Development of Biomolecule-Based In-Silico Models for Cancer Diagnostics and Therapeutics.

By

Shipra Jain

(PhD18204)

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

Department of Computational Biology

Indraprastha Institute of Information Technology, Delhi

New Delhi – 110020

November, 2025

Development of Biomolecule-Based In-Silico Models for Cancer Diagnostics and Therapeutics.

By

Shipra Jain

(PhD18204)

A Thesis

Submitted in Partial Fulfilment of the Requirements for the Degree Of

Doctor of Philosophy

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

Department of Computational Biology

Indraprastha Institute of Information Technology, Delhi

New Delhi – 110020

November, 2025

Certificate

This is to certify that the thesis entitled "**Development of Biomolecule-Based In-Silico Models for Cancer Diagnostics and Therapeutics.**" being submitted by **Ms. Shipra Jain** 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.

November, 2025

Prof. Gajendra P. S. Raghava

Supervisor Name

Indraprastha Institute of Information Technology Delhi

New Delhi - 110020

Declaration

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

November, 2025

Student Name: **Shipra Jain**

Roll Number: **PhD18204**

Advisor Name: **Prof. Gajendra P. S. Raghava**

Indraprastha Institute of Information Technology Delhi

New Delhi 110020

Acknowledgements

I would like to express my profound gratitude to **Professor Gajendra P. S. Raghava** for his invaluable guidance, supervision, and encouragement throughout the course of this research. His exceptional expertise, insightful suggestions, and constant motivation have been instrumental in shaping the direction and quality of this work. The opportunity to work under his mentorship has been a privilege and an immensely enriching experience that has deepened my understanding of computational biology and research methodology.

I am sincerely thankful to **Professor Pankaj Jalote**, Founding Director of IIIT-Delhi, for his inspiration and encouragement to pursue the Ph.D. program alongside my professional responsibilities. I would also like to express my sincere appreciation to all the faculty members of the Department of Computational Biology, Indraprastha Institute of Information Technology Delhi (IIIT-Delhi), for their continuous support and valuable academic guidance. I am especially grateful to **Dr. Sriram** and **Dr. Jaspreet Kaur Dhanjal** for their consistent encouragement, thoughtful advice, and motivation throughout my research journey.

My heartfelt thanks are due to my lab members and research colleagues, especially **Sumeet Patiyal, Anjali Dhall, Neelam Sharma, and Ritu Tomer**, in the Computational Biology Group for their constructive feedback, intellectual engagement, and commitment to scientific exchange have made this research endeavour both productive and fulfilling.

I extend my gratitude to the technical and administrative staff of IIIT-Delhi for maintaining the research infrastructure and providing timely support in all operational and academic matters. I am particularly thankful to **Ms. Risha Dcruz, Ms. Nidhi Yadav, Ms. Priti Patel, Ms. Anshu Dureja, and Mr. Gursevak Singh** for their constant motivation, goodwill, and support as well-wishers throughout this journey.

Finally, I wish to express my heartfelt gratitude to my *parents, brother, and husband*, whose unconditional love, patience, and faith have been the true foundation of my journey. Their unwavering support and countless sacrifices have given me strength and courage to persevere through every challenge, inspiring me to pursue this work with dedication and confidence.

Shipra Jain

Abstract

Cancer remains a major global health challenge, necessitating advanced computational resources to support effective treatment and disease management. These resources can assist researchers in the early detection of cancer and the screening of therapeutic molecules. This study presents a suite of machine learning-based in-silico models and resources developed to facilitate cancer diagnosis and therapy. First, we developed a highly accurate classifier based on a ten-gene panel selected using propensity indices. This model reliably distinguishes prostate cancer from normal samples, offering a potential non-invasive alternative to conventional diagnostic procedures. In recent years, there has been a notable shift in drug discovery from small molecules to peptide-based therapies. To support this transition, we developed THPdb2, an updated repository of FDA-approved therapeutic peptides. Building on this, we introduced THPPred, a machine learning-based tool that predicts novel druggable peptides by analysing their physicochemical, molecular, and sequence features, thereby streamlining therapeutic screening. To further support peptide-based immunotherapy, we developed IL13Pred, a predictive model that identifies IL-13-inducing peptides, aiding in the design of cytokine-based cancer immunotherapies. Additionally, for small-molecule drug discovery, we created NfκBin, an in-silico model that predicts NF-κB inhibitors using molecular fingerprints, accelerating the identification of compounds targeting this key oncogenic pathway. Collectively, these integrated models and databases form a comprehensive computational pipeline for cancer biomarker discovery and therapeutic screening. By combining gene expression analysis, peptide-based drug prediction, and small-molecule screening, this framework advances personalized diagnostics and treatment strategies in oncology. Mostly resources developed in this study are publicly accessible via a web-based platform and GitHub.

Table of Content

List of Abbreviations

Abbreviation	Full Form
AID	Assay ID (PubChem BioAssay identifier)
AAC	Amino Acid Composition
AUC	Area Under the Receiver Operating Characteristic Curve
BAFF	B-cell Activating Factor
CD-HIT	Cluster Database at High Identity with Tolerance
CNN	Convolutional Neural Network
COVID-19	Coronavirus Disease 2019
CV	Cross-Validation
DC	Dendritic Cell
DNA	Deoxyribonucleic Acid
DT	Decision Tree
DPC	Dipeptide Composition
DSSP	Dictionary of Secondary Structure of Proteins
FDA	Food and Drug Administration
FP	False Positive
FN	False Negative
FPR	False Positive Rate
GEO	Gene Expression Omnibus
GO	Gene Ontology
HLA	Human Leukocyte Antigen
IFN-γ	Interferon-gamma
IEDB	Immune Epitope Database
IL	Interleukin
IL-4	Interleukin-4
IL-6	Interleukin-6
IL-10	Interleukin-10
IL-13	Interleukin-13
KNN	K-Nearest Neighbour
LR	Logistic Regression
MAE	Mean Absolute Error

MCC	Matthews Correlation Coefficient
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Squared Error
NB	Naïve Bayes
NCBI	National Center for Biotechnology Information
NF-κB	Nuclear Factor Kappa-light-chain-enhancer of Activated B Cells
NN	Neural Network
OpenBabel	Open Babel Chemical Format Converter
PCA	Principal Component Analysis
PDB	Protein Data Bank
PRAD	Prostate Adenocarcinoma
PSA	Prostate-Specific Antigen
PSSM	Position Specific Scoring Matrix
PubChem	Public Chemical Database (NCBI)
RF	Random Forest
ROC	Receiver Operating Characteristic
RNA	Ribonucleic Acid
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
Th1	T-helper Cell Type 1
Th2	T-helper Cell Type 2
TNF-α	Tumor Necrosis Factor-alpha
TME	Tumor Microenvironment
TP	True Positive
TPR	True Positive Rate
UniProt	Universal Protein Resource
XGB	Extreme Gradient Boosting
WHO	World Health Organization

List of genes and their description

Gene Symbol	Full Gene Name
BAFF (TNFSF13B)	Tumor Necrosis Factor Superfamily Member 13B
CD40	CD40 Molecule, TNF Receptor Superfamily Member 5
DD3 / PCA3	Prostate Cancer Antigen 3 (Non-protein coding RNA)
ERG	ETS-Related Gene
FOXA1	Forkhead Box A1
GSTP1	Glutathione S-Transferase Pi 1
IFNG	Interferon Gamma
IL13	Interleukin-13
IL13RA1	Interleukin-13 Receptor Subunit Alpha 1
IL13RA2	Interleukin-13 Receptor Subunit Alpha 2
IL4R	Interleukin-4 Receptor
KLK3	Kallikrein Related Peptidase 3
MAP3K14 (NIK)	Mitogen-Activated Protein Kinase Kinase Kinase 14
MED12	Mediator Complex Subunit 12
NFKB1	Nuclear Factor Kappa B Subunit 1
NFKB2	Nuclear Factor Kappa B Subunit 2
NFKBIA	NFKB Inhibitor Alpha (IκBα)
PTEN	Phosphatase and Tensin Homolog
RELA	RELA Proto-Oncogene, NF-κB Subunit
RELB	RELB Proto-Oncogene, NF-κB Subunit
SPOP	Speckle-Type BTB/POZ Protein
TNF	Tumor Necrosis Factor
TMPRSS2	Transmembrane Protease, Serine 2
TP53	Tumor Protein P53

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Background

The pathophysiology of cancer is a complex and heterogeneous biologically based disease that develops as a result of genetic changes, immune maladaptation, and cellular proliferation abnormalities[1,2]. Irrespective of the significant progress in molecular oncology, the realization of consistent early diagnosis and successful treatment intervention has been a continuing problem. This weakness is mainly due to inter and intra tumor heterogeneity which makes it difficult to find universal biomarkers and treatment targets[3–5]. Traditional experimental pipelines to discover biomarkers and develop therapeutics are time-consuming, costly and lack scalability, which restrict their applicability in large-scale translational studies[6,7].

Biomarkers are measurable biological surrogates that are linked to physiological or pathological conditions, and play a critical role in the detection, prognosis and stratification of cancer therapy[8–10]. Likewise, precision oncology is based on targeted therapeutic molecules, which allow tailoring the treatment choice to a patient. Nevertheless, experimental screening methods based on experimental screening alone are becoming less viable in the light of growing volumes of biological data and complicated pathophysiology[11,12].

The recent developments in computational biology have made in-silico approach based tools as essential instruments to tackle these issues. Specifically, machine learning-based methods have shown high potential in nonlinear association in high-dimensional biological data and predictive models with translational significance[13–15]. These technologies enable the unification of various modalities of biological data, such as transcriptomic profiles, peptide and protein sequences, and small-molecule chemical descriptors, and are useful to provide an in-depth modelling of cancer-related biological systems[16].

The recent rapid expansion of publicly available biological repositories, particularly due to high-throughput sequencing and large-scale immunological experiments, has emphasised the need of efficient and precise computational frameworks tools. TCGA, GEO, IEDB, UniProt, DrugBank and PubChem are databases have large amounts of experimentally verified data. To derive actionable knowledge out of such resources, it is necessary to have well-planned feature representation schemes, learning algorithms that are statistically sound, and the rigorous validation measures that guarantee reproducibility and clinical applicability[17–22].

Over this background, the current thesis focuses on the development of machine learning-driven computational pipelines in the discovery of cancer biomarkers and prediction of therapeutic molecules. This would create scalable and reproducible predictive models based on transcriptomic, peptide based and chemical compound data, which would help in ranking biologically meaningful candidates. The proposed methodologies are introduced as independent tools and web-based platforms and open-source software to facilitate the transparency and usability.

1.2 Origin of the Proposal

The conceptual foundation of this study lies in the fact that immune regulation and responses is a crucial stage of cancer initiation, progression, and response to treatment. Depending on the molecular and cellular context, immune signalling pathways may have tumor-suppressive or tumor-promoting actions. The main players of these mechanisms are transcriptional regulators like NF-kB and cytokines like IL-6, IL-4, IL-10, and IL-13 that in combination will determine the tumor microenvironment and disease outcome.

Also, genes expression based biomarkers and peptide-based therapeutic approaches have been prominent with their capability to be diagnosed early and targeted. Nevertheless, application of these biological understandings into clinically implementable tools is yet to be achieved due to the lack of automated, precise and scalable computing structures that can utilize heterogeneous biological data. These constraints led to the development of the current research study, which is aimed at closing the gap between generation of experimental data and clinical implementation with the help of systematic in-silico modelling. The thesis was therefore designed to develop, validate, and deploy machine learning-based applications to identify cancer biomarkers and screen therapeutic treatments on rigorously curated public data.

1.3 Objectives of the Thesis

The main aim of this thesis is to design and develop effective machine learning-based computational models for the identification of cancer-related biomarkers and therapeutic molecules. In this direction, the study aims at building predictive models to discover cancer biomarkers using gene expression microarray data through improved feature selection methods and optimized classification algorithms. Simultaneously, a manually curated and well-filtered library of therapeutic peptides and proteins is built, and based on that a predictive tool is developed that can effectively separate therapeutic sequences and non-therapeutic analogs.

This thesis also presents an in-silico model of IL-13-stimulating peptide prediction and methodically test its applicability to the cancer immunology and viral proteome settings. In addition to this, a virtual screening tool that uses machine learning algorithm to screen potential NF-kB inhibitors in large-scale small-molecule libraries. All models and tools developed are available as computational resources, such as web servers and open-source software, to achieve wide usability in the scientific community. These models are properly evaluated in terms of performance and reliability through rigorous cross-validation techniques, as independent test data and application-based case studies to guarantee robustness, generalization and translational applicability.

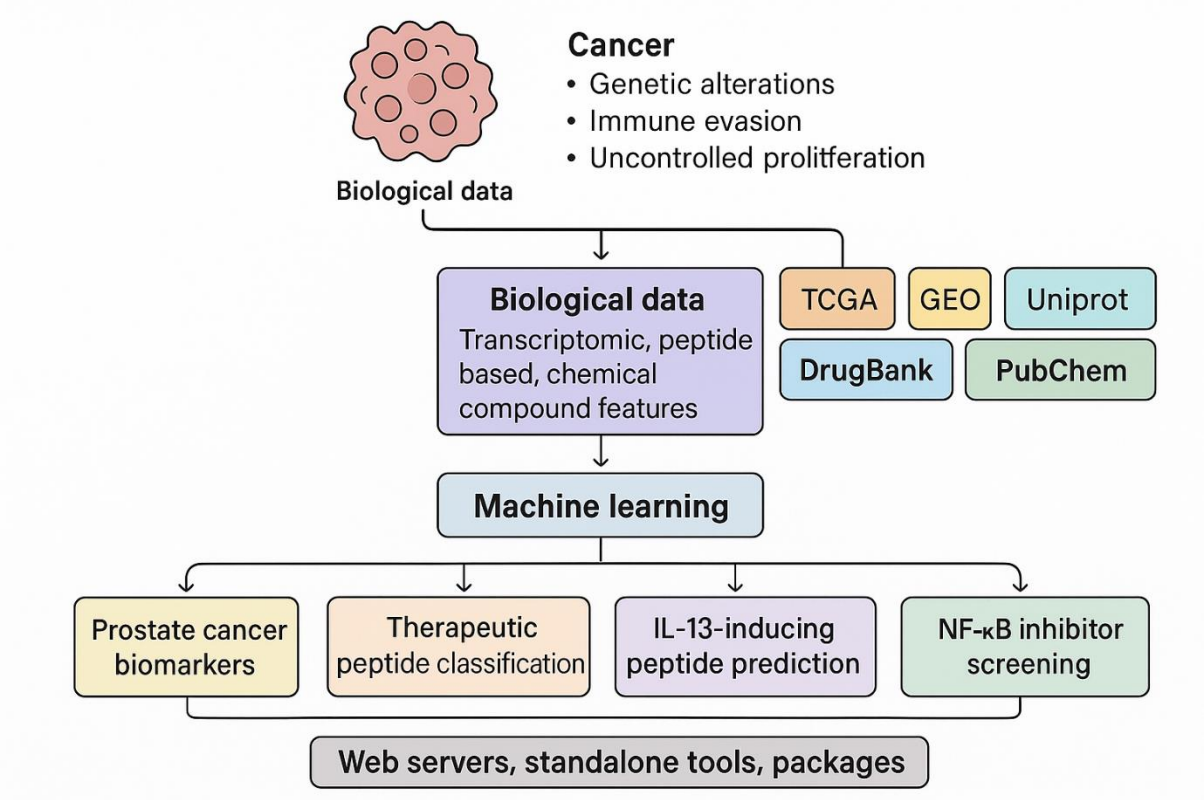


Figure 1.1: Computational tools developed for cancer diagnostics and therapeutics.

1.4 Organization of the Thesis

The thesis is structured into seven chapters, four of which are technical core research chapters dealing with complementary nature of the topics of cancer biomarker discovery and therapeutic prediction through in-silico methodologies. The brief description of the chapters are as follows:

- **Chapter 2** gives a dedicated literature review of the discovery of disease biomarkers, protein and peptide-based therapeutics, and small-molecule-based drug discovery and the importance of in-silico based machine learning methods.
- **Chapter 3** demonstrates a framework of machine learning-based biomarker discovery in prostate cancer based on the gene expression data in the TCGA-PRAD cohort. An innovative feature selection method based on propensity index is introduced, and a ten-gene biomarker panel is proposed that screen cancerous and normal samples with high accuracy.
- **Chapter 4** describes the ThPdb2, repository an updated and manually curated collection and comprehensive data of approved therapeutic peptides/ proteins by FDA. In addition to this, ThPPred a predictor model that is aimed at classifying therapeutic and non-therapeutic peptide sequences according to compositional and motif-based features is also introduced.
- **Chapter 5** highlights the IL13Pred tool, an in-silico predictor of IL-13-stimulating peptides developed utilizing experimentally tested immunological data. The usefulness of the model is evidenced in its effective usage on peptide screening of SARS-CoV-2 proteome.
- **Chapter 6** details the development of NFκBIn, a machine learning-based virtual screening framework for identifying NF-κB inhibitors from large-scale small-molecule bioassay data.
- **Chapter 7** deflects on the importance of the tools presented in this thesis. It further highlights the applicability and usability of the webservers and open source tools developed. Also, it leads to the concluding summary and the future directions to extend this research further.

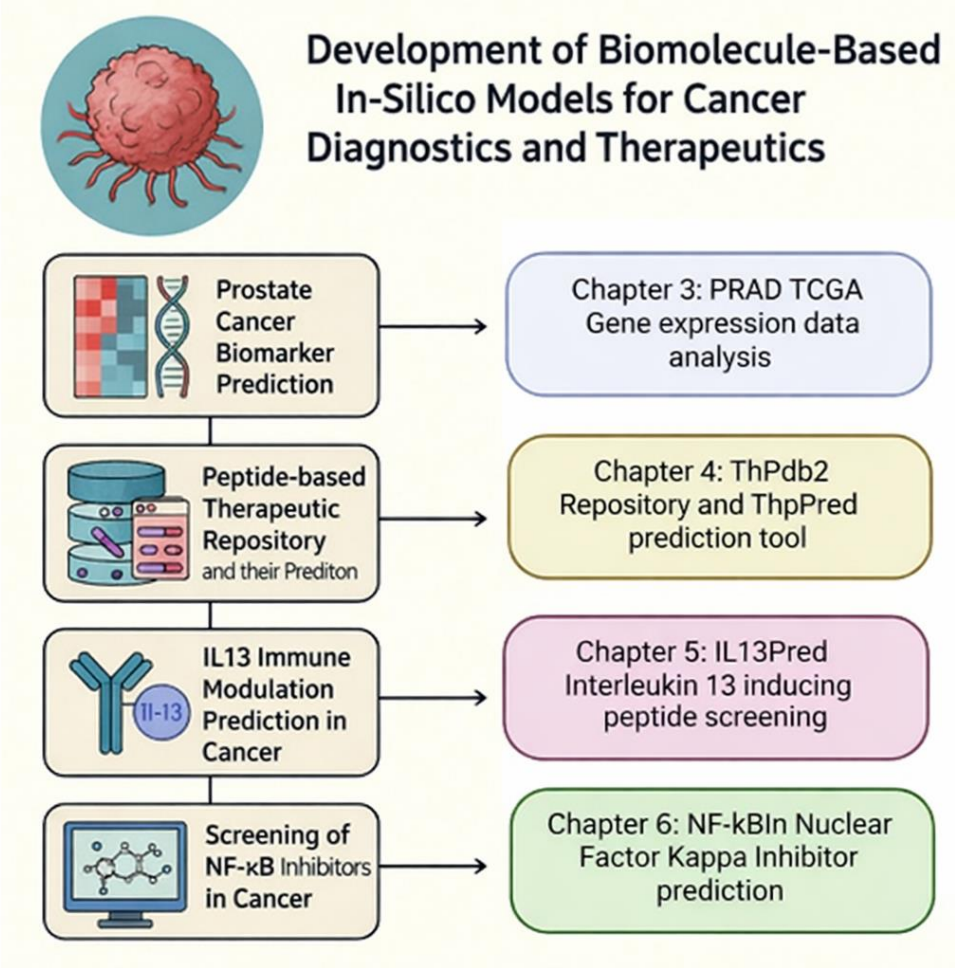


Figure 1.2: Illustration depicting overall architecture of this thesis.

In all chapters, the best practices in machine learning are adhered to, such as robust cross-validation, independent testing, a cautious choice of features (to avoid information leakage), and case study-based external validation (wherever applicable). Building on these computational tools and methods, several promising future research directions are envisioned. Future tools could target multiple immune-modulatory or cancer-related functions, and the incorporation of multi-omics data, e.g. epigenetic, proteomic and single-cell transcriptomics, to increase specificity of the model and its clinical applicability. It is also possible that future work would involve 3D structural modelling and candidate validation by using molecular docking, real-time clinical use models via patient-specific pipelines, and encourage wet-lab collaborations via experimental validation. Furthermore, the development of automated, explainable AI workflows will support transparent, scalable, and clinically actionable predictions.

Chapter 2
Review of Literature

2.1 Overview

Biological systems function are highly regulated molecular processes extending across biological layers, such as transcriptional regulation, protein expression, peptide-mediated signalling and small-molecule interactions[23,24]. Abnormalities in these biological layers can lead to the pathology of disease such as cancer by causing abnormal cellular behaviour, immune escape, and changes in intracellular communication[25,26].

Consequently, the phenotypes of diseases do not arise as a consequence of single molecular incidences, but rather as the resultant breakdown of interrelationships between molecular networks. The abnormal expression patterns, disrupted immune signalling through peptides and cytokines, and sustained activation of intracellular signalling pathways are some of the factors that define cancer progression, which are frequently targeted by therapeutic small molecules[27–29].

All these molecular dimensions, i.e., genes, peptides/proteins, and chemical compounds, provide a specific point of entry into diagnosis or therapeutic intervention[30–32]. Therefore, the study of any single dimension alone implies a one-sided account of the biology of the disease. Even though the experimental approaches have played a key role in determining disease-related genes, therapeutic peptides, and drug candidates, they have marked limited scalability as compared to the volume and complexity of the contemporary biological data. Large-scale transcriptomic experiments, large repositories of peptide sequences, and large data sets of chemical bioassays have tremendously increased the molecular space, and required new methods/ tools for analysis that are not limited to traditional experimental analysis[15,33].

In-silico and computational techniques, especially those that use machine learning, have thus become part and parcel of modern biomedical research[34]. Through these techniques, it is possible to explore high-dimensional data in a systematic way, to identify discriminative molecular features, and to make predictive models that can be generalized across heterogeneous biological contexts[35,36]. As a result, machine learning methods are heavily relied upon in gene expression-based disease biomarker explorations, therapeutic peptide screening, immune regulation-study and virtual screening of small molecules.

In this framework, the current thesis has a multi-domain style of computation which deals with the identification of disease biomarkers, protein and peptide-drug development, and small-molecule drug development in a coordinated way. The knowledge sources that are discussed in this chapter give the theoretical and methodological background of the subsequent sections

that will be presented in the following chapters and outline the unresolved challenges that drive the development of computational approaches developed in this work.

