MACHINE LEARNING - BASED PREDICTION OF INTERLEUKIN-2 INDUCING POTENTIAL OF PEPTIDES

A Dissertation

Submitted in partial fulfillment of the requirement for the degree of

MASTER OF SCIENCE

in

BIOTECHNOLOGY

by:

Anshita

23/MSCBIO/57

Under the supervision of

Prof. Yasha Hasija

Professor and Head of Department

Department of Biotechnology

To the

Department of Biotechnology

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Bawana Road, Delhi-110042. India

May, 2025

ACKNOWLEDGEMENT

At the time of submission of my M.Sc. Dissertation I am grateful to the divine God who has given me the insight, fortitude, and perseverance to complete this endeavour. In addition to work, a lot of other people's support and direction are crucial to this project's success. Thus, I would like to use this opportunity to thank everyone who has contributed to the accomplishment of this project.

I would first want to express my gratitude to my mentor, Prof. Yasha Hasija, Department of Biotechnology, Delhi Technological University, for providing me with the chance to work on a project under her direction. I was able to finish this assignment because of her mysterious oversight, constant encouragement, and knowledgeable direction. I respectfully use this as a chance to thank her.

I would like to thank all the professors from Department of Biotechnology, Delhi Technological University, for their advice and constant assistance during my time at the university. I also want to express my profound appreciation to Ph.D. scholars Khushi Yadav, Neha Kumari, Akansha Bisht, Rashmi and Bhargavi for their insightful observations and constant guidance. Their analytical comments on the project's problems have been crucial to its successful development.

I would like to thank Mr. Jitender Singh and Mr. C.B. Singh, the technical personnel, for their assistance whenever needed. Finally, I would like to thank my family and friends for their unwavering support over this entire process.

Anshita
23/MSCBIO/57

Delhi Technological University

Shahbad Daulatpur, Main Bawana Road, Delhi-110042. India

DECLARATION

I, Anshita, 23/MSCBIO/57 student of M.Sc. Biotechnology hereby declares that the Dissertation Project entitled **“Machine Learning Based Prediction of Interleukin-2 Inducing Potential of Peptides”** is submitted by me to the Department of Biotechnology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Science. This work is original and not copied from any source without paper citation. I have honored the principles of academic integrity and have upheld the normal student code of academic conduct in the completion of this work.

Candidate’s Signature

Delhi Technological University

Shahbad Daulatpur, Main Bawana Road, Delhi-110042. India

CERTIFICATE BY THE SUPERVISOR(S)

This is to certify that the Dissertation Project titled **"Machine Learning Based Prediction of Interleukin-2 Inducing Potential of Peptides"** which is being submitted by Anshita, 23/MSCBIO/57, Department of Biotechnology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Science is a record of the work carried out by the student under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: DTU, Delhi

Date:

Prof. Yasha Hasija
Supervisor
Department of Biotechnology
Delhi Technological University
Delhi -110042

Prof. Yasha Hasija
Head of Department
Department of Biotechnology
Delhi Technological University
Delhi- 110042

Machine Learning Based Prediction of Interleukin-2 Inducing Potential of Peptides

Anshita (23/MSCBIO/57)

ABSTRACT

Interleukins are pivotal cytokines that orchestrate immune regulation, with their dysregulation contributing to a spectrum of diseases ranging from autoimmunity to cancer. Among them, Interleukin-2 (IL-2) plays a crucial role in T-cell proliferation, immune tolerance, and the efficacy of immunotherapies. The precise identification of IL-2-inducing peptides (IIPs) is fundamental for the rational design of vaccines and immunotherapeutic agents. However, experimental discovery of IIPs is inherently laborious, costly, and limited in throughput, underscoring the urgent need for robust computational approaches.

Recent advances in machine learning (ML) have revolutionized peptide immunoinformatics, enabling the high-throughput, accurate, and reproducible prediction of cytokine-inducing peptides directly from sequence data. This thesis provides a comprehensive review and critical assessment of ML methodologies developed for predicting the interleukin-inducing potential of peptides, with a particular emphasis on IL-2. The work begins by contextualizing the immunological significance of interleukins and the central role of IL-2 in immune modulation and therapy. It then systematically addresses the challenges and limitations associated with experimental identification of IIPs, motivating the transition to computational strategies.

A detailed methodology is presented for the extraction, preprocessing, and curation of high-quality peptide datasets from the Immune Epitope Database (IEDB), focusing on experimentally validated IL-2 inducers and non-inducers. Rigorous data cleaning, feature engineering, and exploratory data analysis are performed to uncover sequence-level and physicochemical patterns distinguishing IL-2 inducers. The thesis explores a broad spectrum of ML algorithms, feature selection techniques, and validation strategies, providing insights into model interpretability and performance. Comparative analyses highlight the strengths and limitations of current computational tools for IL-2 prediction, identifying key gaps and opportunities for further innovation.

The review concludes with a forward-looking discussion on the integration of novel ML architectures, multi-omics data, and explainable AI to enhance the predictive power and biological interpretability of IIP models. The findings underscore the transformative potential of ML-driven approaches in immunological research, facilitating the rapid discovery of therapeutic peptides and advancing the frontiers of translational medicine. This thesis not only serves as an authoritative resource on the state-of-the-art in IL-2 peptide prediction but also provides a reproducible framework and actionable recommendations for future computational immunology studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

IL	Interleukin
CD	Cluster of Differentiation
IIP	Interleukin-Inducing Potential
ML	Machine Learning
IEDB	Immune Epitope Database
AIML	Artificial Intelligence and Machine Learning
MAPK	Mitogen-Activated Protein Kinase
ERK	Extracellular Signal-Regulated Kinase
JAK-STAT	Janus Kinase - Signal Transducer and Activator of Transcription
Tregs	Regulatory T Cells
NF	Nuclear Factor
AP	Antigen-Presenting
MHC	Major Histocompatibility Complex
Th	T Helper (Cells)
NK	Natural Killer (Cells)
PSSM	Position-Specific Scoring Matrix
FASTA	Fast-All (sequence format)
AAC	Amino Acid Composition
DPC	Dipeptide Composition
MARCI	Motif-Associated Regions for Cytokine Induction
mRMR	Minimum Redundancy Maximum Relevance
BLAST	Basic Local Alignment Search Tool
CD-HIT	Cluster Database at High Identity with Tolerance
SMOTE	Synthetic Minority Over-sampling Technique
BLOSUM	BLOcks SUbstitution Matrix
SHAP	SHapley Additive exPlanations
PCA	Principal Component Analysis
T SNE	t-distributed Stochastic Neighbor Embedding

CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GNN	Graph Neural Network
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
MCC	Matthews Correlation Coefficient
TPC	Tripeptide Composition
APAAC	Amphiphilic Pseudo Amino Acid Composition
HLA	Human Leukocyte Antigen
XAI	Explainable Artificial Intelligence
A	Alanine
R	Arginine
N	Asparagine
D	Aspartic acid
C	Cysteine
E	Glutamic acid
Q	Glutamine
G	Glycine
H	Histidine
I	Isoleucine
L	Leucine
K	Lysine
M	Methionine
F	Phenylalanine
P	Proline
S	Serine
T	Threonine
W	Tryptophan
Y	Tyrosine
V	Valine

CHAPTER 1. INTRODUCTION

1.1 Background on the Immune System and Cytokines

The human immune system is a highly sophisticated network of organs, cells, proteins, and signaling molecules that collectively defend the body against infections and maintain tissue homeostasis [1]. It operates through two principal arms: innate and adaptive immunity. Innate immunity represents the body's first line of defense, providing immediate, non-specific responses to invading pathogens through barriers such as skin, mucous membranes, and a variety of immune cells including phagocytes and natural killer cells [2]. This system acts rapidly but lacks the ability to remember specific pathogens. In contrast, adaptive immunity develops over time and is characterized by specificity and memory. It is mediated primarily by lymphocytes—T cells and B cells—which recognize specific antigens and mount tailored responses. Upon re-exposure to the same pathogen, adaptive immunity responds more rapidly and robustly due to immunological memory [3].

Central to the coordination and regulation of these immune responses are cytokines—a diverse group of small, secreted proteins that facilitate intercellular communication [4]. Cytokines orchestrate immune cell proliferation, differentiation, migration, and effector functions, acting in complex networks or cascades. Among the various cytokine families, interleukins (ILs) play particularly pivotal roles in modulating both innate and adaptive immunity. Interleukins regulate inflammation, hematopoiesis, and the activation or suppression of specific immune cell subsets, with each interleukin exhibiting pleiotropic and sometimes overlapping biological activities. Their precise regulation is essential for effective immune defense and the prevention of pathological conditions such as chronic inflammation, autoimmunity, or immunodeficiency [5].

1.2 Role of IL-2 in Immunity

Interleukin-2 (IL-2**)** is a prototypical member of the interleukin family and a critical regulator of immune system function [6]. Produced primarily by activated CD4+ and CD8+ T cells, IL-2 exerts its effects through a heterotrimeric receptor complex composed of α (CD25), β (CD122), and γ (CD132) chains [7]. The binding of IL-2 to its high-affinity receptor initiates a cascade of intracellular signaling events—most notably via the JAK-STAT, PI3K/Akt/mTOR, and MAPK/ERK pathways—leading to the transcription of genes involved in cell proliferation, survival, and differentiation [8].

IL-2 is indispensable for T-cell proliferation, particularly during the clonal expansion phase following antigen recognition. It promotes the growth and differentiation of both helper and cytotoxic T cells, as well as the maintenance and function of regulatory T cells (Tregs), which are essential for self-tolerance and immune homeostasis [9]. Notably, IL-2's dual role in stimulating effector T cells and expanding Tregs positions it as a master regulator—capable of amplifying immune responses against infections and tumors, while also preventing autoimmunity by restraining excessive or misdirected immune activity. The therapeutic relevance of IL-2 is underscored by its clinical applications. High-dose IL-2 therapy was among the first immunotherapies approved for metastatic renal cell carcinoma and melanoma, achieving durable responses in a subset of patients. More recently, low-dose IL-2 regimens have shown promise in selectively expanding Tregs to treat autoimmune diseases such as type 1 diabetes and systemic lupus erythematosus [10]. However, the clinical use of IL-2 is complicated by its pleiotropic effects, narrow therapeutic window, and potential for severe toxicities, necessitating more precise approaches for harnessing its immunomodulatory potential.

1.3 Need for Computational Prediction

Despite the central importance of IL-2 in immunology and therapy, the experimental identification of IL-2-inducing peptides remains a significant challenge [11]. Traditional wet-lab approaches—including peptide synthesis, in vitro assays, and animal models—are labor-intensive, costly, and often limited in throughput. These constraints hinder the rapid discovery and optimization of peptide-based immunotherapies or vaccines targeting IL-2 pathways [12]. In recent years, the rise of machine learning (ML) and artificial intelligence has revolutionized peptide immunoinformatics, enabling the prediction of cytokine-inducing peptides directly from sequence data [13]. ML models can learn complex relationships between peptide features and biological activity, facilitating the in silico screening of vast peptide libraries and prioritizing candidates for experimental validation. While robust computational tools exist for predicting peptides that induce other interleukins—such as IL-4, IL-5, and IL-10—there remains a notable gap in resources specifically tailored for IL-2 [14]. This gap is particularly striking given IL-2's therapeutic significance and the growing interest in personalized immunotherapies. The development of accurate, interpretable, and user-friendly ML models for IL-2-inducing peptide prediction would accelerate discovery pipelines, reduce experimental burden, and deepen our understanding of the sequence and structural determinants underlying IL-2-mediated immune responses. Such advances are critical for translating basic immunological insights into novel diagnostics, vaccines, and immunotherapies [15].

In short, the immune system's complexity and the central role of cytokines like IL-2 in orchestrating immune responses highlight the need for advanced computational tools. By leveraging machine learning, researchers can bridge the gap between experimental limitations and the growing demand for precision immunomodulation, ultimately enabling more effective and safer therapeutic strategies. This thesis is dedicated to addressing this unmet need by providing a comprehensive review and data-driven framework for the computational prediction of IL-2-inducing peptides.

CHAPTER 2. REVIEW OF LITERATURE

2.1Introduction

Interleukins (ILs) are central cytokines controlling innate and adaptive immunity through complex signaling cascades that modulate immune cell activation, differentiation, and effector responses [16]. As an illustration, IL-1 signaling turns on NF-κB and AP-1 transcription factors, activating dendritic cells and macrophages, and enabling T cell-dependent antibody responses—testifying to the importance of ILs within host defense mechanisms [17]. The precise identification of interleukin-inducing peptides (IIPs) is therefore pivotal for the development of immunotherapeutic methods, vaccine design, and diagnostics [18]. Although classical assays like cytokine release and ELISpot are axiomatic, they are labor-intensive and low-throughput in nature, which makes the use of computational alternatives a necessity [19]. Advances in machine learning in recent times have made it possible to extract sophisticated sequence features and build accurate predictive models, as evidenced by methods such as IL2pred and IL6Pred, which apply large-scale immunological data to predict IL-2 and IL-6 inducing peptides, respectively [18], [20]. Models like enhanced iIL13Pred demonstrates that the accuracy and generalizability of IIP prediction can be increased by using combination of multi-classifiers and advanced feature selection which facilitates quick in silico screening and logical immunomodulator design [21]. Moreover, the development of comprehensive databases and web servers dedicated to interleukin-inducing peptides has made it simpler for the scientific community to benchmark and access predictive models. Even though all these developments have been made, there is still room for improvement regarding data imbalance, sparse experimental validation, and the interpretable models.

