projects and guiding me throughout my journey. My sincere and heartiest gratitude to my wonderful colleague **Mr. Sumeet Patiyal** for supporting me intellectually and academically, especially during challenging and difficult circumstances. His knowledge and expertise always helped me finish the most challenging portion of my research. I would like to thank my wonderful colleague **Miss. Shipra Jain** for her support and help during my PhD. I had some greatest enjoyable time with my juniors **Ritu Tomer, Akanksha Arora. Shubham Choudhury, Nishant Kumar, Nisha Bajiya, Shivani Malik, Anand Singh Rathore, Mansi Goel and Gaurav.** I would also like to thank my Ph.D. batchmates and seniors along with whom I pursued my pre-Ph.D. course work. I want to thank **Neetesh Pandey, Indra Prakash Jha, Omkar R. Chandra, Dr. Krishan Gupta, Sarita Poonia, Smriti Chawla, Shreya Mishra, Raghava Awasthi, Divya** for being supportive and available at all times when I needed them. The most heart-warming feeling for me is to thank my adorable juniors- **Nishant, Shubham, Ritu, Nisha, Shivani, Anand, Alok, Shruti, Madhu, Samridhi, Vishakha, Sanjay, Aayshi, Sakshi, Shubadeep, Gayatri, Sukriti, Saveena, Pushpendra, Akshya, and Pradeep** who extended me a mealtime and teatime companionship, revitalizing me and reinfusing a carefree attitude of a young learner into me. I'm grateful to all of these incredible people for including me in their own significant and amazing life events, including their birthdays and career triumphs.

I want to express my gratitude to the IIIT-D administrative personnel for providing the hostel amenities. I would especially like to thank **Rajeev Ji and Malti aunty** (the hostel warden), **Tinku bhaiya, Naveen Ji** for being available and doing the task as soon as possible. I must express my gratitude to the mess and canteen for feeding me day and night. These people have given me wholesome food while I was away from

home, and I will be forever grateful. I am appreciative of the "**Ravi Tea Stall**" and its proprietor Rajesh Ji for providing me with energising tea in the morning and evening, which helped my brain's neurons recover after a long day of exhaustion.

Last but not the least, I would like to thank my beloved grandmother **Late. Smt. Shanti Devi**, who encourage me and support me to go for higher studies. I would like to thank my wonderful parents **Smt. Sunita Rani** and **Mr. Ashwani Kumar** for providing me inner strength and continuous support during the toughest phase of my life. I want to thank my big brothers **Mr. Karan Kumar** and **Mr. Honey Dhall** for their love, care and support during this journey. I would like to thank my lovely sister-in-law's **Mrs. Deepika Dhall** and **Mrs. Surbhi Dhall** for their love, support and guidance. My deepest love to my niece and nephew (**Naira & Jaivin**), and youngest member of my family my niece **Ojasvi** who bring happiness in my life. Nevertheless, I would like to thank my Delhi relatives & cousins, family members and friends (**Pia**, **Tanvi, Abhijeet and Gazal**) for their love and support during my PhD tenure.

**Anjali Dhall**

# Abstract

One of the major challenges in designing the cancer vaccine or immunotherapy is to predict the cancer-specific peptides or neopeptides that can stimulate the immune system to fight against the cancer cells. Human leukocyte antigens (HLA) bind and present neopeptides on the cell surface, where these neopeptides are recognized by the T-cells. T-cells activate a wide range of cytokines to provide protection/defence against the cancer cells. Thus, it is important to investigate the role of cytokines and HLA molecules in order to design the cancer immunotherapy. Broadly, this study can be divided in the following four parts; i) Prognostic biomarkers, ii) HLA binders, iii) Cytokine inducing peptides, and iv) Inhibition of STAT3. Firstly, we have investigated the prognostic role of class-I HLA (HLA-I) alleles, HLA-I binders and cytokines with the overall survival of the cancer patients. It was observed that certain HLA-alleles have high impact on the survival of a patients suffering from a specific type of cancer. Based on this observation, a method SKCMhrp has been developed for predicting high-risk cutaneous melanoma patients using HLA-alleles. In the past, numerous methods have been developed for predicting binders of classical HLA alleles. Thus, second part of this thesis describe methods developed for predicting binders of non-classical HLA alleles (HLA-G and HLA-E). Our server $HLA_{nc}$Pred allow users to predict promiscuous binders for non-classical HLA-alleles (HLA-G*01:01, HLA-G*01:03, HLA-G*01:04, HLA-E*01:01, and HLA-E*01:03). Thirdly, methods have been developed to predict peptides or epitopes that can induce following types of cytokine; IL6 (IL6Pred), TNF-α (TNFepitope), and IFN-γ (IFNepitope2). It has been shown in number of studies that STAT3 is a promising therapeutic target for several diseases including cancer. Thus, fourthly, a method has been developed to predict STAT3 inhibitor that can inhibit the STAT3 signaling pathway. In summary, in this thesis a number of in silico tools have been developed, which may play vital role directly or indirectly in developing the cancer vaccine/immunotherapy.

# List of Publications

## *Thesis Related Publications*

- **Dhall A,** Patiyal S, Kaur H, Bhalla S, Arora C, Raghava GPS. Computing skin cutaneous melanoma outcome from the HLA-alleles and clinical characteristics. ***Front Genet.*** 2020; 11:22.
- **Dhall A#,** Patiyal S#, Sharma N, Usmani SS, Raghava GPS. Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. ***Brief Bioinform***. 2021 Mar;22(2):936–45.
- **Dhall A,** Patiyal S, Sharma N, Devi NL, Raghava GPS**.** Computer-aided prediction of inhibitors against STAT3 for managing COVID-19 associated cytokine storm. ***Comput Biol Med.*** 2021;137(October 2021):104780
- **Dhall A#,** Patiyal S#, Raghava GPS. $HLA_{nc}$Pred: A method for predicting promiscuous non-classical HLA binding sites. ***Brief Bioinform***. 2022; bbac192.
- **Anjali Dhall,** Sumeet Patiyal, Shubham Choudhury, Shipra Jain, Kashish Narang, Gajendra PS Raghava. Prediction, scanning and designing of TNF-α inducing epitopes for human and mouse. 2022 (Under Communication)
- **Anjali Dhall,** Sumeet Patiyal, Harpreet Kaur, Gajendra PS Raghava. Risk assessment of cancer patients based on HLA-I alleles, neobinders and expression of cytokines. 2022 (Under Communication)

## *Book Chapter*

- **Dhall A,** Jain S, Sharma N, Naorem LD, Kaur D, Patiyal S, Raghava GPS. In silico tools and databases for designing cancer immunotherapy. ***Adv Protein Chem Struct*** Biol. 2022;129:1-50.

## *Other Publications*

- Patiyal S#, Kaur D#, Kaur H#, Sharma N#, **Dhall A,** et al. A Web-Based Platform on Coronavirus Disease-19 to Maintain Predicted Diagnostic, Drug, and Vaccine Candidates. ***Monoclon Antib Immunodiagn Immunother.***2020;39(6):204–16.
- Bhalla, Sherry; Kaur, Harpreet; **Dhall, Anjali**; Raghava, Gajendra PS; Prediction and analysis of skin cancer progression using genomics profiles of patients. ***Scientific reports,*** Nature Publishing Group, 2019.
- Kaur, Harpreet; **Dhall, Anjali**; Kumar, Rajesh; Raghava, Gajendra PS; Identification of platform-independent diagnostic biomarker panel for hepatocellular carcinoma

using large-scale transcriptomics data. ***Frontiers in genetics***, 2020.

- Sharma N, Patiyal S, **Dhall A,** Pande A, Arora C, Raghava GPS**.** AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. ***Brief Bioinform***. 2021;22(4):bbaa294.
- Kumar V, Patiyal S, **Dhall A,** Sharma N, Raghava GP**.** B3Pred: A Random-Forest-Based Method for Predicting and Designing Blood–Brain Barrier Penetrating Peptides. Vol. 13, ***Pharmaceutics***. 2021
- Sharma N, Patiyal S, **Dhall A,** Naorem DL, Raghava GPS**.** ChAlPred: A web server for prediction of allergenicity of chemical compounds. ***Comput Biol Med.*** 2021;136(September 2021):104746.
- Attila Gabor, Alice Driessen, Jovan Tanevski, Baosen Guo, Wencai Cao, He Shen, et al. Cell-to-cell and type-to-type heterogeneity of signaling networks: insights from the crowd. ***Mol Syst Biol.*** 2021;17(10).
- Tarca AL, Pataki BA, Romero R, Sirota M, Guan Y, Kutum R, et al. Crowdsourcing assessment of maternal blood multi-omics for predicting gestational age and preterm birth. ***Cell Reports Med***. 2021;2(6)
- Jain S, **Dhall A,** Patiyal S, Raghava GPS**.** IL13Pred: A method for predicting immunoregulatory cytokine IL-13 inducing peptides. ***Comput Biol Med.*** 2021;137(October 2021):104780.
- Patiyal S, Agrawal P, Kumar V, **Dhall A,** Kumar R, Mishra G, et al. NAGbinder: An approach for identifying N-acetylglucosamine interacting residues of a protein from its primary sequence. ***Protein Sci.*** 2020 Jan;29(1):201–10.
- Patiyal S[#], **Dhall A[#],** Raghava GPS. Prediction of risk-associated genes and high-risk liver cancer patients from their mutation profile: Benchmarking of mutation calling techniques, ***Biology Methods and Protocols***, 2022;bpac012.
- Pande A[#], Patiyal S[#], Lathwal A, Arora C, Kaur D, **Dhall A,** et al. Computing wide range of protein/peptide features from their sequence and structure. ***Journal of computational biology,*** 2022.
- Patiyal S, **Dhall A,** Raghava GPS**.** DBPred: A deep learning method for the prediction of DNA interacting residues in protein sequences. ***Brief Bioinform*** 2022**.**
- Patiyal S, **Dhall A,** Bajaj K, Sahu H, Raghava GPS**.** Prediction of RNA-interacting residues in a protein using CNN and evolutionary profile. (Under Communication)

# Table of Content

# List of Abbreviations

| Acronym | Full Form |
|---|---|
| **2D** | 2 dimensional |
| **3D** | 3 dimensional |
| **APC** | Antigen Presenting Cell |
| **AJCC** | American Joint Committee on Cancer |
| **AAC** | Amino Acid Composition |
| **Acc** | Accuracy |
| **ACR** | Autocorrelation |
| **APAAC** | Amphiphilic Pseudo Amino Acid Composition |
| **ATC** | Atomic Composition |
| **AUROC** | Area Under Receiver Operating Characteristic |
| **AIDS** | Acquired Immunodeficiency Syndrome |
| **BLAST** | Basic Local Alignment Search Tool |
| **BTC** | Bond Composition |
| **BRCA** | Breast invasive carcinoma |
| **BLCA** | Bladder urothelial carcinoma |
| **BCG** | Bacille Calmette-Guérin |
| **BA** | Binding Affinity |
| **BAM** | Binary Alignment and Map |
| **C-index** | Concordance index |
| **CeTD** | Composition enhanced-Transition Distribution |
| **CI** | Confidence interval |
| **Cox-PH** | Cox proportional hazard |
| **CSS** | Cascading Style Sheets |
| **CTD** | Conjoint Triad Distribution |
| **CV** | Cross-Validation |
| **CTLA-4** | Cytotoxic T-Lymphocyte–associated Antigen 4 |
| **CAR-T** | Chimeric antigen receptor T cells |
| **CRC** | Colorectal Cancer |
| **CLL** | Chronic Lymphocytic Leukemia |
| **CESC** | Cervical squamous cell carcinoma and Endocervical adenocarcinoma |
| **CGD** | chronic granulomatous disease |
| **DDOR** | Distance Distribution Of Residues |
| **DPC** | Di-Peptide Composition |
| **DT** | Decision Tree |
| **E-value** | Expect value |
| **ENT** | Elastic Net Regressor |
| **ET** | Extra Tree |

| | |
|---|---|
| **EGF** | Epidermal Growth Factor |
| **FDA** | Food and Drug Administration |
| **FN** | False Negative |
| **FP** | False Positive |
| **FP** | Finger Print |
| **FGF** | Fibroblast Growth Factor |
| **GBM** | Glioblastoma multiforme |
| **GNB** | Gaussian Naïve Bayes |
| **GDC** | Genome Data Commons |
| **GEO** | Gene Expression Omnibus |
| **HR** | Hazard ratio |
| **HTTP** | Hyper Text Transfer Protocol |
| **HLA** | Human Leukocyte Antigen |
| **HPV** | Human Papilloma Virus |
| **HNSC** | Head and Neck Squamous cell Carcinoma |
| **IEDB** | immune epitope database |
| **JAK** | Janus Kinase |
| **IMGT** | international ImMunoGeneTics project |
| **IGF** | insulin-like growth factor |
| **KICH** | kidney chromophobe |
| **KIRC** | Kidney renal clear cell carcinoma |
| **KIRP** | Kidney renal papillary cell carcinoma |
| **KM** | Kaplan-Meier |
| **KNN** | K Nearest Neighbors |
| **LAS** | Lasso Regressor |
| **LIHC** | Liver hepatocellular carcinoma |
| **LPC** | Ligand Protein Contacts |
| **LR** | Logistic Regression |
| **LR** | Linear Regression |
| **LASSO** | Least Absolute Shrinkage and Selection Operator |
| **LUAD** | Lung Adenocarcinoma |
| **LUSC** | Lung Squamous cell Carcinoma |
| **MAE** | Mean Absolute Error |
| **MCC** | Matthew's Correlation Coefficient |
| **MLP** | Multi-Layer Perceptron |
| **MHC** | Major histocompatibility complex |
| **MOS** | Mean Overall Survival |
| **NAG** | N-acetylglucosamine |
| **NB** | Naive Bayes |
| **NS** | Negative Samples |
| **NK** | Natural Killer |
| **NGS** | Next Generation Sequencing |

| | |
|---|---|
| **OS** | Overall Survival |
| **OV** | Ovarian serous cystadenocarcinoma |
| **PAAC** | Pseudo Amio Acid Composition |
| **PAAD** | Pancreatic Adenocarcinoma |
| **PRAD** | Prostate Adenocarcinoma |
| **PCB** | Physico-chemical Properties Binary Profile |
| **PCP** | Physico-chemical Properties |
| **PDB** | Protein Data Bank |
| **PHP** | Personal Home Page |
| **PS** | Positive Samples |
| **PSI-BLAST** | Position-Specific Iterated BLAST |
| **PD-1** | Programmed cell death protein 1 |
| **PD-L1** | Programmed Cell Death Ligand 1 |
| **PRI** | Property repeat information |
| **QSO** | Quasi Sequence Order |
| **RF** | Random Forest |
| **RFR** | Random Forest Regressor |
| **RID** | Ridge Regressor |
| **RMSE** | Root Mean Square Error |
| **RNA** | Ribose Nucleic Acid |
| **RRI** | Residue Repeats Information |
| **RNA** | Ribonucleic acid |
| **RS** | Risk Score |
| **RA** | Rheumatoid arthritis |
| **READ** | Rectum adenocarcinoma |
| **Sens** | Sensitivity |
| **SU** | Survival Unfavourable |
| **SF** | Survival Favourable |
| **SEP** | Shannon Entropy for Proteins |
| **SER** | Shannon Entropy for Residues |
| **SMILES** | Simplified Molecular Input Line Entry System |
| **SOCN** | Sequence Order Coupling Number |
| **SPC** | Shannon Entropy for Physico-chemical Properties |
| **Spec** | Specificity |
| **SQL** | Structured Query Language |
| **SVC** | Support Vector Classifiers |
| **SVM** | Support Vector Machine |
| **SVR** | Support Vector Regressor |
| **STAT3** | Signal transducer and activator of transcription 3 |
| **SKCM** | Skin Cutaneous Melanoma |
| **STAD** | Stomach Adenocarcinoma |
| **SARS-CoV-2** | Severe acute respiratory syndrome coronavirus 2 |

| | |
|---|---|
| **TCGA** | The Cancer Genome Atlas |
| **TCIA** | The Cancer Immunome Atlas |
| **TN** | True Negative |
| **TP** | True Positive |
| **TS** | Total Samples |
| **TSL** | Two Sample Logo |
| **TPC** | Tri-Peptide Composition |
| **TCR** | T-cell Receptor |
| **THCA** | Thyroid Carcinoma |
| **TNM** | tumor (T), nodes (N), and metastases (M) |
| **TAA** | Tumor Associated Antigen |
| **TAP** | transporter associated with antigen processing |
| **TSL** | Two Sample Logo |
| **UCEC** | Uterine Corpus Endometrial Carcinoma |
| **WGS** | Whole Genome Sequencing |
| **WES** | Whole Exome Sequencing |
| **XGB** | eXtreme Gradient Boosting |
| **CHOL** | Cholangiocarcinoma |

# List of genes and their description

| Gene | Description |
|---|---|
| **AP-1** | Activator Protein 1 |
| **APBB1IP** | Amyloid beta Precursor protein binding family B member 1 Interacting Protein |
| **Bcl-xL** | B-cell lymphoma-extra large |
| **C3** | Complement component 3 |
| **C6orf27** | Chromosome 6 open reading frame 47 |
| **CYP21A1P** | Cytochrome P450 Family 21 Subfamily A Member 1, Pseudogene |
| **CCL** | C-C Motif Chemokine Ligand |
| **CANX** | Calnexin |
| **c-Myc** | Cellular Myelocytomatosis Oncogene |
| **GM-CSF** | Granulocyte-Macrophage Colony-Stimulating Factor |
| **GAL** | Galanin And GMAP Prepropeptide |
| **GNRH1** | Gonadotropin Releasing Hormone 1 |
| **HER2** | Human Epidermal growth factor Receptor 2 |
| **HSPA1B** | Heat Shock Protein Family A (Hsp70) Member 1B |
| **IFN-α** | Interferon alpha |
| **IFN-γ** | Interferon gamma |
| **IL-6** | Interleukin 6 |
| **IL-2** | Interleukin 2 |
| **IL-12** | Interleukin 12 |
| **IL-15** | Interleukin 15 |
| **KLRC2** | Killer Cell Lectin Like Receptor C2 |
| **KIR2DL1** | Killer Cell Immunoglobulin Like Receptor, Two Ig Domains And Long Cytoplasmic Tail 1 |
| **LTB4R** | Leukotriene B4 Receptor |
| **MCL-1** | Myeloid Leukaemia 1 |
| **PSMC6** | Proteasome 26S Subunit, ATPase 6 |
| **PLD3** | Phospholipase D3 |
| **RFXAP** | Regulatory Factor X-Associated Protein |
| **TAP1** | Transporter 1, ATP Binding Cassette Subfamily B Member |
| **TGF** | Transforming Growth Factor |
| **TNF-α** | Tumor Necrosis Factor Alpha |
| **VEGF** | Vascular Endothelial Growth Factor A |

# List of Figures

# List of Tables

# CHAPTER 1

## INTRODUCTION

## 1.1 Overview of immune system

The immune system is a complex network of cells and proteins which provide protection from external invaders such as bacteria, viruses, and parasites that cause infection, sickness, and diseases (Nicholson, 2016). Our immune system evolved to protect the host from a universe of dangerous bacteria that are continually changing itself (Chaplin, 2010). It also aids in the elimination of harmful or allergenic chemicals that enter the body through mucosal surfaces (Belkaid & Hand, 2014; Demberg & Robert-Guroff, 2009). This complicated network of immune system is made up of organs, white blood cells, proteins (antibodies), lymphoid organs, humoral factors, cells, cytokines, and other chemicals (Nicholson, 2016). The immune system is essential to our survival. These specialised cells and immune system components help to protect the body against diseases and termed as immunity. The overall function of the immune system is to prevent or limit infection. When our immune system fails it causes severe infections, immunodeficiency, autoimmune diseases, hypersensitivity, and malignancies. It can also be described as a puzzling biological system that recognises and embraces what belongs to the self while also acknowledges and rejects what does not belong to the self (non-self). Innate, adaptive, and passive immunity are the three main categories of immune system (Parkin & Cohen, 2001).

Innate immunity is often referred to as non-specific immune response or intrinsic immunity. It is a natural immunity and act as a general defence that is present at birth. One such barrier is the skin, which prevents germs from entering the body. The immune system also recognises when to defend against outside intruders that could be harmful. It frequently describes a first-line of protection that is physical, chemical, and biological. Acute-phase proteins, neutrophils, monocytes, cytokines, and macrophages offer the host an immediate line of protection. Their actions are non-specific and non-inclusive (Jain et al., 2011). When innate immune system fails to eliminate the infectious agents, adaptive immunity plays a highly significant role. Adaptive or acquired immunity recognize the foreign antigens and activate specific immunologic effector pathways to eliminate the pathogen or infected cells (Dunkelberger & Song, 2010). An individual can develop adaptive immunity by being exposed to an illness or by receiving a vaccine immunization (Clem, 2011). It also develops the memory which aids to generate a specific immune response against the pathogens on their subsequent encounters. Lymphocytes, a type of white blood cell, are responsible for adaptive immune responses. Such reactions fall into two main categories: antibody reactions and cell-mediated immunological reactions. The major components of adaptive immune system or cell mediated immune reactions are carried by T cells and B lymphocytes. To connect innate and adaptive immune responses, antigen presenting cells (APCs) engage T cells (See Figure 1.1). These APCs directly affect T cell

differentiation by presenting bacterial, viral and tumorigenic fragments of peptides/antigens on their surface via major histocompatibility complex (MHC) or human leucocyte antigens (HLA) system.

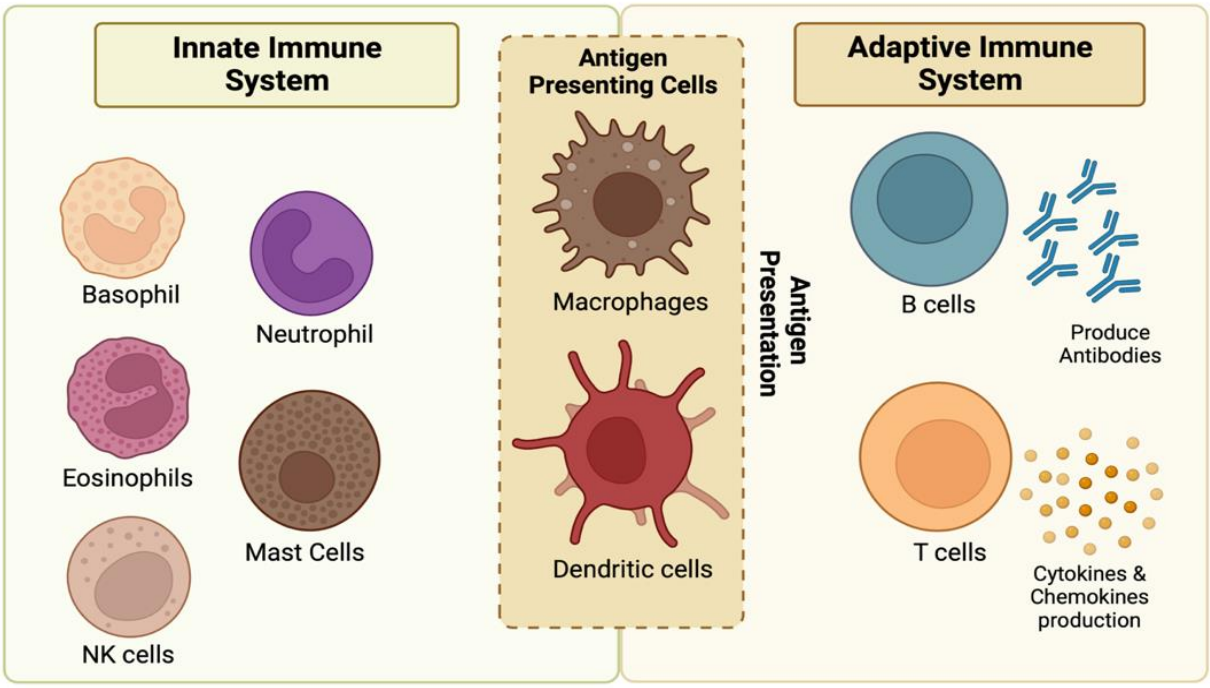



**Figure 1.1 Major cells involved in innate and adaptive immune system**

## *1.2 HLA system -antigen presentation mechanism*

The HLA system is the highly polymorphic genomic region located on human chromosome 6 (6p21.3) and majorly classified into Class-I (HLA-A, B, C, E, F, G) and Class-II (HLA-DP, DQ, DR, DM, DO) genes (Choo, 2007). The Class I and Class II HLA genes are the most polymorphic genes among 200 immune-related genes encoded by the major histocompatibility complex (MHC). These HLA genes produce proteins that act as histocompatibility antigens in transplantation and as important mediators of self-tolerance development and immune responses to infections. Moreover, class-III HLA region composed of 60 immune related genes (such as *TNF, C3, C4, C6orf27, CYP21A1P*, etc.) which encode proteins that play major role in the activation of hormonal synthesis, inflammation, and regulation of immunoregulatory molecules. Figure 1.2 shows the genetic location of major HLA genes on chromosome 6.

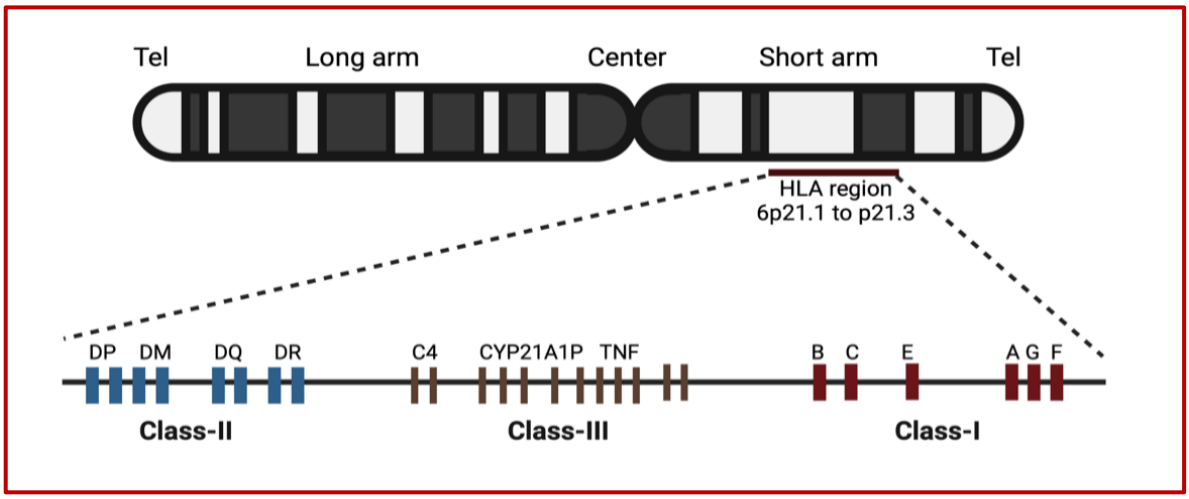



**Figure 1.2 Genetic map of human leucocyte antigen (HLA) region on chromosome 6**

According to IMGT/HLA, more than 34000 variant alleles for Class-I and Class-II HLA molecules are reported. The complete distribution of HLA-alleles and IMGT/HLA statistics is provided in Table 1.1. The major role of HLA-alleles is to bind with the antigenic peptides and present them to the cell surface. HLA-alleles have different binding affinities with antigenic peptides. Where, HLA-antigen complex interacts with T cell receptors and induces cytokines secretion which plays crucial immunoregulatory roles in activating/inhibiting the immune responses. Recent research suggests that the development of diseases including cancer and autoimmune disorders is directly linked to the mutations or changed expression of HLA molecules (including type 1 diabetes, celiac disease, and rheumatoid arteritis).

**Table 1.1: Number of classical and non-classical Class-I/II HLA alleles reported in IMGT/HLA (Robinson et al., 2020)**

| Class | Gene name | Number of HLA alleles |
|---|---|---|
| **HLA Class I (Classical)** | HLA-A | 7644 |
| | HLA-B | 9097 |
| | HLA-C | 7609 |
| **HLA Class I (Non-classical)** | HLA-E | 342 |
| | HLA-F | 59 |
| | HLA-G | 110 |
| **HLA Class II (Classical)** | HLA-DR | 8559 |
| | HLA-DQ | 2896 |
| | HLA-DP | 2728 |
| **HLA Class II (Non-classical)** | HLA-DM | 163 |
| | HLA-DO | 152 |

In addition, studies also reveal that the presence/absence of certain HLA molecules may associate with the adverse drug hypersensitive reactions and also increases the risk factors in cancer patients (Alfirevic A, 2010 Dec 23). With the knowledge of accurate HLA typing, clinicians can design personalized vaccines and immunotherapy-based prognostic biomarkers against cancer (Dhall et al., 2020; Xu et al., 2021). In clinical practices, HLA typing could be used as predictive or diagnostic tests for the drug induced hypersensitivity (Rive et al., 2013). Moreover, non-classical HLA-G and HLA-E molecules act as essential immune checkpoint molecules which mediates the NK-cell lysis, cytotoxicity, cytokine production, tumor proliferations (Cao et al., 2020). Of note, in order to develop better immunotherapeutic candidate against cancer and diseases it is essential to understand the role of HLA-alleles (Sabbatino et al., 2020). In addition, to design novel immunotherapies or subunit vaccine candidates, it is crucial to accurately identify the HLA-peptides or antigen binding regions (Zhao et al., 2013).

### *1.2.1 HLA Class-I presentation*

HLA Class-I molecules are made up of two chains one is polymorphic heavy chain and other is β2-microglobulin chain. Class-I HLA are assembled in endoplasmic reticulum (ER) and expressed in all the nucleated cells and follows endogenous or intercellular mechanism. As shown in the Figure 1.3, the antigenic protein degraded into small antigenic peptides, these peptides are then translocated from cytoplasm to ER via TAP protein and further bound to HLA class-I molecules. HLAs deliver short peptides to the cell surface and interacted with CD8+ (cytotoxic) T cells. These antigenic regions when come in contact with the T-cell receptors it activates several immune responses and induces the production of several cytokines such as IFN-gamma, TNF-alpha, IL6, IL-12, IL-4 etc. (Y. Zhang et al., 2020).

### *1.2.2 HLA Class-II presentation*

HLA Class-II genes are majorly expressed by antigen presenting cells (including dendritic cells, macrophages and B cells). Class-II molecules assembled in ER and made up of α- and β- chains. HLA class-II molecules bound to exogenous peptides which were degraded in the endocytic pathway. They present the antigenic peptides on the cell surface and interacted with CD4+ (T-helper cells). Which further activate B-cells in order to stimulate antibody production against specific antigen. Moreover, T-helper cells generate memory B-cells, plasma cells and increases the production of cytokines in order to kill the pathogen or cancerous cells (See Figure 1.3).

## *1.3 Immunity against cancer*

Our immune system is able to identify a malignant cell as aberrant and eliminate it before it spreads or replicates. In this case the malignant or cancerous cells entirely eradicate and the disease never manifests. Tumor associate antigens (TAAs) are tumor specific peptides presented by HLA molecules and are recognized by our immune system (Restifo et al., 1994; Z. Zhang et al., 2021). Although all cells have antigens on their surfaces, the immune system typically does not respond to a person's own cells. The new antigens or neoantigens that are unfamiliar to the immune system emerge on the surface of cancer cells. These neoantigens, also known as tumors antigens, recognized as foreign peptides by the immune system (Yarchoan et al., 2017). By using this technique, the body eliminates aberrant cells and frequently stops the development of cancerous cells. As shown in the Figure 1.3, mutated peptides or tumor specific antigens are recognized by cytotoxic T cells and helper T cells which further secretes a number of cytokines and generate specific immune responses, in order to kill the malignant cells. These tumor specific antigens can act as tumor markers and can be used to made cancer vaccines. For instance; in the case of melanoma, breast cancer, ovarian cancer, liver and prostate cancer tumor antigens are identified (Feola et al., 2020). The antigen vaccines stimulate the immune response and can be used for the treatment of certain type of cancers (Tagliamonte et al., 2014). Nowadays, due to advancements in technology tumor specific antigens (tumor markers) can be detected in blood tests (Holdenrieder et al., 2016).

However, certain type of cancers are more likely to advance and grow on faster rate in persons with weakened immune systems, such as patient suffering from AIDS (Prakash et al., 2002). Moreover, tumor cells may not present antigens on their cell surface or loss the expression of HLA-I molecules or inhibiting T-cells via producing immunosuppressive chemicals. The onset and progression of cancer may be influenced by immune system disorders such as immunological deficiency and immune suppression (Gonzalez et al., 2018). Patients with immunodeficiency illnesses as well as transplant recipients who have received long-term immunosuppressive medications are more likely to develop specific types of cancer (Gallagher et al., 2010). For instance, individuals having AIDS (acquired immunodeficiency syndrome) are more likely to get tumors like Kaposi sarcoma, which are linked to viruses (Angeletti et al., 2008). In older age, when some immune responses deteriorate, the incidence of cancer also rises significantly. Age-related genetic changes associated to the cancer also accumulate, thus, immune responses may not be the main cause of cancer development in the elderly (Hong et al., 2019; Laconi et al., 2020). In this situation targeted therapy or immunotherapy given to cancer patients to immunize patients against specific type of cancer.

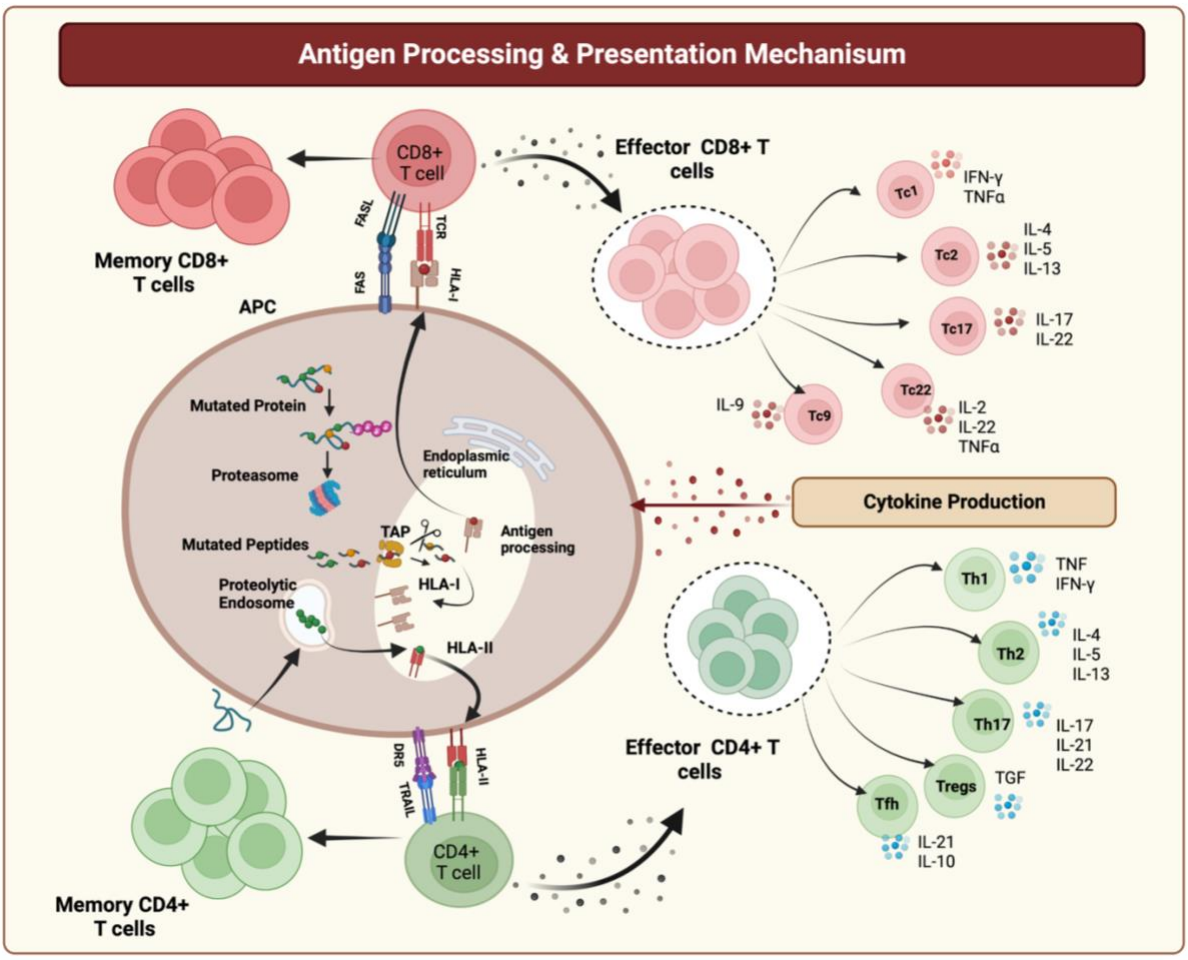



**Figure 1.3 Illustration of antigen presentation and processing mechanism**

## *1.4 Cancer immunotherapy*

Cancer immunotherapy acts as a novel pillar for cancer care and shows significantly increased patient's survival and quality of life as compared to traditional treatment regimens such as chemotherapy, radiation, and surgery (Esfahani et al., 2020). Recently, a number of immunotherapies are available to treat cancer patients as shown in Figure 1.4. Adoptive cell therapy are HLA-dependent immunotherapies and are mainly focused on CD8+ T cells, like tumor-infiltrating lymphocytes (TILs) therapy and TCR-engineered T cells (TCR-Ts) therapy. Immune checkpoint inhibitor therapies such as, CTLA-4 inhibitor (Ipilimumab), PD-1 inhibitors (Pembrolizumab and Nivolumab), PD-L1 inhibitors (Atezolizumab, Avelumab, Durvalumab) are used to treat advance-stage cancers (Wu et al., 2012). In addition, some of the immune checkpoint modulators (*CD70, CD27, CD40, CD47,* and *CD73*) and antagonist antibodies are under clinical trials (Wang et al., 2022).

Recombinant cytokine products are also used for cancer immunotherapies for instance interferon alpha (*IFN-alpha*), proleukin, and interleukin-2 (*IL-2*) for the treatment of hairy Cell leukemia, malignant melanoma, follicular lymphoma, AIDS-Related Kaposi's Sarcoma, metastatic renal cell carcinoma, and metastatic melanoma (Waldmann, 2018). Food and Drug Administration (FDA) has approved vaccines to prevent cancer for example: HPV vaccines protect against human papillomavirus (Thomas, 2016) and can be used to prevent cancers like cervical, vaginal, vulvar, and anal cancer. Sipuleucel-T (Provenge) is used to treat the metastatic prostate cancer (Anassi & Ndefo, 2011) and Bacillus Calmette-Guérin (BCG) vaccine is used for the treatment of early-stage bladder cancer (Guallar-Garrido & Julian, 2020).

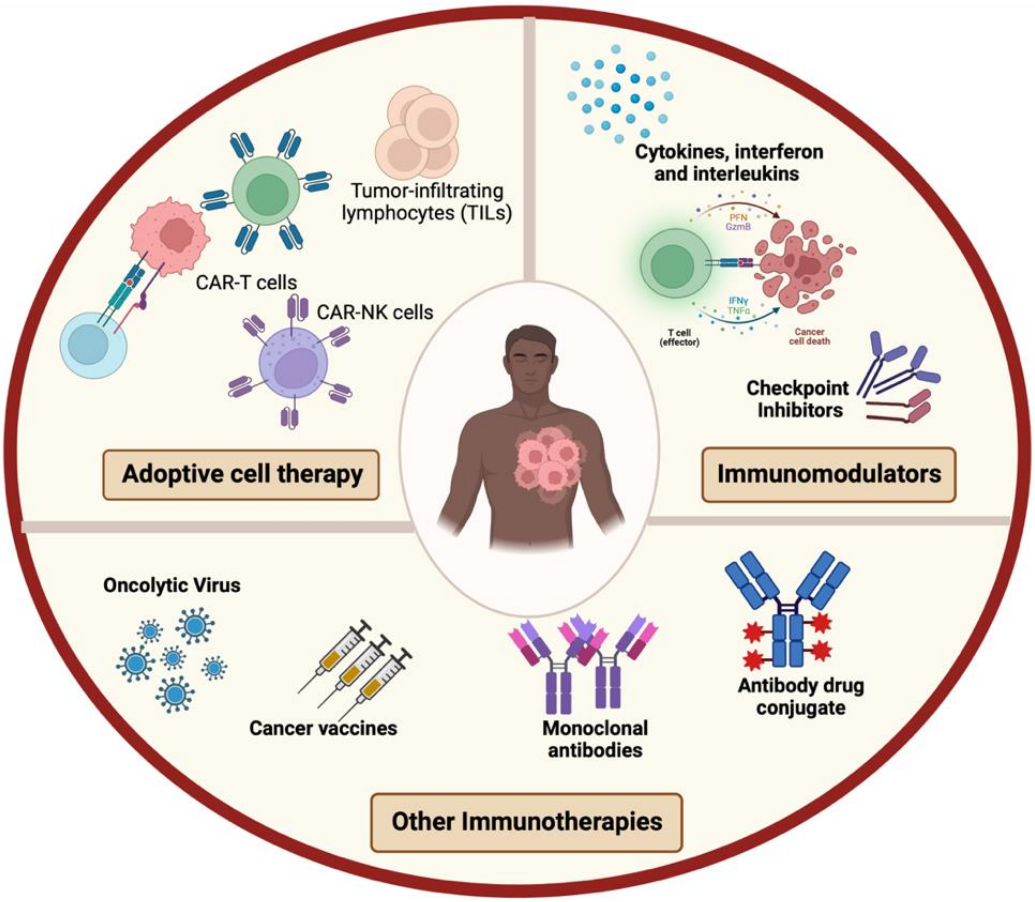



**Figure 1.4 Types of immunotherapies used for cancer treatment**

Subunit or peptide-based vaccines are also used nowadays for the treatment of cancer. The aim of peptide-based anticancer vaccines is to stimulate immune response against the tumor specific antigens. Number of pre-clinical and clinical trials have been initiated to check the efficacy of subunit or peptide-based vaccines (Abd-Aziz & Poh, 2022; Slingluff, 2011). TAA-derived peptides, personalized peptide vaccine, HER2, W3, E6/E7, neoantigens, synthetic long peptide (SLPs) (Chen, Yang, et al., 2020) are

under clinical investigation and can be used for the treatment of bladder carcinoma, breast carcinoma, gastric carcinoma, glioblastoma, and HPV+ tumors (Bezu et al., 2018).

## 1.5 Proposal's origin

Tumors are part of a complex network of tissues, cells, and chemical messengers, including immune cells, stroma, blood, lymphatic, and epithelial cells, as well as cytokines and chemokines. Tumor antigens or neoantigens are used by immune system to distinguish between tumor cells and normal cells. These tumor specific antigens are produced by the extensive genetic changes that are specific to tumor. Neoantigens are significant because they trigger the T cell response, a crucial line of defence against tumorigenesis, via the Human Leucocyte antigen molecules. In contrast to this, tumor cells have created ways to get beyond host immunity in their never-ending struggle for survival and growth. In order to fight against cancer, several initiatives and therapies have been made in the past. These traditional treatment regimens use surgery, radiations, chemotherapy, medications to stop the progression of tumor. However, these conventional treatment causes adverse effects on the health and survival of the cancer patients. On the other side, cancer immunotherapy or biological therapy shows promising outcome and improves the survival of the cancer patients. Adaptive immune system components including (HLA molecules, neoantigens and cytokines) plays important role in designing patient specific immunotherapy. The major step shared by immunotherapies require T-cells to recognize specific antigenic peptides presented by HLA molecules on the infected cell surface. HLAs are essential components of the immune system that stimulate immune cells to provide protection and defence against diseases including cancer. So, it is essential to understand the impact of HLA-alleles, HLA-binders and cytokines. HLA-based biomarkers can be utilized by the researchers to design personalized therapy and to predict the survival and risk in the cancer patients. Furthermore, the peptide based vaccines or subunit vaccines are crucial immunotherapeutic candidate which can elicit an appropriate immune response against cancer. Cytokine inducing peptides and cancer growth blockers (inhibitors) could be utilized in the designing of immunotherapy or subunit against cancer.

## 1.6 Objective of thesis

In the present study, we mainly focus on the components of adaptive immune system. Where, we tried to understand the impact of HLA-alleles, HLA-binding peptides, cytokines and chemokines in the overall survival of the cancer patients. The study is primarily divided into four major categories (i) Prognostic biomarkers for cancer (ii) Non-classical HLA-binder prediction (iii) Designing of cytokine inducing peptides (iv) Inhibition of IL6/STAT3 pathway. For this, we have created a computational

resource (CancerHLA-I) and risk estimation tool (SKCMhrp) for the analysis and prediction of survival rate of cancer patients using the HLA-typing and clinical information. We have also developed a computational tool named ($HLA_{nc}$Pred) for the prediction of non-classical HLA binding peptides. Next, we have created user-friendly tools for the prediction, scanning and designing of IL6 (IL6Pred), TNF-α (TNFepitope) and IFN-γ (IFNepitope2) inducing peptides. In addition, we have generated an in-silico method for the prediction of IL6 mediated STAT3 inhibitors using chemical descriptors. All the brief information is depicted in the Figure 1.5.

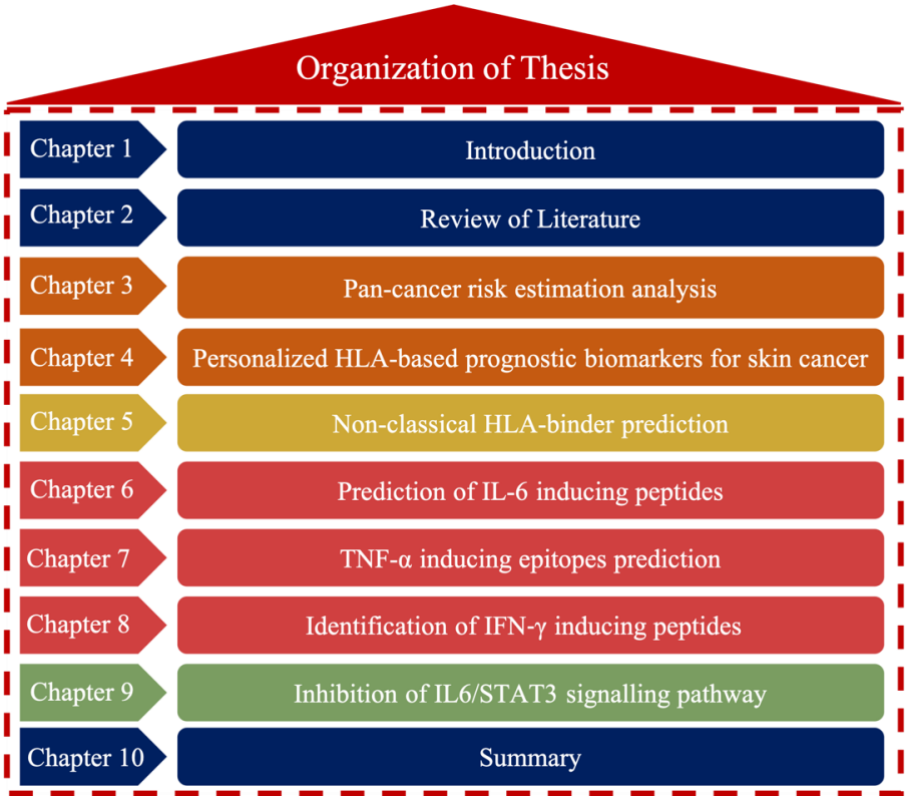



**Figure 1.5: Overall organization of thesis in different chapters**

## *1.7 Organization of chapters*

This thesis is divided into ten chapters and information regarding each chapter is given below:

***Chapter 1:*** In this section, the background information of immune system and its various components is provided. Moreover, the importance of antigen processing and presenting mechanism via HLAs, neoantigens and cytokines in cancer is briefly discussed. Finally, we focused on understanding the mechanism of the immune system to fight against cancer. In conclusion, this chapter emphasis on the importance of immune system components in the development of immunotherapy or subunit vaccine candidates against various type of cancers.

***Chapter 2:*** This chapter is focused on the review of literature on the adaptive immune systems, use of tumor specific antigens, HLAs and cytokines in the cancer immunotherapy. Moreover, this chapter summarize the available tools for HLA-typing, HLA-binder prediction, and cytokine inducing peptides identification methods. In a nut shell, this chapter explains why this study was conducted.

***Chapter 3:*** This chapter is focused on the first objective of the thesis, which is development of a computational resource named "CancerHLA-I" for the risk estimation analysis. This study provided prognostic biomarkers based on HLA-alleles, cancer specific neoantigens and cytokines for 20 types of cancers. The patient-specific HLA-typing and survival datasets for 20 types of cancers is obtained from the TCIA and TCGA repositories. Moreover, expression profiles of cancer patients are used for the identification of cytokines based prognostic biomarkers. In conclusion, the novel HLA-based prognostic biomarkers could be used for designing the cancer immunotherapy.

***Chapter 4:*** This chapter is dedicated for the development of risk estimation prediction method using the TCGA-SKCM dataset. In this objective, we investigate the role of Class-I and Class-II HLA-alleles and clinical characteristics on the overall survival of skin cutaneous melanoma patients. Moreover, machine learning based survival prediction method is generated based on HLA-alleles, patient demographics, and clinical characteristics.

***Chapter 5:*** This chapter is about the non-classical HLA (HLA-G and HLA-E) binding peptide prediction. In the past two decades, a number of HLA-binder prediction methods have been developed. However, there was no specific method for the prediction of non-classical HLA alleles. The prediction models developed using binders for the non-classical HLA-alleles (HLA-G*01:01, HLA-G*01:03, HLA-G*01:04, HLA-E*01:01, and HLA-E*01:03). The experimentally validated datasets obtained from IEDB resource. $HLA_{nc}$Pred, a bioinformatics tool was developed using the highly accurate prediction models.

***Chapter 6:*** This chapter explains the role of pro-inflammatory cytokine interleukin 6 (IL6) in the cancer and other diseases. In this objective, we attempted to create a computational tool for the prediction, scanning and designing of IL6 inducing peptides. The webserver named "IL6Pred" developed for the researcher for predicting IL6 inducing regions while designing the subunit vaccine or peptide-based therapeutics. Here, the dataset is obtained from IEDB and prediction models were

developed using composition based features. The best models were incorporated in the webserver and standalone package.

***Chapter 7:*** This chapter is about the cytokine inducing peptide prediction and designing. Here, we have focused on the most important inflammatory cytokine tumor necrosis factor alpha (TNF-α). We have developed a host-specific prediction method for the identification of TNF-α inducing peptides using primary information. The experimentally validated TNF-α inducing and non-inducing epitopes were obtained from the IEDB resource. Moreover, various classifiers were used to train and evaluate the models using training and independent dataset. Finally, a webserver named "TNFepitope" was developed to serve the scientific community.

***Chapter 8:*** This chapter is also provide a computational tool for the designing and prediction of interferon-gamma (IFN-γ) inducing peptides. IFN-γ is important immunoregulatory cytokine and causes anti-allergic, anti-tumorigenic immune responses. In this study, we have created "IFNepitope2.0" for the prediction, scanning and designing of IFN-γ inducing epitopes for human and mouse hosts. We anticipate this method could be used by experimental biologist in the designing the cytokine based immunotherapy or subunit vaccine candidate.

***Chapter 9:*** The aim of this chapter is to develop an pharmacological tool for the prediction of chemical molecules and drugs which can inhibit the activation of STAT3 signaling pathway. The STAT3 inhibitors and non-inhibitor molecules obtained from PubChem repository. PaDEL software was used for the generation of chemical molecule descriptors and machine learning algorithms were implemented for classification of STAT3 inhibitors and non-inhibitors. Finally, a computer-aided tool named "STAT3In" provided to the scientific community for the prediction of STAT3 inhibitors which can be used for designing the anti-cancer therapies.

***Chapter 10:*** This chapter gives the overall summary of thesis, and quick overview of all the studies conducted in the area of immunotherapy and subunit vaccine designing against cancer development.

# CHAPTER 2

## REVIEW OF LITERATURE

## *2.1 Overview of adaptive immune system*

When innate immunity fails to eradicate the infectious pathogens and the infection becomes establishes, adaptive immunity emerges. The recognition of particular "non-self" antigens in the presence of "self" antigens, the development of pathogen-specific immunologic effector pathways that can kill the particular pathogens, and the creation of an immunologic memory that can quickly eradicate a particular pathogen are the three main functions of the adaptive immune response (Marshall et al., 2018). Moreover, acquired immunity is of two types natural and artificial. In natural acquired immunity the antigen enters the body naturally, whereas in artificial acquired immunity the antigens are introduced in the vaccines, antibodies and immune serum are generated against them (Clem, 2011). Lymphocytes (T-cells and B-cells) specifically recognize the foreign antigens and generate response against them. The major attributes of adaptive immunity are specificity, diversity, specialization, memory and self/non-self-recognition. Where T lymphocytes are activated by antigen presenting cells (APCs), and B cells are among the cells that make up the adaptive immune system. On the surface of APCs, antigenic peptides are presented via HLA class-I and II molecules. HLA-peptide complex interacts with cytotoxic T-cell or helper T-cells and activate the immune responses (Hewitt, 2003; Wieczorek et al., 2017). The major functions of adaptive immune responses are the elimination of specific pathogens or pathogen-infected cells and development of immunological memory (i.e., memory B cells and memory T cells) also known as immunization (Dunkelberger & Song, 2010).

## *2.2 Role of adaptive immunity in cancer*

In cancer cells a wide range of genetic alterations generate mutated peptides also known as tumor specific peptides or neoantigens. These tumor antigens enable the immune system to recognize and differentiate normal cell and cancerous cell. Tumor specific antigens are essential to trigger the immune response, as they are presented on the cell surface via HLA-molecules and recognized by T-cells. The different types of T cells perform specific functions. Helper T (Th) cells and cytotoxic T cells are the two main subtypes of T cells (Zamora et al., 2018). Th cells have an important role as activators of other cells, such as cytotoxic T cells and B cells (Waldman et al., 2020). Killer cells known as cytotoxic T cells target cancerous or malignant cells. However, natural killer cells able to recognize and destroy cancerous cells without looking for HLA receptors (Paul & Lal, 2017). Cytotoxic T cells use perforins, granzymes, proteases, or even FAS ligand signaling to start the caspase cascade and cause the cancerous cell to undergo apoptosis (Chowdhury & Lieberman, 2008; Prager & Watzl, 2019). Of note, tumor specific antigens, HLA molecules and T cell response are significant line of defense against cancer.

## *2.3 Role of HLA and neoantigens in cancer*

Human leukocyte antigen (HLA) is the most polymorphic region of human genome and composed of several genes which play major roles in immune regulations (Choo, 2007; Crux & Elahi, 2017). Due to high polymorphism, HLA genes are encoded by thousands of HLA-alleles reported in IMGT/HLA database (Robinson et al., 2020). It is essential to check the type of HLA in order to identify the immune response because the tumor specific antigens bind to specific HLA-alleles (Crux & Elahi, 2017; Mosaad, 2015). T cell receptor recognize HLA-peptide complex which further activate T cells and trigger the production of cytokines in order to kill the cancer cells (He et al., 2019). However, under some conditions tumor cells may escape the immune attack due to down regulation or mutations in HLA molecules, limited tumor specific peptides binding to HLA and over expression of non-classical HLA genes (Garrido & Aptsiauri, 2019). In order to overcome the HLA downregulation, several immunotherapies are available such as chimeric antigen receptor CAR-T cell therapy, NK cell therapy and CD4+ T cell based immunotherapy (Liu et al., 2021).

In the recent studies, some of the neoepitopes or tumor specific peptides are tested in clinical trials and can be uses in immunotherapy (Hutchison & Pritchard, 2018). Tumor specific peptides restricted to specific HLA-alleles such as HLA-A*02:01, HLA-A*24:02, HLA-A*02, HLA-A*11:01, HLA-A*02:642 and activate the immune system (Boucherma et al., 2013). These studies reveled the importance of HLA-alleles and restricted peptides while designing immunotherapy against specific type of cancer. Researcher also rebelled that the HLA-alleles may impact the survival of cancer patients, for instance in melanoma patients the presence of HLA-B∗55 and HLA-A∗01 increases the survival rate while HLA-B∗50 and HLA-DRB1∗12 significantly reduces the survival rate (Dhall et al., 2020). In addition, HLA-DRB1∗07 shows negative correlation with the survival of lung cancer, cervical cancer, and breast cancer patients. HLA class-II expression improves the survival of leukemia and lymphoma cancer patients (Liu et al., 2015).

## *2.4 Role of cytokines in cancer*

Cytokines are polypeptide or glycoproteins that play pro-inflammatory and anti-inflammatory roles in the immune system. Cytokines trigger intra-cellular signaling and can modulate proliferation, differentiation by activating or suppressing cell functions. Pro-inflammatory cytokines such as interleukin 6 (IL6), tumor necrosis factor alpha (TNF-α) and interferon gamma (IFN-γ) play significant roles in the induction of acute phage responses, inflammation, innate and adaptive immune activation (Cavalcanti et al., 2012; Kany et al., 2019). IFN-γ is primarily secreted by natural killer (NK) and

activated T cells, and it can facilitate the activation of macrophages, mediate immunity against bacteria and viruses, improve antigen presentation, orchestrate the activation of the innate immune system, and regulate lymphocyte-endothelium interaction. The dysregulation in the expression levels or overexpression of interleukin 6 (IL6) and tumor necrosis factor alpha (TNF-α) cytokines increases the severity of several diseases including sepsis, diabetes, rheumatoid arthritis and cancer (Hirano, 2021; Navarro-Gonzalez & Mora-Fernandez, 2008; Stenvinkel et al., 2005).

Most importantly, the cytokine storm syndrome in COVID-19 patients is significantly associated with the elevated levels of IL6 and TNF-α (Kountouri et al., 2021; Remy et al., 2020). Recent, studies showed that, cytokines can control the tumor growth by stimulating anti-proliferative and pro-apoptotic activities. Till now, IL-2 and IFN-α cytokines which are approved by FDA for clinical usage and for the treatment of advanced renal cell carcinoma, metastatic melanoma, hairy cell leukaemia, follicular non-Hodgkin lymphoma, melanoma and AIDS-related Kaposi's sarcoma. However, a number of cytokines such as IL-12, IL-15, granulocyte-macrophage colony-stimulating factor (GM-CSF) and IL-10 are under clinical investigation (Conlon et al., 2019) (see Table 2.1). Recent studies revealed that cytokine interleukin 6 (IL6) plays major role in tumor development. Overexpression of IL6R and gp130 activate JAK/STAT3 pathway, which further induces pro-tumor activities. Moreover, the combination of IL6 and TGF-β induces the proliferation of tumor cells by inducing Th17 cells. Elevated levels of IL6 acts as negative prognostic marker for patients survival (Chonov et al., 2019). Therefore, anti-IL6 targeted therapy is given to the cancer patients with multiple myeloma or metastatic renal cell carcinoma. STAT3 hyperactivation in cancer cells plays a major role as it increases the production of immunosuppressive factors, tumor proliferation, angiogenesis and metastasis (Johnson et al., 2018). Inhibiting STAT3 in cancer immunotherapy is extensively investigated; some of the drugs are under pre-clinical and clinical trials for the inhibition of STAT3. BBI608, celecoxib and pyrimethamine are the FDA-approved drugs are under phase-II/III clinical trial for the treatment of advanced malignancies, CRC, CLL, small lymphocytic lymphoma cancer (S. Zou et al., 2020).

**Table 2.1: List of cytokines used for the treatment of different type of cancers**

| Cytokine | Cancer type (References) |
|---|---|
| **IFN-α** | Metastatic renal cell Carcinoma (Rini et al., 2010), AIDS-related Kaposi's sarcoma (Rokx et al., 2013), Human T cell lymphotropic-1 associated adult T cell leukaemia (Bazarbachi et al., 2010) |
| **IFN alfa-2b** | Stage III or IV high-risk melanoma (Tarhini et al., 2012) |
| **IFN-γ** | Malignant melanoma (Gollob et al., 2000) |

| | |
|---|---|
| **GM-CSF** | Stage III/IV melanoma (Kaufman et al., 2014) |
| **IL-12** | Hodgkin's and non-Hodgkin's lymphoma (Younes et al., 2004) |
| **IL-2** | Metastatic renal cell cancer (Klapper et al., 2008), Metastatic Melanoma (Marabondo & Kaufman, 2017) |
| **IL-21** | Renal cell cancer (Curti, 2006), Metastatic colorectal cancer (Steele et al., 2012) |
| **IL-15** | Metastatic malignant melanoma and metastatic cancer (Chen et al., 2012) |

## *2.5 Immune-related prognostic biomarkers in cancer*

A prognostic biomarker used to identify the likelihood of cancer outcome such as disease recurrence, disease progression or death (Kerr & Yang, 2021). The availability of high throughput techniques such as microarrays and RNA-seq produces huge amount of gene expression, methylation and mutation data (Kukurba & Montgomery, 2015). With the utility of genomic dataset, clinical characteristics and survival information of cancer patients one can identify the prognostic markers (Mehta et al., 2010). In the past, a number of studies reported the prognostic biomarkers based on expression profiles, mutation profile and epigenetic profiles of cancer patients (Herceg & Hainaut, 2007). Guo et al., identified six immune-related genes *CD8A, KIR2DL1, CD79A, APBB1IP, GAL,* and *PLD3* that play significant impact on the overall survival of osteosarcoma patients (Guo et al., 2021). In addition, high expression of *GNRH1* and *LTB4R* immune genes reduces the survival of clear cell renal cell carcinoma patients (Wu et al., 2021).

Researchers also identify that, higher expression levels of *CANX, HSPA1B, KLRC2, PSMC6, RFXAP,* and *TAP1* immune genes reduces the survival rate of lower grade glioma patients (M. Zhang et al., 2020). A recent study reported that the higher expression of HLA-DRA gene is positively correlated with the survival of lower grade glioma patients; and HLA-G higher expression act as negative prognostic marker in colorectal cancer/colon and rectal cancer (CRC), colon cancer (COAD), rectal cancer (RC), stomach cancer/gastric cancer (GC), esophageal cancer (ESCC), pancreatic cancer/pancreatic adenocarcinoma (PC), liver cancer/hepatocellular carcinoma (HCC), small bowel cancer (SBC), gastrointestinal cancer (GI) patients (Peng et al., 2021). In addition, the high protein expression levels of HLA-DQB1 and LIMCH1 genes are significantly associated with the poor survival of cervical cancer patients (Halle et al., 2021).

## *2.6 Available immunological resources*

In the past, a number of repositories have been developed to store the huge amount of immunological and experimental data. For example, MHCBN (Bhasin et al., 2003) is one of the oldest repository and

contains MHC-binding and non-binding epitopes. Designing immunotherapy candidates for the treatment of cancer and other disorders can be done using the immunological data from the IEDB (R et al., 2019). IEDB provides experimentally validated peptides/epitopes of T-cells, B-cells, MHC, cytokines etc. The IPD-IMGT/HLA (Robinson et al., 2016) database was created to store 45 HLA coding genes of the human genome and more than 25000 experimentally confirmed HLA allele sequences. A repository named VDJdb (Shugay et al., 2018) was created to gather antigen-specific TCR sequences. Additionally, it visualises antigen-specific TCR sequence patterns and annotates data on TCR repertoire.