2.2 Disease Biomarker Discovery

The discovery of disease biomarkers is at the heart of the early diagnosis, prognosis evaluation, and individualized approaches to treatment and disease management[37,38]. Biomarkers are vital in early and personalized cancer research. They are commonly the results of molecular changes that initiate and propagate tumors such as gene expression, chromosomal rearrangements, and somatic mutations[8,39]. The targeted genetic events such as aberrant expression of prostate cancer associated genes and recurrent mutations in regulatory pathways have significantly advanced understanding of disease mechanisms[40–43].

Genetic biomarkers particularly gene expression based have been of significant relevance as they harbor the aggregate impact of genetic and epigenetic changes and dynamic disease states[44]. The transcriptomic profiling has facilitated high-throughput comparison of tumor and normal tissues in a larger patient cohorts[45]. Nevertheless, although large datasets of public expressions are available, there is still a gap in translating gene expression signatures to screening clinically useful biomarkers because of tumor heterogeneity, bias in datasets, and cross-platform variability[46]. To address these challenges, machine learning techniques have been used extensively to allow the data-driven discovery of informative sets of genes and model predictive systems[47,48]. The feature selection strategies are vital in the dimension's reduction and enhancement of interpretability especially in high-dimensional transcriptomic data[49]. Models trained on precisely curated datasets i.e. supervised classifiers have shown enhanced diagnostic performance over traditional techniques[50–52].

Table 2.1: Genomic and transcriptomic tools for disease biomarker identification.

Tool Name	Description
CancerLivER[53]	A database of gene expression and biomarker data of liver cancer
ProCanBio[54]	Predicts prostate cancer–associated genes using transcriptomic features.
CancerSPP[55]	Predicts skin cancer-associated secretory proteins using signal peptide features.
CancerCSP[56]	Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer
BBcancer[57]	Blood-based biomarker expression profiling for early cancer diagnosis

oncoPredict[58]	R package for predicting cancer patient drug response and biomarkers from cell line screening data
CancerLSP[59]	Prediction of early- and late-stage liver cancer using transcriptomic and epigenomic biomarkers
CancerTSP[60]	Classification of early- and late-stage thyroid carcinoma using RNA expression profiles
HCCPred[61]	Prediction and discrimination of hepatocellular carcinoma and normal liver samples

However, the majority of reported biomarker panels have low levels of validation, overfitting and are not reproducible in independent cohorts. These constraints highlight the necessity of new feature selection procedures and rigorous model validation approaches. This thesis has created a biomarker discovery strategy that directly tackles these issues with a combination of both innovative feature selection techniques and powerful machine learning classifiers as detailed in the subsequent chapter.

2.3 Protein and Peptide-Based Therapeutics

Proteins and peptides are necessary in cellular regulation, immune signalling and intercellular communication[62–64]. Cytokines and signalling mediators are among the peptides that play a major role in immune reaction and the disease progression[65,66]. Their specificity and structural heterogeneity predisposes them to a specific type of targeted therapy, especially in cancer and immune-related disorders[67–69]. The increasing count of FDA approved peptide and protein-based therapeutics indicates their precise clinical applicability[70–72]. However, the optimization and experimental discovery of the therapeutic peptides is limited by the complexity of biology and phenomena such as sequence variation, post-translational changes, and context-dependent activity[73,74]. These limitations necessitates computational methods/ tools that can effectively screen large peptide sequences data in bulk in a systematic manner[75,76].

In-silico tools developed deploying machine learning algorithms have proven their efficacy in predicting protein properties directly using sequential data. These models facilitate the in-silico screening of large peptide libraries and can identify the specific physicochemical property of peptides using the compositional, physicochemical, and motif information[77]. High quality manually curated datasets are very important to develop precise and with improved reproducible models.

Table 2.2: Computational tools for protein and peptide-based therapeutic discovery.

Tool Name	Description
THPdb[78]	Curated database of FDA-approved therapeutic proteins and peptides.
AntiCP[79]	Predicts anticancer peptides using sequence-based machine learning models.
AntiCP 2.0[80]	Improved anticancer peptide prediction with enhanced datasets.
ACPred[81]	Classifies anticancer peptides using machine learning-based features.
CancerPPD[82]	Repository and predictor for experimentally validated anticancer peptides.
CellPPD[83]	Predicts cell-penetrating therapeutic peptides.
HemoPI[84]	Identifies hemolytic peptides to support safe therapeutic design.
ToxinPred2[85]	Predicts toxic peptides to aid therapeutic peptide safety screening.
BepiPred3.0[86]	Predicts linear B-cell epitopes from protein sequences.
IL6pred[87]	Predicts IL-6 inducing peptides involved in immune modulation.
IL4pred[88]	Predicts IL-4 inducing peptides involved in immune modulation.
PEPred-Suite[89]	Generic framework for therapeutic peptide prediction and screening.

The creation of new repositories and platforms with integrated prediction tools is a solution to the major limitations of the past tools developed. By integrating curated knowledgebase with robust predictor models, candidate prioritization becomes efficient and experimental load is alleviated. Such integrated frameworks are needed and are ultimately actualized in the peptide-based therapeutic tool and database that have been developed in this thesis.

2.4 Small-Molecule–Based Drug Discovery

Small-molecule compounds remain a cornerstone of cancer therapy as they can regulate intracellular signalling pathways that mediate inflammation, proliferation, and survival[90–92]. Pathways of signalling that are dysregulated, especially those involving transcription factors like the TGF-beta, Jak-STAT, Wnt, Notch, Hippo, MAPK and NF-κB are vital players in tumor progression and immune modulation[93–95]. Conventional small-molecule discovery pipelines relies on experimental high-throughput screening and structure-based design, neither of which is typically cost-effective or efficient when used on large chemical libraries screening[96,97].

Computational methods, such as QSAR modelling, molecular docking, and machine learning-enhanced virtual screening, have thus gained a vital role in rankings of assessment of candidate compounds[98,99].

Machine learning models that are trained on chemical descriptors and bioassays data allow quick identification of potential inhibitors and represent complex structure-activity associations. Such models can be trained on experimental data available in public bioassay repositories, while taking special care of heterogeneity issues between assays, imbalance between classes and specificity to pathways. These in-silico models developed would aid in effective and potential drug candidates.

Table 2.3: Computational tools for small-molecule drug discovery and screening.

Tool Name	Description
ChAlPred[100]	Predicts chemical allergenicity and adverse reactions.
DrugMint[101]	Web server for drug-likeness prediction
EGFRPred[102]	Web-based prediction and design of EGFR-targeted anticancer inhibitors
AlgPred 2.0[103]	Prediction of allergenic proteins using computational approaches
MetaPred[104]	Prediction of cytochrome P450 isoforms involved in drug metabolism
CancerIN[105]	Web-based prediction of molecular anticancer activity
ntEGFR[106]	QSAR-based models for designing inhibitors targeting wild-type and mutant EGFR in cancer
DrugPred[107]	Predicts drug-likeness and binding potential of small molecules.

Pathway-directed computational screening, especially of immune and inflammatory signalling cascades, is a promising way to enhance the translational relevance. However, the existing tools lacks rigorous validation or are not designed to address pathway-specific mechanisms. The literature review, highlights the need to have scalable, pathway-specific computational models, using which the small-molecule screening of larger libraries can be done.

2.5 Research Gap

The current state of literature is evidenced by significant advances in the use of computational methods in the disease biomarker discovery field, peptide-based therapeutics, and in small-molecule drug screening. However, there are still a number of inherent setbacks. The major limitation reliance of numerous computational works on limited or outdated datasets, often accompanied by insufficient external validation. These practices reduce confidence on models developed, their reproducibility and restrict the trust of their derived analysis. Moreover, the discontinuation or poor maintenance of web-based tools, which also restricts their future viability in the research community.

In addition, despite the fact that immune regulation and modulation of signalling pathways are key to cancer progression and therapy, the computational frameworks that explicitly take into consideration immune context or pathways specificity remains limited. Predictive tools are often based on generic activity classification and do not sufficiently consider biological processes, e.g. the induction of cytokines or signalling cascades by transcription factors.

These gaps collectively indicate that the computational frameworks that are integrative, rigorously validated, and biologically informed are a need of the hour. Addressing these unmet needs forms the central motivation of this thesis, which aims to develop a cohesive suite of machine learning–based in-silico tools encompassing gene expression–driven biomarker discovery, therapeutic peptide and protein prediction, immune modulation analysis, and pathway-focused small-molecule screening. The following chapters presents the design, validation and use of these frameworks in detail.

Chapter 3
Prostate Cancer Biomarker Prediction

3.1 Introduction

Prostate adenocarcinoma (PRAD) is one of the most common malignant disorders in men all over the world [108,109]. Although the diagnostics has improved, the challenge of early detection is still a clinical problem since the disease is asymptomatic at the early stages [110,111]. Widely used prostate-specific antigen (PSA), the test limits itself due to a higher percentage of false positives and usually leads to unnecessary biopsies and overtreatment [112,113]. Consequently, there is an urgent need to establish trustworthy non-invasive biomarkers in order to enhance early detection and lessen diagnostic load.

In the chapter, we introduce a machine learning (ML)-based prostate cancer screening model that uses gene expression data to determine a set of highly discriminative biomarkers. The proposed study combines various feature selection methods and a new approach to feature engineering to improve the accuracy of models and contribute to the creation of more efficient diagnostic tools.

3.2 Background

The recent discoveries of molecular biology have brought new insights into the development of prostate cancer, and it has been shown that genetic and epigenetic modifications play a role during the genesis of tumor [114,115]. Markedly, the detection of PCA3, TMPRSS2:ERG fusion genes, and certain mutations in TP53, FOXA1, PTEN, and SPOP have enhanced the knowledge of the disease progression [40–43,116]. A number of proposed RNA-based biomarker panels, including SelectMDx and ConfirmMDx, have been suggested, although their application in clinical studies is still limited because of data limitations [117,118]. Combining machine learning methods with gene expression data is a possibility to enhance the accuracy of prostate cancer detection with a minimal number of unnecessary invasive procedures [115,119].

3.3 Materials and Methods

3.3.1 Dataset Acquisition and Preprocessing: The prostate cancer gene expression data was downloaded using the GDC TCGA Prostate Cancer (PRAD) dataset via Xena Browser [120]. This data consisted of 500 prostate cancer samples and 51 normal tissue samples and expression profile of 20,530 genes. To reduce variability, normalization of FPKM values was done through a log2 transformation.

3.3.2 Feature Selection Techniques: In order to determine the most discriminative gene biomarkers, three methods of feature selection were used:

- Mean Difference Analysis: Genes whose mean difference in expression were increased in cancerous and normal samples were ranked and selected.
- Significance Difference in Mean: Genes were screened by their statistical significance including standard deviation to narrow down the choice.
- Area Under the Curve (AUC-ROC) Analysis: The classification ability of each gene was measured by computing AUC scores, with the top-performing genes chosen for model development.

3.3.3 Propensity Index-Based Feature Engineering: A novel approach, the propensity index matrix, was introduced to enhance classification performance. The expression range for each gene was divided into ten bins, and the proportion of cancerous and non-cancerous samples in each bin was calculated. Gene expression values were then replaced with their respective propensity scores, creating a refined dataset for machine learning models.

3.3.4 *Machine Learning Models Implementation:* In this study, we have divided the dataset in 80:20 ratios i.e. 80% dataset was used for training of models and 20% was marked as independent validation dataset [121,122]. Several machine learning classifiers were employed to develop predictive models for identifying PRAD cancer and normal samples. We have applied Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Linear Regression (LR), Gaussian Naïve Bayes (GNB), XGBoost (XGB) Each model was trained using five-fold cross-validation over 80% dataset, and performance evaluation markers such as sensitivity, specificity, accuracy, Matthew's correlation coefficient (MCC), and AUC-ROC values were recorded. The optimum models were then evaluated using independent validation dataset.

$$Sensitivity = \frac{TP}{TP+FN} * 100 \quad (1)$$

$$Specificity = \frac{TN}{TN+FP} * 100 \quad (2)$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} * 100 \quad (3)$$

$$MCC = \frac{(TP*TN) - (FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4)$$

Where, FP is false positive, FN is false negative, TP is true positive and TN is true negative.

3.4 Results

3.4.1 *Model Performance with Gene Expression-Based Features:* The top 10 genes identified using three feature selection techniques can be referred in Figure 3.1. Models trained using the top ten genes from mean based analysis achieved an AUC of 0.91 with logistic regression over validation dataset, highlighting the predictive potential of these biomarkers.

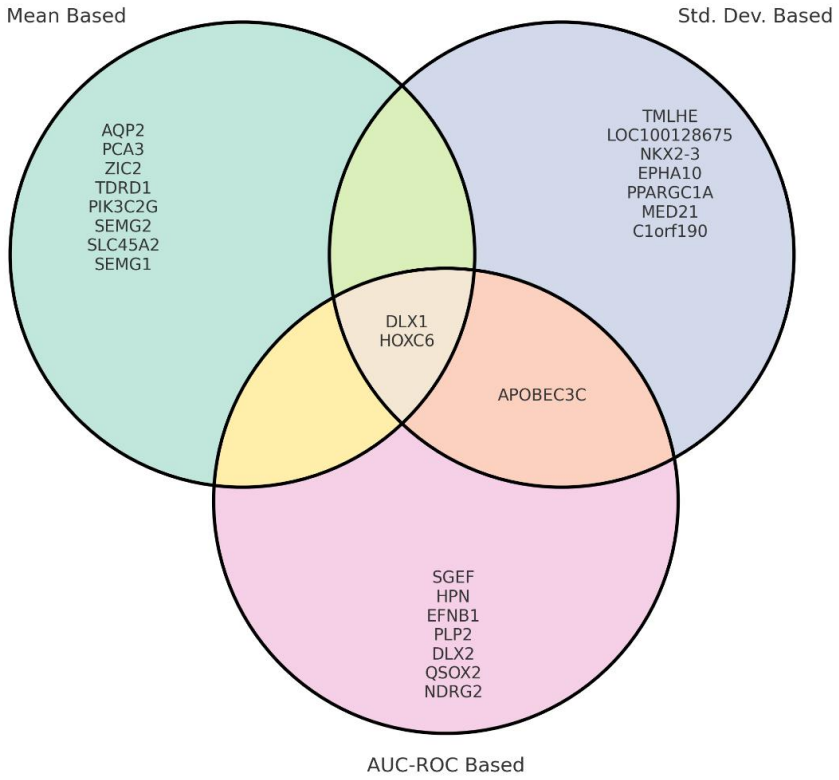


Figure 3.1: Venn Diagram showing top 10 genes identified by each feature selection technique.

Similarly, the SVM model using genes selected via Significance Difference in Mean achieved an AUC of 0.92 on training data and 0.89 on validation data. AUC-based feature selection further enhanced predictive accuracy, with KNN obtaining an AUC of 0.91 on validation datasets. The detailed performance parameters for training and validation dataset for all three-feature selection approaches can be referred in Table 3.1.

Table 3.1: ML model performance for training and validation dataset for all feature selection approaches.

Features	Model	Training Dataset					Validation Dataset				
		Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
	GNB	97.41	83.33	96.10	0.90	0.78	94.63	68.75	92.12	0.82	0.59

Top 10 genes selected using mean based approach.	**KNN**	98.85	75.00	96.63	0.87	0.79	97.32	75.00	95.15	0.86	0.73
	SVM	94.82	88.88	94.29	0.92	0.73	85.91	75.00	84.85	0.80	0.45
	DT	98.56	83.33	96.36	0.90	0.78	97.99	62.50	94.54	0.80	0.66
	RF	99.42	72.22	96.62	0.86	0.80	97.98	62.50	94.54	0.80	0.66
	XGB	98.56	72.22	96.10	0.85	0.76	95.95	62.50	92.73	0.79	0.58
	LR	93.96	88.89	93.50	0.91	0.70	83.22	75.00	82.42	0.91	0.70
Top 10 genes selected using significance difference in mean.	**GNB**	97.41	91.97	96.88	0.95	0.83	95.30	81.25	93.94	0.88	0.70
	KNN	98.56	86.11	97.40	0.92	0.85	95.30	62.50	92.12	0.79	0.56
	SVM	95.40	88.89	94.80	0.92	0.74	90.60	87.50	90.30	0.89	0.62
	DT	97.70	77.77	96.09	0.88	0.75	97.32	56.25	93.33	0.76	0.58
	RF	97.98	77.78	96.35	0.88	0.77	97.31	75.00	95.15	0.86	0.72
	XGB	97.99	80.56	96.36	0.89	0.79	96.64	81.25	95.15	0.89	0.74
	LR	95.40	97.22	95.57	0.96	0.80	91.95	87.50	91.15	0.88	0.65
Top 10 genes selected using AUC based approach.	**GNB**	94.54	94.45	94.53	0.94	0.75	93.29	87.50	92.73	0.90	0.68
	KNN	98.27	86.11	97.14	0.92	0.83	95.30	87.50	94.55	0.91	0.74
	SVM	90.51	94.45	90.90	0.92	0.65	88.59	87.50	88.48	0.88	0.58
	DT	97.99	83.34	96.10	0.91	0.80	96.64	68.75	93.94	0.83	0.65
	RF	98.85	72.22	97.40	0.86	0.77	97.31	81.25	95.76	0.89	0.76
	XGB	98.27	75.00	96.09	0.87	0.76	97.31	81.25	95.76	0.89	0.76
	LR	92.53	94.45	92.72	0.93	0.70	88.59	87.50	88.48	0.88	0.58

3.4.2 *Model Performance with Propensity Index*: However, in prominent features we replaced gene expression values with propensity index scores, which significantly improved ML model performance. The RF model achieved an AUC of 1.00 on training data and 0.91 on validation data. Similarly, LR and SVM models demonstrated enhanced classification accuracy when trained on propensity-indexed feature data. The detailed performance parameters for training and validation dataset for all three-feature selection approaches using propensity index can be referred in Table 3.2.

Table 3.2: ML model performance for training and validation dataset for all feature selection approaches using Propensity Index.

Features	**Model**	**Training Dataset**					**Validation Dataset**				
		Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
Top 10 genes selected using mean based approach.	**GNB**	98.50	100.00	98.64	1.00	0.93	100.00	80.00	98.18	0.90	0.89
	KNN	98.50	100.00	98.64	1.00	0.93	100.00	80.00	98.18	0.89	0.89
	SVM	98.75	97.62	98.64	1.00	0.93	100.00	60.00	96.36	0.90	0.76
	DT	99.25	92.86	98.64	0.96	0.92	100.00	30.00	93.64	0.65	0.53
	RF	98.50	100.00	98.64	1.00	0.93	100.00	70.00	97.27	0.91	0.82
	XGB	97.75	97.62	97.74	0.99	0.88	100.00	80.00	98.18	0.93	0.89
	LR	98.75	100.00	98.87	1.00	0.94	100.00	70.00	97.27	0.87	0.82
Top 10 genes	**GNB**	99.50	100.00	99.55	1.00	0.97	100.00	70.00	97.27	0.90	0.82
	KNN	99.75	100.00	99.77	1.00	0.99	100.00	70.00	97.27	0.95	0.82

selected using significance difference in mean.	**SVM**	99.00	100.00	99.10	1.00	0.95	100.00	80.00	98.18	0.95	0.89
	DT	97.75	90.48	97.06	0.94	0.84	99.00	60.00	95.45	0.80	0.69
	RF	99.75	95.24	99.32	1.00	0.96	100.00	80.00	98.18	0.94	0.89
	XGB	98.75	97.62	98.64	1.00	0.93	99.00	80.00	97.27	0.93	0.83
	LR	99.00	100.00	99.10	1.00	0.95	100.00	80.00	98.18	0.97	0.89
Top 10 genes selected using AUC based approach.	**GNB**	99.50	100.00	99.55	1.00	0.97	100.00	50.00	95.45	0.85	0.69
	KNN	99.50	100.00	99.55	1.00	0.97	100.00	50.00	95.45	0.84	0.69
	SVM	99.25	100.00	99.32	1.00	0.96	100.00	70.00	97.27	0.95	0.82
	DT	97.75	90.48	97.06	0.94	0.84	99.00	60.00	95.45	0.80	0.69
	RF	98.25	100.00	98.42	1.00	0.92	99.00	70.00	96.36	0.99	0.76
	XGB	98.50	95.24	98.19	0.99	0.90	99.00	80.00	97.27	0.97	0.83
	LR	98.75	100.00	98.87	1.00	0.94	100.00	70.00	97.27	0.78	0.82

3.4.3 *Single-Gene Classification Analysis:* In order to map the individual predictive power of each selected gene, threshold-based classification ML models were developed. As evident from Figure 3.2, out of these 30 selected genes, 13 demonstrated predictive probabilities exceeding 0.97. Notably, HOXC6, DLX1, TDRD1, and PCA3 emerged as key discriminatory biomarkers, aligning with previous studies on prostate cancer biomarker discovery.

Gene	Minimum	Maximum	Threshold	Sensitivity	Specificity	Accuracy	Prob_of_CP
DLX2	0.000	10.990	2.66	90.40	90.38	90.40	0.989
HPN	3.510	14.726	11.05	89.00	88.46	88.95	0.987
QSOX2	7.435	10.686	8.59	88.80	88.46	88.77	0.987
HOXC6	0.777	11.083	6.24	88.40	88.46	88.41	0.987
DLX1	0.000	11.545	4.93	87.60	88.46	87.68	0.986
SGEF	7.060	12.735	9.84	88.20	86.54	88.04	0.984
SLC45A2	0.000	12.989	2.39	87.20	86.54	87.14	0.984
EPHA10	2.774	9.523	6.66	87.00	86.54	86.96	0.984
NKX2-3	0.000	8.208	2.30	85.20	84.62	85.14	0.982
LOC100128675	0.000	6.911	2.06	84.80	84.62	84.78	0.981
ZIC2	0.000	9.725	2.37	82.60	82.69	82.61	0.979
TDRD1	0.000	11.333	3.36	77.80	76.92	77.72	0.970
PCA3	0.872	15.567	9.93	77.00	76.92	76.99	0.970

Figure 3.2: Matrix depicting predictive probabilities of top 14 selected genes.

3.5 Conclusion and Applications

The current research highlights the possibilities of machine learning in the detection of prostate cancer using the gene expression data and by applying new feature engineering methods. The propensity index-based feature introduced in this study depicts great improvement in the

accuracy of classification model compared with the old traditional methods developed using gene expression.

A comparison with existing biomarker panels, such as SelectMDx, highlights the robustness of our approach. Previous studies reported AUC values of 0.85 for SelectMDx and 0.77 for a urinary three-gene panel (HOXC6, TDRD1, DLX1), whereas our model achieved highest AUC as 0.91 for validation set [117,118,123]. The application of a wide range of machine learning classifiers also enhanced the power of predictions, with RF and SVM turning out as the best models.