2.2 Biological Significance of Interleukins

Innate immunity acts as the first line of defense in our body. B and T lymphocytes play a very pivotal role in innate immunity. The foreign antigens expressed by MHC II stimulate the release of various cytokines by T helper cells. Interleukins are one of the extensive group of cytokines which are secreted by Th2 helper cells and promotes proliferation of lymphocytes, natural killer cells and macrophages. Chemically interleukins are glycoproteins having a length range of 99 to 1332 amino acids [22]. They were first discovered in 1970s and till now scientists have characterized almost 40 different interleukins. They play pivotal role in immune cell system like cell proliferation, differentiation, maturation, chemotaxis, phagocytosis adhesion and migration to the affected region. The action mechanism of interleukins involved binding to high affinity receptors expressed on cell surfaces and the inducing cascade of responses [23]. This action mechanism can be achieved by three different ways which are autocrine, paracrine and endocrine release of interleukins and this helps them to act throughout the body like metabolic, neuroendocrine and cardiovascular systems ultimately contributing to homeostasis of body. They are able to perform so many functions because of some key properties they possess like redundancy, pleiotropy, synergism and antagonism [24]. Based on their function, Interleukins can be divided into three categories which includes i) inflammatory mediators, largest category including interleukins like IL-1, IL-4, IL5, IL-6 etc. and they act by initiating immune reactions against pathogens by activating and attracting immune cells ii) anti-inflammatory biomarkers like IL-10, IL-30, IL-37 which act to contain immune reactions and avoid excessive tissue destruction, thus ensuring immune balance and iii) the last category includes IL-2,IL-11, IL-12 which can act as both and result in inflammation and anti-inflammation depending upon the situation [22]. Pathological conditions like persistent inflammation, autoimmune conditions, allergies, and even cancer can result from a distortion of the delicate balance between these two opposing interleukins. Moreover, interleukins are also responsible for the management of disease by creating a favorable microenvironment and boosting cytotoxic T cells and natural killer (NK) cells [25]. Interleukins can also provide anti-tumor immunity in cancer Targeted therapies, such as recombinant IL-2 in cancer immunotherapy and monoclonal antibodies that act against IL-6 in rheumatoid arthritis, have been developed as a result of this dualism, highlighting their therapeutic use [26]. Additionally, activity of non-immune cells like endothelial and epithelial cells can be controlled by interleukins by promoting tissue repair and regeneration after damage [27].

By influencing insulin sensitivity and the breakdown of lipids, interleukins also contribute to metabolic regulation, and provide us a link between immune function to metabolic homeostasis [28]. The intricacy of interleukin's signaling pathways and pleiotropic effects points out both their importance as therapeutic targets in a variety of diseases and their vital role in sustaining health.

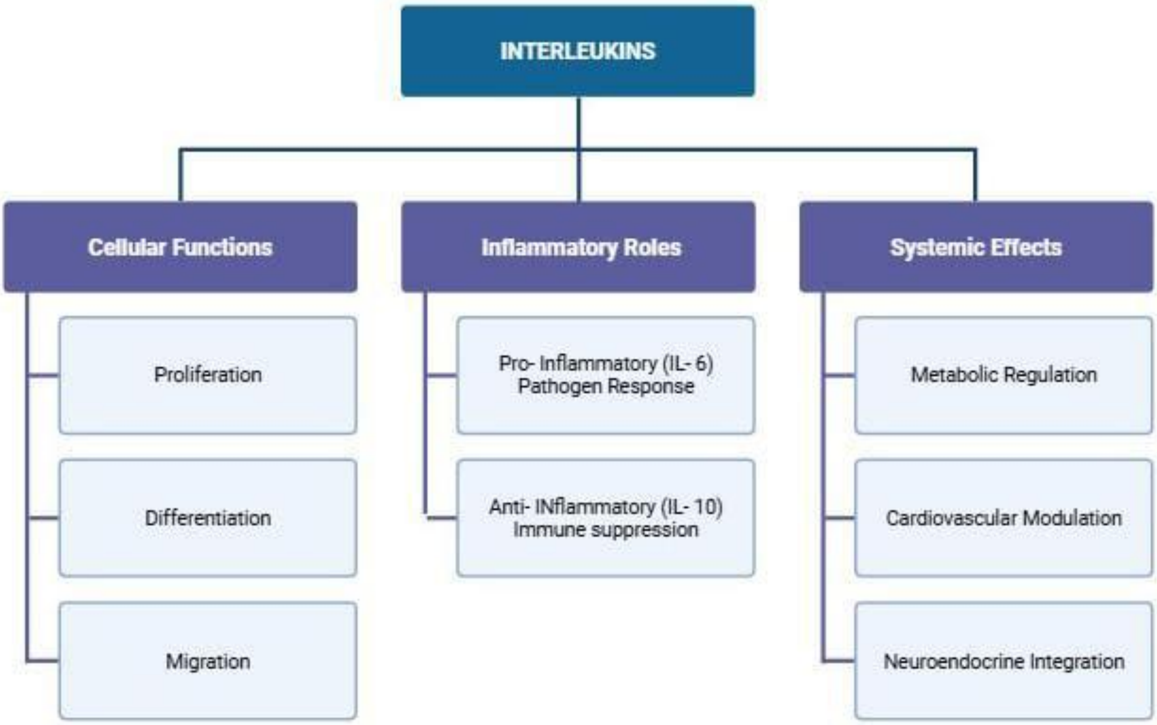

INTERLEUKINS
Cellular Functions
Inflammatory Roles
Systemic Effects
Proliferation
Differentiation
Migration
Pro- Inflammatory (IL- 6) Pathogen Response
Anti- INflammatory (IL- 10) Immune suppression
Metabolic Regulation
Cardiovascular Modulation
Neuroendocrine Integration

Figure 1: Functions of Interleukins

2.3 Interleukins and Peptide Immunogenecity

The communication and behavior of immune cells depend extensively on interleukins such as IL-2, IL-4, IL-6, IL-10, IL-12, and IL-17. For instance, T-cell proliferation necessitate IL-2, Th2 differentiation calls for IL-4, acute phase response involves IL-6, immune suppression demands IL-10, and pro-inflammatory response demands IL-17.The immunogenicity of peptides—that is, their capacity to elicit a cytokine response—is determined by their primary sequence, conformational structure, and physicochemical characteristics. Physical and structural features such as specific amino acid motifs, hydrophobicity, charge, and secondary structure tendency can determine immune receptor recognition of peptides and cytokine induction. Experimental discovery of IIPs is normally achieved by in vitro immune cell stimulation followed by ELISA, flow cytometry, or multiplex bead assay for cytokine quantification. All of these wet-lab methods are, however, impaired by limited scalability, variability in biological samples, and the intricate nature of immune signaling networks. Consequently, computational prediction has also become an essential adjunct to experimental discovery, facilitating efficient hypothesis generation and selection of candidate peptides for confirmation [29].