The most important targets in the detection and therapy of different carcinomas are tumor-associated antigens (TAAs). They are also used in the creation of immunotherapies for the treatment of various malignancies. Moreover, few tumour antigen databases were created in the past for the treatment of many disorders, including cancer. One of the most effective repositories, the Human Possible Tumor-Associated Antigen Database (HPtaa) (Wang et al., 2006), contains 3518 potential TAAs that can target different types of malignancies. The TANTIGEN knowledge base has also been updated with TANTIGEN 2.0 (G. Zhang et al., 2021). It is a comprehensive data repository for neoepitopes and tumor-associated T cell antigens. It has around 1500 T cell epitopes, 4296 antigen variations, and 403 distinct tumour antigens. Immunoglobins or antibodies are still another crucial component in fighting cancer. The immune responses to immunotherapy depend heavily on the tumor-epitope-binding immunoglobins. CIG-DB, which contains 2081 genes for immunoglobulins related with cancer and T-cell receptors, is the most important public resource for immunoglobulins. Table 2.2 provide complete list of available immunological resources/databases which can be used for designing immunotherapy or subunit vaccines.

**Table 2.2: List of the immunological databases with their brief description and weblink**

| Name & Description | Resource Link |
|---|---|
| **MHCBN:** A resource of MHC-binding and non-binding peptides (Bhasin et al., 2003) | https://webs.iiitd.edu.in/raghava/mhcbn/ |
| **JenPep 2.0:** Immunobiology and vaccinology database (McSparron et al., 2003) | http://www.jenner.ac.uk/JenPep |
| **Bcipep:** A repository of B-Cell epitopes (Saha & Raghava, 2006) | https://webs.iiitd.edu.in/raghava/bcipep/ |
| **Epitome:** Resource of proteins with structurally inferred antigenic epitopes (Schlessinger et al., 2006) | https://www.rostlab.org/services/epitome/ |
| **SuperHapten:** Immunogenic compound database (Wang et al., 2017) | https://bioinformatics.charite.de/superhapten/ |
| **Ctdatabase:** A resource of cancer specific testis antigens (Almeida et al., 2009) | http://www.cta.lncc.br |

| | |
|---|---|
| **AntigenDB:** Experimentally validated antigens database (Ansari et al., 2010) | https://webs.iiitd.edu.in/raghava/antigendb/ |
| **Protegen:** A database for protective antigens (Yang et al., 2011) | http://www.violinet.org/protegen/ |
| **AgAbDb:** A database of antigen-antibody interactions (Kulkarni-Kale et al., 2014) | http://bioinfo.net.in/AgAbDb.htm |
| **VDJdb:** A repository of T-cell receptor sequences (Shugay et al., 2018) | https://vdjdb.cdr3.net |
| **IEDB-AR:** Immune epitope database—analysis resource in 2019 (Dhanda et al., 2019) | http://tools.iedb.org/ |
| **IPD-IMGT/HLA:** A database of human leukocyte antigens sequences (Robinson et al., 2020) | https://www.ebi.ac.uk/ipd/imgt/hla/ |
| **TANTIGEN 2.0:** A database of tumor T cell antigens & epitopes (G. Zhang et al., 2021) | http://projects.met-hilab.org/tadb/ |

## *2.7 Cancer associated repositories*

Genome profiles can help in the pre-screening of patients who will respond to the immunotherapies with the greatest potential for the benefit and the fewest possible negative effects. According to the National Cancer Institute, "Genome profiling is a technique for deciphering genetic information about a single person or cell type as well as how their genes interact with one another and with the environment". Existing sequencing technologies including WGS, WES, RNA-seq, and ChIP-seq have an inverse relationship between cost and accuracy due to the rapid advancement of technology. These methods provide single-cell RNA seq data as well as geographical information. It aids in the development of effective immunotherapies by assisting researchers in better comprehending the diseases.

Today's genomic profile data can be a gold mine for finding predictive and diagnostic markers. In the past, several databases with genomic profiles have been developed at an exponential rate. For example; The Cancer Genome Atlas (TCGA (Zhu et al., 2014)) is the most complete, effective, and commonly used tool for cancer genomics. TCGA project generated, examined, and disseminated clinical, microsatellite instability, miRNA, mRNA, and protein expression data on more than 20,000 samples spanning 33 different cancer types. Genomic Data Commons (GDC (Jensen et al., 2017)) data portal is the most important portal or site to obtain the multi-omics data connected to cancer. Data from about 68 projects, including TCGA, are included in it. Gene Expression Omnibus (GEO), a significant database that openly distributes high-throughput gene expression data and functional genomics data to public. GEO includes more than 4000 datasets and information for more than 1.5 lakh studies involving over 45 lakh samples. Additionally, GEO offers the tools for data analysis and visualisation. We provide list of all major cancer associated repositories in Table 2.3.

**Table 2.3: List of cancer associated resources with description and weblink**

| Name & Description | Resource Link |
|---|---|
| **dbGap:** A repository of genotype and phenotype (Mailman et al., 2007) | https://www.ncbi.nlm.nih.gov/gap/ |
| **caBIG:** Cancer Biomedical Informatics Grid (ca, 2007) | https://biospecimens.cancer.gov/caBigTools.asp |
| **SRA:** High-throughput sequencing reads database (Leinonen et al., 2011) | http://www.ncbi.nlm.nih.gov/Traces/sra |
| **CCLE:** Genomic profiles of human cancer cell lines (Barretina et al., 2012) | https://sites.broadinstitute.org/ccle/ |
| **cBioPortal:** Exploration of cancer genomics data (Cerami et al., 2012) | https://www.cbioportal.org/ |
| **Survexpress:** Cancer gene expression and survival analysis database (Aguirre-Gamboa et al., 2013) | http://bioinformatica.mty.itesm.mx:8080/Biomatec/SurvivaX.jsp |
| **GTEx:** A database for tissue-specific gene expression (Consortium, 2013) | https://gtexportal.org/home/ |
| **TCGA:** A comprehensive resource on cancer (Tomczak et al., 2015) | https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga |
| **GEO:** Gene expression data sets (Clough & Barrett, 2016) | http://www.ncbi.nlm.nih.gov/geo/ |
| **GDC Data portal:** Multi-omics and clinical database of cancer patients (Jensen et al., 2017) | https://portal.gdc.cancer.gov/ |
| **TCIA:** Immunogenomic analyses repository of cancer patients (Feng et al., 2018) | https://tcia.at |
| **CancerEnD:** Enhancer information for various cancer types (Kumar et al., 2020) | https://webs.iiitd.edu.in/raghava/cancerend/ |
| **NGDC:** Genomics data centre (National Genomics Data Center & Partners, 2020) | https://ngdc.cncb.ac.cn |

## *2.8 HLA-typing tools*

Due to the advancement in sequencing technologies a number of in-silico tools and computational pipelines have been generated for HLA typing. HLA genotype can be utilised as a biomarker in immunotherapy. A clinician can create an appropriate tailored therapy or immunotherapy for cancer patients with a better understanding of HLA types. Several computer pipelines and techniques have been created in the past for the reliable and exact genotyping of HLA alleles utilising the human genome. These tools utilized whole genome, whole exome and RNA-sequencing data of the patients and performed in-silico typing of HLA-alleles (Boegel et al., 2012; Hosomichi et al., 2015; Wittig et al., 2015). For instance, seq2HLA determine HLA-alleles using the RNA-seq reads (Boegel et al., 2012), HLAminer perform class-I, II typing using shotgun sequencing data (Warren et al., 2012), Optitype (Szolek et al., 2014) and xHLA (Xie et al., 2017) uses NGS data for HLA-typing. List of pipelines and computational tools for HLA-typing is provided in Table 2.4.

**Table 2.4: List of in-silico HLA-typing pipelines and computational tools**

| Name & Description | Year | Weblink |
|---|---|---|
| **seq2HLA:** HLA-typing using RNA-seq reads (Boegel et al., 2012) | 2012 | https://github.com/TRON-Bioinformatics/seq2HLA |
| **HLAminer:** Class-I,II HLA-typing using shotgun sequencing reads (Warren et al., 2012) | 2012 | https://github.com/bcgsc/HLAminer |
| **ATHLATES:** HLA-typing using whole exome sequencing (Liu et al., 2013) | 2013 | https://www.broadinstitute.org/viral-genomics/athlates |
| **Optitype:** Class-I typing using NGS dataset (Szolek et al., 2014) | 2014 | https://github.com/FRED-2/OptiType |
| **HLAreporter:** A tool for HLA-typing from NGS data (Huang et al., 2015) | 2015 | http://paed.hku.hk/genome/ |
| **xHLA:** Four digit HLA-tying using NGS dataset (Xie et al., 2017) | 2017 | https://github.com/humanlongevity/HLA |
| **Kourami:** HLA discovery using whole genome sequencing (Lee & Kingsford, 2018) | 2018 | https://github.com/Kingsford-Group/kourami |
| **HLA*LA:** HLA-genotyping using whole genome sequencing & whole exome sequencing (Dilthey et al., 2019) | 2018 | https://github.com/DiltheyLab/HLA-LA |
| **HISAT-genotype:** Identification of HLA from whole genome sequencing (Kim et al., 2019) | 2019 | https://daehwankimlab.github.io/hisat-genotype/ |

## *2.9 HLA Class-I binder*

Short, linear protein fragments known as major histocompatibility complex (MHC) binders or HLA binders attach to HLA molecules so that T-cell receptors may examine them (TCRs). Non-self-antigens are recognised by T lymphocytes as peptide fragments linked to MHC molecules and displayed on the cell surface. The outer extracellular domains of MHC molecules, which are membrane proteins, create a gap in which a peptide fragment is bound. HLA class I (HLA-I) molecules that bind intracellular short peptides are derived from the degradation of ubiquitinated cytosolic proteins in proteasomes and interacts with CD8+ T cells. Prediction of binding peptides corresponding to class-I alleles is very crucial for designing peptide-based therapeutics (Meydan et al., 2013; Vang & Xie, 2017). In the last two decades, huge number of computational tools have been generated for the accurate prediction of HLA-binding peptides (See Table 2.5). Studies shows that, the binding groove of HLA-I alleles is well-defined and closed from both sides (Kosaloglu-Yalcin et al., 2021). Therefore, a number of HLA-I binder prediction tools have been purposed by researchers. Table 2.5 enlists major HLA-I binder prediction tools. ProPred1 (Singh & Raghava, 2003) is the oldest and highly accurate in-silico method for the MHC-I binder prediction. However, NetMHCpan 4.0 (Jurtz et al., 2017) and

MHCflurry 2.0 (O'Donnell et al., 2020) software are recently developed for the prediction of larger number of HLA-I alleles binding peptides.

## *2.10 HLA Class-II binder*

HLA class II (HLA-II) molecules bind extracellular peptides and present them to the cell surface for recognition by T-cells with receptors. During pathogen infection and tumour development, CD4+ helper T lymphocytes play crucial roles in the immune response by detecting antigenic peptides presented by class II major histocompatibility complexes (MHC-II). It is difficult to predict binders corresponding to class-II HLA-alleles as the binding groove of HLA-II alleles is open from both sides and not well-defined. Although several computer techniques have been published for predicting peptide binding to HLA-II proteins, however, their effectiveness differs substantially. HLA-DR4Pred (Bhasin & Raghava, 2004) is the in-silico method used for the prediction of binders corresponding to HLA-DRB1*0401 binding peptides. With the advancements of computational algorithms, it is possible to predict binders corresponding to number of alleles, MULTIPRED2 (Zhang et al., 2011) makes it simple to predict peptide binding to several alleles of HLA class I and class II DR molecules. It allows for the prediction of peptide binding to products made by a single HLA allele, a group of alleles, or a supertype of HLA. Prediction engines NetMHCIIpan (Reynisson et al., 2020) is employed for the prediction of hundreds of MHC-II alleles binder. Table 2.5 shows the description of major HLA-II binder prediction methods.

**Table 2.5: Computational tools for Class-I & Class-II HLA-binder prediction**

| Name & Description | Year | Weblink |
|---|---|---|
| **Class-I HLA-binder prediction tools** | | |
| **MHCPred:** MHC-peptide binding prediction (Guan et al., 2003) | 2003 | http://www.ddg-pharmfac.net/mhcpred/MHCPred/ |
| **ProPred1:** MHC-I binder prediction method (Singh & Raghava, 2003) | 2003 | http://webs.iiitd.edu.in/raghava/propred1/ |
| **nHLAPred:** MHC Class I binders prediction tool (Bhasin & Raghava, 2007) | 2004 | http://webs.iiitd.edu.in/raghava/nhlapred/ |
| **POPI: P**redicting immunogenicity of MHC-I binding peptides (Tung & Ho, 2007) | 2007 | http://iclab.life.nctu.edu.tw/POPI |
| **NetCTLpan:** MHC class-I epitope prediction (Stranzl et al., 2010) | 2010 | http://www.cbs.dtu.dk/services/NetCTLpan/ |
| **NetMHCcons:** Consensus method for predicting MHC class I binders (Karosiene et al., 2012) | 2012 | http://www.cbs.dtu.dk/services/NetMHCcons/ |

| | | |
|---|---|---|
| **NetMHCpan 4.0:** HLA-neoepitope prediction tool (Jurtz et al., 2017) | 2017 | http://www.cbs.dtu.dk/services/NetMHCpan/ |
| **MHCflurry 2.0:** MHC-I binding peptide prediction (O'Donnell et al., 2020) | 2020 | https://github.com/openvax/mhcflurry |
| **Class-II HLA-binder prediction tools** | | |
| **ProPred:** HLA-DR binding peptide prediction (Singh & Raghava, 2001) | 2001 | https://webs.iiitd.edu.in/raghava/propred/ |
| **HLA-DR4Pred:** Prediction of MHC Class II alleles (HLA-DRB1*0401) binding peptides (Bhasin & Raghava, 2004) | 2004 | http://webs.iiitd.edu.in/raghava/hladr4pred/ |
| **MHCMIR:** Prediction of the binding affinity of MHC-II peptides (Nielsen et al., 2007) | 2007 | http://ailab.ist.psu.edu/mhcmir/predict.html |
| **EpiTOP:** HLA-DRB1 alleles binder prediction (Dimitrov et al., 2010) | 2010 | http://www.pharmfac.net/EpiTOP |
| **MULTIPRED2:** Class-I and Class-II HLA supertype binder prediction (Zhang et al., 2011) | 2010 | http://cvc.dfci.harvard.edu/multipred2/ |
| **EpiDOCK:** Prediction of MHC-II binders (Atanasova et al., 2013) | 2013 | http://www.ddg-pharmfac.net/epidock/EpiDockPage.html |
| **Consensus:** A tool for MHC-II binder prediction | 2013 | http://tools.iedb.org/mhcii/ |
| **NetMHCII - 2.3:** Binders of MHC-II molecules (Jensen et al., 2018) | 2018 | https://services.healthtech.dtu.dk/service.php?NetMHCII-2.3 |
| **DeepHLApan:** Neoantigen prediction using deep learning (Wu et al., 2019) | 2019 | https://github.com/jiujiezz/deephlapan |
| **MHCnuggets:** HLA-neoantigen binding prediction (Shao et al., 2020) | 2020 | https://github.com/KarchinLab/mhcnuggets |

## *2.11 Cytokine prediction tools*

It is not always desirable to identify the HLA-binding peptides or their immunogenicity. Due to the T cells' varying responses to various antigens and cytokine release patterns, identification of cytokine release-specific T cell epitopes are crucial because protective immunity against various infectious agents varies (Sidney et al., 2020). Numerous scientists have worked to create prediction methods that can categories specific cytokine-inducing antigen epitopes. These cytokines inducing peptides may act as potential therapeutic target while designing subunit vaccines which can elicit the appropriate immune response against cancer and immunological disorders (Kumai et al., 2017). Due to the availability of huge amount of experimentally validated epitope data for most of the cytokines in the immune epitope database IEDB (R et al., 2019), a number of computational tools have been developed for the prediction of cytokine inducing peptides. These machine learning based methods used by experimental biologist while designing subunit vaccine or peptide based cancer immunotherapies. In

Table 2.6, we enlist some cytokine specific tools which can be used for the prediction of cytokine inducing peptides.

**Table 2.6: In-silico methods for the prediction of cytokines inducing peptides**

| Name & Description | Year | Weblink |
|---|---|---|
| **IFNepitope:** Interferon-gamma inducing peptides prediction (Dhanda, Vir, et al., 2013) | 2013 | https://webs.iiitd.edu.in/raghava/ifnepitope/ |
| **IL-4Pred:** IL-13 inducing peptides prediction (Dhanda, Vir, et al., 2013) | 2013 | https://webs.iiitd.edu.in/raghava/il4pred/ |
| **ProInflam:** Proinflammatory cytokines prediction method (Gupta et al., 2016) | 2016 | http://metabiosys.iiserb.ac.in/proinflam/ |
| **IL10Pred:** IL-10 inducing peptides prediction (Nagpal et al., 2017) | 2017 | https://webs.iiitd.edu.in/raghava/il10pred/ |
| **IL17eScan:** IL-17 inducing peptides prediction (Gupta, Mittal, et al., 2017) | 2017 | http://metagenomics.iiserb.ac.in/IL17eScan/ |
| **AntiInflam:** Anti-inflammatory peptides prediction (Gupta, Sharma, et al., 2017) | 2017 | http://metagenomics.iiserb.ac.in/antiinflam/ |
| **PIP-EL:** Proinflammatory peptide prediction (Manavalan et al., 2018) | 2018 | http://www.thegleelab.org/PIP-EL/ |
| **IL2Pred:** Identification of IL-2 inducing peptides (Anjali Lathwal, 2021) | 2021 | https://webs.iiitd.edu.in/raghava/il2pred/ |
| **IL13Pred:** Prediction of IL-13 inducing epitopes (Jain et al., 2022) | 2021 | https://webs.iiitd.edu.in/raghava/il13pred/ |

## *2.12 Conclusion*

Human leukocyte antigens (HLA) molecules are plays significant role in the regulation of immune system and provide right defence and protection against the cancer or other diseases. In the IMGT/HLA, thousands of class-I and class-II HLA-alleles have been reported, however a specific type of alleles are present in an individual. This specific set of HLA-alleles plays an important role and impacts on the survival of the cancer patients. A number of past studies reported the prognostic biomarkers based on the gene expression and mutation profiles of cancer patients. However, with the knowledge of accurate HLA-typing one can design personalized vaccines and immunotherapy based prognostic biomarkers against cancer. Moreover, HLA-binding peptides are very crucial for eliciting the immune response against cancer cells. In the past, a number of computational tool developed for the prediction of classical HLA binding peptides. However, there is no specific method for the non-classical HLA-binder prediction. Non-classical HLA (HLA-G and HLA-E) are important immunoregulatory molecules; therefore, it is the need of the hour to develop computational tool for

the prediction of binders corresponding to non-classical HLA alleles. Cytokines inducing peptides or epitopes prediction methods are necessary for the prediction of antigenic regions or potential subunit vaccine candidates. Moreover, it is very crucial to develop a computational tool for the prediction or designing of anti-cancer drugs or molecules that can inhibit the IL6-mediated STAT3 signalling pathway in order to reduce the tumor progression and proliferation.

# CHAPTER 3

## PAN-CANCER RISK ESTIMATION ANALYSIS

## 3.1 Introduction

According to the American Cancer Society an estimation of 1,918,030 new cancer cases and 609,360 cancer deaths had occurred in the United States by the year 2022. Over the past few decades, researchers working very hard to find new therapies and solutions for the treatment of cancer (Pucci et al., 2019). The most widely utilised treatments include traditional therapies like chemotherapy, radiation, and surgery (Arruebo et al., 2011). The patient's health and survival are adversely affected by these radiation-based treatments (Altun & Sonkaya, 2018; Dilalla et al., 2020; Pucci et al., 2019). New treatment modalities, such as targeted cancer therapies, adoptive T cell therapy, immune checkpoint inhibitor-based therapies, immunomodulators, and oncolytic viruses based therapies have been created to overcome the limitations of conventional drugs (Dine et al., 2017; Esfahani et al., 2020; Franzin et al., 2020; Hemminki et al., 2020; Padma, 2015). Improvements in immunotherapy have produced notable results and improve the survival of many patients with a variety of solid tumours (Amin et al., 2020; Ruiz-Patino et al., 2020). Immune checkpoint inhibitors and chimeric antigen receptor (CAR) T cells are the two main foundations of immunotherapy. T-lymphocytes (T cells), which recognise tumor-associated peptides expressed on the infected cell surface by human leukocyte antigens, are completely necessary for these treatments (HLA) (Waldman et al., 2020).

As seen in Figure 3.1, when cells display antigenic peptides, the immune system is triggered to respond. The HLA genes, which are found on chromosome 6, are the most intricate and variable genes in the human genome. To start a sequence of immune responses aimed at removing the tumour cells from our system, CD8+ T cell receptors (TCR) interact with antigenic peptides presented by HLA class I alleles (Buhrman & Slansky, 2013; Chan et al., 2018; Engels et al., 2013; He et al., 2019). Recently, research has concentrated on HLA-dependent medicines for the treatment of cancer patients, including neoantigen-based therapy, tumor-infiltrating lymphocytes (TILs) therapy, and CD8+ T cell therapy (Sun et al., 2021; Yarmarkovich et al., 2021). Determined by HLA-peptide binding, cancer immunogenicity. HLA genotyping, neoantigens, and binding affinity, must be found in order to stratify patient-specific therapy. With the use of cutting-edge technologies and the accessibility of sequencing data, it is now possible to identify patient-specific HLA alleles. The integration of genomic datasets from cancer patients has been made possible in recent years by the development of several repositories and bioinformatics tools.

Genetic data such as HLA-alleles, neoantigens, HLA-peptide binding affinity, and immune response must be found in order to create patient-specific therapies. In the pilot study, we gathered patient-specific data from databases like the TCGA and TCIA, analysed patient survival based on HLA-alleles,

as well as the relationship between the frequency of neoantigens specific to HLA-alleles and overall survival in different cancer types. In addition, correlational analysis helped us comprehend how chemokines, cytokines, and their receptors affect the prognosis of cancer patients. User-friendly website named "CancerHLA-I" is accessible at https://webs.iiitd.edu.in/raghava/cancerhla1/ , we combined the aforementioned information for 20 different types of cancer. In Figure 3.1, the overall process of the current investigation is shown.

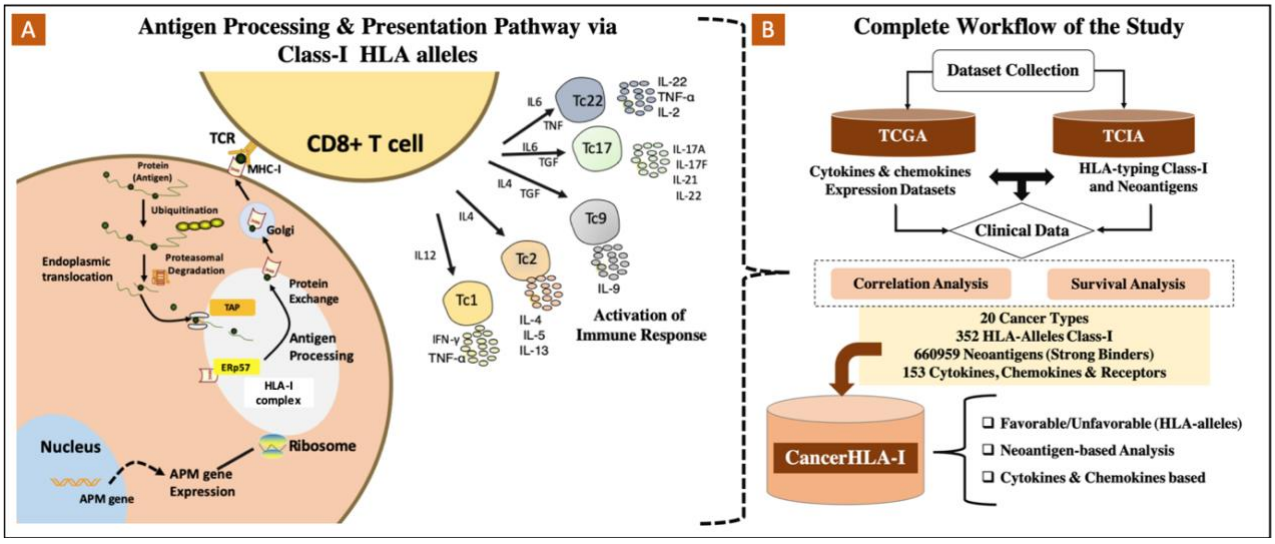



**Figure 3.1 Overall design of the study: (A) Presentation and processing of neobinders via Class-I HLA molecules (B) Pipeline of CancerHLA-I resource**

## *3.2 Material and methods*

### *3.2.1 Dataset collection*

We gathered genomic and clinical data for this investigation from The Cancer Genome Atlas (TCGA (Tomczak et al., 2015)) and The Cancer Immunome Atlas (TCIA) (Charoentong et al., 2017)) repositories. We obtain the control excess dataset from TCIA [with the approval of dbGap (Project No. 17674)], which contain class-I HLA-tying data and corresponding neoantigens for 20 type of cancer patients. We build patient specific class-I HLA typing and neoantigens data for each cancer type. Additionally, normalised RNA-seq data of cytokines, chemokines, and their receptors for each cancer type were downloaded using TCGA Assembler 2.0. After that, the expression profiles were converted into log2 values. Vital status and overall survival time are included in the survival information (OS). All of the research was done on 8346 cancer patients with 20 distinct cancer kinds.

### *3.2.2 HLA-binder prediction*

Using the MHCflurry 2.0 tool, we were able to identify the strong binding neoantigens/epitopes associated with each HLA-allele for each cancer (O'Donnell et al., 2020). Using the binding affinity (BA) percentile of the MHCflurry software, we categorise neoepitopes as strong or weak binders, while neoantigens with BA<2 are thought of as strong binders and vice versa. The amount of binders matching to each HLA-allele and tumour type was then put into a count matrix.

### *3.2.3 Mean-overall survival analysis*

We first created a binary matrix based on the presence or absence of HLA-alleles for each form of cancer. Each row represents samples/patients, and each column represents HLA-alleles. Based on the presence or absence of an HLA-allele, we calculated mean overall survival (MOS) using each person's survival data. The difference in MOS (based on presence/absence) is then calculated.

### *3.2.4 Univariate survival analysis*

In the current study, Cox-PH regression models were utilised to identify HLA-alleles associated with cancer patient survival. For the univariate analysis, the R package "survival" was utilised (V.3.5.1). The existence of an HLA allele has an adverse effect on survival (cox regression coefficient > 0), but the presence of alleles improves survival (cox regression coefficient < 0). We determined the Hazard Ratio (HR) and 95% Confidence Interval (CI) for each HLA-allele. While HR =1 has no effect on survival, HR >1 indicates high-risk HLA alleles while HR <1 indicates low-risk alleles. In order to assess the significant distribution of low-risk and high-risk patients, the log-rank test and p-value were also performed. We utilised the Concordance index (C) to determine how well each model predicted outcomes.

### *3.2.5 Correlation analysis*

#### *3.2.5.1 HLA-neoantigen*

After combining the survival data, we determine the Pearson correlation between the survival and the number of strong binders for each individual HLA-allele. The relevance of the quantity of HLA-binding neoepitopes on cancer patient survival is shown by the correlation coefficient (r) and p-value. Based on 20 cancers, we conduct association analyses for each HLA allele.

#### *3.2.5.2 Cytokines & chemokines*

The impact of cytokines, chemokines, and their receptor genes on cancer patient survival was examined using the Pearson correlation test. Data on survival as well as the expression of 153 cytokines expression profiles. The association analyses was conducted for each gene based on both the integration of expression across all malignancies and individual cancer type.

## 3.3 Results

### 3.3.1 Distribution of dataset

We first examine the distribution of HLA-alleles associated with each form of cancer. Table 3.1, details the descriptions of 20 different cancer kinds along with the total number of samples, HLA-alleles, and neoantigens associated with each type of cancer. We noticed that the most HLA-alleles were found in cases of uterine corpus endometrial cancer (UCEC) and kidney chromophobe (KICH) respectively, while fewer HLA-alleles were found in cases of other cancers.

**Table 3.1: Distribution of samples in twenty type of cancers**

| Cancer Type | Number of Samples |
|---|---|
| **Bladder urothelial carcinoma (BLCA)** | 407 |
| **Beast invasive carcinoma (BRCA)** | 1093 |
| **Cervical squamous cell carcinoma and Endocervical adenocarcinoma (CESC)** | 304 |
| **Colorectal Cancer (CRC)** | 455 |
| **Glioblastoma Multiforme (GBM)** | 154 |
| **Head and Neck Squamous cell Carcinoma (HNSC)** | 501 |
| **Kidney chromophobe (KICH)** | 65 |
| **Kidney renal clear cell carcinoma (KIRC)** | 533 |
| **Kidney renal papillary cell carcinoma (KIRP)** | 289 |
| **Liver Hepatocellular Carcinoma (LIHC)** | 370 |
| **Lung Adenocarcinoma (LUAD)** | 507 |
| **Lung Squamous cell Carcinoma (LUSC)** | 495 |
| **Ovarian serous cystadenocarcinoma (OV)** | 420 |
| **Pancreatic Adenocarcinoma (PAAD)** | 178 |
| **Prostate Adenocarcinoma (PRAD)** | 497 |
| **Rectum adenocarcinoma (READ)** | 165 |
| **Skin Cutaneous Melanoma (SKCM)** | 454 |
| **Stomach Adenocarcinoma (STAD)** | 410 |
| **Thyroid Carcinoma (THCA)** | 505 |
| **Uterine Corpus Endometrial Carcinoma (UCEC)** | 544 |

For the prediction of strong and weak neoantigen binders corresponding to each HLA-allele, we employed the MHCflurry 2.0 software. The total number of strong and weak binders corresponding to each cancer type is shown in Figure 3.2. For each cancer type, we have included the total number of both strong and weak binders. For the SKCM, UCEC and LUAD cancer types, the greatest number of strongly binding neoantigens was obtained. For the cancer types KICH, KIRP, LIHC, and THCA, we obtain less number of strong binders.

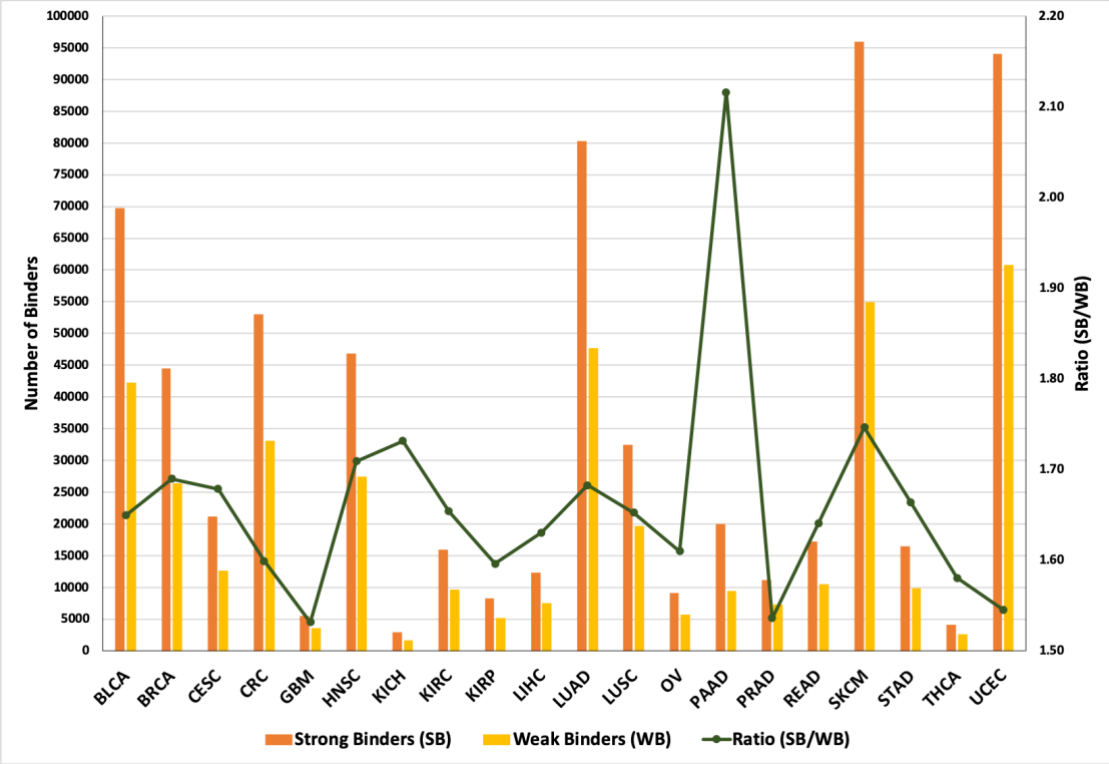



**Figure 3.2 Distributions and ratio of strong and weak Class-I HLA-binders in 20 types of cancer**

## *3.3.2 HLA-based biomarkers*

Based on the presence or absence of HLA alleles, we created binary matrices for each cancer patient. We calculated the Hazard ratio (HR), p-value, and concordance index for each allele in 20 cancer types using the utility of survival program. Some of the HLA-alleles with HR>1 that negatively affect cancer patients' survival are displayed in Table 3.2. In KICH, THCA, and PRAD cancer patients, we found that the presence of HLA-A*02:01, HLA-B*50:01, and HLA-B*52:01, HLA-B*50:01 was substantially related with poor survival (with HR>4). Additionally, the survival rate of cancer patients is considerably decreased by alleles such HLA-B*53:01, HLA-B*52:01, HLA-C*05:01, HLA-A*26:15 with an HR>2 and p-value less than or equal to 0.05. Some alleles are prevalent in many

cancer types and are linked to a bad prognosis. Additionally, the presence of HLA-alleles increases the likelihood that cancer patients will survive. For example, in certain types of cancer, HLA-C*14:02, HLA-B*07:02, HLA-C*12:03, HLA-A*23:01, HLA-B*27:05, and HLA-C*02:02 significantly act as good prognostic markers and improve the survival rate of cancer patients (See Table 3.2).

**Table 3.2: List of cancer types with best HLA-alleles based prognostic biomarkers obtained using univariable survival analysis**

| Cancer | HLA-allele | Present (No. of patients) | Absent (No. of patients) | Hazard (95%CI) | P-value | Concordance |
|---|---|---|---|---|---|---|
| **BLCA** | HLA-C*14:02 | 15 | 392 | 0.14(0.02-1.00) | 0.05 | 0.517 |
| **BRCA** | HLA-B*53:01 | 44 | 1049 | 2.32(1.25-4.30) | 0.007 | 0.524 |
| **CESC** | HLA-B*57:01 | 20 | 284 | 1.97(0.89-4.34) | 0.009 | 0.591 |
| **CRC** | HLA-B*07:02 | 84 | 371 | 0.57(0.32-0.98) | 0.045 | 0.54 |
| **GBM** | HLA-C*12:03 | 22 | 132 | 0.52(0.29-0.93) | 0.029 | 0.532 |
| **HNSC** | HLA-B*52:01 | 18 | 483 | 2.11(1.14-3.88) | 0.016 | 0.514 |
| **KICH** | HLA-A*02:01 | 26 | 39 | 5.46(1.13-26.29) | 0.034 | 0.72 |
| **KIRC** | HLA-A*23:01 | 32 | 501 | 0.43(0.19-0.97) | 0.044 | 0.519 |
| **KIRP** | HLA-A*03:01 | 74 | 215 | 1.88(1.01-3.52) | 0.044 | 0.531 |
| **LIHC** | HLA-B*44:03 | 46 | 324 | 1.66(1.04-2.66) | 0.033 | 0.529 |
| **LUAD** | HLA-B*08:01 | 94 | 413 | 1.62(1.15-2.28) | 0.005 | 0.544 |
| **LUSC** | HLA-C*07:01 | 139 | 356 | 1.36(1.02-1.83) | 0.037 | 0.526 |
| **OV** | HLA-C*02:02 | 46 | 374 | 0.65(0.43-0.90) | 0.041 | 0.517 |
| **PAAD** | HLA-B*50:01 | 5 | 173 | 3.66(1.33-10.11) | 0.002 | 0.52 |
| **PRAD** | HLA-B*50:01 | 10 | 487 | 10.09(1.92-53.10) | 0.006 | 0.574 |
| **READ** | HLA-C*05:01 | 22 | 143 | 2.21(0.80-5.56) | 0.009 | 0.597 |
| **SKCM** | HLA-B*27:05 | 45 | 409 | 0.52(0.30-0.92) | 0.025 | 0.52 |
| **STAD** | HLA-C*14:02 | 20 | 390 | 0.32(0.1-0.98) | 0.048 | 0.516 |
| **THCA** | HLA-B*52:01 | 25 | 480 | 4.05(1.15-14.25) | 0.029 | 0.62 |
| **UCEC** | HLA-A*26:15 | 8 | 536 | 2.68(0.84-8.49) | 0.009 | 0.514 |

### *3.3.3 Neoepitope based biomarkers*

To comprehend how HLA-binders affect cancer patients' chances of survival, we employed the Pearson correlation test. The association between the number of neoantigens and the overall survival of each cancer type has been calculated. As shown in Figure 3.3, we found that the majority of HLA that have a detrimental influence on survival also have a negative correlation with survival. A few

alleles, such as HLA-A*01:01, HLA-B*15:03, HLA-B*44:03, HLA-C*02:10, etc., are provided in Figure 3.3.

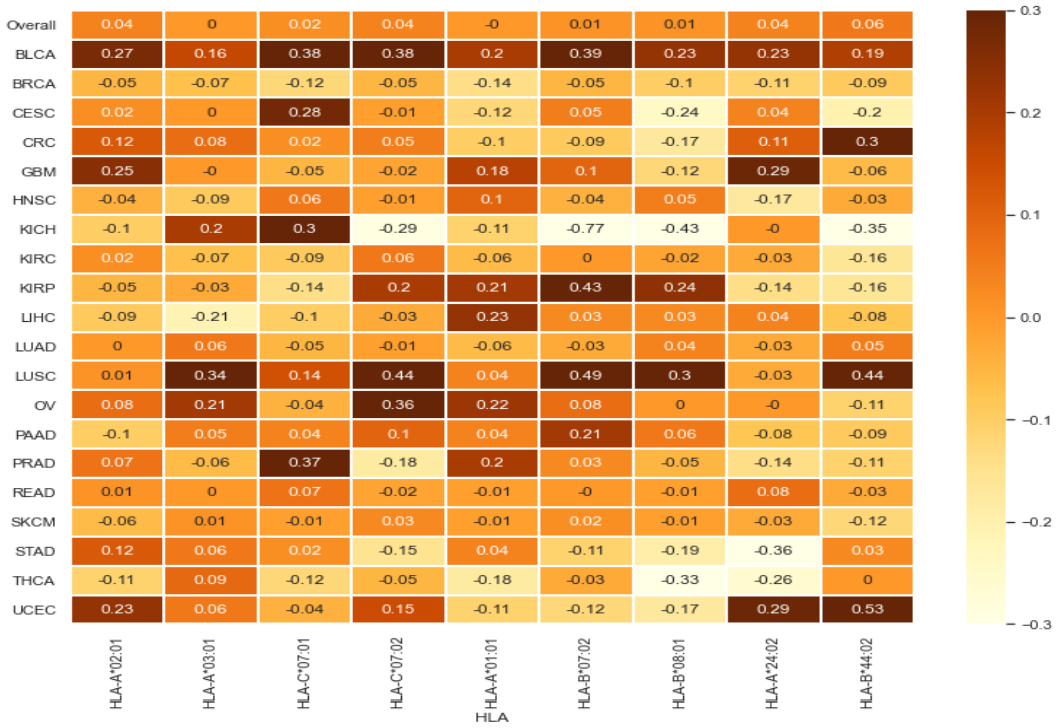



**Figure 3.3 Heatmap shows correlation between number of neobinders (Class-I HLA) and overall survival of cancer patients. Where, light colour depicts negative correlation and dark colour shows positive correlation**

## *3.3.4 Cytokines-based prognostic biomarkers*

We have carried out univariate survival analysis employing the expression of these immune genes in order to find survival favourable and unfavourable cytokines and chemokines. We have included the cytokines and chemokines in Figure 3.4 whose expression has a significant impact on cancer patient's survival rates. We found that high expression levels of the cytokines IL2, IFNB1, IFNA8, and IL5 had a significant impact on the survival of various cancer patients (HR 0.4 and p-value 0.05). However, in KICH, READ, and GBM patients, elevated levels of IL5RA, TGFBR3, CCR4, TGFB2, and IL17A are strongly linked to a poor survival rate (HR >4 and p-value 0.05).

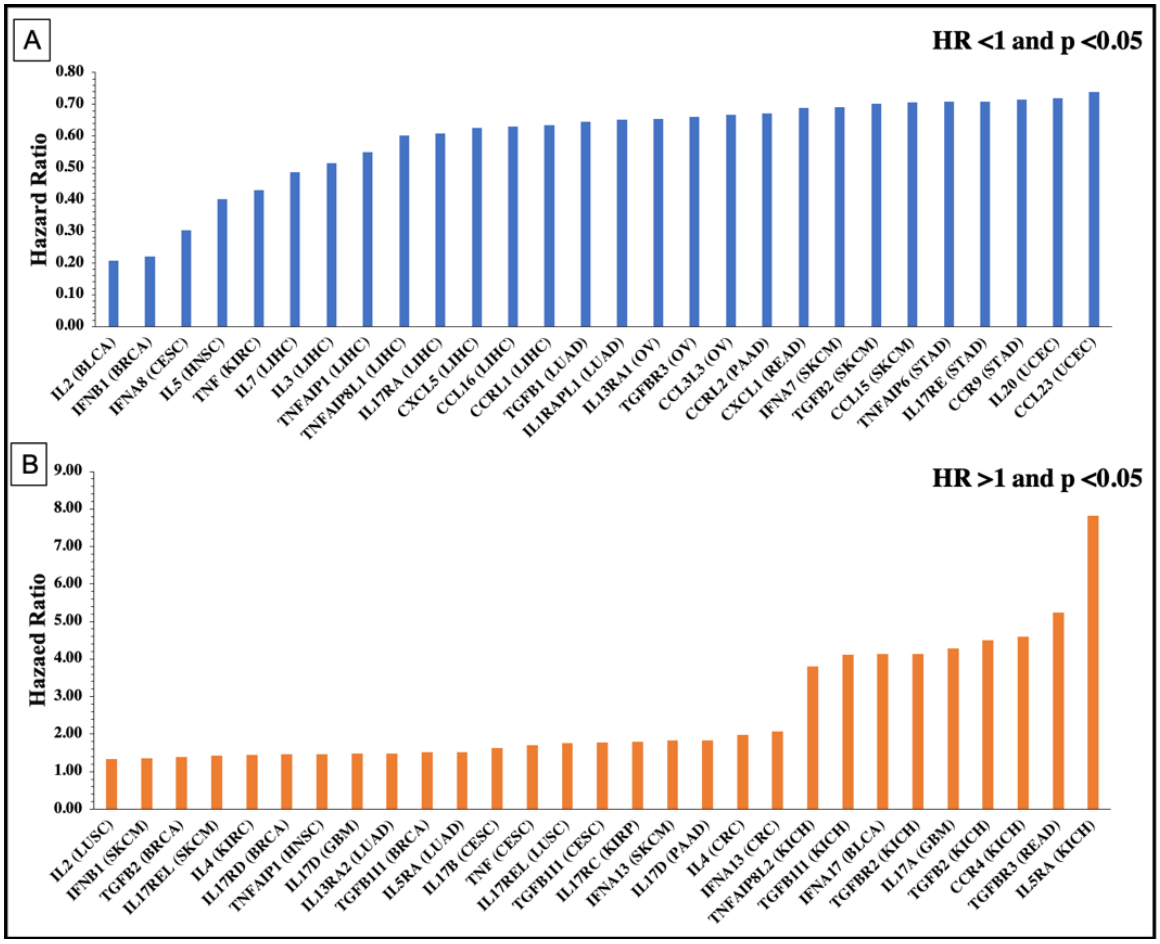



**Figure 3.4 Shows Hazard ratio for different cytokines whose expression plays significant role (p<0.05) with the survival of cancer patients obtained using univariate survival analysis. A) Survival favourable cytokines/chemokines (higher expression increases the survival) B) Survival unfavourable cytokines/chemokines (higher expression decreases the survival of cancer patients)**

Additionally, by taking into account the gene expression of cytokines, chemokines, and their receptors, we conducted association analysis. Figure 3.5, heatmap depicts the relationship between overall survival and the expression of several genes in 20 different cancer types. While pale yellow colour illustrates the negative correlation, and darker blue colour indicates the positive correlation. We found that increased expression of the cytokine IL9 is linked to a substantial positive association in cancer patients with BLCA, KIRC, and OV, and that cytokine IFNG has a very high and significant correlation with the survival rate of HNSC cancer patients. Contrarily, the cytokines IL2, IL5, IL12A, TNFA1P8, and TNF are linked to the opposite correlation ($p < 0.05$). We found a strong positive correlation between the expression rates of CCL1 in (colorectal cancer and kidney renal clear cell carcinoma), CCL20 (bladder urothelial carcinoma), and CCL27 (prostate adenocarcinoma)

chemokines. The correlation of some of the cytokines and chemokines with the survival of 20 types of cancer patients in demonstrated in Figure 3.5.

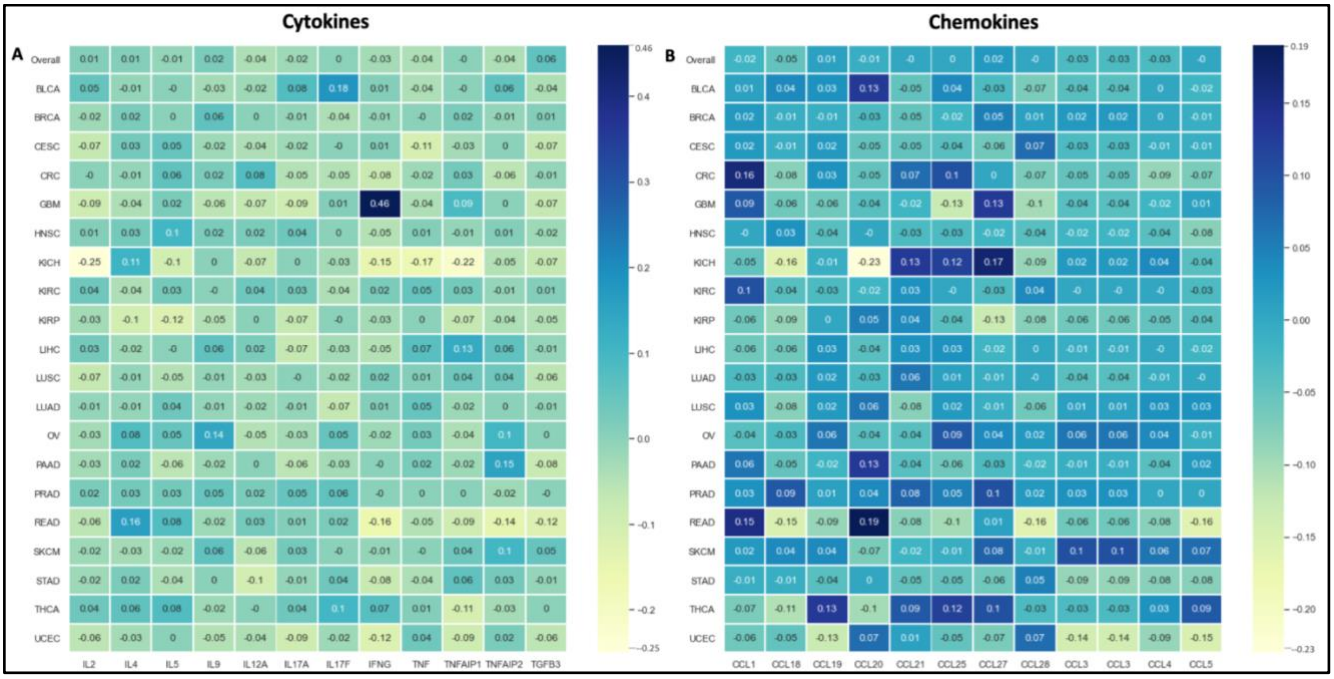



**Figure 3.5 Heatmap shows the correlation of expression of cytokines and chemokines with the overall survival of cancer patients A) Cytokines B) Chemokines and, where pale yellow depicts the negative correlation with survival, darker blue colour shows positive correlation with survival of cancer patients**

## *3.4 Web-server Implementation*

The web-interface of CancerHLA-I is developed using HTML, CSS, JavaScript, MySQL and PHP. The webserver is responsive and can be browsed/searched on various web browsers such as google chrome, Firefox, safari and variety of devises (smartphones, tablets, desktops and laptops). CancerHLA-I resource provides a simple search page, where users can search query in the database for specific cancer type, HLA-allele, neoantigens, cytokine/chemokine, and its survival association. The home page of webserver is provided in Figure 3.6.

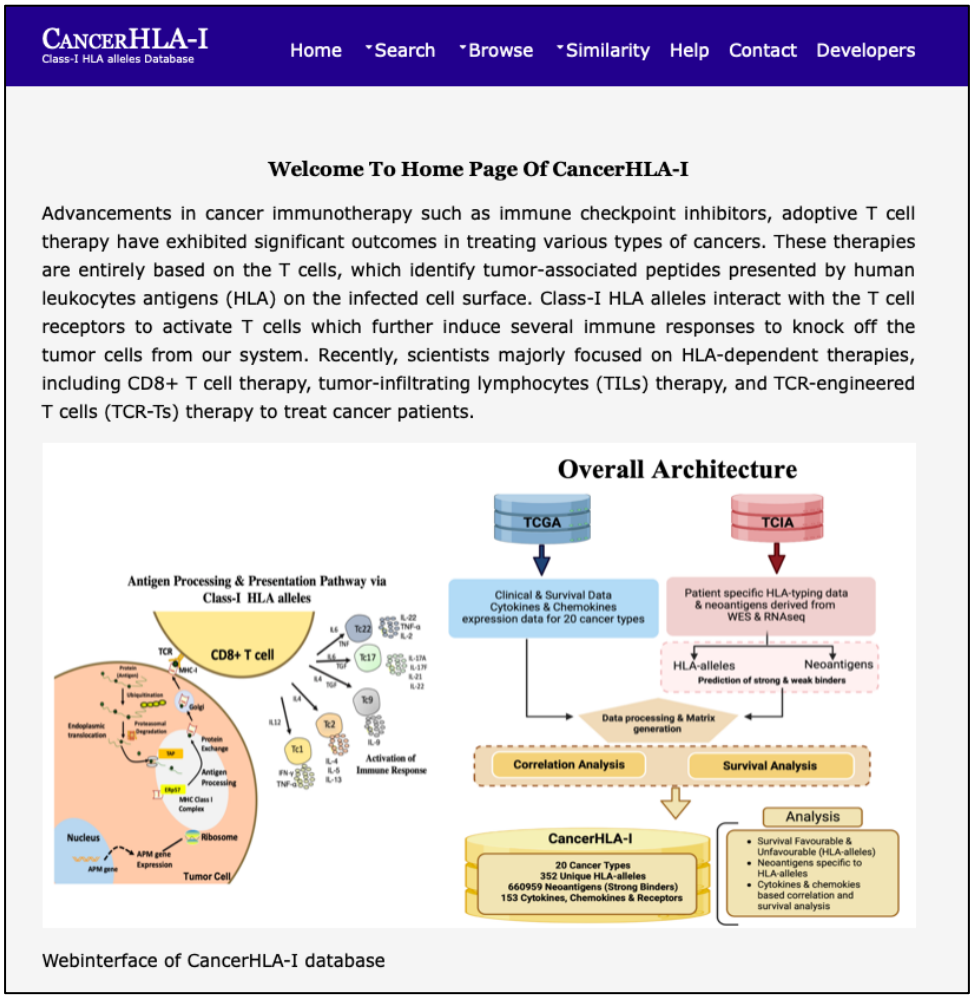



**Figure 3.6: Homepage of CancerHLA-I webserver**
**(https://webs.iiitd.edu.in/raghava/cancerhla1/ )**

## *3.5 Discussion*

Class-I (HLA-A, HLA-B, and HLA-C) molecules are crucial for cancer immunotherapy and immunosurveillance. It is essential to deliver tumor-specific peptides or neoantigens via HLA-alleles for our immune system to recognise and destroy tumour cells. However, several cancer types demonstrate escape mechanisms due to the loss of class-I HLA molecules' activities. Numerous studies also claim that the overexpression of class-I non-classical HLA molecules is crucial for the immunological escape of malignancy. Class-I HLA alleles interact with T cell receptors to activate T cells, which then trigger a series of immunological responses to eliminate tumour cells from our system. To treat cancer patients, scientists have recently focused on HLA-dependent therapies such as CD8+ T cell therapy, tumor-infiltrating lymphocytes (TILs) therapy, and TCR-engineered T cells (TCR-Ts). HLA-dependent immunotherapies are more successful and efficient than standard chemotherapies. Patients with non-small cell lung carcinoma and colorectal cancer have loss of heterozygosity in the HLA genes on chromosome 6 as a result of alterations at the genetic and

epigenetic levels. Additionally, changes in the type-I and type-II interferon pathway genes affect the prognosis of cancer patients. Interleukins including *IL6, IL-11, IL-1,* and *TGF* promote the growth and advancement of cancer cells (Esquivel-Velazquez et al., 2015). According to a new study, the presence of particular HLA-alleles can predict the drug response or therapy response on cancer patients. They found that patients with kidney cancer who carried the HLA-A*03 allele had lower survival rates and responded poorly to immune checkpoint inhibitor (ICI) therapy (Naranbhai et al., 2022). According to studies, cytokines are crucial in controlling the tumour microenvironment. Therefore, in order to grasp the impact and effectiveness of cancer immunotherapy, it is essential to comprehend the prognostic significance of HLA-alleles and cytokines.

In this study we investigate the connections of class-I HLA alleles with cancer patient survival in order to aid cancer researchers. Using the HLA data of cancer patients, we do a pan-cancer study. HLA typing and clinical information were obtained using the cancer genome atlas (TCGA) and the cancer immunome atlas. To ascertain the relationship between the presence or absence of the 352 HLA-alleles and the prognosis for cancer, we employed survival analysis. Our findings show that the sample distributions for the various cancer types are skewed. For instance, we found only 10 PRAD cancer patients are dead, while the remaining 487 patients are either alive or censored. As a result, in this instance the hazard index is very large and sample discrimination is relatively straightforward. On the other hand, in the example of BLCA cancer, where 228 patients are alive and censored but 178 patients are dead, we achieved a very low hazard index. Additionally, in the instance of BLCA cancer patients have large number of neoepitopes against HLA-alleles, which leads to positive correlation values with the overall survival. Contrarily, cancer patients with PRAD have extremely few neobinders, leading to negative correlation values. Additionally, we look into the connection between cancer patients' overall survival and the expression of cytokines and chemokines. We found that the IFNG cytokine has a positive correlation coefficient of $r = +0.46$; this suggests that higher levels of IFNG expression increase the survival of cancer patients. Our study is anticipated to produce potential new HLA-biomarkers for enhanced cancer immunotherapy and treatment.

## *3.6 Conclusion*

This study reveals that survival of cancer patients depends upon the type of HLA-allele. Correlation and univariate survival analysis shows class-I HLA alleles, HLA-I neobinders and cytokines are significantly associated with the survival of cancer patients. Moreover, we have provided a user-friendly web portal for the identification of cancer specific biomarkers. The cancer specific peptides also provided in the CancerHLA-I (https://webs.iiitd.edu.in/raghava/cancerhla1/) database, which can be further examined by the experimental biologist in order to design cancer specific immunotherapies.

# CHAPTER 4

## PERSONALIZED HLA-BASED PROGNOSTIC BIOMARKERS FOR SKIN CANCER

## *4.1 Introduction*

Melanoma cancer accounts for 0.6% of cancer-related fatalities and 1.6% of newly diagnosed cancer cases worldwide (Sung et al., 2021). The American Cancer Society estimates that there will be 7,650 fatalities and 99,780 new cases of melanoma in the United States in 2022. Males are more likely than females to get melanoma. Melanoma develops when healthy human epithelial melanocytes, which are found in the skin's basement membrane, undergo malignant transformation (Soura et al., 2016). Environmental and genetic variables include excessive UV radiation exposure, indoor tanning beds, and interaction with certain chemicals are some of the major causes (Volkovova et al., 2012). Previous research examined multi-omics markers for the advancement of malignancy and found that cutaneous melanoma is one of the most dangerous and lethal types of skin cancer (Bhalla et al., 2019; Li et al., 2015; Ossio et al., 2017). Furthermore, it has been demonstrated in the past that melanoma has a 95% OS rate if it is discovered at an early stage; but, once it has spread (lesion thickness > 4 mm), it is difficult to treat and the survival rate drops to less than 50% (Bristow et al., 2010; Buttner et al., 1995). Therefore, tumour staging is essential to give clinicians the basic prognostic information they need and information regarding tumour stage grouping and tumor-nodes-metastasis (TNM) classification is provided by the American Joint Committee on Cancer (AJCC) and the Melanoma Staging Committee. Primary tumours (stages I and II) are divided into T1, T2, T3, and T4 categories, with corresponding tumour thicknesses of 1.00 mm, 1.01 – 2.0 mm, 2.01 – 4.0 mm, and >4.0 mm. Regional lymph nodes (stage III) are divided into N0, N1, N2, and N3, which stand for the number of metastatic tumour nodes (i.e., 0, 1, 2, 3, and 4+) and distant metastasis (stage IV) further divided in four categories—M0, M1a, M1b and M1c (Dickson & Gershenwald, 2011; Gershenwald et al., 1998).

Prior research has shown that melanoma tumour cells are able to bypass immunological checkpoints and multiply more quickly than normal tissue cells (Khair et al., 2019). Tumor resistance to apoptosis has been linked to HLA class I, II alleles. Immune responses are induced and regulated directly via HLA molecules. Recent research indicates that the poor prognosis and metastatic progression may be related to the altered expression of HLA molecules. Some of the key escape strategies employed by tumour cells to circumvent the immune response include the modification of surface molecules, the absence of co-stimulatory molecules, the creation of immunosuppressive cytokines, and changes to HLA molecules (Aptsiauri et al., 2007; Johansen et al., 2016; Mendez et al., 2009; Sabapathy & Nam, 2008). Melanoma is also classified as an immunogenic tumour since its lesions have been reported to exhibit markers for a number of immune escape strategies, including the downregulation of HLA molecule expression, the release of cytokines like IL-10, and the loss of tumor-specific antigens. The

poor prognosis and ineffective treatment in melanoma cases have been significantly linked with the downregulation of class I HLA molecules (Cabrera et al., 2007; Nestle et al., 1997). Furthermore, current research highlights the significance of HLA alleles in melanoma prognosis. One instance is the loss of heterozygosity in the HLA class I allele (HLA-B*15:01), which has been linked to a poor prognosis for survival. Additionally, it has been demonstrated that HLA-C alleles and the HLA-B44 supertype improve overall survival rate (Campillo et al., 2006; Chowell et al., 2018; Gogas et al., 2010).

Therefore, understanding how class I and class II antigens affect melanoma patients' survival is crucial. Accurate HLA typing allows for the creation of tailored cancer vaccines and prognostic biomarkers for immunotherapy. In the current study, we have used The Cancer Genome Atlas (TCGA-SKCM) dataset, we have attempted to investigate the function of HLA (class I and II) alleles and superalleles in the survival of cutaneous melanoma patients. Here, we first performed class I and class II HLA typing on the patients before assigning them to superallele (low-resolution HLA allele) groups. We next divided the HLA superalleles into groups that were survival-favourable and unfavourable depending on how significantly their presence affected patient survival. Additionally, using various machine learning techniques, we have created survival prediction models that incorporate important HLA superalleles, patient demographics, and clinical characteristics. As a further service to the scientific community, we created the "SKCMhrp" webserver for computing the survival rate of high-risk skin cutaneous melanoma patients using the clinical and HLA-typing information.

## *4.2 Materials and methods*

### *4.2.1 Pipeline of the study*

The entire study's workflow, including data collection and compilation, survival analysis, model construction, and webserver implementation, is depicted in Figure 4.1.

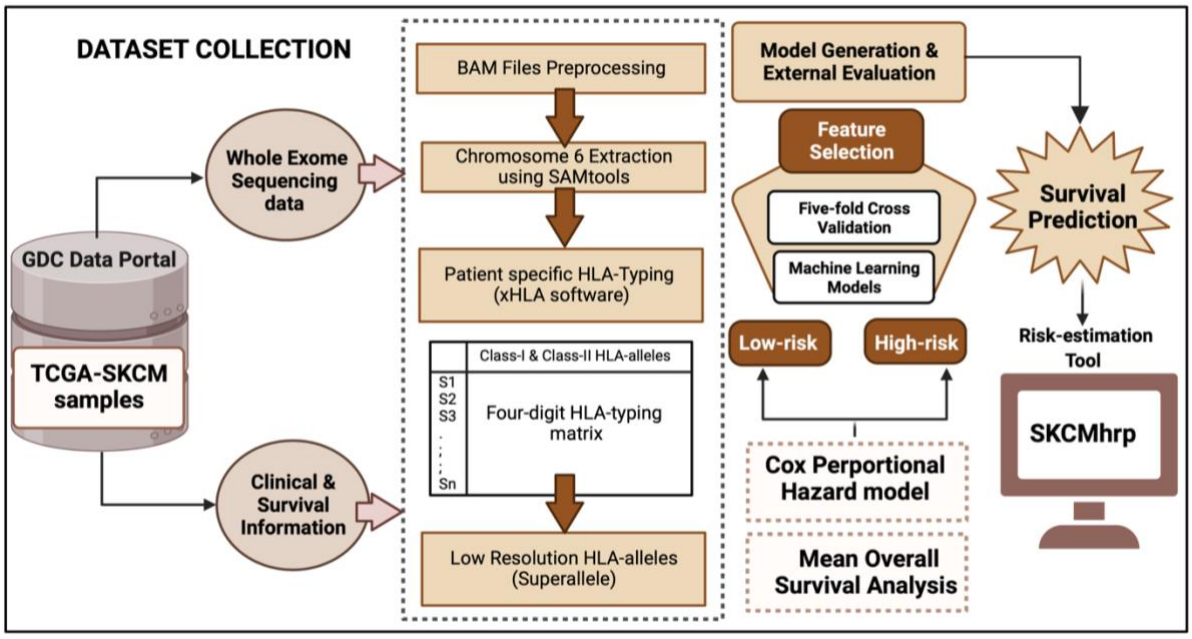



**Figure 4.1 Steps involved in the development of SKCMhrp; including the pre-processing of clinical and genomic data, building of prediction models and webserver**

## *4.2.2 Collection of dataset*

We accessed the Genome Data Commons (GDC) data portal to retrieve the TCGA-SKCM controlled access dataset. With the aid of an internal high-performance computing (HPC) facility and scripts, the whole-exome sequencing (WXS) BAM files of distinct melanoma patients were specifically downloaded [with the consent of dbGap (Project No. 17674)] in accordance with the GDC protocols (Grossman et al., 2016). Using TCGA Assembler 2, clinical data for 470 patients was also gathered, including age, gender, stage, tumour status, therapy status, Breslow depth, vital status, overall survival (OS), etc (J. Liu et al., 2018; Zhu et al., 2014). After deleting irrelevant BAM file errors, we were only able to retrieve the HLA type information for 415 of the 470 TCGA-SKCM patients, 14 patients out of 415 samples lacked OS data. In summary, we used 401 patients with cutaneous melanoma for whom complete survival statistics with exome sequencing data were available. The clinical details of the patients, such as the type of melanoma, tumour stage, tumour site, Breslow depth, and treatments, are displayed in Table 4.1.