We also highlight the weaknesses of PSA-based screening in our findings. KLK3 gene (linked with PSA) was not considered one of the most effective biomarkers, which supports the issue of specificity of PSA [124,125]. The other biomarkers like DLX1 and HOXC6 showed a greater potential in classification (See Figure 3.3), which is consistent with the existing literature on the use of these biomarkers in diagnosing prostate cancer.

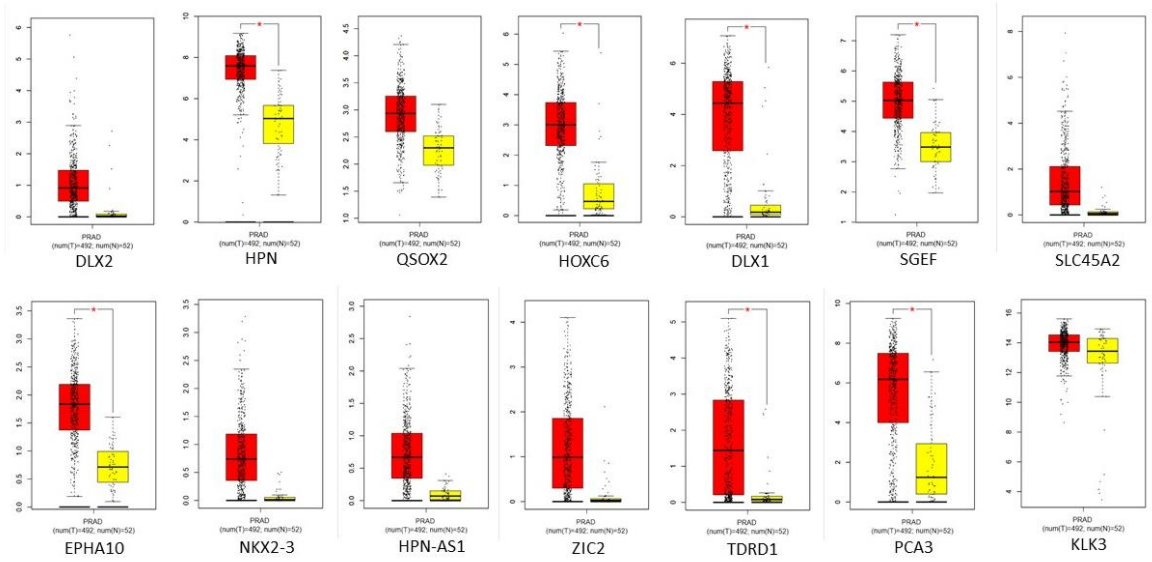


Figure 3.3: Box plots of top ranked genes; red box plot depicts cancer samples and yellow depicts normal samples.

The machine learning model proposed successfully manages to identify patients with prostate cancer with high accuracy using gene expression-based biomarkers. The implementation of the propensity index method contributed greatly to the performance of the model, and it can provide a novel model that employs the use of biomarkers in classifying the data. Future studies must aim at increasing the datasets to incorporate various patient groups with which to generalize the models, and incorporating other omics data (e.g., proteomics, metabolomics) to

speed up precise biomarker discovery. Besides this, experimental and clinical studies that could be used to verify the computational findings would help in the real-world implementation. This research offers one of the potential solutions that can ensure a reduction in the use of PSA testing and avoidable biopsies through a promising framework of non-invasive prostate cancer screening.

Chapter 4

Peptide-based Therapeutic repository and their Prediction

4.1 Introduction

Therapeutic peptides and proteins have evolved to be essential targets in drug development in the last 20 years due to their outstanding target specificity, favourable safety and variable mechanisms of action [126,127]. In contrast to small molecules, which can react with multiple off-target sites, and can induce unwanted side effects, peptide-based drugs can take advantage of the inherent binding affinities to be potent and less toxic [128,129]. In addition, they can be modified to have modular sequences that are more intrinsic to derive stability, bioavailability and tissue specificity [130–132].

Despite these advantages, the experimental identification and clinical development of new +peptide therapeutics nevertheless pose significant challenges: the traditional high-throughput screening campaigns are not only labor-intensive and costly, but also most promising drug candidates are eliminated at later stages because of inappropriate pharmacokinetics or immunogenicity [127,133,134]. In order to solve these bottlenecks, computational systems that combine all-encompassing details about approved therapeutics with predictive algorithms could become strong instruments in hastening drug finding and directing rational design endeavour's [135].

4.2 Background

This chapter provides two complementary elements of such an in-silico framework as illustrated in Figure 4.1. First, ThPdb2- a human curated database that summarizes the information on 894 FDA-approved peptides and proteins with more than 5,500 sequence variations, physicochemical annotations, pharmacological categories, and routes of administration. It represents a revised form of ThPdb repository that was built in 2017, listing 852 entries, of 239 US-FDA approved therapeutic peptides and proteins and of 380 drug versions of them [136]. Second ThPPred, a prediction machine learning platform which separates druggable proteins and peptides against non-therapeutic sequences; it utilizes both the descriptors of global protein sequences, as well as, motif patterns. This combined strategy empowers researchers to scan known pharmacological space, search large peptide libraries in silico, and refine candidate sequences through repeated cycles of experiments significantly simplifies the task of experimental science and can enable more rapid application in the clinic.

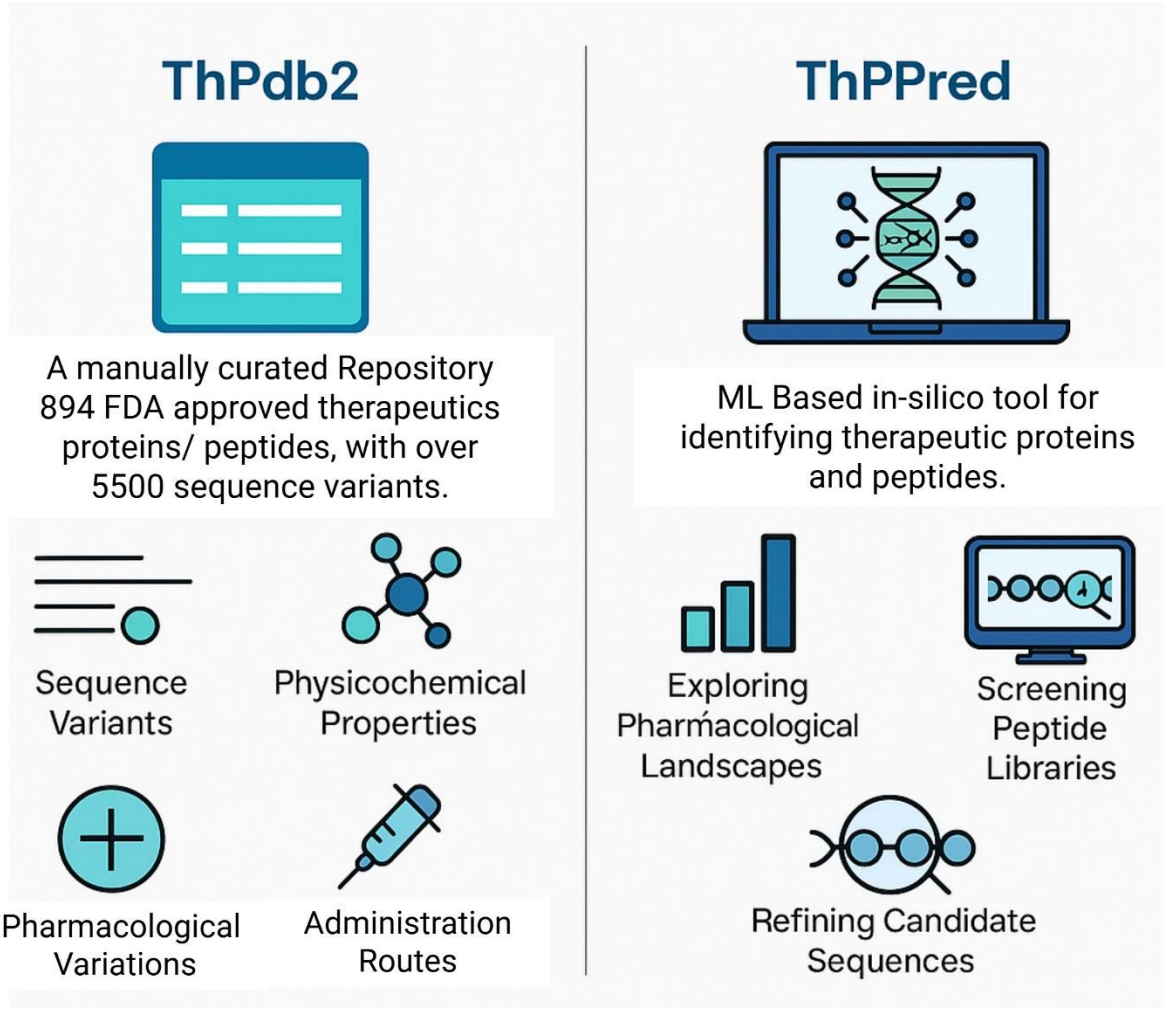


Figure 4.1: Illustration depicts overview of ThPdb2 repository and ThPPred prediction tool.

4.3 Therapeutic Protein/ Peptide Repository

We employed a data collection approach similar to that used for THPdb. DrugBank was first screened using the keywords "biotech drugs" and filtered for "Approved" and "Investigational" entries, yielding 831 therapeutic protein and peptide drugs [137]. Of these, 387 had sequences available in databases like UniProt, PubChem, and DrugBank [22,137,138]. To gather comprehensive information, we screened additional public repositories for details such as name, molecular weight, formula, isoelectric point, hydrophobicity, melting point, half-life, mechanism of action, diseases treated, pharmacodynamics, metabolism, toxicity, and administration route. We also collected data on brand names, formulations, dosages, side effects, and patents. Missing details were retrieved from PubMed and patent databases using relevant keywords. The final compilation integrates 239 drugs from THPdb and 655 newly approved therapeutics. Thus, provides manually aggregated data on 894 distinct therapeutic

molecules, resulting in a total of 5,504 entries including multiple drug variants, formulation differences, brand nomenclature etc.

4.3.1. Webserver Implementation

THPdb2 webserver was built using Apache HTTP server (version 2.4.7) and MySQL (version 5.5.62) to store and manage the data (https://webs.iiitd.edu.in/raghava/thpdb2/). This web-based user-friendly platform has 4 major modules such as: Search, Browse, Similarity Search and Data Statistics. Both basic and advances search options have been implemented in Search module. Advanced browsing options allow exploration by target class—receptors, enzymes, transporters—or by drug formulation, facilitating quick identification of analogs or scaffolds of interest. It enables several browse modules to explore therapeutic peptides/proteins based on disease category, formulation, pharmacological class, route of administration, target type, and sequence length. These features enable efficient filtering and retrieval of relevant therapeutic data for research and analysis.

Besides this, similarity searches using sequences driven by BLAST and Smith-Waterman algorithm allows scientists to map new sequences against the repository whereby sub-sequence matching (sub-searches) and super-sequence matching (super-searches) are differentiated. This feature assists in the discovery of repurposed or lead optimization due to partial overlap of therapeutics in their sequence of action. Overview of the repository in statistics also indicate important trends in peptide therapeutics. As depicted in Figure 4.2, oncology represents the largest category, with 579 entries targeting various tumour types, followed by infectious diseases (355), immunological disorders (350), and cardiovascular indications (173). Antibodies constitute the most prevalent pharmacological class (590 entries), while enzymes (382), antineoplastic agents (344), and hormones (238) also feature prominently. All search results can be downloaded in multiple formats (.csv, .xls, .pdf), empowering users to integrate ThPdb2 data seamlessly into downstream analysis.

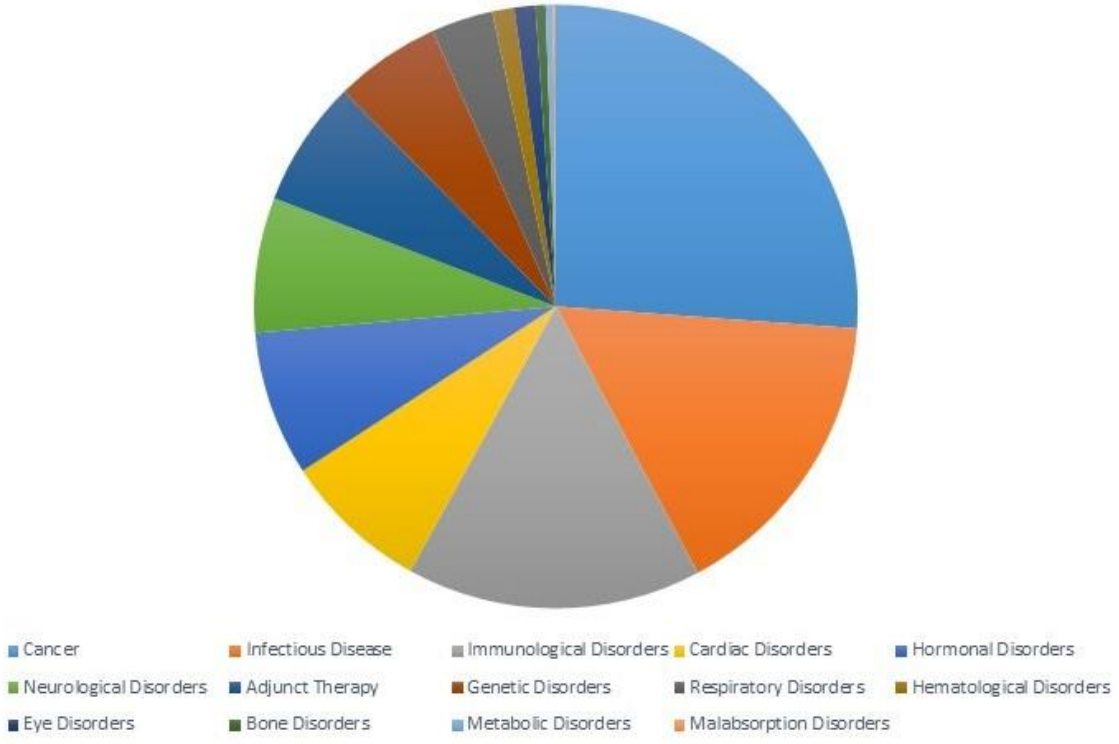


Figure 4.2: Represents the diseases against which therapeutic peptides/proteins are present in ThPDB2.

4.4 Machine Learning based Prediction tool for Druggable Sequences

With the explosion of peptide sequence data, effective therapeutic targeted drug development depends on in silico tools that can systematically pinpoint clinically viable candidates from large libraries. Numerous computational tools like PTPD, PrMFTP, TPpred-ATMV, TPpred-LE, and others have been developed for predicting only specific therapeutic peptides/ proteins properties, such as anti-cancer, anti-bacterial, anti-inflammatory, cell-penetrating peptides., etc [89,139–147]. This highlights the need for a predictive approach grounded in clinically validated therapeutics. To address this gap, we developed ThPPred—an in-silico model trained on experimentally validated therapeutic peptides and proteins employing supervised machine learning methods to discriminate FDA-approved therapeutics from non-therapeutic proteins.

4.4.1 Materials and Methods

4.4.1.1 Dataset Compilation: In order to develop our prediction model, we compiled datasets of therapeutic (positive) and non-therapeutic (negative) proteins. The positive dataset included 356 FDA-approved therapeutic proteins/peptides from THPdb2, which were filtered to retain sequences between 30 and 1500 amino acids length without non-natural residues. The negative

set included an equal number of length-matched Swiss-Prot sequences lacking any therapeutic annotation. Also, in addition to this for simulating real-world class imbalance, a secondary "realistic" dataset was developed which incorporated ten times as many negatives i.e. 3,560 sequences against the same 356 positives.

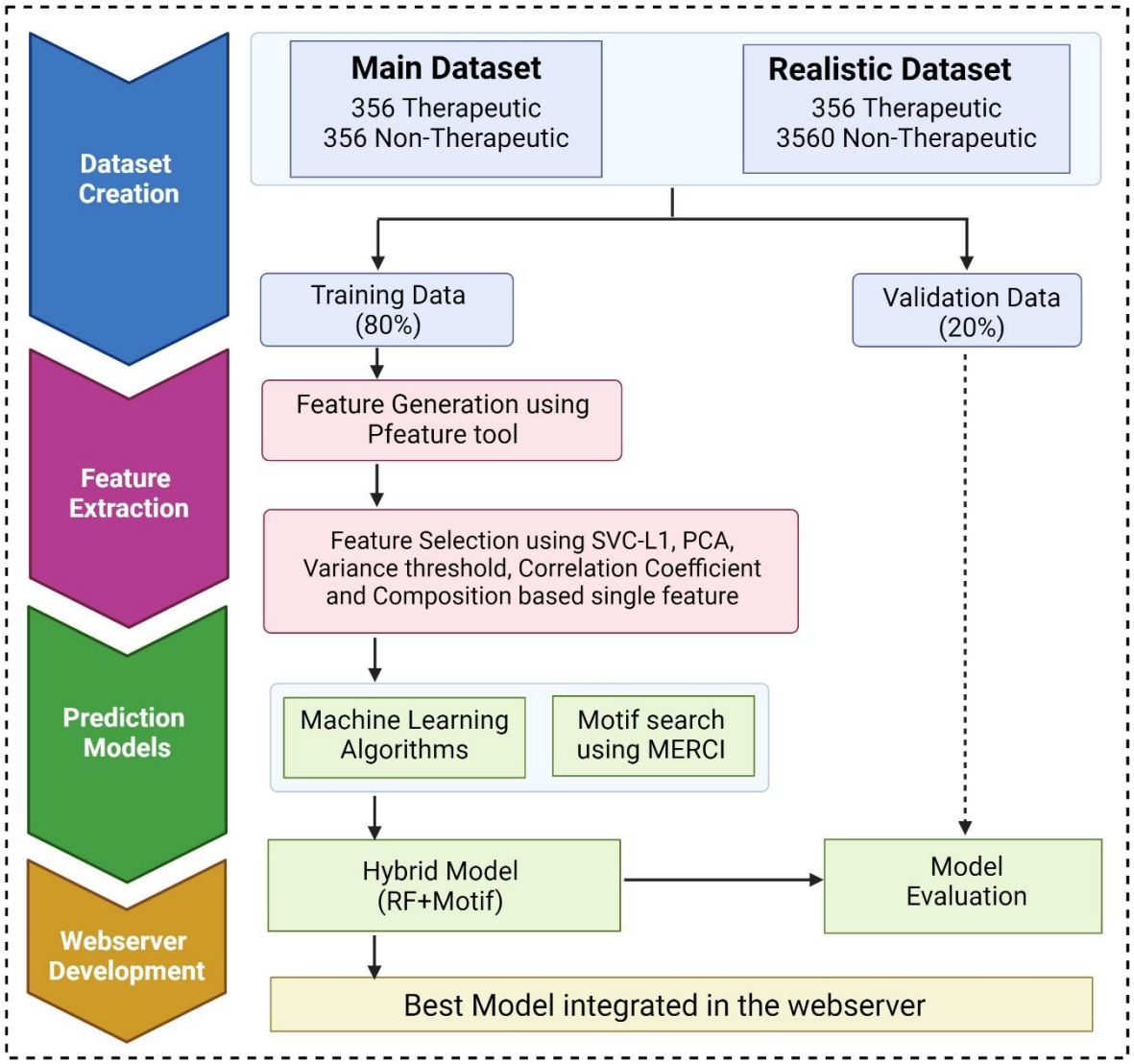


Figure 4.3: Overall workflow of ThPPred tool.

Further to reduce bias and redundancy in model evaluation, we applied CD-HIT clustering (40% cutoff) to both datasets [148]. This yielded sequence clusters which were further spliced into 80% training and 20% independent validation sets, ensuring no sequence similarity above 40% between them. This clustering-based stratification preserved dataset integrity while maximizing coverage of diverse protein sequences [103,149].

4.4.1.2 *Feature Engineering:* In this study, we have deployed Pfeature to compute approximately 9,189 features per sequence, encompassing amino acid composition metrics, physicochemical scales, and higher-order constructs such as pseudo-amino acid composition etc [150]. Given the high dimensionality of generated feature matrix, feature reduction was essential. We employed several feature selection techniques such as L1-regularized linear SVM (SVC-L1), Variance Threshold, Correlation Coefficient, and principal component analysis (PCA). The result of this process narrowed the feature space to about 60-90 informative features, trading off predictive power and model interpretability.

4.4.1.3 *Machine learning model evaluation:* In the current study, we have applied six machine learning models such as the Random Forest, Support Vector Classifier, XGBoost, Logistic Regression, Decision Tree, and k-Nearest Neighbours. When carrying out the development of the machine learning based models, five-fold cross-validation was implemented within the training set. Training set models that worked best were further tested on independent validation set and the performance parameters were documented. This ensured that performance estimates reflect genuine generalization to novel sequences.

4.4.1.4 *Alignment based approach with Machine learning model evaluation:* In order to enhance predictive accuracy, we incorporated motif analysis using the MERCI tool, which identifies exclusive sequence patterns in proteins/peptide dataset. Twenty-three motifs exclusive to the therapeutic protein/peptide were discovered, as listed in Table 4.1. In the ensemble framework, a motif hit contributed a positive score, which was combined with the machine learning model's prediction score to produce a hybrid decision metric.

Table 4.1: List of exclusive motifs found in therapeutic and non-therapeutic proteins dataset.

Main Dataset				
Motifs exclusively present in positive dataset			**Motifs exclusively present in negative main dataset**	
Motifs	*Occurrence*		*Motifs*	*Occurrence*
sulfurbasicgapSneutral	135		PVchargedP	135
YaromaticC	133		PVchargedPsurface	132
neutralKNneutral	133		PVchargedPpolar	121
WgapQG	132		PVacidicP	108
Qaliphaticaromaticneutralaliphatic	132		PVchargedPneutral	106
basicVneutralN	132		TPaliphaticPaliphatic	105
CbasicgapSneutral	132		aliphaticPaliphaticbasicP	104
aromaticgapCchargedmedium	131		LacidicLaromatic	100
VSW	129		LacidicLaromaticaliphatic	100

PSVF	129		ILDL	92
CSV	126		GFPV	92
NWY	124		FPDW	91
KGQP	121		LDLW	91
neutralVpolarWsurface	115		PMgapT	14
SgapWneutralneutral	94		PVchargedgapP	12
cyclicgapCchargedmedium	19		ValiphaticFgapP	11
surfacecycliclargeCcharged	13		PVgapchargedP	7
YgapCgapN	13		PVgapacidicP	6
CgapSV	12		ILgapDL	1
VTgapC	8			
WgapYgapD	7			
cyclicneutralCpolarmedium	5			
YgaparomaticC	5			

Note: sulphur: C,M; basic: R,H,K; gap: single amino acid gap; aromatic: H,F,W,Y ; aliphatic: A,G,I, L,V; neutral: A,N,C,Q,G,I,L,M,F,P,S,T,W,Y,V; charged: D,E,R,H,K; surface: R,N,D,E,Q,G,H,K,P,S, T,Y; polar: R,N,D,C,E,Q,H,K,S,T; acidic: D,E.