2.4 Biological Datasets and Features

The root of ML-based IIP prediction stems from the presence of high-quality, annotated datasets [21], [30]. Databases like the Immune Epitope Database (IEDB) contain experimentally confirmed interleukin-inducing and non-inducing peptides for different cytokines like IL-2, IL-4, IL-6, IL-10, IL-13, and IL-17. Dedicated resources and webservers like IL13Pred and iIL13Pred for IL-13, and IL17eScan for IL-17, offer pre-curated datasets and predictive tools [21], [16]. Peptide information are usually provided in formats like FASTA for sequence data and PDB for structural information. Feature extraction involves a variety of sequence-derived and physicochemical descriptors such as amino acid composition (AAC), dipeptide composition (DPC), position-specific scoring matrices (PSSM), and motifs of higher-order [18], [20], [25], [30]. Sophisticated feature engineering strategies, including the application of physicochemical property indices, secondary structure prediction, and pre-trained model-based structural embeddings (e.g., ESM-1b), also add more richness to the feature set and enhance the performance of the model [30]. Apart from the applications of extensive repositories and sophisticated feature extraction, recent works have also stressed the significance of curating non-redundant sets to limit bias and overfitting in the ML models for IIP prediction. For instance, the IL2pred tool contains modules that not only enable prediction of IL-2-inducing peptides but also scanning proteins for IL-2-inducing peptides and peptide analog ranking according to predicted activity, thus facilitating rational peptide design [27]. Likewise, iIL13Pred has shown that incorporating multivariate feature selection techniques, for instance, minimum redundancy maximum relevance (mRMR), can substantially improve model performance by determining the most informative and least redundant features from high-dimensional peptide data. Multimodal modeling approaches combining alignment-based strategies (e.g., BLAST), motif discovery tools (e.g., MERCI), and machine learning classifiers have also shown to improve IIP prediction accuracy and robustness, as evident in recent IL-17 research [30]. These combined methods, coupled with easy-to-use web servers, have become available to the wider scientific community due to ML-based IIP prediction, allowing for rapid and high-throughput in silico screening of candidate immunomodulatory peptides.

2.5 Data Preprocessing Strategies

Strong data pre-processing is essential for developing trustworthy ML models. Removal of redundancy, usually carried out through means such as CD-HIT, eliminates redundant datasets through clustering similar sequences and keeping representatives [30]. Balancing class imbalance—frequent in biological datasets—is obtained with methods like Synthetic Minority Over-sampling Technique (SMOTE) and random under-sampling, equating the number of positive (inducing) and negative (non-inducing) instances [31]. Feature scaling and normalization, i.e., min-max scaling and z-score standardization, are used to standardize feature distributions and ensure model convergence. Encoding techniques like one-hot encoding, BLOSUM matrices, and learned embeddings are used to encode peptide sequences in compatible formats with ML algorithms [32].

Moreover, aggressive cross-validation techniques, including k-fold and stratified cross-validation, are applied systematically to validate model generalizability and avoid overfitting, particularly for the case of small or unbalanced datasets. Utilization of external independent test sets, derived from recent experimental findings or withheld during training, also confirms model robustness and real-world validity. Recent research has also emphasized the strength of ensemble learning techniques, where predictions from several models are combined to increase overall accuracy and minimize variance. Interpretability methods—e.g., feature importance ranking and SHAP (SHapley Additive exPlanations) analysis—are increasingly used to clarify the biological significance of important features and to build confidence in ML-generated predictions. These robust pre-processing and validation approaches are the foundations of leading-edge ML pipelines for the prediction of interleukin-inducing peptides [33].

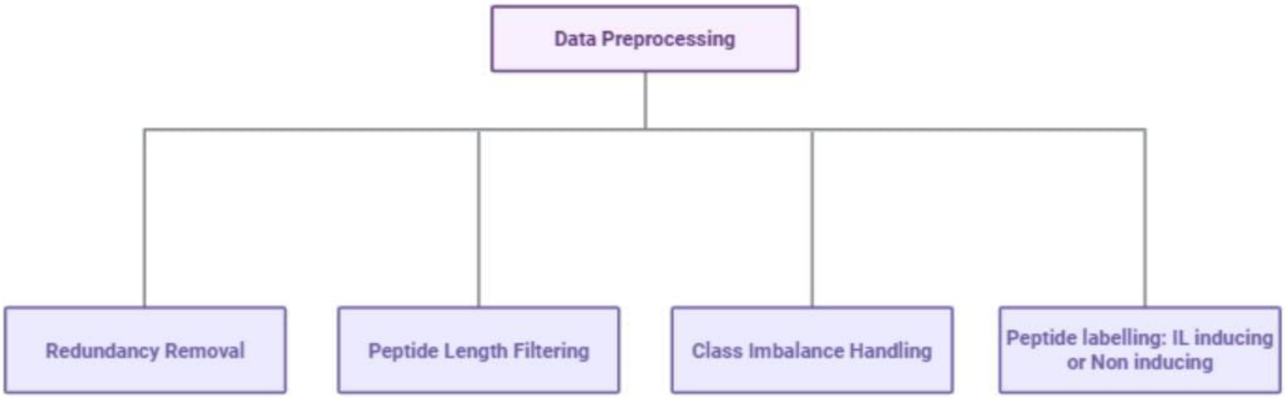


Figure 2: Steps of Data Preprocessing

2.6 Feature Extraction and Selection

In IIP prediction, feature extraction takes both manual and computational forms. Domain Expertise is used to calculate features like AAC, DPC, and physicochemical descriptors in conventional manual methods [21], [30], [31], [34]. While newer computational developments have incorporated embedding methods based on pre-trained language models (e.g., ESM-1b, ProtBert) that is based on the principle of learning context and structure from peptide sequences. Feature extraction plays a pivotal role to further enhance model explainability and reduce dimensionality [35]. Methods such as LASSO regression, analysis of variance (ANOVA), and minimum redundancy maximum relevance (mRMR) are also used frequently to get the features containing best information. Principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) are two dimensionality reduction techniques that helps in visualization and lower the likelihood of overfitting [34], [35]. Moreover, because hybrid feature sets provide better performance than descriptor or embedding alone because they can capture both global sequence features and subtle contextual information. Scientists have proved in their research that combining different feature selection techniques, such as mRMR with ANOVA or LASSO, can improve feature space refinement and produce more stable predictive models. Estimating class separability and identifying probable outliers or mislabeled samples received assistance from the visualization of high-dimensional embeddings using t-SNE and PCA In addition to improving classification accuracy, deep learning-based feature extraction from ESM-1b and ProtBert makes it easier to find novel sequence motifs relevant to interleukin induction. When building state-of-the-art IIP prediction models, these advancements emphasize the necessity of an integrated and iterative approach for feature engineering along with selection [36].