**Table 4.1: Distribution of TCGA-SKCM samples based on clinical and demographic characteristics**

| Clinical Parameter | Description | No. of samples |
|---|---|---|
| **Age** | Age <=58 | 197 |
| | Age > 58 | 211 |
| **Gender** | Male | 256 |
| | Female | 159 |
| **Tumor Stage** | Stage 0 | 7 |
| | Stage I | 67 |
| | Stage II | 134 |
| | Stage III | 151 |
| | Stage IV | 21 |
| **Tumor Status** | With Tumor | 219 |
| | Without Tumor | 184 |
| **Breslow Depth** | <=1.0mm | 53 |
| | >1.0-2.0mm | 69 |
| | >2.0-4.0mm | 65 |
| | >4.0mm | 130 |

The prediction models were trained using the TCGA-SKCM dataset, and the effectiveness of our models was evaluated using a specific collection of variables, including HLA alleles and clinical traits. Finally, the performance was assessed using an external validation dataset. We obtained HLA-typing and clinical data of 121 cutaneous melanoma patients from several studies (Hugo et al., 2016; Riaz et al., 2017; Snyder et al., 2014; Van Allen et al., 2015). These data included 145 distinct class I and II HLA alleles with two demographics (age and gender) and one clinical feature (tumor stage). Finally, we have used the TCGA-SKCM dataset to train our machine learning model and the external dataset with a comparable collection of attributes to evaluate it.

### *4.2.3 Typing of HLA-alleles*

Chromosome 6 region was extracted from the BAM files using the SAMtools software after receiving the whole exome BAM files of cutaneous patients from TCGA (Li et al., 2009). After that, we identify HLA molecules from the region of chromosome 6 using the xHLA software (Xie et al., 2017). Four-digit HLA typing information was determine with their class I (-A, -B, -C) and class II (-DP, -DQ, -

DR) HLA alleles for each TCGA-SKCM patient. Each HLA-allele is given a distinct name in accordance with the IMGT/HLA nomenclature, which is then followed by an asterisk (*) and separated by colons (Marsh, 2003; Robinson et al., 2016). According to Listgarten et al. (2008), the first two digits identify an allele group (field1), third and fourth digits identify a particular HLA protein (field2). Due to the limited frequency of high-resolution HLA alleles in SKCM patients, we merged field1 HLA alleles which correspond to the historical serological antigen group (or allele family) to create low-resolution HLA alleles. For the first time, low-resolution HLA alleles were referred in this study as "superalleles" and a high resolution (i.e., four-digit typing) was given to a low resolution (i.e., two-digit typing) HLA allele.

### *4.2.4 HLA-superalleles*

Based on the effect of HLA superalleles on patient survival, i.e., whether the presence of the superallele enhances or degrades the survival rate. We divided HLA superalleles into favourable and unfavourable groups in this study. First, all patients were split into two groups; those who carried a certain HLA allele and those who did not; and the mean survival of patients was calculated for each group. Additionally, an allele was designated as a survival-favourable allele if the patients who carried it having significantly (p-value 0.05) longer survival than those who did not. Similar to this, an allele is designated as an unfavourable allele if patients with that allele have a poor mean survival rate than those with another allele. A single allele has only been found in a small subset of patients, so classifying patients based on the frequency of alleles will be biased. As a result, we classified patients into groups with survival-favourable (SF) and survival-unfavourable (SU) superalleles based on the presence or lack of HLA superalleles in the patients. Here, we used a two-sample t-test to determine whether these superalleles were statistically significant (p-value <0.05).

Notably, we only took into account superalleles that could belong to any of these groupings if they were present in at least 10 samples. We merged SF and SU superalleles and created a matrix to analyse the overall effects of their existence. If an SF or SU superallele was present, we gave it a score of +1 if a favourable superallele was present in an SKCM patient, and a score of -1 if an unfavourable superallele was, otherwise 0. Finally, the sum of all the alleles was calculated to produce a single score known as the risk score (RS). Threshold-based techniques have subsequently been created employing these superalleles as features. In the end, we classed a patient as high-risk if their score above the RS cut-off; otherwise, they were categorised as low-risk.

### *4.2.5 Statistical analysis*

Cox proportional hazard (Cox PH) models were used in the current investigation to perform "univariate" and "multivariate" survival analyses, which were then implemented by the "survival" package in R. (V.3.5.1). The impact of each variable, including age, tumour stage, tumour status, gender, class I, class II HLA alleles, HLA superalleles, and RS, on the prognosis of cutaneous melanoma patients was examined using univariate analysis. In addition, a multivariate survival analysis was carried out to comprehend the independent clinical impact of these HLA superalleles in the presence of additional numerous factors, including age, tumour stage, tumour status, gender, and class I, II HLA superalleles (Bradburn et al., 2003). The significant survival distributions between the high-risk and low-risk groups were estimated using the log-rank test in terms of the p-value. High-risk and low-risk groups were represented graphically using Kaplan-Meier (KM) survival curves (Goel et al., 2010).

### *4.2.6 Machine learning models*

The objective of the current study was to create regression models for OS time prediction in patients with cutaneous melanoma using a variety of machine learning techniques. We used a variety of features, such as HLA superalleles, as well as clinical and demographic aspects of the patients, such as age, gender, stage, tumour status, Breslow depth, and their interactions, to construct prediction models. Regression algorithms such as Decision tree (DT), random forest (RF), ridge, and lasso were used for the development of models utilizing the python-based scikit-learn library. DT is supervised machine learning model, which predicts the response variable by learning the decision rules from the predictor variables, is produced by regression using the decision tree approach. It is a top-down, tree-based method in which a decision tree is built by employing recursive partitioning. RF is a supervised machine learning technique that uses ensemble learning. When a model is being trained, it works by creating a number of decision trees, and then predicts the response variable using the average prediction of each tree. The shrinkage methodology is used by the linear regression technique known as least absolute shrinkage and selection operator, or LASSO. It applies L1 regularisation, creating a model with predictor variable coefficients that aid in predicting the response variable. Conversely, the L2 regularisation is carried out in ridge regression to determine the coefficients.

### *4.2.7 Feature selection techniques*

In this study we have used wrapper method for the selection of best-set of HLAs having poor impact on the survival of cancer patients. Here, HLA superalleles were individually added to the clinical and

demographic characteristics to create a recursive feature selection model. Following the prediction of survival time, the hazard ratio (HR) for each combination was calculated. In a nutshell, each time the input matrix was changed, a new column containing an HLA superallele was added; this superalleles HR was marginally higher than that of the input matrix before it. Until there was no more improvement in the HR, we kept repeating this technique. We were eventually left with the matrix that had the highest HR. This matrix was then employed to create the ultimate prediction model for the estimation of OS time.

### *4.2.8 Performance evaluation*

We used the five-fold cross-validation method to prevent over-optimization during model training (Patiyal et al., 2020). In a nutshell, each instance is separated into five sets at random, four of which are utilised for training and one for testing. This procedure run through five times so that each set is tested at least once. The performance on all five sets is averaged to determine the final performance. Choosing the right parameters to assess the performance of models is the main obstacle in these kinds of investigations. In this study, the performance of the models was evaluated using the standard parameter HR. The impact of an intervention on a desired outcome over time is measured by HR. By using a median cut-off, our regression models divide patients into high-risk and low-risk categories. We compute HR from the anticipated OS time for the group of patients to assess our model (high-risk or low-risk patients). We measured the confidence interval (CI) along with the HR and computed the HR at a 95% CI as well. We also computed the p-value using the log-rank test to assess the significance of the prediction.

## *4.3 Results*

### *4.3.1 Frequency of HLA-alleles*

Utilizing the xHLA software (Xie et al., 2017), we were able to extract 4711 HLA alleles from 415 TCGA-SKCM patients, 367 of which were unique. We identified that 237 HLAs are belonged to HLA class I genes i.e., HLA-A, HLA-B, and HLA-C, and 130 were class II genes i.e., HLA-DPB1, HLA-DQB1, HLA-DRB1. We calculated the patient population's frequency distribution of various alleles. All alleles were not identified in all individuals due to the variability of the HLA genes, therefore the frequency of alleles varies from patient to patient. We analysed that only 357 of the 415 individuals possessed all six HLA class I gene alleles. However, 264 patients possessed all six alleles of the HLA class II gene. We identified the most abundant class-I and class-II HLA alleles (found in more than

20% of the population) as shown in Figure 4.2. HLA-A*02:01 is the most abundant class-I HLA-allele present in more than 160 samples, whereas in class-II HLA-DPB1*04:01 is the most frequent allele present in 250 cutaneous melanoma samples. The distribution of all the class-I and class-II HLA-alleles is provided in Figure 4.2.

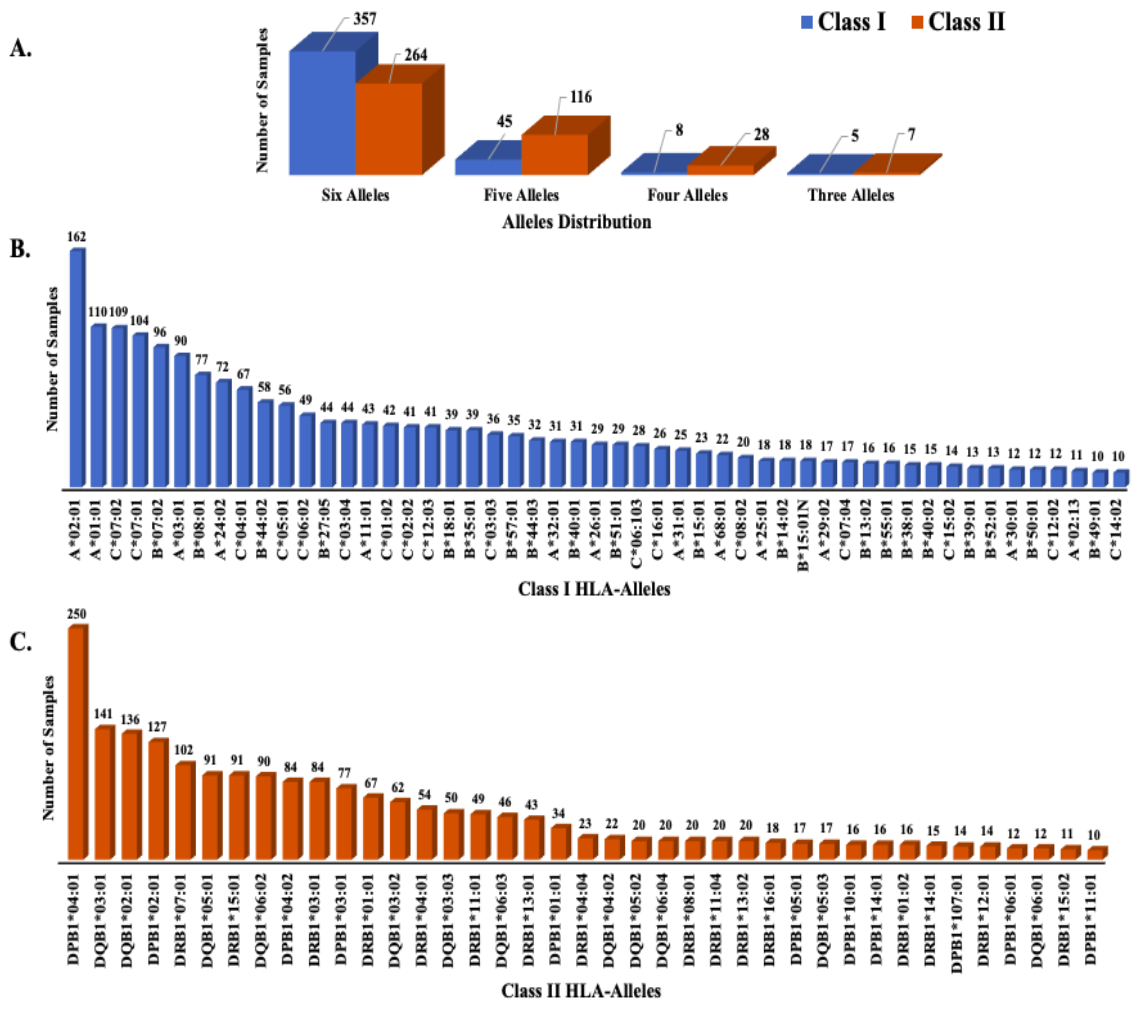



**Figure 4.2 Distribution of HLA-alleles in SKCM samples, (A) Number of samples having Class-I/II HLA-alleles (B) Number of samples having different types of Class-I HLA-alleles (C) Number of samples having different types of Class-II HLA-alleles**

## *4.3.2 Mean overall survival analysis*

We estimated the difference in mean overall survival (MOS) of patients to determine if an allele is beneficial for the patient's survival or not. If the difference in MOS is positive, the HLA allele is categorised as favourable; if not, it is unfavourable. These alleles can be used to estimate the likelihood of survival; regrettably, this statistic may be skewed because the majority of the alleles have very few

patients that carry them. As a result, we used field1 to convert the high-resolution HLA alleles into the HLA superalleles (low-resolution HLA alleles) (F1). Here, 121 superalleles were created out of 367 alleles. 60 and 61 of the 121 superalleles fall into classes I and II, respectively. Additionally, on the basis of statistical test we divide the HLA superalleles into two categories, SF and SU. We observed only 24 HLA-superalleles are significantly impact the survival where 9 were SF and 15 were SU (Table 4.2).

**Table 4.2: List of 9 favourable and 15 unfavourable HLA-alleles which play significant role in the survival of skin cancer patients**

| HLA-alleles | No. of Samples | | Mean Overall Survival (OS) | | Mean Diff OS (P-A) | P-value |
|---|---|---|---|---|---|---|
| | Present (P) | Absent (A) | Present (P) | Absent (A) | | |
| Survival Favourable HLA alleles | | | | | | |
| HLA-B*55 | 16 | 385 | 94.58 | 58.46 | 36.12 | 0.002 |
| HLA-DPB1*01 | 34 | 367 | 87.51 | 57.34 | 30.17 | 6.82E-07 |
| HLA-B*08 | 80 | 321 | 81.09 | 54.62 | 26.47 | 6.36E-14 |
| HLA-DRB1*03 | 85 | 316 | 80.14 | 54.46 | 25.69 | 2.29E-14 |
| HLA-B*49 | 11 | 390 | 77.87 | 59.39 | 18.48 | 0.037 |
| HLA-A*01 | 115 | 286 | 72.88 | 54.68 | 18.2 | 1.24E-17 |
| HLA-C*05 | 61 | 340 | 72.74 | 57.6 | 15.15 | 1.82E-12 |
| HLA-DPB1*10 | 16 | 385 | 72.87 | 59.36 | 13.51 | 0.0004 |
| HLA-C*07 | 217 | 184 | 66.01 | 52.7 | 13.31 | 3.65E-31 |
| Survival Unfavourable HLA alleles | | | | | | |
| HLA-B*14 | 27 | 374 | 48.34 | 60.74 | –12.39 | 2.20E-05 |
| HLA-A*24 | 81 | 320 | 48.59 | 62.77 | –14.18 | 5.61E-13 |
| HLA-DPB1*05 | 17 | 384 | 46.26 | 60.51 | –14.25 | 0.001 |
| HLA-A*31 | 26 | 375 | 46.34 | 60.84 | –14.5 | 1.76E-05 |
| HLA-DPB1*11 | 10 | 391 | 45.32 | 60.27 | –14.95 | 0.003 |
| HLA-DRB1*07 | 103 | 298 | 48.37 | 63.89 | –15.51 | 4.31E-14 |
| HLA-DPB1*06 | 12 | 389 | 43.68 | 60.4 | –16.72 | 0.014 |
| HLA-C*14 | 10 | 391 | 43.44 | 60.32 | –16.88 | 0.003 |
| HLA-B*18 | 39 | 362 | 44.41 | 61.57 | –17.16 | 1.07E-08 |
| HLA-C*01 | 42 | 359 | 44.35 | 61.72 | –17.37 | 9.08E-07 |
| HLA-B*13 | 19 | 382 | 41.94 | 60.79 | –18.86 | 0.03 |
| HLA-A*30 | 26 | 375 | 42.14 | 61.13 | –19 | 5.22E-06 |
| HLA-DRB1*16 | 23 | 378 | 29.53 | 61.75 | –32.22 | 7.00E-06 |
| HLA-B*50 | 12 | 389 | 25.03 | 60.98 | –35.95 | 6.33E-05 |
| HLA-DRB1*12 | 19 | 382 | 23.46 | 61.71 | –38.26 | 9.43E-05 |

### *4.3.3 Univariate survival analysis*

We first performed univariate survival analysis using the HLA-alleles, superalleles and clinical characteristics. We identified certain alleles/superalleles which had the significant impact on the survival of SKCM patients. For instance, presence of HLA-B∗50 alleles associated with the poor survival rate with an HR of 2.77 (95% CI 1.284 to 5.941) and p-value 0.009. In addition, HLA-DRB1∗12 reduces the survival rate with HR 3.13 (95% CI 1.687–5.826) and p-value < 0.001. We also identified the combined effect of both HLA-B∗50/DRB1*12 and observed that the patients are at high-risk with HR 3.15, 95% (CI 1.906–5.194) and significant p-value. In addition age, gender, tumour stage, tumour status, and Breslow depth are clinical and demographic characteristics that have historically demonstrated a considerable impact on the prevalence of skin cancer and a bias against a certain population. We investigated the relationship between these clinical characteristics and patient survival. We therefore used these clinical and demographic characteristics in a univariate survival analysis. According to this investigation, the tumour status is a key prognostic factor in the estimation of melanoma patients' survival times. Here, we have achieved maximum HR of 8.293 with p-value< 0.0001. Additional characteristics that have a strong correlation with patient prognosis include age, tumour stage, and Breslow depth. However, depending on gender, samples cannot be divided into high-risk and low-risk groups (See Figure 4.4).

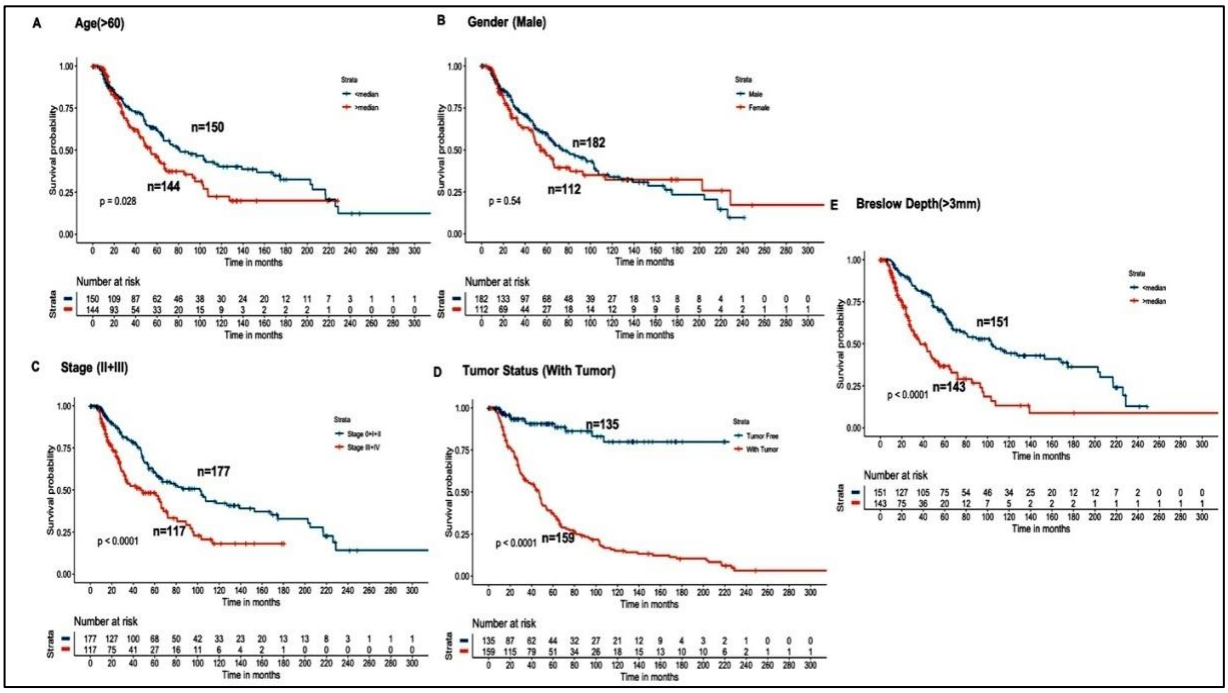



**Figure 4.3 Survival curves for risk estimation using clinical characteristics - Adopted from (Dhall et al., 2020)**

HLA superalleles that play a substantial influence in the prognosis of melanoma patients have been discovered from the aforementioned univariate analysis. The creation of prediction methods was our next objective, using them as features. Therefore, using RS, which was created by combining several HLA superalleles, we created a threshold-based technique. A survival analysis was run using this RS as an input feature to evaluate how well RS based on several superalleles categorised risk-groups of cutaneous melanoma patients. As shown in Figure 4.5, the patients are significantly split into high-risk and low-risk groups if the threshold value is 2, with HR 2.18 (95% CI 1.441-3.297) and p-value of 0.000223. Finally, we discovered that RS thresholds can be served as a prognostic indicator as shown in Figure 4.4, which was further used to divide melanoma patients into high-risk and low-risk categories. Additionally, KM survival plots indicate how melanoma patients are divided into risk groups based on various RS threshold values (shown in Figure 4.4).

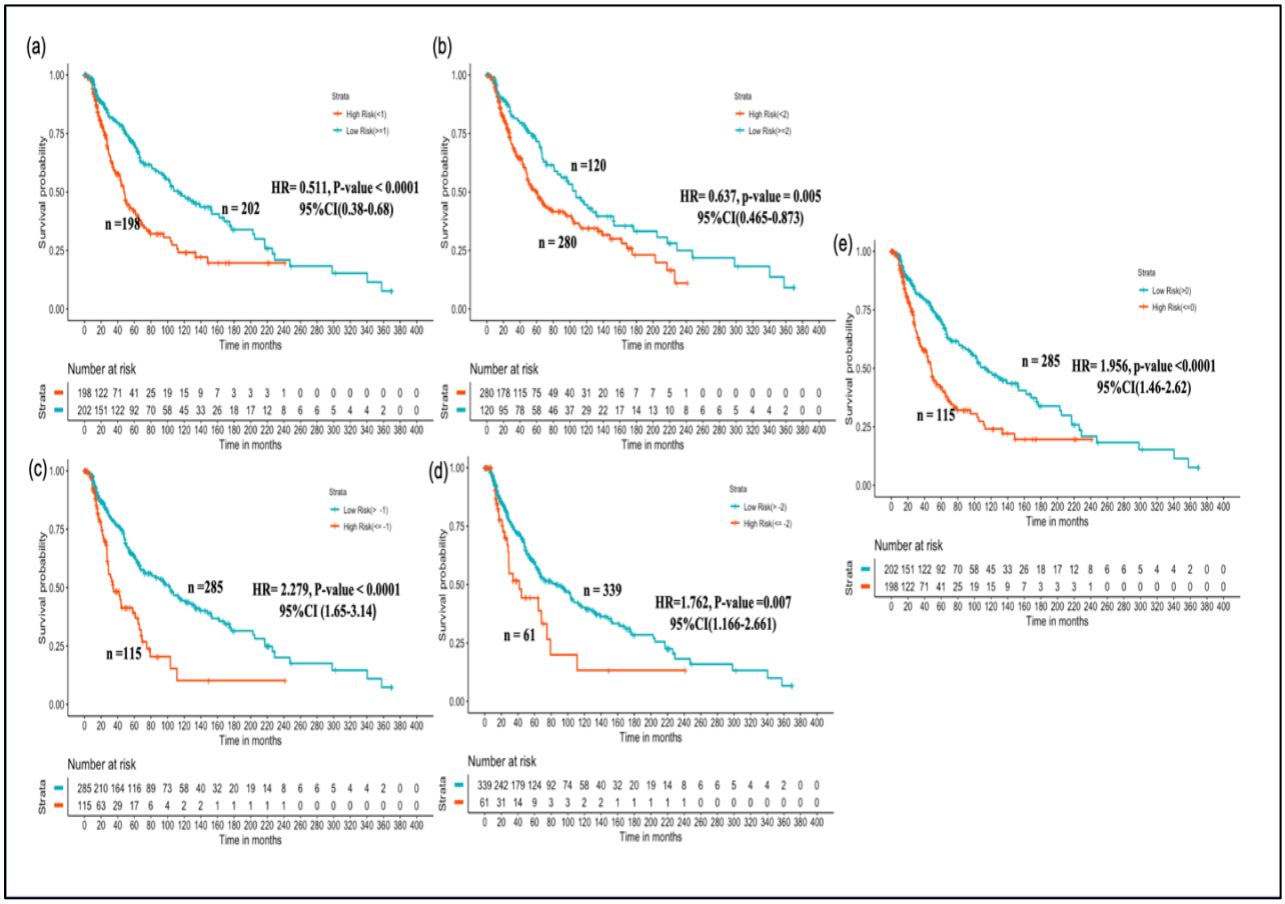



**Figure 4.4 Kaplan Meier survival curves for the risk estimation of melanoma patient cohort based on the Risk score (RS) - Adopted from (Dhall et al., 2020)**

## *4.3.4 Performance-based on prediction models*

The above mentioned results demonstrate that in order to identify high-risk patients, HLA superalleles, clinical, and demographic characteristics (such as age, gender, tumour stage, tumour status, and Breslow depth) are crucial. The threshold-based approach, however, is straightforward yet ineffective when numerous indicators are present. So, in order to further enhance performance, we developed prediction models using a variety of machine learning techniques (such as lasso, RF, ridge, and DT). To create machine learning models, we have developed various feature sets as shown in Table 4.3.

**Table 4.3: The performance of machine learning based models developed using different set of features**

| Feature Type | LASSO | | RIDGE | | Random Forest | | Decision Tree | |
|---|---|---|---|---|---|---|---|---|
| | HR | P-value | HR | P-value | HR | P-value | HR | P-value |
| All clinical features | 3.17 | 3.50E-11 | 3.01 | 1.76E-10 | 3.09 | 2.87E-11 | 2.25 | 6.93E-07 |
| Clinical features without tumor status | 3.5 | 3.93E-13 | 3.49 | 3.93E-13 | 3.74 | 3.01E-14 | 2.15 | 2.24E-06 |
| Clinical features without tumor stage | 2.8 | 9.96E-10 | 2.43 | 4.68E-08 | 2.81 | 2.05E-10 | 2.5 | 1.64E-08 |
| Clinical features without tumor stage and tumor status | 2.4 | 4.41E-08 | 2.4 | 4.41E-08 | 2.99 | 9.37E-12 | 2.54 | 1.06E-08 |

In order to avoid over-optimization and for practical implementation in daily life, it is crucial to have a minimal amount of features. Therefore, wrapper method was applied to iteratively reduce the number of characteristics. Finally, using various machine learning technique, prediction models were created utilising five clinical and demographic characteristics (age, gender, tumour stage, tumour status, and Breslow depth) and various HLA superalleles. The LASSO technique, based on five clinical characteristics and 14 HLA alleles (HLA-A*31, HLA-A*24, HLA-DPB1*10, HLA-B*08, HLA-DRB1* 03, HLA-DRB1*07, HLA-B*18, HLA-B*55, HLA-A*01, HLA-C* 05, HLA-DRB1*16, HLA-DRB1*12, HLA-B*49, HLA-DPB1*11, achieved highest performance, with an HR of 4.52 and a p-value of 8.01E-15.

## 4.4 Utility of webserver

We created the "SKCMhrp" web server to support the scientific community, available at https://webs.iiitd.edu.in/raghava/skcmhrp/ . HTML, PHP 5.2.9, and JAVA scripts were used to create

the “SKCMhrp” web server. We used an HTML5 web template to make the website mobile and tablet friendly. The technologies described above that have been used are open source and cross-platform. The goal of SKCMhrp is to estimate risk using clinical, demographic, and HLA superalleles data. The two modules are based on clinical characteristics and superalleles, respectively. Based on their clinical and demographic parameters, such as age, gender, tumour stage, tumour status, and Breslow depth, the first module forecasts the risk status of melanoma patients. By selecting just one clinical parameter, a user can forecast the particular sample’s survival time (in months) in this case (See Figure 4.5). A regression model receives input values to assess the risk status. The second module uses all 121 superalleles and 14 superalleles with five clinical and demographic characteristics to determine the risk status of melanoma patients (See Figure 4.6).

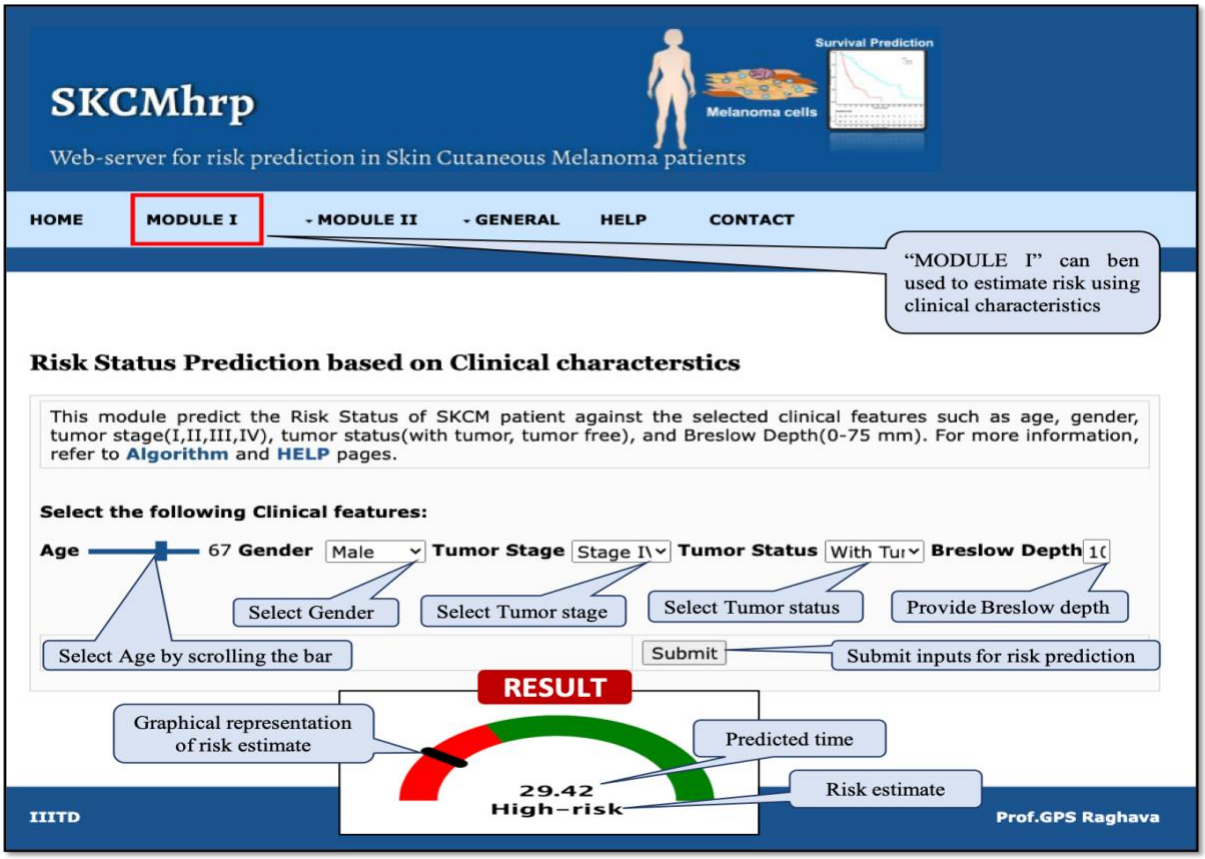



**Figure 4.5 Utility of Module I of SKCMhrp server**

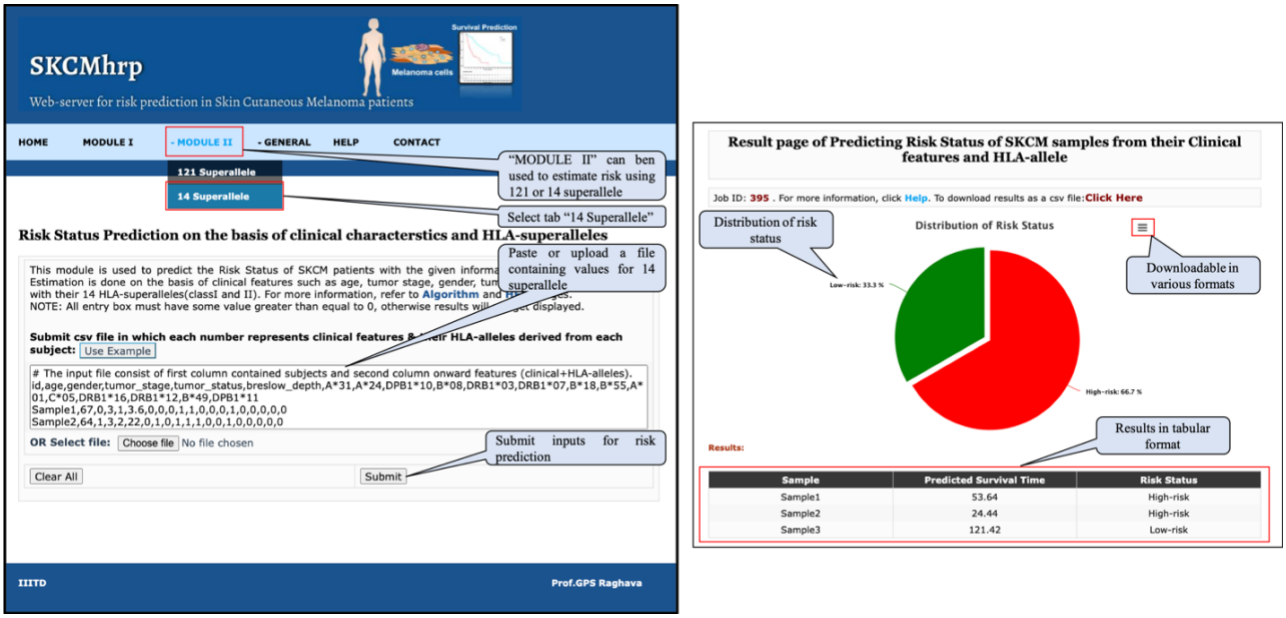



**Figure 4.6 Utility of Module II of SKCMhrp server**

## *4.5 Discussion*

The growth in melanoma incidence indicates that skin cutaneous melanoma is a fatal cancer. Over the past few years, the FDA (Food and Drug Administration) has approved a number of treatments and preventative measures for melanoma. However, information regarding the tumour, such as its location, stage, etc., is necessary before selecting a treatment among the available possibilities. It might be difficult to accurately and precisely identify the tumour stage in many cancers. According to recent research, antigenic repertoire diversity plays a critical role in tumour development and immunosurveillance. For instance, it has been demonstrated that HLA-class I and II proteins have a crucial role in the development of melanoma. It is crucial to comprehend which specific HLA alleles from class I and II may have an impact on the patients' prognosis. In order to better understand how class-I/II alleles affect melanoma patients' prognoses, the current study is an organised effort. Studies revealed that HLA-DRB1*07 has been demonstrated to be unfavourably correlated with patient survival in additional cancers, including lung cancer, cervical cancer, and breast cancer. HLA-A*01, HLA-C*05, and HLA-C*07 have been demonstrated in the literature to be favourably linked with the survival of melanoma patients. However, HLA-A*31, HLA-B*14, HLA-C*14, HLA-A*24, and HLA-B*13, have a negative correlation with melanoma patients' survival rates. In this study, we understand the impact of HLA-alleles and clinical characteristics on the survival of skin cancer patients. Overall, our results demonstrate that HLA-class I and II alleles have both positive and negative effects on the OS of TCGA-SKCM patients. The categorization of high-risk and low-risk survival groups and the calculation of OS time using survival analysis and recursive machine learning

regression models indicated the prognostic significance of 14 HLA-A*31, HLA-A*24, HLA-DPB1*10, HLA-B*08, HLA-DRB1* 03, HLA-DRB1*07, HLA-B*18, HLA-B*55, HLA-A*01, HLA-C*05, HLA-DRB1*16, HLA-DRB1*12, HLA-B*49, HLA-DPB1*11 superalleles, clinical, and demographic variables. We have created a website named “SKCMhrp” to help the scientific community predict high-risk patients.

## *4.6 Conclusion*

In this study, we have developed a survival prediction method based on Class-I & Class-II HLA-alleles and clinical characteristics. HLA-based markers may be taken into account for creating tailored vaccinations for a number of clinical populations. The further investigation regarding the role of these superalleles in additional cohorts will help to further confirm this for clinical utility.

# CHAPTER 5

## NON-CLASSICAL HLA-BINDER PREDICTION

## 5.1 Introduction

Our immune system depends on human leukocyte antigens (HLAs), which are expressed on cell surfaces for antigen presentation and to elicit immunological responses (Chaplin, 2010; Marshall et al., 2018). The most polymorphic genomic region of the human genome is the major histocompatibility complex, or HLA, which is found at chromosome 6 (6p21.3) in humans (Beck & Trowsdale, 2000; Choo, 2007). According to the IMGT/HLA database, 2020 edition (Robinson et al., 2020), more than 23000 class-I and 8600 class-II HLA alleles have previously been recorded in various ethnic groups worldwide. The two main groupings of HLA class-I genes are classical (HLA-A, -B, -C) and non-classical (HLA-G, -E, -F). The classical genes induce CD8+ T cells to produce an immunological response by presenting antigenic peptide ligands on infected cells. Contrarily, non-classical class-I alleles control the immune response by activating/inhibiting CD8+ T cells and natural killer cells (Uzhachenko & Shanker, 2019). By triggering and controlling immunological responses, HLA alleles defend humans against a number of diseases (Blackwell et al., 2009; Crux & Elahi, 2017; Tavasolian et al., 2020). At the same time, negative consequences like the onset of autoimmune diseases, the growth of cancer, the advancement of metastases, and poor prognosis have been observed in a variety of ethnic groups (Aptsiauri et al., 2007; Johansen et al., 2016; Mendez et al., 2009; Sabapathy & Nam, 2008).

Recent research indicates that both the innate and adaptive immune systems are modulated by the non-classical alleles (HLA-G and HLA-E) (Amiot et al., 2014; Crux & Elahi, 2017; Murdaca et al., 2016; Rouas-Freiss et al., 1999) (See Figure 5.1). It is noteworthy that HLA-G has four membrane-bound isoforms and three soluble isoforms, and that they interact with the natural killer cell receptors (NKG2A/CD94), killer cell immunoglobulin-like receptor (KIR2DL4), and immunoglobulin-like transcript (ILT2 and ILT4) (Ho et al., 2020; Rizzo et al., 2013; Tronik-Le Roux et al., 2017). Until recently, scientists thought that HLA-G alleles could only be detected at the maternal-fetal interface. But according to current research, the expression of HLA-G is noticeably higher in a number of illness conditions, including cancer, COVID-19 infection, auto-immune, and inflammatory diseases (Amiot et al., 2011; Carosella et al., 2011; Kovats et al., 1990; Schmidt & Orr, 1993; Shih Ie, 2007; Zidi, 2020). HLA-G also prevents the activation of immune cells such as CD8+ T, dendritic, and natural killer cells during parasitic and viral infections (including those caused by the influenza A virus, herpes, and coronavirus) (Catamo et al., 2014; Dias et al., 2015; Sabbagh et al., 2018). These viral infections increase HLA-G expression and create an environment that is tolerant to the immune system. On the other hand, HLA-E has little variation and is linked to highly conserved peptides and

epitopes. Through interactions with inhibitory receptors (NKG2A/CD94, NKG2B/CD94, and activating receptor (NKG2C/CD94), HLA-E controls immune cells (natural killer and cytotoxic T cells) (Kraemer et al., 2014).

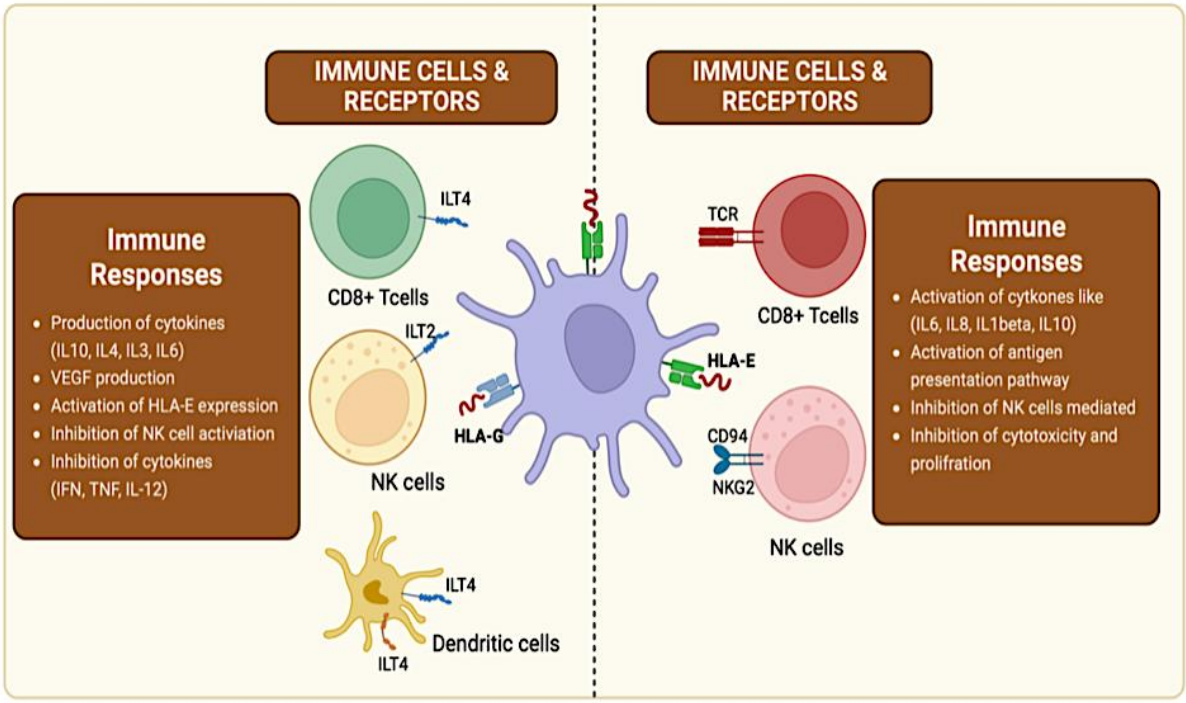



**Figure 5.1 Representation of non-classical HLA with their immunoregulatory functions**

HLA-E alleles, also control cell fate via representing antigens through two recognised processes. Peptide fragments derived from the signal sequence of other class Ia HLA-alleles. By interacting with the NKG2A/CD94 receptors, this representation inhibits the activity of NK cells. Some research, however, has shown that the viral peptides (including those from SARS-CoV-2, Epstein-Barr virus, cytomegalovirus, and hepatitis C virus) are presented by HLA-E on the cell surface and recognised by virus-specific immune cells, which further activates the immune responses (Crew et al., 2005; Garcia et al., 2002; Joosten et al., 2016; Pietra et al., 2003; Romagnani et al., 2004; Romagnani et al., 2002). The production of anti-inflammatory cytokines including transforming growth factor (TGF-β), interleukin 4 (IL4), and interleukin 10 (IL10), which is in turn responsible for the down-regulation of pro-inflammatory cytokine production, is also a result of HLA-E restricted CD8+ T-cells. It also preventing the cytokine storm, which is essential for the development of COVID-19. The level of tissue damage is also reduced by inhibiting the cytokine storm (Caccamo et al., 2020). As shown in Figure 5.1, several investigations have shown that HLA-E impairs NK-mediated lysis, cytotoxicity, cytokine production, and tumour growth. According to these findings, immunological checkpoint

molecules HLA-G and HLA-E may be crucial for developing innovative immunotherapies or subunit vaccines against a variety of disorders. Therefore, techniques for prediction of non-classical HLA binders must be developed. Although many computational techniques for predicting HLA binders have been developed in the past, they have mostly focused on classical HLAs (Chen et al., 2019; Jurtz et al., 2017; Mei et al., 2021; O'Donnell et al., 2020; Singh & Raghava, 2001, 2003; Ye et al., 2021).

Models for predicting binders for non-classical HLA alleles are only few and developed on limited set of dataset. To the best of our knowledge, no computational tool has been created specifically for identification of non-classical HLA binders. In this study, which is specifically focused on non-classical HLA, an organised effort has been made to create models for anticipating non-classical HLA binders. From the immune epitope database (IEDB), we gathered and evaluated each experimentally verified non-classical HLA binder. We created models for predicting binders for the non-classical alleles HLA-G*01:01, HLA-G*01:03, HLA-G*01:04, HLA-E*01:01 based on the experimentally validated dataset. To more accurately predict the non-classical HLA binders, we have used a variety of machine learning methods.

## 5.2 Material and methods

### 5.2.1 Dataset generation & pre-processing

We have gathered the non-classical class-I HLA-binding peptides for the current study from the IEDB, obtained on October 26, 2021. 1135 HLA-E and 5151 HLA-G binding peptides in total were obtained. Then, in order to create non-redundant datasets, we delete identical peptides from each dataset. Additionally, we removed from each dataset any peptides with a length of more than 15 or less than 8 residues. Finally, for the HLA-E*01:01 and -E*01:03 alleles, respectively, we were able to collect 142 and 723 distinct peptides. Likewise, for the HLA-G*01:01, -G*01:03, and -G*01:04 alleles, we obtain 2633, 751, and 812 distinct binding peptides, respectively. The binding peptides linked with HLA-G alleles derived from the mass spectrometry experiments. On the other hand, HLA-E alleles linked binders were primarily generated using mass spectrometry and fluorescence based (biophysical) approaches. In the case of HLA-E*01:03, the majority of the data (i.e., 632 distinct positive binders with 8–15 residues range) came from mass spectrometry. In addition, 87 peptides were produced using fluorescence-based techniques, and 4 peptides came from X-ray crystallography. We exclusively take into account mass spectrometry-derived peptides for HLA-G*01:01, -G*01:03, -G*01:04, and -E*01:03 in order to retain the homogeneity in the datasets. However, HLA-E*01:01 only has a small number of mass spectrometry-derived experimentally validated binders, thus we have taken into

account the entire dataset of 142 binding peptides (114 derived from fluorescence based and 28 peptides derived from mass spectrometry).

We randomly created the HLA-G and HLA-E non-binding peptides with lengths of 8 to 15 residues from the Swiss-Prot [54] database because to the IEDB's dearth of negative peptides (March 2021 release). In this case, we have produced two distinct datasets, one of which is balanced and contains an equal amount of negative and positive peptides for each allele. The other dataset is the unbalanced/realistic dataset, which contains ten times as much negative data as positive data.

### *5.2.2 Amino-acid composition*

To comprehend the compositional similarities in various peptide sequences, the amino acid composition (AAC) of the positive and negative dataset for each allele is computed. The AAC for binder/non-binder peptides for the HLA-G and HLA-E alleles is calculated using the following equation.

$$AAC_i = \frac{AAR_i}{Total\ number\ of\ residues} \times 100$$

where $AAC_i$ and $AAR_i$ are the percentage composition and number of residues of type i in a peptide, respectively.

### *5.2.3 Sequence logo*

With the aid of the TSL programme, we created sequence logos for each HLA-allele. In our dataset, the minimum length of peptide was eight, and hence, we created the fixed-length peptides having sixteen residues. In order to create a fixed-length vector, we picked eight residues from N-terminal and eight residues from C-terminal; further, we merged the two sequences and got the final sixteen residue peptides for each positive and negative dataset.

### *5.2.4 Binary profile generation*

To represent the amino acid sequence in the numerical vector, we have implemented the binary profile module of Pfeature (Pande et al., 2019). Binary profile is the binary representation of the sequences, where each amino acid represented by the vector of length 20. In the binary vector each position belongs to 20 different amino acids, where each element represents the presence/absence of the residues, presence was signified by "1" and absence of residues was signified by "0" at that particular

position. For instance, residue "A" was represented by the vector "1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0". In order to generate the vector of fixed length to train the models, it is important to fix the sequence length. Since, the length varies from 8 to 15, we have generated patterns N8, C8, N8C8, and AA15. In case of N8 patterns, eight residues were selected from N-terminal of the sequences, whereas in C8 patterns, eight residues were taken from the C-terminal. In case of N8C8, patterns of length 16 were generated by joining the eight residues from N- and C-terminal. Therefore, pattern N8 and C8 generated the vector size of 160 (8*20) and pattern N8C8 was represented by vector of length 320 (16*20). Similarly, patterns with length 15 (i.e. maximum length) was generated and called as AA15. In order to make up the length for sequences having length less than 15, a dummy variables "X" was padded and then the binary profile was generated. In this case, each amino acid is represented by length 21 instead of 20, where 21st element represents the presence/absence of dummy variable "X". Therefore, the generated vector for these patterns have the length of 315 (15*21).

### *5.2.5 Machine learning*

To build the prediction models to classify the peptides into non-classical HLA-binding peptides, we have implemented several machine learning classifiers using the scikit-learn library of Python. We have used Decision Tree (DT), Random Forest (RF), Support Vector Classifier (SVC), eXtreme Gradient Boosting (XGB), Gaussian Naïve Bayes (GNB), Logistic Regression (LR), K-Nearest Neighbor (KNN), and randomized Extra Tree (ET) classifier to develop the prediction models.

### *5.2.6 Cross validation technique*

To prevent bias and overfitting in the derived models, we used a 5-fold cross-validation procedure. The evaluation of the prediction model is one of the most important processes. This method divides the complete dataset into five segments, of which four are utilised for training, and the final model is tested on the left segment. Five repetitions of the precise procedure are duplicated in order to provide each component a chance to serve as the testing dataset. The average of the performances of the five models that emerged from the five iterations ultimately serves as the representation of the final performance.

### *5.2.7 Performance measures*

The performance evaluation parameters are broadly categorised into threshold-dependent and threshold-independent parameters, can be used to evaluate the prediction models. The threshold-dependent metrics in this investigation were identified as sensitivity, specificity, accuracy, F1-score, and Matthews correlation coefficient (MCC). As a threshold-independent metric, Area Under Receiver Operating Characteristics (AUC) curve is determined. Equation 1 quantifies the model's sensitivity, while equation 2 calculates its specificity, which is the proportion of correctly predicted non-binders. Equation 3 shows the percentage of binders and non-binders that were successfully predicted, equation 4 shows the balance between precision and recall, and equation 5 shows the relationship between predicted and observed values. The capacity of the model to differentiate between the classes is captured by the area under the curve (AUC), which is a plot between sensitivity and 1-specificity.

$$Sensitivity = \frac{T_P}{T_P + F_N} \quad [1]$$

$$Specificity = \frac{T_N}{T_N + F_P} \quad [2]$$

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \quad [3]$$

$$F1 - Score = \frac{2T_P}{2T_P + F_P + F_N} \quad [4]$$

$$MCC = \frac{(T_P * T_N) - (F_P * F_N)}{\sqrt{(T_P + F_P)(T_P + F_N)(T_N + F_P)(T_N + F_N)}} \quad [5]$$

Where, $T_P$, $T_N$, $F_P$ and $F_N$ stands for true positive, true negative, false positive and false negative, respectively.

## 5.3 Results

### 5.3.1 Overall study design

Figure 5.2 incorporates the overall workflow of the present study and display the collection of dataset, feature generation method, machine learning and web-server implementation.

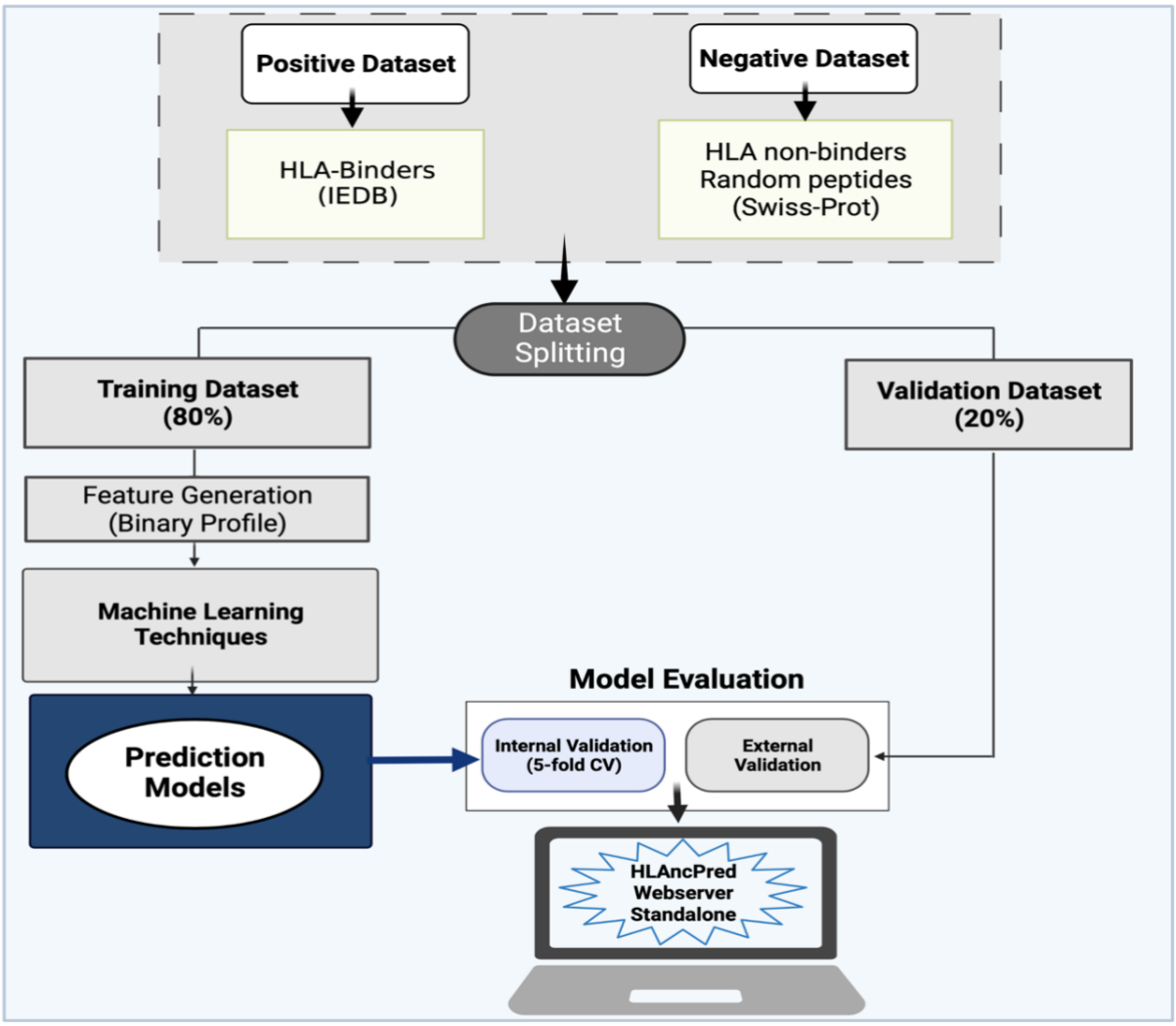
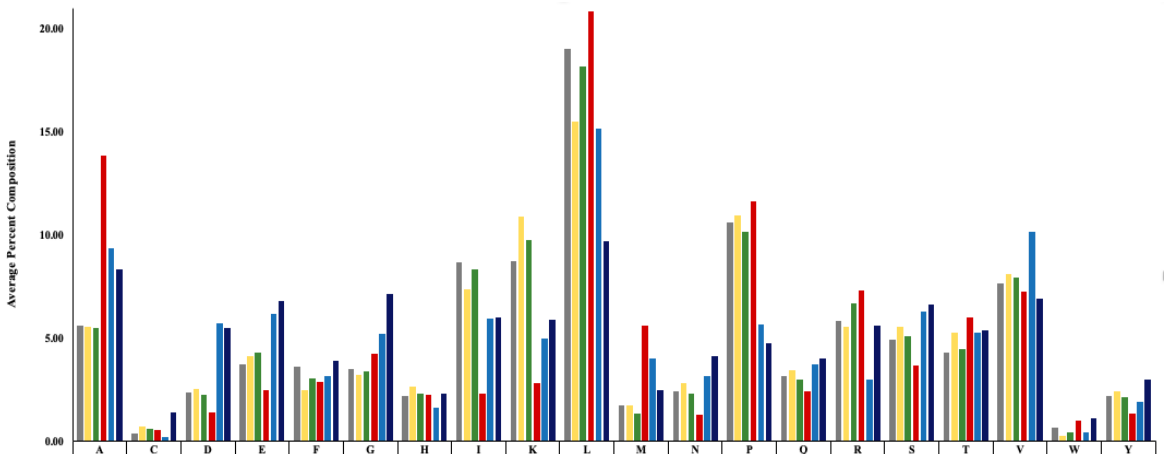



**Figure 5.2 Show the flow chart of algorithm used for the building of HLAncPred, where models are trained on training dataset and validated on independent dataset**

## *5.3.2 Amino-acid composition*

Figure 5.3 illustrates the average amino-acid composition of HLA-G and HLA-E binding and non-binding peptides. The compositional difference between the positive and negative datasets is evident in the graphs, as illustrated. Figure 5.3 shows that compared to non-binding peptides, HLA-G*01:01, -G*01:03, and -G*01:04 binders (i.e., positive peptides) have a higher composition of residues like isoleucine (I), lysine (K), leucine (L), and proline (P). In contrast to the negative dataset, HLA-E*01:01, -E*01:03 binding peptides have a larger average composition of Alanine (A), Leucine (L), Methionine (M), Proline (P), and Valine (V) residues.




| | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HLA-G*01:01 | 5.53 | 0.30 | 2.31 | 3.65 | 3.54 | 3.42 | 2.17 | 8.59 | 8.69 | 18.95 | 1.70 | 2.37 | 10.54 | 3.08 | 5.76 | 4.86 | 4.26 | 7.57 | 0.58 | 2.13 |
| HLA-G*01:03 | 5.50 | 0.65 | 2.49 | 4.05 | 2.44 | 3.16 | 2.60 | 7.30 | 10.84 | 15.46 | 1.67 | 2.75 | 10.90 | 3.41 | 5.49 | 5.47 | 5.21 | 8.06 | 0.20 | 2.36 |
| HLA-G*01:04 | 5.46 | 0.54 | 2.20 | 4.23 | 3.00 | 3.31 | 2.25 | 8.26 | 9.68 | 18.08 | 1.29 | 2.25 | 10.09 | 2.94 | 6.61 | 5.01 | 4.41 | 7.90 | 0.37 | 2.10 |
| HLA-E*01:0I | 13.76 | 0.47 | 1.33 | 2.45 | 2.80 | 4.16 | 2.21 | 2.26 | 2.74 | 20.83 | 5.56 | 1.21 | 11.58 | 2.35 | 7.23 | 3.64 | 5.97 | 7.22 | 0.94 | 1.29 |
| HLA-E*01:03 | 9.27 | 0.14 | 5.64 | 6.11 | 3.09 | 5.16 | 1.57 | 5.90 | 4.94 | 15.10 | 3.94 | 3.10 | 5.62 | 3.67 | 2.93 | 6.24 | 5.18 | 10.10 | 0.35 | 1.86 |
| General Proteome | 8.25 | 1.37 | 5.45 | 6.75 | 3.86 | 7.07 | 2.27 | 5.96 | 5.84 | 9.66 | 2.42 | 4.06 | 4.70 | 3.93 | 5.53 | 6.56 | 5.34 | 6.87 | 1.08 | 2.92 |

**Figure 5.3 Average amino acid composition of different non-classical HLA-alleles (HLA-G*01:01, HLA-G*01:03, HLA-G*01:04, HLA-E*01:01, and HLA-E*01:03) & general proteome**

## *5.3.3 Position-wise conservation*

Here, we have used two sample logo to depict the sequence logo for each non-classical HLA-allele (HLA-G*01:01, HLA-G*01:03, HLA-G*01:04, HLA-E*01:01, and HLA-E*01:03). As depicted in Figure 5.4, each logo is used to identify the conserved residues and their precise location in the nonameric sequences. In case of HLA-G alleles amino-acid residue ‘P’ is highly conserved at position P3 and ‘L’ at position P9 and display very high abundance, whereas ‘K/R’ anchor residues placed at initial anchor position (P1). When it comes to HLA-E alleles the amino-acid residues are conserved and primarily found at positions (P2 and P9) with hydrophobic residues predominating. The anchor residues for HLA-E are M/L at the second anchor position (P2) and sixteenth position (P16). However, residue ‘L’ highly conserved at the ninth position (P9) for HLA-E*01:01, and V/L for HLA-E*01:03.