4.4.2 Results

4.4.2.1 *Machine learning models using compositional features*

Among the compositional features evaluated, Tripeptide Composition (TPC) demonstrated the best predictive performance on the main validation dataset. The Random Forest model developed using TPC achieved an AUC-ROC of 0.78 and an MCC of 0.56, outperforming the corresponding AAC- and DPC-based models. On the realistic validation dataset, Gaussian Naïve Bayes (GNB) using DPC features yielded the highest performance with an AUC-ROC of 0.77 and an MCC of 0.57. Overall, the results suggest that sequence-order information captured by dipeptide and tripeptide compositions is more informative for distinguishing therapeutic and non-therapeutic peptides than amino acid composition alone. The detailed performance of all machine learning models developed using compositional features on both the main and realistic datasets is presented in Table 4.2.

Table 4.2: Performance of ML models using composition-based features over main and realistic dataset.

		Main validation data					**Realistic validation data**				
Features	**Model**	**Acc**	**Sens**	**Spec**	**AUC**	**MCC**	**Acc**	**Sens**	**Spec**	**AUC**	**MCC**
	RF	0.63	0.75	0.67	0.63	0.26	0.93	0.30	0.43	0.64	0.46
	DT	0.61	0.80	0.67	0.61	0.23	0.88	0.48	0.43	0.70	0.36
AAC	**KNN**	0.58	0.70	0.63	0.58	0.16	0.94	0.42	0.55	0.71	0.55

	GNB	0.57	0.66	0.61	0.57	0.14	0.93	0.32	0.45	0.66	0.45
	XGB	0.61	0.75	0.65	0.61	0.22	0.83	0.45	0.32	0.66	0.25
	RF	0.72	0.68	0.71	0.72	0.44	0.92	0.10	0.18	0.55	0.28
	DT	0.55	0.59	0.57	0.55	0.10	0.70	0.24	0.13	0.49	-0.01
	KNN	0.55	0.82	0.64	0.55	0.12	0.92	0.31	0.41	0.64	0.39
	GNB	0.51	0.63	0.57	0.51	0.03	**0.93**	**0.58**	**0.61**	**0.77**	**0.57**
DPC	**XGB**	0.73	0.73	0.73	0.73	0.46	0.93	0.23	0.36	0.61	0.41
	RF	**0.78**	**0.79**	**0.78**	**0.78**	**0.56**	0.92	0.07	0.13	0.54	0.25
	DT	0.56	0.59	0.58	0.56	0.13	0.59	0.35	0.14	0.48	-0.02
	KNN	0.58	0.77	0.65	0.58	0.17	0.90	0.23	0.30	0.60	0.27
	GNB	0.61	0.76	0.66	0.61	0.24	0.93	0.25	0.39	0.62	0.44
TPC	**XGB**	0.58	0.77	0.65	0.58	0.17	0.91	0.15	0.23	0.57	0.23

4.4.2.2 *Machine learning models using feature selection approaches*

To further improve predictive performance and reduce feature redundancy, several feature selection approaches, including SVC-L1, Variance Threshold, Correlation Coefficient, and Principal Component Analysis (PCA), were applied to the generated feature set. As shown in Table 4.3, the SVC-L1 based feature selection approach yielded the best performance on the main validation dataset, where the Random Forest classifier achieved an AUC-ROC of 0.80 and an MCC of 0.61. On the realistic validation dataset, the Support Vector Classifier (SVC) developed using SVC-L1 selected features achieved the highest AUC-ROC of 0.78 with an MCC of 0.60. Overall, the SVC-L1 feature selection approach consistently outperformed the other feature selection methods on both datasets, indicating its effectiveness in identifying discriminative features for distinguishing therapeutic and non-therapeutic peptides.

Table 4.3: Performance of ML models using features selection approaches over main and realistic dataset.

	Main validation data					**Realistic validation data**				
Model	**Acc**	**Sens**	**Spec**	**AUC**	**MCC**	**Acc**	**Sens**	**Spec**	**AUC**	**MCC**
SVC-L1 based feature selection approach										
RF	**0.80**	**0.75**	**0.79**	**0.80**	**0.61**	0.95	0.49	0.64	0.74	0.64
DT	0.72	0.75	0.73	0.72	0.44	0.92	0.49	0.51	0.73	0.47
KNN	0.58	0.65	0.61	0.58	0.17	0.91	0.25	0.35	0.62	0.34
SVC	0.65	0.66	0.66	0.65	0.31	**0.94**	**0.58**	**0.63**	**0.78**	**0.60**
XGB	0.71	0.77	0.73	0.71	0.43	0.95	0.51	0.63	0.75	0.63
Variance Threshold based feature selection approach										
RF	0.75	0.68	0.73	0.75	0.51	0.93	0.32	0.47	0.66	0.51
DT	0.73	0.73	0.73	0.73	0.46	0.92	0.48	0.52	0.72	0.47
KNN	0.58	0.65	0.61	0.58	0.17	0.92	0.34	0.43	0.66	0.41
SVC	0.54	0.63	0.58	0.54	0.07	0.92	0.31	0.42	0.65	0.41

XGB	0.75	0.80	0.76	0.75	0.50	0.95	0.49	0.63	0.74	0.62
Correlation Coefficient based feature selection approach										
RF	0.54	0.48	0.51	0.54	0.09	0.85	0.06	0.06	0.49	-0.02
DT	0.52	0.44	0.48	0.52	0.04	0.85	0.06	0.06	0.49	-0.02
KNN	0.61	0.49	0.56	0.61	0.23	0.79	0.07	0.06	0.47	-0.06
SVC	0.55	0.72	0.61	0.55	0.10	0.91	0.00	0.00	0.50	0.00
XGB	0.45	0.52	0.49	0.45	-0.10	0.85	0.03	0.03	0.48	-0.04
Principal Component analysis-based feature selection approach										
RF	0.58	0.69	0.62	0.58	0.16	0.94	0.34	0.49	0.67	0.53
DT	0.58	0.66	0.61	0.58	0.16	0.92	0.48	0.51	0.72	0.47
KNN	0.72	0.72	0.72	0.72	0.44	0.93	0.32	0.47	0.66	0.51
SVC	0.61	0.77	0.66	0.61	0.22	0.92	0.15	0.27	0.58	0.36
XGB	0.60	0.65	0.62	0.60	0.20	0.86	0.42	0.35	0.66	0.28

4.4.2.3 *Ensemble model performance*

While feature-based models capture global sequence characteristics, alignment-based methods provide local sequence similarity information. As shown in Table 4.4, the Random Forest-based ensemble achieved the best performance on the main validation dataset with an AUC-ROC of 0.92 and an MCC of 0.83. On the realistic dataset, the SVC-based ensemble attained the highest AUC-ROC of 0.78 with an MCC of 0.62. These results demonstrate that integrating complementary sources of information improves the discrimination of therapeutic and non-therapeutic peptides. The improved performance of the ensemble models highlights the benefit of combining alignment-based information with machine learning-derived sequence features.

Table 4.4: Performance of ensemble method combining alignment based and ML models over main and realistic dataset.

	Main validation data					**Realistic validation data**				
Model	**Acc**	**Sens**	**Spec**	**AUC**	**MCC**	**Acc**	**Sens**	**Spec**	**AUC**	**MCC**
RF	0.92	0.92	0.92	0.92	0.83	0.95	0.49	0.64	0.74	0.64
DT	0.89	0.94	0.90	0.89	0.79	0.92	0.49	0.54	0.73	0.51
KNN	0.85	0.90	0.86	0.85	0.71	0.92	0.25	0.36	0.62	0.37
SVC	0.86	0.86	0.86	0.86	0.72	0.94	0.58	0.65	0.78	0.62
XGB	0.91	0.92	0.91	0.91	0.82	0.95	0.51	0.64	0.75	0.64

4.4.3 Webserver Implementation

ThPPred has been deployed as both a web server and a standalone Python package at https://webs.iiitd.edu.in/raghava/thppred/. Users can submit single or batch FASTA files, choose between the main or realistic dataset models, adjust classification thresholds, and

optionally perform motif scans or even propose single-point mutations to increase predicted druggability. The responsive HTML interface makes it compatible between desktops and mobile devices that enable high-throughput virtual screening of peptide libraries.

4.5 Integration and Applications

The synergy created by ThPdb2 and ThPPred reflects a closed pipeline loop in the search of peptide therapeutics. The researchers start their research by investigating the ThPdb2 to learn about known drug scaffolds, target classes, and physicochemical trends. Novel sequences derived from genome mining, combinatoric libraries or rational design are then screened using ThPPred to select the best novel sequence candidate. Iterative sequence modification is informed by motif analyses and rankings of feature importance, and analogs are put into place relative to approved therapeutics by similarity searches against ThPdb2. This combined workflow can greatly decrease the workload of the experiments by focusing on high-value candidates to be synthesized and evaluated biologically.

4.6 Conclusion and Future Directions

We have described the creation of ThPdb2, a highly annotated collection of FDA-approved peptide and protein drugs, and ThPPred, a machine learning system of therapeutic potential prediction by primary sequence, in this chapter. A combination of large volumes of manually curated data with rigorous modelling methods and motif analytics approach results in a high-performance end-to-end computational platform to the scientific community. In the future, it is possible that researchers will incorporate into ThPPred models structural and multi-omics features of predicted tertiary conformations, post-translational modifications, and interactome contexts. They are also able to investigate automated pipelines in order to revise ThPdb2 as new approvals are generated so that both data and forecasts are up to date. Finally, this evolving framework aims to accelerate the translation of peptide and protein therapeutics from in silico hypotheses to clinical realities.

Chapter 5

IL13 Immune modulation prediction in cancer

5.1 Introduction

Interleukin 13 (IL-13) is an immunoregulatory cytokine that is primarily released by activated T helper type 2 (Th2) cells [151]. It is important in regulating immune responses, whereby it majorly suppresses the synthesis of inflammatory cytokines [152]. In addition to Th2 cells, IL-13 is also released by other cell types eosinophils, mast cells, basophils, smooth muscle cells, natural killer cells, and fibroblasts, all of which participate in its wide range of biological activities [153,154]. Although IL-4 and IL-13 share functional similarities, IL-13 has emerged as a more promising therapeutic target [154]. IL-13 is implicated in several key biological functions such as regulation of airway hyperresponsiveness, allergic inflammation, mastocytosis, goblet cell hyperplasia, tissue eosinophilia, IgE antibody production, tumor cell growth, tissue remodelling, parasitism, and fibrosis [155,156].

Several studies have highlighted the role of IL-13 in cancer immunology [157,158]. Alterations in IL-13 effector pathways are linked to malignancies such as B-cell chronic lymphocytic leukemia and Hodgkin's disease [158,159]. IL-13 inhibits tumor immunosurveillance and can thus be targeted by immunotherapies aiming to activate type-1 anti-cancer responses [160,161]. Overexpression of IL-13 receptors has been observed in multiple cancer types including pancreatic, gastric, colon, and notably, glioblastoma multiforme, where IL-13 binding receptors are densely populated in tumor tissues but scarcely in normal cells [162–166].

5.2 Background

The immunomodulatory influence of IL-13 positions it as a central node in a variety of disease states, including cancer, asthma, infectious diseases like COVID-19 [167,168]. Despite its crucial immunological role, no computational method existed to predict IL-13-inducing peptides prior to this study. Predicting peptides that induce IL-13 can aid in vaccine design and immunotherapeutic development. To address this gap, we developed IL-13Pred, the first in-silico tool to predict IL-13-inducing and non-inducing peptides. This study presents the systematic development of IL-13Pred using machine learning models trained on curated datasets of experimentally validated peptides.

5.3 Material and Method

5.3.1 Dataset Preparation

In this study, 343 experimentally validated IL-13-inducing human peptides were retrieved from IEDB [169]. Post pre-processing (removal of peptides shorter than 8 or longer than 35 residues, and duplicates), 313 unique IL-13 inducers were retained as positive dataset. For negative data, peptides known to induce other cytokines (e.g., IL1α, IL1β, TNFα) were compiled and filtered similarly, resulting in 2908 non-inducing peptides.

5.3.2 Feature Generation and selection

We utilized the Pfeature tool, and computed 9165 compositional features per peptide including AAC, DPC, TPC, ATC, and others [150]. These features capture sequence-level attributes which are required for developing classification tool. We used SVC with L1 penalty (SVC-L1) to select 95 key features from large feature matrix. Upon that a feature-selector tool (LightGBM-based) was deployed to further ranked these based on their classification utility.

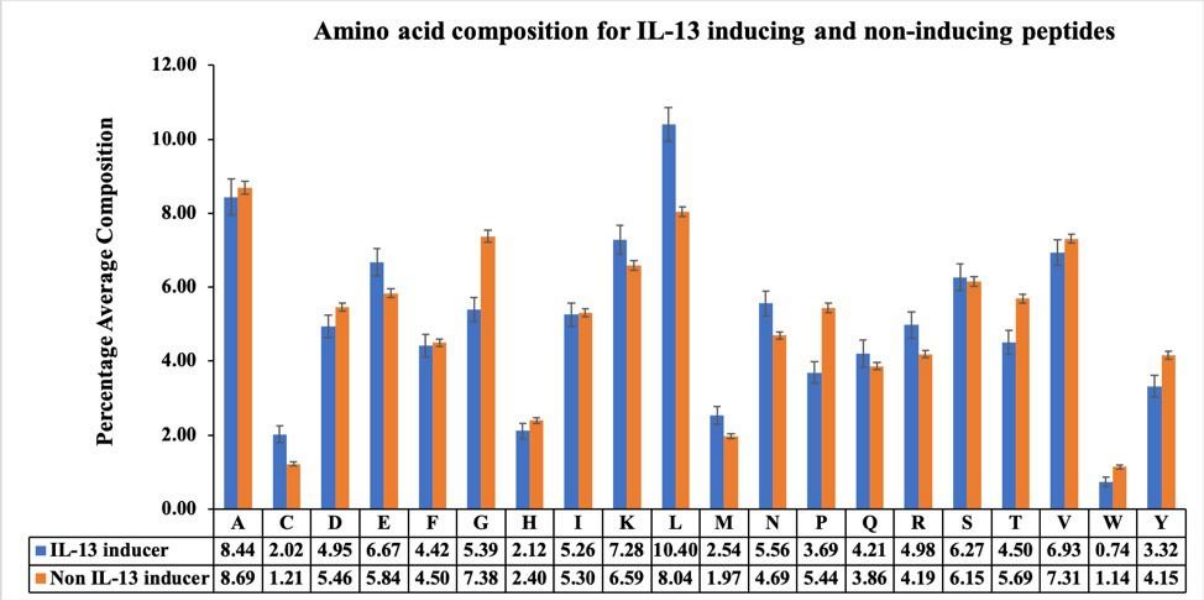
5.3.3 Machine Learning Models

In order to develop prediction models, several machine learning-based algorithms including Random Forest (RF), eXtreme Gradient Boosting (XGB), K-nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Gaussian Naive Bayes (GNB), and Decision Tree (DT) were implemented. We deployed these using Python's Scikit-learn library. In order to avoid overfitting, we adopted best practices such as 5-fold cross-validation and 80:20 external validation for robust model evaluation. Standard performance metrics were used such as Sensitivity, Specificity, Accuracy, MCC (Matthews Correlation Coefficient), and AUC (Area Under the ROC Curve).

5.4 Results

5.4.1 Compositional and Positional Analysis

IL-13 inducing peptides exhibited a compositional bias towards residues like Leucine (L), Lysine (K), Methionine (M), and Asparagine (N), suggesting their potential role in immune activation. In contrast as shown in Figure 5.1, non-inducing peptides were enriched in Glycine (G), Proline (P), Tyrosine (Y), and Threonine (T), indicating a distinct structural or functional preference.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
IL-13 inducer	8.44	2.02	4.95	6.67	4.42	5.39	2.12	5.26	7.28	10.40	2.54	5.56	3.69	4.21	4.98	6.27	4.50	6.93	0.74	3.32
Non IL-13 inducer	8.69	1.21	5.46	5.84	4.50	7.38	2.40	5.30	6.59	8.04	1.97	4.69	5.44	3.86	4.19	6.15	5.69	7.31	1.14	4.15

Figure 5.1: Amino acid composition distribution among IL-13 inducing and non-inducing peptides.

Positional analysis revealed that certain residues, particularly Leucine and Asparagine, were frequently enriched at specific N- and C-terminal positions in IL-13 inducing peptides. This highlights that not only the overall composition but also the spatial arrangement of key amino acids plays a critical role in modulating peptide immunogenicity and IL-13 induction potential.

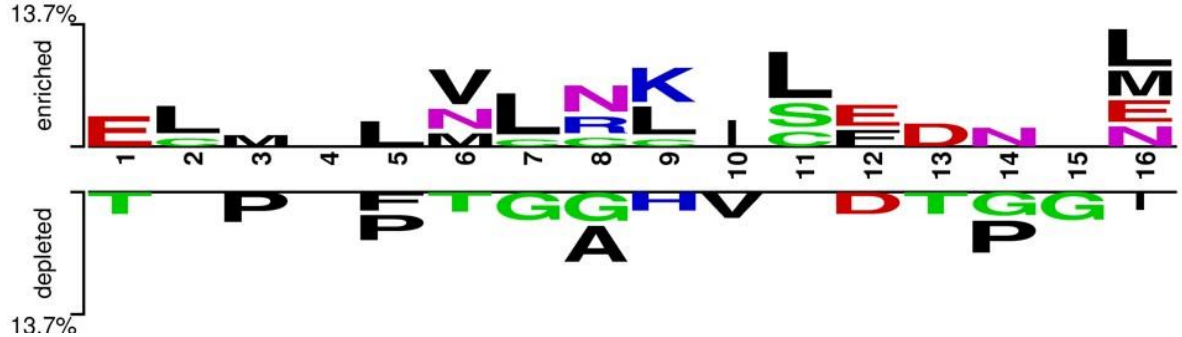


Figure 5.2: Positional analysis distribution among IL-13 inducing and non-inducing peptides.

5.4.2 Machine learning model performance

We developed prediction models using 313 IL-13 inducing and 2,908 non-inducing unique peptides. Various machine learning methods SVM, KNN, LR, DT, RF, XGB, and GNB were applied. As depicted in Table 5.1, random forest model developed using 95 feature set outperformed others, models by achieving AUCs of 0.87 and 0.83 on training and validation sets. Using the Python feature-selector library, we ranked 95 features and built models with top

10 feature as well. Models with the top 10 features performed comparably well, with the XGB model achieving AUCs of 0.83 and 0.80 on training and validation dataset respectively.

Table 5.1: Machine learning models performance parameters using various classifiers.

Classifier	Sens	Spec	Accuracy	AUC	MCC
Validation Dataset performance using top 10 features					
DT	60.26	71.43	70.08	0.72	0.22
GNB	64.10	66.31	66.05	0.73	0.21
KNN	60.26	55.73	56.28	0.61	0.11
LR	56.41	59.61	59.23	0.63	0.11
RF	64.10	75.13	73.80	0.77	0.28
XGB	71.80	73.02	72.87	0.80	0.30
SVC	46.15	74.25	70.85	0.64	0.15
Validation Dataset performance using top 95 features					
DT	51.28	66.31	64.50	0.60	0.12
GNB	38.46	77.43	72.71	0.61	0.12
KNN	50.00	69.31	66.98	0.62	0.13
LR	58.97	68.25	67.13	0.68	0.19
RF	70.51	77.07	76.28	0.83	0.34
XGB	69.23	73.19	72.71	0.80	0.30
SVC	51.28	68.08	66.05	0.62	0.13

In order to check the robustness of the developed models, we reshuffled the dataset 10 times and calculated the performance of models on top-10 features. After reshuffling the data 10 times, XGB maintained robust performance (AUC 0.82 ± 0.03), refer Table 5.2.

Table 5.2: Performance of machine learning models developed after reshuffling dataset 10 times using top 10 features.

Classifier	Sens	Spec	Accuracy	AUC	MCC
DT	63.65 ± 4.76	67.94 ± 2.83	67.52 ± 2.52	0.71 ± 0.03	0.20 ± 0.03
GNB	72.38 ± 4.92	63.81 ± 2.52	64.65 ± 2.30	0.74 ± 0.02	0.22 ± 0.03
KNN	64.44 ± 5.35	51.25 ± 2.00	52.54 ± 1.83	0.62 ± 0.04	0.09 ± 0.03
LR	62.22 ± 4.66	60.76 ± 1.87	60.90 ± 1.62	0.66 ± 0.02	0.14 ± 0.03
RF	74.13 ± 6.17	73.95 ± 1.41	73.97 ± 1.08	0.81 ± 0.03	0.30 ± 0.03
XGB	76.98 ± 4.98	71.13 ± 1.75	71.71 ± 1.51	0.82 ± 0.03	0.31 ± 0.03
SVC	63.81 ± 8.46	63.20 ± 3.35	63.26 ± 2.70	0.68 ± 0.04	0.16 ± 0.04

5.4.3 Webserver implementation

In this study, we developed a user-friendly webserver and standalone tool named as IL-13Pred, for screening of IL13 inducing peptides, which can be accessed using https://webs.iiitd.edu.in/raghava/il13pred. As shown in Figure 5.3, the webserver has featuring modules as Prediction, Peptide design, Protein scanning, and BLAST scanning. This platform is compatible with most of modern devices including mobile, tablets, desktop and laptops.

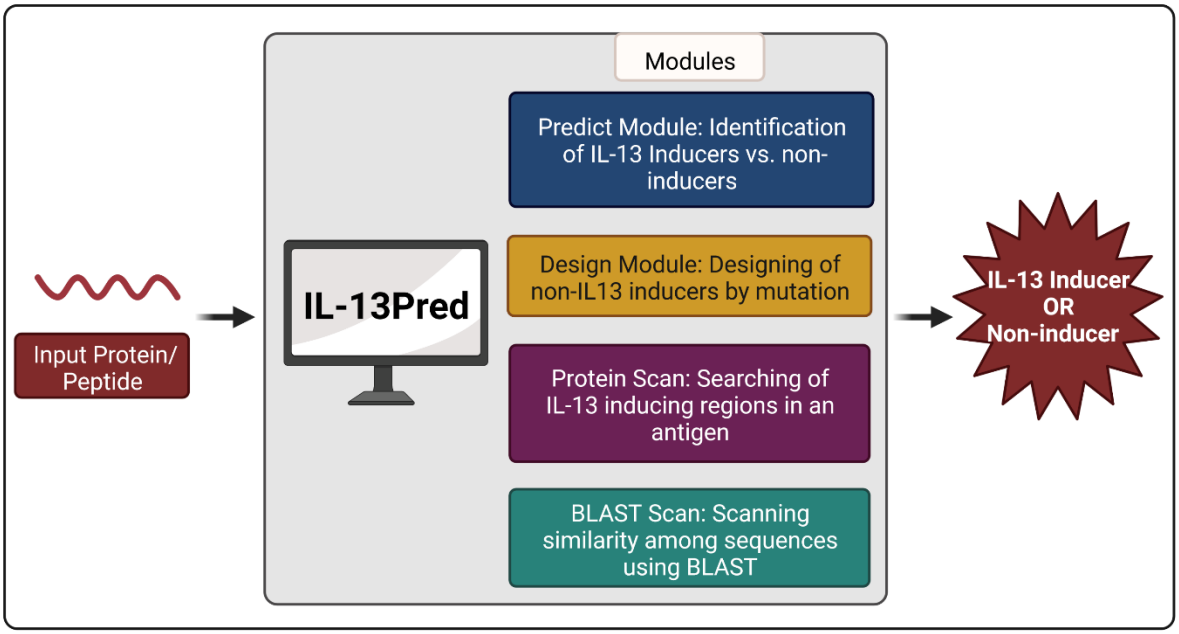


Figure 5.3: Overall architecture highlighting important webserver modules of IL13Pred.