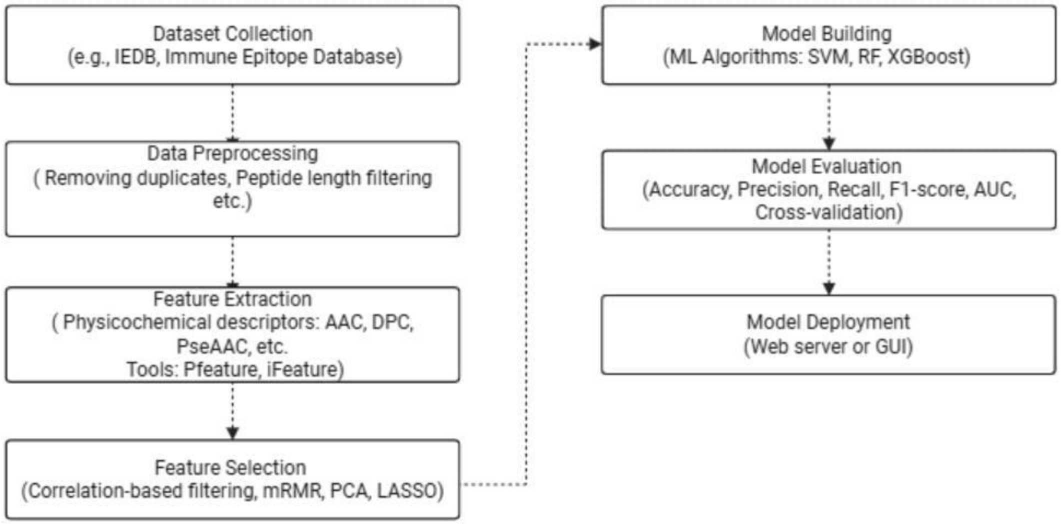

Dataset Collection
(e.g., IEDB, Immune Epitope Database)
Data Preprocessing
(Removing duplicates, Peptide length filtering etc.)
Feature Extraction
(Physicochemical descriptors: AAC, DPC, PseAAC, etc.
Tools: Pfeature, iFeature)
Feature Selection
(Correlation-based filtering, mRMR, PCA, LASSO)
Model Building
(ML Algorithms: SVM, RF, XGBoost)
Model Evaluation
(Accuracy, Precision, Recall, F1-score, AUC, Cross-validation)
Model Deployment
(Web server or GUI)

Figure 3: Worlflow for Developing AIML Model

2.7 Machine Learning Models for IIP Prediction

A wide variety of supervised learning models has been used to predict IIP [21], [37]. SVM, RF, k-NN, and logistic regression are commonly used because they are strong and easy to interpret [30], [37]. Ensemble learning methods, such as gradient boosting, XGBoost, and LightGBM, provide better predictive accuracy through the combination of multiple base learners [26], [34]. Deep learning algorithms have come to the forefront, where convolutional neural networks (CNNs) are outstanding at identifying local sequence motifs and recurrent neural networks (RNNs, LSTMs) at modeling sequential relationships [33], [38]. Transformer-based systems, including ProtBert, make use of attention mechanisms to represent long-range interactions and have been useful in peptide classification problems [34], [38]. Interestingly, the latest research has also investigated graph neural networks (GNNs) with three-dimensional (3D) structural information, as in the case of DGIL-6 for IL-6-inducing peptide prediction, which considers sequence-based and structure-based features for improved accuracy [38]. Recent comparative studies have shown that hybrid models, which combine both traditional machine learning and deep learning methods, tend to perform better than single-model strategies in terms of precision and transferability. For example, DGIL-6 amalgamates graph neural networks with deep embeddings to efficiently leverage both sequence and structural data, leading to better performance in independent test sets. In addition, architectures based on GNN have also demonstrated specific potential in representing spatial relationships between amino acids, which are of utmost importance for the functional activity of interleukin-inducing peptides. Transfer learning has also been put forward through studies by using pre-trained transformer models such as ProtBert, which is fine-tuned against IIP datasets, and which allows for quick adaptation across novel cytokine targets under minimal labeled data. Ensemble stacking techniques, which combine the predictions of various base learners, have been utilized to enhance predictive reliability even further as well as to overcome the shortcomings of single algorithms. These methodological improvements highlight the dynamic and changing nature of computational IIP prediction, opening the door to increasingly precise and biologically driven models [34], [39].

2.8 Performance Matrices and Validation Techniques

The assessment of ML models for IIP prediction uses conventional classification measures, such as accuracy, precision, recall, and F1 score, which together measure the balance between true positive and false positive rates [40]. Receiver Operating Characteristic Area under the Curve (ROC-AUC) and Precision-Recall AUC (PR-AUC) offer threshold-independent estimations of model discrimination, especially useful in imbalanced data [41].

Cross-validation techniques, including k-fold and stratified cross-validation, are commonly applied to estimate model generalizability and prevent overfitting. Independent test of a dataset by the use of external, unseen data is the gold standard for the evaluation of real-world predictive accuracy [42]. Apart from these baseline metrics, Matthews Correlation Coefficient (MCC) is also commonly suggested for imbalanced datasets since it offers balanced assessment even when class distributions are imbalanced. Brier scores and calibration plots are also used to measure the reliability of predicted probabilities so that the model's confidence is consistent with actual cases. Current guidelines prefer the reporting of multiple measures—over accuracy alone—to present a complete performance profile, particularly in biomedical contexts where false positives or false negatives may result in serious consequences. In addition, testing protocols that are independent recommend the application of temporally or experimentally separated datasets to critically assess model robustness and avoid the potential for data leakage [43]. These best practices, in conjunction with transparent reporting and extensive validation, are essential for establishing the credibility and translational value of ML-based IIP prediction tools.

. Table 1: A comparative summary of recent ML-based tools for IIP prediction

Tool/Server	Interleukin Target	Interleukin Target	Features Used	Performance (AUC/Accuracy)	Reference
iIL13Pred	IL-13	IL-13	mRMR-selected, DPC, AAC	AUC: 0.83, MCC: 0.33	[44]
IL13Pred	IL-13	IL-13	DPC, AAC	Lower than iIL13Pred	[45]
DGIL-6	IL-6	IL-6	Structural + sequence	Noted improved accuracy	[46]
IL17eScan/iIL17Pred	IL-17	IL-17	DPC, BLAST, MERCI	AUC: 0.88, MCC: 0.68	[47]
IL10Pred	IL-10	IL-10	DPC, AAC	Accuracy: 81.24%, MCC: 0.59	[48]
ILeukin10Pred	IL-10	IL-10	Sequence-based, hybrid features	Accuracy: 87.5%, MCC: 0.755	[49]
StackIL10	IL-10	IL-10	AAC, TPC, APAAC, DPC, etc.	Accuracy: 81.24%, MCC: 0.59	[50]
IL2pred	IL-2	IL-2	Dipeptide composition, peptide length	AUC: 0.84 (ensemble), MCC: 0.51	[51]
IL4Pred2	IL-4	IL-4	Sequence-derived, similarity, motifs	AUC: 0.80, MCC: 0.45 (human/mouse)	[52]

2.9 Challenges and Limitations

Even with considerable advancements, a number of issues still hinder ML-based IIP prediction. Relatively low access to experimentally verified data limits model training and external validation, especially for less-well-studied interleukins. Biases in training sets, e.g., an overabundance of some peptide motifs or experimental error, can reduce model generalizability. Immune response complexity as well as peptide immunogenicity's multifactorial nature pre sent further challenges.