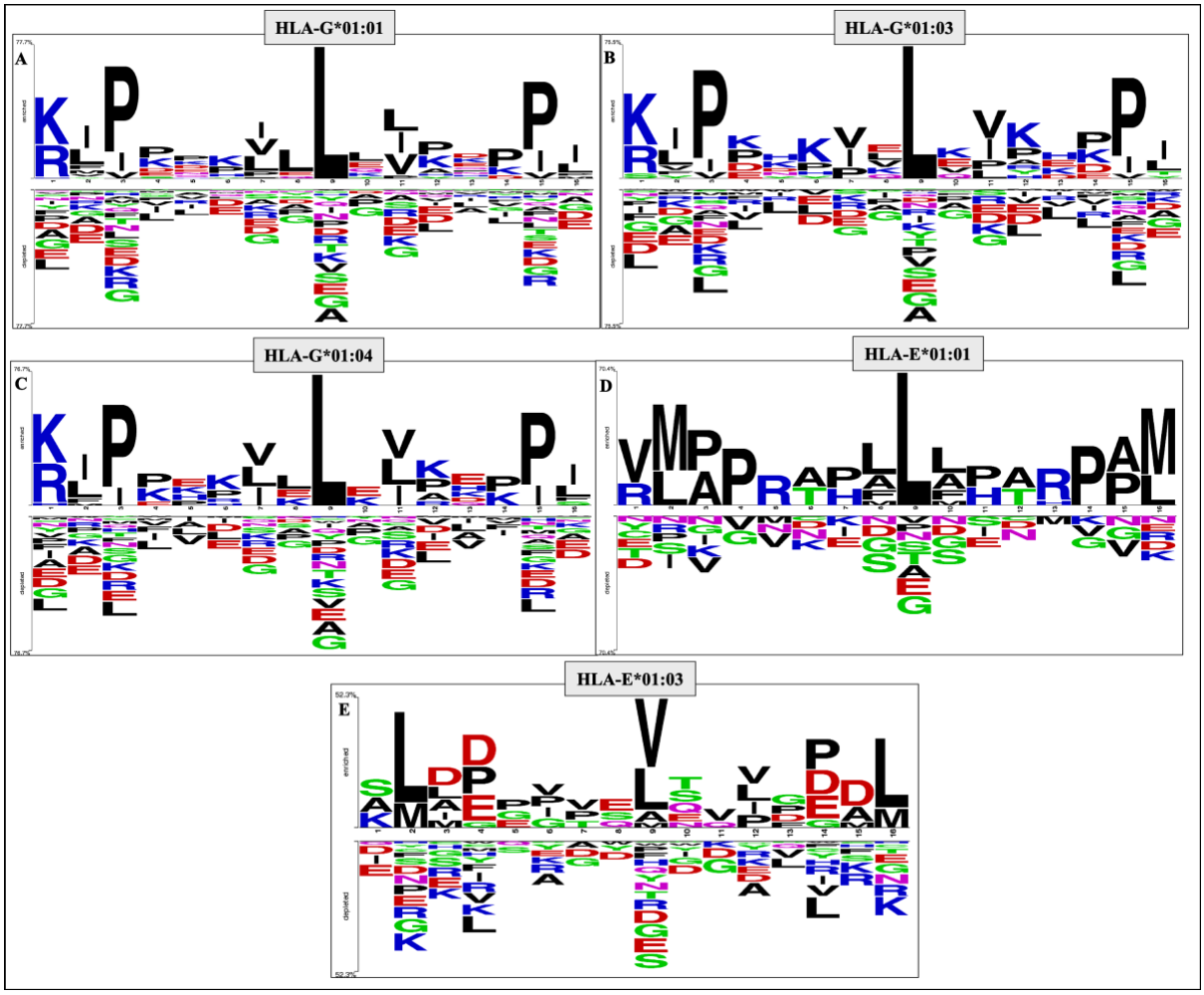



**Figure 5.4 Two sample logo generated for non-classical HLA-alleles; where, upper portion shows non-classical HLA binders and lower part shows non-binders**

## *5.3.4 Performance of classification models*

To create prediction models for this work, we used a variety of classifiers, including GNB, XGB, RF, DT, SVC, ET, KNN, and LR. For positive and negative datasets (i.e., HLA-G*01:01, -G*01:03, -G*01:04, -E*01:01, and -E*01:03 binding and non-binding peptides), we compute binary profile-based features. Using the Pfeature standalone package, we first create four feature sets (i.e., the N8, C8, N8C8, and AA15 binary profiles). Then, using each feature set for the HLA-G and HLA-E alleles, we created a number of machine learning models.

### *5.3.4.1 HLA-G based models*

We have developed various models using N8 and C8 binary profiles-based features. As shown in Table 5.1, HLA-G*01:01 and HLA-G*01:04 achieved maximum AUC of 0.98 on validation dataset using

C8 binary profiles. HLA-G*01:03 performed quite less and achieved an AUC of 0.95 on validation dataset. However, N8 based features perform less and achieved maximum AUC of 0.97, 0.93 and 0.95 for HLA-G*01:01, HLA-G*01:03 and HLA-G*01:04 alleles.

**Table 5.1: The performance of machine learning based models developed using N8 and C8 binary profile-based features of HLA-G alleles on validation datasets**

| | N8 | | | | | C8 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Name** | **Sens** | **Spec** | **Acc** | **AUC** | **MCC** | **Sens** | **Spec** | **Acc** | **AUC** | **MCC** |
| **HLA-G*01:01** | | | | | | | | | | |
| **DT** | 83.11 | 82.92 | 83.02 | 0.90 | 0.66 | 89.18 | 88.62 | 88.90 | 0.93 | 0.78 |
| **RF** | 89.94 | 92.03 | 90.99 | 0.96 | 0.82 | 92.60 | 92.79 | 92.69 | 0.98 | 0.85 |
| **LR** | 89.94 | 92.60 | 91.27 | 0.95 | 0.83 | 93.55 | 91.65 | 92.60 | 0.97 | 0.85 |
| **XGB** | 90.32 | 92.03 | 91.18 | 0.96 | 0.82 | 92.98 | 92.79 | 92.88 | 0.98 | 0.86 |
| **KNN** | 87.67 | 91.46 | 89.56 | 0.95 | 0.79 | 92.79 | 92.03 | 92.41 | 0.97 | 0.85 |
| **GBM** | 90.32 | 80.65 | 85.48 | 0.92 | 0.71 | 88.05 | 92.79 | 90.42 | 0.94 | 0.81 |
| **ET** | 91.08 | 91.84 | 91.46 | 0.97 | 0.83 | 93.17 | 92.98 | 93.07 | 0.98 | 0.86 |
| **SVC** | 90.13 | 93.17 | 91.65 | 0.96 | 0.83 | 94.12 | 92.98 | 93.55 | 0.98 | 0.87 |
| **HLA-G*01:03** | | | | | | | | | | |
| **DT** | 74.67 | 74.83 | 74.75 | 0.83 | 0.50 | 89.33 | 79.47 | 84.39 | 0.93 | 0.69 |
| **RF** | 86.00 | 92.05 | 89.04 | 0.93 | 0.78 | 88.67 | 95.36 | 92.03 | 0.95 | 0.84 |
| **LR** | 86.00 | 92.05 | 89.04 | 0.93 | 0.78 | 88.67 | 92.72 | 90.70 | 0.95 | 0.82 |
| **XGB** | 81.33 | 90.73 | 86.05 | 0.93 | 0.72 | 90.00 | 91.39 | 90.70 | 0.95 | 0.81 |
| **KNN** | 85.33 | 92.72 | 89.04 | 0.93 | 0.78 | 88.67 | 93.38 | 91.03 | 0.95 | 0.82 |
| **GBM** | 88.67 | 76.82 | 82.72 | 0.84 | 0.66 | 86.67 | 84.11 | 85.38 | 0.86 | 0.71 |
| **ET** | 86.00 | 93.38 | 89.70 | 0.93 | 0.80 | 90.00 | 94.04 | 92.03 | 0.95 | 0.84 |
| **SVC** | 86.00 | 94.04 | 90.03 | 0.94 | 0.80 | 90.67 | 92.05 | 91.36 | 0.95 | 0.83 |
| **HLA-G*01:04** | | | | | | | | | | |
| **DT** | 75.93 | 79.76 | 77.85 | 0.85 | 0.56 | 85.185 | 84.663 | 84.923 | 0.891 | 0.698 |
| **RF** | 92.59 | 88.34 | 90.46 | 0.95 | 0.81 | 96.296 | 96.319 | 96.308 | 0.98 | 0.926 |
| **LR** | 91.98 | 86.50 | 89.23 | 0.95 | 0.79 | 95.679 | 94.479 | 95.077 | 0.975 | 0.902 |
| **XGB** | 90.12 | 87.73 | 88.92 | 0.94 | 0.78 | 96.296 | 95.092 | 95.692 | 0.978 | 0.914 |
| **KNN** | 91.36 | 89.57 | 90.46 | 0.95 | 0.81 | 96.296 | 95.706 | 96 | 0.976 | 0.92 |
| **GBM** | 82.72 | 87.12 | 84.92 | 0.88 | 0.70 | 91.975 | 86.503 | 89.231 | 0.902 | 0.786 |
| **ET** | 92.59 | 88.96 | 90.77 | 0.95 | 0.82 | 96.296 | 96.933 | 96.615 | 0.978 | 0.932 |
| **SVC** | 91.36 | 84.66 | 88.00 | 0.95 | 0.76 | 96.296 | 95.092 | 95.692 | 0.976 | 0.914 |

**#DT, Decision tree; GNB, Gaussian Naive Bayes; KNN, k-nearest neighbor; LR, Logistic Regression; RF, Random Forest; XGB, *XGBoost*; Sens, ET, Extra Tree; SVC, Support vector classifier; Sensitivity; Spec, Specificity; Acc, Accuracy; AUROC, Area Under Receiver Operating Curve**

We found that models based on the AA15 binary profile perform better than others with balanced sensitivity and specificity. According to Table 5.2, the HLA-G*01:01 dataset had an accuracy of more than 95% on both the training and validation datasets, with a maximum AUC of 0.99. On training and validation datasets, ET-based models exhibit comparable outcomes, with an AUC of 0.99 and accuracy greater than 95%. (Table 4.2). On the HLA-G*01:03 dataset, the XGB classifier performs similarly, with a maximum AUC of 0.98 and accuracy of 91.69%. On the HLA-G*01:04 dataset, however, the performance of the RF, ET, and SVC classifiers surpasses that of the other models.

**Table 5.2: The performance of machine learning based models developed using N8C8 and AA15 binary profile-based features of HLA-G alleles on validation datasets**

| | N8C8 | | | | | AA15 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Name** | **Sens** | **Spec** | **Acc** | **AUC** | **MCC** | **Sens** | **Spec** | **Acc** | **AUC** | **MCC** |
| **HLA-G*01:01** | | | | | | | | | | |
| **DT** | 87.86 | 82.16 | 85.01 | 0.90 | 0.70 | 89.37 | 89.75 | 89.56 | 0.94 | 0.79 |
| **RF** | 94.31 | 95.26 | 94.78 | 0.98 | 0.90 | 93.93 | 95.83 | 94.88 | 0.98 | 0.9 |
| **LR** | 94.12 | 94.50 | 94.31 | 0.98 | 0.89 | 92.79 | 95.07 | 93.93 | 0.98 | 0.88 |
| **XGB** | 94.12 | 95.45 | 94.78 | 0.98 | 0.90 | 94.12 | 92.98 | 93.55 | 0.98 | 0.87 |
| **KNN** | 93.17 | 93.93 | 93.55 | 0.98 | 0.87 | 91.27 | 94.12 | 92.69 | 0.97 | 0.85 |
| **GBM** | 92.41 | 92.79 | 92.60 | 0.96 | 0.85 | 91.08 | 86.34 | 88.71 | 0.9 | 0.78 |
| **ET** | 95.07 | 95.83 | 95.45 | 0.98 | 0.91 | 93.93 | 96.02 | 94.97 | 0.99 | 0.9 |
| **SVC** | 94.88 | 95.45 | 95.16 | 0.98 | 0.90 | 94.5 | 95.83 | 95.16 | 0.99 | 0.9 |
| **HLA-G*01:03** | | | | | | | | | | |
| **DT** | 78.67 | 93.38 | 86.05 | 0.90 | 0.73 | 80.67 | 82.78 | 81.73 | 0.88 | 0.64 |
| **RF** | 88.00 | 96.69 | 92.36 | 0.96 | 0.85 | 87.33 | 94.04 | 90.7 | 0.97 | 0.82 |
| **LR** | 90.00 | 96.03 | 93.02 | 0.96 | 0.86 | 88 | 94.7 | 91.36 | 0.97 | 0.83 |
| **XGB** | 89.33 | 95.36 | 92.36 | 0.96 | 0.85 | 90 | 93.38 | 91.69 | 0.98 | 0.83 |
| **KNN** | 88.67 | 96.69 | 92.69 | 0.95 | 0.86 | 85.33 | 95.36 | 90.37 | 0.94 | 0.81 |
| **GBM** | 88.67 | 78.81 | 83.72 | 0.85 | 0.68 | 91.33 | 65.56 | 78.41 | 0.78 | 0.59 |
| **ET** | 89.33 | 96.03 | 92.69 | 0.96 | 0.86 | 89.33 | 94.04 | 91.69 | 0.97 | 0.84 |
| **SVC** | 88.67 | 96.69 | 92.69 | 0.96 | 0.86 | 88 | 95.36 | 91.69 | 0.97 | 0.84 |
| **HLA-G*01:04** | | | | | | | | | | |
| **DT** | 85.19 | 84.66 | 84.92 | 0.89 | 0.70 | 86.42 | 76.69 | 81.54 | 0.87 | 0.63 |
| **RF** | 96.30 | 96.32 | 96.31 | 0.98 | 0.93 | 96.3 | 93.87 | 95.08 | 0.98 | 0.9 |
| **LR** | 95.68 | 94.48 | 95.08 | 0.98 | 0.90 | 96.3 | 92.64 | 94.46 | 0.98 | 0.89 |
| **XGB** | 96.30 | 95.09 | 95.69 | 0.98 | 0.91 | 95.06 | 94.48 | 94.77 | 0.98 | 0.9 |
| **KNN** | 96.30 | 95.71 | 96.00 | 0.98 | 0.92 | 95.06 | 90.8 | 92.92 | 0.97 | 0.86 |
| **GBM** | 91.98 | 86.50 | 89.23 | 0.90 | 0.79 | 93.21 | 67.49 | 80.31 | 0.8 | 0.63 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **ET** | 96.30 | 96.93 | 96.62 | 0.98 | 0.93 | 96.3 | 93.87 | 95.08 | 0.98 | 0.9 |
| **SVC** | 96.30 | 95.09 | 95.69 | 0.98 | 0.91 | 96.91 | 93.87 | 95.39 | 0.98 | 0.91 |

**#DT, Decision tree; GNB, Gaussian Naive Bayes; KNN, k-nearest neighbor; LR, Logistic Regression; RF, Random Forest; XGB, *XGBoost*; Sens, ET, Extra Tree; SVC, Support vector classifier; Sensitivity; Spec, Specificity; Acc, Accuracy; AUROC, Area Under Receiver Operating Curve**

### *5.3.4.2 HLA-E based models*

To achieve this, we created a number of prediction models using both positive and negative datasets for the HLA-E*01:01 and -E*01:03 alleles. As shown in Table 5.3, we have achieved maximum AUC of 0.90 and 0.89 using RF-based classifier on N8 and C8 binary profile-based features in the case of HLA-E*01:01. However, we observer a significant difference in the sensitivity and specificity.

**Table 5.3: The performance of machine learning based models developed using N8 and C8 binary profile-based features of HLA-E alleles on validation datasets**

| | **N8** | | | | | **C8** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Name** | **Sens** | **Spec** | **Acc** | **AUC** | **MCC** | **Sens** | **Spec** | **Acc** | **AUC** | **MCC** |
| **HLA-E*01:01** | | | | | | | | | | |
| **DT** | 89.29 | 79.31 | 84.21 | 0.82 | 0.69 | 61.54 | 92.59 | 77.36 | 0.81 | 0.57 |
| **RF** | 85.71 | 68.97 | 77.19 | 0.90 | 0.55 | 76.92 | 85.19 | 81.13 | 0.89 | 0.62 |
| **LR** | 82.14 | 68.97 | 75.44 | 0.86 | 0.52 | 84.62 | 77.78 | 81.13 | 0.87 | 0.63 |
| **XGB** | 78.57 | 79.31 | 78.95 | 0.86 | 0.58 | 80.77 | 85.19 | 83.02 | 0.88 | 0.66 |
| **KNN** | 75.00 | 79.31 | 77.19 | 0.87 | 0.54 | 73.08 | 88.89 | 81.13 | 0.85 | 0.63 |
| **GBM** | 92.86 | 55.17 | 73.68 | 0.74 | 0.52 | 80.77 | 44.44 | 62.26 | 0.63 | 0.27 |
| **ET** | 82.14 | 75.86 | 78.95 | 0.91 | 0.58 | 69.23 | 85.19 | 77.36 | 0.86 | 0.55 |
| **SVC** | 82.14 | 79.31 | 80.70 | 0.90 | 0.62 | 0.00 | 100.00 | 50.94 | 0.15 | 0.00 |
| **HLA-E*01:03** | | | | | | | | | | |
| **DT** | 70.35 | 70.35 | 70.35 | 0.76 | 0.41 | 89.29 | 68.97 | 78.95 | 0.77 | 0.59 |
| **RF** | 77.93 | 78.62 | 78.28 | 0.86 | 0.57 | 82.14 | 86.21 | 84.21 | 0.95 | 0.68 |
| **LR** | 76.55 | 75.17 | 75.86 | 0.84 | 0.52 | 85.71 | 96.55 | 91.23 | 0.98 | 0.83 |
| **XGB** | 73.10 | 78.62 | 75.86 | 0.83 | 0.52 | 96.43 | 86.21 | 91.23 | 0.97 | 0.83 |
| **KNN** | 73.10 | 78.62 | 75.86 | 0.82 | 0.52 | 82.14 | 82.76 | 82.46 | 0.90 | 0.65 |
| **GBM** | 82.07 | 62.76 | 72.41 | 0.74 | 0.46 | 71.43 | 82.76 | 77.19 | 0.77 | 0.55 |
| **ET** | 75.86 | 78.62 | 77.24 | 0.86 | 0.55 | 92.86 | 79.31 | 85.97 | 0.95 | 0.73 |
| **SVC** | 74.48 | 77.24 | 75.86 | 0.84 | 0.52 | 89.29 | 82.76 | 85.97 | 0.96 | 0.72 |

**#DT, Decision tree; GNB, Gaussian Naive Bayes; KNN, k-nearest neighbor; LR, Logistic Regression; RF, Random Forest; XGB, *XGBoost*; Sens, ET, Extra Tree; SVC, Support vector classifier; Sensitivity; Spec, Specificity; Acc, Accuracy; AUROC, Area Under Receiver Operating Curve**

In order to improve the performance, we have computed performances on N8C8 and AA15 binary profile-based features. As shown in the results of the preceding section, binary profile-based features outperform other classifiers for this dataset AA15. According to Table 5.4, ET-based models outperform other classifiers for the HLA-E*01:01 allele, with accuracy values of 87.67% and 89.47% and an AUC of 0.96 on the training and validation datasets, respectively. Additionally, models based on RF and XGB function admirably with balanced sensitivity and specificity (Table 5.4). However, SVC worked admirably on the HLA-E*01:03 dataset, with AUC of 0.93 and 0.94; accuracy of 84.08% and 84.98%, respectively, on the training and validation dataset, as indicated in Table 5.4.

**Table 5.4: The performance of machine learning based models developed using N8C8 and AA15 binary profile-based features of HLA-E alleles on validation datasets**

| | N8C8 | | | | | AA15 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Name** | **Sens** | **Spec** | **Acc** | **AUC** | **MCC** | **Sens** | **Spec** | **Acc** | **AUC** | **MCC** |
| **HLA-E*01:01** | | | | | | | | | | |
| **DT** | 68.28 | 81.38 | 74.83 | 0.79 | 0.50 | 75 | 75.86 | 75.44 | 0.81 | 0.51 |
| **RF** | 80.69 | 83.45 | 82.07 | 0.90 | 0.64 | 89.29 | 86.21 | 87.72 | 0.96 | 0.76 |
| **LR** | 80.69 | 80.69 | 80.69 | 0.89 | 0.61 | 85.71 | 89.66 | 87.72 | 0.97 | 0.76 |
| **XGB** | 82.07 | 81.38 | 81.72 | 0.90 | 0.63 | 89.29 | 86.21 | 87.72 | 0.96 | 0.76 |
| **KNN** | 78.62 | 80.00 | 79.31 | 0.86 | 0.59 | 82.14 | 82.76 | 82.46 | 0.93 | 0.65 |
| **GBM** | 89.66 | 58.62 | 74.14 | 0.74 | 0.51 | 89.29 | 79.31 | 84.21 | 0.84 | 0.69 |
| **ET** | 79.31 | 83.45 | 81.38 | 0.90 | 0.63 | 92.86 | 86.21 | 89.47 | 0.96 | 0.79 |
| **SVC** | 82.76 | 80.00 | 81.38 | 0.89 | 0.63 | 85.71 | 86.21 | 85.97 | 0.96 | 0.72 |
| **HLA-E*01:03** | | | | | | | | | | |
| **DT** | 68.28 | 81.38 | 74.83 | 0.79 | 0.50 | 71.43 | 66.93 | 69.17 | 0.76 | 0.38 |
| **RF** | 80.69 | 83.45 | 82.07 | 0.90 | 0.64 | 92.06 | 77.95 | 84.98 | 0.93 | 0.71 |
| **LR** | 80.69 | 80.69 | 80.69 | 0.89 | 0.61 | 88.89 | 77.17 | 83 | 0.9 | 0.67 |
| **XGB** | 82.07 | 81.38 | 81.72 | 0.90 | 0.63 | 82.54 | 77.95 | 80.24 | 0.9 | 0.61 |
| **KNN** | 78.62 | 80.00 | 79.31 | 0.86 | 0.59 | 88.1 | 72.44 | 80.24 | 0.9 | 0.61 |
| **GBM** | 89.66 | 58.62 | 74.14 | 0.74 | 0.51 | 90.48 | 46.46 | 68.38 | 0.69 | 0.41 |
| **ET** | 79.31 | 83.45 | 81.38 | 0.90 | 0.63 | 93.65 | 77.95 | 85.77 | 0.93 | 0.73 |
| **SVC** | 82.76 | 80.00 | 81.38 | 0.89 | 0.63 | 90.48 | 79.53 | 84.98 | 0.94 | 0.7 |

**#DT, Decision tree; GNB, Gaussian Naive Bayes; KNN, k-nearest neighbor; LR, Logistic Regression; RF, Random Forest; XGB, *XGBoost*; Sens, ET, Extra Tree; SVC, Support vector classifier; Sensitivity; Spec, Specificity; Acc, Accuracy; AUROC, Area Under Receiver Operating Curve**

## 5.4 Comparison with existing methods

In order to understand the advantages/dis-advantages of this method, it is crucial to compare and validate our method with existing tools. Currently we have compared $HLA_{nc}$Pred with MHCflurry 2.0 and NetMHCpan 4.1 existing methods. Here, we have trained our models on the dataset used in MHCflurry 2.0 and NetMHCpan 4.1 tools and validate the performance of all the methods including $HLA_{nc}$Pred on the updated dataset provided in IEDB database. As shown in the Table 5.5, HLAncPred outperform all existing methods.

**Table 5.5: The comparison of performance of HLAncPred and other methods on the updated IEDB dataset - Adopted from (Dhall et al., 2022)**

| HLA-allele | $HLA_{nc}$Pred | | | | MHCflurry 2.0 | | | | NetMHCpan | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sens | Spec | Acc | MCC | Sens | Spec | Acc | MCC | Sens | Spec | Acc | MCC |
| **HLA-G∗01:01** | 92.60 | 94.30 | 93.40 | 0.87 | 88.70 | 93.30 | 91.00 | 0.82 | 47.20 | 98.70 | 72.90 | 0.53 |
| **HLA-G∗01:03** | 72.20 | 61.10 | 66.70 | 0.33 | 27.80 | 94.40 | 61.10 | 0.29 | 8.30 | 97.20 | 52.80 | 0.12 |
| **HLA-G∗01:04** | 73.50 | 70.60 | 72.10 | 0.44 | 32.40 | 97.10 | 64.70 | 0.39 | 11.80 | 100.00 | 55.90 | 0.25 |
| **HLA-E∗01:01** | 92.10 | 88.10 | 90.10 | 0.80 | 85.70 | 84.90 | 85.40 | 0.71 | 82.50 | 92.10 | 87.30 | 0.75 |
| **HLA-E∗01:03** | 71.30 | 83.80 | 77.60 | 0.56 | 61.20 | 91.00 | 76.50 | 0.55 | 50.50 | 95.30 | 72.90 | 0.51 |

## 5.5 Webserver & standalone package

In the current study, we created a web-based tool called "HLAncPred" to provide facility to the researchers for the predictions and scanning of non-classical HLA-binder and non-binder peptides (https://webs.iiitd.edu.in/raghava/hlancpred/) (See Figure 5.5). In order to more accurately predict non-classical HLA-binders, we have used our best models in this web server. We have provided two major modules (1) PREDICT and (2) SCAN in our website. The prediction module enables users to determine which HLA-G (-G*01:01, -G*01:02, -G*01:03) and HLA-E (-E*01:01, -E*01:03) peptides are the most promiscuous binders and non-binders. Users can upload the input files or upload numerous peptides in the usual FASTA format and choose whether to predict binding for just one allele or several alleles. The results are presented by the server in tabular format, including the input sequence, score, and prediction (binder/non-binder). By utilizing their binary profiles, this module enables the facility to identify the protein areas that might bind to non-classical alleles such HLA-G*01:01, -G*01:02, -G*01:03, -E*01:01, and -E*01:03.

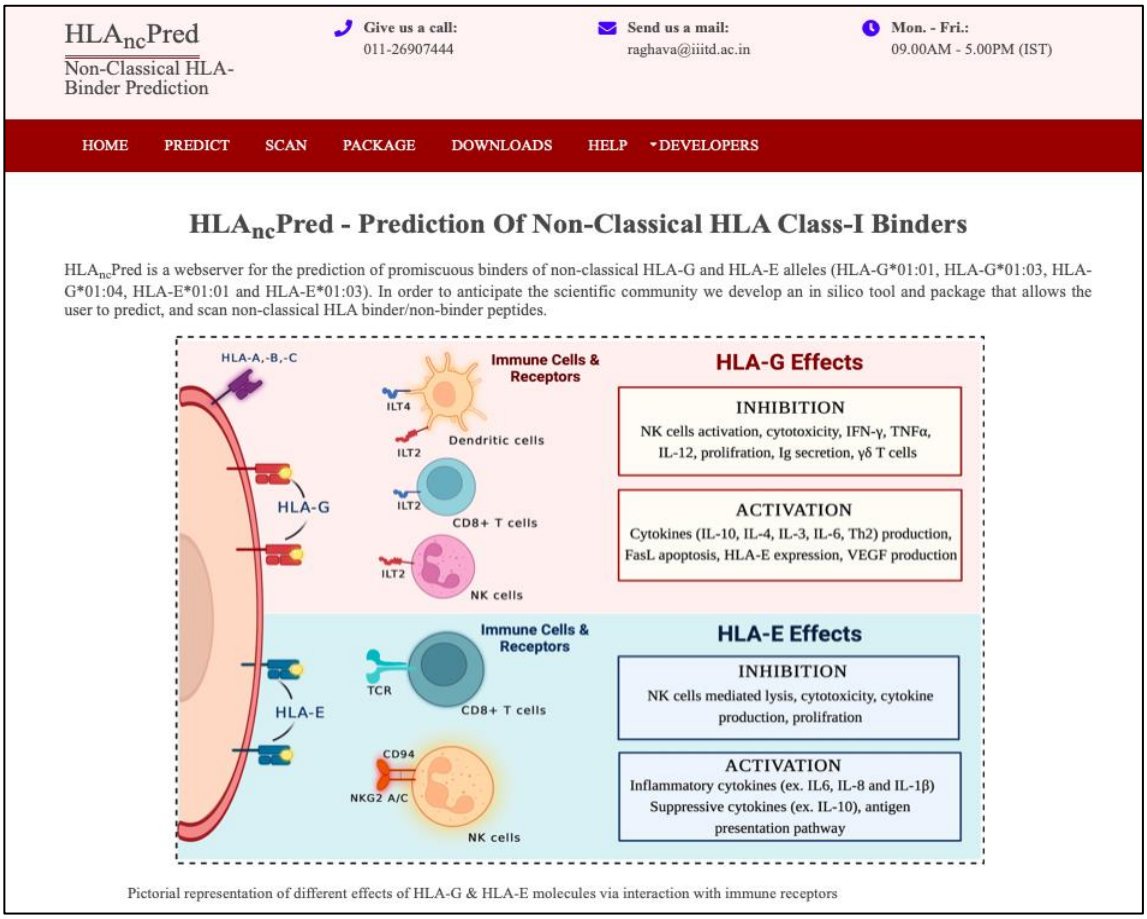



**Figure 5.5: Home page of HLA$_{nc}$Pred webserver**

**(https://webs.iiitd.edu.in/raghava/hlancpred/index.html )**

Additionally, it enables users to anticipate binders using any length of sequence. Users can also search a protein sequence in the opposite direction to discover new peptides that can bind to HLA-G and HLA-E alleles. It will produce fragments with the length the users specify and forecast their behavior. The user can select the allele(s) for the prediction and provide one or more protein sequences in FASTA format. We have created our tool using HTML, PHP, and JAVA scripts and is compatible with a variety of gadgets (including the iPhone, iPad, computers, and android mobile phones). The utility of 'PREDICT' module of the server is provided in Figure 5.6 and 5.7.

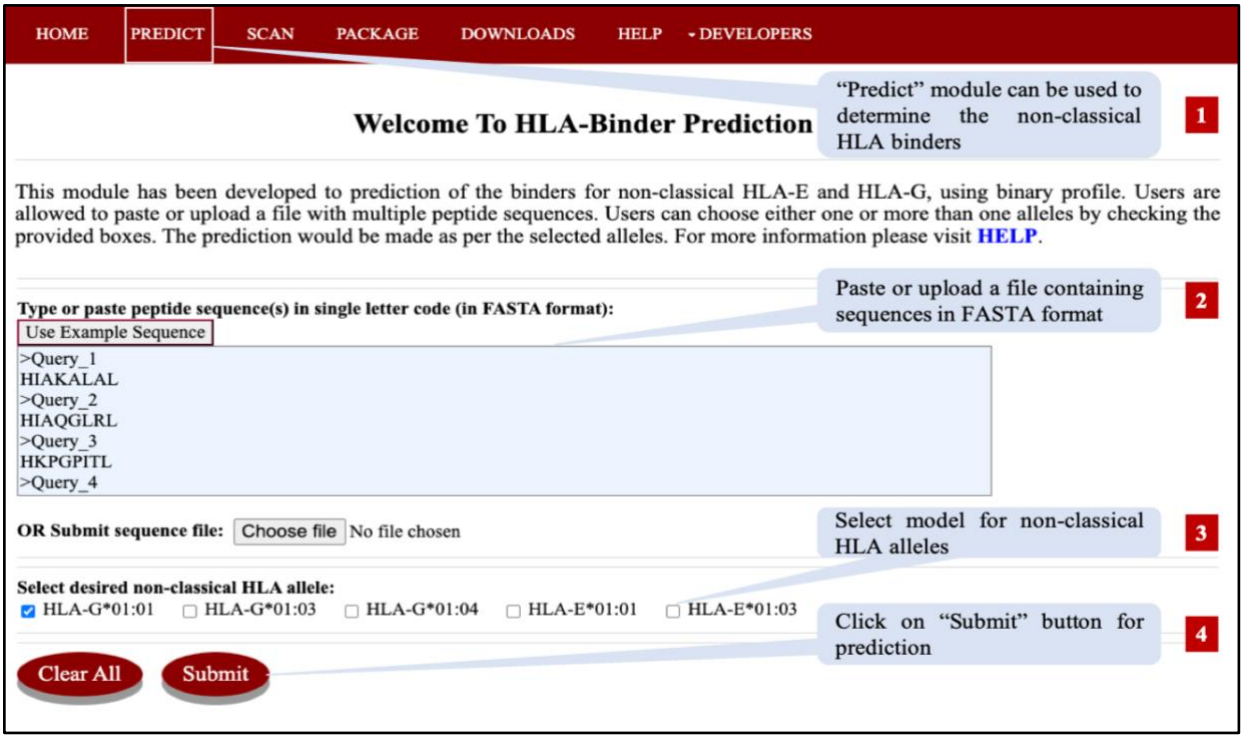



**Figure 5.6 Steps involved in submitting a sequence for predicting binders for non-classical HLA-alleles using 'PREDICT' module of $HLA_{nc}$Pred server**

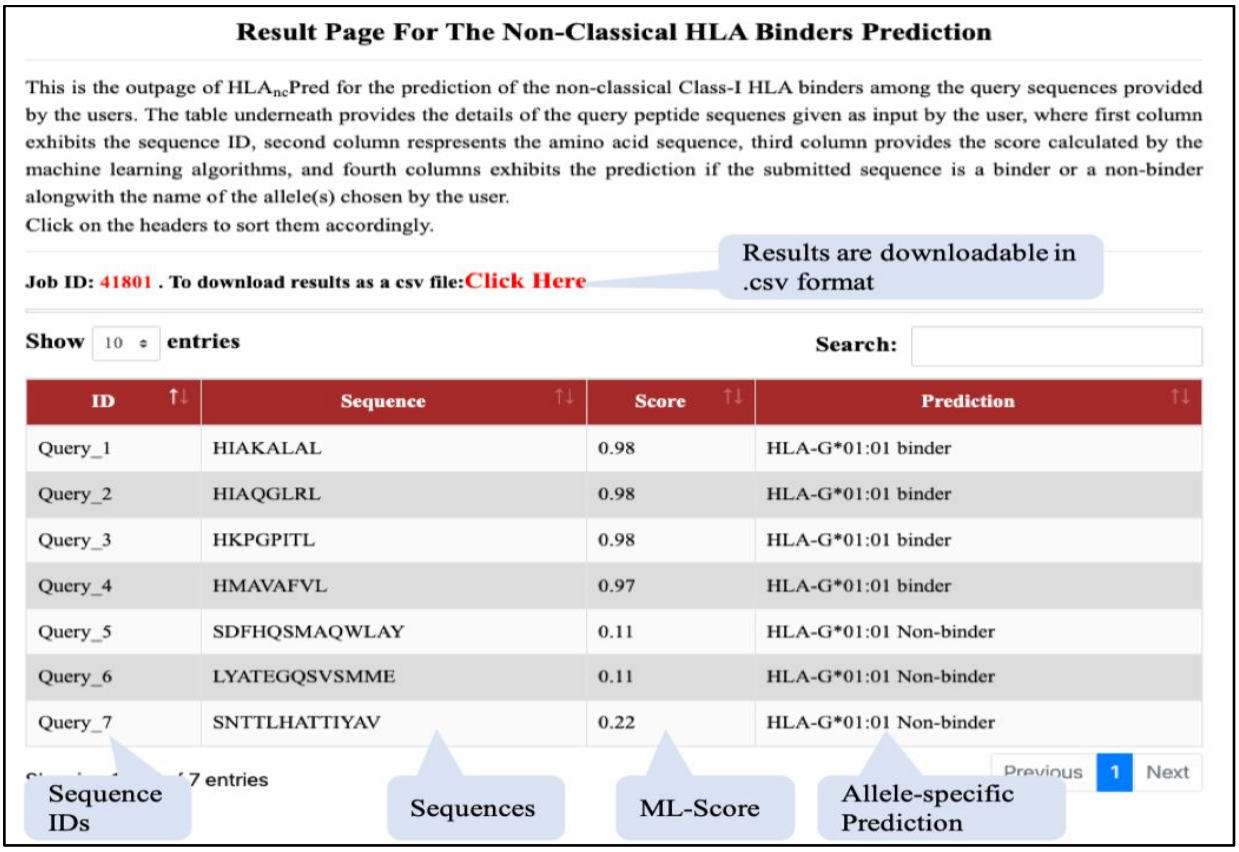



**Figure 5.7 Output page of 'PREDICT' module provides query sequence, score and prediction**

## 5.6 Discussion

During the development of the fetus, the non-classical HLA, such as HLA-G, functions as an immunomodulatory molecule and a natural defense. During viral infections, HLA-E activates inflammatory cytokines to cause immunological responses. It should be noted that excessive HLA-G expression may create an immuno-suppressive milieu, which could aid tumor cells in avoiding our innate and adaptive immune systems. Studies have also revealed that immune-mediated diseases including multiple sclerosis and systemic lupus erythematosus are caused by the excessive and abnormal expression of HLA-G. HLA-E based T cell immunotherapy may be administered to a heterogenic population due to the low polymorphism of non-classical HLA, which may have numerous advantages over traditional HLA-based therapies. Additionally, by interacting with CD8+ T-cells via HLA-E, anti-inflammatory immune response is activated, which inhibits cytokine storm. On the other hand, immunotherapies based on HLA-G have been demonstrated to have encouraging outcomes in the management of solid cancers. An anti-HLA-G CAR-T cell immunotherapy has been developed by researchers to treat acute lymphoblastic leukemia and B-cell malignancies. Therefore, the creation of an accurate prediction approach for the detection of non-classical HLA-binder peptides is absolutely necessary. Numerous HLA binding peptide prediction algorithms have been created in recent years, but only a small number of them have been used for non-classical binder prediction. In the current study, we developed a prediction method, which can be utilized by scientific community to develop a vaccine against the cancer. Researchers can also use this approach to forecast the peptides that non-classical HLA-alleles will bind to in order to fight off various viral, autoimmune, and pathogenic diseases. We believe that the community engaged in designing vaccines and HLA-based immunotherapies will profit from our approach. Identifying the promiscuous binders or antigenic areas that can bind to a large variety of HLA alleles is one of the key difficulties in the field of immunotherapy. We created a standalone software and web server called HLAncPred to forecast the promiscuous binders for non-classical HLA alleles.

## 5.7 Conclusion

The scientific community can use these findings to develop a vaccine against the deadly virus and cancers in order to predict the peptides that binds to non-classical HLA-alleles. Moreover, this tool can be extended by adding more information and new models for other Class-I and Class-II non-classical HLA-alleles.

# CHAPTER 6

## PREDICTION OF IL6 INDUCING PEPTIDES

## *6.1 Introduction*

The pleiotropic cytokine interleukin 6 is produced by the interleukin 6 gene (IL6). It is also known by some other names, including plasmacytoma growth factor, interferon-beta (IFN- β2) and B cell stimulatory factor-2 (Ataie-Kachoie et al., 2014). It is a multifunctional cytokine and play crucial role in both innate and adaptive immune responses, rheumatoid arthritis, haematopoiesis, acute phase reactions, and organ development (Su et al., 2017), among other inflammatory illnesses. Infections and tissue injury are the main triggers for its production (Tanaka et al., 2014; Velazquez-Salinas et al., 2019). Numerous cell types, including macrophages, dendritic cells, mast cells, fibroblasts, endothelial cells, T cells, and B cells are associated with the generation of IL6 (Mauer et al., 2015; Velazquez-Salinas et al., 2019) (See Figure 6.1). IL6 is essential for controlling numerous physiological processes, including those of the immune system, central neurological system, and cardiovascular system. Recent research has been shown that IL6 dysregulation contributes to the onset, progression, and metastasis of a number of diseases, including different forms of cancer (Hong et al., 2007).

Numerous studies have shown that elevated levels of IL6 are associated with a higher risk of developing cancer as well as other diseases like insulin resistance, asthma, coronary heart disease, and cancer. They have also shown that elevated levels of IL6 can serve as a prognostic marker for cancer (Ujiie et al., 2012; Zarogoulidis et al., 2013). A cytokine storm or cytokine release syndrome (CRS), which is the abnormal release of circulating cytokines, may have contributed to the recent outbreak of coronavirus disease (COVID-19), which is caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), also known as the 2019 novel coronavirus (2019-nCoV). The health of COVID-19 patients has significantly declined as a result of the dramatically increasing levels/high levels of IL6 and other pro-inflammatory cytokines, such as IL-1, IL-8, IL-12, IL-18, interferon (IFN), and tumour necrosis factor (TNF). The progression of COVID-19 infection from pneumonia to respiratory failure (L. Zou et al., 2020) and acute respiratory distress syndrome (ARDS) ultimately results in multi-system organ failure and significant mortality. Since the severity of the disease's effects is worsened by the larger cytokine storm caused by the elevated IL6 concentration, IL6 may be exploited as a possible therapeutic target or biomarker for severe COVID-19 cases (Chen, Zhao, et al., 2020).

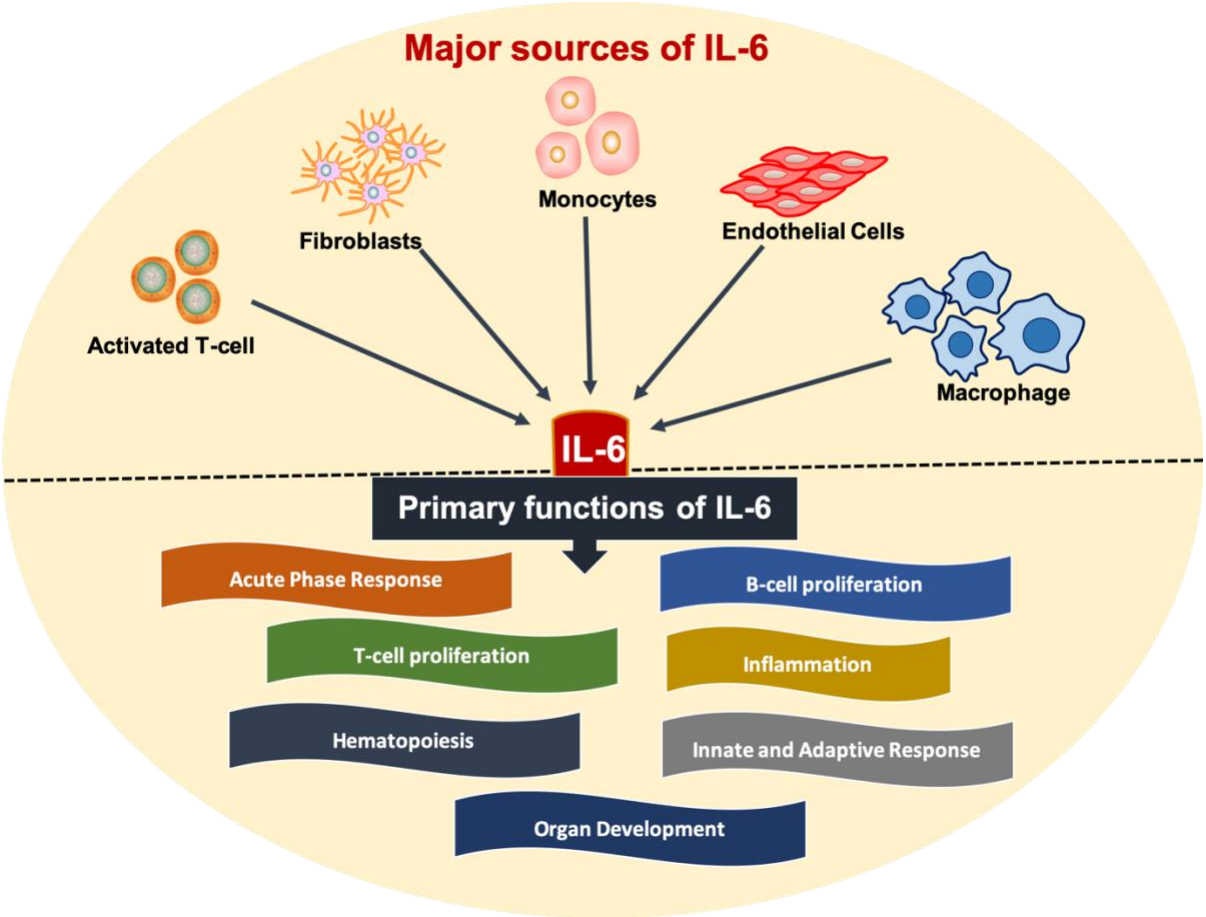



**Figure 6.1: Depicts the mode of IL6 secretion by different cells and its main roles in our immune system (i.e., T-cell, B-cell proliferation, organ development, etc.)**

Through tightly regulated transcriptional and post-transcriptional mechanisms, IL6 is quickly produced as an immunological response to infection and tissue damage. However, IL6 expression that is dysregulated has a detrimental impact on autoimmune disease and chronic inflammation. In numerous disorders, including Alzheimer's disease, atherosclerosis, Behçet's disease, diabetes, depression, multiple myeloma, prostate cancer, rheumatoid arthritis, and systemic lupus erythematosus, IL6 increases the inflammatory and auto-immune processes. Numerous COVID-19 verified cases have been identified with elevated serum IL6 levels. Therefore, it must be highlighted that certain disorders either require anti-IL6 treatment or require checking for the existence of IL6 triggering factors.

In recent years, a number of computational methods for cytokine prediction and classification have been established. A cytokine-specific approach called CytoPred (Lata & Raghava, 2008) predicts and further categorises the cytokine into its family and sub-family. IFNepitope (Dhanda, Vir, et al., 2013) is a technique created to predict and create peptides that induce IFN-gamma (IFN-gamma). In order to

predict the peptides that induce IL-4, IL-10, and IL-17, respectively, some techniques, such as IL4Pred (Dhanda, Gupta, et al., 2013), IL-10Pred (Nagpal et al., 2017), and IL17eScan (Gupta, Mittal, et al., 2017), were created. ProInflam (Gupta et al., 2016) and PIP-EL (Manavalan et al., 2018), which predict the pro-inflammatory nature of the peptides and proteins, which causes the generation of pro-inflammatory cytokines, are two methods that have been developed for the prediction of specific cytokines. The peptides or proteins that trigger the generation of anti-inflammatory cytokines are predicted by AntiInfam (Gupta, Sharma, et al., 2017). Of note, there is no dedicated method or computational tool which can predict the IL6 inducing peptides. An effort has been made to create computational models for prediction, scanning and designing of peptides that can cause the release of the cytokine IL6 in order to benefit the scientific community. We created the positive and negative dataset from IEDB and applied various machine learning algorithm for the development of classification models. The overall architecture of the study is depicted in Figure 6.2.

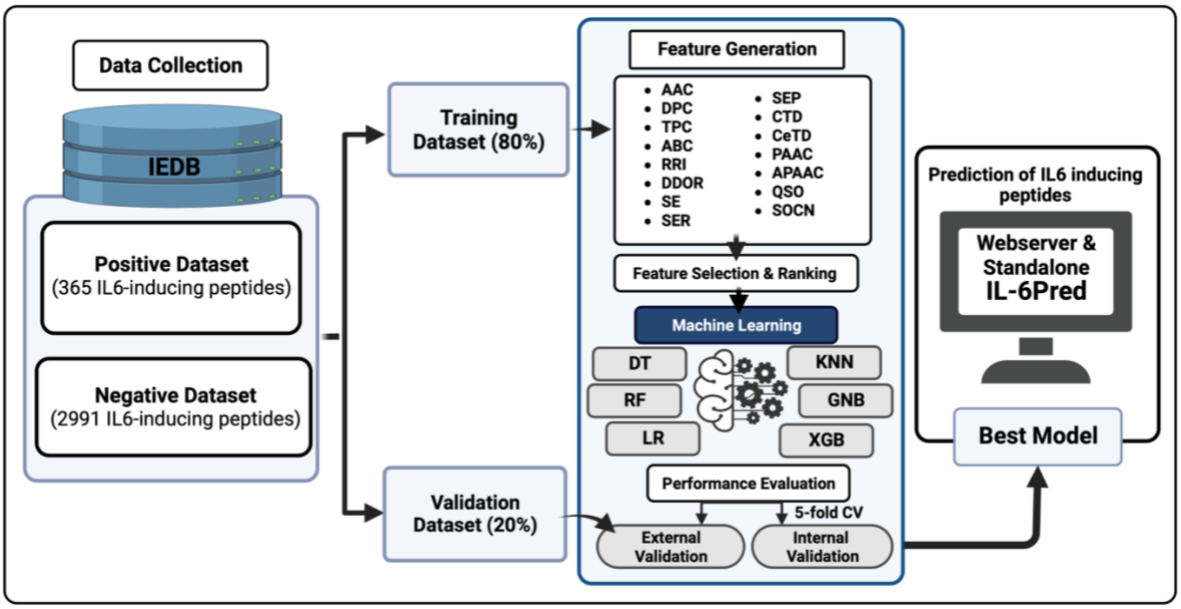



**Figure 6.2 Shows the complete workflow of the study, including dataset collection from IEDB, feature generation and selection, machine learning algorithms and webserver development**

## *6.2 Material and methods*

### *6.2.1 Compilation of data*

In order to create machine learning models, we have to select a clean, well-annotated dataset. In this study, we compiled the dataset from immune epitope database (IEDB), which is available to the public,

and extracted experimentally validated linear epitopes (R et al., 2019). In the T-cell assay of human and mouse, the positive dataset includes IL6-inducing peptides, while the negative dataset includes proinflammatory cytokines that do not induce IL6 (e.g., IL1, IL1, TNF, IL8, IL12, IL17, IL18, and IL23). We eliminate identical sequences from peptides that induce IL6 and those that do not. Then, peptides with lengths less then 8 residues or more and more than 25 residues were eliminated. Finally, we had a major dataset with 365 distinct epitopes that could induce IL6 and 2991 peptides are IL6-non inducing that couldn't, known as the positive dataset and the negative dataset, respectively.

### *6.2.2 Data analysis*

To compute the amino-acid composition of positive and negative dataset we have used Pfeature tool. To determine the ideal length for both positive and negative peptides, we first examine the IL6-inducing and non-inducing sequences. Since the a two-sample logo (TSL) technique requires a predetermined length of input sequence vector criterion, we develop sequence logo to comprehend the preference of particular amino acids at a given position. We need eight residues from the N-terminal and eight residues from the C-terminal to build a fixed-length feature vector because the minimum length of the peptide in our dataset is eight residues. This results in a profile with 16 residue positions.

### *6.2.3 Feature generation*

This work uses Pfeature to compute a variety of characteristics from the peptide sequences. 15 different types of features were generated using the composition-based feature module. The Pfeature composition-based features module produced a vector with 9149 features. Amino acid composition, dipeptide composition, tripeptide composition, and atom & bond composition are used to construct simple composition-based characteristics. Computing physio-chemical properties, Residue repeat information, Property repeat information, and Distance distribution of residue have been used to identify residue and distribution information. Conjoint Triad Descriptors, Composition Improved Transition and Distribution, Shannon entropy Quasi-Sequence Order, Amphiphilic Pseudo Amino Acid Composition, Sequence Order Coupling Number, and Pseudo Amino Acid Composition.

### *6.2.4 Development of prediction models*

In this study, a classification model for the prediction of IL6 or non-IL6 inducing peptides has been developed using a number of machine learning methods. Decision tree (DT), Random Forest (RF), Logistic Regression (LR), XGBoost (XGB), k-nearest neighbor (KNN) and Gaussian Naive Bayes

(GNB). scikit-Learn, a Python library, was used to build these classification methods. We employed a number of protein features produced by the Pfeature package to create prediction models.

### *6.2.5 Feature selection/ranking techniques*

Finding a relevant set of features from the enormous dimension of features is one of the study's biggest obstacles. There are various techniques for feature selection; in order to remove unimportant features from the training dataset, we employed the SVC with the L1 penalty and the feature selection methodology from the Scikit package. The L1 penalty is applied on the non-zero coefficients chosen by the SVC-L1 approach before the relevant features are chosen to minimise the dimensions. From the 9149 features in the feature set, we choose 186 for SVC-L1. Furthermore, we rank the feature using the feature selector tool (https://pypi.org/project/feature-selector/ ) in order to decrease the number of protein features from the chosen collection of features. In this procedure, characteristics were ordered based on their normalised and cumulative importance, respectively. To comprehend the nature of IL6 inducing peptides, these major features were investigated. We also used machine learning to analyse a set of features, computing the performance on the top 10, 20, 30, …. 186 features, respectively.

### *6.2.6 Parameters of evaluation*

In order to evaluate and test our prediction models, we used the five-fold cross-validation method. One of the most used evaluation methods is five-fold cross-validation. The complete training dataset is first split into five equal sets, or folds, each of which has the same number of positive and negative examples. The fifth fold was then used for testing after using the first four for training. Each set is tested after the operation has been repeated several times. In addition, we employed established evaluation metrics to assess the effectiveness of various prediction models. We employed both threshold-dependent and independent parameters in this work, and we used the following equations to assess threshold-dependent characteristics including sensitivity, specificity, and accuracy. In order to evaluate the effectiveness of the models, we additionally employed the typical threshold-independent parameter Area under the Receiver Operating Characteristic (AUROC) curve. Plotting sensitivity versus (1-specificity) on various thresholds results in the creation of the AUROC curve.

## *6.3 Results*

In this investigation, we employed 365 peptides as a positive dataset that can stimulate the production of the cytokine IL6. The primary dataset has 2991 peptides that do not trigger the IL6 cytokine, called

as negative dataset. All the calculations and predictions made for IL6-inducing and non-inducing peptides.

### *6.3.1 Conservation and compositional analysis*

To explore the preference of individual amino acids at a particular location in the peptide string. We generate a sequence logo for the IL6 inducing peptides using WebLogo software (http://weblogo.threeplusone.com), as shown in Figure 6.3. The relative abundance of the sequence is represented by the most important amino acid residue. WebLogo shows residue positions on x-axis, and bit-score on y-axis. Here, bit-score signifying the conservation of residues. Each position exhibits the stack of amino acids which are conserved at that position, where the height of each residue signifies the relative frequency. We observed that the leucine, alanine and isoleucine amino acid residue is highly conserved in IL6 inducing peptides.

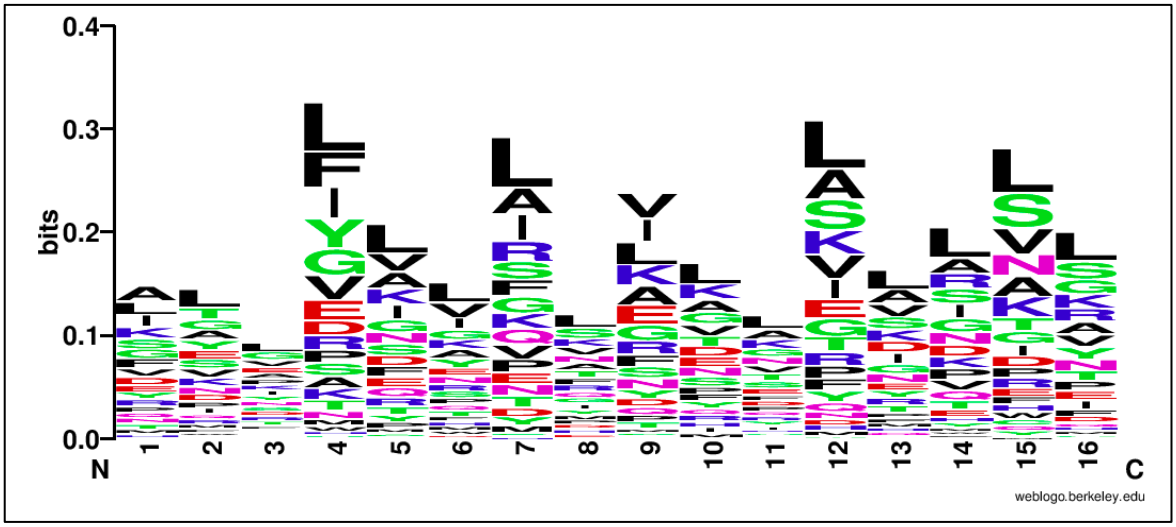



**Figure 6.3 WebLogo represent the conserved amino-acid residues**

Moreover, we calculated the amino acid composition (AAC) in this analysis for both positive and negative datasets. Figure 6.4 depicts the typical composition of IL6-inducing and non-inducing peptides. In contrast to non-IL6 peptides, IL6 inducing peptides have a higher average composition of residues like (I, L, and S). Additionally, non-IL6 peptides include more of the residues (A, D, and G) than IL6 inducing peptides.

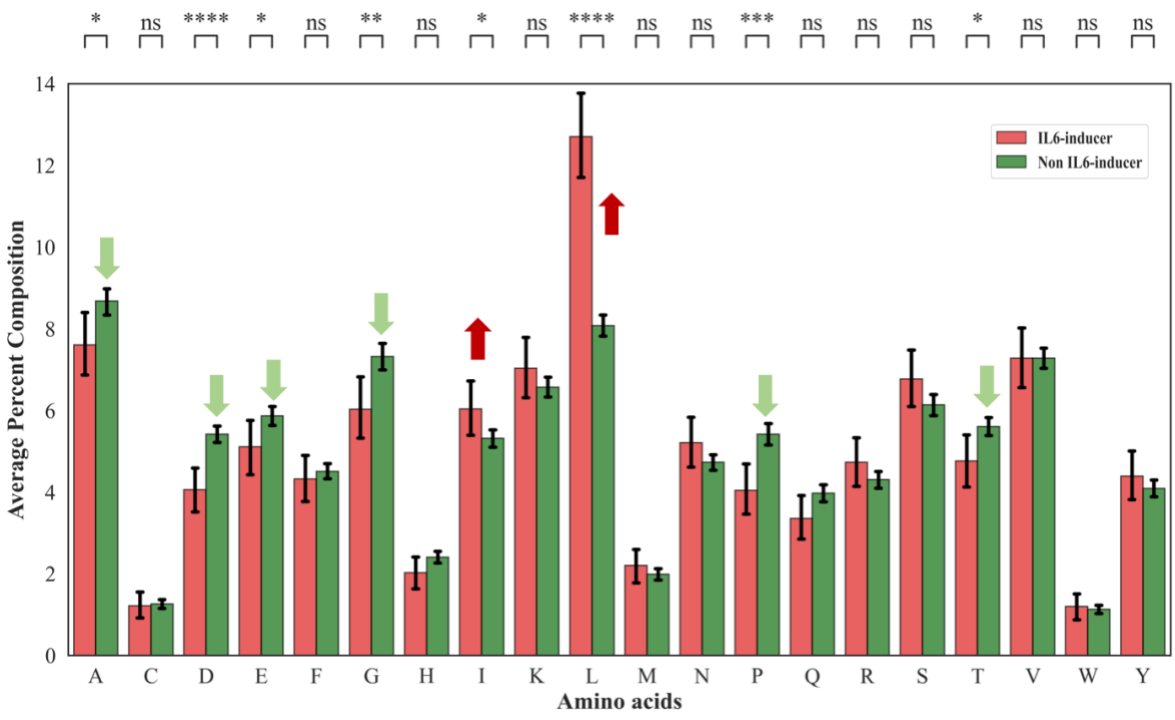



**Figure 6.4 Illustrate average amino-acid composition of IL6 inducing and non-inducing peptides; where, up-arrow represents the average composition of residue is higher in IL6 inducing peptides and down-arrow represents the average composition of residue is lower in IL6 inducing peptides**

## *6.3.2 Preformation of prediction models*

Utilizing a variety of classifiers, including RF, DT, GNB, XGB, and LR, we create prediction models. First, using the Pfeature compositional based module, we compute the features of the IL6 inducers and non-inducers. Pfeature generates a total of 9149 features and we used the SVC-L1 feature selection technique to choose the most important features. After selecting 186 features, we rank the features using feature-selector algorithm. Then, we have computed performance on top-10, 20, 30, 40, …… 186 features, as shown in Table 6.1. We observed that the performance of top-10 and top-186 features is highly accurate with balanced sensitivity and specificity. While, the models based on other features achieved high accuracy and AUC on training dataset and poor performance on validation dataset. Moreover the performance of other features is highly imbalanced.

**Table 6.1: Evaluation of machine learning based models on training and validation dataset; developed using top-10, 20, 30, …….. 186 features**

| Number of Features | Method | Dataset | Sens | Spec | Acc | AUC |
|---|---|---|---|---|---|---|
| **TOP-10** | **RF** | **Training** | 77.40 | 77.39 | 77.39 | 0.84 |
| | | **Validation** | 75.34 | 73.24 | 73.47 | 0.83 |
| **TOP-20** | **XGB** | **Training** | 71.58 | 83.12 | 81.86 | 0.84 |
| | | **Validation** | 71.23 | 81.44 | 80.33 | 0.86 |
| **TOP-30** | **XGB** | **Training** | 62.67 | 87.88 | 85.14 | 0.84 |
| | | **Validation** | 64.38 | 87.46 | 84.95 | 0.84 |
| **TOP-40** | **XGB** | **Training** | 60.62 | 90.47 | 87.23 | 0.84 |
| | | **Validation** | 61.64 | 89.30 | 86.29 | 0.83 |
| **TOP-50** | **XGB** | **Training** | 65.41 | 88.30 | 85.81 | 0.86 |
| | | **Validation** | 65.75 | 86.96 | 84.65 | 0.85 |
| **TOP-60** | **XGB** | **Training** | 63.36 | 85.17 | 82.79 | 0.84 |
| | | **Validation** | 67.12 | 85.28 | 83.31 | 0.84 |
| **TOP-70** | **XGB** | **Training** | 68.15 | 87.05 | 84.99 | 0.86 |
| | | **Validation** | 63.01 | 86.62 | 84.05 | 0.84 |
| **TOP-80** | **XGB** | **Training** | 69.86 | 87.51 | 85.59 | 0.87 |
| | | **Validation** | 68.49 | 83.95 | 82.27 | 0.84 |
| **TOP-90** | **XGB** | **Training** | 72.26 | 85.50 | 84.06 | 0.87 |
| | | **Validation** | 68.49 | 83.61 | 81.97 | 0.83 |
| **TOP-100** | **XGB** | **Training** | 69.18 | 88.97 | 86.82 | 0.87 |
| | | **Validation** | 60.27 | 87.12 | 84.20 | 0.82 |
| **TOP-110** | **RF** | **Training** | 81.16 | 73.46 | 74.30 | 0.87 |
| | | **Validation** | 78.08 | 71.24 | 71.98 | 0.82 |
| **TOP-120** | **RF** | **Training** | 83.56 | 71.38 | 72.70 | 0.87 |
| | | **Validation** | 79.45 | 70.07 | 71.09 | 0.82 |
| **TOP-130** | **RF** | **Training** | 83.56 | 72.13 | 73.37 | 0.87 |
| | | **Validation** | 76.71 | 71.07 | 71.68 | 0.83 |
| **TOP-140** | **XGB** | **Training** | 65.41 | 90.97 | 88.19 | 0.86 |
| | | **Validation** | 63.01 | 89.30 | 86.44 | 0.85 |
| **TOP-150** | **XGB** | **Training** | 66.78 | 87.59 | 85.33 | 0.87 |
| | | **Validation** | 60.27 | 87.46 | 84.50 | 0.81 |
| **TOP-160** | **XGB** | **Training** | 66.44 | 87.13 | 84.88 | 0.88 |
| | | **Validation** | 67.12 | 87.79 | 85.54 | 0.84 |
| **TOP-170** | **XGB** | **Training** | 60.27 | 92.35 | 88.86 | 0.88 |
| | | **Validation** | 56.16 | 91.14 | 87.33 | 0.86 |
| **TOP-180** | **RF** | **Training** | 87.67 | 73.76 | 75.27 | 0.89 |
| | | **Validation** | 82.19 | 72.07 | 73.17 | 0.86 |
| **TOP-186** | **RF** | **Training** | 85.96 | 74.55 | 75.79 | 0.89 |
| | | **Validation** | 83.56 | 72.07 | 73.32 | 0.86 |

**#RF, Random Forest; XGB, *XGBoost*; Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; AUROC, Area Under Receiver Operating Curve**

### *6.3.2.1 Top-10 features based model*

Using the feature selector tool, all 186 features were ranked in order of relevance using their normalised and cumulative scores. We also assess how well the various feature sets perform. We determined the feature set with the smallest amount of characteristics that will accurately and with high AUROC distinguish between IL6 inducers and non-inducers. So, using 10, 20, 30,……..,186 characteristics, respectively, we develop several models and assess their performance using the training and validation datasets. We calculated the average values of the top 10 characteristics of IL6 inducing and non-inducing peptides, as shown in Table 6.2, to better grasp the distinction between the positive and negative datasets. In terms of AUROC and accuracy, the top-10 traits chosen have high discriminatory power. As shown in Table 6.2 RF-based models achieve maximum performance with accuracy (77.39 and 73.47), AUROC (0.84 and 0.83) on training and validation datasets, respectively.