5.5 Case studies

5.5.1 COVID-19 Analysis: Studies have shown elevated IL-13 levels in COVID-19 patients, as coronavirus proteins strongly induce this cytokine. Using IL-13Pred's protein scan module, we identified IL-13 inducers in SARS-CoV-2 proteins downloaded using NCBI of India, China, USA, Brazil, and Japan. As evident from Table 5.3, a total of 213 inducing peptides were found in the spike protein of the USA (California) strain. In addition, the IL-13 inducing/non-inducing peptides in other coronavirus proteins such as envelope protein, ORF6, ORF1ab, ORF3a, ORF7a/7b, ORF8, nucleocapsid phosphoprotein, ORF10 and membrane glycoprotein of USA strain were also screened. These insights would aid in vaccine designing against COVID-19.

Table 5.3: Prediction of IL-13 inducing peptides in five SARS-CoV-2 strains.

	INDIA (MZ340539.1)		**USA (MZ319836.1)**		**BRAZIL (MZ169911.1)**		**JAPAN (LC528233.2)**		**CHINA (MT291829.1)**	
	Ind	Non-Ind	Ind	Non-Ind	Ind	Non-Ind	Ind	Non-Ind	Ind	Non-Ind
ORF1ab	1288	5791	1288	5794	1294	5788	1484	5598	1484	5598
ORF1a	841	3547	834	3557	845	3546	955	3436	955	3436
Spike	215	1043	213	1046	213	1046	35	72	261	998
ORF3a	77	184	77	184	77	184	85	175	85	176
Envelop	7	54	7	54	7	54	10	51	10	51
Membrane	68	140	65	143	68	140	74	134	74	134
ORF6	25	22	25	22	25	22	27	20	27	20
ORF7a	31	76	31	76	31	76	35	72	35	72
ORF7b	17	12	17	12	17	12	17	12	17	12
ORF8	19	88	19	88	19	88	23	84	23	84
Nucleocapsid	56	349	55	350	59	346	66	339	66	339
ORF10	10	14	10	14	10	10	11	13	11	13

Note: *Ind means IL-13 inducers and Non-Ind means IL-13 non-inducers.*

5.5.2. Variant Analysis: In this study, we systematically analysed IL-13 inducing peptides across SARS-CoV-2 variants Alpha (B.1.1.7), Beta (B.1.351), and Delta (B.1.617.2). We obtained the reference spike protein sequence from NCBI and introduced variant-specific mutations reported by the CDC and Indian SARS-CoV-2 Genomics Consortium using an in-house script. As depicted in Table 4.4, the Alpha variant has seven substitutions, Beta has nine, and Delta also carries nine mutations. Using the "Protein Scan" module of IL-13Pred with default parameters, we predicted IL-13 inducing and non-inducing peptides in these mutated spike proteins. These functional alterations are dependent on the composition and location of the amino acids such as Arginine and Asparagine which are preferred in IL-13 inducing peptides.

Table 5.4: Changes in IL-13 inducing properties by mutation in Spike protein in SARS-CoV-2 variant strains.

SARS-Cov2 Variants	Mutation	Reference Peptide	Mutated Peptide	Score_R	Score_M	Prediction_R	Prediction_M
Alpha (B.1.1.7)	**A570D**	NKKFLPFQQFGRDIA	NKKFLPFQQFGRDID	0.05	0.11	IL-13 non-inducer	IL-13 inducer
		KFLPFQQFGRDIADT	KFLPFQQFGRDIDDT	0.03	0.06	IL-13 non-inducer	IL-13 inducer
	T716I	SNNSIAIPTNFTISV	SNNSIAIPINFTISV	0.04	0.06	IL-13 non-inducer	IL-13 inducer
	S980A	NFGAISSVLNDILSR	NFGAISSVLNDILAR	0.04	0.06	IL-13 non-inducer	IL-13 inducer
		GAISSVLNDILSRLD	GAISSVLNDILARLD	0.03	0.08	IL-13 non-inducer	IL-13 inducer
		VLNDILSRLDKVEAE	VLNDILARLDKVEAE	0.03	0.07	IL-13 non-inducer	IL-13 inducer
		LNDILSRLDKVEAEV	LNDILARLDKVEAEV	0.05	0.08	IL-13 non-inducer	IL-13 inducer
		NDILSRLDKVEAEVQ	NDILARLDKVEAEVQ	0.04	0.09	IL-13 non-inducer	IL-13 inducer
	D1118H	TQRNFYEPQIITTDN	TQRNFYEPQIITTHN	0.06	0.04	IL-13 inducer	IL-13 non-inducer
		QRNFYEPQIITTDNT	QRNFYEPQIITTHNT	0.07	0.03	IL-13 inducer	IL-13 non-inducer
		YEPQIITTDNTFVSG	YEPQIITTHNTFVSG	0.04	0.06	IL-13 non-inducer	IL-13 inducer
Beta (B.1.351)	**L18F**	LVLLPLVSSQCVNLT	LVLLPLVSSQCVNFT	0.06	0.04	IL-13 inducer	IL-13 non-inducer
		LLPLVSSQCVNLTTR	LLPLVSSQCVNFTTR	0.07	0.02	IL-13 inducer	IL-13 non-inducer
	D80A	VSGTNGTKRFDNPVL	VSGTNGTKRFANPVL	0.05	0.12	IL-13 non-inducer	IL-13 inducer
		GTNGTKRFDNPVLPF	GTNGTKRFANPVLPF	0.02	0.07	IL-13 non-inducer	IL-13 inducer
		SGTNGTKRFDNPVLP	SGTNGTKRFANPVLP	0.09	0.02	IL-13 inducer	IL-13 non-inducer
Delta (B.1.617.2)	**T19R**	LVLLPLVSSQCVNLT	LVLLPLVSSQCVNLR	0.06	0.04	IL-13 inducer	IL-13 non-inducer
	R158G	FRVYSSANNCTFEYV	FGVYSSANNCTFEYV	0.03	0.06	IL-13 non-inducer	IL-13 inducer
	P618R	TQTNSPRRARSVASQ	TQTNSRRRARSVASQ	0.02	0.08	IL-13 non-inducer	IL-13 inducer
		QTNSPRRARSVASQS	QTNSRRRARSVASQS	0.02	0.08	IL-13 non-inducer	IL-13 inducer
	D950N	DSLSSTASALGKLQD	DSLSSTASALGKLQN	0.05	0.06	IL-13 non-inducer	IL-13 inducer

Note: 1) Score_R and Score_M is the score of Reference Spike protein and score of mutated Spike protein respectively. 2) Prediction_R and Prediction_M is the IL-13 Inducing/non-inducing prediction of the Reference Spike protein and mutated Spike protein respectively.

5.6 Conclusion

IL-13Pred is the first computational method developed to predict IL-13-inducing peptides. Its evolution accommodates a significant gap in immunoinformatics providing a solid therapeutic and vaccine design tool. The IL-13 biological significance on disease progression and severity is supported by our results, and relevant to cancer, asthma, and COVID-19. We encourage the scientific community to use IL-13Pred in immunotherapeutic studies and in translational studies.

Chapter 6
Screening of NF-κB Inhibitors in Cancer

6.1 Introduction

Nuclear factor kappa B (NF-κB) is an important transcription factor that regulates genes involved in immune and inflammatory responses [170]. Since its identification in 1986, NF-κB has been known to be at the heart of the host defense to adverse stimuli, such as pathogens and damage to cells [171–174]. Dysregulated NF-κB signalling plays a crucial role in several disease progressions such as chronic inflammation disease, autoimmune disorders, and cancers [175–177]. Constant activation enhances inflammation, tissue injury, tumor growth, and metastasis via increasing cell survival and establishing a pro-tumor microenvironment [178–182]. Because of its centrality, NF-κB is a promising therapeutic target, with inhibitors ranging from small molecules to peptides [175,183–185]. Nevertheless, the process of developing drugs is expensive and time-consuming, which brings the necessity to enhance the process of finding the desired inhibitors with the help of computational approaches. This paper presents a novel in-silico predictor of NF-κB inhibitors, which we term "NFκBIn", using experimentally validated drugs. This tool deals with the shortfall of computational tools in predicting the inhibitors efficiently and accurately.

6.2 Background

NF-κB activation occurs via two main pathways: canonical and non-canonical, as shown in Figure 6.1 [186–188]. TNF-α and IL-1stimuli trigger the canonical pathway, which is marked by phosphorylation and degradation of IkB, which allows NF-κB to access the nucleus and transcription of genes that control inflammatory and immune reaction [170,189]. The non-canonical pathway, activated by receptors like CD40 and BAFF, relies on NIK-mediated processing of p100 into its active form (p52), which pairs with RelB to promote immune system development and adaptive immune responses [187].

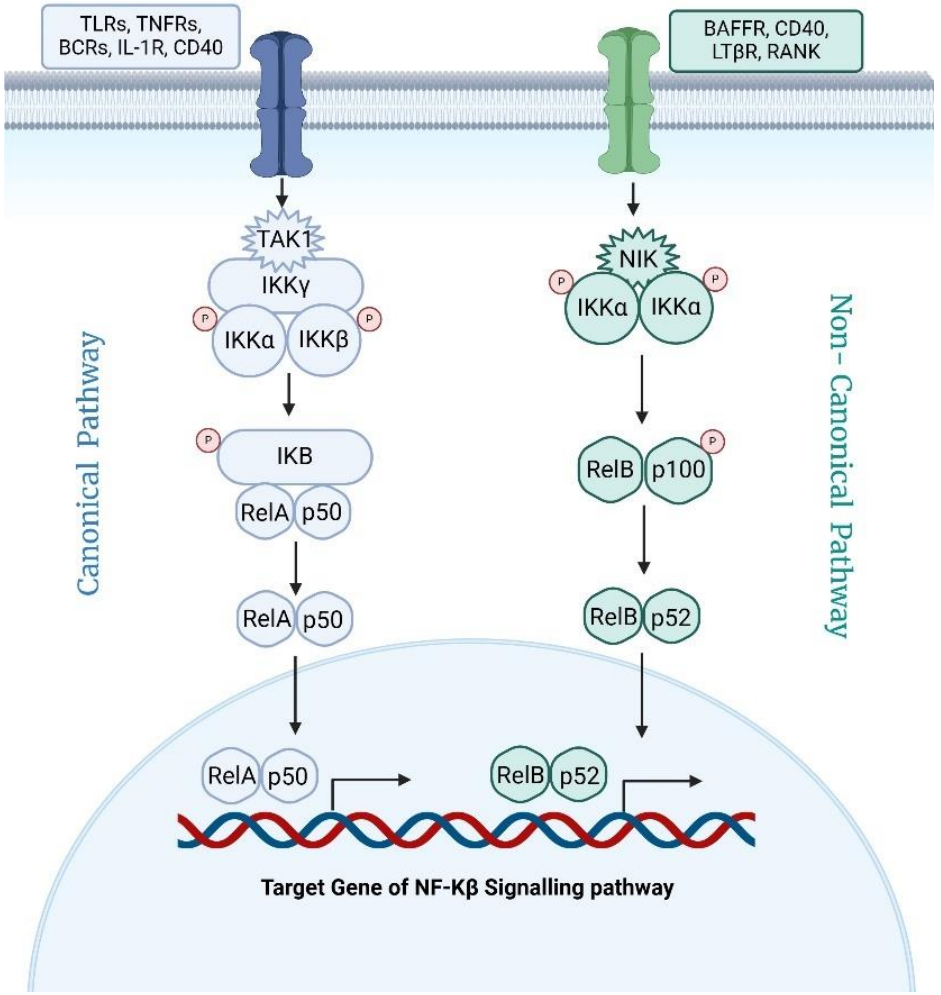


Figure 6.1: Illustration showing NF-κB activation pathways i.e. canonical and non-canonical pathways.

6.3 Materials and Methods

6.3.1 Dataset Collection and Pre-processing

We curated compounds from the PubChem database focusing on TNF-α-induced NF-κB inhibition assays [22]. An extensive search using keywords "(TNF AND NF-κB) inhibitors" yielded 90 bioassays, from which the high-throughput assay AID 1852 was selected for its specificity and data quality. This assay utilized HEK-293-T NF-κB-Luc cells and applied a tiered activity scoring system classifying compounds with >50% activity as active inhibitors. The dataset comprised 2,481 compounds: 1,149 inhibitors and 1,332 non-inhibitors. The dataset was spliced into 80% training (936 inhibitors, 1,048 non-inhibitors) and 20%

independent validation sets (213 inhibitors, 284 non-inhibitors) following established machine learning protocols.

6.3.2 Molecular Descriptors and Fingerprints features

Molecular descriptors represent chemical compounds quantitatively, capturing structural and physicochemical properties. We employed the PaDEL software to compute 17,967 descriptors per compound, including 1D, 2D, 3D descriptors, and various fingerprint types (binary vectors representing substructures) [190]. Descriptors with over 80% missing values were discarded, leaving 10,862 features. Data normalization was performed using z-score standardization to homogenize scale.

6.3.3 Feature Selection techniques

In order to develop a robust model for predicting NF-κB inhibitors and non-inhibitors we applied two feature selection techniques in this study. To reduce dimensionality, two feature selection strategies were implemented:

- **Correlation-based selection:** Low-variance features were eliminated followed by correlation analysis (threshold 0.6), yielding 2,365 descriptors. A support vector classifier with L1 regularization (SVC-L1) further refined this set to 383 features (32 2D, 3 3D, 348 fingerprints).
- **Univariate analysis:** Using Student's t-test, top 2,000 descriptors were selected based on significant p-value and univariate score. Subsequently SVC-L1 and Recursive Feature Elimination (RFE) selected 266 and 50 features, respectively.

6.3.4 Machine Learning Model Development

In this systematic attempt, of developing an in-silico tool for screening NF-κB inhibitors and non-inhibitors five classification machine learning algorithms were implemented via the Scikit-learn library. These were Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), eXtreme Gradient Boosting (XGB). Models were trained using five-fold cross-validation on the training set, with hyperparameter tuning for optimal performance. Best performing model on training dataset were then evaluated over independent dataset using parameters such as Sensitivity, Specificity, Accuracy, MCC and AUC. As depicted in Figure 6.2, the best performing model was implemented in NF- κB tool.

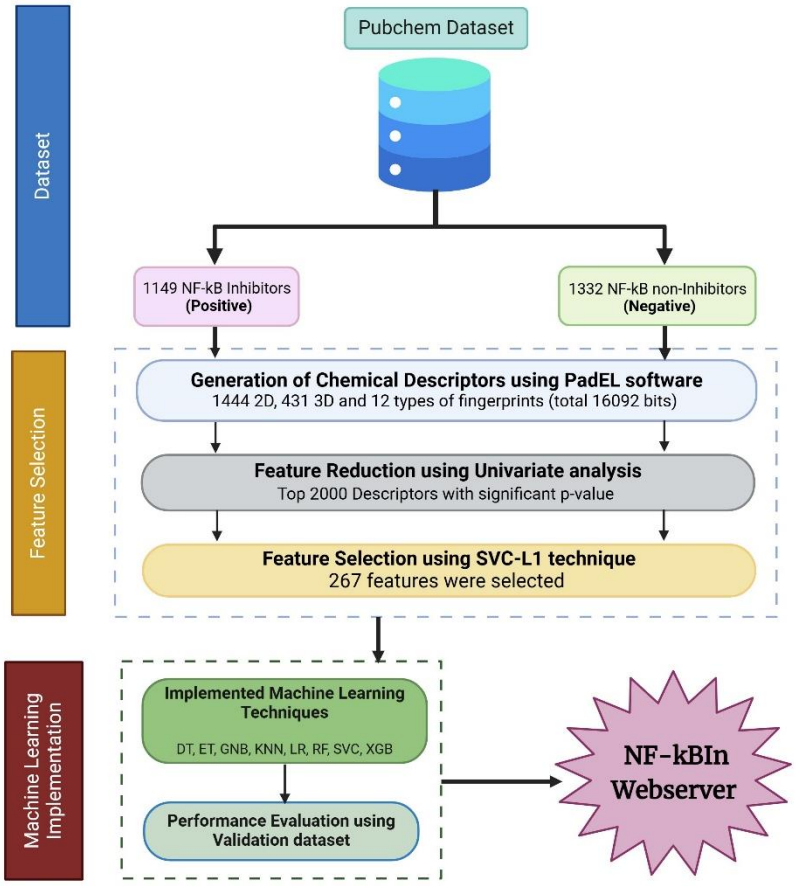


Figure 6.2: Workflow demonstrating NF-κBIn tool pipeline.

6.4 Results

6.4.1 Functional Group Analysis

In this study, to gain deeper insights into the structural features distinguishing NF-κB inhibitors from non-inhibitors, functional group analysis was performed using the RDKit[191] cheminformatics package. The frequency of various functional groups was computed in both classes, and the differences in their occurrence were analysed. As shown in Figure 6.3, bicyclic, Ar_N (aromatic nitrogen), pyridine, halogen, and allylic oxide groups were enriched in NF-κB inhibitors, whereas amide and aniline groups were more prevalent in non-inhibitors. Several other groups, including alcohols, alkyl halides, ethers, methoxy, and carbonyl-containing moieties, also exhibited differential distributions between the two classes. These findings suggest that specific structural motifs and functional groups may contribute to NF-κB inhibitory activity and could serve as useful features for distinguishing inhibitors from non-inhibitors.

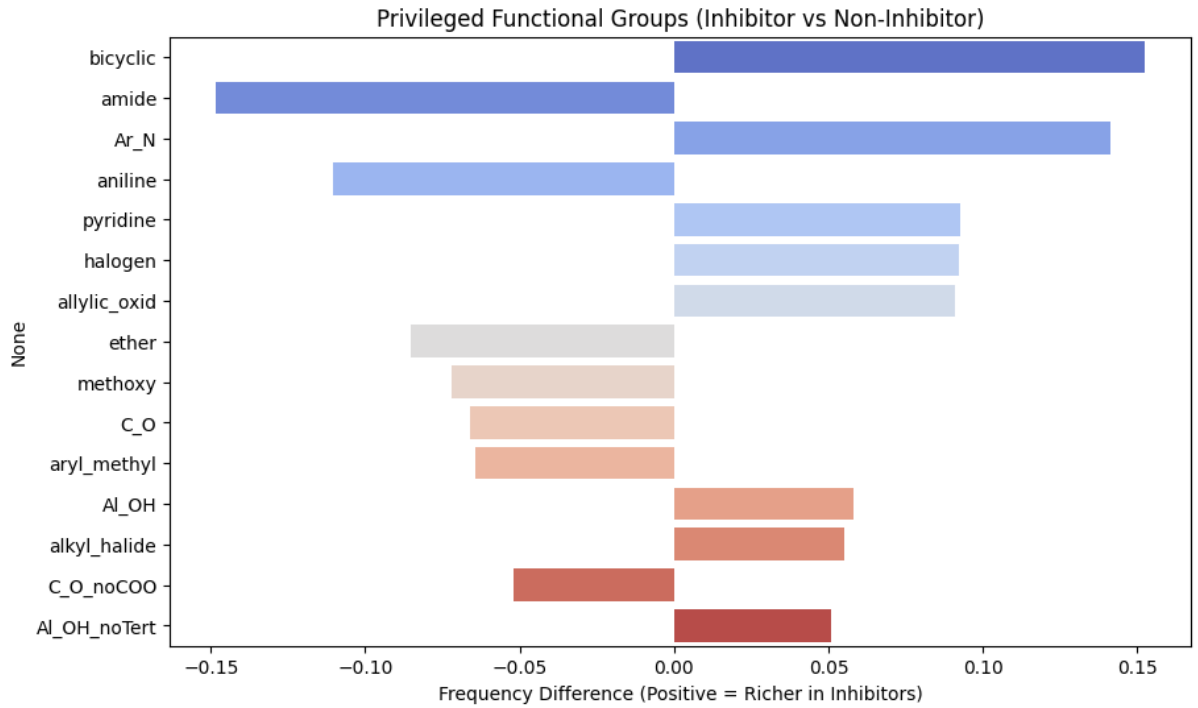


Figure 6.3: Functional group analysis for NF-κB pathway inhibitors and non-inhibitors using RDKit package.

6.4.2 ML Model Performance

6.4.2.1 Correlation-Based Descriptor Model Performance

In order to develop powerful prediction models for NF-κB inhibitors, we filtered out the highly correlated features and used the SVC-L1 feature selection method. This gave a smaller set of descriptors of 32 two-dimensional (2D), 3 three-dimensional (3D), and 348 fingerprint (FP) descriptors. On these characteristics, we created separate machine learning (ML) models, and an ensemble model as follows. Table 6.1 shows the results of the numerous ML classifiers.

I. ML Model based on 2D Descriptors:

We trained several classification models with the usage of 32 chosen 2D descriptors. K-Nearest Neighbors (KNN) classifier performed best in terms of AUC of 0.62 and accuracy of 64.85% on the validation set.

II. ML Model based on 3D Descriptors:

ML model was constructed on the 3 descriptors chosen as 3D had a modest predictive power. Random Forest classifier gave the best validation AUC of 0.56 and an accuracy of 56.14% on the validation dataset.

III. ML Model Based on Fingerprint Descriptors:

There were 348 fingerprint descriptors that used to implement several ML algorithms. Random Forest classifier outperformed others and recorded the highest AUC of 0.66 and accuracy of 66.40% on the validation data. The XGBoost classifier had a similar result with the AUC value of 0.66 and the accuracy of 65.79% on validation dataset.

IV. Ensemble-Based ML Model:

In this study, we have also developed an ensemble method to improve the accuracy of prediction, that is, an integrated method of the chosen 2D, 3D, and FP descriptors (a total of 383 descriptors). The ensemble model with a random forest classifier achieved the highest validation AUC of 0.67 and 67.20% accuracy to the validation dataset.

Table 6.1: Performance parameters of various ML classifiers developed using 32 2D, 3 3D, 348 Fingerprint and Ensemble approach over validation dataset.