In addition, interpretability of deep models is still limited, hampering biological insight and translation to the clinic. Overcoming such limitations calls for sustained efforts in data curation, model explainability and fusion of multi-modal information.

2.10 Future Research Directions

Future IIP prediction research stands to gain from the combination of multi-omics data, including genomics, transcriptomics, and proteomics, to capture a comprehensive picture of immune regulation [53]. Host genetic variability like HLA alleles need to be accounted for in personalized prediction models for precision immunotherapy and vaccine design [54]. The advancement of explainable AI (XAI) methods has the potential to improve the interpretability of ML models, supporting biological discovery as well as hypothesis generation. Moreover, automated integration of IIP prediction pipelines with vaccine design software will expedite the translation of computational insights into clinic-approved applications [55].

2.11 Conclusion

Machine learning has revolutionized the field of interleukin-inducing peptide prediction as it provides scalable, accurate, and stable alternatives to experimental approaches. Innovation in feature engineering, model design, and validation techniques has led to the creation of cutting-edge tools for various interleukins, with potential applications in immunotherapy and vaccine development. Future advancements in this field will depend on increasing high-quality data, enhancing model explainability, and multi-disciplinary collaboration. Finally, ML- and immunology convergence promises a new era of translational research, with important implications for human health and disease control.

CHAPTER 3. METHEDOLOGY

3.1 Data Extraction from IEDB

i. **Protein Epitope Type Selection**: *Linear peptide* was chosen, ensuring that only continuous amino acid sequences are included in the dataset. Discontinuous and non- peptidic epitopes are excluded, focusing the analysis on linear peptide antigens.

ii. **Assay Type and Outcome:** *T Cell* and *B Cell* assays, as well as *MHC Ligand* data, were selected. This comprehensive approach allows collection of peptides tested in various immune contexts. The assay filter specifically includes "IL-2 release" or "biological assay," ensuring that only peptides evaluated for their ability to induce IL-2 production are retrieved. Both *Positive* and *Negative* outcomes were selected, allowing for the creation of datasets of IL-2 inducers and non-inducers, which is essential for supervised machine learning model development.

iii. **MHC Restriction**: The *Any* option is selected for MHC restriction, meaning peptides restricted by any MHC class (I, II, or non-classical) are included. This broadens the dataset and allows later filtering by specific MHC types if required for downstream analysis.

iv. **Epitope Source**: No specific organism, antigen, or host was pre-selected, but the interface allows for further filtering by organism (e.g., *Homo sapiens*) or antigen if needed.

v. **Host and Disease***: All* host and *Any* disease state were selected, capturing peptides tested in all available host organisms and disease contexts. This maximizes dataset size and diversity, with the option to restrict to human data or specific disease states during preprocessing.

3.2 Data Preprocessing

i. **Removal of Amino Acids After '+' Sign and the '+' itself:** For any peptide containing a '+' character, both the '+' and the amino acid immediately following it were removed. This step addresses potential post-translational modification annotations or concatenated peptide artifacts.

ii. **Exclusion of Peptides with Unnatural Amino Acids**: Peptides containing non-standard or ambiguous amino acids (B, J, O, U, X, and Z) were filtered out. This ensures only naturally occurring amino acid sequences remain for analysis.

iii. **Removal of Redundancy**: Duplicate peptide sequences were eliminated from each dataset, resulting in a non-redundant list of unique peptides.

iv. **Length Filtering**: Only peptides with lengths of 8 to 30 amino acids were retained. Sequences shorter than 8 or longer than 30 residues were excluded, focusing the dataset on peptides most relevant for MHC binding and immunogenicity studies.

v. **Frequency Calculation**: The frequency (occurrence count) of each unique peptide was determined, although after deduplication, each peptide appears only once in the filtered dataset.

vi. **Saving Results**: The final lists of filtered peptides for both positive and negative datasets were saved as new files for further analysis.

vii. **Removal of Overlapping Peptides**: After preprocessing, peptides present in both the positive and negative datasets were identified as common. These overlapping sequences were removed from both datasets, ensuring that each peptide is exclusive to either the positive or negative set, eliminating class ambiguity.

3.3 Exploratory Data Analysis and Visualization

i. **Peptide Length Distribution Analysis:** The peptide length distribution was analyzed for both positive and negative datasets. Histograms and boxplots were generated to visualize the range and central tendency of peptide lengths.

ii. **Amino Acid Composition (AAC) Analysis :** The relative frequency of each amino acid was calculated for both classes. Bar plots were created to compare the enrichment or depletion of specific residues.

iii. **Dipeptide Composition (DPC) Analysis:** The occurrence of all 400 possible dipeptide pairs was computed, normalized by the total number of dipeptides in each class. Heatmaps were generated to visualize the most and least frequent dipeptides, highlighting sequence motifs potentially relevant for IL-2 induction.

iv. **Sequence Logo Generation:** To assess positional preferences, sequence logos were generated for both N- and C-terminal residues using R's ggseqlogo package or Python's WebLogo. This approach revealed enrichment of certain residues (e.g., hydrophobic or positively charged) at specific positions in IL-2 inducers, and depletion of others (e.g., hydrophilic or acidic residues).

v. **Motif Discovery:** Motif analysis was performed using MERCI software, identifying short sequence patterns unique to either IL-2 inducers or non-inducers. Motifs were reported in tabular format, including their frequency and exclusivity, and visualized as motif logos for interpretability

vi. **Physicochemical Property Analysis**: Using Python packages such as peptides.py and modlAMP, a suite of physicochemical descriptors was computed for each peptide, including:
   a. Hydrophobicity (Kyte-Doolittle scale)
   b. Isoelectric point
   c. Molecular weight

3.4 Software and Tools used

Table 2. Overview of software tools used in this study, including their primary functions and official access links.