**Table 6.2: Evaluation of machine learning based models on training and validation dataset; developed using top-10 features**

| Classifier | Dataset | Sens | Spec | Acc | AUC |
|---|---|---|---|---|---|
| **DT** | **Training** | 70.55 | 69.12 | 69.27 | 0.74 |
| | **Validation** | 69.86 | 68.23 | 68.41 | 0.72 |
| **GNB** | **Training** | 70.21 | 66.15 | 66.59 | 0.74 |
| | **Validation** | 67.12 | 64.38 | 64.68 | 0.72 |
| **KNN** | **Training** | 58.56 | 42.29 | 44.06 | 0.52 |
| | **Validation** | 64.38 | 48.16 | 49.93 | 0.58 |
| **LR** | **Training** | 61.64 | 58.63 | 58.96 | 0.64 |
| | **Validation** | 64.38 | 57.19 | 57.97 | 0.64 |
| **RF** | **Training** | 77.4 | 77.39 | 77.39 | 0.84 |
| | **Validation** | 75.34 | 73.24 | 73.47 | 0.83 |
| **XGB** | **Training** | 71.23 | 72.71 | 72.55 | 0.8 |
| | **Validation** | 71.23 | 67.56 | 67.96 | 0.8 |

**#DT, Decision tree; GNB, Gaussian Naive Bayes; KNN, k-nearest neighbor; LR, Logistic Regression; RF, Random Forest; XGB, *XGBoost*; Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; AUROC, Area Under Receiver Operating Curve**

### *6.3.2.2 Top-186 features based model*

Performance of top-186 features provided in Table 6.3. We achieved an AUROC of 0.893 and 0.863; accuracy 75.79 and 73.32 on training and validation, and balanced sensitivity and specificity, random forest (RF) achieves optimal performance. With AUROC values of 0.87 and 0.82 and accuracy values

of 86.29 and 84.65, XGboost also performs well on training and validation datasets; however, there is a significant variation in sensitivity and specificity. As shown in Table 6.3, other classifiers like DT, LR, KNN, and GNB perform badly on training and validation data.

**Table 6.3: Evaluation of machine learning based models on training and validation dataset; developed using top-186 features**

| Classifier | Dataset | Sens | Spec | Accuracy | AUC |
|---|---|---|---|---|---|
| **DT** | **Training** | 40.068 | 89.887 | 84.469 | 0.662 |
| | **Validation** | 39.726 | 89.632 | 84.203 | 0.65 |
| **GNB** | **Training** | 57.534 | 88.884 | 85.475 | 0.815 |
| | **Validation** | 53.425 | 88.294 | 84.501 | 0.782 |
| **KNN** | **Training** | 47.603 | 59.967 | 58.622 | 0.534 |
| | **Validation** | 52.055 | 56.187 | 55.738 | 0.542 |
| **LR** | **Training** | 69.178 | 78.103 | 77.132 | 0.803 |
| | **Validation** | 68.493 | 76.087 | 75.261 | 0.783 |
| **RF** | **Training** | 85.959 | 74.551 | 75.791 | 0.893 |
| | **Validation** | 83.562 | 72.074 | 73.323 | 0.863 |
| **XGB** | **Training** | 66.096 | 88.759 | 86.294 | 0.870 |
| | **Validation** | 58.904 | 87.793 | 84.650 | 0.823 |

**#DT, Decision tree; GNB, Gaussian Naive Bayes; KNN, k-nearest neighbor; LR, Logistic Regression; RF, Random Forest; XGB, *XGBoost*; Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; AUROC, Area Under Receiver Operating Curve**

## *6.4 Computational resource*

We create a user-friendly prediction web server that combines various modules to predict IL6 inducing peptides in order to serve the scientific community. The web server incorporates the prediction models that were utilised in the study. Using the prediction models' score at various thresholds, users can forecast whether the provided query peptide would induce IL6 or not. There are five crucial modules on the web server: (1) Prediction, (2) Design, (3) Protein scan, (4) Motif scan, and (5) BLAST scan (See Figure 6.5). The “Predict” module gives the user the ability to distinguish between peptides that induce IL6 and those that do not. The user can design every potential analogue of the input sequence using the “Design” module. The supplied amino-acid sequence was scanned using the “Protein Scan” module to look for IL6-inducing areas.

Users of the “Motif Scan” module can map or scan IL6 motifs in the query sequence. We extracted motifs from IL6 inducing peptides that have been experimentally verified using the MEME/MAST

and MERCI tools. The “Blast Scan” module is based on the Basic Local Alignment Search Tool, a similarity search technique (BLAST). The database of recognised IL6 inducing peptides is searched against the given query sequence. If a query sequence matches or hits in the database, it is anticipated to be an IL6 inducer; otherwise, it is expected to be a non-IL6 inducer peptide. The peptide sequence (positive and negative datasets) utilised in this work are both available for download by users. Figure 6.6 and 6.7 depicts the usage of Predict module of IL6Pred server.

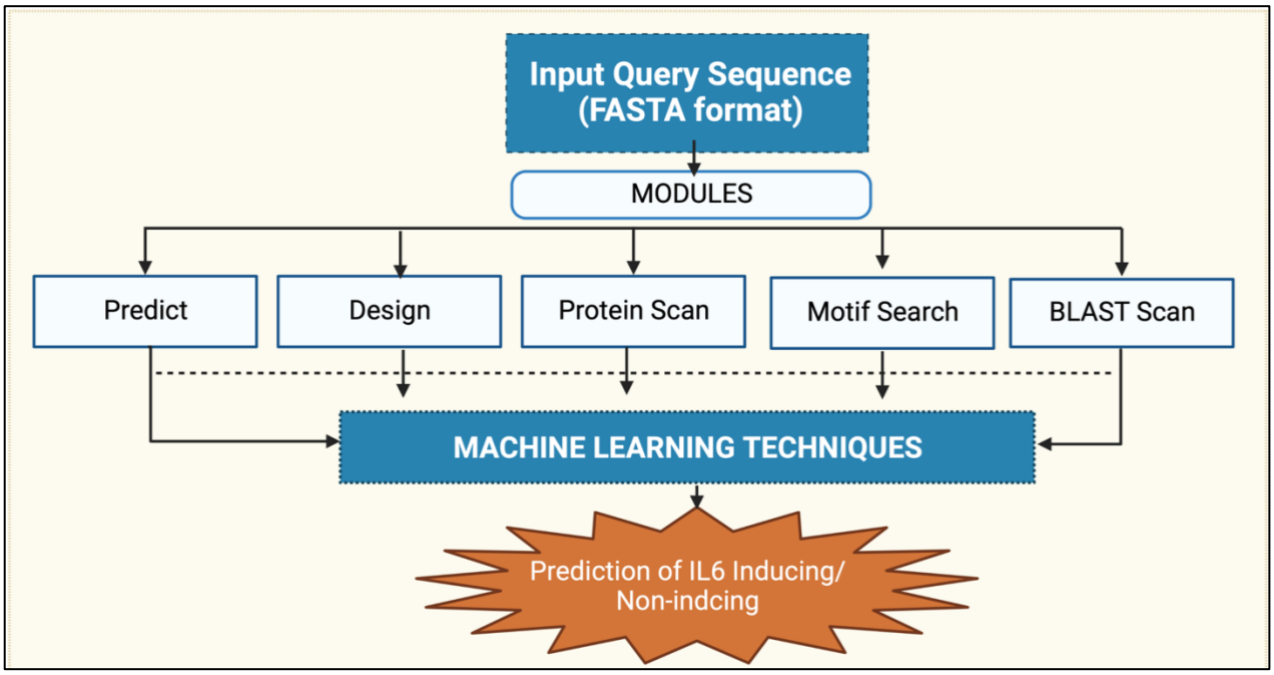



**Figure 6.5 Different modules of IL6pred webserver; where, ‘Predict’ module used for the prediction of IL6 inducing peptides, ‘Design’ module used for the designing of IL6-inducing peptides, ‘Protein Scan’ module identify IL6 inducing regions in protein sequence, ‘Motif Search’ used for the scanning of IL6 specific motifs and ‘BLAST Scan’ utilized for the similarity search**

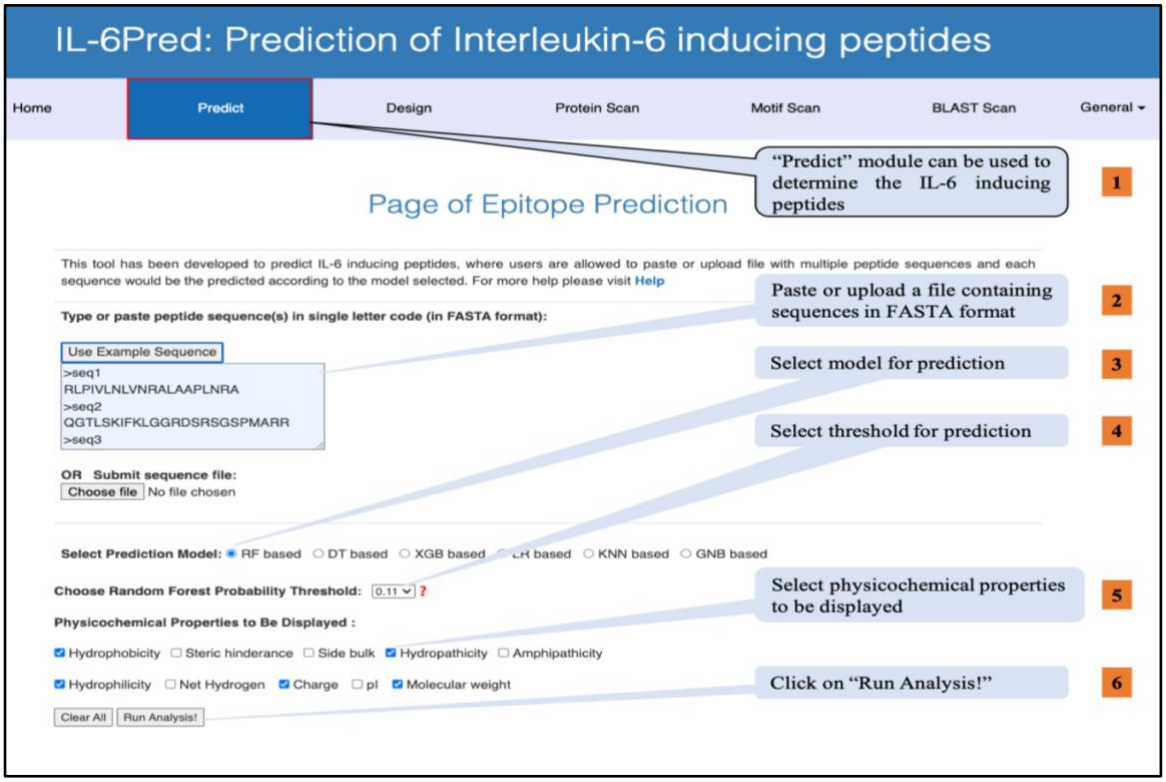
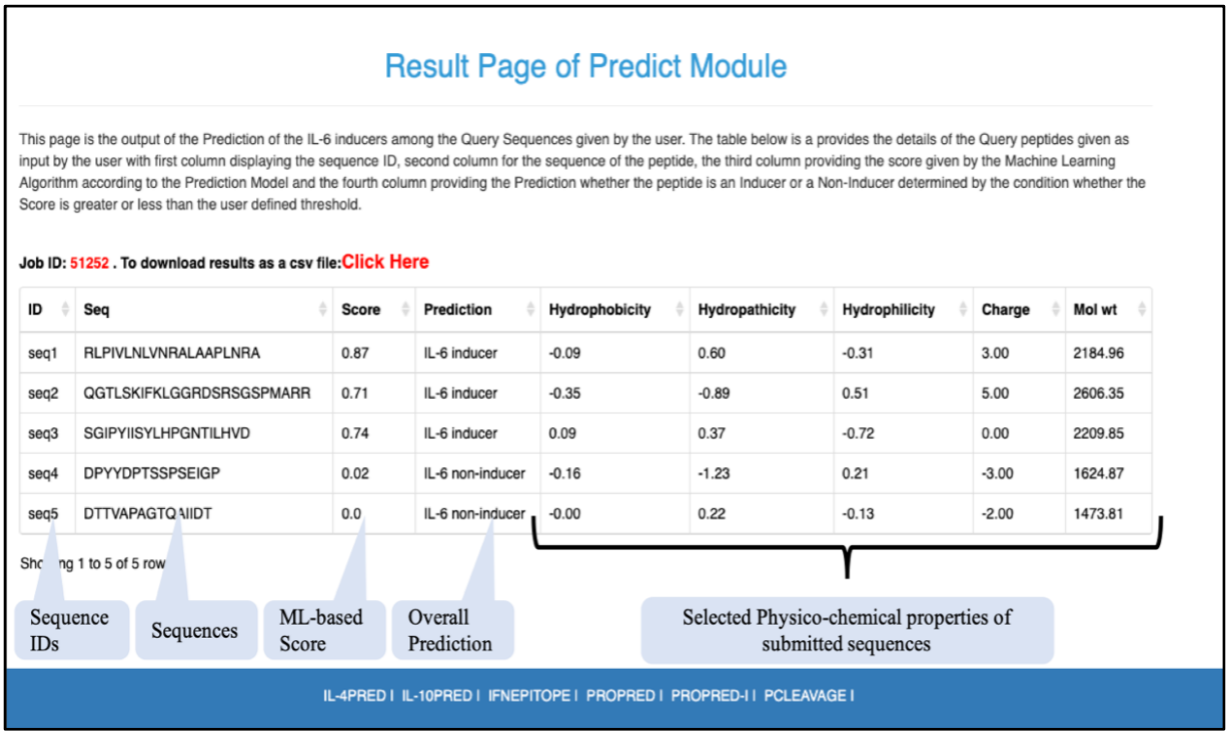



**Figure 6.6 Shows the sequence submission form of IL6Pred, where user can submit query sequence for prediction of IL6 inducing peptides**




| ID | Seq | Score | Prediction | Hydrophobicity | Hydropathicity | Hydrophilicity | Charge | Mol wt |
|---|---|---|---|---|---|---|---|---|
| seq1 | RLPIVLNLVNRALAAPLNRA | 0.87 | IL-6 inducer | -0.09 | 0.60 | -0.31 | 3.00 | 2184.96 |
| seq2 | QGTLSKIFKLGGRDSRSGSPMARR | 0.71 | IL-6 inducer | -0.35 | -0.89 | 0.51 | 5.00 | 2606.35 |
| seq3 | SGIPYIISYLHPGNTILHVD | 0.74 | IL-6 inducer | 0.09 | 0.37 | -0.72 | 0.00 | 2209.85 |
| seq4 | DPYYDPTSSPSEIGP | 0.02 | IL-6 non-inducer | -0.16 | -1.23 | 0.21 | -3.00 | 1624.87 |
| seq5 | DTTVAPAGTQAIIDT | 0.0 | IL-6 non-inducer | -0.00 | 0.22 | -0.13 | -2.00 | 1473.81 |

**Figure 6.7 Output of prediction module of IL6pred server, which shows query sequence, score and prediction as IL6 inducer or IL6 non-inducer**

## 6.5 Discussion

Several vaccines have been created in the past to safely elicit an immune response against disease. Subunit vaccines are being explored as an alternative to traditional attenuation procedures in the current vaccination efforts. Subunit vaccines are made up of protein or peptide fragments from the pathogen that can trigger an immune response to protect against infectious illnesses. These therapeutic peptide subunit vaccines are intriguing prospects for creating vaccines against a variety of illnesses, including cancer, hepatitis B, COVID-19, and tuberculosis. Finding antigenic areas that might cause the appropriate immune response is the main problem in vaccine creation. It would be ideal to experimentally verify the immune response to each conceivable peptide or fragment of the pathogen proteome, but this would be exceedingly costly and time-consuming. Designing subunit vaccines and immunotherapies requires the identification of antigenic areas that bind to MHC and activate T helper cells, which then release cytokines. Several prediction techniques have been created in the past for cytokine detection.

An important pro-inflammatory cytokine known as interleukin 6 (IL6) is essential for both innate and adaptive immune responses. Previous research has shown that elevated levels of IL6 in COVID-19 patients encourage the growth of cancer, autoimmune diseases, and cytokine storm. Through tightly regulated transcriptional and post-transcriptional pathways, IL6 is quickly produced as an immune response in cases of infection and tissue damage. However, IL6 expression that is dysregulated has a detrimental impact on autoimmune disease and chronic inflammation. Multiple disorders, including Alzheimer's disease, atherosclerosis, Behçet's disease, diabetes, depression, multiple myeloma, prostate cancer, rheumatoid arthritis, and systemic lupus erythematosus, are affected by the auto-immune and inflammatory processes that are stimulated by IL6. Numerous COVID-19 verified individuals have reported having high levels of IL6. Therefore, in order to treat a variety of diseases, either anti-IL6 therapy is required, or IL6 stimulating substances must be looked for. Therefore, it is crucial to spot and eliminate antigenic areas in therapeutic proteins or vaccine candidates that could lead to IL6-related immunotoxicity. We have created the computational tool IL6pred to find IL6 inducing peptides in a vaccine candidate in order to solve this difficulty.

In this study, we have built models to recognise the IL6 producing capability of peptides and have attempted to understand the nature of IL6 inducing peptides. This is, as far as we are aware, the first attempt at creating an IL6 inducing peptide prediction tool. We created the dataset using IEDB since the dataset is crucial to machine learning. To investigate the composition and positional preference, TSL and compositional analytical experiments were conducted. We found that IL6 inducing peptides are concentrated in the amino acid leucine (L). 9149 features have been generated from sequencing

data using the programme “Pfeature”. Relevant features were chosen using SVC-L1 from the Scikit package, and they were then sorted using feature selection tools. According to our compositional study, a particular residues such as L, I, and S are preferred types of residues in IL6 peptides, but A, D, and G are not favoured types of residues in IL6 inducing peptides. It’s interesting to notice that 186 features chosen by contemporary feature selection methods SVC-L1 also incorporate these residues' composition (i.e. L, I, A, D, G). This suggests that straightforward compositional-based approaches can recognise crucial traits. In our investigation, we created classification models using these 186 features. On the training and validation datasets, RF achieves its best performance with AUROC values of 0.893 and 0.863, respectively. Additionally, different models were created based on the highest-ranked features, and the performance was validated using a 5-fold cross-validation technique. We wish to have a minimum set of models to avoid over-optimization. We chose the top ten features for the final classification models since there is less of a difference in performance between models based on 10 features and those based on 186 features, as measured by AUROC (0.84 and 0.83) on training and validation, respectively.

## *6.6 Conclusion*

We created IL6Pred, a web server for the scientific community, along with a standalone version that included our top models (https://webs.iiitd.edu.in/raghava/il6pred/ ). We have used all state-of-art methods for the development of prediction models. We identify certain amino-acid residues are highly abundant in IL6 inducing peptides. Our tool can be easily used by scientific community for the prediction, scanning or designing of IL6 inducing peptides. Before moving further with clinical trials and study, experimental biologists and researchers can use this tool to assess the therapeutic peptide’s ability to induce IL6. We believe that the researcher who is involved in vaccine designing and wants to include or remove IL6 producing regions will undoubtedly profit from this work.

# CHAPTER 7

## TNF-α INDUCING PEPTIDE PREDICTION

## *7.1 Introduction*

Tumor Necrosis Factor alpha (TNF-α) is a pleiotropic pro-inflammatory cytokine that promotes cellular signal activation and leukocyte trafficking to inflammatory regions (Sethi & Hotamisligil, 2021). During acute inflammation, macrophages/monocytes or other cell types (e.g., B cells, T cells, mast cells, fibroblasts) produce TNF-α cytokine, which affects haematopoiesis, immunological responses, tumour regression, and other infections (Adams et al., 2002; Aggarwal, 2003; Holbrook et al., 2019; Idriss & Naismith, 2000; Wang et al., 2014). TNF-α plays an important part in a variety of biological processes, including immunomodulation, fever, inflammatory response, tumour formation inhibition, and viral replication inhibition (You et al., 2021). TNF-α molecule occurs as a homotrimer in its active state, where it interacts to homo-trimeric TNFRs receptors to elicit signalling. The majority of TNF-α downstream actions are carried out by binding to two different receptors: TNFR1 and TNFR2 (Locksley et al., 2001). The receptors of TNF-α occurs as circulating and membrane bound molecule and interaction of TNF-α with its receptor is responsible for the diverse biological function (Idriss & Naismith, 2000). Ample of signaling pathways get elicited due to the interaction between TNF-α and its receptors such as, transcription factor activation, protein kinase and proteases activation, which overall regulate the immune response (Pasparakis & Vandenabeele, 2015).

TNF-α has been implicated in a variety of physiological consequences, including the generation of pro-inflammatory interleukins such as *IL-1* and *IL6*, according to recent research (Grivennikov & Karin, 2011; Old, 1988; Saklatvala et al., 1996). In the past it has been shown that *TNF-α* and *IL-1β* have also been implicated in the aetiology of myocardial dysfunction in ischemia-reperfusion damage, sepsis, chronic heart failure, viral myocarditis, and cardiac allograft rejection (Bryant et al., 1998; Cain et al., 1999; Muller-Werdan et al., 2006). Moreover, the interaction of TNF-α with other cytokines is responsible for regulating signaling transduction in various other disease states (Parameswaran & Patial, 2010). In the recent COVID-19 pandemic, it has been shown that its pathogenesis is associated with the cytokine storm in which the levels of cytokines such as *TNF-α, IL6, IL-2, IL-7,* and *IL-10* increased (Guo et al., 2022). Recent studies also established the strong relationship between the levels of TNF-α and IL6 with the severity of COVID-19 patients (Del Valle et al., 2020; Halim et al., 2022; Santa Cruz et al., 2021). As a result, multiple anti-TNF medications are available on the market that can reduce TNF overproduction in various illness states. Anti-TNF medication has been widely used in trials to treat rheumatoid arthritis (RA), spondyloarthropathy, psoriasis, and inflammatory bowel disease (Dreyer et al., 2009; Menegatti et al., 2019; Peyrin-Biroulet, 2010; Plasencia et al., 2013).

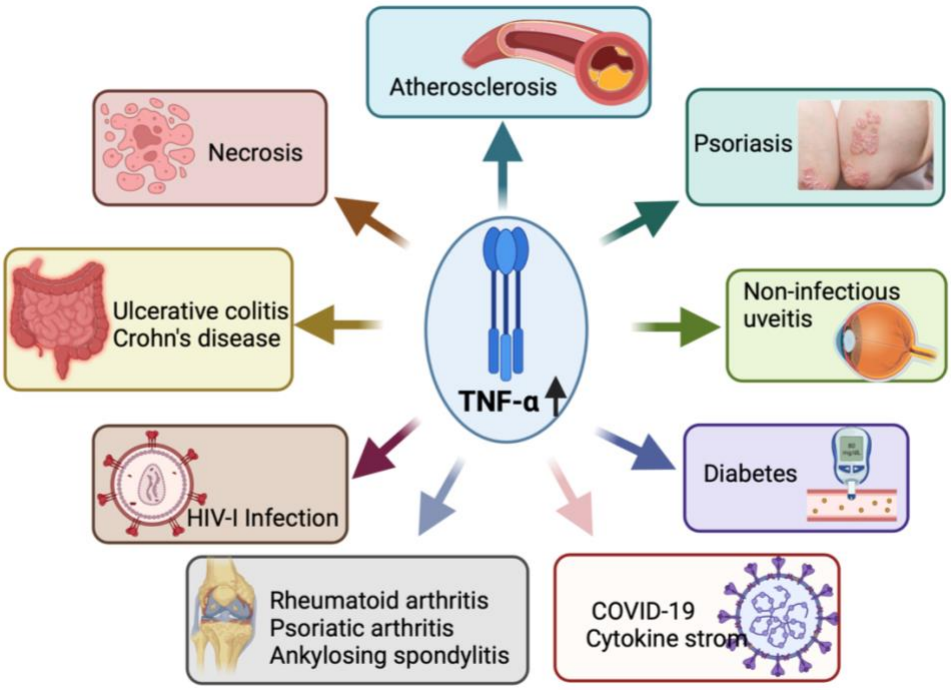



**Figure 7.1 Roles of TNF-α in various diseases, where overproduction of TNF-α cytokine found in acute and chronic inflammatory conditions**

Anti-TNF medication has recently been shown to be advantageous by not only correcting dysfunctional TNF-mediated immune systems, but also by deactivating harmful fibroblast-like mesenchymal cells (Evangelatos et al., 2022). TNF is a major cytokine implicated in various illnesses and their growing severity, according to the research. As a result, it has the potential to be a main target cytokine in disease progression. Therefore, it is the need of the hour to develop a computational approach to classify the TNF-α inducing peptides using primary structure information. In this study, we made a systematic attempt to develop a bioinformatic-ware to predict the TNF-α inducing and non-inducing epitopes. We have developed the method for human and mouse host using the experimentally validated TNF-α inducing and non-inducing peptides. Moreover, we have also used the random peptides generated using SwissProt database (Bairoch & Apweiler, 2000) to be treated as another negative dataset. We have implemented various classifiers to train and evaluate the models using training and independent dataset.

## *7.2 Material and Methods*

### *7.2.1 Overall architecture*

The complete architecture adapted in this study is exhibited in Figure 7.2.

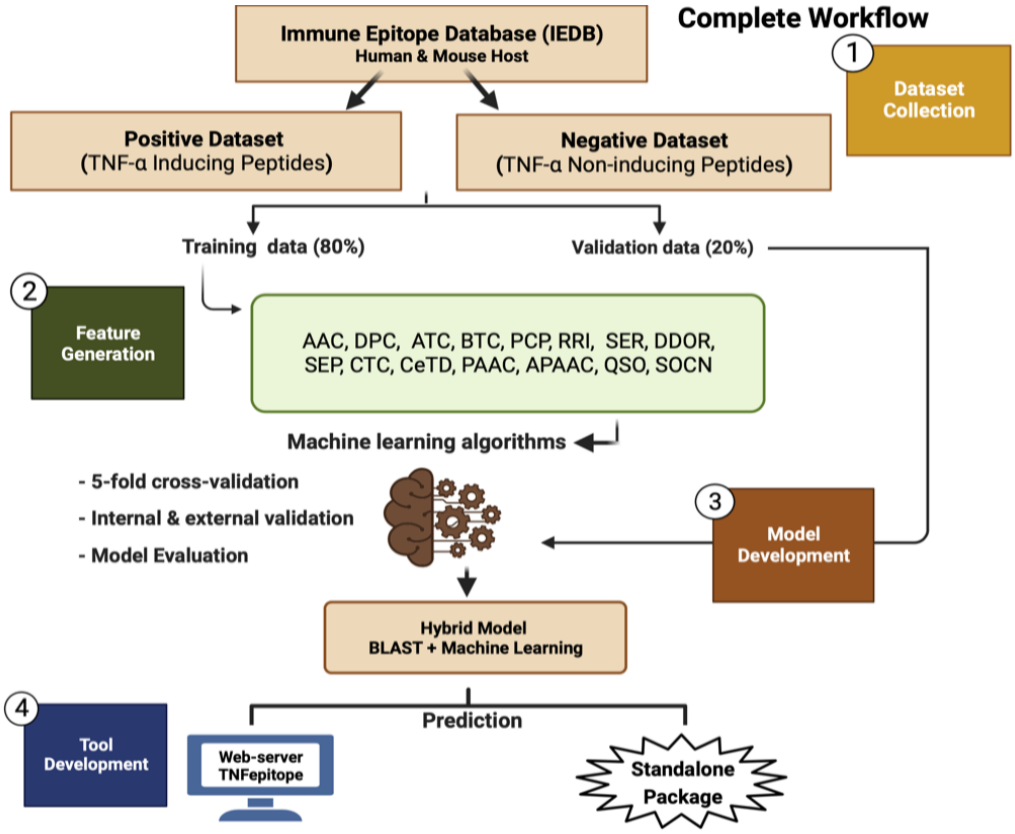



**Figure 7.2 Step-by-step representation of overall workflow of the study, including datasets collection from IEDB, feature generation using Pfeature, model evaluation and TNFepitope tool development**

## *7.2.2 Datasets*

We have downloaded 3635 experimentally validated TNF-α inducing epitopes from the immune epitope database (IEDB) (R et al., 2019), out of which 3177 belonged to human and mouse hosts. From the IEDB we have collected experimentally validated negative assays peptides and from SwissProt we have generated random peptides as negative datasets. On investigating the length distribution of these peptides, it was found that most of the peptides have length between 8-20 amino acids. Further, we removed the duplicate sequences and left with 1215 and 539 TNF-α inducing peptides belong to human and mouse host, respectively. One of the major challenge in the classification task is to choose the accurate negative instances. To overcome the issue, we have used two different negative dataset for both hosts i.e. human and mouse. The first negative dataset was compiled using IEDB dataset with 2383 experimentally validated TNF-α non-inducing epitopes. After preprocessing the peptides, we were left with 1312 peptides for human and 539 peptides for mouse. Thus, the main dataset comprises

of 1215 TNF-α inducing and 1312 non-inducing peptides for human host, and 539 TNF-α inducing and 539 non-inducing peptides for mouse host. The other negative dataset was created using the SwissProt database, by generating equal number of random peptides of length 8-20 as per the positive dataset. Therefore, in the alternate dataset for human, we have 1215 TNF-α inducing and 1215 randomly generated non-inducing peptides, whereas for mouse host, its 539 TNF-α inducing and 539 randomly generated peptides. Further, each dataset was divided in the 80:20 ratio, where 80% data was used for training purpose and remaining 20% data was kept aside for the external validation.

### *7.2.3 Analysis of peptides*

In order to understand the abundance of amino acids in TNF-α inducing peptides in comparison to the non-inducing peptides, we have calculated average amino acid composition for TNF-α inducing and non-inducing peptides in main and alternate dataset using Pfeature tool. Equation 1 was implemented to calculate the amino acids composition of each peptide in both the dataset.

$$Composition_i = \frac{NR_i}{NT} x\ 100 \qquad [1]$$

Where, $Composition_i$ signifies the amino acid composition of residue of type $i$, $NR_i$ number of residue of type $i$, $NT$ stands for total number of residues in a peptide.

### *7.2.4 WebLogo*

In order to explore the preference of amino acid residue at each position of TNF-α inducing peptides, we have developed the logo using WebLogo (Crooks et al., 2004) webserver. Generation of sequence logo require the peptides of fixed length; therefore, we have taken the eight residues (as eight is the minimum length of the peptides) from each terminus and join them to create a peptide of length 16 for each peptide in each dataset. Finally, logos were generated with positions on x-axis and bit score on y-axis signifying the preference of amino acids at each position.

### *7.2.5 Peptide features*

To develop the prediction models, the amino acid sequences should be represented by the numerical vectors of fixed length. In order to do that, we have implemented the composition module of Pfeature, which computed total of 1163 attributes for each sequence in main and alternate dataset. We have calculated 12 different types of compositional features such as, amino acid composition, dipeptide composition, atomic composition, physico-chemical properties based composition, pseudo- and

amphiphilic pseudo amino acid composition, composition enhanced transition and distribution, conjoint triad composition, residue repeat information, distance distribution of residues, Shannon-entropy based on physico-chemical properties, and quasi-sequence order. We have built models using each type of features as well as their combination.

### *7.2.6 Building of model*

Once the features were generated, the next step is to use them to build the prediction model to classify the TNF-α inducing peptides. To develop the prediction models, we have implemented the various machine learning classifiers, such as, decision tree (DT), random forest (RF), K-nearest neighbor (KNN), Gaussian Naïve Bayes (GNB), randomized extra tree (ET), logistic regression (LR), and support vector classifier (SVC), using scikit-learn library of Python. In each classifier, the parameters were hyper-tuned using grid-search approach on range of parameters. Final model was developed on the combination of parameters on which the highest performance was attained.

### *7.2.7 Similarity Search*

Further, to explore the potential of similarity search to classify the peptides into TNF-α inducing and non-inducing peptides, we have implemented BLAST (McGinnis & Madden, 2004). We have used the makeblastdb suite of NCBI-BLAST+ version 2.2.29 to create the custom database using training dataset of main and alternate dataset for human and mouse hosts. Further, the query sequences in the validation dataset hit against the customized dataset using blastp suite and record the top hit to assign the class to each query sequence. Such that, if the top-hit of the BLAST is TNF-α inducing then the query sequence was assigned as TNF-α inducing, otherwise non-inducing. To determine the ideal e-value cut-off, we run the BLAST at several e-value cut-offs ranging from 1e-6 to 1e+3.

### *7.2.8 Hybrid Model*

In order to improve the performance of prediction models, we have combined the two approaches such as alignment-based approach i.e. similarity search and alignment-free approach i.e. machine learning. In this approach, first we tried to classify the peptides using machine learning based models and calculated their probabilities. Then, similarity search based prediction were made using BLAST on optimal e-value and scores were assigned based on the hit found. If the top-hit is found out to be positive then score of 0.5 is assigned, if the top-hit is negative then score of -0.5 is assigned, other

score of 0 is assigned to the query sequence. Further, the scores from alignment-free and alignment-based method were added to get the new score, based on which the overall prediction were made.

### *7.2.10 Cross-validation*

To avoid the curse of overfitting and biasness while training the model, we have implemented five-fold cross validation on the training dataset for each dataset, as done in the previous studies (Dhall et al., 2022; Dhall, Patiyal, Sharma, Devi, et al., 2021; Dhall, Patiyal, Sharma, Usmani, et al., 2021). In this method, the entire dataset was first divided into five possible equal parts, out of which four were used for training purpose and tested on the remaining one. The same procedure was iterated five times so that each part gets the chance to be act as testing dataset. Eventually, the final performance was calculated by taking the average of performances achieved in the five iterations.

### *7.2.11 Model evaluation parameters*

To compare the generated prediction models, we have used the standard threshold-dependent and threshold-independent parameters. In threshold-dependent parameters, we have calculated sensitivity, specificity, accuracy, F1-score, and Matthews correlation coefficient (MCC). Whereas, in case of threshold-independent measures, we have computed Area Under Receiver Operating Characteristics (AUROC) curve.

## *7.3 Results*

### *7.3.1 Analysis of TNF-inducing peptides*

We investigate the preference of residues at certain positions in the TNF-α inducing epitopes for human and mouse datasets in this work. In the case of human host, TNF-α inducing epitopes, residues 'L' are highly conserved at the majority of places, although 'V' is favoured at the 9th and 16th positions; 'A' is found on the $7^{th}$ , $9^{th}$ , $10^{th}$ , $11^{th}$ , $12^{th}$ , $13^{th}$ , and $16^{th}$ positions (See Figure 7.3A). 'L' are greatly dominated on the $2^{nd}$ , $3^{rd}$ , $8^{th}$ , $9^{th}$ , $12^{th}$ , $13^{th}$ , and $16^{th}$ places in TNF-α inducing epitopes of mouse host; similarly, residue 'N' is largely conserved on the 5th and 13th positions; nevertheless, 'A' is predominated on the $5^{th}$ , $8^{th}$ , $9^{th}$ , $13^{th}$ , and $16^{th}$ positions, as illustrated in Figure 7.3B.

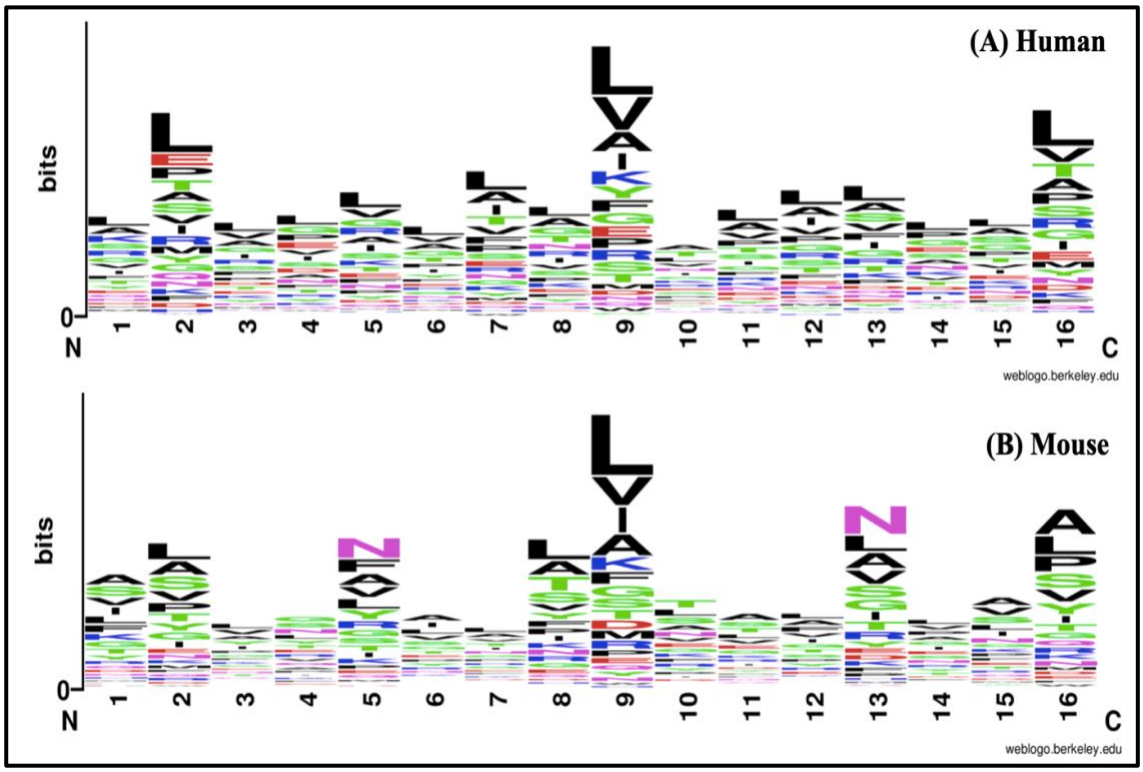



**Figure 7.3 Sequence logo generated by WebLogo tool, shows preference of different type of residues at particular positions (A) TNF-α inducing peptides in human dataset (B) TNF-α inducing peptides in mouse dataset**

For human and mouse hosts, we estimated amino acid composition from the main and alternative datasets. The average compositions of TNF-α inducing and non-inducing peptides were then computed. As shown in Figure 7.4A, amino acids such as L, V, Y, and W have a richer composition in TNF-α inducing peptides than in non-inducing and random peptides in the human dataset. Similarly, the average composition of residues such as A, I, N, and S is more prevalent in mouse TNF-α inducing peptides (See Figure 7.4B). In case of negative datasets, the average composition of D, E, G and K is higher in case of both human and mouse dataset.

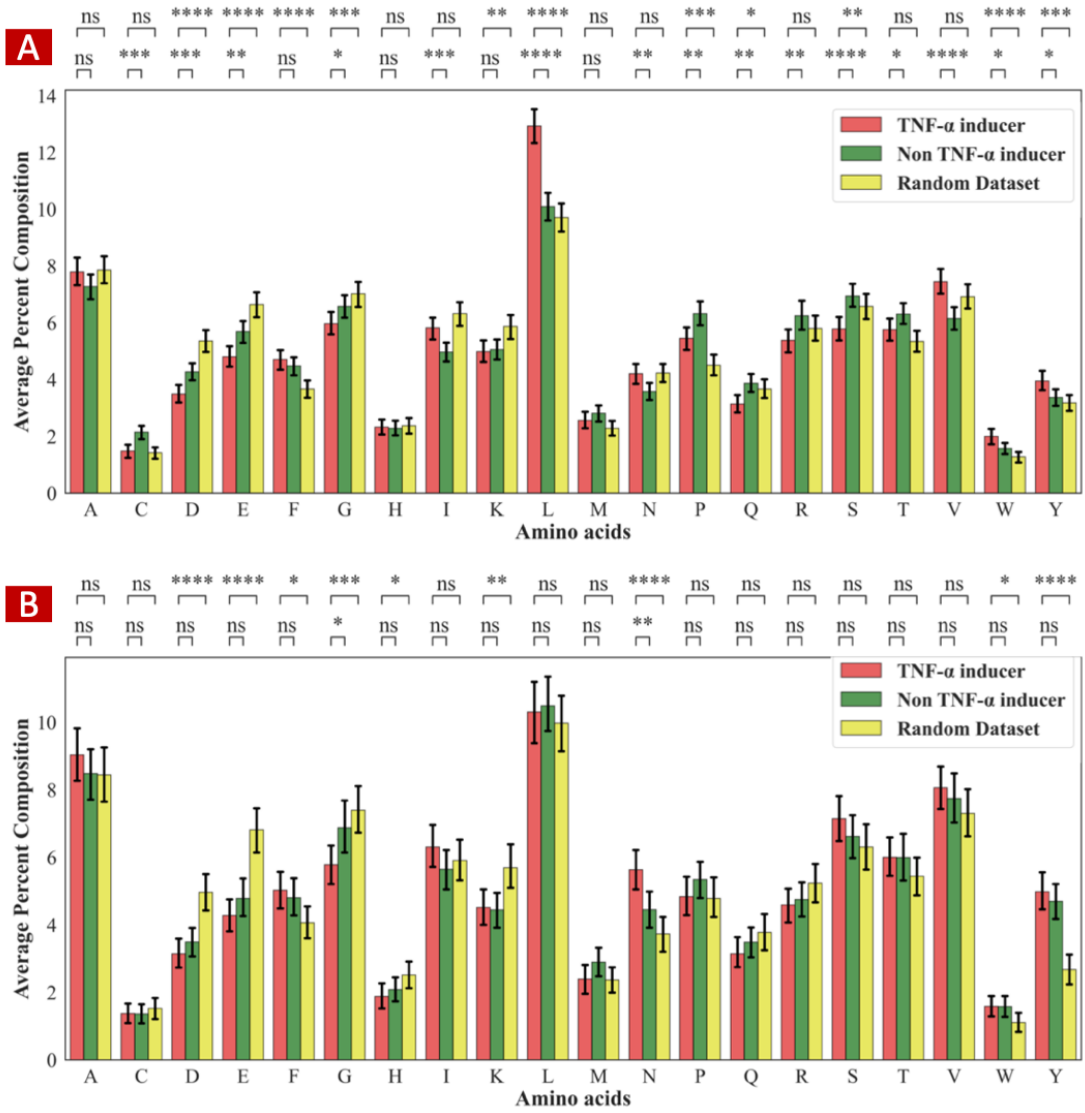



**Figure 7.4 Depicts average amino-acid composition of TNF-α inducer, non-inducer, and random peptides; where, (A) shows composition of human dataset (B) shows composition mouse datasets**

## *7.3.2 Performance of ML-based model*

We computed performance on 15 distinct descriptors in this case. We discovered that the RF and ET classifiers outperformed the other classifiers. As indicated in Table 7.1, we achieved maximum performance on the main dataset with an AUROC of 0.79 and MCC of 0.45 on the independent dataset utilising DPC-based features in the case of human hosts. APAAC and SER-based features worked well on an independent dataset as well, with an AUROC of 0.78 and an AUPRC of 0.75. Using DPC-based features, we get a maximum AUROC of 0.71, AUPRC of 0.73, and MCC of 0.31 in the case of the alternate dataset. When we combine all of the attributes, we get (0.77 and 0.71) AUROC on the main and alternate datasets, respectively. On both the main and alternate datasets, other composition-based

features perform poorly. In case of mouse dataset, on the alternate dataset with DPC as input feature, RF-based classifier performs well with an AUROC of 0.74, AUPRC of 0.76, and MCC of 0.34, as shown in Table 7.2). Similarly, employing AAC-based features on the alternate dataset, we achieved comparable results (i.e., AUROC = 0.72, MCC = 0.30, and AUPRC = 0.73). Furthermore, RRI, DDR, and APAAC perform well on alternate dataset, with AUROC>0.72. However, the performance of machine learning models on the main dataset is fairly poor.

### *7.3.3 Performance of hybrid model*

In this work, we created a hybrid model to distinguish between TNF-α inducing and non-inducing peptides. Initially, we employed the similarity search strategy (BLAST) to predict positive and negative peptides. DPC-based features outperformed other feature types in human and mouse prediction models, as demonstrated in Tables 7.1. Therefore, to create the final predictions, we blended BLAST similarity scores with machine learning scores derived using DPC features. The RF and ET-based models performed well on the main and alternate human datasets. We utilised models developed on DPC feature to compute the performance of hybrid models on separate datasets at different e-value cut-offs, as shown in Table 7.1 for human host. Aside from that, the RF-based model outperforms the other classifier on both the main and alternate mouse datasets with DPC-based features. Using the hybrid approach, on the main and alternate datasets, we achieved the best performance at e-value (1.00E-01) with AUROC of (0.70 and 0.77), AUPRC of (0.69 and 0.81), and MCC of (0.28 and 0.34), respectively (See Table 7.2).

**Table 7.1: The performance of machine learning based models on independent dataset developed using composition-based features for the main and alternate human datasets**

| Feature | Main Dataset | | | | | Alternate Dataset | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Sens | Spec | Acc | AUROC | MCC | Sens | Spec | Acc | AUROC | MCC |
| **AAC** | 55.97 | 58.56 | 57.31 | 0.63 | 0.15 | 63.37 | 66.26 | 64.82 | 0.7 | 0.3 |
| **DPC** | 72.02 | 72.62 | 72.33 | 0.79 | 0.45 | 68.72 | 61.73 | 65.23 | 0.71 | 0.31 |
| **ATC** | 55.97 | 58.56 | 57.31 | 0.63 | 0.15 | 59.67 | 58.03 | 58.85 | 0.61 | 0.18 |
| **APAAC** | 68.31 | 74.91 | 71.74 | 0.78 | 0.43 | 63.37 | 67.49 | 65.43 | 0.7 | 0.31 |
| **BTC** | 69.55 | 68.82 | 69.17 | 0.69 | 0.38 | 55.97 | 50.62 | 53.29 | 0.55 | 0.07 |
| **CETD** | 66.67 | 70.34 | 68.58 | 0.74 | 0.37 | 61.32 | 61.32 | 61.32 | 0.64 | 0.23 |
| **CTD** | 61.32 | 66.92 | 64.23 | 0.7 | 0.28 | 62.14 | 61.73 | 61.93 | 0.66 | 0.24 |
| **DDR** | 72.02 | 73.76 | 72.93 | 0.77 | 0.46 | 62.55 | 64.61 | 63.58 | 0.7 | 0.27 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **PAAC** | 68.31 | 74.14 | 71.34 | 0.78 | 0.43 | 65.02 | 65.43 | 65.23 | 0.7 | 0.31 |
| **PCP** | 64.61 | 67.68 | 66.21 | 0.73 | 0.32 | 62.96 | 63.37 | 63.17 | 0.67 | 0.26 |
| **QSO** | 62.55 | 71.86 | 67.39 | 0.72 | 0.35 | 63.79 | 65.43 | 64.61 | 0.69 | 0.29 |
| **RRI** | 62.55 | 68.06 | 65.42 | 0.73 | 0.31 | 62.96 | 57.2 | 60.08 | 0.66 | 0.2 |
| **SEP** | 63.37 | 60.84 | 62.06 | 0.69 | 0.24 | 43.62 | 57.61 | 50.62 | 0.51 | 0.01 |
| **SER** | 67.08 | 73.38 | 70.36 | 0.78 | 0.41 | 64.61 | 67.9 | 66.26 | 0.7 | 0.33 |
| **SCP** | 66.67 | 73.38 | 70.16 | 0.74 | 0.4 | 65.02 | 62.14 | 63.58 | 0.68 | 0.27 |
| **ALL_COMP** | 68.31 | 74.91 | 71.73 | 0.77 | 0.433 | 65.43 | 65.02 | 65.22 | 0.71 | 0.3 |
| **Hybrid model (DPC+ BLAST)** | 76.54 | 75.95 | 76.24 | 0.83 | 0.53 | 68.72 | 67.9 | 68.31 | 0.77 | 0.37 |

**#Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; AUROC, Area Under Receiver Operating Curve; MCC, Matthews correlation coefficient**

**Table 7.2: The performance of machine learning based models on independent dataset developed using composition-based features for the main and alternate mouse datasets**

| **Feature** | **Main Dataset** | | | | | **Alternate Dataset** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Sens** | **Spec** | **Acc** | **AUROC** | **MCC** | **Sens** | **Spec** | **Acc** | **AUROC** | **MCC** |
| **AAC** | 62.18 | 60.56 | 61.37 | 0.67 | 0.23 | 64.82 | 64.82 | 64.82 | 0.72 | 0.3 |
| **DPC** | 58.47 | 59.86 | 59.17 | 0.63 | 0.18 | 66.67 | 67.59 | 67.13 | 0.74 | 0.34 |
| **ATC** | 51.97 | 50.35 | 51.16 | 0.54 | 0.02 | 55.56 | 62.04 | 58.8 | 0.65 | 0.18 |
| **APAAC** | 62.18 | 60.09 | 61.14 | 0.65 | 0.22 | 63.89 | 65.74 | 64.82 | 0.72 | 0.3 |
| **BTC** | 51.51 | 52.44 | 51.97 | 0.55 | 0.04 | 51.85 | 58.33 | 55.09 | 0.56 | 0.1 |
| **CETD** | 56.15 | 58.24 | 57.19 | 0.62 | 0.14 | 63.89 | 66.67 | 65.28 | 0.7 | 0.31 |
| **CTD** | 51.51 | 53.13 | 52.32 | 0.56 | 0.05 | 65.74 | 63.89 | 64.82 | 0.68 | 0.3 |
| **DDR** | 56.85 | 59.86 | 58.35 | 0.62 | 0.17 | 69.44 | 67.59 | 68.52 | 0.74 | 0.37 |
| **PAAC** | 60.79 | 61.02 | 60.91 | 0.65 | 0.22 | 67.59 | 65.74 | 66.67 | 0.72 | 0.33 |
| **PCP** | 57.77 | 61.49 | 59.63 | 0.61 | 0.19 | 56.48 | 69.44 | 62.96 | 0.7 | 0.26 |
| **QSO** | 58.01 | 58.47 | 58.24 | 0.6 | 0.17 | 61.11 | 70.37 | 65.74 | 0.73 | 0.32 |
| **RRI** | 59.86 | 60.79 | 60.33 | 0.63 | 0.21 | 65.74 | 66.67 | 66.2 | 0.75 | 0.32 |
| **SEP** | 55.68 | 54.06 | 54.87 | 0.57 | 0.1 | 36.11 | 51.85 | 43.98 | 0.45 | -0.12 |
| **SER** | 60.56 | 62.41 | 61.49 | 0.67 | 0.23 | 67.59 | 69.44 | 68.52 | 0.73 | 0.37 |
| **SCP** | 57.77 | 58.47 | 58.12 | 0.61 | 0.16 | 60.19 | 69.44 | 64.82 | 0.69 | 0.3 |
| **ALL_COMP** | 62.96 | 62.96 | 62.96 | 0.67 | 0.26 | 64.81 | 68.51 | 66.67 | 0.73 | 0.33 |
| **Hybrid model (DPC+ BLAST)** | 62.62 | 65.42 | 64.02 | 0.7 | 0.28 | 66.36 | 67.29 | 66.82 | 0.77 | 0.34 |

**#Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; AUROC, Area Under Receiver Operating Curve; MCC, Matthews correlation coefficient**

## *7.4 Service to scientific community*

We created the ‘TNFepitope’ web service for the scientific community for the prediction of TNF-α inducing and non-inducing epitopes based on sequencing information (See Figure 7.5). The website now includes the best prediction models for human and mouse hosts. The server has five primary modules: (i) Predict; (ii) Design; (iii) Scan; (iv) Blast Search; and (v) Standalone. The ‘Predict’ feature assists users in distinguishing TNF-inducing peptides from non-inducing peptides. Figure 7.5, 7.6 and 7.7 depicts the homepage and usage of “Predict” module of TNFepitope server. The ‘Design’ module allows the user to design/create all conceivable mutations of the query sequence and forecast whether or not they may cause TNF-α release. The ‘Scan’ module enables the user to map/scan the TNF-α secretion section of a protein sequence. The 'BLAST Search' module is solely based on a similarity search method, and the input sequence is compared to a specific database of known TNF-α inducing and non-inducing peptides. Based on the similarities, the provided amino-acid sequence is anticipated to be a TNF-α inducer/non-inducer. The ‘TNFepitope’ server was created with HTML, JAVA, and PHP scripts and is compatible with a variety of devices including laptops, iPhones, and phones. The webserver (https://webs.iiitd.edu.in/raghava/tnfepitope), standalone package (https://webs.iiitd.edu.in/raghava/tnfepitope/package.php), and GitLab (https://gitlab.com/raghavalab/tnfepitope) are all available for free use.

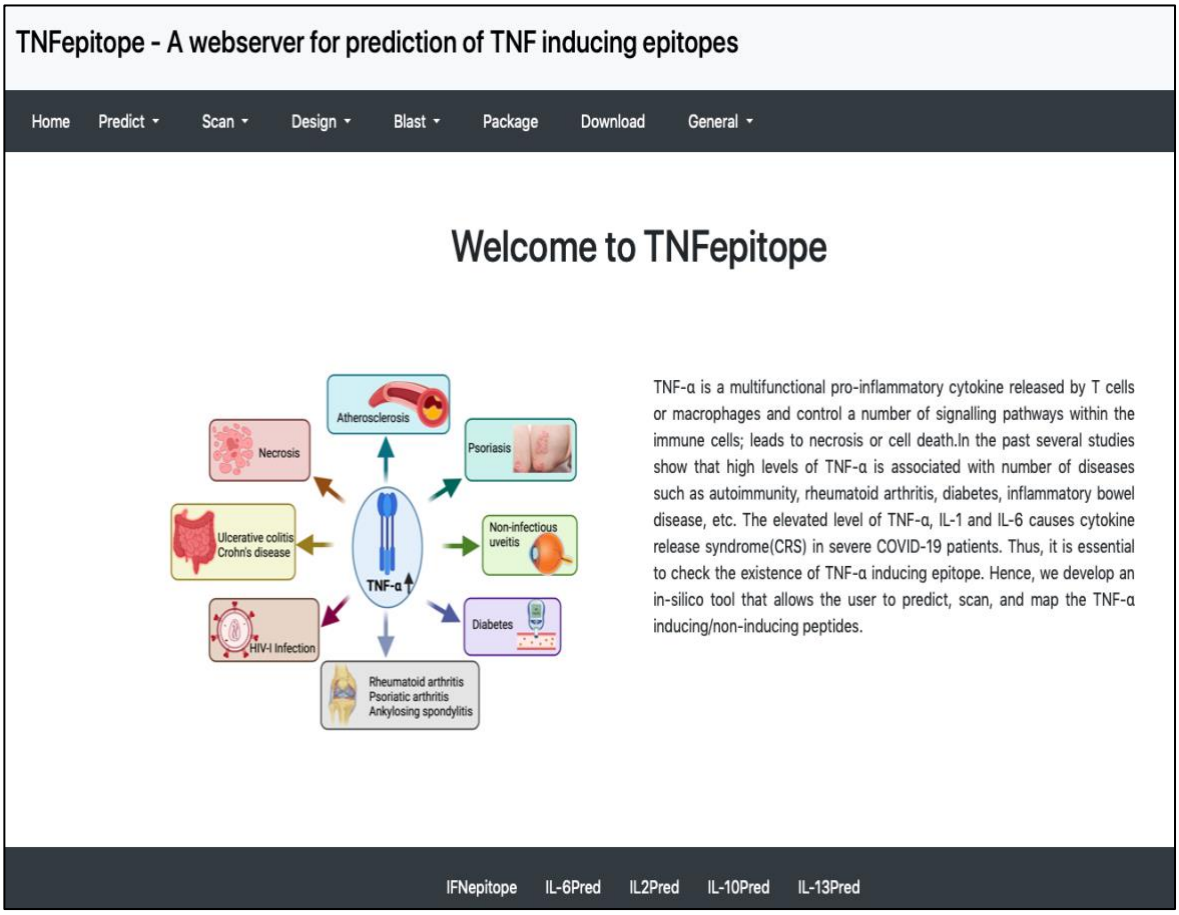


**Figure 7.5: Homepage of TNFepitope Webserver**

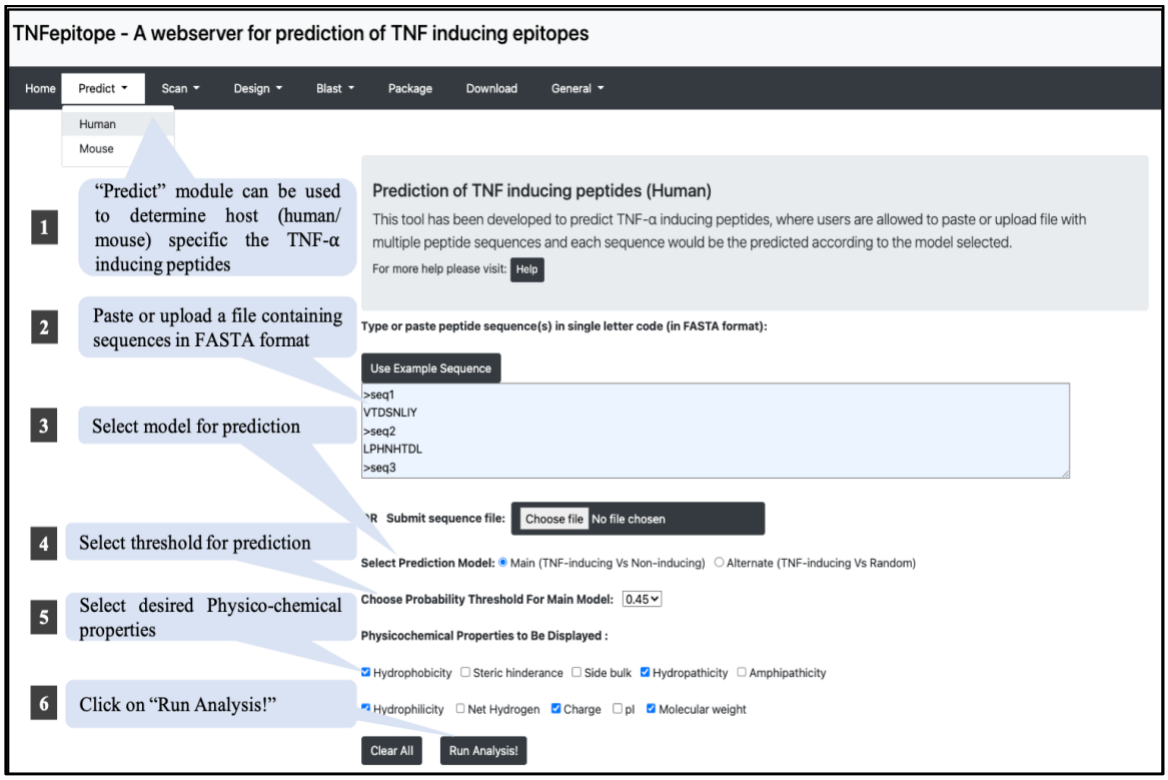



**Figure 7.6: Shows data submission page of "Predict" module of TNFepitope server**

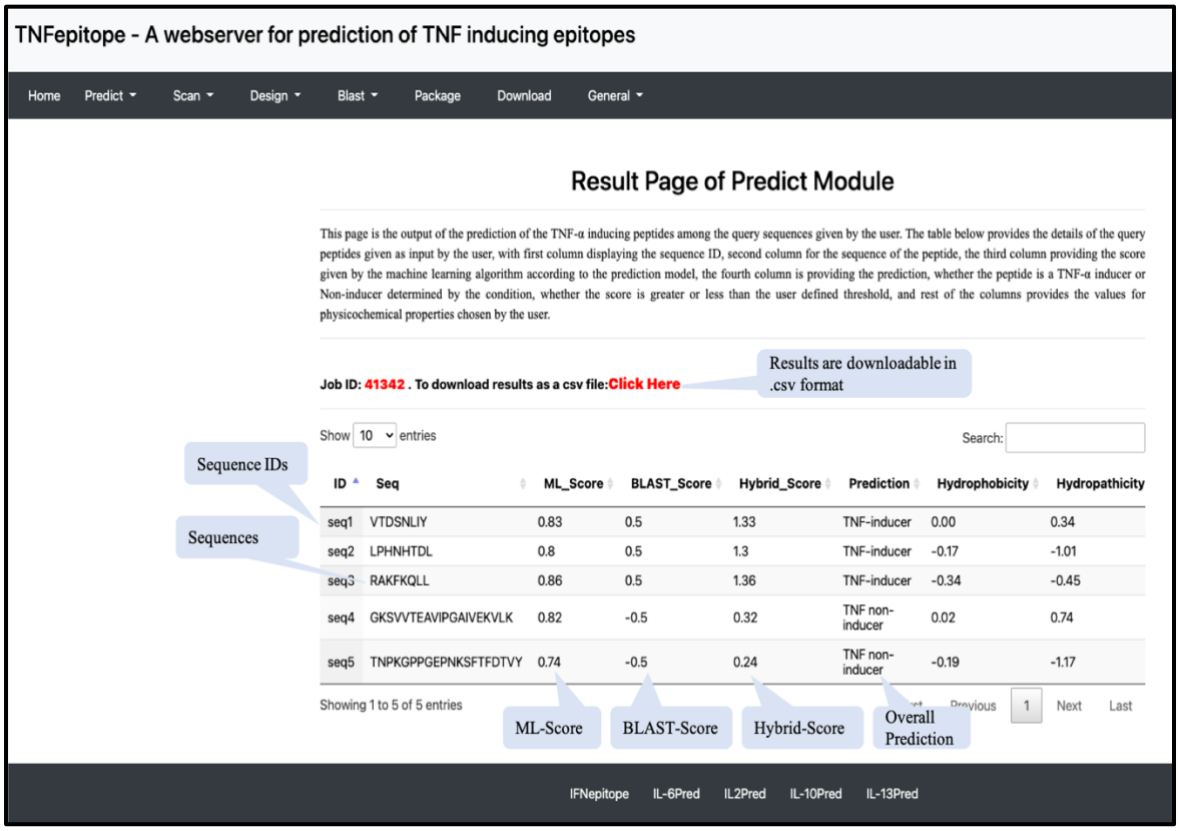


**Figure 7.7 Result page of "Predict" module, which provides query sequence, machine learning, BLAST and Hybrid model scores with prediction as TNF-inducer/non-inducer**

## *7.5 Discussion*

The major histocompatibity region of chromosome 6, encode number of HLA molecules which are required for peptide binding and presentation and cytokines genes such as TNF, LTA and LTB which are important for inflammation (Shiina et al., 2009). Whereas, TNF or tumor necrosis factor is a significant inflammatory cytokine that is generated by T cells or macrophages and regulates a number of immune cell signaling pathways. The major role of TNF is to cause necrosis or cell death (Gershenwald et al., 1998; Shen et al., 2018). A variety of biological responses, including cell proliferation, differentiation, and survival, are managed by these pathways. TNF cytokine is used to treat cancer and has anti-cancer properties by generating immune response, inflammation, and tumors cell apoptosis. However, incorrect or overzealous activation of the TNF signaling pathway can lead to the development of pathological conditions like HIV-1, anorexia, cachexia, obesity, and autoimmune diseases such rheumatoid arthritis, diabetes, and inflammatory bowel disease (Adegbola et al., 2018; Lane et al., 1999; Montfort et al., 2019). Numerous proteins, including, are encoded within the major histocompatibility complex area. Several TNF-inhibitors have been created and given the green light for clinical usage to treat disorders linked to aberrant or excessive TNF-secretion, including infliximab, etanercept, golimumab, certolizumab, and adalimumab. Studies show that COVID-19 patients have greater levels of soluble TNF than the healthy control group. Therefore, it is necessary to apply anti-TNF medication or to look for TNF-inducing epitopes in a variety of disorders.