2D Features								
Model	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Sens**	**Spec**	**AUC**	**MCC**
RF	65.05	0.65	0.49	0.74	0.49	0.56	0.61	0.24
DT	54.23	0.54	0.44	0.63	0.44	0.49	0.54	0.07
KNN	64.85	0.65	0.53	0.72	0.53	0.58	0.62	0.25
SVC	61.33	0.61	0.37	0.77	0.37	0.46	0.57	0.15
XGB	61.33	0.61	0.45	0.72	0.45	0.52	0.58	0.18
3D Features								
RF	56.14	0.57	0.49	0.64	0.49	0.52	0.56	0.12
DT	52.31	0.52	0.55	0.50	0.55	0.53	0.52	0.05
KNN	53.52	0.53	0.55	0.52	0.55	0.54	0.54	0.07
SVC	51.51	0.52	0.32	0.71	0.32	0.40	0.51	0.03
XGB	51.91	0.68	0.06	0.97	0.06	0.11	0.52	0.08
Fingerprint Features								
RF	66.40	0.74	0.51	0.82	0.51	0.60	0.66	0.34
DT	57.95	0.59	0.50	0.66	0.50	0.54	0.58	0.16
KNN	65.19	0.66	0.62	0.69	0.62	0.64	0.65	0.30
SVC	59.96	0.61	0.54	0.66	0.54	0.57	0.60	0.20
XGB	65.79	0.68	0.59	0.73	0.59	0.63	0.66	0.32
Ensemble method								

RF	67.20	0.73	0.53	0.81	0.53	0.62	0.67	0.36
DT	56.54	0.57	0.51	0.62	0.51	0.54	0.57	0.13
KNN	64.79	0.66	0.62	0.68	0.62	0.63	0.65	0.30
SVC	60.97	0.62	0.54	0.68	0.54	0.58	0.61	0.22
XGB	64.59	0.68	0.54	0.75	0.54	0.60	0.65	0.30

6.4.2.1 Univariate Analysis-Based Descriptor Model Performance

In an alternate strategy, we applied univariate analysis to screen the top 2000 (1000 positive and 1000 negative) descriptors. For the top 20 descriptors, we computed individual descriptor-based AUC scores and mean differences between positive and negative classes. Notably, KRFP605 achieved the highest AUC of 0.62 and a mean difference of 2.60, refer Table 6.2.

Table 6.2: List of top 20 descriptors with highest AUC and their mean difference among inhibitors and non-inhibitors.

Descriptors	Type	Mean_Pos	Mean_Neg	Mean Diff	Sens	Spec	AUC
KRFP605	FP	-0.140	0.121	-0.260	0.810	0.447	0.628
GATS1i	2D	-0.138	0.119	-0.256	0.613	0.644	0.628
ExtFP627	FP	-0.134	0.115	-0.249	0.775	0.477	0.626
MACCSFP86	FP	-0.132	0.114	-0.246	0.683	0.568	0.626
PubchemFP615	FP	-0.130	0.112	-0.242	0.718	0.530	0.624
RDF20u	3D	-0.130	0.112	-0.242	0.697	0.545	0.621
RDF20i	3D	-0.129	0.112	-0.241	0.704	0.538	0.621
RDF20e	3D	-0.129	0.112	-0.241	0.655	0.583	0.619
ExtFP129	FP	-0.126	0.109	-0.235	0.556	0.682	0.619
MACCSFP138	FP	-0.126	0.108	-0.234	0.662	0.576	0.619
ATSC1i	2D	0.139	-0.120	0.259	0.676	0.553	0.615
MATS1i	2D	0.139	-0.120	0.258	0.528	0.697	0.613
GraphFP786	FP	0.134	-0.116	0.250	0.754	0.470	0.612
AATSC1i	2D	0.129	-0.111	0.241	0.563	0.659	0.611
GraphFP171	FP	0.128	-0.111	0.239	0.711	0.508	0.609
FP584	FP	0.123	-0.106	0.229	0.535	0.682	0.609
FP964	FP	0.122	-0.105	0.228	0.535	0.682	0.609
AATS3v	2D	0.116	-0.100	0.217	0.789	0.424	0.606
FP785	FP	0.116	-0.100	0.216	0.782	0.424	0.603
FP282	FP	0.116	-0.100	0.216	0.585	0.621	0.603

Further, we applied SVC-L1 and Recursive Feature Elimination (RFE) on the top 2000 descriptors. The SVC-L1 approach selected 266 descriptors, while RFE yielded a refined subset of 50 features. As shown in Table 6.3, a Support Vector Classifier (SVC) trained on the 266 SVC-L1-selected features outperformed all models, achieving an AUC of 0.80 on the

training and 0.75 on the validation set. Meanwhile, the KNN model using the 50 RFE-selected descriptors achieved a training AUC of 0.66 and validation AUC of 0.65.

Table 6.3: Machine learning models performance developed using SVC-L1 and RFE based feature selection techniques over Univariate analysis-based descriptor selection.

Model	Accuracy	FPR	Kappa	F1-score	Sens	Spec	AUC	MCC
SVC L1 feature selection technique								
RF	65.59	29.60	0.31	0.64	60.73	70.40	0.71	0.31
DT	60.36	33.60	0.21	0.58	54.25	66.40	0.63	0.21
KNN	65.19	30.40	0.30	0.63	60.73	69.60	0.70	0.31
SVC	**67.61**	**28.00**	**0.35**	**0.66**	**63.16**	**72.00**	**0.75**	**0.35**
XGB	62.78	34.00	0.26	0.61	59.51	66.00	0.69	0.26
RFE feature selection technique								
RF	62.98	35.60	0.26	0.62	61.54	64.40	0.66	0.26
DT	57.55	42.40	0.15	0.57	57.49	57.60	0.59	0.15
KNN	62.17	38.00	0.24	0.62	62.35	62.00	0.65	0.24
SVC	59.56	38.00	0.19	0.58	57.09	62.00	0.64	0.19
XGB	57.95	36.40	0.16	0.55	52.23	63.60	0.62	0.16

6.4.3 Drug Repurposing Screening

Using NFκBIn standalone version prediction module, 2,577 FDA-approved drugs (from DrugBank) were screened for potential NF-κB inhibitory activity. Top candidates included Bleomycin, Nitroprusside, Guanidine, Ivermectin, Tobramycin, Pentosan polysulfate, and Gentamicin, many of which have literature support for NF-κB pathway inhibition, refer Table 6.4.

Table 6.4: List of seven FDA-approved drug candidates as potential Nf-κB pathway inhibitors.

DrugBank ID	FDA approved Drugs	Prediction Label	Literature Remarks
DB00290	Bleomycin	Inhibitor	Co-administration of ginsenoside with BLM resulted in marked improvement in lung structure and a significant reduction in Nf-κB expression. [192]
DB00325	Nitroprusside	Inhibitor	Nitroprusside is a vasodilator that releases nitric oxide (NO) upon metabolism. NO can influence various signalling pathways, including the Nf-κB pathway. [193]
DB00536	Guanidine	Inhibitor	The guanidine compound ME10092 inhibits Nf-κB activation and the upregulation of inflammatory mediators in vivo. [194]
DB00602	Ivermectin	Inhibitor	Ivermectin may inhibit LPS-induced production of inflammatory cytokines by blocking NF-kB pathway and improve LPS-induced

			survival in mice. FDA-approved antiparasitic drug, could potentially be used in combination with chemotherapeutic agents to treat cancers. [195–197]
DB00684	Tobramycin	Inhibitor	Tobramycin suppresses Nf-κB activation, reducing pro-inflammatory cytokines and controlling excessive inflammation in lung infections, thus helping prevent further lung damage in conditions like cystic fibrosis. [198,199]
DB00686	Pentosan Polysulfate	Inhibitor	Pentosan polysulfate sodium (PPS), an inhibitor of NF-kB activation. [200,201]

6.4.4 Webserver and Standalone Package

In order to facilitate the scientific community access, NFκBIn was implemented as a webserver and standalone package. The webserver supports high-throughput screening with modules for prediction, chemical structure drawing, and analog design. The platform is accessible at https://webs.iiitd.edu.in/raghava/nfkbin/.

6.5 Conclusion

Nuclear factor kappa B (NF-κB) plays a central role in regulating immune and inflammatory responses and is critically involved in the pathogenesis of several chronic diseases, including cancer and autoimmune disorders. Traditional drug discovery approaches, such as molecular docking, often lack the ability to capture complex biological interactions, limiting their success in identifying potent NF-κB inhibitors. Machine learning has the potential to become a potent alternative in this regard, since the diverse molecular descriptors can be investigated to more accurately predict the inhibitor potential. In this study, we present a machine learning-powered predictive model of NF-κB inhibitors induced by TNF-α, called NFκBIn. The platform also incorporates 2D, 3D and fingerprint descriptors with enhanced features selection algorithms, such as SVC-L1 and RFE. The ensemble approach was the most accurate in prediction of all the models that were developed, so the synergistic importance of multiple types of descriptors cannot be underestimated.

Notably, the model could recognize FDA approved drugs that had known NF-κB inhibitive activity, which is a contributing factor to its possible drug repurposing method. NFκBIn is also offered as a webserver and as an independent tool so that, enabling users to effectively screen larger compound libraries and create new analogs within several minutes. This resource helps in accelerated cost-efficient discovery of drugs against NF-κB in cancer and inflammation.

Chapter 7
Summary

7.1 Overview and Rationale of the Thesis

Oncogenic diseases are still one of the most challenging and heterogeneous diseases with genetic mutations, immune regulation, and dys-adjustment of signalling pathways resulting in uncontrolled cellular proliferation. Although there has been significant progress in experimental and clinical oncology, early diagnosis and therapeutic intervention at high specificity remain very challenging. The traditional method of revealing biomarkers and therapeutics is usually time-consuming, expensive, and constrained by the viability of experiments. Computational and machine learning-based methods, in this case, provide transformative opportunities to expedite the processes of discovery, combine vast amounts of biological data, and inform the rational construction of diagnostic and therapeutic options.

The current thesis project is called Development of Biomolecule-Based In-Silico Models of Cancer Diagnostics and Therapeutics and is intended to create, test, and apply effective computational models of predicting cancer biomarkers and therapeutic candidates based on publicly available genomic, proteomic, and chemical data. The thesis combines multiple sources of biological data, such as transcriptomic and peptide sequencing and chemical descriptors, with state-of-the-art feature engineering and machine learning algorithms. It is this work that leads to the development of four distinct, although interrelated, computational models that cover a key aspect of cancer biology each, which are biomarker discovery, predictive therapeutic peptides, regulatory cytokine and screens of small molecule inhibitors. Taken together, these models help to create a holistic, modular pipeline to discover biomarkers and translate them into therapeutic programs in the field of oncology.

7.2 Major Scientific Contributions of the Thesis

The initial section of this thesis focuses on the discovery of biomarkers in prostate cancer which is one of the most common malignancies in males all over the world. With the assistance of gene expression data on The Cancer Genome Atlas (TCGA-PRAD) cohort, an in-silico pipeline was constructed to discover that could be used to distinguish tumor samples and normal tissue using discriminative gene signatures. To construct the predictive models to be used in the classification, a combination of statistical filtering, feature selection, and supervised learning was used. To increase the strength and biological relevance of the model, a novel feature selection method was suggested to be based on the propensity index. The discriminative power of each gene is quantitatively assessed by this approach and this approach resulted in the discovery of a small ten-gene biomarker panel, which showed a high predictive validity. The

last classification model has been in excellent performance in cross-validation as well as in independent testing which indicates that it is generalizable. The identified genes are associated with the previously known prostate cancer-related pathways related to androgen signalling, apoptosis, and cellular proliferation, which supports their biological validity. This paper illustrates how integrative methods of transcriptomic and computational analysis could be used to better come up with clinically useful biomarkers that can be used as an addition or a supplement to the current diagnostic techniques including the PSA (KLK3) test. The resultant gene panel offers a bright future on non-invasive prostate cancer screening and risk stratification.

The second part of the thesis highlights the relevance of therapeutic peptide by creation of a manually curated repository of therapeutic peptides and proteins (ThPdb2) and a predictive framework, ThPPred, that would identify potential therapeutic peptides and proteins, using global and sequence-derived feature engineering. The new version of ThPdb is called ThPdb2 and is an improvement of its predecessor (ThPdb, 2017). It lists 894 approved therapeutic peptides and proteins approved by FDA and more than 5,500 records containing physicochemical properties, pharmacological categories, routes of administration and molecular annotations. The data curation procedure combined several open sources and hand checked the records to verify the accuracy and consistency. To supplement this repository, ThPPred model was designed as classification system by using machine learning that was able to differentiate between therapeutic and non-therapeutic peptides/proteins in water. The predictor makes use of sequence compilation characteristics, motif insight, and pattern signified depictors to educate ensemble classifiers. The model was strongly supported by deep cross-validation and external testing which confirmed the efficiency and predictability of the model. ThPdb2 and ThPPred fill an important gap in the peptide therapeutics field by offering a bi-fold platform-data repository system, prediction system-that is open to the scientific community through a webserver and GitHub repository. This framework simplifies the process of discovering and ranking of peptides as drug discovery, thus saving time and cost involved in the experimental screening.

The third section of this work is based on immune modulation which targets Interleukin-13 (IL-13) which is a cytokine in the development of cancers, allergic responses as well as chronic inflammation. The IL-13 dysregulation affects the tumor microenvironment and immune evasion processes; therefore, IL-13 signalling is an effective therapeutic target. A predictive IL-13 in-silico model IL13Pred was designed based on experimentally verified data of Immune

Epitope Database (IEDB). The data was comprised of experimentally validated IL-13-inducing, and non-inducing peptides of human hosts. Compositional and positional analysis showed that there were specific preferences of the amino acids in strategic positions, which informed the design of features and model formation. Several machine learning models such as the Random Forest, Extra Trees, and Support Vector Machine have been tested. The resultant best model was very precise and recalling, which was confirmed by a rigorous cross-validation and independent testing. Besides predictive modelling, the motif analysis with MERCI detected specific sequence motifs that IL-13 inducers share, which makes the results easier to interpret. The IL13Pred webserver provides the ability to screen peptides based on their ability to cause IL-13, which is then used in vaccine design, immunotherapy and the study of cytokine modulation. The model was applied in screen immunogenic peptides of the SARS-CoV-2 proteome as a case study, and it proved to be useful in practice with the identification of cytokine-associated epitopes in the context of viral pathogenesis and immunogenicity.

The last part of this thesis aims at exploring small molecule screening which determines the potential inhibitors of the Nuclear Factor Kappa B (NF-κB) which is a master transcription factor that controls the genes related to inflammation, immunity and oncogenesis. NF-κB is an aberrant induced activation in many cancers, including chronic inflammatory-driven cancers. The paper presents NFκBIn, a machine learning-based system that can be used to screen and rank NF-κB inhibitors virtually by using experimentally valid experimental data on PubChem BioAssays. The data set included small molecules that were labeled on the basis of their inhibitory property against the action of NF-κB triggered by TNF-a. Molecular descriptors and fingerprints were calculated with PaDEL software and different strategies of feature selection were employed to narrow down on the set of descriptors. A number of classifiers were trained and benchmarked, such as the Random Forest, Logistic Regression, and XGBoost. The optimized model had high accuracy and AUC values and was highly discriminative, which makes it a reliable model in terms of the use of virtual screening.

There was also independent test set and external case study validation of the performance, which has the potential to fast-track early-stage drug discovery. The NFκBIn system is a computing service enabling the screening of large chemical libraries in a short period of time to discover hits as NF-κB inhibitors and which is a complement to high-throughput experiments.

7.3 Integrated Contributions and Applications

Combined, the four computational models that are introduced in the present thesis form a unified and scalable pipeline to cancer research, incorporating multi-domain biological data, i.e. gene expression to peptides and small molecules. All modules are independent, but used as a concept with each other to fill in the gap between raw biological data and applications in translational research. The prostate cancer prognosis model expands diagnostics using transcriptomic profiling; ThPdb2 and ThPPred are used to screen and identify therapeutic peptides, IL13Pred is used to screen and discover immunomodulatory peptides, and NFκBIn is used to assist in rational small molecule inhibitors. All these tools are illustrations of machine learning and data-based modelling in revolutionizing the systems biology of cancer and personalized medicine.

The uses of the tools that have been created in this paper are set to be spread in various fields within the biomedical and pharmaceutical research. Transcriptomic data in public repositories (including TCGA and GEO) can be used with the prostate cancer biomarker model to diagnose, sub-type, and predict prognosis of other types of cancer early in life. The ThPPred platform offers a feasible platform where peptide libraries are screened in-silico to find therapeutically relevant peptides, and as such, it adds to peptide-based drug design and repurposing. Equally IL13Pred can be used to identify peptides that can regulate the response of cytokines, which is applicable to the design of vaccines, immunotherapy, and insight into tumor microenvironments mediated by cytokines. In contrast, the NFκBIn tool is useful in high-throughput screening of small molecules to determine the NF-κB inhibitors, which is a valuable part of anti-inflammatory and anticancer drug discovery programs. In addition, the computational models and web-based tools are useful educational and research resources, which will encourage the use of the machine learning techniques in the circles of biological and pharmaceutical research.

7.4 Future Scope and Concluding Remarks

The future of this study falls in a number of directions that would elaborate and improve the derived computational frameworks. A potential direction of improvement is the combination of multi-omics data, including proteomic, epigenomic, and metabolomic data to create a more detailed picture of cancer biology. Deep learning architectures, including convolutional, recurrent, and graph neural networks, may also be the key to improving the predictive model accuracy and generalization, especially in situations involving intricate biological interactions

and 3D molecular structures. The machine learning model can be combined with structural bioinformatics methods such as molecular dynamics simulations and molecular docking to confirm the accuracy of predicted peptide and ligand-binding effects at the atomic scale, enhancing the accuracy of in-silico predictions. The other important direction is to make these predictive frameworks applicable in personalized medicine, where a personalized approach of identifying biomarkers and therapeutic agents can be done based on the individual needs of a patient, using personalized molecular profiles. Last, it is possible to increase partnerships with experimental laboratories to empirically validate the computational predictions which would significantly increase the translational relevance of such findings and reduce the gap between in-silico discovery and clinical application.

Overall, this thesis illustrates how computational modelling with machine learning can be used to speed up the discovery of biomarkers and the design of therapeutics in the field of oncology. Four independent but interconnecting tools were created and proven using the systematic combination of biological information and algorithmic innovation: transcriptomic biomarkers, peptide therapeutics, immune modulators, and small molecule inhibitors. The contributions do not only develop the discipline of computational cancer biology, but also make the resources available to the rest of the scientific community. The frameworks developed herein are the foundation of the future of explainable artificial intelligence, systems immunology, and precision cancer therapeutics, which is a step in the direction of data-driven, personalized medicine.

List of Publications and Book Chapters

➢ *Thesis Related Publications*

Peer Reviewed Journals

1. Jain S, Tomer R, Patiyal S, Raghava GPS. NfκBin: a machine learning based method for screening TNF-α induced NF-κB inhibitors. *Front Bioinforma*. 2025;5. doi:10.3389/FBINF.2025.1573744. Pub Link
2. Jain, S., Gupta, S., Patiyal, S., & Raghava, G. P. S. (2024). THPdb2: compilation of FDA approved therapeutic peptides and proteins. Drug Discovery Today, 29(7). Pub Link
3. Jain, S., Dhall, A., Patiyal, S., & Raghava, G. P. S. (2022). IL13Pred: A method for predicting immunoregulatory cytokine IL-13 inducing peptides. Computers in Biology and Medicine, 143. Pub Link
4. Jain, S., Malhotra, K. P. K., Patiyal, S., & Raghava, G. P. S. (2023). A Highly Accurate Model for Screening Prostate Cancer Using Propensity Index Panel of Ten Genes. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology, 30(12), 1305–1314. Pub Link

Preprints and Paper under review/preparation

1. Jain S, Gupta S, Raghava GPS, Raghava GPS. An ensemble method for predicting and designing of druggable proteins. doi:10.1101/2024.10.17.618792. Preprint Link.

❖ *Book Chapters*

1. Jain, S., Dhall, A., Patiyal, S., & Raghava, G. P. S. (2023). In Silico Tool for Identification, Designing, and Searching of IL13-Inducing Peptides in Antigens. Methods in Molecular Biology, 2673, 329–338. Book Chapter Link
2. Dhall, A., Jain, S., Sharma, N., Naorem, L. D., Kaur, D., Patiyal, S., & Raghava, G. P. S. (2022). In silico tools and databases for designing cancer immunotherapy. Advances in Protein Chemistry and Structural Biology, 129, 1–50. Book Chapter Link

❖ ***Other Publications***

Peer Reviewed Journals

1. Tomer R, Jain S, Gahlot PS, Bajiya N, Raghava GPS. In-silico tool for predicting and scanning rheumatoid arthritis-inducing peptides in an antigen. Front Immunol. 2025;16. doi:10.3389/FIMMU.2025.1630863. Pub Link
2. 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. (2023). Pfeature: A Tool for Computing Wide Range of Protein Features and Building Prediction Models. Journal of Computational Biology, 30(2), 204–222. Pub Link
3. Dhall, A., Patiyal, S., Choudhury, S., Jain, S., Narang, K., & Raghava, G. P. S. (2023). TNFepitope: A webserver for the prediction of TNF-α inducing epitopes. Computers in Biology and Medicine, 160. Pub Link
4. Sharma, N., Naorem, L. D., Jain, S., & Raghava, G. P. S. (2022). ToxinPred2: an improved method for predicting toxicity of proteins. Briefings in Bioinformatics, 23(5). Pub Link
5. Gupta, S., Sharma, N., Naorem, L. D., Jain, S., & Raghava, G. P. S. (2022). Collection, compilation and analysis of bacterial vaccines. Computers in Biology and Medicine, 149. Pub Link

Preprints and Paper under review/preparation

1. Mathur, M., Patiyal, S., Dhall, A., Jain, S., Tomer, R., Arora, A., Raghava, G. P. S. Nfeature: A platform for computing features of nucleotide sequences. doi:10.1101/2021.12.14.472723 Preprint Link
2. Tomer, R., Mehta, N. K., Malik, S., Jain, S., & Raghava, G. P. S. IL4Pred2: Prediction of Interleukin-4 Inducing Peptides in Human and Mouse 1. https://doi.org/10.1101/2025.04.23.650150 Preprint Link
3. Mathur, K., Jain, S., Bajiya, N., Kumar, N., & Raghava, G. P. S. In silico tool for identification of colorectal cancer from cell-free DNA biomarkers. https://orcid.org/0000-0002-7045-5188 Preprint Link

URL of Computational Resources

Following resource have been developed during this study.