Purpose	Tool/Package	Language	Reference/Functionality	Reference (Link)
Data handling and statistics	pandas	Python	Data import, cleaning, descriptive statistics, plotting	pandas.pydata.org [56]
Sequence feature extraction	modlAMP, peptides.py	Python	Compute AAC, DPC, and physicochemical descriptors	modlAMP, peptides.py [57]
Visualization	matplotlib, seaborn	Python	Generate histograms, boxplots, bar charts, and heatmaps	matplotlib.org, seaborn.pydata.org [58]
Sequence logo generation	ggseqlogo, RWebLogo	R	Create sequence logos for positional amino acid analysis	ggseqlogo CRAN, RWebLogo [59]
Motif discovery	MERCI	Standalone	Discover conserved motifs exclusive to IL-2 inducing peptides	MERCI Tool [60]
Machine learning (optional)	scikit-learn	Python	Feature selection and modeling (not used in this study)	scikit-learn.org [61]
Notebook environment	Jupyter Notebooks	Python	Reproducible interactive workflow for data exploration and analysis	jupyter.org [62]
Data management	Microsoft Excel	-	Manual curation and summary table creation	Microsoft Excel [63]

CHAPTER 4. RESULT AND DISCUSSION

4. 1 Peptide length analysis

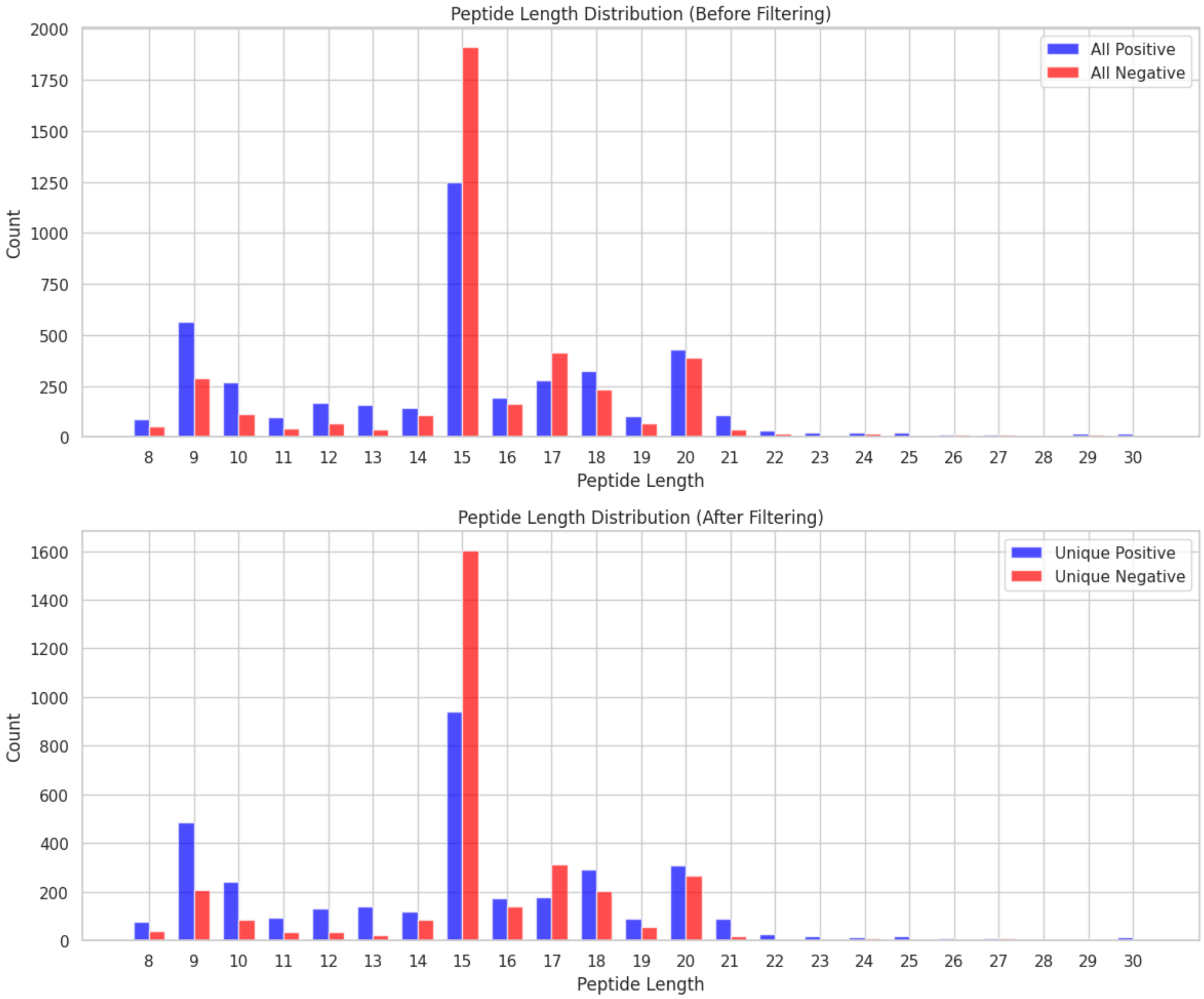


Figure 4: Positive and Negative Peptide Length Distribution before and after filtering

Before filtering, both positive and negative datasets show a pronounced peak at 15 amino acids, with additional smaller peaks at 9, 17, and 20 residues. This reflects the experimental design and biological relevance, as peptides of these lengths are commonly tested in immunological assays. While, after filtering, the overall shape of the distribution is preserved, but the total counts decrease, especially for the most frequent lengths (e.g., 15-mers). This reduction indicates that many redundant or overlapping sequences have been removed, resulting in a cleaner, non-redundant dataset. The negative class consistently shows a higher count of 15-mer peptides compared to the positive class, both before and after filtering. This suggests that 15-mers are more frequently tested but not necessarily more likely to induce IL-2, highlighting a potential experimental bias. The positive class maintains a broader distribution across multiple lengths, indicating that IL-2 inducers are not confined to a single peptide length.

Filtering for uniqueness and removing overlaps ensures that the dataset used for model training is free from bias caused by repeated or ambiguous sequences. This step is crucial to prevent overfitting and to ensure that the model learns generalizable features rather than memorizing redundant examples. The persistence of length distribution patterns after filtering suggests that peptide length remains a relevant feature for distinguishing between IL-2 inducers and non-inducers, though it should be used in conjunction with other sequence-based features. The dominance of 15-mers likely reflects their use in overlapping peptide libraries for T-cell epitope mapping, particularly for MHC class II studies. The presence of unique peptides across a range of lengths in the positive set supports the biological diversity of IL-2 inducing sequences and suggests that models should not be restricted to a narrow length window.