In the present work, we have made an effort to comprehend the characteristics of TNF-inducing peptides and have developed a prediction model to identify the epitopes that can cause TNF-secretion. Datasets are crucial for creating machine learning models, thus we have gathered peptides for both human and mouse TNF-inducing and non-inducing reactions that have undergone experimental validation. We created random peptides using the Swiss-Prot database for the alternate negative dataset. To learn more about compositional analysis, positioning preference, and sequence logo. Our analyses were carried out on both human and mouse datasets, we discovered that TNF-inducing epitopes are abundant in the amino acid residue (L). Moreover, we observed that 105 (28%) out of 365 IL6 inducing peptides also induce TNF-α. Composition of IL6 and TNF-α inducing peptides shows reasonably good similarity but not-identical. Number of studies revealed that leucine amino-acid

controls the production of inflammatory cytokines including (IL6 and TNF-alpha) (Cruz et al., 2017; Kubo et al., 2020; S. Q. Liu et al., 2018).

Then, using the standalone software, we used "Pfeature" to compute 15 different types of compositional features. We have created prediction models using a variety of machine-learning classifiers. According to our findings, di-peptide composition-based characteristics outperformed other features for the mouse and human models. On the independent human and mouse dataset, we have obtained the maximum AUROC of 0.79 and 0.74 using di-peptide composition-based features. Notably, on the independent datasets for humans and mice, our hybrid model (BLAST + machine learning) beat others with an AUROC of 0.83 and 0.77. However, our models' accuracy is just about 70%, which is quite low. Creating HLA-specific prediction models in the future could increase the accuracy of models. These models could predict TNF-inducing peptides that were specific to HLA alleles.

## *7.6 Conclusion*

In this study, we have developed a variety of machine learning based models to classify the host specific TNF-α inducing peptides using sequence information for human and mouse, in this study. To differentiate TNF-α inducing peptides from non-inducing peptides, we developed machine learning-based models using diverse composition based features. One of the study's main objectives is to aid the scientific community. We developed a user-friendly web server (https://webs.iiitd.edu.in/raghava/tnfepitope) that allows users to determine whether or not a particular peptide sequence has the potential to induce TNF-α release. We have also provided the Python- and Perl-based standalone package which can be used to predict the TNF-α inducing regions in the large dataset such as entire proteome or in the absence of internet. We hope that our study will benefit researchers in the development of computer-aided vaccine design, allowing them to construct subunit vaccines that elicit the optimal immune response against a variety of TNF-α associated disorders. We develop a standalone software and a web server called TNFepitope for the scientific community using the best models available. Furthermore, we have provided a web platform named TNFepitope (https://webs.iiitd.edu.in/raghava/tnfepitope ) offers tools for predicting, designing, and scanning the TNF-inducing regions.

# CHAPTER 8

## IDENTIFICATION OF IFN-γ INDUCING PEPTIDE

## *8.1 Introduction*

Cytokines are molecular messengers of innate and adaptive immunity that allow immune cells to communicate in paracrine and autocrine (Conlon et al., 2019). When the immune system functions, both innate and adaptive components are engaged in identifying the stress and cytokines providing effective response (Kursunel & Esendagli, 2016). Interferons (IFNs) are pleiotropic cytokines (Castro et al., 2018) that belong to a protein family (Farrar & Schreiber, 1993) and play an essential role in innate and acquired immune responses, (Zaidi & Merlino, 2011) with antiviral, anticancer, and immunomodulatory activities, and serve as central immune response coordinators (Castro et al., 2018). Interferons are agents or substances that inhibit viral replication and protect cells from viral infection (Castro et al., 2018; Schroder et al., 2004). IFNs are classified into three types: (Castro et al., 2018) Type I IFNs (IFN α and IFN β), type II IFNs (IFN-γ), and the newly found type III IFNs are distinguished by their ability to bind certain receptors (Conlon et al., 2019).

IFN-γ is a tiny protein that occurs as a 34-kDa homodimer that can increase host defence and immunopathologic processes (Reljic, 2007). Its receptor can be found on all nucleated cells (Reljic, 2007). IFN-γ is produced by a diverse range of lymphocytes, primarily T and NK cells such as CD4+ and CD8+ T cells, Treg cells, and FoxP3+ cells. Monocytes, macrophages, dendritic cells, and neutrophil granulocytes all generate this cytokine (Costela-Ruiz et al., 2020). IFN-γ is involved in intracellular communication, tumour cell identification and eradication (Zaidi & Merlino, 2011) as well as various immune, adaptive immunological functions and inflammatory processes (Costela-Ruiz et al., 2020). IFN-γ and IFN-α/β both boost MHC class I protein expression, but only IFN-γ is an efficient inducer of MHC class II expression (Shtrichman & Samuel, 2001). The pro-inflammatory and anti-inflammatory properties of IFN-γ is shown in Figure 8.1. IFN-γ primary role is to upregulate MHC class I molecules, which aid in antigen priming and presentation in professional antigen-presenting cells (Zaidi & Merlino, 2011). It was discovered that the serum of COVID-19 patients has greater IFN-γ levels than that of healthy individuals, and it was postulated that this and other cytokines may be elevated due to Th1 and Th2 cell activation. Increased IFN-γ levels have been linked to increased viral load and lung injury (Costela-Ruiz et al., 2020).

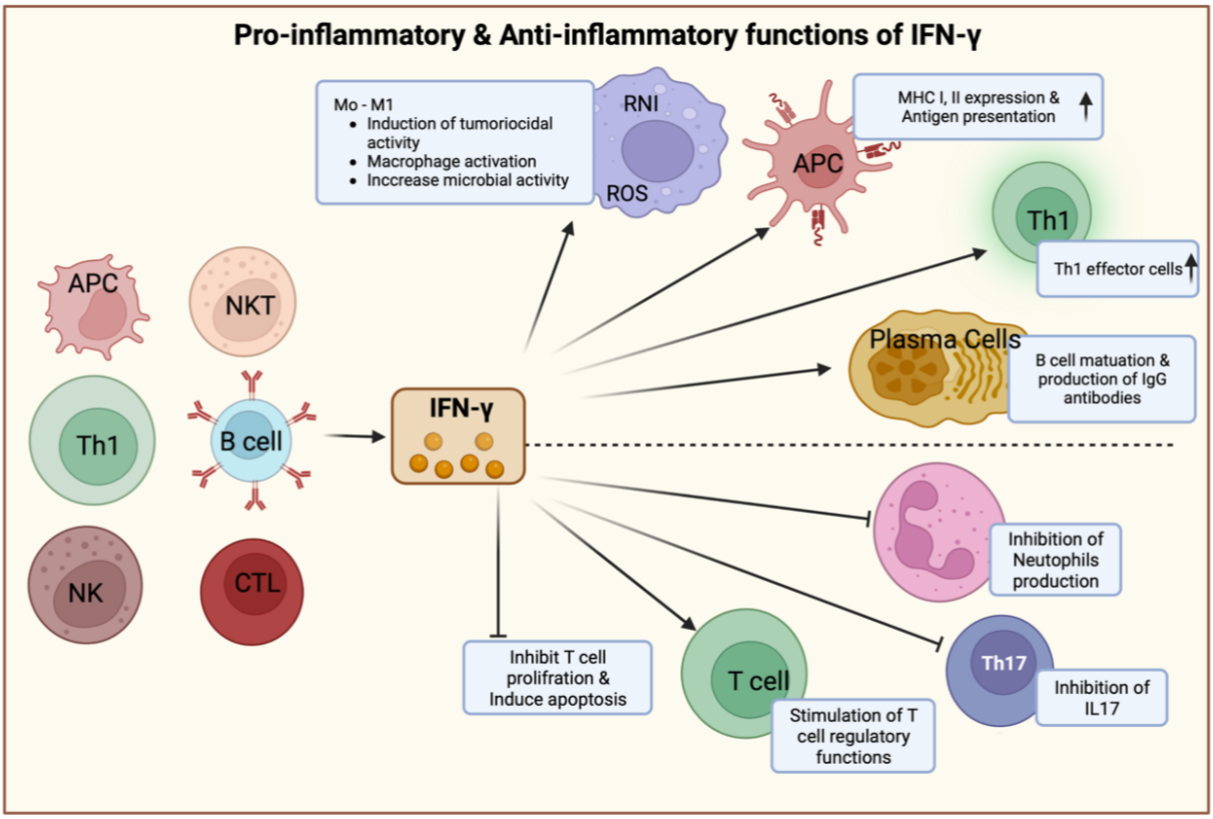



**Figure 8.1 Schematic representation of production of IFN-γ and its functions**

In the tumour microenvironment (TME), IFN-γ consistently orchestrates both pro-tumorigenic and antitumorigenic immune responses. Secreted pro-inflammatory cytokines bind to their receptors on IFN-producing cells and activate transcription elements such as members of the signal transducer and activator of transcription (STAT) family, specifically STAT4, T-box transcription factor (T-bet), activator protein 1 (AP-1), or Eomes, which further drive IFN- production (Jorgovanovic et al., 2020). Furthermore, IFN may cause apoptosis in tumor-specific T-cells, impairing antitumor immunity. Inhibiting IFN is a strategy for disrupting immunosuppressive tumour microenvironments or suppressing IFN-induced epigenomic and transcriptome alterations in tumour cells that allow immune escape (Mojic et al., 2017). Therefore, it is important to identify IFN-γ inducing peptides or epitopes in order to develop subunit vaccines against number of diseases and cancer. In the current, study we attempted to develop an updated method of IFNepitope tool. Here, we have used huge sequence datasets and generated host-specific tool for human and mouse. The experimentally validated peptides selected from IEDB database and machine learning algorithms were used for the development of prediction models.

## *8.2 Material and methods*

### 8.2.1 Creation of dataset

From the immune epitope database, we have extracted IFN-γ inducing peptides/epitopes. We then sorted the dataset by host and discovered that most of the peptides have been experimentally validated on human or mouse hosts, with only a few epitopes available for other hosts. As a result, we only chose two significant hosts (i.e., human and mouse). Similarly, we have collected experimentally validated negative assays datasets for human and mouse species from IEDB database. We examined the length distribution of epitopes and discovered that the majority of peptides have 8-20 amino-acid residues. We obtained 25492 and 7983 IFN-γ inducing epitopes for human and mouse, respectively, after deleting redundancy. We have negative datasets for both humans and mice in this investigation. The human negative dataset, encompassing 61681 experimentally confirmed epitopes with a range of (8-20 amino acids). In the instance of the mouse host, we obtain 27837 distinct IFN-γ non-inducing epitopes with a range of (8-20 amino acids). Finally, the main human dataset contains 25492 IFN-γ inducing and 61681 IFN-γ non-inducing peptides. In the case of mouse host, we obtain a total of 7983 IFN-γ inducing and 27837 non-inducing peptides. Following the generation of final datasets for human and mouse hosts, each dataset was separated into a training and an independent/validation set. The entire dataset was divided into an 80:20 ratio, with 80% data used to train the models and 20% data used for validation.

### 8.2.2 Analysis of IFN- γ inducing peptides

Pfeature was used to compute the amino acid composition (AAC). Using compositional analysis, we can see how similar distinct peptide sequences from positive and negative samples are. We built a feature vector of length 20 using the following equation 1, which specifies the percent composition of 20 amino-acid residues.

$$AAC_i = \frac{AAR_i}{Total\ number\ of\ residues} \times 100$$

where $AAC_i$ and $AAR_i$ are the percentage composition and number of residues of type i in a peptide, respectively.

### 8.2.3 Two sample logo

To understand the preference of individual amino acids at a specific position, we develop a two-sample logo (TSL). The TSL tool requires a defined length criteria for the input sequence vector. In both datasets, the peptide must be at least eight residues long. As a result, we extract eight residues from a peptide's N-terminus and eight residues from its C-terminus. These sections were linked to form a 16-residue sequence that corresponded to each sequence in the negative and positive datasets.

### *8.2.4 Feature extraction*

In the current study, we estimated a wide range of characteristics utilising peptide sequence information. To calculate the composition-based features for our datasets, we used the Pfeature [31] standalone software. In all positive and negative datasets, we computed a total of 1163 characteristics for each epitope/peptide sequence. We calculated twelve different types of descriptors/features, including AAC (Amino acid composition), DPC (Di-peptide composition), APAAC (Amphiphilic pseudo amino acid composition), ATC (Atomic composition), CETD (Composition-enhanced transition distribution), DDR (Residue distance distribution), PAAC (Pseudo amino acid composition), PCP (Physico-chemical properties composition), QSO (Quasi-se We developed prediction in this investigation.

### *8.2.5 Model building techniques*

We employed a variety of machine learning methods to create the prediction models, including Random Forest (RF), Decision Tree (DT), Gaussian Naive Bayes (GNB), Logistic Regression (LR), Support Vector Classifier (SVC), K-Nearest Neighbor (KNN), and Extra Tree (ET). The parameters were trained on the training dataset, and predictions were performed on the independent dataset. The python library scikit-learn was utilised in the study to create multiple classifiers. To avoid the curses of bias and overfitting, we used a five-fold cross validation technique. The training dataset was partitioned into five equal sets for five-fold cross-validation, with four sets used for training and the fifth set used for testing. This procedure is performed several times.

### *8.2.6 Evaluation of model*

The sensitivity, specificity, accuracy, Area Under Receiver Operating Characteristics (AUROC) curve, Matthews Correlation Coefficient (MCC), and F1-score were used to evaluate the performance of

various models. We calculated both threshold-dependent metrics (such as sensitivity, specificity, accuracy, and MCC) and independent parameters such as AUROC and AUPRC.

## 8.3 Results

### 8.3.1 Composition analysis

We computed amino acid composition using the human and mouse datasets. The average compositions of IFN-inducing and non-inducing peptides were computed. After that, the difference in the composition of each amino-acid is computed for human and mouse dataset. As illustrated in Figure 8.2, amino acids such as K, M, N, P, and Q are more abundant in IFN-inducing peptides than in non-inducing in the human dataset. Similarly, the average composition of residues such as A, E, G, and P is higher in mouse IFN-inducing peptides.

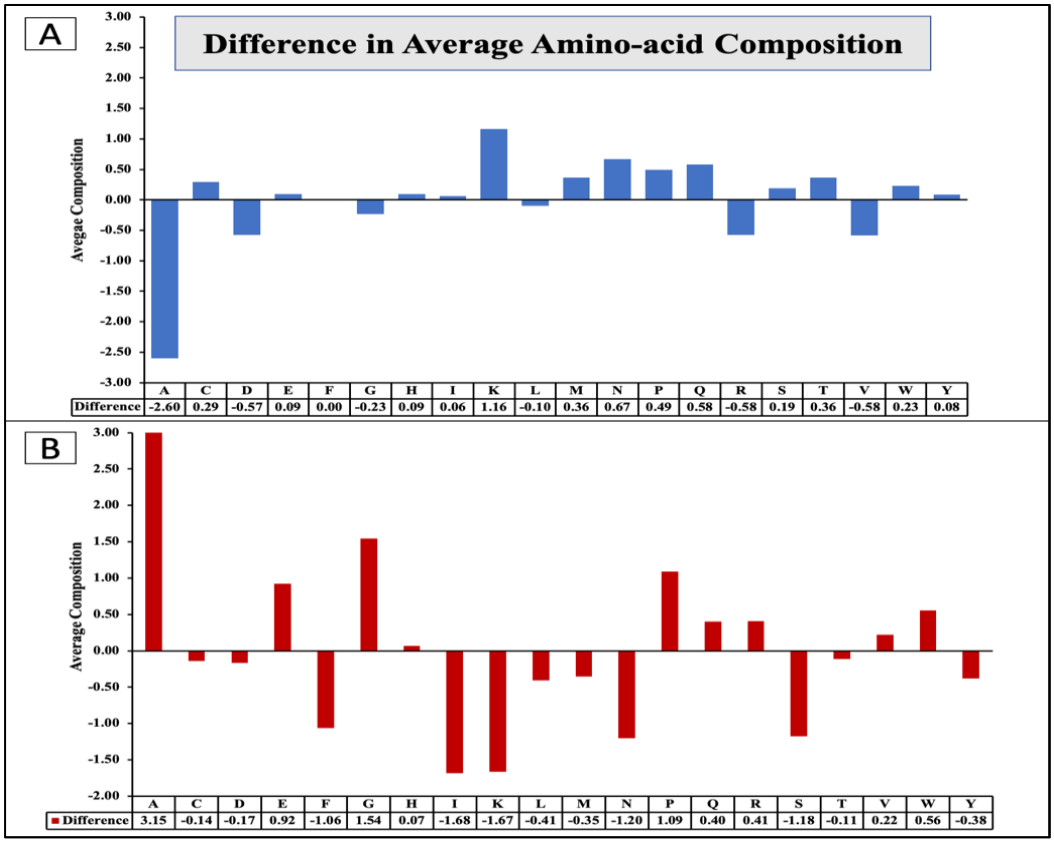



**Figure 8.2: Difference in average amino-acid composition IFN-γ inducing and Non IFN-γ inducing epitopes (A) for human dataset and (B) for mouse dataset**

### 8.3.2 Positional analysis

In this paper, we look at the preference of residues at specific places in IFN-inducing epitopes for human and mouse datasets. In the case of human host IFN-inducing epitopes, residues 'K' are highly conserved at the majority of positions, though 'P' is preferred at the 6$^{th}$ and 7$^{th}$ positions; 'A' is

preferred in most of the positions (See Figure 8.3). In IFN-inducing epitopes of mouse host, residues 'P' are greatly dominated on the 4th, 6th, 7th, 14th, and 16th positions; similarly, residue 'Y' is largely conserved on the 7th and 15th positions; as illustrated in Figure 8.3.

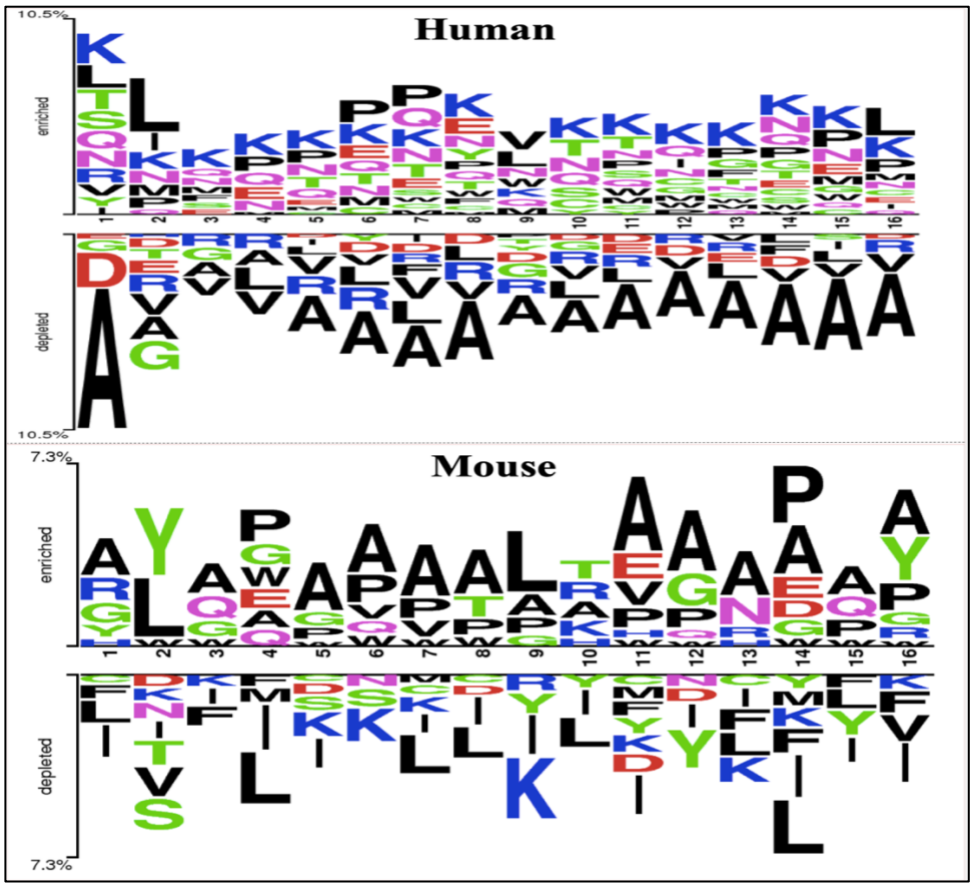



**Figure 8.3 Representation of two sample logo of IFN-γ inducing and IFN-γ non-inducing peptides for human and mouse hosts**

## *8.3.3 Performance of machine-learning models*

### *8.3.3.1 Model for human*

In this scenario, we calculated performance using AAC and DPC based descriptors. The RF and ET classifiers outperformed the other classifiers, as shown in Table 8.1, we were able to maximise performance using the independent dataset using AAC-based features for human hosts with an AUROC of 0.79 and MCC of 0.43. However, we achieved maximum AUCROC of 0.83 on independent dataset using DPC based features.

**Table 8.1: The performance of machine learning based models developed on various composition-based features using human independent dataset**

| Feature Type | Sensitivity | Specificity | Accuracy | AUROC | MCC |
|---|---|---|---|---|---|
| **AAC** | 72.58 | 73.43 | 73.18 | 0.79 | 0.43 |
| **DPC** | 74.46 | 76.06 | 75.6 | 0.83 | 0.47 |
| **CTC** | 63.42 | 62 | 62.41 | 0.68 | 0.23 |
| **ATC** | 55.26 | 58.91 | 57.84 | 0.6 | 0.13 |
| **RRI** | 63.32 | 61.48 | 62.02 | 0.68 | 0.23 |
| **SER** | 71.95 | 71.87 | 71.9 | 0.78 | 0.41 |
| **SOC** | 50.79 | 49.57 | 49.92 | 0.51 | 0 |
| **APAAC** | 73.44 | 71.69 | 72.2 | 0.79 | 0.42 |
| **PAAC** | 72.6 | 71.71 | 71.97 | 0.79 | 0.41 |
| **QSO** | 66.77 | 65.75 | 66.05 | 0.72 | 0.3 |
| **BTC** | 57.75 | 55.93 | 56.46 | 0.59 | 0.13 |
| **DDR** | 71.38 | 67.49 | 68.63 | 0.76 | 0.36 |
| **CETD** | 68.56 | 71.55 | 70.67 | 0.76 | 0.37 |
| **SPC** | 70.48 | 68.71 | 69.23 | 0.76 | 0.36 |
| **PCP** | 70.69 | 69.51 | 69.86 | 0.76 | 0.37 |

**# AUROC, Area Under Receiver Operating Curve; MCC, Matthews correlation coefficient**

### *8.3.3.2 Model for mouse*

In addition, we computed performance using AAC and DPC based descriptors on mouse dataset. As shown in previous results, RF and ET classifiers outperformed, results provided in Table 8.2. The models build using mouse dataset perform poor on AAC based features and achieved an AUROC 0.71 independent datasets. Whereas, ET based models achieved 0.756 AUROC on independent datasets using DPC based features.

**Table 8.2: The performance of machine learning based models developed on various composition-based features using mouse independent dataset**

| Feature | Sensitivity | Specificity | Accuracy | AUROC | MCC |
|---|---|---|---|---|---|
| **AAC** | 66.479 | 63.697 | 64.317 | 0.710 | 0.254 |
| **DPC** | 68.860 | 69.014 | 68.979 | 0.756 | 0.323 |
| **ATC** | 56.328 | 55.021 | 55.312 | 0.573 | 0.095 |
| **APAAC** | 68.734 | 62.853 | 64.163 | 0.717 | 0.265 |
| **BTC** | 60.401 | 52.039 | 53.902 | 0.591 | 0.104 |
| **CETD** | 63.596 | 60.895 | 61.497 | 0.677 | 0.205 |
| **CTC** | 63.596 | 61.577 | 62.027 | 0.673 | 0.211 |
| **DDR** | 63.596 | 61.415 | 61.901 | 0.676 | 0.21 |

| | | | | | |
|---|---|---|---|---|---|
| **PAAC** | 67.982 | 63.984 | 64.875 | 0.713 | 0.269 |
| **PCP** | 63.910 | 58.990 | 60.087 | 0.654 | 0.191 |
| **QSO** | 66.855 | 61.721 | 62.865 | 0.693 | 0.239 |
| **RRI** | 63.409 | 59.709 | 60.533 | 0.65 | 0.193 |
| **SER** | 65.602 | 64.200 | 64.512 | 0.706 | 0.251 |
| **SOC** | 52.130 | 52.955 | 52.771 | 0.534 | 0.042 |
| **SPC** | 60.840 | 60.374 | 60.477 | 0.642 | 0.178 |

**# AUROC, Area Under Receiver Operating Curve; MCC, Matthews correlation coefficient**

## *8.4 Web-implementation*

We have developed IFNepitope 2.0 for the identification of peptides that induce and do not induce IFN-gamma. The web server's front end was created utilising HTML5, JAVA, CSS3, and PHP scripts. It is built using responsive templates that change the screen size to fit the device. It works with practically all current gadgets, including smartphones, tablets, iMacs, and desktop computers. Three main modules, including Predict, Design, Protein Scan, are included in the web server. The “Predict” module allow the user to identify the IFN-gamma inducing and non-inducing peptides. User can submit or paste multiple sequences in FASTA format. The “Design” module of our server provide the facility to the user to modulate the sequence from IFN-inducer to non-inducer via incorporating minimum mutations in the query sequence. The third module is “Scan”, which is used for the screening of interferon inducing peptides in the input protein sequences. The results generated by all three modules exhibited in the tabular format which is downloadable in the “.csv” format. We anticipate these module can be used for the prediction of vaccine candidates in the antigenic sequences or can be used for designing subunit vaccine which have the capacity to induce interferon gamma. The homepage of our server and the example utility of our server is provided in Figure 8.4, 8.5 and 8.6.

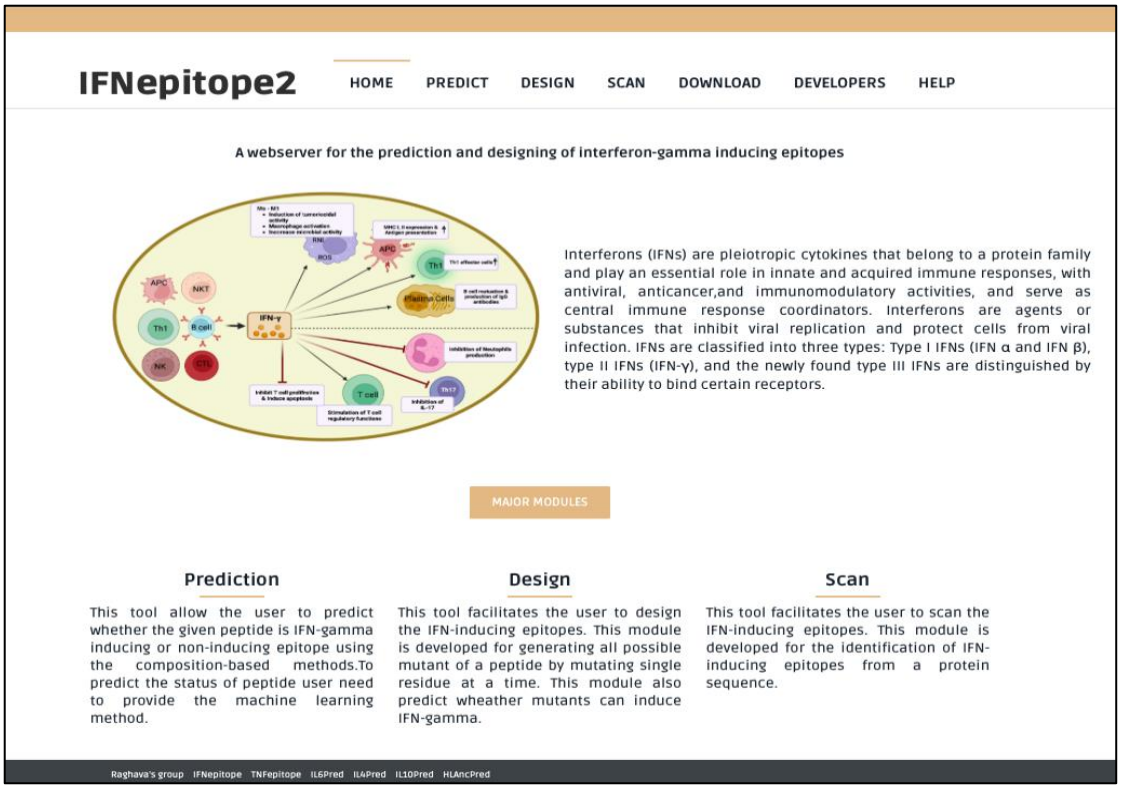



**Figure 8.4 Home-page of IFNepitope 2.0 website**

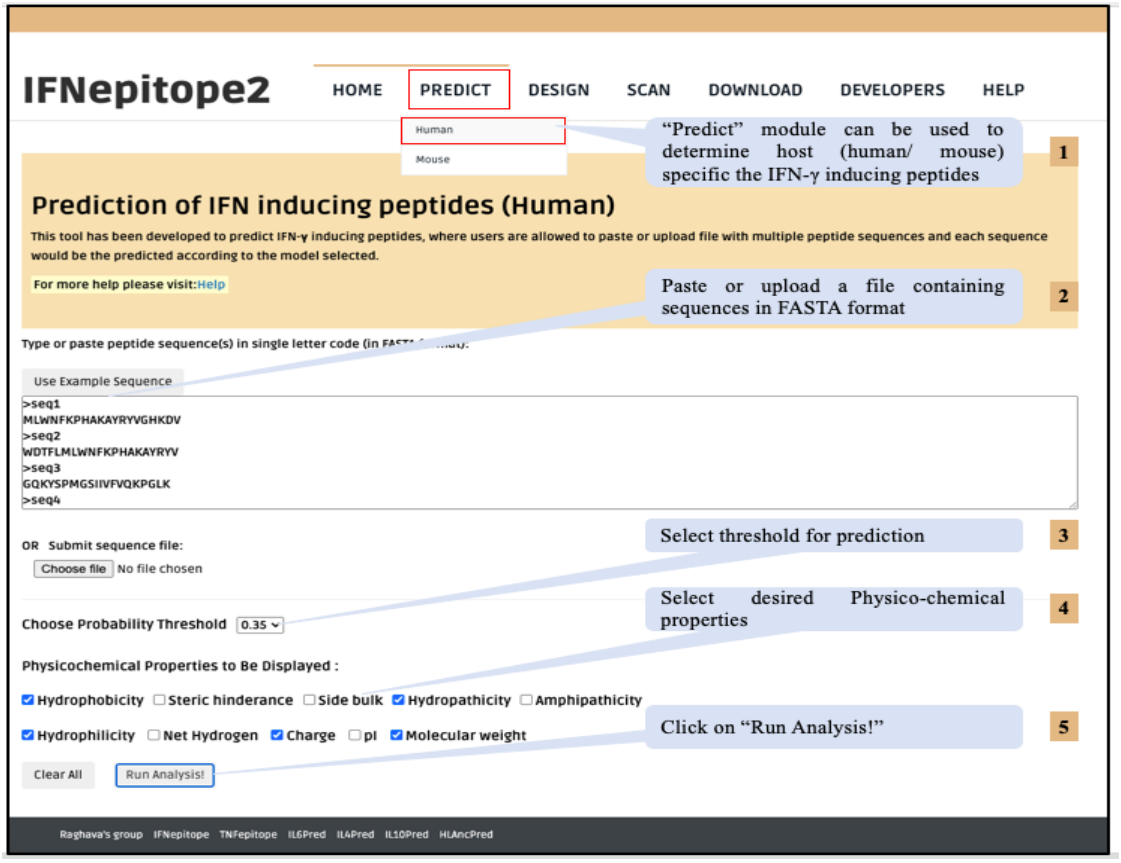



**Figure 8.5 Steps involved in the submission of sequence using 'Predict' module of IFNepitope 2.0 website**

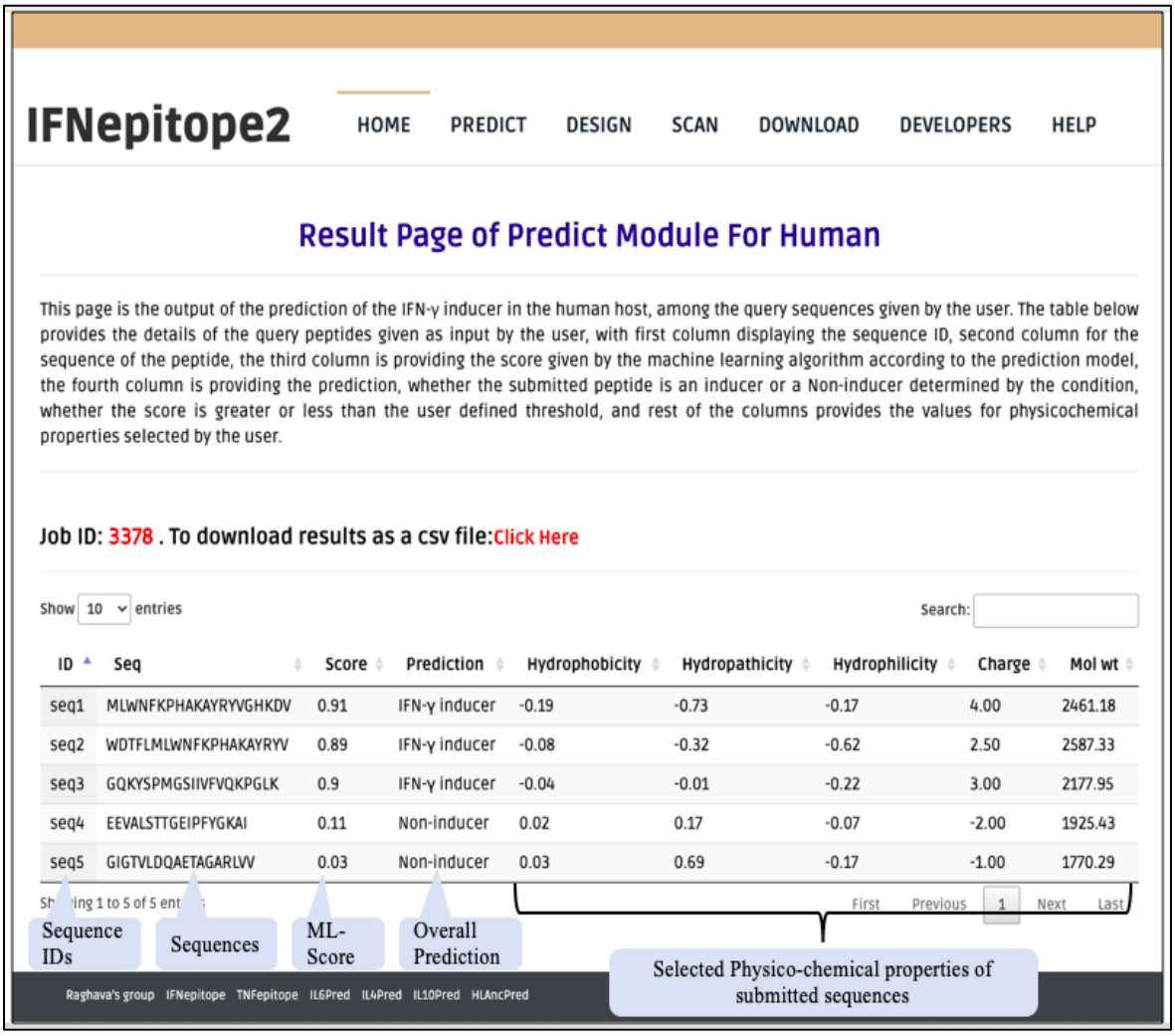



**Figure 8.6: Output page of prediction module; provide query sequence, prediction score and prediction as IFN-γ inducer and non-inducer**

## *8.5 Discussion*

IFN-gamma also known as type II interferon, is an essential cytokine for both innate and adaptive immunity against protozoan, bacterial, and viral infections. IFN-gamma is a crucial macrophage activator and inducer of the production of class II molecules from the major histocompatibility complex (Tau & Rothman, 1999). IFN-gamma is primarily produced by natural killer and natural killer T cells during the innate immune response, and CD4 and CD8 cells during the development of antigen-specific immunity during the adaptive immunological response (Castro et al., 2018; Schoenborn & Wilson, 2007). T helper cells, particularly Th1 cells, cytotoxic T cells, macrophages, mucosal epithelial cells, and NK cells all release IFN-gamma. IFN-gamma is a crucial paracrine signal in the early innate immune response and a crucial autocrine signal for professional APCs in the adaptive

immune response. The cytokines IL-12, IL-15, IL-18, and type I IFN all contribute to the induction of IFN-gamma expression. The single Type II interferon is IFN-gamma, which differs from Type I interferons serologically by being acid-labile as opposed to Type I variations' acid-stability (Burke & Young, 2019; Jorgovanovic et al., 2020). Numerous autoimmune and autoinflammatory disorders have abnormal IFN-gamma expression. In addition to its direct capacity to prevent viral replication, IFN is significant for the immune system due to its immunostimulatory and immunomodulatory properties.

The U.S. Food and Drug Administration has given interferon-1b approval to treat osteopetrosis and chronic granulomatous disease (CGD). IFN-gamma improves neutrophil activity against catalase-positive bacteria by regulating patients' oxidative metabolism, which is how it helps CGD (Ahlin et al., 1999). Children's hospital of Philadelphia has undertaken preliminary research on the use of IFN-gamma in the treatment of Friedreich's ataxia (FA), and found that patients' gait and stance had significantly improved (YetkIn & M, 2020). Interferon has also been demonstrated to be successful in treating individuals with moderate to severe atopic dermatitis, while not yet receiving formal approval. Recombinant IFN-therapy has particularly showed potential in children and patients with decreased IFN-expression, such as those at risk for herpes simplex virus (Brar & Leung, 2016). IFN-gamma upregulates MHC I and MHC II expression, which improves immunorecognition and the expulsion of harmful cells, while increasing an anti-proliferative state in cancer cells (Zhou, 2009). IFN-gamma also inhibits tumour spread by upregulating fibronectin, which has a detrimental effect on tumour architecture (Jorgovanovic et al., 2020). Hence, it is very important to identify the epitopes or peptides which can secrete the IFN-gamma.

In this study, we have developed a prediction method for the prediction of IFN-gamma inducing and non-inducing peptides for human and mouse hosts. We have computed composition based features for both IFN-gamma inducing and non-inducing peptides. We observed certain amino-acid residues (K, L, P and Q) and (A, P, G and V) are highly conserved in case of human and mouse IFN-gamma inducing peptides, respectively. Moreover, it was observer that dipeptide (QP, PQ, KL, KK, LK) and (AA, PA, GP, AV, AG) are the most abundant residue pair motifs in human and mouse IFN-inducing peptides in comparison with non-inducing peptides. We computed di-peptide composition based features, extra-tree based classifier we achieve maximum AUROC of 0.83 and 0.76 on human and mouse models respectively. We have incorporated the best models in the website IFNepitope 2.0 (https://webs.iiitd.edu.in/raghava/ifnepitope2/). We hope our study aid the scientific community in order design novel therapeutic candidate against deadly diseases and cancer.

## *8.6 Conclusion*

Subunit or peptide-based vaccines are more safely elicit immune response against infections caused by different pathogens. Peptide subunit vaccines can act as promising candidates for developing immunization against number of diseases including cancer. To serve the scientific community we have developed a computational method for the prediction of IFN-gamma inducing peptides or regions in human and mouse host. IFNepitope 2.0 is an updated version of IFNepitope witch is developed for the prediction of MHC-II binding peptides which can induce the interferon production. We have generated the latest method on the largest dataset obtained from immune epitope database. We integrated best models in the webserver and can be used for the prediction, scanning and designing of IFN-gamma inducing peptides in human and mouse models.

# CHAPTER 9

## INHIBITION OF IL6/STAT3 SIGNALLING PATHWAY

## *9.1 Introduction*

The Janus kinase (JAK)/Signal Transducer and Activator of Transcription (STAT) signalling system, also referred to as the JAK/STAT signalling route, is crucial in directing signals to numerous cytokines, hormones, and growth factors. Seven mammalian members of the STAT family, including STAT1, STAT2, STAT3, STAT4, STAT5a, STAT5b, and STAT6, are cytoplasmic transcription factors. They take part in cellular and biological processes such as differentiation, proliferation, apoptosis, and angiogenesis (Calo et al., 2003). The STAT3 gene encodes STAT3, a pleiotropic transcription factor belonging to the STATs family. Growth factors include fibroblast growth factor (FGF), epidermal growth factor (EGF), and insulin-like growth factor (IGF); they are activated in response to a variety of cytokines, including interleukin 6 (IL6) and interleukin 10 (IL-10) (Levy & Lee, 2002). The addition of the phosphate group to JAKs causes phosphorylation as a result of these factors' interaction to the cell surface receptor. STAT3 was phosphorylated at Serine 727 and Tyrosine 705. Additionally, STAT3 monomers combine to create a homodimer that interacts with one another via the SH2 domain. In order to control the transcription of genes, the homodimer STAT3 molecule later translocate into the nucleus and attaches to the specific target gene promoters with the aid of different coactivators, such as p68 (Ma et al., 2020). The STAT3 signalling pathway, however, is altered in a number of pathogenic processes that promote the development of cancer and other disorders. Specifically, upregulating STAT3 inhibits anticancer immune responses while promoting tumour cell growth, proliferation, invasion, migration, angiogenesis, and multidrug resistance (Corvinus et al., 2005; Kamran et al., 2013; Lee et al., 2019) (See Figure 9.1). By raising the mRNA levels of various genes involved in apoptosis, cell proliferation, and angiogenesis, such as Bcl-xL, Mcl-1, cyclin D1/D2, c-Myc, and VEGF (Banerjee & Resat, 2016; Furqan et al., 2013; Weerasinghe et al., 2007), aberration of STAT3 contributes to oncogenesis. For instance, STAT3 up-regulates the production of the anti-apoptotic protein Bcl-xL, whereas inhibiting STAT3 causes Bcl-xL expression to be down-regulated.

According to the research by Sateesh Kunigal et al., STAT3 expression was knocked down by small interfering RNA (siRNA), which decreased the expression of Bcl-xL and survivin in MDA-MB-231 breast cancer cells and increased the expression of Fas, Fas-L, and cleaved Caspase 3, which induced apoptosis and tumour suppression. Therefore, using siRNA to target STAT3 will aid in the treatment of breast cancer patients (Kunigal et al., 2009). Growing evidence suggests that STAT3 gene mutations are linked to a number of inflammatory diseases, including pulmonary fibrosis and acute lung injury (Forbes et al., 2016; Pedroza et al., 2016). By hindering the growth of regulatory T (Treg) cells and encouraging the multiplication and activation of Th17 [interleukin-17 (IL-17)-producing helper T

(TH) cells, also known as TH(IL-17), TH17, or inflammatory TH cells], STAT3 activation produces autoimmunity (Yang et al., 2007). When Th17 is activated and dysregulated, it plays a crucial role in the emergence of autoimmune diseases including Type 1 diabetes (T1D) (Shao et al., 2012). Additionally, STAT3 plays a significant role in coronavirus infection that contributed to the pathogenesis of COVID-19, such as promoting SARS-COV-2 replication, amplifying inflammatory responses, promoting lung fibrosis and injury, and lymphopenia (Gubernatorova et al., 2020; Jafarzadeh, Jafarzadeh, et al., 2021). Additionally, the STAT3-mediated signalling pathway stimulates the formation of M2-like macrophages, production of an inflammatory response, and immunopathological reactions (Chen, Tang, et al., 2020; Deenick et al., 2018; Jafarzadeh et al., 2020).

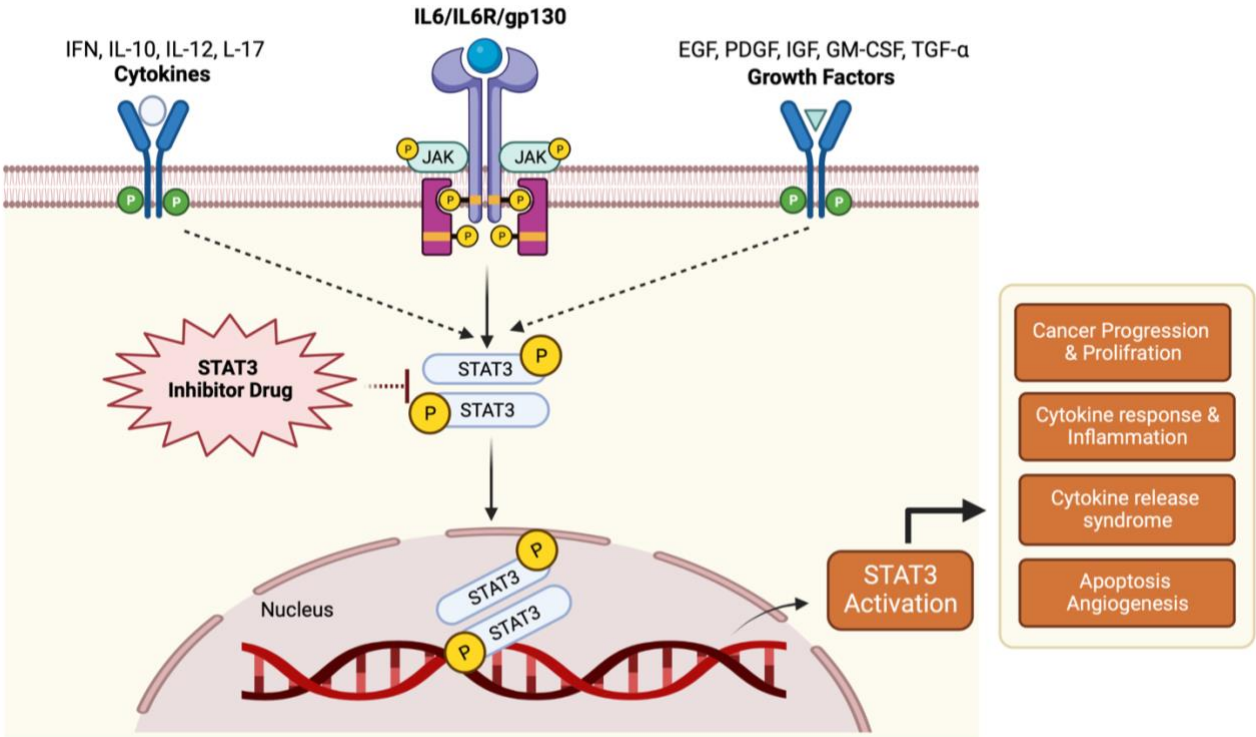



**Figure 9.1 Representation of IL6-mediated STAT3 signalling pathway, where IL6/IL6R/gp130 activate the phosphorylation of JAK and STAT3. In addition, several growth factors and cytokines activates the STAT3 phosphorylation and STAT3 hyperactivation leads to development of several diseases**

Furthermore, STAT-3 hyperactivation boosted cytokine storm production, which is important in the pathophysiology of COVID-19. As a result, targeting STAT-3 may have superior therapeutic potentials in COVID-19 (Jafarzadeh, Nemati, et al., 2021). STAT3 inhibitor development has arisen as an important subject of study because they have not yet been licenced for cancer treatment and a number of STAT3 inhibitors are in clinical testing. Researchers have sought to target STAT3 for the

development and application of new medications to date. STAT3 inhibitors work by suppressing STAT3 phosphorylation to impede the IL6/JAK/STAT3 signalling cascade. For example, JSI-124 (cucurbitacin I), a selective inhibitor, blocks STAT3 phosphorylation at serine 727, leading to death and cell-cycle arrest in B cell leukaemia. One of the pyrrolidinesulhonylaryl compounds (6a) selectively inhibits STAT3 phosphorylation and has promising anti-IL6/STAT3 signalling activity in IL6 driven MDA-MB-231 breast cancer and HeLa cell lines. Celecoxib* (FDA approved), BBI608* (FDA approved), Pyrimethamine* (FDA approved), and other STAT3 direct inhibitors are being tested in clinical studies for cancer immunotherapy (S. Zou et al., 2020). Despite the fact that the number of STAT3 inhibitor molecules is continually increasing, discovering novel STAT3 inhibitors remains a significant scientific issue.

There is currently no computational approach that can distinguish STAT3 inhibiting drugs from non-inhibitors. Based on these concepts, we aimed to create a prediction tool that can predict STAT3 inhibitors and non-inhibitors using various machine learning methods. Furthermore, by screening out inactive compounds in silico, fewer compounds will need to be produced or evaluated in vitro/in vivo. Machine learning has the potential to significantly accelerate the process and reduce the costs of developing novel treatments from previously tested and authorised chemical substances. The current study aimed to create machine learning-based models for predicting STAT3 inhibitor and non-inhibitor chemicals. To assist the scientific community, we present STAT3In (https://webs.iiitd.edu.in/raghava/stat3in/ ) a computational tool for the prediction and design of novel STAT3 inhibitor drugs.

## *9.2 Material and methods*

### *9.2.1 Curation of dataset*

In this investigation, the data for active and inactive STAT3 inhibitors were collected from the PubChem bioassay record (AID 862) [Primary cell-based high throughput screening assay to evaluate STAT3 inhibition]. A total of 194,698 chemicals were evaluated in this bioassay to see if they might inhibit or diminish IL6-mediated STAT3 transcription. This bioassay yielded a total of 194,698 chemical compounds with STAT3 inhibition and non-inhibition activity, including 1724 active and 192974 inactive chemical inhibitors. We choose 1724 molecules at random from a pool of 192974 inactive chemical inhibitors. 1724 chemical compounds with the IL6-mediated STAT3 inhibition property were regarded positive and named active inhibitors, while 1724 chemical compounds with the IL6-mediated STAT3 inhibition property were judged negative and called inactive inhibitors.

Then, using PubChem substance IDs and compound IDs, the 2D and 3D structural files for 1724 active (positive) and inactive (negative) chemical compounds were downloaded. However, only 1565 active and 1671 inactive compound structures were accessible out of 1724 compounds. As a result, the final dataset contains 1565 active chemical compounds and 1671 inactive chemical compounds. To assess the model's performance, we divided the entire dataset in an 80:20 ratio. 80% of the data was used as a training set, which included 1323 inactive and 1265 active chemical compounds, while the remaining 20% was used as a validation set, which included 300 active and 348 inactive chemical compounds.

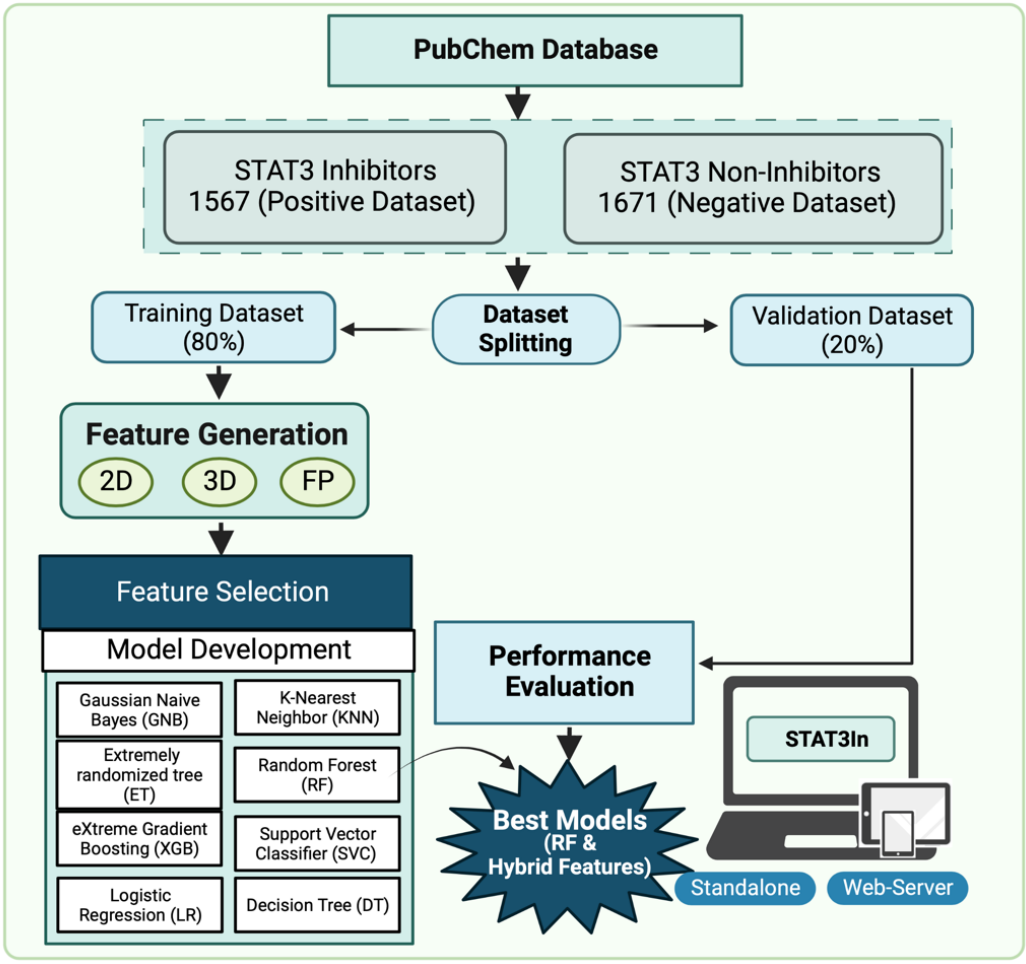



**Figure 9.2: Complete workflow of STAT3In, including data collection, model development and webserver implementation**

## *9.2.2 Chemical descriptors*

Chemical descriptors are the characteristics of chemical molecules that contribute to their activity. In this investigation, we calculated the descriptors of the molecules using the PaDEL software (Yap, 2011). For a single chemical substance, this software may compute a number of molecular descriptors. It generates a variety of 1D/2D/3D and binary fingerprints (FP) (e.g., Fingerprinter, Extended,

SubStructure, Substructure count, PubChem FP, MACCS keys, KlekotaRoth, KlekotaRoth count, Estate). We calculated 1444 2D descriptors, 136 3D descriptors, and 14532 binary fingerprint-based (FP) descriptors for 1564 active and 1671 inactive inhibitor drugs in this work. Various machine learning models were created using these 2D, 3D, and FP descriptors.

### *9.2.3 Pre-processing of data*

The generated descriptors were in a varied range, therefore to pre-process the dataset, we normalised each descriptor file using scikit learn's standard scaler module, sklearn.preprocessing. StandardScaler is a method for normalising data that uses the z-score algorithm. After normalising the data, we eliminated the null values from each descriptor file, if any existed. The 2D and FP descriptor files contain no null values, but the 3D descriptor file has a few null values. After we removed the null values, we had 1444 2D, 116 3D, and 14532 FP descriptors/features for the entire dataset. Previous research has revealed that most of the descriptors derived with PaDEL are meaningless (Dhanda, Singla, et al., 2013; Singh et al., 2015; Svetnik et al., 2003). As a result, selecting the most important descriptors is a critical step in developing any prediction model (Garg et al., 2010; Singla et al., 2011).

### *9.2.4 Feature selection techniques*

We employed three feature selection strategies in this study: first is the VarianceThreshold-based method, second is the correlation-based method, and third is the SVC-L1-based method. To remove low-variance features from all descriptors, we utilised scikit's VarianceThreshold package (sklearn.feature selection). After deleting low variance features, we were left with 622 2D, 66 3D, and 2251 FP descriptors instead of 1444 2D, 116 3D, and 14532 FP descriptors. Following that, a correlation-based feature selection method was utilised to choose those features that correlate with each other by less than 0.6 with each other. As a result, we excluded the features with a correlation more than or equal to 0.6 (>=0.6). After that, we were left with 73 2D, 9 3D, and 1622 FP descriptors out of a total of 622 2D, 66 3D, and 2251 FP descriptors. Finally, the SVC-L1 feature selection technique was utilised to obtain the most significant feature set. This is a typical strategy for reducing the size of the feature vector. Using the SVC-L1 technique, we were left with the most important feature set of 162 features, which includes 41 2D, 5 3D, and 116 FP descriptors. Using the feature-selector algorithm, these 162 traits were prioritised according to their importance in distinguishing active and inactive inhibitors. Gradient Boosting Decision Tree (GBDT) is used in this software. LightGBM, a prominent machine learning technique, was used to rank the characteristics. It calculates how many times a feature is used to split the data across all trees to estimate its rank. The features

picked and rated by this method were utilised to create several machine learning models, and the models' performance was computed on the top 10, 20, 30,...., 116 features, respectively.

### *9.2.5 Machine learning-based classifiers*

We used different machine learning techniques to construct prediction models for the classification of STAT3 inhibitors and non-inhibitors chemical substances in this study. In order to create models, we used random forest (RF), Support Vector Classifier (SVC), decision tree (DT), K-nearest neighbour (KNN), Logistic Regression (LR), Gaussian Naive Bayes (GNB), and XGBoost (XGB). Scikit's sklearn package was used to implement all of these machine learning algorithms (Pedregosa et al., 2011).

### *9.2.6 Performance evaluation*

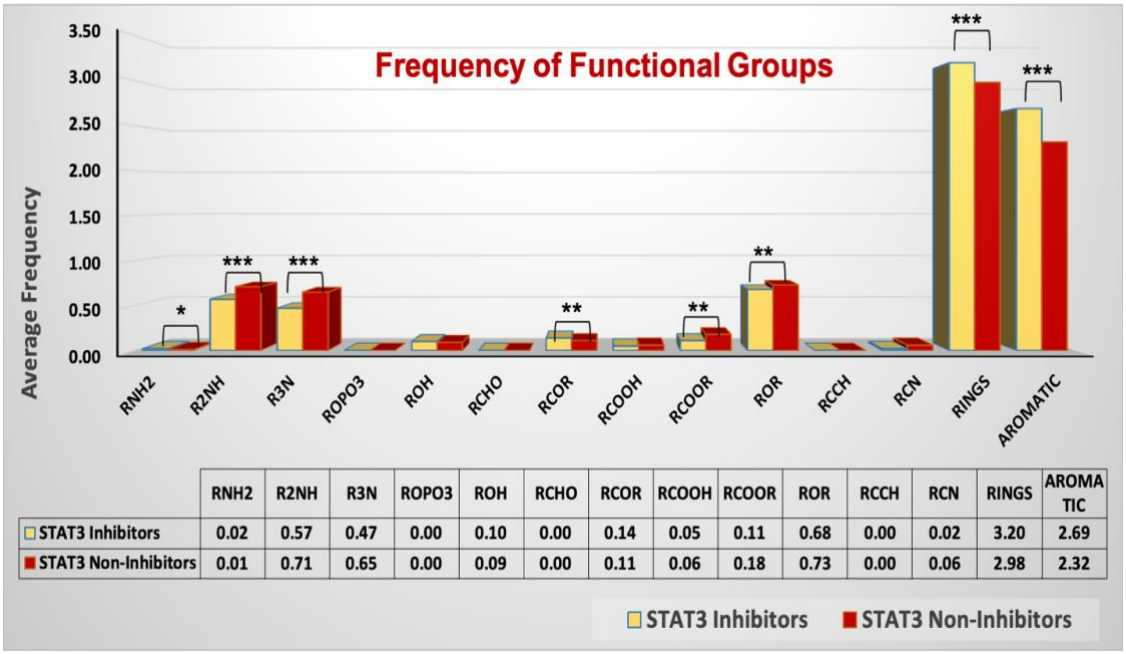
The model's performance was assessed using the leave one out cross-validation (LOOCV) technique. To analyse our prediction model in this work, we employed the usual 5-fold cross-validation technique. The entire dataset was divided in an 80:20 ratio, resulting in an 80% training dataset and a 20% external validation dataset. The training dataset was subjected to five-fold cross-validation. The 80% training dataset was divided into five equal-sized sets, each with an equal number of positive and negative chemicals. Four of these five sets will be utilised for training, while the last fifth set will be used for testing. The same procedure is repeated five times to ensure that each of the five sets is used at least once for model testing. The prediction models were built using these five training and testing sets. The model's overall performance was then assessed using the 20% external validation dataset.

We used conventional evaluation metrics to assess the performance of various prediction models. We employed both threshold-dependent and independent factors in this analysis. The model's performance was assessed using sensitivity (Sens), specificity (Spec), accuracy (Acc), and the Matthews correlation coefficient (MCC), all of which are threshold-dependent characteristics. The threshold-independent parameter, i.e., the area under the receiver operating characteristic curve (AUROC), was used to evaluate the model's performance.

## *9.3 Results*

### *9.3.1 Analysis of functional groups*

We used the chemmineR package to compute the frequency of distinct functional groups of IL6-mediated STAT3 inhibitors (positive dataset) and non-inhibitors (negative dataset). We can see from the average frequency values that the abundance of rings and aromatic groups is much larger in the positive sample than in the non-inhibitors. Inactive substances, i.e., STAT3 non-inhibitors, have a higher frequency of secondary amines (R2NH), tertiary amines (R3N), and ester (ROR) groups, as seen in Figure 9.3.




| | RNH2 | R2NH | R3N | ROPO3 | ROH | RCHO | RCOR | RCOOH | RCOOR | ROR | RCCH | RCN | RINGS | AROMATIC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| STAT3 Inhibitors | 0.02 | 0.57 | 0.47 | 0.00 | 0.10 | 0.00 | 0.14 | 0.05 | 0.11 | 0.68 | 0.00 | 0.02 | 3.20 | 2.69 |
| STAT3 Non-Inhibitors | 0.01 | 0.71 | 0.65 | 0.00 | 0.09 | 0.00 | 0.11 | 0.06 | 0.18 | 0.73 | 0.00 | 0.06 | 2.98 | 2.32 |

**Figure 9.3 Average frequency distribution of different functional groups of STAT3 inhibitors and non-inhibitors chemical compounds**

We also discovered the presence of rings and aromatic groups in the STAT3 inhibitors Napabucasin (BBI608), an FDA-approved medicine used to treat advanced malignancies (Ref), and STAT3 Inhibitor VII (STAT3-IN-8) drug, which is utilised for STAT3 inhibition and the treatment of head and neck cancer. Some FDA-approved indirect STAT3 inhibitors, such as AZD-1480 and Ruxolitinib, have comparable tendencies. These findings imply that the researcher can use this study to develop innovative medications that can be employed as active inhibitors of STAT3.