Name	Webserver	Dataset	Standalone
ThPDB2	*thpdb2	-	-
ThPPred	*thppred	*thppred/download.php	#thppred
IL13Pred	*il13pred	*il13pred/dataset.php	#il13pred
NFκBIn	*nfkbin	*nfkbin/dataset.php	#nfkbin

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

Bibliography

1. Brown JS, Amend SR, Austin RH, Gatenby RA, Hammarlund EU, Pienta KJ. Updating the Definition of Cancer. *Mol Cancer Res*. 2023;21(11):1142-1147. doi:10.1158/1541-7786.MCR-23-0411,

2. Upadhyay A. Cancer: An unknown territory; rethinking before going ahead. *Genes Dis*. 2021;8(5):655-661. doi:10.1016/j.gendis.2020.09.002

3. Anand U, Dey A, Chandel AKS, et al. Cancer chemotherapy and beyond: Current status, drug candidates, associated risks and progress in targeted therapeutics. *Genes Dis*. 2023;10(4):1367-1401. doi:10.1016/j.gendis.2022.02.007

4. Liu Y, Zhang Y, Li H, Hu TY. Recent advances in the bench-to-bedside translation of cancer nanomedicines. *Acta Pharm Sin B*. 2025;15(1). doi:10.1016/J.APSB.2024.12.007,

5. Ingham J, Ruan JL, Coelho MA. Breaking barriers: we need a multidisciplinary approach to tackle cancer drug resistance. *BJC reports*. 2025;3(1):11. doi:10.1038/S44276-025-00129-2,

6. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. *Drug Discov Today*. 2021;26(1):80-93. doi:10.1016/j.drudis.2020.10.010

7. Jarallah SJ, Almughem FA, Alhumaid NK, et al. Artificial intelligence revolution in drug discovery: A paradigm shift in pharmaceutical innovation. *Int J Pharm*. 2025;680. doi:10.1016/J.IJPHARM.2025.125789,

8. Passaro A, Al Bakir M, Hamilton EG, et al. Cancer biomarkers: Emerging trends and clinical implications for personalized treatment. *Cell*. 2024;187(7):1617-1635. doi:10.1016/j.cell.2024.02.041

9. Califf RM. Biomarker definitions and their applications. *Exp Biol Med*. 2018;243(3):213-221. doi:10.1177/1535370217750088,

10. Zhou Y, Tao L, Qiu J, et al. Tumor biomarkers for diagnosis, prognosis and targeted therapy. *Signal Transduct Target Ther*. 2024;9(1). doi:10.1038/S41392-024-01823-2,

11. Yang Y, Adelstein SJ, Kassis AI. Target discovery from data mining approaches. *Drug Discov Today*. 2012;17(SUPPL.). doi:10.1016/j.drudis.2011.12.006

12. Melandri G, R-Radohery G, Beaumont C, et al. Artificial intelligence: the human response to approach the complexity of big data in biology. *Gigascience*. 2025;14. doi:10.1093/GIGASCIENCE/GIAF057

13. Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, et al. A review on machine learning approaches and trends in drug discovery. *Comput Struct Biotechnol J*. 2021;19:4538-4558. doi:10.1016/j.csbj.2021.08.011

14. Rehman AU, Li M, Wu B, et al. Role of Artificial Intelligence in Revolutionizing Drug Discovery. *Fundam Res*. 2024;5(3). doi:10.1016/j.fmre.2024.04.021

15. You Y, Lai X, Pan Y, et al. Artificial intelligence in cancer target identification and drug discovery. *Signal Transduct Target Ther*. 2022;7(1). doi:10.1038/S41392-022-00994-0,

16. Goles M, Daza A, Cabas-Mora G, et al. Peptide-based drug discovery through artificial intelligence: towards an autonomous design of therapeutic peptides. *Brief Bioinform*. 2024;25(4). doi:10.1093/BIB/BBAE275,

17. Wang Z, Jensen MA, Zenklusen JC. A practical guide to The Cancer Genome Atlas (TCGA). *Methods Mol Biol*. 2016;1418:111-141. doi:10.1007/978-1-4939-3578-9_6,

18. Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol*. 2006;411:352-369. doi:10.1016/S0076-6879(06)11019-8

19. Vita R, Blazeska N, Marrama D, et al. The Immune Epitope Database (IEDB): 2024 update. *Nucleic Acids Res*. 2025;53(D1):D436-D443. doi:10.1093/NAR/GKAE1092,

20. Rhizobium G promoting E. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res*. 2025;53(D1):13-14. doi:10.1093/NAR/GKAE1010

21. Knox C, Wilson M, Klinger CM, et al. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Res*. 2024;52(D1):D1265-D1275. doi:10.1093/NAR/GKAD976,

22. Wang Y, Xiao J, Suzek TO, et al. PubChem's BioAssay database. *Nucleic Acids Res*. 2012;40(D1). doi:10.1093/NAR/GKR1132,

23. Jeon S, Jeon Y, Lim JY, Kim Y, Cha B, Kim W. Emerging regulatory mechanisms and functions of biomolecular condensates: implications for therapeutic targets. *Signal*

Transduct Target Ther. 2025;10(1). doi:10.1038/S41392-024-02070-1

24. Miller RA, Kutmon M, Bohler A, Waagmeester A, Evelo CT, Willighagen EL. Understanding signaling and metabolic paths using semantified and harmonized information about biological interactions. *PLoS One*. 2022;17(4). doi:10.1371/JOURNAL.PONE.0263057

25. Turlej E, Domaradzka A, Radzka J, Drulis-Fajdasz D, Kulbacka J, Gizak A. Cross-Talk Between Cancer and Its Cellular Environment-A Role in Cancer Progression. *Cells*. 2025;14(6). doi:10.3390/CELLS14060403

26. Albonici L, Giganti MG, Modesti A, Manzari V, Bei R. Multifaceted Role of the Placental Growth Factor (PlGF) in the Antitumor Immune Response and Cancer Progression. *Int J Mol Sci*. 2019;20(12). doi:10.3390/IJMS20122970

27. Zhao H, Wu L, Yan G, et al. Inflammation and tumor progression: signaling pathways and targeted intervention. *Signal Transduct Target Ther*. 2021;6(1). doi:10.1038/S41392-021-00658-5

28. Hargadon KM. Genetic dysregulation of immunologic and oncogenic signaling pathways associated with tumor-intrinsic immune resistance: a molecular basis for combination targeted therapy-immunotherapy for cancer. *Cell Mol Life Sci*. 2023;80(2). doi:10.1007/S00018-023-04689-9

29. Tonello S, Rolla R, Tillio PA, Sainaghi PP, Colangelo D. Microenvironment and Tumor Heterogeneity as Pharmacological Targets in Precision Oncology. *Pharmaceuticals (Basel)*. 2025;18(6). doi:10.3390/PH18060915

30. You M, Xie Z, Zhang N, et al. Signaling pathways in cancer metabolism: mechanisms and therapeutic targets. *Signal Transduct Target Ther*. 2023;8(1). doi:10.1038/S41392-023-01442-3

31. Fu D, Hu Z, Xu X, Dai X, Liu Z. Key signal transduction pathways and crosstalk in cancer: Biological and therapeutic opportunities. *Transl Oncol*. 2022;26. doi:10.1016/j.tranon.2022.101510

32. Huang S, Zhu J, Yu L, Huang Y, Hu Y. Cancer-nervous system crosstalk: from biological mechanism to therapeutic opportunities. *Mol Cancer*. 2025;24(1). doi:10.1186/S12943-025-02336-4

33. Rehman AU, Li M, Wu B, et al. Role of artificial intelligence in revolutionizing drug discovery. *Fundam Res*. 2025;5(3):1273-1287. doi:10.1016/j.fmre.2024.04.021

34. Agamah FE, Mazandu GK, Hassan R, et al. Computational/in silico methods in drug target and lead prediction. *Brief Bioinform*. 2020;21(5):1663-1675. doi:10.1093/BIB/BBZ103

35. Sharma A, Lysenko A, Jia S, Boroevich KA, Tsunoda T. Advances in AI and machine learning for predictive medicine. *J Hum Genet*. 2024;69(10):487-497. doi:10.1038/S10038-024-01231-Y

36. Zhang Y, Liu C, Liu M, et al. Attention is all you need: utilizing attention in AI-enabled drug discovery. *Brief Bioinform*. 2023;25(1). doi:10.1093/BIB/BBAD467

37. AlDoughaim M, AlSuhebany N, AlZahrani M, et al. Cancer Biomarkers and Precision Oncology: A Review of Recent Trends and Innovations. *Clin Med Insights Oncol*. 2024;18. doi:10.1177/11795549241298541

38. Das S, Dey MK, Devireddy R, Gartia MR. Biomarkers in Cancer Detection, Diagnosis, and Prognosis. *Sensors (Basel)*. 2023;24(1). doi:10.3390/S24010037

39. Quraish R ul, Hirahata T, Quraish A ul, ul Quraish S. An Overview: Genetic Tumor Markers for Early Detection and Current Gene Therapy Strategies. *Cancer Inform*. 2023;22. doi:10.1177/11769351221150772

40. Hessels D, Klein Gunnewiek JMT, Van Oort I, et al. DD3PCA3-based molecular urine analysis for the diagnosis of prostate cancer. *Eur Urol*. 2003;44(1):8-16. doi:10.1016/S0302-2838(03)00201-X

41. Laxman B, Tomlins SA, Mehra R, et al. Noninvasive detection of TMPRSS2:ERG fusion transcripts in the urine of men with prostate cancer. *Neoplasia*. 2006;8(10):885-888. doi:10.1593/neo.06625

42. Tomlins SA, Bjartell A, Chinnaiyan AM, et al. ETS Gene Fusions in Prostate Cancer: From Discovery to Daily Clinical Practice. *Eur Urol*. 2009;56(2):275-286. doi:10.1016/j.eururo.2009.04.036

43. Barbieri CE, Baca SC, Lawrence MS, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat Genet*. 2012;44(6):685-689. doi:10.1038/ng.2279

44. Herceg Z, Hainaut P. Genetic and epigenetic alterations as biomarkers for cancer detection, diagnosis and prognosis. *Mol Oncol*. 2007;1(1):26-41. doi:10.1016/J.MOLONC.2007.01.004

45. Morrison GJ, Cunha AT, Jojo N, et al. Cancer transcriptomic profiling from rapidly enriched circulating tumor cells. *Int J cancer*. 2020;146(10):2845-2854. doi:10.1002/IJC.32915

46. Jiang Z, Zhang H, Gao Y, Sun Y. Multi-omics strategies for biomarker discovery and application in personalized oncology. *Mol Biomed*. 2025;6(1). doi:10.1186/S43556-025-00340-0

47. Vadapalli S, Abdelhalim H, Zeeshan S, Ahmed Z. Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine. *Brief Bioinform*. 2022;23(5). doi:10.1093/BIB/BBAC191

48. Upstill-Goddard R, Eccles D, Fliege J, Collins A. Machine learning approaches for the discovery of gene-gene interactions in disease data. *Brief Bioinform*. 2013;14(2):251-260. doi:10.1093/BIB/BBS024

49. Labory J, Njomgue-Fotso E, Bottini S. Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning algorithms for patient classification using metabolomics biomedical data. *Comput Struct Biotechnol J*. 2024;23:1274-1287. doi:10.1016/j.csbj.2024.03.016

50. Maleki Varnosfaderani S, Forouzanfar M. The Role of AI in Hospitals and Clinics: Transforming Healthcare in the 21st Century. *Bioeng (Basel, Switzerland)*. 2024;11(4). doi:10.3390/BIOENGINEERING11040337

51. Islam MA, Hasan Majumder MZ, Hussein MA, Hossain KM, Miah MS. A review of machine learning and deep learning algorithms for Parkinson's disease detection using handwriting and voice datasets. *Heliyon*. 2024;10(3). doi:10.1016/j.heliyon.2024.e25469

52. Lanka P, Rangaprakash D, Dretsch MN, Katz JS, Denney TS, Deshpande G. Supervised machine learning for diagnostic classification from large-scale neuroimaging datasets. *Brain Imaging Behav*. 2020;14(6):2378-2416. doi:10.1007/S11682-019-00191-8

53. Kaur H, Bhalla S, Kaur D, Raghava GPS. CancerLivER: a database of liver cancer gene

expression resources and biomarkers. *Database (Oxford)*. 2020;2020. doi:10.1093/DATABASE/BAAA012

54. Sapra D, Kaur H, Dhall A, Raghava GPS. ProCanBio: A Database of Manually Curated Biomarkers for Prostate Cancer. *J Comput Biol*. 2021;28(12):1248-1257. doi:10.1089/CMB.2021.0348

55. Bhalla S, Kaur H, Dhall A, Raghava GPS. Prediction and Analysis of Skin Cancer Progression using Genomics Profiles of Patients. *Sci Rep*. 2019;9(1). doi:10.1038/S41598-019-52134-4

56. Bhalla S, Chaudhary K, Kumar R, et al. Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer. *Sci Rep*. 2017;7(1):44997. doi:10.1038/srep44997

57. Zuo Z, Hu H, Xu Q, et al. BBCancer: an expression atlas of blood-based biomarkers in the early diagnosis of cancers. *Nucleic Acids Res*. 2020;48(D1):D789-D796. doi:10.1093/NAR/GKZ942

58. Maeser D, Gruener RF, Huang RS. oncoPredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data. *Brief Bioinform*. 2021;22(6). doi:10.1093/BIB/BBAB260

59. Kaur H, Bhalla S, Raghava GPS. Classification of early and late stage liver hepatocellular carcinoma patients from their genomics and epigenomics profiles. *PLoS One*. 2019;14(9). doi:10.1371/JOURNAL.PONE.0221476

60. Bhalla S, Kaur H, Kaur R, Sharma S, Raghava GPS. Expression based biomarkers and models to classify early and late stage samples of Papillary Thyroid Carcinoma. *bioRxiv*. Published online August 20, 2018:393975. doi:10.1101/393975

61. Kaur H, Dhall A, Kumar R, Raghava GPS. Identification of Platform-Independent Diagnostic Biomarker Panel for Hepatocellular Carcinoma Using Large-Scale Transcriptomics Data. *Front Genet*. 2020;10. doi:10.3389/FGENE.2019.01306

62. Biemann K. Peptides and proteins: overview and strategy. *Methods Enzymol*. 1990;193(C):351-360. doi:10.1016/0076-6879(90)93426-L

63. Alternative delivery systems for peptides and proteins as drugs - PubMed. Accessed January 10, 2026. https://pubmed.ncbi.nlm.nih.gov/3052876/

64. Kent SBH. The critical role of peptide chemistry in the life sciences. *J Pept Sci*. 2015;21(3):136-138. doi:10.1002/PSC.2754

65. Arenas EJ. Cytokine-driven cancer immune evasion mechanisms. *Int Rev Cell Mol Biol*. 2025;396:1-54. doi:10.1016/bs.ircmb.2025.01.004

66. Lin JX, Leonard WJ. Fine-Tuning Cytokine Signals. *Annu Rev Immunol*. 2019;37:295-324. doi:10.1146/ANNUREV-IMMUNOL-042718-041447

67. Goodwin D, Simerska P, Toth I. Peptides as therapeutics with enhanced bioactivity. *Curr Med Chem*. 2012;19(26):4451-4461. doi:10.2174/092986712803251548

68. Sahu KK, Minz S, Kaurav M, Pandey RS. Proteins and peptides: The need to improve them as promising therapeutics for ulcerative colitis. *Artif cells, nanomedicine, Biotechnol*. 2016;44(2):642-653. doi:10.3109/21691401.2014.975239

69. ElBeblawy R, Malla M, Akce M. Role of cytokine-based immunotherapy approaches in gastrointestinal cancers. *Cytokine Growth Factor Rev*. 2025;86:149-166. doi:10.1016/J.CYTOGFR.2025.10.005

70. Zafar S, Beg S, Panda SK, et al. Novel therapeutic interventions in cancer treatment using protein and peptide-based targeted smart systems. *Semin Cancer Biol*. 2021;69:249-267. doi:10.1016/j.semcancer.2019.08.023

71. Wang Z, Zhi K, Ding Z, et al. Emergence in protein derived nanomedicine as anticancer therapeutics: More than a tour de force. *Semin Cancer Biol*. 2021;69:77-90. doi:10.1016/j.semcancer.2019.11.012

72. Zhang M, Xu H. Peptide-assembled nanoparticles targeting tumor cells and tumor microenvironment for cancer therapy. *Front Chem*. 2023;11. doi:10.3389/FCHEM.2023.1115495

73. Wang M, Liu J, Xia M, et al. Peptide-drug conjugates: A new paradigm for targeted cancer therapy. *Eur J Med Chem*. 2024;265. doi:10.1016/j.ejmech.2023.116119

74. Sikder S, Gote V, Alshamrani M, Sicotte J, Pal D. Long-term delivery of protein and peptide therapeutics for cancer therapies. *Expert Opin Drug Deliv*. 2019;16(10):1113-1131. doi:10.1080/17425247.2019.1662785

75. Liu M, Svirskis D, Proft T, et al. Progress in peptide and protein therapeutics: Challenges

and strategies. *Acta Pharm Sin B*. 2025;15(12). doi:10.1016/j.apsb.2025.10.026

76. Rossino G, Marchese E, Galli G, et al. Peptides as Therapeutic Agents: Challenges and Opportunities in the Green Transition Era. *Molecules*. 2023;28(20). doi:10.3390/MOLECULES28207165

77. Basith S, Manavalan B, Hwan Shin T, Lee G. Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. *Med Res Rev*. 2020;40(4):1276-1314. doi:10.1002/MED.21658

78. Usmani SS, Bedi G, Samuel JS, et al. THPdb: Database of FDA-approved peptide and protein therapeutics. *PLoS One*. 2017;12(7). doi:10.1371/JOURNAL.PONE.0181748

79. Tyagi A, Kapoor P, Kumar R, Chaudhary K, Gautam A, Raghava GPS. In silico models for designing and discovering novel anticancer peptides. *Sci Rep*. 2013;3. doi:10.1038/SREP02984

80. Agrawal P, Bhagat D, Mahalwal M, Sharma N, Raghava GPS. AntiCP 2.0: an updated model for predicting anticancer peptides. *Brief Bioinform*. 2021;22(3). doi:10.1093/BIB/BBAA153

81. Schaduangrat N, Nantasenamat C, Prachayasittikul V, Shoombuatong W. ACPred: A Computational Tool for the Prediction and Analysis of Anticancer Peptides. *Molecules*. 2019;24(10). doi:10.3390/MOLECULES24101973

82. Tyagi A, Tuknait A, Anand P, et al. CancerPPD: a database of anticancer peptides and proteins. *Nucleic Acids Res*. 2015;43(Database issue):D837-D843. doi:10.1093/NAR/GKU892

83. Gautam A, Chaudhary K, Kumar R, et al. In silico approaches for designing highly effective cell penetrating peptides. *J Transl Med*. 2013;11(1). doi:10.1186/1479-5876-11-74

84. Chaudhary K, Kumar R, Singh S, et al. A Web Server and Mobile App for Computing Hemolytic Potency of Peptides. *Sci Rep*. 2016;6(1):22843. doi:10.1038/srep22843

85. Sharma N, Naorem LD, Jain S, Raghava GPS. ToxinPred2: an improved method for predicting toxicity of proteins. *Brief Bioinform*. 2022;23(5). doi:10.1093/BIB/BBAC174

86. Clifford JN, Høie MH, Deleuran S, Peters B, Nielsen M, Marcatili P. BepiPred-3.0: Improved B-cell epitope prediction using protein language models. *Protein Sci*. 2022;31(12). doi:10.1002/PRO.4497

87. A D, S P, N S, SS U, GPS R. Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. *Brief Bioinform*. 2021;22(2):936-945. doi:10.1093/BIB/BBAA259

88. Tomer R, Mehta NK, Malik S, Jain S, Raghava GPS. IL4Pred2: Prediction of Interleukin-4 Inducing Peptides in Human and Mouse 1. doi:10.1101/2025.04.23.650150

89. Wei L, Zhou C, Su R, Zou Q. PEPred-Suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning. *Bioinformatics*. 2019;35(21):4272-4280. doi:10.1093/BIOINFORMATICS/BTZ246

90. Wu X, Feng N, Wang C, Jiang H, Guo Z. Small molecule inhibitors as adjuvants in cancer immunotherapy: enhancing efficacy and overcoming resistance. *Front Immunol*. 2024;15. doi:10.3389/FIMMU.2024.1444452

91. Osipov A, Saung MT, Zheng L, Murphy AG. Small molecule immunomodulation: the tumor microenvironment and overcoming immune escape. *J Immunother cancer*. 2019;7(1). doi:10.1186/S40425-019-0667-0

92. Zhu Q, Zhang R, Zhao Z, Xie T, Sui X. Harnessing phytochemicals: Innovative strategies to enhance cancer immunotherapy. *Drug Resist Updat*. 2025;79. doi:10.1016/j.drup.2025.101206

93. Xue C, Chu Q, Shi Q, Zeng Y, Lu J, Li L. Wnt signaling pathways in biology and disease: mechanisms and therapeutic advances. *Signal Transduct Target Ther*. 2025;10(1). doi:10.1038/S41392-025-02142-W

94. Wang H, Liu C, Jin K, Li X, Zheng J, Wang D. Research advances in signaling pathways related to the malignant progression of HSIL to invasive cervical cancer: A review. *Biomed Pharmacother*. 2024;180. doi:10.1016/J.BIOPHA.2024.117483

95. Feng Y, Gan X, Rong X, Han Q. The components and regulation of the Hippo pathway and its relationships with the progression and treatment of Non-small cell lung cancer (NSCLC). *Cancer Cell Int*. 2025;25(1). doi:10.1186/S12935-025-03946-0

96. Singh S, Kumar R, Payra S, Singh SK. Artificial Intelligence and Machine Learning in Pharmacological Research: Bridging the Gap Between Data and Drug Discovery. *Cureus*. 2023;15(8). doi:10.7759/CUREUS.44359

97. Ambreen S, Umar M, Noor A, Jain H, Ali R. Advanced AI and ML frameworks for transforming drug discovery and optimization: With innovative insights in polypharmacology, drug repurposing, combination therapy and nanomedicine. *Eur J Med Chem*. 2025;284. doi:10.1016/j.ejmech.2024.117164

98. Bai H, Cheng Y, Jia S, et al. Machine learning-based QSAR and structure-based virtual screening guided discovery of novel mIDH1 inhibitors from natural products. *J Comput Aided Mol Des*. 2025;39(1). doi:10.1007/S10822-025-00624-1

99. Pitakbut T, Munkert J, Xi W, Wei Y, Fuhrmann G. Utilizing machine learning-based QSAR model to overcome standalone consensus docking limitation in beta-lactamase inhibitors screening: a proof-of-concept study. *BMC Chem*. 2024;18(1). doi:10.1186/S13065-024-01324-X

100. Sharma N, Patiyal S, Dhall A, Devi NL, Raghava GPS. ChAlPred: A web server for prediction of allergenicity of chemical compounds. *Comput Biol Med*. 2021;136. doi:10.1016/J.COMPBIOMED.2021.104746