## *9.3.2 Classification model performance*

One important problem in this type of investigation is classifying STAT3 inhibitors and non-inhibitors using 2D, 3D, and FP descriptors. We employed different feature selection strategies to obtain the

optimal collection of features that may be used for categorization. Following that, we created many prediction models using classifiers such as RF, DT, LR, XGB, SVM, and GBM.

### *9.3.2.1 2D-based models*

For the positive and negative datasets, we compute 1444 2D descriptors. We get 74 features after deleting low variance and highly correlated characteristics. We created classification models using this feature set. On the training and validation (AUC = 0.84) datasets, RF achieves maximum performance with balanced sensitivity and specificity. Using the SVC-L1 method, we were able to obtain 41 2D-descriptors. The AUC 0.83 and 0.84; accuracy 76.35% and 75.46% on training and validation datasets with the RF classifier vary somewhat after feature reduction. SVM also works well on training and testing datasets, with accuracy values of 74.27 and 72.99, respectively, as shown in Table 9.1.

**Table 9.1: Performance measures of 2D-based descriptors developed on training dataset and testing dataset**

| Method | Dataset | Sensitivity | Specificity | Accuracy | AUC |
|---|---|---|---|---|---|
| **DT** | **Training** | 64.2 | 64.3 | 64.2 | 0.69 |
| | **Testing** | 72.2 | 59.7 | 66.2 | 0.73 |
| **RF** | **Training** | 76.1 | 76.6 | 76.4 | 0.83 |
| | **Testing** | 74.6 | 76.4 | 75.5 | 0.84 |
| **LR** | **Training** | 69.7 | 69 | 69.3 | 0.75 |
| | **Testing** | 71.6 | 69 | 70.4 | 0.77 |
| **XGB** | **Training** | 71.6 | 71.8 | 71.7 | 0.78 |
| | **Testing** | 72.5 | 70.9 | 71.8 | 0.8 |
| **KNN** | **Training** | 70.3 | 70.4 | 70.4 | 0.77 |
| | **Testing** | 70.8 | 70.9 | 70.8 | 0.79 |
| **GNB** | **Training** | 65.2 | 66.1 | 65.7 | 0.7 |
| | **Testing** | 69.6 | 68.1 | 68.8 | 0.73 |
| **SVM** | **Training** | 74.8 | 73.8 | 74.3 | 81 |
| | **Testing** | 71.3 | 74.8 | 73 | 81 |

**#DT, Decision tree; GNB, Gaussian Naive Bayes; KNN, k-nearest neighbor; LR, Logistic Regression; RF, Random Forest; SVM, Support Vector Machine; XGB, *XGBoost*; AUROC, Area Under Receiver Operating Curve**

### *9.3.2.2 3D-based models*

With SVC-L1, we selected top-5 features of 3D descriptors and computed the performance. In this situation, RF surpasses all other classifiers on training and testing data, with the greatest AUC (0.741

and 0.729). XGB and SVM, on the other hand, perform pretty well, with AUC 0.73 on training data and AUC 0.71 on validation data, as shown in Table 9.2.

**Table 9.2: Performance measures of 3D-based descriptors developed on training dataset and testing dataset**

| Method | Dataset | Sensitivity | Specificity | Accuracy | AUC |
|---|---|---|---|---|---|
| **DT** | **Training** | 64.80 | 62.00 | 63.33 | 0.68 |
| | **Testing** | 67.16 | 51.76 | 59.72 | 0.66 |
| **RF** | **Training** | 67.15 | 66.35 | 66.73 | 0.74 |
| | **Testing** | 66.27 | 65.18 | 65.74 | 0.73 |
| **LR** | **Training** | 65.77 | 65.54 | 65.65 | 0.71 |
| | **Testing** | 65.67 | 64.54 | 65.12 | 0.70 |
| **XGB** | **Training** | 65.29 | 66.94 | 66.15 | 0.73 |
| | **Testing** | 65.67 | 66.13 | 65.90 | 0.72 |
| **KNN** | **Training** | 68.21 | 67.01 | 67.58 | 0.74 |
| | **Testing** | 69.85 | 62.62 | 66.36 | 0.73 |
| **GNB** | **Training** | 65.85 | 65.69 | 65.77 | 0.71 |
| | **Testing** | 67.46 | 61.98 | 64.82 | 0.70 |
| **SVM** | **Training** | 66.91 | 66.50 | 66.69 | 0.73 |
| | **Testing** | 66.87 | 65.18 | 66.05 | 0.71 |

**#DT, Decision tree; GNB, Gaussian Naive Bayes; KNN, k-nearest neighbor; LR, Logistic Regression; RF, Random Forest; SVM, Support Vector Machine; XGB, *XGBoost*; AUROC, Area Under Receiver Operating Curve**

### *9.3.2.3 FP-based models*

Models based on FP outperform models based on 2D and 3D characteristics. On both the training and validation datasets, the RF algorithm achieves maximum performance, i.e., AUC (0.86) with balanced sensitivity and specificity. SVM achieves comparable performance in this scenario, i.e., AUC (training data = 0.84 and testing data = 0.85), and results of XGB, GBM, LR, DT, and KNN are reported in Table 9.3.

**Table 9.3: Performance measures of FP-based descriptors developed on training dataset and testing dataset**

| Method | Dataset | Sensitivity | Specificity | Accuracy | AUC |
|---|---|---|---|---|---|
| **DT** | **Training** | 64.96 | 65.24 | 65.11 | 0.71 |
| | **Testing** | 67.46 | 61.66 | 64.66 | 0.70 |
| **RF** | **Training** | 78.46 | 77.61 | 78.01 | 0.86 |

| | | | | | |
|---|---|---|---|---|---|
| | **Testing** | 79.40 | 77.96 | 78.7 | 0.86 |
| **LR** | **Training** | 75.85 | 76.66 | 76.28 | 0.83 |
| | **Testing** | 72.84 | 76.68 | 74.69 | 0.81 |
| **XGB** | **Training** | 77.32 | 77.54 | 77.43 | 0.84 |
| | **Testing** | 77.91 | 80.83 | 79.32 | 0.86 |
| **KNN** | **Training** | 76.18 | 75.04 | 75.58 | 0.83 |
| | **Testing** | 77.02 | 73.80 | 75.46 | 0.83 |
| **GNB** | **Training** | 73.98 | 74.08 | 74.03 | 0.81 |
| | **Testing** | 69.55 | 73.8 | 71.61 | 0.79 |
| **SVM** | **Training** | 78.62 | 78.35 | 78.48 | 0.86 |
| | **Testing** | 77.31 | 80.19 | 78.70 | 0.86 |

**#DT, Decision tree; GNB, Gaussian Naive Bayes; KNN, k-nearest neighbor; LR, Logistic Regression; RF, Random Forest; SVM, Support Vector Machine; XGB, *XGBoost*; AUROC, Area Under Receiver Operating Curve**

### *9.3.2.4 Hybrid models*

Then, to increase performance, we combined 2D (41 features), 3D (5 features), and FP (116 features) descriptors and built models with 162 descriptors. The accuracy (79.48 and 81.02) and AUC (0.87 and 0.88) of RF models employing integrated features are quite high on training and validation datasets. We discovered that integrating 2D+3D+FP characteristics had no discernible effect on the performance of ML-based models. As a result, we use the feature selector algorithm to perform feature ranking on the combined 162 features. Finally, we achieved a minimal set of features that perform almost as well as the hybrid model (2D+3D+FP) features. By ranking the features and then examine the performance of the top-10, 20, 30,.....162 features. Finally, we choose the best 50 descriptors (14 2D, 1 3D, and 35 FP) from a set of 162 features. The top-50 features perform remarkably identically to the 162 features. On both the training and validation datasets, RF achieved a maximum AUC of 0.87 and accuracy greater than 78.5 with the smallest sensitivity and specificity difference (See Table 9.4).

**Table 9.4: The performance of machine learning models using hybrid (2D+3D+FP) descriptors on training dataset and testing dataset**

| Method | Dataset | Sensitivity | Specificity | Accuracy | AUC |
|---|---|---|---|---|---|
| **DT** | **Training** | 68.22 | 68.03 | 68.12 | 0.74 |
| | **Testing** | 66.67 | 72.70 | 69.91 | 0.74 |
| **RF** | **Training** | 78.42 | 78.61 | 78.52 | 0.87 |
| | **Testing** | 79.00 | 78.16 | 78.55 | 0.87 |
| **LR** | **Training** | 77.00 | 76.34 | 76.66 | 0.84 |
| | **Testing** | 75.67 | 77.87 | 76.85 | 0.83 |

| | | | | | |
|---|---|---|---|---|---|
| **XGB** | **Training** | 77.31 | 77.10 | 77.20 | 0.85 |
| | **Testing** | 80.00 | 75.29 | 77.47 | 0.85 |
| **KNN** | **Training** | 74.94 | 75.89 | 75.43 | 0.83 |
| | **Testing** | 78.00 | 75.58 | 76.70 | 0.83 |
| **GNB** | **Training** | 74.23 | 74.00 | 74.11 | 0.81 |
| | **Testing** | 75.33 | 72.99 | 74.07 | 0.80 |
| **SVM** | **Training** | 77.71 | 77.55 | 77.63 | 0.86 |
| | **Testing** | 78.33 | 76.72 | 77.47 | 0.85 |

**#DT, Decision tree; GNB, Gaussian Naive Bayes; KNN, k-nearest neighbor; LR, Logistic Regression; RF, Random Forest; SVM, Support Vector Machine; XGB, *XGBoost*; AUROC, Area Under Receiver Operating Curve**

## *9.4 Web-based platform*

In order to help the scientific community, we created "STAT3In" (https://webs.iiitd.edu.in/raghava/stat3in/) a webserver that can classify STAT3 inhibitors. We built the web server's front and back ends with HTML5, JAVA, CSS3, and PHP scripts. The STAT3In web server is compatible with a variety of platforms, including mobile, iPad, tablet, and desktop computers, as well as multiple browsers. In the server's backend, we applied the random forest model developed with hybrid chemical descriptors as input features. The web server is divided into three key modules: "Predict", "Draw" and "Analog design". The "Predict" module aids the user in determining if a chemical substance is a STAT3 inhibitor or not. The module accepts chemical compounds from users in a variety of forms, including SDF, SMILES, and MOL, and also lets users choose the desired threshold. The users can upload a file with numerous chemical compounds or insert a single molecule or multiple molecules. The result page includes the machine learning score and the class(es) of the provided compound(s) as either a STAT3 inhibitor or non-inhibitor. To search or sort the output table, the result is supplied in comma-separated value (CSV) format. The "Draw" module allow the user to create or modify the chemical molecule structure and this module then import the structure into the prediction model to determine whether the molecule is a STAT3 inhibitor or not. In the third, module users can create the analogues in the "Analogue design" module by combining submitted scaffolds, building blocks, and linkers. The homepage of website and utility of prediction module in Figure 9.4.

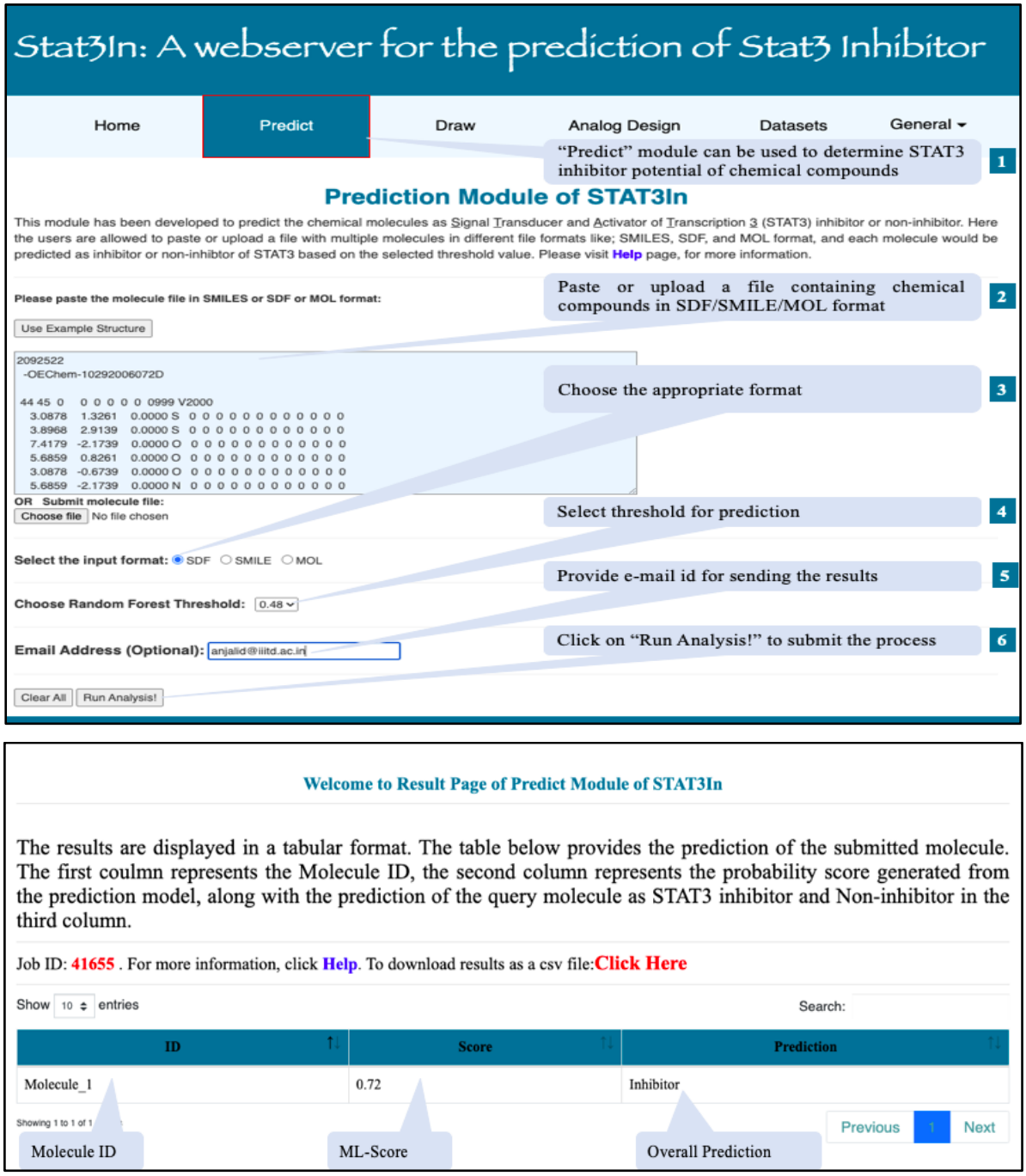



**Figure 9.4 Input and output page of 'Prediction' module of STAT3In webserver, provides molecule ID, machine learning score and prediction**

## *9.5 Case Study: Repurposing of FDA-approved drugs*

In order to find out the applicability of STAT3In server, we have performed a case study. In this we have find the possible therapeutic candidates for inhibiting the STAT3 pathway, we have used 1102 FDA-approved pharmacological compounds from the Drug Bank database. At first, we determined the PubChem CID from the FDA-approved drugs. A total of 842 drugs, out of the 1102 drugs, compose

the 2-D structures. We have used the Predict module of our STAT3In server with default settings, i.e., Random Forest Threshold = 0.48. Out of 842 FDA-approved drugs we find out 8 possible pharmacological candidates for STAT3 inhibition using our prediction model. The drugs predicted by our server previously used by number of studies for the treatment of cancer, inhibition of tumor progression, angiogenesis, and COVID-19 progression. The complete description and functions of predicted FDA-approved drugs is given in Table 9.5.

**Table 9.5: Predicted FDA-approved drug candidates for STAT3 inhibition (Adopted from-Dhall et. al., 2021)**

| Drug Bank ID | FDA-Approved (Drugs) | STAT3In (Prediction) | Functions |
|---|---|---|---|
| **DB00682** | Warfarin | Inhibitor | Inhibition of IL6/STAT3-dependent fibrin production in severe listeriosis. |
| **DB09357** | Dexpanthenol | Inhibitor | Inhibition of LPS-induced neutrophils influx, protein leakage, and release of TNF-α and IL6 in bronchoalveolar lavage fluid in acute lung injury. |
| **DB00790** | Perindopril | Inhibitor | It regulates the inflammatory mediators, NF-κB/TNF-α/IL6, and apoptosis in renal diseases and inhibit the activation of STAT3. ACE inhibitor perindopril-inhibited tumor growth was associated with the suppression of angiogenesis. |
| **DB00675** | Tamoxifen | Inhibitor | Treatment of ER-positive breast cancer with tamoxifen by inhibiting the IL6/STAT3 signal pathway, inhibition of tumor growth and angiogenesis. Anticancer drugs that have shown potential activity in both MERS and SARS-CoV. |
| **DB00183** | Pentagastrin | Inhibitor | Anti-malarial, anti-fungal, anti-bacterial, and anti-inflammatory. |
| **DB00476** | Duloxetine | Inhibitor | Inhibit overexpression of IL6 mRNA in anxiety- and major depressive disorder, anti-inflammatory action against IL6. |
| **DB09027** | Ledipasvir | Inhibitor | Anti-viral activity against COVID-19, (sofosbuvir, and ledipasvir) inhibited STAT3 protein levels to cure HCV infections. |
| **DB00768** | Olopatadine | Inhibitor | Inhibit CHMCs activation and release of IL6, tryptase, and histamine and use as anti-allergy drug. |

## *9.6 Discussion*

One of the most important transcription factors and an oncogene, STAT3 play major role in the development and spread of tumours. It may therefore provide a great therapeutic target for a variety of cancer treatments because of its flexible regulatory pathways and significant biological functions in cancer. Additionally, it has been documented in the literature that coronavirus-infected patients, whose numbers are rising rapidly all over the world, have highly higher levels of IL6. The JAK/STAT3 pathway is how the cytokine IL6 mediates its effects, hence it is imperative to create computational algorithms that can anticipate how effective a chemical molecule will be as a STAT3 inhibitor. Numerous techniques have been developed in the past that take use of the link between the structure and activity of chemical compounds to use machine learning techniques to predict whether a chemical molecule has the potential to be an inhibitor. For example, EGFRPred predicts whether a molecule has

the potential to be an EGFR inhibitor, and DrugMint determines whether a molecule has the potential to be a potential drug candidate.

In this study, we tried to create a computational approach that could distinguish between STAT3 inhibitors and non-inhibitors. In STAT3 inhibitor compounds, we noted a high frequency of rings and a low frequency of R2NH, R3N, and ROR groups. The high prevalence of these functional groups in STAT3 medications such AZD-1480, Ruxolitinib, Napabucasin, and STAT3-In-8 is further supported by literature (Furqan et al., 2013). For the purpose of creating the prediction models, we take into account STAT3 inhibitors and non-inhibitors as the positive and negative datasets. On the validation dataset using hybrid descriptors, random forest-based models perform best (AUC=0.87 and accuracy=78.55). To further identify possible therapeutic candidates against STAT3 activation, we took 842 FDA-approved medications. We have predicted eight drugs "Warfarin, dexpanthenol, perindopril, tamoxifen, pentagastrin, duloxetine, ledipasvir, and olopatadine" as potential medications that we have found to be effective in treating severe diseases like tumour progression, angiogenesis, COVID-19 progression, and the ability to inhibit the IL6/STAT3 pathway. IL6/STAT3 activation and may be employed as a therapeutic candidate to combat the COVID-19-related cytokine storm. A website called STAT3In is created to anticipate and design probable STAT3 inhibitors using machine learning techniques and basic information derived from chemical compounds. The user-friendly web-server is freely available at https://webs.iiitd.edu.in/raghava/stat3in/ . This method will aid researchers working in the field of cancer therapy and infectious diseases.

## *9.7 Conclusion*

In the current study, we developed a prediction method to distinguish between chemical compounds that are STAT3 inhibitors and non-inhibitors. We have provided a webserver for the prediction of STAT3 inhibiting chemical compounds, which can utilized by experimental biologist for the identification of STAT3 inhibiting molecules. However, our work is limited by the fact that the models were created using the chemical that were only tested on the "human U3A fibrosarcoma" cell line. In order to build a rigorous methodology, the investigation should be carried out on animal models or on a variety of cell lines.

# CHAPTER 10

## SUMMARY

Cancer is one of the leading cause of death globally, according to GLOBOCAN approx. 10.3 million deaths and 19.3 million new cases of cancer occurred in the United States. Over the past few decades, researchers have work tirelessly for finding new therapies and solutions for the devastating disease. The most widely utilised treatments include traditional therapies like chemotherapy, radiation, and surgery. The patient's health and survival are adversely affected by these radiation-based treatments. New treatment modalities, such as targeted cancer therapies, adoptive T cell therapy, immune checkpoint inhibitor-based therapies, immunomodulators, and oncolytic viruses based therapy, have been created to overcome the limitations of conventional treatments. Immunotherapy is a type of cancer treatment, which uses the body's own immune cells to boost the immune system and assist the body in locating and eliminating cancer cells. Numerous forms of cancer can be treated using immunotherapy. It may be used alone or in conjunction with other cancer treatments such as chemotherapy. Improvements in immunotherapy have showed notable results and improves the lives of many patients with a variety of solid tumours. Our immune system recognize the mutated peptides (tumour specific peptides or neoantigens), which are produced by a variety of genetic changes in cancer cells. The immune system can distinguish between malignant and normal cells with the help of tumour specific antigens. Since tumor-specific antigens are displayed on cell surfaces via Human leukocyte antigen (HLA) molecules and are identified by T cells. Adaptive immunity is mediated by CD8+ T cells, a crucial subset of HLA class I-restricted T cells. They consist of CD8+ suppressor T cells, which control certain immune responses, and cytotoxic T cells, which are crucial for eliminating malignant or virally infected cells. Cytotoxic T cells initiate the production of cytokines majorly TNF-α and IFN-γ, which causes anti-tumor and anti-viral responses.

Antigen-presenting cells have class II HLA molecules which present mutated or tumorigenic peptides which are recognized by CD4+ T lymphocytes. They all significantly contribute to initiating and directing adaptive immune responses. CD4+ T lymphocytes activate T-helper cells and secrete number of cytokines (IL-12, IFN-γ, IL-4, IL-5, IL6, TNF-α and IL-13) in order kill or eradicate the pathogen or cancer cells. Moreover, the overproduction of cytokines (IL6) leads to the activation of STAT3 signaling pathway which further proliferates the production of oncogenes, tumor metastasis, angiogenesis and development of tumor. So, it is crucial to inhibit the IL6/STAT3 signalling pathway in order to suppress the tumor growth. Human leukocyte antigens (HLA), HLA-binding peptides (neobinders) and cytokines are the most crucial components of our immune system. These molecules play a vital role directly or indirectly in developing cancer vaccine or immunotherapy. In this study, we investigated the role of cytokines and HLA molecules, in order to design better therapeutics against cancer. We majorly divided our study in four sections: (i) Prognostic biomarkers for cancer, (ii) Non-

classical HLA-binder prediction, (iii) Designing of cytokine inducing peptides, (iv) Inhibition of IL6/STAT3 pathway.

In the first part of the study we tried to investigate the role of HLA-alleles, neobinders and cytokines on the survival of cancer patients. This section is further subdivided into two categories: (i) Pan-cancer risk estimation analysis (ii) Personalized HLA-based prognostic biomarkers for skin cancer. These sub-sections are explained in details in Chapter 3 and Chapter 4. In the first section we investigated the importance of class-I HLA, neobinders and cytokines expressions with the survival of cancer patients. Here, we used HLA-typing information, tumor specific neoantigens and expression profiles of twenty types of cancer patients in order to perform univariate survival analysis and correlation analysis. We have incorporated all the analysis in a user friendly web-resource "CancerHLA-I" (https://webs.iiitd.edu.in/raghava/cancerhla1/ ). We anticipate this web-based platform could be utilized for the analysis and identification of cancer-specific biomarkers. This study may provide promising HLA-biomarkers for designing cancer immunotherapy. In the second part of the study we have developed a risk estimation tool "SKCMhrp" for skin cutaneous melanoma patients. Here, we performed patient-specific HLA-typing for class-I and class-II alleles and use the clinical information to derive the prognostic biomarker. We have used machine learning algorithms to develop survival prediction models and web-tool SKCMhrp which is freely accessible at (https://webs.iiitd.edu.in/raghava/skcmhrp/ ).

In the second part of the study, we have developed a computational tool for the prediction of HLA-binding peptides. We have explained the details of this section in Chapter 5. In the past number of HLA-binder prediction methods have been developed, however there is not a single platform for non-classical HLA i.e., HLA-G and HLA-E. Hence, we have developed an in-silico tool for the identification of binding peptides corresponding to HLA-G*01:01, HLA-G*01:02, HLA-G*03:01, HLA-E*01:01, and HLA-E*01:03. We have also developed a highly accurate and easy to use web platform "HLA$_{nc}$Pred" which is available at (https://webs.iiitd.edu.in/raghava/hlancpred/ ). Moreover, we developed a standalone version of HLAncPred (https://webs.iiitd.edu.in/raghava/hlancpred/stand.html ).

In the third part of the study, we have developed three prediction tools for the major cytokines (IL6, TNF-α and IFN-γ). We have divided this section into three sub-sections: (i) Prediction of IL6 inducing peptides (ii) TNF-α inducing epitopes prediction and (iii) Identification of IFN-γ inducing peptides. The complete description of all these studies is given in Chapter 6, Chapter 7 and Chapter 8. In the first part, we have developed a tool for the prediction, scanning and designing of IL6 inducing peptides.

We have used the experimentally validated datasets from IEDB resource and developed classification models using several machine learning techniques. Finally, the best models incorporated in the website IL6Pred (https://webs.iiitd.edu.in/raghava/il6pred/) and standalone package (https://webs.iiitd.edu.in/raghava/il6pred/stand.html). In the next part, we have generated a host-specific prediction method for the identification of TNF-α inducing epitopes or peptides. The models were trained and tested on experimentally validated TNF-α inducing and non-inducing peptides. Finally, the best prediction models integrated in the user-friendly web tool named "TNFepitope" (https://webs.iiitd.edu.in/raghava/tnfepitope/). In the third sub-section, we have developed an updated method for the prediction of interferon-gamma inducing and non-inducing peptides. This method can be utilized in the identification IFN inducing regions in the subunit or peptide based vaccines. The method is easy to use and available at the (https://webs.iiitd.edu.in/raghava/ifnepitope2/).

In the fourth part, we have conducted a study for the identification of inhibitors against IL6 mediated STAT3 signalling pathway. The complete details of the study is provided in Chapter 9. We tried to develop an computational tool for the prediction of molecules which can inhibit the activation of STAT3. As shown in literature, the production of IL6 activate the JAK/STAT3 signalling pathway. The overactivation of STAT3 leads to the proliferation of tumor cells. To assist the scientific community, we purpose a computational tool for the prediction and design of novel STAT3 inhibitor drugs. We have used the dataset from PubChem repository and generate chemical descriptors using PaDEL software. These numerical features are provided to machine learning algorithms for training and validated on the external datasets. Finally the best prediction models integrated in the web-based platform named "STAT3In" (https://webs.iiitd.edu.in/raghava/stat3in/). Overall, the study done in this thesis addresses various aspects of the immunology and use of genomic profiles to identify the prognostic biomarkers for cancer patients. Moreover, we anticipate that experimental biologist and clinicians will use these findings of our investigations to develop novel subunit vaccines and immunotherapies to treat cancer patients.

# BIBLIOGRAPHY

Abd-Aziz, N., & Poh, C. L. (2022). Development of Peptide-Based Vaccines for Cancer. *J Oncol*, *2022*, 9749363. https://doi.org/10.1155/2022/9749363

Adams, A. B., Larsen, C. P., Pearson, T. C., & Newell, K. A. (2002). The role of TNF receptor and TNF superfamily molecules in organ transplantation. *Am J Transplant*, *2*(1), 12-18. https://doi.org/10.1034/j.1600-6143.2002.020104.x

Adegbola, S. O., Sahnan, K., Warusavitarne, J., Hart, A., & Tozer, P. (2018). Anti-TNF Therapy in Crohn's Disease. *Int J Mol Sci*, *19*(8). https://doi.org/10.3390/ijms19082244

Aggarwal, B. B. (2003). Signalling pathways of the TNF superfamily: a double-edged sword. *Nat Rev Immunol*, *3*(9), 745-756. https://doi.org/10.1038/nri1184

Aguirre-Gamboa, R., Gomez-Rueda, H., Martinez-Ledesma, E., Martinez-Torteya, A., Chacolla-Huaringa, R., Rodriguez-Barrientos, A., Tamez-Pena, J. G., & Trevino, V. (2013). SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PLoS One*, *8*(9), e74250. https://doi.org/10.1371/journal.pone.0074250

Ahlin, A., Larfars, G., Elinder, G., Palmblad, J., & Gyllenhammar, H. (1999). Gamma interferon treatment of patients with chronic granulomatous disease is associated with augmented production of nitric oxide by polymorphonuclear neutrophils. *Clin Diagn Lab Immunol*, *6*(3), 420-424. https://doi.org/10.1128/CDLI.6.3.420-424.1999

Alfirevic A, P. M. (2010 Dec 23). Drug Induced Hypersensitivity and the HLA Complex. Pharmaceuticals (Basel). . *;4(1):69–90.* . https://doi.org/doi: 10.3390/ph4010069.

Almeida, L. G., Sakabe, N. J., deOliveira, A. R., Silva, M. C., Mundstein, A. S., Cohen, T., Chen, Y. T., Chua, R., Gurung, S., Gnjatic, S., Jungbluth, A. A., Caballero, O. L., Bairoch, A., Kiesler, E., White, S. L., Simpson, A. J., Old, L. J., Camargo, A. A., & Vasconcelos, A. T. (2009). CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. *Nucleic Acids Res*, *37*(Database issue), D816-819. https://doi.org/10.1093/nar/gkn673

Altun, I., & Sonkaya, A. (2018). The Most Common Side Effects Experienced by Patients Were Receiving First Cycle of Chemotherapy. *Iran J Public Health*, *47*(8), 1218-1219. https://www.ncbi.nlm.nih.gov/pubmed/30186799

Amin, S., Baine, M. J., Meza, J. L., & Lin, C. (2020). Association of Immunotherapy With Survival Among Patients With Brain Metastases Whose Cancer Was Managed With Definitive Surgery of the Primary Tumor. *JAMA Netw Open*, *3*(9), e2015444. https://doi.org/10.1001/jamanetworkopen.2020.15444

Amiot, L., Ferrone, S., Grosse-Wilde, H., & Seliger, B. (2011). Biology of HLA-G in cancer: a candidate molecule for therapeutic intervention? *Cell Mol Life Sci*, *68*(3), 417-431. https://doi.org/10.1007/s00018-010-0583-4

Amiot, L., Vu, N., & Samson, M. (2014). Immunomodulatory properties of HLA-G in infectious diseases. *J Immunol Res*, *2014*, 298569. https://doi.org/10.1155/2014/298569

Anassi, E., & Ndefo, U. A. (2011). Sipuleucel-T (provenge) injection: the first immunotherapy agent (vaccine) for hormone-refractory prostate cancer. *P T*, *36*(4), 197-202. https://www.ncbi.nlm.nih.gov/pubmed/21572775

Angeletti, P. C., Zhang, L., & Wood, C. (2008). The viral etiology of AIDS-associated malignancies. *Adv Pharmacol*, *56*, 509-557. https://doi.org/10.1016/S1054-3589(07)56016-3

Anjali Lathwal, R. K., Dilraj kaur, Gajendra P.S. Raghava. (2021). In silico model for predicting IL-2 inducing peptides in human.

Ansari, H. R., Flower, D. R., & Raghava, G. P. (2010). AntigenDB: an immunoinformatics database of pathogen antigens. *Nucleic Acids Res*, *38*(Database issue), D847-853. https://doi.org/10.1093/nar/gkp830

Aptsiauri, N., Cabrera, T., Mendez, R., Garcia-Lora, A., Ruiz-Cabello, F., & Garrido, F. (2007). Role of altered expression of HLA class I molecules in cancer progression. *Adv Exp Med Biol*, *601*, 123-131. https://doi.org/10.1007/978-0-387-72005-0_13

Arruebo, M., Vilaboa, N., Saez-Gutierrez, B., Lambea, J., Tres, A., Valladares, M., & Gonzalez-Fernandez, A. (2011). Assessment of the evolution of cancer treatment therapies. *Cancers (Basel)*, *3*(3), 3279-3330. https://doi.org/10.3390/cancers3033279

Ataie-Kachoie, P., Pourgholami, M. H., Richardson, D. R., & Morris, D. L. (2014). Gene of the month: Interleukin 6 (IL-6). *J Clin Pathol*, *67*(11), 932-937. https://doi.org/10.1136/jclinpath-2014-202493

Atanasova, M., Patronov, A., Dimitrov, I., Flower, D. R., & Doytchinova, I. (2013). EpiDOCK: a molecular docking-based tool for MHC class II binding prediction. *Protein Eng Des Sel*, *26*(10), 631-634. https://doi.org/10.1093/protein/gzt018

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*, *28*(1), 45-48. https://doi.org/10.1093/nar/28.1.45

Banerjee, K., & Resat, H. (2016). Constitutive activation of STAT3 in breast cancer cells: A review. *Int J Cancer*, *138*(11), 2570-2578. https://doi.org/10.1002/ijc.29923

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehar, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jane-Valbuena, J., Mapa, F. A., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I. H., Cheng, J., Yu, G. K., Yu, J., Aspesi, P., Jr., de Silva, M., Jagtap, K., Jones, M. D., Wang, L., Hatton, C., Palescandolo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R. C., Liefeld, T., MacConaill, L., Winckler, W., Reich, M., Li, N., Mesirov, J. P., Gabriel, S. B., Getz, G., Ardlie, K., Chan, V., Myer, V. E., Weber, B. L., Porter, J., Warmuth, M., Finan, P., Harris, J. L., Meyerson, M., Golub, T. R., Morrissey, M. P., Sellers, W. R., Schlegel, R., & Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, *483*(7391), 603-607. https://doi.org/10.1038/nature11003

Bazarbachi, A., Plumelle, Y., Carlos Ramos, J., Tortevoye, P., Otrock, Z., Taylor, G., Gessain, A., Harrington, W., Panelatti, G., & Hermine, O. (2010). Meta-analysis on the use of zidovudine and interferon-alfa in adult T-cell leukemia/lymphoma showing improved survival in the leukemic subtypes. *J Clin Oncol*, *28*(27), 4177-4183. https://doi.org/10.1200/JCO.2010.28.0669

Beck, S., & Trowsdale, J. (2000). The human major histocompatability complex: lessons from the DNA sequence. *Annu Rev Genomics Hum Genet*, *1*, 117-137. https://doi.org/10.1146/annurev.genom.1.1.117

Belkaid, Y., & Hand, T. W. (2014). Role of the microbiota in immunity and inflammation. *Cell*, *157*(1), 121-141. https://doi.org/10.1016/j.cell.2014.03.011

Bezu, L., Kepp, O., Cerrato, G., Pol, J., Fucikova, J., Spisek, R., Zitvogel, L., Kroemer, G., & Galluzzi, L. (2018). Trial watch: Peptide-based vaccines in anticancer therapy. *Oncoimmunology*, *7*(12), e1511506. https://doi.org/10.1080/2162402X.2018.1511506

Bhalla, S., Kaur, H., Dhall, A., & Raghava, G. P. S. (2019). Prediction and Analysis of Skin Cancer Progression using Genomics Profiles of Patients. *Sci Rep*, *9*(1), 15790. https://doi.org/10.1038/s41598-019-52134-4

Bhasin, M., & Raghava, G. P. (2004). SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics*, *20*(3), 421-423. https://doi.org/10.1093/bioinformatics/btg424

Bhasin, M., & Raghava, G. P. (2007). A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. *J Biosci*, *32*(1), 31-42. https://doi.org/10.1007/s12038-007-0004-5

Bhasin, M., Singh, H., & Raghava, G. P. (2003). MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics*, *19*(5), 665-666. https://doi.org/10.1093/bioinformatics/btg055

Blackwell, J. M., Jamieson, S. E., & Burgner, D. (2009). HLA and infectious diseases. *Clin Microbiol Rev*, *22*(2), 370-385, Table of Contents. https://doi.org/10.1128/CMR.00048-08

Boegel, S., Lower, M., Schafer, M., Bukur, T., de Graaf, J., Boisguerin, V., Tureci, O., Diken, M., Castle, J. C., & Sahin, U. (2012). HLA typing from RNA-Seq sequence reads. *Genome Med*, *4*(12), 102. https://doi.org/10.1186/gm403

Boucherma, R., Kridane-Miledi, H., Bouziat, R., Rasmussen, M., Gatard, T., Langa-Vives, F., Lemercier, B., Lim, A., Berard, M., Benmohamed, L., Buus, S., Rooke, R., & Lemonnier, F. A. (2013). HLA-A*01:03, HLA-A*24:02, HLA-B*08:01, HLA-B*27:05, HLA-B*35:01, HLA-B*44:02, and HLA-C*07:01 monochain transgenic/H-2 class I null mice: novel versatile preclinical models of human T cell responses. *J Immunol*, *191*(2), 583-593. https://doi.org/10.4049/jimmunol.1300483

Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003). Survival analysis part II: multivariate data analysis--an introduction to concepts and methods. *Br J Cancer*, *89*(3), 431-436. https://doi.org/10.1038/sj.bjc.6601119

Brar, K., & Leung, D. Y. (2016). Recent considerations in the use of recombinant interferon gamma for biological therapy of atopic dermatitis. *Expert Opin Biol Ther*, *16*(4), 507-514. https://doi.org/10.1517/14712598.2016.1135898

Bristow, I. R., de Berker, D. A., Acland, K. M., Turner, R. J., & Bowling, J. (2010). Clinical guidelines for the recognition of melanoma of the foot and nail unit. *J Foot Ankle Res*, *3*, 25. https://doi.org/10.1186/1757-1146-3-25

Bryant, D., Becker, L., Richardson, J., Shelton, J., Franco, F., Peshock, R., Thompson, M., & Giroir, B. (1998). Cardiac failure in transgenic mice with myocardial expression of tumor necrosis factor-alpha. *Circulation*, *97*(14), 1375-1381. https://doi.org/10.1161/01.cir.97.14.1375

Buhrman, J. D., & Slansky, J. E. (2013). Improving T cell responses to modified peptides in tumor vaccines. *Immunol Res*, *55*(1-3), 34-47. https://doi.org/10.1007/s12026-012-8348-9

Burke, J. D., & Young, H. A. (2019). IFN-gamma: A cytokine at the right time, is in the right place. *Semin Immunol*, *43*, 101280. https://doi.org/10.1016/j.smim.2019.05.002

Buttner, P., Garbe, C., Bertz, J., Burg, G., d'Hoedt, B., Drepper, H., Guggenmoos-Holzmann, I., Lechner, W., Lippold, A., Orfanos, C. E., & et al. (1995). Primary cutaneous melanoma. Optimized cutoff points of tumor thickness and importance of Clark's level for prognostic classification. *Cancer*, *75*(10), 2499-2506. https://doi.org/10.1002/1097-0142(19950515)75:10<2499::aid-cncr2820751016>3.0.co;2-8

ca, B. I. G. S. P. W. (2007). The Cancer Biomedical Informatics Grid (caBIG): infrastructure and applications for a worldwide research community. *Stud Health Technol Inform*, *129*(Pt 1), 330-334. https://www.ncbi.nlm.nih.gov/pubmed/17911733

Cabrera, T., Lara, E., Romero, J. M., Maleno, I., Real, L. M., Ruiz-Cabello, F., Valero, P., Camacho, F. M., & Garrido, F. (2007). HLA class I expression in metastatic melanoma correlates with tumor development during autologous vaccination. *Cancer Immunol Immunother*, *56*(5), 709-717. https://doi.org/10.1007/s00262-006-0226-7

Caccamo, N., Sullivan, L. C., Brooks, A. G., & Dieli, F. (2020). Harnessing HLA-E-restricted CD8 T lymphocytes for adoptive cell therapy of patients with severe COVID-19. *Br J Haematol*, *190*(4), e185-e187. https://doi.org/10.1111/bjh.16895

Cain, B. S., Meldrum, D. R., Dinarello, C. A., Meng, X., Joo, K. S., Banerjee, A., & Harken, A. H. (1999). Tumor necrosis factor-alpha and interleukin-1beta synergistically depress human myocardial function. *Crit Care Med*, *27*(7), 1309-1318. https://doi.org/10.1097/00003246-199907000-00018

Calo, V., Migliavacca, M., Bazan, V., Macaluso, M., Buscemi, M., Gebbia, N., & Russo, A. (2003). STAT proteins: from normal control of cellular events to tumorigenesis. *J Cell Physiol*, *197*(2), 157-168. https://doi.org/10.1002/jcp.10364

Campillo, J. A., Martinez-Escribano, J. A., Muro, M., Moya-Quiles, R., Marin, L. A., Montes-Ares, O., Guerra, N., Sanchez-Pedreno, P., Frias, J. F., Lozano, J. A., Garcia-Alonso, A. M., & Alvarez-Lopez, M. R. (2006). HLA class I and class II frequencies in patients with cutaneous

malignant melanoma from southeastern Spain: the role of HLA-C in disease prognosis. *Immunogenetics*, *57*(12), 926-933. https://doi.org/10.1007/s00251-005-0065-2

Cao, Y., Wang, X., Jin, T., Tian, Y., Dai, C., Widarma, C., Song, R., & Xu, F. (2020). Immune checkpoint molecules in natural killer cells as potential targets for cancer immunotherapy. *Signal Transduct Target Ther*, *5*(1), 250. https://doi.org/10.1038/s41392-020-00348-8

Carosella, E. D., Gregori, S., Rouas-Freiss, N., LeMaoult, J., Menier, C., & Favier, B. (2011). The role of HLA-G in immunity and hematopoiesis. *Cell Mol Life Sci*, *68*(3), 353-368. https://doi.org/10.1007/s00018-010-0579-0

Castro, F., Cardoso, A. P., Goncalves, R. M., Serre, K., & Oliveira, M. J. (2018). Interferon-Gamma at the Crossroads of Tumor Immune Surveillance or Evasion. *Front Immunol*, *9*, 847. https://doi.org/10.3389/fimmu.2018.00847

Catamo, E., Zupin, L., Crovella, S., Celsi, F., & Segat, L. (2014). Non-classical MHC-I human leukocyte antigen (HLA-G) in hepatotropic viral infections and in hepatocellular carcinoma. *Hum Immunol*, *75*(12), 1225-1231. https://doi.org/10.1016/j.humimm.2014.09.019

Cavalcanti, Y. V., Brelaz, M. C., Neves, J. K., Ferraz, J. C., & Pereira, V. R. (2012). Role of TNF-Alpha, IFN-Gamma, and IL-10 in the Development of Pulmonary Tuberculosis. *Pulm Med*, *2012*, 745483. https://doi.org/10.1155/2012/745483

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., & Schultz, N. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*, *2*(5), 401-404. https://doi.org/10.1158/2159-8290.CD-12-0095

Chan, K. F., Gully, B. S., Gras, S., Beringer, D. X., Kjer-Nielsen, L., Cebon, J., McCluskey, J., Chen, W., & Rossjohn, J. (2018). Divergent T-cell receptor recognition modes of a HLA-I restricted extended tumour-associated peptide. *Nat Commun*, *9*(1), 1026. https://doi.org/10.1038/s41467-018-03321-w

Chaplin, D. D. (2010). Overview of the immune response. *J Allergy Clin Immunol*, *125*(2 Suppl 2), S3-23. https://doi.org/10.1016/j.jaci.2009.12.980

Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., Hackl, H., & Trajanoski, Z. (2017). Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep*, *18*(1), 248-262. https://doi.org/10.1016/j.celrep.2016.12.019

Chen, B., Khodadoust, M. S., Olsson, N., Wagar, L. E., Fast, E., Liu, C. L., Muftuoglu, Y., Sworder, B. J., Diehn, M., Levy, R., Davis, M. M., Elias, J. E., Altman, R. B., & Alizadeh, A. A. (2019). Predicting HLA class II antigen presentation through integrated deep learning. *Nat Biotechnol*, *37*(11), 1332-1343. https://doi.org/10.1038/s41587-019-0280-2

Chen, J., Petrus, M., Bamford, R., Shih, J. H., Morris, J. C., Janik, J. E., & Waldmann, T. A. (2012). Increased serum soluble IL-15Ralpha levels in T-cell large granular lymphocyte leukemia. *Blood*, *119*(1), 137-143. https://doi.org/10.1182/blood-2011-04-346759

Chen, X., Tang, J., Shuai, W., Meng, J., Feng, J., & Han, Z. (2020). Macrophage polarization and its role in the pathogenesis of acute lung injury/acute respiratory distress syndrome. *Inflamm Res*, *69*(9), 883-895. https://doi.org/10.1007/s00011-020-01378-2

Chen, X., Yang, J., Wang, L., & Liu, B. (2020). Personalized neoantigen vaccination with synthetic long peptides: recent advances and future perspectives. *Theranostics*, *10*(13), 6011-6023. https://doi.org/10.7150/thno.38742

Chen, X., Zhao, B., Qu, Y., Chen, Y., Xiong, J., Feng, Y., Men, D., Huang, Q., Liu, Y., Yang, B., Ding, J., & Li, F. (2020). Detectable Serum Severe Acute Respiratory Syndrome Coronavirus 2 Viral Load (RNAemia) Is Closely Correlated With Drastically Elevated Interleukin 6 Level in Critically Ill Patients With Coronavirus Disease 2019. *Clin Infect Dis*, *71*(8), 1937-1942. https://doi.org/10.1093/cid/ciaa449

Chonov, D. C., Ignatova, M. M. K., Ananiev, J. R., & Gulubova, M. V. (2019). IL-6 Activities in the Tumour Microenvironment. Part 1. *Open Access Maced J Med Sci*, *7*(14), 2391-2398. https://doi.org/10.3889/oamjms.2019.589

Choo, S. Y. (2007). The HLA system: genetics, immunology, clinical testing, and clinical implications. *Yonsei Med J*, *48*(1), 11-23. https://doi.org/10.3349/ymj.2007.48.1.11

Chowdhury, D., & Lieberman, J. (2008). Death by a thousand cuts: granzyme pathways of programmed cell death. *Annu Rev Immunol*, *26*, 389-420. https://doi.org/10.1146/annurev.immunol.26.021607.090404

Chowell, D., Morris, L. G. T., Grigg, C. M., Weber, J. K., Samstein, R. M., Makarov, V., Kuo, F., Kendall, S. M., Requena, D., Riaz, N., Greenbaum, B., Carroll, J., Garon, E., Hyman, D. M., Zehir, A., Solit, D., Berger, M., Zhou, R., Rizvi, N. A., & Chan, T. A. (2018). Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science*, *359*(6375), 582-587. https://doi.org/10.1126/science.aao4572

Clem, A. S. (2011). Fundamentals of vaccine immunology. *J Glob Infect Dis*, *3*(1), 73-78. https://doi.org/10.4103/0974-777X.77299

Clough, E., & Barrett, T. (2016). The Gene Expression Omnibus Database. *Methods Mol Biol*, *1418*, 93-110. https://doi.org/10.1007/978-1-4939-3578-9_5

Conlon, K. C., Miljkovic, M. D., & Waldmann, T. A. (2019). Cytokines in the Treatment of Cancer. *J Interferon Cytokine Res*, *39*(1), 6-21. https://doi.org/10.1089/jir.2018.0019

Consortium, G. T. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, *45*(6), 580-585. https://doi.org/10.1038/ng.2653

Corvinus, F. M., Orth, C., Moriggl, R., Tsareva, S. A., Wagner, S., Pfitzner, E. B., Baus, D., Kaufmann, R., Huber, L. A., Zatloukal, K., Beug, H., Ohlschlager, P., Schutz, A., Halbhuber, K. J., & Friedrich, K. (2005). Persistent STAT3 activation in colon cancer is associated with enhanced cell proliferation and tumor growth. *Neoplasia*, *7*(6), 545-555. https://doi.org/10.1593/neo.04571

Costela-Ruiz, V. J., Illescas-Montes, R., Puerta-Puerta, J. M., Ruiz, C., & Melguizo-Rodriguez, L. (2020). SARS-CoV-2 infection: The role of cytokines in COVID-19 disease. *Cytokine Growth Factor Rev*, *54*, 62-75. https://doi.org/10.1016/j.cytogfr.2020.06.001

Crew, M. D., Cannon, M. J., Phanavanh, B., & Garcia-Borges, C. N. (2005). An HLA-E single chain trimer inhibits human NK cell reactivity towards porcine cells. *Mol Immunol*, *42*(10), 1205-1214. https://doi.org/10.1016/j.molimm.2004.11.013

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res*, *14*(6), 1188-1190. https://doi.org/10.1101/gr.849004

Crux, N. B., & Elahi, S. (2017). Human Leukocyte Antigen (HLA) and Immune Regulation: How Do Classical and Non-Classical HLA Alleles Modulate Immune Response to Human Immunodeficiency Virus and Hepatitis C Virus Infections? *Front Immunol*, *8*, 832. https://doi.org/10.3389/fimmu.2017.00832

Cruz, B., Oliveira, A., & Gomes-Marcondes, M. C. C. (2017). L-leucine dietary supplementation modulates muscle protein degradation and increases pro-inflammatory cytokines in tumour-bearing rats. *Cytokine*, *96*, 253-260. https://doi.org/10.1016/j.cyto.2017.04.019

Curti, B. D. (2006). Immunomodulatory and antitumor effects of interleukin-21 in patients with renal cell carcinoma. *Expert Rev Anticancer Ther*, *6*(6), 905-909. https://doi.org/10.1586/14737140.6.6.905

Deenick, E. K., Pelham, S. J., Kane, A., & Ma, C. S. (2018). Signal Transducer and Activator of Transcription 3 Control of Human T and B Cell Responses. *Front Immunol*, *9*, 168. https://doi.org/10.3389/fimmu.2018.00168

Del Valle, D. M., Kim-Schulze, S., Huang, H. H., Beckmann, N. D., Nirenberg, S., Wang, B., Lavin, Y., Swartz, T. H., Madduri, D., Stock, A., Marron, T. U., Xie, H., Patel, M., Tuballes, K., Van Oekelen, O., Rahman, A., Kovatch, P., Aberg, J. A., Schadt, E., Jagannath, S., Mazumdar, M.,

Charney, A. W., Firpo-Betancourt, A., Mendu, D. R., Jhang, J., Reich, D., Sigel, K., Cordon-Cardo, C., Feldmann, M., Parekh, S., Merad, M., & Gnjatic, S. (2020). An inflammatory cytokine signature predicts COVID-19 severity and survival. *Nat Med*, *26*(10), 1636-1643. https://doi.org/10.1038/s41591-020-1051-9

Demberg, T., & Robert-Guroff, M. (2009). Mucosal immunity and protection against HIV/SIV infection: strategies and challenges for vaccine design. *Int Rev Immunol*, *28*(1), 20-48. https://doi.org/10.1080/08830180802684331

Dhall, A., Patiyal, S., Kaur, H., Bhalla, S., Arora, C., & Raghava, G. P. S. (2020). Computing Skin Cutaneous Melanoma Outcome From the HLA-Alleles and Clinical Characteristics. *Front Genet*, *11*, 221. https://doi.org/10.3389/fgene.2020.00221

Dhall, A., Patiyal, S., & Raghava, G. P. S. (2022). HLAncPred: a method for predicting promiscuous non-classical HLA binding sites. *Brief Bioinform*. https://doi.org/10.1093/bib/bbac192

Dhall, A., Patiyal, S., Sharma, N., Devi, N. L., & Raghava, G. P. S. (2021). Computer-aided prediction of inhibitors against STAT3 for managing COVID-19 associated cytokine storm. *Comput Biol Med*, *137*, 104780. https://doi.org/10.1016/j.compbiomed.2021.104780

Dhall, A., Patiyal, S., Sharma, N., Usmani, S. S., & Raghava, G. P. S. (2021). Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. *Brief Bioinform*, *22*(2), 936-945. https://doi.org/10.1093/bib/bbaa259

Dhanda, S. K., Gupta, S., Vir, P., & Raghava, G. P. (2013). Prediction of IL4 inducing peptides. *Clin Dev Immunol*, *2013*, 263952. https://doi.org/10.1155/2013/263952

Dhanda, S. K., Mahajan, S., Paul, S., Yan, Z., Kim, H., Jespersen, M. C., Jurtz, V., Andreatta, M., Greenbaum, J. A., Marcatili, P., Sette, A., Nielsen, M., & Peters, B. (2019). IEDB-AR: immune epitope database-analysis resource in 2019. *Nucleic Acids Res*, *47*(W1), W502-W506. https://doi.org/10.1093/nar/gkz452

Dhanda, S. K., Singla, D., Mondal, A. K., & Raghava, G. P. (2013). DrugMint: a webserver for predicting and designing of drug-like molecules. *Biol Direct*, *8*, 28. https://doi.org/10.1186/1745-6150-8-28

Dhanda, S. K., Vir, P., & Raghava, G. P. (2013). Designing of interferon-gamma inducing MHC class-II binders. *Biol Direct*, *8*, 30. https://doi.org/10.1186/1745-6150-8-30

Dias, F. C., Castelli, E. C., Collares, C. V., Moreau, P., & Donadi, E. A. (2015). The Role of HLA-G Molecule and HLA-G Gene Polymorphisms in Tumors, Viral Hepatitis, and Parasitic Diseases. *Front Immunol*, *6*, 9. https://doi.org/10.3389/fimmu.2015.00009

Dickson, P. V., & Gershenwald, J. E. (2011). Staging and prognosis of cutaneous melanoma. *Surg Oncol Clin N Am*, *20*(1), 1-17. https://doi.org/10.1016/j.soc.2010.09.007

Dilalla, V., Chaput, G., Williams, T., & Sultanem, K. (2020). Radiotherapy side effects: integrating a survivorship clinical lens to better serve patients. *Curr Oncol*, *27*(2), 107-112. https://doi.org/10.3747/co.27.6233

Dilthey, A. T., Mentzer, A. J., Carapito, R., Cutland, C., Cereb, N., Madhi, S. A., Rhie, A., Koren, S., Bahram, S., McVean, G., & Phillippy, A. M. (2019). HLA*LA-HLA typing from linearly projected graph alignments. *Bioinformatics*, *35*(21), 4394-4396. https://doi.org/10.1093/bioinformatics/btz235

Dimitrov, I., Garnev, P., Flower, D. R., & Doytchinova, I. (2010). EpiTOP--a proteochemometric tool for MHC class II binding prediction. *Bioinformatics*, *26*(16), 2066-2068. https://doi.org/10.1093/bioinformatics/btq324

Dine, J., Gordon, R., Shames, Y., Kasler, M. K., & Barton-Burke, M. (2017). Immune Checkpoint Inhibitors: An Innovation in Immunotherapy for the Treatment and Management of Patients with Cancer. *Asia Pac J Oncol Nurs*, *4*(2), 127-135. https://doi.org/10.4103/apjon.apjon_4_17

Dreyer, L., Mellemkjaer, L., & Hetland, M. L. (2009). [Cancer in arthritis patients after anti-tumour necrosis factor therapy]. *Ugeskr Laeger*, *171*(7), 506-511.

https://www.ncbi.nlm.nih.gov/pubmed/19210932 (Cancer blandt gigtpatienter behandlet med tumornekrosefaktor-alfa-haemmere.)

Dunkelberger, J. R., & Song, W. C. (2010). Complement and its role in innate and adaptive immune responses. *Cell Res*, *20*(1), 34-50. https://doi.org/10.1038/cr.2009.139

Engels, B., Engelhard, V. H., Sidney, J., Sette, A., Binder, D. C., Liu, R. B., Kranz, D. M., Meredith, S. C., Rowley, D. A., & Schreiber, H. (2013). Relapse or eradication of cancer is predicted by peptide-major histocompatibility complex affinity. *Cancer Cell*, *23*(4), 516-526. https://doi.org/10.1016/j.ccr.2013.03.018

Esfahani, K., Roudaia, L., Buhlaiga, N., Del Rincon, S. V., Papneja, N., & Miller, W. H., Jr. (2020). A review of cancer immunotherapy: from the past, to the present, to the future. *Curr Oncol*, *27*(Suppl 2), S87-S97. https://doi.org/10.3747/co.27.5223

Esquivel-Velazquez, M., Ostoa-Saloma, P., Palacios-Arreola, M. I., Nava-Castro, K. E., Castro, J. I., & Morales-Montor, J. (2015). The role of cytokines in breast cancer development and progression. *J Interferon Cytokine Res*, *35*(1), 1-16. https://doi.org/10.1089/jir.2014.0026

Evangelatos, G., Bamias, G., Kitas, G. D., Kollias, G., & Sfikakis, P. P. (2022). The second decade of anti-TNF-a therapy in clinical practice: new lessons and future directions in the COVID-19 era. *Rheumatol Int*. https://doi.org/10.1007/s00296-022-05136-x

Farrar, M. A., & Schreiber, R. D. (1993). The molecular cell biology of interferon-gamma and its receptor. *Annu Rev Immunol*, *11*, 571-611. https://doi.org/10.1146/annurev.iy.11.040193.003035

Feng, X., Li, L., Wagner, E. J., & Li, W. (2018). TC3A: The Cancer 3' UTR Atlas. *Nucleic Acids Res*, *46*(D1), D1027-D1030. https://doi.org/10.1093/nar/gkx892

Feola, S., Chiaro, J., Martins, B., & Cerullo, V. (2020). Uncovering the Tumor Antigen Landscape: What to Know about the Discovery Process. *Cancers (Basel)*, *12*(6). https://doi.org/10.3390/cancers12061660

Forbes, L. R., Milner, J., & Haddad, E. (2016). Signal transducer and activator of transcription 3: a year in review. *Curr Opin Hematol*, *23*(1), 23-27. https://doi.org/10.1097/MOH.0000000000000206

Franzin, R., Netti, G. S., Spadaccino, F., Porta, C., Gesualdo, L., Stallone, G., Castellano, G., & Ranieri, E. (2020). The Use of Immune Checkpoint Inhibitors in Oncology and the Occurrence of AKI: Where Do We Stand? *Front Immunol*, *11*, 574271. https://doi.org/10.3389/fimmu.2020.574271

Furqan, M., Mukhi, N., Lee, B., & Liu, D. (2013). Dysregulation of JAK-STAT pathway in hematological malignancies and JAK inhibitors for clinical application. *Biomark Res*, *1*(1), 5. https://doi.org/10.1186/2050-7771-1-5

Gallagher, M. P., Kelly, P. J., Jardine, M., Perkovic, V., Cass, A., Craig, J. C., Eris, J., & Webster, A. C. (2010). Long-term cancer risk of immunosuppressive regimens after kidney transplantation. *J Am Soc Nephrol*, *21*(5), 852-858. https://doi.org/10.1681/ASN.2009101043

Garcia, P., Llano, M., de Heredia, A. B., Willberg, C. B., Caparros, E., Aparicio, P., Braud, V. M., & Lopez-Botet, M. (2002). Human T cell receptor-mediated recognition of HLA-E. *Eur J Immunol*, *32*(4), 936-944. https://doi.org/10.1002/1521-4141(200204)32:4<936::AID-IMMU936>3.0.CO;2-M

Garg, A., Tewari, R., & Raghava, G. P. (2010). KiDoQ: using docking based energy scores to develop ligand based model for predicting antibacterials. *BMC Bioinformatics*, *11*, 125. https://doi.org/10.1186/1471-2105-11-125

Garrido, F., & Aptsiauri, N. (2019). Cancer immune escape: MHC expression in primary tumours versus metastases. *Immunology*, *158*(4), 255-266. https://doi.org/10.1111/imm.13114

Gershenwald, J. E., Buzaid, A. C., & Ross, M. I. (1998). Classification and staging of melanoma. *Hematol Oncol Clin North Am*, *12*(4), 737-765. https://doi.org/10.1016/s0889-8588(05)70021-6

Goel, M. K., Khanna, P., & Kishore, J. (2010). Understanding survival analysis: Kaplan-Meier estimate. *Int J Ayurveda Res*, *1*(4), 274-278. https://doi.org/10.4103/0974-7788.76794

Gogas, H., Kirkwood, J. M., Falk, C. S., Dafni, U., Sondak, V. K., Tsoutsos, D., Stratigos, A., Markopoulos, C., Pectasides, D., & Spyropoulou-Vlachou, M. (2010). Correlation of molecular human leukocyte antigen typing and outcome in high-risk melanoma patients receiving adjuvant interferon. *Cancer*, *116*(18), 4326-4333. https://doi.org/10.1002/cncr.25211

Gollob, J. A., Mier, J. W., Veenstra, K., McDermott, D. F., Clancy, D., Clancy, M., & Atkins, M. B. (2000). Phase I trial of twice-weekly intravenous interleukin 12 in patients with metastatic renal cell cancer or malignant melanoma: ability to maintain IFN-gamma induction is associated with clinical response. *Clin Cancer Res*, *6*(5), 1678-1692. https://www.ncbi.nlm.nih.gov/pubmed/10815886

Gonzalez, H., Hagerling, C., & Werb, Z. (2018). Roles of the immune system in cancer: from tumor initiation to metastatic progression. *Genes Dev*, *32*(19-20), 1267-1284. https://doi.org/10.1101/gad.314617.118

Grivennikov, S. I., & Karin, M. (2011). Inflammatory cytokines in cancer: tumour necrosis factor and interleukin 6 take the stage. *Ann Rheum Dis*, *70 Suppl 1*, i104-108. https://doi.org/10.1136/ard.2010.140145

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med*, *375*(12), 1109-1112. https://doi.org/10.1056/NEJMp1607591

Guallar-Garrido, S., & Julian, E. (2020). Bacillus Calmette-Guerin (BCG) Therapy for Bladder Cancer: An Update. *Immunotargets Ther*, *9*, 1-11. https://doi.org/10.2147/ITT.S202006

Guan, P., Doytchinova, I. A., Zygouri, C., & Flower, D. R. (2003). MHCPred: A server for quantitative prediction of peptide-MHC binding. *Nucleic Acids Res*, *31*(13), 3621-3624. https://doi.org/10.1093/nar/gkg510

Gubernatorova, E. O., Gorshkova, E. A., Polinova, A. I., & Drutskaya, M. S. (2020). IL-6: Relevance for immunopathology of SARS-CoV-2. *Cytokine Growth Factor Rev*, *53*, 13-24. https://doi.org/10.1016/j.cytogfr.2020.05.009

Guo, J., Li, X., Shen, S., & Wu, X. (2021). Expression of immune-related genes as prognostic biomarkers for the assessment of osteosarcoma clinical outcomes. *Sci Rep*, *11*(1), 24123. https://doi.org/10.1038/s41598-021-03677-y

Guo, Y., Hu, K., Li, Y., Lu, C., Ling, K., Cai, C., Wang, W., & Ye, D. (2022). Targeting TNF-alpha for COVID-19: Recent Advanced and Controversies. *Front Public Health*, *10*, 833967. https://doi.org/10.3389/fpubh.2022.833967

Gupta, S., Madhu, M. K., Sharma, A. K., & Sharma, V. K. (2016). ProInflam: a webserver for the prediction of proinflammatory antigenicity of peptides and proteins. *J Transl Med*, *14*(1), 178. https://doi.org/10.1186/s12967-016-0928-3

Gupta, S., Mittal, P., Madhu, M. K., & Sharma, V. K. (2017). IL17eScan: A Tool for the Identification of Peptides Inducing IL-17 Response. *Front Immunol*, *8*, 1430. https://doi.org/10.3389/fimmu.2017.01430

Gupta, S., Sharma, A. K., Shastri, V., Madhu, M. K., & Sharma, V. K. (2017). Prediction of anti-inflammatory proteins/peptides: an insilico approach. *J Transl Med*, *15*(1), 7. https://doi.org/10.1186/s12967-016-1103-6

Halim, C., Mirza, A. F., & Sari, M. I. (2022). The Association between TNF-alpha, IL-6, and Vitamin D Levels and COVID-19 Severity and Mortality: A Systematic Review and Meta-Analysis. *Pathogens*, *11*(2). https://doi.org/10.3390/pathogens11020195

Halle, M. K., Sodal, M., Forsse, D., Engerud, H., Woie, K., Lura, N. G., Wagner-Larsen, K. S., Trovik, J., Bertelsen, B. I., Haldorsen, I. S., Ojesina, A. I., & Krakstad, C. (2021). A 10-gene prognostic signature points to LIMCH1 and HLA-DQB1 as important players in aggressive cervical

cancer disease. *Br J Cancer*, *124*(10), 1690-1698. https://doi.org/10.1038/s41416-021-01305-0

He, Q., Jiang, X., Zhou, X., & Weng, J. (2019). Targeting cancers through TCR-peptide/MHC interactions. *J Hematol Oncol*, *12*(1), 139. https://doi.org/10.1186/s13045-019-0812-8

Hemminki, O., Dos Santos, J. M., & Hemminki, A. (2020). Oncolytic viruses for cancer immunotherapy. *J Hematol Oncol*, *13*(1), 84. https://doi.org/10.1186/s13045-020-00922-1

Herceg, Z., & Hainaut, P. (2007). Genetic and epigenetic alterations as biomarkers for cancer detection, diagnosis and prognosis. *Mol Oncol*, *1*(1), 26-41. https://doi.org/10.1016/j.molonc.2007.01.004

Hewitt, E. W. (2003). The MHC class I antigen presentation pathway: strategies for viral immune evasion. *Immunology*, *110*(2), 163-169. https://doi.org/10.1046/j.1365-2567.2003.01738.x

Hirano, T. (2021). IL-6 in inflammation, autoimmunity and cancer. *Int Immunol*, *33*(3), 127-148. https://doi.org/10.1093/intimm/dxaa078

Ho, G. T., Celik, A. A., Huyton, T., Hiemisch, W., Blasczyk, R., Simper, G. S., & Bade-Doeding, C. (2020). NKG2A/CD94 Is a New Immune Receptor for HLA-G and Distinguishes Amino Acid Differences in the HLA-G Heavy Chain. *Int J Mol Sci*, *21*(12). https://doi.org/10.3390/ijms21124362

Holbrook, J., Lara-Reyna, S., Jarosz-Griffiths, H., & McDermott, M. (2019). Tumour necrosis factor signalling in health and disease. *F1000Res*, *8*. https://doi.org/10.12688/f1000research.17023.1

Holdenrieder, S., Pagliaro, L., Morgenstern, D., & Dayyani, F. (2016). Clinically Meaningful Use of Blood Tumor Markers in Oncology. *Biomed Res Int*, *2016*, 9795269. https://doi.org/10.1155/2016/9795269

Hong, D. S., Angelo, L. S., & Kurzrock, R. (2007). Interleukin-6 and its receptor in cancer: implications for translational therapeutics. *Cancer*, *110*(9), 1911-1928. https://doi.org/10.1002/cncr.22999

Hong, H., Wang, Q., Li, J., Liu, H., Meng, X., & Zhang, H. (2019). Aging, Cancer and Immunity. *J Cancer*, *10*(13), 3021-3027. https://doi.org/10.7150/jca.30723

Hosomichi, K., Shiina, T., Tajima, A., & Inoue, I. (2015). The impact of next-generation sequencing technologies on HLA research. *J Hum Genet*, *60*(11), 665-673. https://doi.org/10.1038/jhg.2015.102

Huang, Y., Yang, J., Ying, D., Zhang, Y., Shotelersuk, V., Hirankarn, N., Sham, P. C., Lau, Y. L., & Yang, W. (2015). HLAreporter: a tool for HLA typing from next generation sequencing data. *Genome Med*, *7*(1), 25. https://doi.org/10.1186/s13073-015-0145-3

Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., Seja, E., Lomeli, S., Kong, X., Kelley, M. C., Sosman, J. A., Johnson, D. B., Ribas, A., & Lo, R. S. (2016). Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell*, *165*(1), 35-44. https://doi.org/10.1016/j.cell.2016.02.065

Hutchison, S., & Pritchard, A. L. (2018). Identifying neoantigens for use in immunotherapy. *Mamm Genome*, *29*(11-12), 714-730. https://doi.org/10.1007/s00335-018-9771-6

Idriss, H. T., & Naismith, J. H. (2000). TNF alpha and the TNF receptor superfamily: structure-function relationship(s). *Microsc Res Tech*, *50*(3), 184-195. https://doi.org/10.1002/1097-0029(20000801)50:3<184::AID-JEMT2>3.0.CO;2-H

Jafarzadeh, A., Chauhan, P., Saha, B., Jafarzadeh, S., & Nemati, M. (2020). Contribution of monocytes and macrophages to the local tissue inflammation and cytokine storm in COVID-19: Lessons from SARS and MERS, and potential therapeutic interventions. *Life Sci*, *257*, 118102. https://doi.org/10.1016/j.lfs.2020.118102

Jafarzadeh, A., Jafarzadeh, S., Nozari, P., Mokhtari, P., & Nemati, M. (2021). Lymphopenia an important immunological abnormality in patients with COVID-19: Possible mechanisms. *Scand J Immunol*, *93*(2), e12967. https://doi.org/10.1111/sji.12967

Jafarzadeh, A., Nemati, M., & Jafarzadeh, S. (2021). Contribution of STAT3 to the pathogenesis of COVID-19. *Microb Pathog*, *154*, 104836. https://doi.org/10.1016/j.micpath.2021.104836

Jain, S., Dhall, A., Patiyal, S., & Raghava, G. P. S. (2022). IL13Pred: A method for predicting immunoregulatory cytokine IL-13 inducing peptides. *Comput Biol Med*, *143*, 105297. https://doi.org/10.1016/j.compbiomed.2022.105297

Jain, S., Gautam, V., & Naseem, S. (2011). Acute-phase proteins: As diagnostic tool. *J Pharm Bioallied Sci*, *3*(1), 118-127. https://doi.org/10.4103/0975-7406.76489

Jensen, K. K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J. A., Yan, Z., Sette, A., Peters, B., & Nielsen, M. (2018). Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology*, *154*(3), 394-406. https://doi.org/10.1111/imm.12889

Jensen, M. A., Ferretti, V., Grossman, R. L., & Staudt, L. M. (2017). The NCI Genomic Data Commons as an engine for precision medicine. *Blood*, *130*(4), 453-459. https://doi.org/10.1182/blood-2017-03-735654

Johansen, L. L., Lock-Andersen, J., & Hviid, T. V. (2016). The Pathophysiological Impact of HLA Class Ia and HLA-G Expression and Regulatory T Cells in Malignant Melanoma: A Review. *J Immunol Res*, *2016*, 6829283. https://doi.org/10.1155/2016/6829283

Johnson, D. E., O'Keefe, R. A., & Grandis, J. R. (2018). Targeting the IL-6/JAK/STAT3 signalling axis in cancer. *Nat Rev Clin Oncol*, *15*(4), 234-248. https://doi.org/10.1038/nrclinonc.2018.8

Joosten, S. A., Sullivan, L. C., & Ottenhoff, T. H. (2016). Characteristics of HLA-E Restricted T-Cell Responses and Their Role in Infectious Diseases. *J Immunol Res*, *2016*, 2695396. https://doi.org/10.1155/2016/2695396

Jorgovanovic, D., Song, M., Wang, L., & Zhang, Y. (2020). Roles of IFN-gamma in tumor progression and regression: a review. *Biomark Res*, *8*, 49. https://doi.org/10.1186/s40364-020-00228-x

Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., & Nielsen, M. (2017). NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol*, *199*(9), 3360-3368. https://doi.org/10.4049/jimmunol.1700893

Kamran, M. Z., Patil, P., & Gude, R. P. (2013). Role of STAT3 in cancer metastasis and translational advances. *Biomed Res Int*, *2013*, 421821. https://doi.org/10.1155/2013/421821

Kany, S., Vollrath, J. T., & Relja, B. (2019). Cytokines in Inflammatory Disease. *Int J Mol Sci*, *20*(23). https://doi.org/10.3390/ijms20236008

Karosiene, E., Lundegaard, C., Lund, O., & Nielsen, M. (2012). NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics*, *64*(3), 177-186. https://doi.org/10.1007/s00251-011-0579-8

Kaufman, H. L., Ruby, C. E., Hughes, T., & Slingluff, C. L., Jr. (2014). Current status of granulocyte-macrophage colony-stimulating factor in the immunotherapy of melanoma. *J Immunother Cancer*, *2*, 11. https://doi.org/10.1186/2051-1426-2-11

Kerr, D. J., & Yang, L. (2021). Personalising cancer medicine with prognostic markers. *EBioMedicine*, *72*, 103577. https://doi.org/10.1016/j.ebiom.2021.103577

Khair, D. O., Bax, H. J., Mele, S., Crescioli, S., Pellizzari, G., Khiabany, A., Nakamura, M., Harris, R. J., French, E., Hoffmann, R. M., Williams, I. P., Cheung, A., Thair, B., Beales, C. T., Touizer, E., Signell, A. W., Tasnova, N. L., Spicer, J. F., Josephs, D. H., Geh, J. L., MacKenzie Ross, A., Healy, C., Papa, S., Lacy, K. E., & Karagiannis, S. N. (2019). Combining Immune Checkpoint Inhibitors: Established and Emerging Targets and Strategies to Improve Outcomes in Melanoma. *Front Immunol*, *10*, 453. https://doi.org/10.3389/fimmu.2019.00453

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*, *37*(8), 907-915. https://doi.org/10.1038/s41587-019-0201-4

Klapper, J. A., Downey, S. G., Smith, F. O., Yang, J. C., Hughes, M. S., Kammula, U. S., Sherry, R. M., Royal, R. E., Steinberg, S. M., & Rosenberg, S. (2008). High-dose interleukin-2 for the

treatment of metastatic renal cell carcinoma : a retrospective analysis of response and survival in patients treated in the surgery branch at the National Cancer Institute between 1986 and 2006. *Cancer*, *113*(2), 293-301. https://doi.org/10.1002/cncr.23552

Kosaloglu-Yalcin, Z., Sidney, J., Chronister, W., Peters, B., & Sette, A. (2021). Comparison of HLA ligand elution data and binding predictions reveals varying prediction performance for the multiple motifs recognized by HLA-DQ2.5. *Immunology*, *162*(2), 235-247. https://doi.org/10.1111/imm.13279

Kountouri, A., Korakas, E., Ikonomidis, I., Raptis, A., Tentolouris, N., Dimitriadis, G., & Lambadiari, V. (2021). Type 1 Diabetes Mellitus in the SARS-CoV-2 Pandemic: Oxidative Stress as a Major Pathophysiological Mechanism Linked to Adverse Clinical Outcomes. *Antioxidants (Basel)*, *10*(5). https://doi.org/10.3390/antiox10050752

Kovats, S., Main, E. K., Librach, C., Stubblebine, M., Fisher, S. J., & DeMars, R. (1990). A class I antigen, HLA-G, expressed in human trophoblasts. *Science*, *248*(4952), 220-223. https://doi.org/10.1126/science.2326636

Kraemer, T., Blasczyk, R., & Bade-Doeding, C. (2014). HLA-E: a novel player for histocompatibility. *J Immunol Res*, *2014*, 352160. https://doi.org/10.1155/2014/352160

Kubo, M., Nagashima, R., Kurihara, M., Kawakami, F., Maekawa, T., Eshima, K., Ohta, E., Kato, H., & Obata, F. (2020). Leucine-Rich Repeat Kinase 2 Controls Inflammatory Cytokines Production through NF-kappaB Phosphorylation and Antigen Presentation in Bone Marrow-Derived Dendritic Cells. *Int J Mol Sci*, *21*(5). https://doi.org/10.3390/ijms21051890

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harb Protoc*, *2015*(11), 951-969. https://doi.org/10.1101/pdb.top084970

Kulkarni-Kale, U., Raskar-Renuse, S., Natekar-Kalantre, G., & Saxena, S. A. (2014). Antigen-Antibody Interaction Database (AgAbDb): a compendium of antigen-antibody interactions. *Methods Mol Biol*, *1184*, 149-164. https://doi.org/10.1007/978-1-4939-1115-8_8

Kumai, T., Kobayashi, H., Harabuchi, Y., & Celis, E. (2017). Peptide vaccines in cancer-old concept revisited. *Curr Opin Immunol*, *45*, 1-7. https://doi.org/10.1016/j.coi.2016.11.001

Kumar, R., Lathwal, A., Kumar, V., Patiyal, S., Raghav, P. K., & Raghava, G. P. S. (2020). CancerEnD: A database of cancer associated enhancers. *Genomics*, *112*(5), 3696-3702. https://doi.org/10.1016/j.ygeno.2020.04.028

Kunigal, S., Lakka, S. S., Sodadasu, P. K., Estes, N., & Rao, J. S. (2009). Stat3-siRNA induces Fas-mediated apoptosis in vitro and in vivo in breast cancer. *Int J Oncol*, *34*(5), 1209-1220. https://www.ncbi.nlm.nih.gov/pubmed/19360334

Kursunel, M. A., & Esendagli, G. (2016). The untold story of IFN-gamma in cancer biology. *Cytokine Growth Factor Rev*, *31*, 73-81. https://doi.org/10.1016/j.cytogfr.2016.07.005

Laconi, E., Marongiu, F., & DeGregori, J. (2020). Cancer as a disease of old age: changing mutational and microenvironmental landscapes. *Br J Cancer*, *122*(7), 943-952. https://doi.org/10.1038/s41416-019-0721-1

Lane, B. R., Markovitz, D. M., Woodford, N. L., Rochford, R., Strieter, R. M., & Coffey, M. J. (1999). TNF-alpha inhibits HIV-1 replication in peripheral blood monocytes and alveolar macrophages by inducing the production of RANTES and decreasing C-C chemokine receptor 5 (CCR5) expression. *J Immunol*, *163*(7), 3653-3661. https://www.ncbi.nlm.nih.gov/pubmed/10490959

Lata, S., & Raghava, G. P. (2008). CytoPred: a server for prediction and classification of cytokines. *Protein Eng Des Sel*, *21*(4), 279-282. https://doi.org/10.1093/protein/gzn006

Lee, H., Jeong, A. J., & Ye, S. K. (2019). Highlighted STAT3 as a potential drug target for cancer therapy. *BMB Rep*, *52*(7), 415-423. https://www.ncbi.nlm.nih.gov/pubmed/31186087

Lee, H., & Kingsford, C. (2018). Kourami: graph-guided assembly for novel human leukocyte antigen allele discovery. *Genome Biol*, *19*(1), 16. https://doi.org/10.1186/s13059-018-1388-2

Leinonen, R., Sugawara, H., Shumway, M., & International Nucleotide Sequence Database, C. (2011). The sequence read archive. *Nucleic Acids Res*, *39*(Database issue), D19-21. https://doi.org/10.1093/nar/gkq1019

Levy, D. E., & Lee, C. K. (2002). What does Stat3 do? *J Clin Invest*, *109*(9), 1143-1148. https://doi.org/10.1172/JCI15650

Li, Y., Krahn, J. M., Flake, G. P., Umbach, D. M., & Li, L. (2015). Toward predicting metastatic progression of melanoma based on gene expression data. *Pigment Cell Melanoma Res*, *28*(4), 453-463. https://doi.org/10.1111/pcmr.12374

Liu, C., Yang, X., Duffy, B., Mohanakumar, T., Mitra, R. D., Zody, M. C., & Pfeifer, J. D. (2013). ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. *Nucleic Acids Res*, *41*(14), e142. https://doi.org/10.1093/nar/gkt481

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., Benz, C. C., Levine, D. A., Lee, A. V., Omberg, L., Wolf, D. M., Shriver, C. D., Thorsson, V., Cancer Genome Atlas Research, N., & Hu, H. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*, *173*(2), 400-416 e411. https://doi.org/10.1016/j.cell.2018.02.052

Liu, S., Galat, V., Galat, Y., Lee, Y. K. A., Wainwright, D., & Wu, J. (2021). NK cell-based cancer immunotherapy: from basic biology to clinical development. *J Hematol Oncol*, *14*(1), 7. https://doi.org/10.1186/s13045-020-01014-w

Liu, S. Q., Wang, L. Y., Liu, G. H., Tang, D. Z., Fan, X. X., Zhao, J. P., Jiao, H. C., Wang, X. J., Sun, S. H., & Lin, H. (2018). Leucine alters immunoglobulin a secretion and inflammatory cytokine expression induced by lipopolysaccharide via the nuclear factor-kappaB pathway in intestine of chicken embryos. *Animal*, *12*(9), 1903-1911. https://doi.org/10.1017/S1751731117003342

Liu, Y., Zhang, J., Jia, Z. M., Li, J. C., Dong, C. H., & Li, Y. M. (2015). The correlation between HLA-DRB1 and HLA-DQB1 gene polymorphisms and cytokines in HPV16 infected women with advanced cervical cancer. *Int J Clin Exp Med*, *8*(7), 11490-11495. https://www.ncbi.nlm.nih.gov/pubmed/26379968

Locksley, R. M., Killeen, N., & Lenardo, M. J. (2001). The TNF and TNF receptor superfamilies: integrating mammalian biology. *Cell*, *104*(4), 487-501. https://doi.org/10.1016/s0092-8674(01)00237-9

Ma, J. H., Qin, L., & Li, X. (2020). Role of STAT3 signaling pathway in breast cancer. *Cell Commun Signal*, *18*(1), 33. https://doi.org/10.1186/s12964-020-0527-z

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., Popova, N., Pretel, S., Ziyabari, L., Lee, M., Shao, Y., Wang, Z. Y., Sirotkin, K., Ward, M., Kholodov, M., Zbicz, K., Beck, J., Kimelman, M., Shevelev, S., Preuss, D., Yaschenko, E., Graeff, A., Ostell, J., & Sherry, S. T. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*, *39*(10), 1181-1186. https://doi.org/10.1038/ng1007-1181

Manavalan, B., Shin, T. H., Kim, M. O., & Lee, G. (2018). PIP-EL: A New Ensemble Learning Method for Improved Proinflammatory Peptide Predictions. *Front Immunol*, *9*, 1783. https://doi.org/10.3389/fimmu.2018.01783

Marabondo, S., & Kaufman, H. L. (2017). High-dose interleukin-2 (IL-2) for the treatment of melanoma: safety considerations and future directions. *Expert Opin Drug Saf*, *16*(12), 1347-1357. https://doi.org/10.1080/14740338.2017.1382472

Marshall, J. S., Warrington, R., Watson, W., & Kim, H. L. (2018). An introduction to immunology and immunopathology. *Allergy Asthma Clin Immunol*, *14*(Suppl 2), 49. https://doi.org/10.1186/s13223-018-0278-1

Mauer, J., Denson, J. L., & Bruning, J. C. (2015). Versatile functions for IL-6 in metabolism and cancer. *Trends Immunol*, *36*(2), 92-101. https://doi.org/10.1016/j.it.2014.12.008

McGinnis, S., & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*, *32*(Web Server issue), W20-25. https://doi.org/10.1093/nar/gkh435

McSparron, H., Blythe, M. J., Zygouri, C., Doytchinova, I. A., & Flower, D. R. (2003). JenPep: a novel computational information resource for immunobiology and vaccinology. *J Chem Inf Comput Sci*, *43*(4), 1276-1287. https://doi.org/10.1021/ci030461e

Mehta, S., Shelling, A., Muthukaruppan, A., Lasham, A., Blenkiron, C., Laking, G., & Print, C. (2010). Predictive and prognostic molecular markers for cancer medicine. *Ther Adv Med Oncol*, *2*(2), 125-148. https://doi.org/10.1177/1758834009360519

Mei, S., Li, F., Xiang, D., Ayala, R., Faridi, P., Webb, G. I., Illing, P. T., Rossjohn, J., Akutsu, T., Croft, N. P., Purcell, A. W., & Song, J. (2021). Anthem: a user customised tool for fast and accurate prediction of binding between peptides and HLA class I molecules. *Brief Bioinform*, *22*(5). https://doi.org/10.1093/bib/bbaa415

Mendez, R., Aptsiauri, N., Del Campo, A., Maleno, I., Cabrera, T., Ruiz-Cabello, F., Garrido, F., & Garcia-Lora, A. (2009). HLA and melanoma: multiple alterations in HLA class I and II expression in human melanoma cell lines from ESTDAB cell bank. *Cancer Immunol Immunother*, *58*(9), 1507-1515. https://doi.org/10.1007/s00262-009-0701-z

Menegatti, S., Bianchi, E., & Rogge, L. (2019). Anti-TNF Therapy in Spondyloarthritis and Related Diseases, Impact on the Immune System and Prediction of Treatment Responses. *Front Immunol*, *10*, 382. https://doi.org/10.3389/fimmu.2019.00382

Meydan, C., Otu, H. H., & Sezerman, O. U. (2013). Prediction of peptides binding to MHC class I and II alleles by temporal motif mining. *BMC Bioinformatics*, *14 Suppl 2*, S13. https://doi.org/10.1186/1471-2105-14-S2-S13

Mojic, M., Takeda, K., & Hayakawa, Y. (2017). The Dark Side of IFN-gamma: Its Role in Promoting Cancer Immunoevasion. *Int J Mol Sci*, *19*(1). https://doi.org/10.3390/ijms19010089

Montfort, A., Colacios, C., Levade, T., Andrieu-Abadie, N., Meyer, N., & Segui, B. (2019). The TNF Paradox in Cancer Progression and Immunotherapy. *Front Immunol*, *10*, 1818. https://doi.org/10.3389/fimmu.2019.01818

Mosaad, Y. M. (2015). Clinical Role of Human Leukocyte Antigen in Health and Disease. *Scand J Immunol*, *82*(4), 283-306. https://doi.org/10.1111/sji.12329

Muller-Werdan, U., Buerke, M., Ebelt, H., Heinroth, K. M., Herklotz, A., Loppnow, H., Russ, M., Schlegel, F., Schlitt, A., Schmidt, H. B., Soffker, G., & Werdan, K. (2006). Septic cardiomyopathy - A not yet discovered cardiomyopathy? *Exp Clin Cardiol*, *11*(3), 226-236. https://www.ncbi.nlm.nih.gov/pubmed/18651035

Murdaca, G., Contini, P., Negrini, S., Ciprandi, G., & Puppo, F. (2016). Immunoregulatory Role of HLA-G in Allergic Diseases. *J Immunol Res*, *2016*, 6865758. https://doi.org/10.1155/2016/6865758

Nagpal, G., Usmani, S. S., Dhanda, S. K., Kaur, H., Singh, S., Sharma, M., & Raghava, G. P. (2017). Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. *Sci Rep*, *7*, 42851. https://doi.org/10.1038/srep42851

Naranbhai, V., Viard, M., Dean, M., Groha, S., Braun, D. A., Labaki, C., Shukla, S. A., Yuki, Y., Shah, P., Chin, K., Wind-Rotolo, M., Mu, X. J., Robbins, P. B., Gusev, A., Choueiri, T. K., Gulley, J. L., & Carrington, M. (2022). HLA-A*03 and response to immune checkpoint blockade in cancer: an epidemiological biomarker study. *Lancet Oncol*, *23*(1), 172-184. https://doi.org/10.1016/S1470-2045(21)00582-9

National Genomics Data Center, M., & Partners. (2020). Database Resources of the National Genomics Data Center in 2020. *Nucleic Acids Res*, *48*(D1), D24-D33. https://doi.org/10.1093/nar/gkz913

Navarro-Gonzalez, J. F., & Mora-Fernandez, C. (2008). The role of inflammatory cytokines in diabetic nephropathy. *J Am Soc Nephrol*, *19*(3), 433-442. https://doi.org/10.1681/ASN.2007091048

Nestle, F. O., Burg, G., Fah, J., Wrone-Smith, T., & Nickoloff, B. J. (1997). Human sunlight-induced basal-cell-carcinoma-associated dendritic cells are deficient in T cell co-stimulatory molecules and are impaired as antigen-presenting cells. *Am J Pathol*, *150*(2), 641-651. https://www.ncbi.nlm.nih.gov/pubmed/9033277

Nicholson, L. B. (2016). The immune system. *Essays Biochem*, *60*(3), 275-301. https://doi.org/10.1042/EBC20160017

Nielsen, M., Lundegaard, C., & Lund, O. (2007). Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics*, *8*, 238. https://doi.org/10.1186/1471-2105-8-238

O'Donnell, T. J., Rubinsteyn, A., & Laserson, U. (2020). MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell Syst*, *11*(1), 42-48 e47. https://doi.org/10.1016/j.cels.2020.06.010

Old, L. J. (1988). Tumor necrosis factor. *Sci Am*, *258*(5), 59-60, 69-75. https://doi.org/10.1038/scientificamerican0588-59

Ossio, R., Roldan-Marin, R., Martinez-Said, H., Adams, D. J., & Robles-Espinoza, C. D. (2017). Melanoma: a global perspective. *Nat Rev Cancer*, *17*(7), 393-394. https://doi.org/10.1038/nrc.2017.43

Padma, V. V. (2015). An overview of targeted cancer therapy. *Biomedicine (Taipei)*, *5*(4), 19. https://doi.org/10.7603/s40681-015-0019-4

Pande, A., Patiyal, S., Lathwal, A., Arora, C., Kaur, D., Dhall, A., Mishra, G., Kaur, H., Sharma, N., Jain, S., Usmani, S. S., Agrawal, P., Kumar, R., Kumar, V., & Raghava, G. P. S. (2019). Computing wide range of protein/peptide features from their sequence and structure. *BioRxiv*, 599126-599126. https://doi.org/10.1101/599126

Parameswaran, N., & Patial, S. (2010). Tumor necrosis factor-alpha signaling in macrophages. *Crit Rev Eukaryot Gene Expr*, *20*(2), 87-103. https://doi.org/10.1615/critreveukargeneexpr.v20.i2.10

Parkin, J., & Cohen, B. (2001). An overview of the immune system. *Lancet*, *357*(9270), 1777-1789. https://doi.org/10.1016/S0140-6736(00)04904-7

Pasparakis, M., & Vandenabeele, P. (2015). Necroptosis and its role in inflammation. *Nature*, *517*(7534), 311-320. https://doi.org/10.1038/nature14191

Patiyal, S., Agrawal, P., Kumar, V., Dhall, A., Kumar, R., Mishra, G., & Raghava, G. P. S. (2020). NAGbinder: An approach for identifying N-acetylglucosamine interacting residues of a protein from its primary sequence. *Protein Sci*, *29*(1), 201-210. https://doi.org/10.1002/pro.3761

Paul, S., & Lal, G. (2017). The Molecular Mechanism of Natural Killer Cells Function and Its Importance in Cancer Immunotherapy. *Front Immunol*, *8*, 1124. https://doi.org/10.3389/fimmu.2017.01124

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Pedroza, M., Le, T. T., Lewis, K., Karmouty-Quintana, H., To, S., George, A. T., Blackburn, M. R., Tweardy, D. J., & Agarwal, S. K. (2016). STAT-3 contributes to pulmonary fibrosis through epithelial injury and fibroblast-myofibroblast differentiation. *FASEB J*, *30*(1), 129-140. https://doi.org/10.1096/fj.15-273953

Peng, Y., Xiao, J., Li, W., Li, S., Xie, B., He, J., & Liu, C. (2021). Prognostic and Clinicopathological Value of Human Leukocyte Antigen G in Gastrointestinal Cancers: A Meta-Analysis. *Front Oncol*, *11*, 642902. https://doi.org/10.3389/fonc.2021.642902

Peyrin-Biroulet, L. (2010). Anti-TNF therapy in inflammatory bowel diseases: a huge review. *Minerva Gastroenterol Dietol*, *56*(2), 233-243. https://www.ncbi.nlm.nih.gov/pubmed/20485259

Pietra, G., Romagnani, C., Mazzarino, P., Falco, M., Millo, E., Moretta, A., Moretta, L., & Mingari, M. C. (2003). HLA-E-restricted recognition of cytomegalovirus-derived peptides by human

CD8+ cytolytic T lymphocytes. *Proc Natl Acad Sci U S A*, *100*(19), 10896-10901. https://doi.org/10.1073/pnas.1834449100

Plasencia, C., Pascual-Salcedo, D., Garcia-Carazo, S., Lojo, L., Nuno, L., Villalba, A., Peiteado, D., Arribas, F., Diez, J., Lopez-Casla, M. T., Martin-Mola, E., & Balsa, A. (2013). The immunogenicity to the first anti-TNF therapy determines the outcome of switching to a second anti-TNF therapy in spondyloarthritis patients. *Arthritis Res Ther*, *15*(4), R79. https://doi.org/10.1186/ar4258

Prager, I., & Watzl, C. (2019). Mechanisms of natural killer cell-mediated cellular cytotoxicity. *J Leukoc Biol*, *105*(6), 1319-1329. https://doi.org/10.1002/JLB.MR0718-269R

Prakash, O., Gill, J., & Farr, G. (2002). Immune disorders and susceptibility to neoplasms. *Ochsner J*, *4*(2), 107-111. https://www.ncbi.nlm.nih.gov/pubmed/22822327

Pucci, C., Martinelli, C., & Ciofani, G. (2019). Innovative approaches for cancer treatment: current perspectives and new challenges. *Ecancermedicalscience*, *13*, 961. https://doi.org/10.3332/ecancer.2019.961

R, V., S, M., Ja, O., Sk, D., S, M., Jr, C., Dk, W., A, S., & B, P. (2019). The Immune Epitope Database (IEDB): 2018 update. *Nucleic acids research*, *47*(D1), D339-D343. https://doi.org/10.1093/NAR/GKY1006

Reljic, R. (2007). IFN-gamma therapy of tuberculosis and related infections. *J Interferon Cytokine Res*, *27*(5), 353-364. https://doi.org/10.1089/jir.2006.0103

Remy, K. E., Mazer, M., Striker, D. A., Ellebedy, A. H., Walton, A. H., Unsinger, J., Blood, T. M., Mudd, P. A., Yi, D. J., Mannion, D. A., Osborne, D. F., Martin, R. S., Anand, N. J., Bosanquet, J. P., Blood, J., Drewry, A. M., Caldwell, C. C., Turnbull, I. R., Brakenridge, S. C., Moldwawer, L. L., & Hotchkiss, R. S. (2020). Severe immunosuppression and not a cytokine storm characterizes COVID-19 infections. *JCI Insight*, *5*(17). https://doi.org/10.1172/jci.insight.140329

Restifo, N. P., Minev, B. R., Taggarse, A. S., McFarland, B. J., Wang, M., & Irvine, K. R. (1994). Enhancing the recognition of tumour associated antigens. *Folia Biol (Praha)*, *40*(1-2), 74-88. https://www.ncbi.nlm.nih.gov/pubmed/7958066

Reynisson, B., Alvarez, B., Paul, S., Peters, B., & Nielsen, M. (2020). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res*, *48*(W1), W449-W454. https://doi.org/10.1093/nar/gkaa379

Riaz, N., Havel, J. J., Makarov, V., Desrichard, A., Urba, W. J., Sims, J. S., Hodi, F. S., Martin-Algarra, S., Mandal, R., Sharfman, W. H., Bhatia, S., Hwu, W. J., Gajewski, T. F., Slingluff, C. L., Jr., Chowell, D., Kendall, S. M., Chang, H., Shah, R., Kuo, F., Morris, L. G. T., Sidhom, J. W., Schneck, J. P., Horak, C. E., Weinhold, N., & Chan, T. A. (2017). Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell*, *171*(4), 934-949 e916. https://doi.org/10.1016/j.cell.2017.09.028

Rini, B. I., Halabi, S., Rosenberg, J. E., Stadler, W. M., Vaena, D. A., Archer, L., Atkins, J. N., Picus, J., Czaykowski, P., Dutcher, J., & Small, E. J. (2010). Phase III trial of bevacizumab plus interferon alfa versus interferon alfa monotherapy in patients with metastatic renal cell carcinoma: final results of CALGB 90206. *J Clin Oncol*, *28*(13), 2137-2143. https://doi.org/10.1200/JCO.2009.26.5561

Rive, C. M., Bourke, J., & Phillips, E. J. (2013). Testing for drug hypersensitivity syndromes. *Clin Biochem Rev*, *34*(1), 15-38. https://www.ncbi.nlm.nih.gov/pubmed/23592889

Rizzo, R., Trentini, A., Bortolotti, D., Manfrinato, M. C., Rotola, A., Castellazzi, M., Melchiorri, L., Di Luca, D., Dallocchio, F., Fainardi, E., & Bellini, T. (2013). Matrix metalloproteinase-2 (MMP-2) generates soluble HLA-G1 by cell surface proteolytic shedding. *Mol Cell Biochem*, *381*(1-2), 243-255. https://doi.org/10.1007/s11010-013-1708-5

Robinson, J., Barker, D. J., Georgiou, X., Cooper, M. A., Flicek, P., & Marsh, S. G. E. (2020). IPD-IMGT/HLA Database. *Nucleic Acids Res*, *48*(D1), D948-D955. https://doi.org/10.1093/nar/gkz950

Robinson, J., Soormally, A. R., Hayhurst, J. D., & Marsh, S. G. E. (2016). The IPD-IMGT/HLA Database - New developments in reporting HLA variation. *Hum Immunol*, *77*(3), 233-237. https://doi.org/10.1016/j.humimm.2016.01.020

Rokx, C., van der Ende, M. E., Verbon, A., & Rijnders, B. J. (2013). Peginterferon alfa-2a for AIDS-associated Kaposi sarcoma: experience with 10 patients. *Clin Infect Dis*, *57*(10), 1497-1499. https://doi.org/10.1093/cid/cit517

Romagnani, C., Pietra, G., Falco, M., Mazzarino, P., Moretta, L., & Mingari, M. C. (2004). HLA-E-restricted recognition of human cytomegalovirus by a subset of cytolytic T lymphocytes. *Hum Immunol*, *65*(5), 437-445. https://doi.org/10.1016/j.humimm.2004.02.001

Romagnani, C., Pietra, G., Falco, M., Millo, E., Mazzarino, P., Biassoni, R., Moretta, A., Moretta, L., & Mingari, M. C. (2002). Identification of HLA-E-specific alloreactive T lymphocytes: a cell subset that undergoes preferential expansion in mixed lymphocyte culture and displays a broad cytolytic activity against allogeneic cells. *Proc Natl Acad Sci U S A*, *99*(17), 11328-11333. https://doi.org/10.1073/pnas.172369799

Rouas-Freiss, N., Khalil-Daher, I., Riteau, B., Menier, C., Paul, P., Dausset, J., & Carosella, E. D. (1999). The immunotolerance role of HLA-G. *Semin Cancer Biol*, *9*(1), 3-12. https://doi.org/10.1006/scbi.1998.0103

Ruiz-Patino, A., Arrieta, O., Cardona, A. F., Martin, C., Raez, L. E., Zatarain-Barron, Z. L., Barron, F., Ricaurte, L., Bravo-Garzon, M. A., Mas, L., Corrales, L., Rojas, L., Lupinacci, L., Perazzo, F., Bas, C., Carranza, O., Puparelli, C., Rizzo, M., Ruiz, R., Rolfo, C., Archila, P., Rodriguez, J., Sotelo, C., Vargas, C., Carranza, H., Otero, J., Pino, L. E., Ortiz, C., Laguado, P., Rosell, R., & ClicaP. (2020). Immunotherapy at any line of treatment improves survival in patients with advanced metastatic non-small cell lung cancer (NSCLC) compared with chemotherapy (Quijote-CLICaP). *Thorac Cancer*, *11*(2), 353-361. https://doi.org/10.1111/1759-7714.13272

Sabapathy, K., & Nam, S. Y. (2008). Defective MHC class I antigen surface expression promotes cellular survival through elevated ER stress and modulation of p53 function. *Cell Death Differ*, *15*(9), 1364-1374. https://doi.org/10.1038/cdd.2008.55

Sabbagh, A., Sonon, P., Sadissou, I., Mendes-Junior, C. T., Garcia, A., Donadi, E. A., & Courtin, D. (2018). The role of HLA-G in parasitic diseases. *HLA*, *91*(4), 255-270. https://doi.org/10.1111/tan.13196

Sabbatino, F., Liguori, L., Polcaro, G., Salvato, I., Caramori, G., Salzano, F. A., Casolaro, V., Stellato, C., Col, J. D., & Pepe, S. (2020). Role of Human Leukocyte Antigen System as A Predictive Biomarker for Checkpoint-Based Immunotherapy in Cancer Patients. *Int J Mol Sci*, *21*(19). https://doi.org/10.3390/ijms21197295

Saha, S., & Raghava, G. P. (2006). Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins*, *65*(1), 40-48. https://doi.org/10.1002/prot.21078

Saklatvala, J., Davis, W., & Guesdon, F. (1996). Interleukin 1 (IL1) and tumour necrosis factor (TNF) signal transduction. *Philos Trans R Soc Lond B Biol Sci*, *351*(1336), 151-157. https://doi.org/10.1098/rstb.1996.0011

Santa Cruz, A., Mendes-Frias, A., Oliveira, A. I., Dias, L., Matos, A. R., Carvalho, A., Capela, C., Pedrosa, J., Castro, A. G., & Silvestre, R. (2021). Interleukin-6 Is a Biomarker for the Development of Fatal Severe Acute Respiratory Syndrome Coronavirus 2 Pneumonia. *Front Immunol*, *12*, 613422. https://doi.org/10.3389/fimmu.2021.613422

Schlessinger, A., Ofran, Y., Yachdav, G., & Rost, B. (2006). Epitome: database of structure-inferred antigenic epitopes. *Nucleic Acids Res*, *34*(Database issue), D777-780. https://doi.org/10.1093/nar/gkj053

Schmidt, C. M., & Orr, H. T. (1993). Maternal/fetal interactions: the role of the MHC class I molecule HLA-G. *Crit Rev Immunol*, *13*(3-4), 207-224. https://www.ncbi.nlm.nih.gov/pubmed/8110376

Schoenborn, J. R., & Wilson, C. B. (2007). Regulation of interferon-gamma during innate and adaptive immune responses. *Adv Immunol*, *96*, 41-101. https://doi.org/10.1016/S0065-2776(07)96002-2

Schroder, K., Hertzog, P. J., Ravasi, T., & Hume, D. A. (2004). Interferon-gamma: an overview of signals, mechanisms and functions. *J Leukoc Biol*, *75*(2), 163-189. https://doi.org/10.1189/jlb.0603252

Sethi, J. K., & Hotamisligil, G. S. (2021). Metabolic Messengers: tumour necrosis factor. *Nat Metab*, *3*(10), 1302-1312. https://doi.org/10.1038/s42255-021-00470-z

Shao, S., He, F., Yang, Y., Yuan, G., Zhang, M., & Yu, X. (2012). Th17 cells in type 1 diabetes. *Cell Immunol*, *280*(1), 16-21. https://doi.org/10.1016/j.cellimm.2012.11.001

Shao, X. M., Bhattacharya, R., Huang, J., Sivakumar, I. K. A., Tokheim, C., Zheng, L., Hirsch, D., Kaminow, B., Omdahl, A., Bonsack, M., Riemer, A. B., Velculescu, V. E., Anagnostou, V., Pagel, K. A., & Karchin, R. (2020). High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets. *Cancer Immunol Res*, *8*(3), 396-408. https://doi.org/10.1158/2326-6066.CIR-19-0464

Shen, J., Xiao, Z., Zhao, Q., Li, M., Wu, X., Zhang, L., Hu, W., & Cho, C. H. (2018). Anti-cancer therapy with TNFalpha and IFNgamma: A comprehensive review. *Cell Prolif*, *51*(4), e12441. https://doi.org/10.1111/cpr.12441

Shih Ie, M. (2007). Application of human leukocyte antigen-G expression in the diagnosis of human cancer. *Hum Immunol*, *68*(4), 272-276. https://doi.org/10.1016/j.humimm.2007.01.010

Shiina, T., Hosomichi, K., Inoko, H., & Kulski, J. K. (2009). The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet*, *54*(1), 15-39. https://doi.org/10.1038/jhg.2008.5

Shtrichman, R., & Samuel, C. E. (2001). The role of gamma interferon in antimicrobial immunity. *Curr Opin Microbiol*, *4*(3), 251-259. https://doi.org/10.1016/s1369-5274(00)00199-5

Shugay, M., Bagaev, D. V., Zvyagin, I. V., Vroomans, R. M., Crawford, J. C., Dolton, G., Komech, E. A., Sycheva, A. L., Koneva, A. E., Egorov, E. S., Eliseev, A. V., Van Dyk, E., Dash, P., Attaf, M., Rius, C., Ladell, K., McLaren, J. E., Matthews, K. K., Clemens, E. B., Douek, D. C., Luciani, F., van Baarle, D., Kedzierska, K., Kesmir, C., Thomas, P. G., Price, D. A., Sewell, A. K., & Chudakov, D. M. (2018). VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res*, *46*(D1), D419-D427. https://doi.org/10.1093/nar/gkx760

Sidney, J., Peters, B., & Sette, A. (2020). Epitope prediction and identification- adaptive T cell responses in humans. *Semin Immunol*, *50*, 101418. https://doi.org/10.1016/j.smim.2020.101418

Singh, H., & Raghava, G. P. (2001). ProPred: prediction of HLA-DR binding sites. *Bioinformatics*, *17*(12), 1236-1237. https://doi.org/10.1093/bioinformatics/17.12.1236

Singh, H., & Raghava, G. P. (2003). ProPred1: prediction of promiscuous MHC Class-I binding sites. *Bioinformatics*, *19*(8), 1009-1014. https://doi.org/10.1093/bioinformatics/btg108

Singh, H., Singh, S., Singla, D., Agarwal, S. M., & Raghava, G. P. (2015). QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest. *Biol Direct*, *10*, 10. https://doi.org/10.1186/s13062-015-0046-9

Singla, D., Anurag, M., Dash, D., & Raghava, G. P. (2011). A web server for predicting inhibitors against bacterial target GlmU protein. *BMC Pharmacol*, *11*, 5. https://doi.org/10.1186/1471-2210-11-5

Slingluff, C. L., Jr. (2011). The present and future of peptide vaccines for cancer: single or multiple, long or short, alone or in combination? *Cancer J*, *17*(5), 343-350. https://doi.org/10.1097/PPO.0b013e318233e5b2

Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J. M., Desrichard, A., Walsh, L. A., Postow, M. A., Wong, P., Ho, T. S., Hollmann, T. J., Bruggeman, C., Kannan, K., Li, Y., Elipenahli, C., Liu, C., Harbison, C. T., Wang, L., Ribas, A., Wolchok, J. D., & Chan, T. A. (2014). Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N Engl J Med*, *371*(23), 2189-2199. https://doi.org/10.1056/NEJMoa1406498

Soura, E., Eliades, P. J., Shannon, K., Stratigos, A. J., & Tsao, H. (2016). Hereditary melanoma: Update on syndromes and management: Genetics of familial atypical multiple mole melanoma syndrome. *J Am Acad Dermatol*, *74*(3), 395-407; quiz 408-310. https://doi.org/10.1016/j.jaad.2015.08.038

Steele, N., Anthony, A., Saunders, M., Esmarck, B., Ehrnrooth, E., Kristjansen, P. E., Nihlen, A., Hansen, L. T., & Cassidy, J. (2012). A phase 1 trial of recombinant human IL-21 in combination with cetuximab in patients with metastatic colorectal cancer. *Br J Cancer*, *106*(5), 793-798. https://doi.org/10.1038/bjc.2011.599

Stenvinkel, P., Ketteler, M., Johnson, R. J., Lindholm, B., Pecoits-Filho, R., Riella, M., Heimburger, O., Cederholm, T., & Girndt, M. (2005). IL-10, IL-6, and TNF-alpha: central factors in the altered cytokine network of uremia--the good, the bad, and the ugly. *Kidney Int*, *67*(4), 1216-1233. https://doi.org/10.1111/j.1523-1755.2005.00200.x

Stranzl, T., Larsen, M. V., Lundegaard, C., & Nielsen, M. (2010). NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics*, *62*(6), 357-368. https://doi.org/10.1007/s00251-010-0441-4

Su, H., Lei, C. T., & Zhang, C. (2017). Interleukin-6 Signaling Pathway and Its Role in Kidney Disease: An Update. *Front Immunol*, *8*, 405. https://doi.org/10.3389/fimmu.2017.00405

Sun, Y., Li, F., Sonnemann, H., Jackson, K. R., Talukder, A. H., Katailiha, A. S., & Lizee, G. (2021). Evolution of CD8(+) T Cell Receptor (TCR) Engineered Therapies for the Treatment of Cancer. *Cells*, *10*(9). https://doi.org/10.3390/cells10092379

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*, *71*(3), 209-249. https://doi.org/10.3322/caac.21660

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci*, *43*(6), 1947-1958. https://doi.org/10.1021/ci034160g

Szolek, A., Schubert, B., Mohr, C., Sturm, M., Feldhahn, M., & Kohlbacher, O. (2014). OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*, *30*(23), 3310-3316. https://doi.org/10.1093/bioinformatics/btu548

Tagliamonte, M., Petrizzo, A., Tornesello, M. L., Buonaguro, F. M., & Buonaguro, L. (2014). Antigen-specific vaccines for cancer treatment. *Hum Vaccin Immunother*, *10*(11), 3332-3346. https://doi.org/10.4161/21645515.2014.973317

Tanaka, T., Narazaki, M., & Kishimoto, T. (2014). IL-6 in inflammation, immunity, and disease. *Cold Spring Harb Perspect Biol*, *6*(10), a016295. https://doi.org/10.1101/cshperspect.a016295

Tarhini, A. A., Cherian, J., Moschos, S. J., Tawbi, H. A., Shuai, Y., Gooding, W. E., Sander, C., & Kirkwood, J. M. (2012). Safety and efficacy of combination immunotherapy with interferon alfa-2b and tremelimumab in patients with stage IV melanoma. *J Clin Oncol*, *30*(3), 322-328. https://doi.org/10.1200/JCO.2011.37.5394

Tau, G., & Rothman, P. (1999). Biologic functions of the IFN-gamma receptors. *Allergy*, *54*(12), 1233-1251. https://doi.org/10.1034/j.1398-9995.1999.00099.x

Tavasolian, F., Rashidi, M., Hatam, G. R., Jeddi, M., Hosseini, A. Z., Mosawi, S. H., Abdollahi, E., & Inman, R. D. (2020). HLA, Immune Response, and Susceptibility to COVID-19. *Front Immunol*, *11*, 601886. https://doi.org/10.3389/fimmu.2020.601886

Thomas, T. L. (2016). Cancer Prevention: HPV Vaccination. *Semin Oncol Nurs*, *32*(3), 273-280. https://doi.org/10.1016/j.soncn.2016.05.007

Tomczak, K., Czerwinska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)*, *19*(1A), A68-77. https://doi.org/10.5114/wo.2014.47136

Tronik-Le Roux, D., Renard, J., Verine, J., Renault, V., Tubacher, E., LeMaoult, J., Rouas-Freiss, N., Deleuze, J. F., Desgrandschamps, F., & Carosella, E. D. (2017). Novel landscape of HLA-G isoforms expressed in clear cell renal cell carcinoma patients. *Mol Oncol*, *11*(11), 1561-1578. https://doi.org/10.1002/1878-0261.12119

Tung, C. W., & Ho, S. Y. (2007). POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. *Bioinformatics*, *23*(8), 942-949. https://doi.org/10.1093/bioinformatics/btm061

Ujiie, H., Tomida, M., Akiyama, H., Nakajima, Y., Okada, D., Yoshino, N., Takiguchi, Y., & Tanzawa, H. (2012). Serum hepatocyte growth factor and interleukin-6 are effective prognostic markers for non-small cell lung cancer. *Anticancer Res*, *32*(8), 3251-3258. https://www.ncbi.nlm.nih.gov/pubmed/22843899

Uzhachenko, R. V., & Shanker, A. (2019). CD8(+) T Lymphocyte and NK Cell Network: Circuitry in the Cytotoxic Domain of Immunity. *Front Immunol*, *10*, 1906. https://doi.org/10.3389/fimmu.2019.01906

Van Allen, E. M., Miao, D., Schilling, B., Shukla, S. A., Blank, C., Zimmer, L., Sucker, A., Hillen, U., Foppen, M. H. G., Goldinger, S. M., Utikal, J., Hassel, J. C., Weide, B., Kaehler, K. C., Loquai, C., Mohr, P., Gutzmer, R., Dummer, R., Gabriel, S., Wu, C. J., Schadendorf, D., & Garraway, L. A. (2015). Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science*, *350*(6257), 207-211. https://doi.org/10.1126/science.aad0095

Vang, Y. S., & Xie, X. (2017). HLA class I binding prediction via convolutional neural networks. *Bioinformatics*, *33*(17), 2658-2665. https://doi.org/10.1093/bioinformatics/btx264

Velazquez-Salinas, L., Verdugo-Rodriguez, A., Rodriguez, L. L., & Borca, M. V. (2019). The Role of Interleukin 6 During Viral Infections. *Front Microbiol*, *10*, 1057. https://doi.org/10.3389/fmicb.2019.01057

Volkovova, K., Bilanicova, D., Bartonova, A., Letasiova, S., & Dusinska, M. (2012). Associations between environmental factors and incidence of cutaneous melanoma. Review. *Environ Health*, *11 Suppl 1*, S12. https://doi.org/10.1186/1476-069X-11-S1-S12

Waldman, A. D., Fritz, J. M., & Lenardo, M. J. (2020). A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nat Rev Immunol*, *20*(11), 651-668. https://doi.org/10.1038/s41577-020-0306-5

Waldmann, T. A. (2018). Cytokines in Cancer Immunotherapy. *Cold Spring Harb Perspect Biol*, *10*(12). https://doi.org/10.1101/cshperspect.a028472

Wang, B., Song, N., Yu, T., Zhou, L., Zhang, H., Duan, L., He, W., Zhu, Y., Bai, Y., & Zhu, M. (2014). Expression of tumor necrosis factor-alpha-mediated genes predicts recurrence-free survival in lung cancer. *PLoS One*, *9*(12), e115945. https://doi.org/10.1371/journal.pone.0115945

Wang, D., Yang, L., Zhang, P., LaBaer, J., Hermjakob, H., Li, D., & Yu, X. (2017). AAgAtlas 1.0: a human autoantigen database. *Nucleic Acids Res*, *45*(D1), D769-D776. https://doi.org/10.1093/nar/gkw946

Wang, X., Zhao, H., Xu, Q., Jin, W., Liu, C., Zhang, H., Huang, Z., Zhang, X., Zhang, Y., Xin, D., Simpson, A. J., Old, L. J., Na, Y., Zhao, Y., & Chen, W. (2006). HPtaa database-potential target genes for clinical diagnosis and immunotherapy of human carcinoma. *Nucleic Acids Res*, *34*(Database issue), D607-612. https://doi.org/10.1093/nar/gkj082

Wang, Y., Zhang, H., Liu, C., Wang, Z., Wu, W., Zhang, N., Zhang, L., Hu, J., Luo, P., Zhang, J., Liu, Z., Peng, Y., Liu, Z., Tang, L., & Cheng, Q. (2022). Immune checkpoint modulators in cancer

immunotherapy: recent advances and emerging concepts. *J Hematol Oncol*, *15*(1), 111. https://doi.org/10.1186/s13045-022-01325-0

Warren, R. L., Choe, G., Freeman, D. J., Castellarin, M., Munro, S., Moore, R., & Holt, R. A. (2012). Derivation of HLA types from shotgun sequence datasets. *Genome Med*, *4*(12), 95. https://doi.org/10.1186/gm396

Weerasinghe, P., Garcia, G. E., Zhu, Q., Yuan, P., Feng, L., Mao, L., & Jing, N. (2007). Inhibition of Stat3 activation and tumor growth suppression of non-small cell lung cancer by G-quartet oligonucleotides. *Int J Oncol*, *31*(1), 129-136. https://www.ncbi.nlm.nih.gov/pubmed/17549413

Wieczorek, M., Abualrous, E. T., Sticht, J., Alvaro-Benito, M., Stolzenberg, S., Noe, F., & Freund, C. (2017). Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. *Front Immunol*, *8*, 292. https://doi.org/10.3389/fimmu.2017.00292

Wittig, M., Anmarkrud, J. A., Kassens, J. C., Koch, S., Forster, M., Ellinghaus, E., Hov, J. R., Sauer, S., Schimmler, M., Ziemann, M., Gorg, S., Jacob, F., Karlsen, T. H., & Franke, A. (2015). Development of a high-resolution NGS-based HLA-typing and analysis pipeline. *Nucleic Acids Res*, *43*(11), e70. https://doi.org/10.1093/nar/gkv184

Wu, H. H., Yan, X., Chen, Z., Du, G. W., Bai, X. J., Tuoheti, K., & Liu, T. Z. (2021). GNRH1 and LTB4R might be novel immune-related prognostic biomarkers in clear cell renal cell carcinoma (ccRCC). *Cancer Cell Int*, *21*(1), 354. https://doi.org/10.1186/s12935-021-02052-1

Wu, J., Wang, W., Zhang, J., Zhou, B., Zhao, W., Su, Z., Gu, X., Wu, J., Zhou, Z., & Chen, S. (2019). DeepHLApan: A Deep Learning Approach for Neoantigen Prediction Considering Both HLA-Peptide Binding and Immunogenicity. *Front Immunol*, *10*, 2559. https://doi.org/10.3389/fimmu.2019.02559

Wu, R., Forget, M. A., Chacon, J., Bernatchez, C., Haymaker, C., Chen, J. Q., Hwu, P., & Radvanyi, L. G. (2012). Adoptive T-cell therapy using autologous tumor-infiltrating lymphocytes for metastatic melanoma: current status and future outlook. *Cancer J*, *18*(2), 160-175. https://doi.org/10.1097/PPO.0b013e31824d4465

Xie, C., Yeo, Z. X., Wong, M., Piper, J., Long, T., Kirkness, E. F., Biggs, W. H., Bloom, K., Spellman, S., Vierra-Green, C., Brady, C., Scheuermann, R. H., Telenti, A., Howard, S., Brewerton, S., Turpaz, Y., & Venter, J. C. (2017). Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. *Proc Natl Acad Sci U S A*, *114*(30), 8059-8064. https://doi.org/10.1073/pnas.1707945114

Xu, Y., Su, G. H., Ma, D., Xiao, Y., Shao, Z. M., & Jiang, Y. Z. (2021). Technological advances in cancer immunity: from immunogenomics to single-cell analysis and artificial intelligence. *Signal Transduct Target Ther*, *6*(1), 312. https://doi.org/10.1038/s41392-021-00729-7

Yang, B., Sayers, S., Xiang, Z., & He, Y. (2011). Protegen: a web-based protective antigen database and analysis system. *Nucleic Acids Res*, *39*(Database issue), D1073-1078. https://doi.org/10.1093/nar/gkq944

Yang, X. O., Panopoulos, A. D., Nurieva, R., Chang, S. H., Wang, D., Watowich, S. S., & Dong, C. (2007). STAT3 regulates cytokine-mediated generation of inflammatory helper T cells. *J Biol Chem*, *282*(13), 9358-9363. https://doi.org/10.1074/jbc.C600321200

Yap, C. W. (2011). PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*, *32*(7), 1466-1474. https://doi.org/10.1002/jcc.21707

Yarchoan, M., Johnson, B. A., 3rd, Lutz, E. R., Laheru, D. A., & Jaffee, E. M. (2017). Targeting neoantigens to augment antitumour immunity. *Nat Rev Cancer*, *17*(4), 209-222. https://doi.org/10.1038/nrc.2016.154

Yarmarkovich, M., Marshall, Q. F., Warrington, J. M., Premaratne, R., Farrel, A., Groff, D., Li, W., di Marco, M., Runbeck, E., Truong, H., Toor, J. S., Tripathi, S., Nguyen, S., Shen, H., Noel, T., Church, N. L., Weiner, A., Kendsersky, N., Martinez, D., Weisberg, R., Christie, M.,

Eisenlohr, L., Bosse, K. R., Dimitrov, D. S., Stevanovic, S., Sgourakis, N. G., Kiefel, B. R., & Maris, J. M. (2021). Cross-HLA targeting of intracellular oncoproteins with peptide-centric CARs. *Nature*, *599*(7885), 477-484. https://doi.org/10.1038/s41586-021-04061-6

Ye, Y., Wang, J., Xu, Y., Wang, Y., Pan, Y., Song, Q., Liu, X., & Wan, J. (2021). MATHLA: a robust framework for HLA-peptide binding prediction integrating bidirectional LSTM and multiple head attention mechanism. *BMC Bioinformatics*, *22*(1), 7. https://doi.org/10.1186/s12859-020-03946-z

YetkIn, M. F., & M, G. U. (2020). Efficacy and Tolerability of Interferon Gamma in Treatment of Friedreich's Ataxia: Retrospective Study. *Noro Psikiyatr Ars*, *57*(4), 270-273. https://doi.org/10.29399/npa.25047

You, K., Gu, H., Yuan, Z., & Xu, X. (2021). Tumor Necrosis Factor Alpha Signaling and Organogenesis. *Front Cell Dev Biol*, *9*, 727075. https://doi.org/10.3389/fcell.2021.727075

Younes, A., Pro, B., Robertson, M. J., Flinn, I. W., Romaguera, J. E., Hagemeister, F., Dang, N. H., Fiumara, P., Loyer, E. M., Cabanillas, F. F., McLaughlin, P. W., Rodriguez, M. A., & Samaniego, F. (2004). Phase II clinical trial of interleukin-12 in patients with relapsed and refractory non-Hodgkin's lymphoma and Hodgkin's disease. *Clin Cancer Res*, *10*(16), 5432-5438. https://doi.org/10.1158/1078-0432.CCR-04-0540

Zaidi, M. R., & Merlino, G. (2011). The two faces of interferon-gamma in cancer. *Clin Cancer Res*, *17*(19), 6118-6124. https://doi.org/10.1158/1078-0432.CCR-11-0482

Zamora, A. E., Crawford, J. C., & Thomas, P. G. (2018). Hitting the Target: How T Cells Detect and Eliminate Tumors. *J Immunol*, *200*(2), 392-399. https://doi.org/10.4049/jimmunol.1701413

Zarogoulidis, P., Yarmus, L., Darwiche, K., Walter, R., Huang, H., Li, Z., Zaric, B., Tsakiridis, K., & Zarogoulidis, K. (2013). Interleukin-6 cytokine: a multifunctional glycoprotein for cancer. *Immunome Res*, *9*(62), 16535. https://doi.org/10.1186/2090-5009-9-1

Zhang, G., Chitkushev, L., Olsen, L. R., Keskin, D. B., & Brusic, V. (2021). TANTIGEN 2.0: a knowledge base of tumor T cell antigens and epitopes. *BMC Bioinformatics*, *22*(Suppl 8), 40. https://doi.org/10.1186/s12859-021-03962-7

Zhang, G. L., DeLuca, D. S., Keskin, D. B., Chitkushev, L., Zlateva, T., Lund, O., Reinherz, E. L., & Brusic, V. (2011). MULTIPRED2: a computational system for large-scale identification of peptides predicted to bind to HLA supertypes and alleles. *J Immunol Methods*, *374*(1-2), 53-61. https://doi.org/10.1016/j.jim.2010.11.009

Zhang, M., Wang, X., Chen, X., Zhang, Q., & Hong, J. (2020). Novel Immune-Related Gene Signature for Risk Stratification and Prognosis of Survival in Lower-Grade Glioma. *Front Genet*, *11*, 363. https://doi.org/10.3389/fgene.2020.00363

Zhang, Y., Guan, X. Y., & Jiang, P. (2020). Cytokine and Chemokine Signals of T-Cell Exclusion in Tumors. *Front Immunol*, *11*, 594609. https://doi.org/10.3389/fimmu.2020.594609

Zhang, Z., Lu, M., Qin, Y., Gao, W., Tao, L., Su, W., & Zhong, J. (2021). Neoantigen: A New Breakthrough in Tumor Immunotherapy. *Front Immunol*, *12*, 672356. https://doi.org/10.3389/fimmu.2021.672356

Zhao, L., Zhang, M., & Cong, H. (2013). Advances in the study of HLA-restricted epitope vaccines. *Hum Vaccin Immunother*, *9*(12), 2566-2577. https://doi.org/10.4161/hv.26088

Zhou, F. (2009). Molecular mechanisms of IFN-gamma to up-regulate MHC class I antigen processing and presentation. *Int Rev Immunol*, *28*(3-4), 239-260. https://doi.org/10.1080/08830180902978120

Zhu, Y., Qiu, P., & Ji, Y. (2014). TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat Methods*, *11*(6), 599-600. https://doi.org/10.1038/nmeth.2956

Zidi, I. (2020). Puzzling out the COVID-19: Therapy targeting HLA-G and HLA-E. *Hum Immunol*, *81*(12), 697-701. https://doi.org/10.1016/j.humimm.2020.10.001

Zou, L., Ruan, F., Huang, M., Liang, L., Huang, H., Hong, Z., Yu, J., Kang, M., Song, Y., Xia, J., Guo, Q., Song, T., He, J., Yen, H. L., Peiris, M., & Wu, J. (2020). SARS-CoV-2 Viral Load in

Upper Respiratory Specimens of Infected Patients. *N Engl J Med*, *382*(12), 1177-1179. https://doi.org/10.1056/NEJMc2001737

Zou, S., Tong, Q., Liu, B., Huang, W., Tian, Y., & Fu, X. (2020). Targeting STAT3 in Cancer Immunotherapy. *Mol Cancer*, *19*(1), 145. https://doi.org/10.1186/s12943-020-01258-7