101. Dhanda SK, Singla D, Mondal AK, Raghava GPS. DrugMint: a webserver for predicting and designing of drug-like molecules. *Biol Direct*. 2013;8(1). doi:10.1186/1745-6150-8-28

102. Singh H, Singh S, Singla D, Agarwal SM, Raghava GPS. QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest. *Biol Direct*. 2015;10(1). doi:10.1186/S13062-015-0046-9

103. 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). doi:10.1093/BIB/BBAA294

104. Mishra NK, Agarwal S, Raghava GPS. Prediction of cytochrome P450 isoform responsible for metabolizing a drug molecule. *BMC Pharmacol*. 2010;10. doi:10.1186/1471-2210-10-8

105. Do D, Bozdag S. Cancerin: A computational pipeline to infer cancer-associated ceRNA

interaction networks. *PLoS Comput Biol*. 2018;14(7). doi:10.1371/JOURNAL.PCBI.1006318

106. Chauhan JS, Dhanda SK, Singla D, Open Source Drug Discovery Consortium SM, Agarwal SM, Raghava GPS. QSAR-based models for designing quinazoline/imidazothiazoles/pyrazolopyrimidines based inhibitors against wild and mutant EGFR. Xue Y, ed. *PLoS One*. 2014;9(7):e101079. doi:10.1371/journal.pone.0101079

107. Rekand IH, Brenk R. DrugPred_RNA-A Tool for Structure-Based Druggability Predictions for RNA Binding Sites. *J Chem Inf Model*. 2021;61(8):4068-4081. doi:10.1021/ACS.JCIM.1C00155

108. Rawla P. Epidemiology of Prostate Cancer. *World J Oncol*. 2019;10(2):63-89. doi:10.14740/WJON1191

109. Ceci F, Airò Farulla LS, Bonatto E, et al. New target therapies in prostate cancer: from radioligand therapy, to PARP-inhibitors and immunotherapy. *Q J Nucl Med Mol Imaging*. 2024;68(2):101-115. doi:10.23736/S1824-4785.24.03575-1,

110. Cani AK, Hu K, Liu CJ, et al. Development of a Whole-urine, Multiplexed, Next-generation RNA-sequencing Assay for Early Detection of Aggressive Prostate Cancer. *Eur Urol Oncol*. 2022;5(4):430-439. doi:10.1016/j.euo.2021.03.002

111. Benesová M, Schäfer M, Bauder-Wüst U, et al. Preclinical evaluation of a tailor-made DOTA-conjugated PSMA inhibitor with optimized linker moiety for imaging and endoradiotherapy of prostate cancer. *J Nucl Med*. 2015;56(6):914-920. doi:10.2967/JNUMED.114.147413,

112. Kirby R. The role of PSA in detection and management of prostate cancer - PubMed. Accessed May 20, 2021. https://pubmed.ncbi.nlm.nih.gov/27337755/

113. Mengual L, Musquera M, Ciudin A, Ribal JM. [Non- PSA serum markers for the diagnosis of PCa] - PubMed. Accessed May 20, 2021. https://pubmed.ncbi.nlm.nih.gov/25948796/

114. CG island methylation changes near the GSTP1 gene in prostatic intraepithelial neoplasia - PubMed. Accessed May 20, 2021. https://pubmed.ncbi.nlm.nih.gov/9641498/

115. Cucchiara V, Cooperberg MR, Dall'Era M, et al. Genomic Markers in Prostate Cancer Decision Making. *Eur Urol*. 2018;73(4):572-582. doi:10.1016/j.eururo.2017.10.036

116. Isaacs MJGBA van BGWVFPSHFMKJASFMJDNRWB. DD3: a new prostate-specific gene, highly overexpressed in prostate cancer - PubMed. Accessed May 20, 2021. https://pubmed.ncbi.nlm.nih.gov/10606244/

117. Hendriks RJ, van der Leest MMG, Dijkstra S, et al. A urinary biomarker-based risk score correlates with multiparametric MRI for prostate cancer detection. *Prostate*. 2017;77(14):1401-1407. doi:10.1002/PROS.23401,

118. Reduced Rate of Repeated Prostate Biopsies Observed in ConfirmMDx Clinical Utility Field Study - PubMed. Accessed June 29, 2025. https://pubmed.ncbi.nlm.nih.gov/24991397/

119. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-Generation Machine Learning for Biological Networks. *Cell*. 2018;173(7):1581-1592. doi:10.1016/j.cell.2018.05.015

120. Goldman MJ, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol*. 2020;38(6):675-678. doi:10.1038/S41587-020-0546-8,

121. Sharma N, Naorem LD, Jain S, Raghava GPS. ToxinPred2: an improved method for predicting toxicity of proteins. *Brief Bioinform*. 2022;23(5). doi:10.1093/BIB/BBAC174

122. Dhall A, Patiyal S, Choudhury S, Jain S, Narang K, Raghava GPS. TNFepitope: A webserver for the prediction of TNF-α inducing epitopes. *Comput Biol Med*. 2023;160. doi:10.1016/J.COMPBIOMED.2023.106929

123. Leyten GHJM, Hessels D, Smit FP, et al. Identification of a candidate gene panel for the early diagnosis of prostate cancer. *Clin Cancer Res*. 2015;21(13):3061-3070. doi:10.1158/1078-0432.CCR-14-3334

124. Kote-Jarai Z, Amin Al Olama A, Leongamornlert D, et al. Identification of a novel prostate cancer susceptibility variant in the KLK3 gene transcript. *Hum Genet*. 2011;129(6):687-694. doi:10.1007/S00439-011-0981-1,

125. Parikh H, Wang Z, Pettigrew KA, et al. Fine mapping the KLK3 locus on chromosome

19q13.33 associated with prostate cancer susceptibility and PSA levels. *Hum Genet*. 2011;129(6):675-685. doi:10.1007/S00439-011-0953-5,

126. Advances in protein therapeutics - PubMed. Accessed January 8, 2024. https://pubmed.ncbi.nlm.nih.gov/18946840/

127. Fosgerau K, Hoffmann T. Peptide therapeutics: current status and future directions. *Drug Discov Today*. 2015;20(1):122-128. doi:10.1016/J.DRUDIS.2014.10.003

128. Cai L, Wang L, Fu X, Xia C, Zeng X, Zou Q. ITP-Pred: an interpretable method for predicting, therapeutic peptides with fused features low-dimension representation. *Brief Bioinform*. 2021;22(4). doi:10.1093/BIB/BBAA367

129. Dimitrov DS. Therapeutic proteins. *Methods Mol Biol*. 2012;899:1-26. doi:10.1007/978-1-61779-921-1_1/COVER

130. Zaman R, Islam RA, Chowdhury EH. Evolving therapeutic proteins to precisely kill cancer cells. *J Control Release*. 2022;351:779-804. doi:10.1016/J.JCONREL.2022.09.066

131. Liu X, Cheng Y, Mu Y, et al. Diverse drug delivery systems for the enhancement of cancer immunotherapy: an overview. *Front Immunol*. 2024;15. doi:10.3389/FIMMU.2024.1328145

132. Dean TT, Jelu-́Reyes J, Allen AC, Moore TW. Peptide-Drug Conjugates: An Emerging Direction for the Next Generation of Peptide Therapeutics. *J Med Chem*. 2024;67(3):1641-1661. doi:10.1021/ACS.JMEDCHEM.3C01835

133. Sharma K, Sharma KK, Sharma A, Jain R. Peptide-based drug discovery: Current status and recent advances. *Drug Discov Today*. 2023;28(2). doi:10.1016/J.DRUDIS.2022.103464

134. Vlieghe P, Lisowski V, Martinez J, Khrestchatisky M. Synthetic therapeutic peptides: science and market. *Drug Discov Today*. 2010;15(1-2):40-56. doi:10.1016/J.DRUDIS.2009.10.009

135. Usmani SS, Kumar R, Bhalla S, Kumar V, Raghava GPS. In Silico Tools and Databases for Designing Peptide-Based Vaccine and Drugs. *Adv Protein Chem Struct Biol*. 2018;112:221-263. doi:10.1016/BS.APCSB.2018.01.006

136. Usmani SS, Bedi G, Samuel JS, et al. THPdb: Database of FDA-approved peptide and protein therapeutics. *PLoS One*. 2017;12(7). doi:10.1371/journal.pone.0181748

137. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res*. 2018;46(D1):D1074-D1082. doi:10.1093/NAR/GKX1037,

138. Bateman A, Martin MJ, Orchard S, et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021;49(D1):D480-D489. doi:10.1093/NAR/GKAA1100,

139. Wu C, Gao R, Zhang Y, De Marinis Y. PTPD: predicting therapeutic peptides by deep learning and word2vec. *BMC Bioinformatics*. 2019;20(1). doi:10.1186/S12859-019-3006-Z

140. Yan K, Lv H, Guo Y, Chen Y, Wu H, Liu B. TPpred-ATMV: therapeutic peptide prediction by adaptive multi-view tensor learning model. *Bioinformatics*. 2022;38(10):2712-2718. doi:10.1093/BIOINFORMATICS/BTAC200

141. Yan W, Tang W, Wang L, Bin Y, Xia J. PrMFTP: Multi-functional therapeutic peptides prediction based on multi-head self-attention mechanism and class weight optimization. *PLoS Comput Biol*. 2022;18(9). doi:10.1371/JOURNAL.PCBI.1010511

142. Lv H, Yan K, Liu B. TPpred-LE: therapeutic peptide function prediction based on label embedding. *BMC Biol*. 2023;21(1). doi:10.1186/S12915-023-01740-W

143. Yan K, Guo Y, Liu B. PreTP-2L: identification of therapeutic peptides and their types using two-layer ensemble learning framework. *Bioinformatics*. 2023;39(4). doi:10.1093/BIOINFORMATICS/BTAD125

144. Yan K, Lv H, Wen J, Guo Y, Xu Y, Liu B. PreTP-Stack: Prediction of Therapeutic Peptides Based on the Stacked Ensemble Learing. *IEEE/ACM Trans Comput Biol Bioinforma*. 2023;20(2):1337-1344. doi:10.1109/TCBB.2022.3183018

145. Xu C, Ge L, Zhang Y, Dehmer M, Gutman I. Computational prediction of therapeutic peptides based on graph index. *J Biomed Inform*. 2017;75:63-69. doi:10.1016/J.JBI.2017.09.011

146. Zhang YP, Zou Q. PPTPP: a novel therapeutic peptide prediction method using physicochemical property encoding and adaptive feature representation learning. *Bioinformatics*. 2020;36(13):3982-3987. doi:10.1093/BIOINFORMATICS/BTAA275

147. Guo Y, Yan K, Lv H, Liu B. PreTP-EL: prediction of therapeutic peptides based on ensemble learning. *Brief Bioinform*. 2021;22(6). doi:10.1093/BIB/BBAB358

148. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22(13):1658-1659. doi:10.1093/bioinformatics/btl158

149. Kaur D, Arora C, Raghava GPS. A Hybrid Model for Predicting Pattern Recognition Receptors Using Evolutionary Information. *Front Immunol*. 2020;11. doi:10.3389/FIMMU.2020.00071

150. Pande A, Patiyal S, Lathwal A, et al. Pfeature: A Tool for Computing Wide Range of Protein Features and Building Prediction Models. *J Comput Biol*. 2023;30(2):204-222. doi:10.1089/CMB.2022.0241,

151. AN M, JA C, R de WM, et al. Interleukin 13, a T-cell-derived cytokine that regulates human monocyte and B-cell function. *Proc Natl Acad Sci U S A*. 1993;90(8):3735-3739. doi:10.1073/PNAS.90.8.3735

152. A M, P C, JM D, et al. Interleukin-13 is a new human lymphokine regulating inflammatory and immune responses. *Nature*. 1993;362(6417):248-250. doi:10.1038/362248A0

153. Mannon P, Reinisch W. Interleukin 13 and its role in gut defence and inflammation. *Gut*. 2012;61(12):1765-1773. doi:10.1136/GUTJNL-2012-303461,

154. Wynn TA. IL-13 effector functions. *Annu Rev Immunol*. 2003;21:425-456. doi:10.1146/ANNUREV.IMMUNOL.21.120601.141142,

155. Marone G, Granata F, Pucino V, et al. The intriguing role of interleukin 13 in the pathophysiology of asthma. *Front Pharmacol*. 2019;10. doi:10.3389/FPHAR.2019.01387,

156. Zhu Z, Homer RJ, Wang Z, et al. Pulmonary expression of interleukin-13 causes inflammation, mucus hypersecretion, subepithelial fibrosis, physiologic abnormalities, and eotaxin production. *J Clin Invest*. 1999;103(6):779-788. doi:10.1172/JCI5909,

157. Jaén M, Martín-Regalado Á, Bartolomé RA, Robles J, Casal JI. Interleukin 13 receptor alpha 2 (IL13Rα2): Expression, signaling pathways and therapeutic applications in cancer. *Biochim Biophys Acta - Rev Cancer*. 2022;1877(5).

doi:10.1016/j.bbcan.2022.188802

158. Oshima Y, Puri RK. Suppression of an IL-13 autocrine growth loop in a human Hodgkin/Reed-Sternberg tumor cell line by a novel IL-13 antagonist. *Cell Immunol*. 2001;211(1):37-42. doi:10.1006/CIMM.2001.1828,

159. Skinnider BF, Kapp U, Mak TW. Interleukin 13: A growth factor in Hodgkin lymphoma. *Int Arch Allergy Immunol*. 2001;126(4):267-276. doi:10.1159/000049523,

160. M T, JM P, JA B. Role of IL-13 in regulation of anti-tumor immunity and tumor growth. *Cancer Immunol Immunother*. 2004;53(2):79-85. doi:10.1007/S00262-003-0445-0

161. Su T, Mi Y, Zhang L, et al. Association between IL13 gene polymorphisms and susceptibility to cancer: A meta-analysis. *Gene*. 2013;515(1):56-61. doi:10.1016/J.GENE.2012.11.035,

162. Traub B, Sun L, Ma Y, et al. Endogenously expressed IL-4Rα promotes the malignant phenotype of human pancreatic cancer in vitro and in vivo. *Int J Mol Sci*. 2017;18(4). doi:10.3390/IJMS18040716,

163. Song X, Traub B, Shi J, Kornmann M. Possible roles of interleukin-4 and-13 and their receptors in gastric and colon cancer. *Int J Mol Sci*. 2021;22(2):1-20. doi:10.3390/IJMS22020727,

164. S OR. Immune surveillance: a balance between protumor and antitumor immunity. *Curr Opin Genet Dev*. 2008;18(1):11-18. doi:10.1016/J.GDE.2007.12.007

165. Cho YA, Kim J. Association of IL4, IL13, and IL4R polymorphisms with gastrointestinal cancer risk: A meta-analysis. *J Epidemiol*. 2017;27(5):215-220. doi:10.1016/j.je.2016.06.002

166. Receptor for interleukin 13 is a marker and therapeutic target for human high-grade gliomas - PubMed. Accessed June 29, 2025. https://pubmed.ncbi.nlm.nih.gov/10353730/

167. AN D, TE S, C M, et al. IL-13 is a driver of COVID-19 severity. *medRxiv Prepr Serv Heal Sci*. Published online March 1, 2021. doi:10.1101/2020.06.18.20134353

168. Yang Y, Shen C, Li J, et al. Plasma IP-10 and MCP-3 levels are highly associated with disease severity and predict the progression of COVID-19. *J Allergy Clin Immunol*.

2020;146(1):119-127.e4. doi:10.1016/j.jaci.2020.04.027

169. Vita R, Mahajan S, Overton JA, et al. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res*. 2019;47(D1):D339-D343. doi:10.1093/NAR/GKY1006

170. Lawrence T. The nuclear factor NF-kappaB pathway in inflammation. *Cold Spring Harb Perspect Biol*. 2009;1(6). doi:10.1101/CSHPERSPECT.A001651

171. Lim KH, Staudt LM. Toll-like receptor signaling. *Cold Spring Harb Perspect Biol*. 2013;5(1). doi:10.1101/CSHPERSPECT.A011247

172. Tak PP, Firestein GS. NF-kappaB: a key role in inflammatory diseases. *J Clin Invest*. 2001;107(1):7-11. doi:10.1172/JCI11830

173. Li Q, Verma IM. NF-kappaB regulation in the immune system. *Nat Rev Immunol*. 2002;2(10):725-734. doi:10.1038/NRI910

174. Hayden MS, Ghosh S. Shared principles in NF-kappaB signaling. *Cell*. 2008;132(3):344-362. doi:10.1016/J.CELL.2008.01.020

175. Hoesel B, Schmid JA. The complexity of NF-κB signaling in inflammation and cancer. *Mol Cancer*. 2013;12(1). doi:10.1186/1476-4598-12-86

176. Zinatizadeh MR, Schock B, Chalbatani GM, Zarandi PK, Jalali SA, Miri SR. The Nuclear Factor Kappa B (NF-kB) signaling in cancer development and immune diseases. *Genes Dis*. 2020;8(3):287-297. doi:10.1016/J.GENDIS.2020.06.005

177. Lee CH, Jeon YT, Kim SH, Song YS. NF-kappaB as a potential molecular target for cancer therapy. *Biofactors*. 2007;29(1):19-35. doi:10.1002/BIOF.5520290103

178. Rasmi RR, Sakthivel KM, Guruvayoorappan C. NF-κB inhibitors in treatment and prevention of lung cancer. *Biomed Pharmacother*. 2020;130. doi:10.1016/J.BIOPHA.2020.110569

179. Tang D, Tao D, Fang Y, Deng C, Xu Q, Zhou J. TNF-Alpha Promotes Invasion and Metastasis via NF-Kappa B Pathway in Oral Squamous Cell Carcinoma. *Med Sci Monit Basic Res*. 2017;23:141-149. doi:10.12659/MSMBR.903910

180. Pikarsky E, Porat RM, Stein I, et al. NF-kappaB functions as a tumour promoter in inflammation-associated cancer. *Nature*. 2004;431(7007):461-466. doi:10.1038/NATURE02924

181. Dong QM, Ling C, Chen X, Zhao L. Inhibition of tumor necrosis factor-α enhances apoptosis induced by nuclear factor-κB inhibition in leukemia cells. *Oncol Lett*. 2015;10(6):3793-3798. doi:10.3892/OL.2015.3786

182. Karin M. Nuclear factor-kappaB in cancer development and progression. *Nature*. 2006;441(7092):431-436. doi:10.1038/NATURE04870

183. Baldwin AS. The NF-kappa B and I kappa B proteins: new discoveries and insights. *Annu Rev Immunol*. 1996;14:649-681. doi:10.1146/ANNUREV.IMMUNOL.14.1.649

184. Wu Y, Zhou BP. TNF-alpha/NF-kappaB/Snail pathway in cancer cell migration and invasion. *Br J Cancer*. 2010;102(4):639-644. doi:10.1038/SJ.BJC.6605530

185. Baud V, Karin M. Is NF-kappaB a good target for cancer therapy? Hopes and pitfalls. *Nat Rev Drug Discov*. 2009;8(1):33-40. doi:10.1038/NRD2781

186. Bonizzi G, Karin M. The two NF-κB activation pathways and their role in innate and adaptive immunity. *Trends Immunol*. 2004;25(6):280-288. doi:10.1016/j.it.2004.03.008

187. Hayden MS, Ghosh S. Regulation of NF-κB by TNF family cytokines. *Semin Immunol*. 2014;26(3):253-266. doi:10.1016/J.SMIM.2014.05.004

188. Hayden MS, Ghosh S. Signaling to NF-kappaB. *Genes Dev*. 2004;18(18):2195-2224. doi:10.1101/GAD.1228704

189. Schütze S, Wiegmann K, Machleidt T, Krönke M. TNF-induced activation of NF-kappa B. *Immunobiology*. 1995;193(2-4):193-203. doi:10.1016/S0171-2985(11)80543-7

190. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*. 2011;32(7):1466-1474. doi:10.1002/JCC.21707

191. RDKit. Accessed June 3, 2026. https://www.rdkit.org/

192. El-Bassouny DR, Omar NM, Khalaf HA, Al-Salam RAA. Role of nuclear factor-kappa B in bleomycin induced pulmonary fibrosis and the probable alleviating role of ginsenoside: histological, immunohistochemical, and biochemical study. *Anat Cell Biol*. 2021;54(4):448-464. doi:10.5115/ACB.21.068

193. Colasanti M, Persichini T. Nitric oxide: an inhibitor of NF-kappaB/Rel system in glial cells. *Brain Res Bull*. 2000;52(3):155-161. doi:10.1016/S0361-9230(00)00262-8

194. Dambrova M, Zvejniece L, Skapare E, et al. The anti-inflammatory and antinociceptive

effects of NF-kappaB inhibitory guanidine derivative ME10092. *Int Immunopharmacol*. 2010;10(4):455-460. doi:10.1016/J.INTIMP.2010.01.006

195. Noori T, Dehpour AR, Alavi SD, et al. Synthesis and evaluation of the effects of solid lipid nanoparticles of ivermectin and ivermectin on cuprizone-induced demyelination via targeting the TRPA1/NF-kB/GFAP signaling pathway. *Iran J Basic Med Sci*. 2023;26(11):1272-1282. doi:10.22038/IJBMS.2023.71309.15493

196. Jiang L, Wang P, Sun YJ, Wu YJ. Ivermectin reverses the drug resistance in cancer cells through EGFR/ERK/Akt/NF-κB pathway. *J Exp Clin Cancer Res*. 2019;38(1). doi:10.1186/S13046-019-1251-7

197. Zhang X, Song Y, Ci X, et al. Ivermectin inhibits LPS-induced production of inflammatory cytokines and improves LPS-induced survival in mice. *Inflamm Res*. 2008;57(11):524-529. doi:10.1007/S00011-008-8007-8

198. Sun Y, Lu T, Pan J, et al. Dual tobramycin and docosahexaenoic acid loaded nanoemulsions combating Pseudomonas aeruginosa-induced pulmonary infection. *Colloids Surf B Biointerfaces*. 2024;242. doi:10.1016/J.COLSURFB.2024.114088

199. Nguyen T, Louie SG, Beringer PM, Gill MA. Potential role of macrolide antibiotics in the management of cystic fibrosis lung disease. *Curr Opin Pulm Med*. 2002;8(6):521-528. doi:10.1097/00063198-200211000-00007

200. Sunaga T, Oh N, Hosoya K, Takagi S, Okumura M. Inhibitory effects of pentosan polysulfate sodium on MAP-kinase pathway and NF-κB nuclear translocation in canine chondrocytes in vitro. *J Vet Med Sci*. 2012;74(6):707-711. doi:10.1292/JVMS.11-0511

201. Krishnan R, Stapledon CJM, Mostafavi H, et al. Anti-inflammatory actions of Pentosan polysulfate sodium in a mouse model of influenza virus A/PR8/34-induced pulmonary inflammation. *Front Immunol*. 2023;14. doi:10.3389/FIMMU.2023.1030879