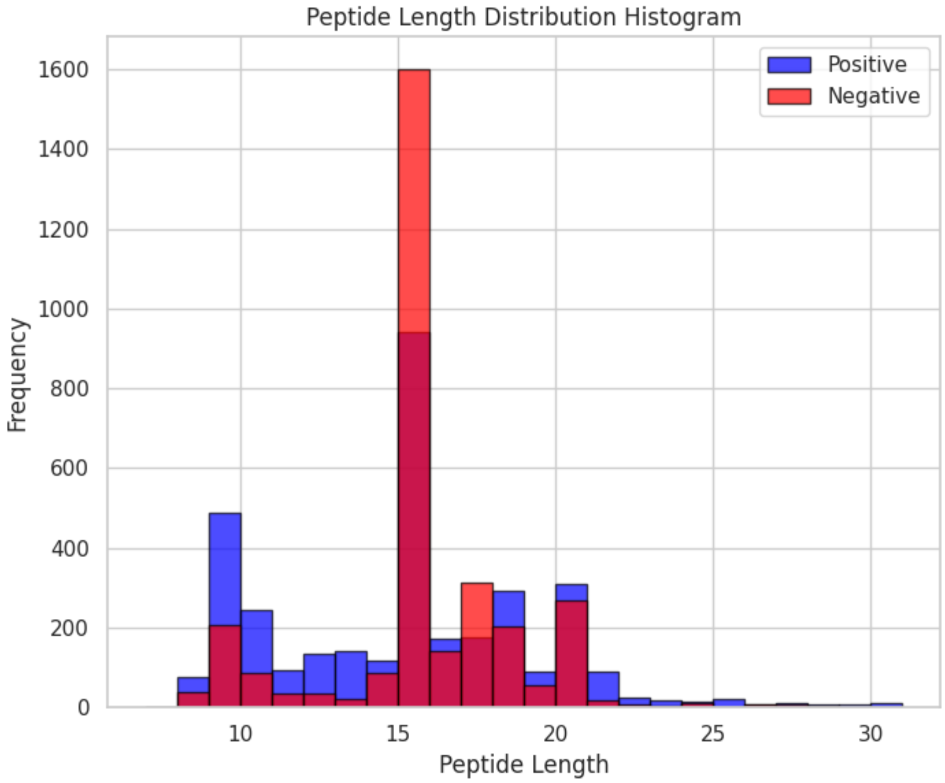


Figure 5: Peptide Length Distribution of positive and negative unique dataset

Both positive and negative peptides exhibit a broad range of lengths, but the majority cluster between 8 and 20 amino acids, with a pronounced peak around 15 residues. The boxplot shows that positive peptides have a slightly wider length distribution, including more outliers at both shorter and longer lengths, while negative peptides are more tightly clustered around the median. The overlap in length distributions suggests that length alone may not be a strong predictor of IL-2 induction. However, the presence of subtle differences—such as a higher median and greater variability in the positive class—indicates that peptide length could still contribute as a supportive feature in machine learning models. Including peptide length as a feature may help models capture sequence characteristics associated with IL-2 induction, especially when combined with other sequence-derived descriptors. The enrichment of peptides around 15 amino acids aligns with the optimal length for MHC class II binding, which is relevant for T-cell mediated IL-2 responses. The presence of longer and shorter peptides among inducers may reflect biological diversity in antigen processing or presentation pathways. These findings support the inclusion of peptide length as a basic feature in predictive modeling, but also highlight the need for more complex features (e.g., sequence composition, motif presence) to improve classification performance. The visualization also justifies preprocessing steps that filter out extreme lengths, focusing the dataset on biologically relevant peptides.

4.2 Amino Acid Composition (AAC) Analysis

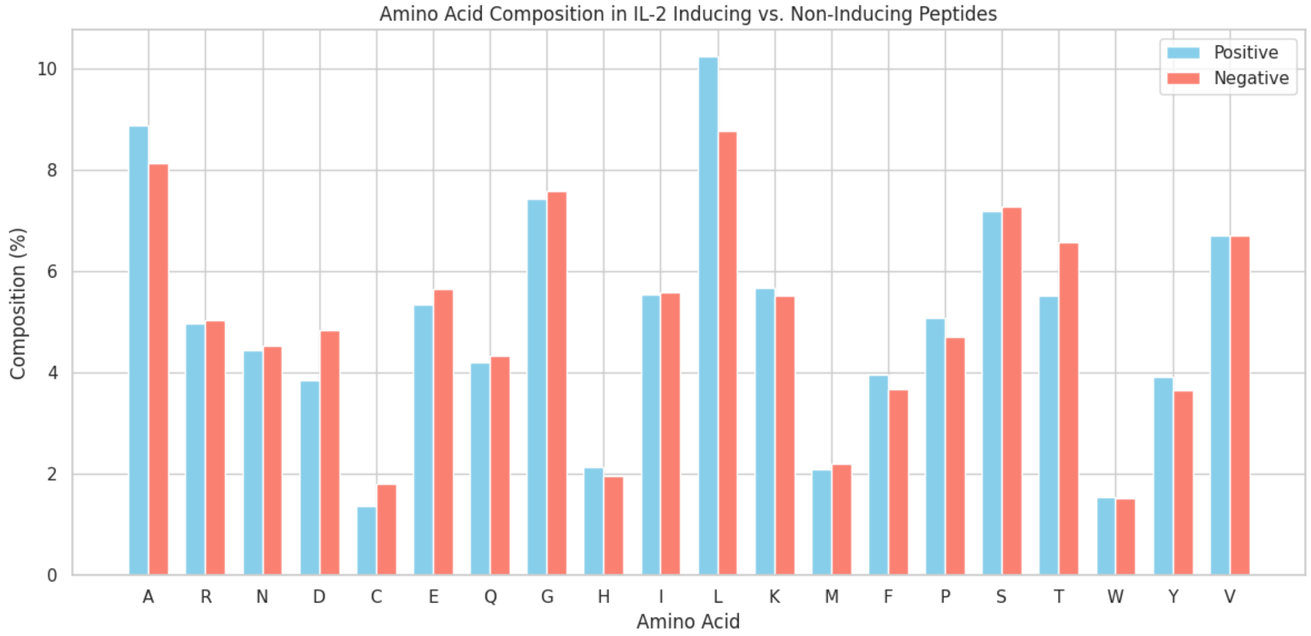


Figure 6: Amino Acid Composition of positive and negative dataset

Certain amino acids, such as alanine (A) and leucine (L), are more abundant in IL-2 inducing peptides compared to non-inducers. For example, leucine shows the highest enrichment in the positive class, suggesting a potential role in IL-2 induction. Conversely, some residues like serine (S) and glutamine (Q) display similar or slightly higher frequencies in the negative class, indicating they may not be as strongly associated with IL-2 induction. The observed differences in amino acid frequencies between the two classes highlight the potential of amino acid composition as a discriminative feature for machine learning models. Amino acids with the largest frequency differences (e.g., L, A, I) can serve as important input variables, helping models distinguish between IL-2 inducers and non-inducers. These compositional biases may reflect underlying structural or functional motifs relevant to T-cell activation and cytokine release, further justifying their inclusion as features. The enrichment of hydrophobic residues (e.g., L, I, A) in IL-2 inducers could indicate a preference for certain physicochemical properties that facilitate MHC binding or T-cell receptor recognition, both critical for cytokine induction. The relatively uniform distribution of other residues suggests that while some amino acids are important, IL-2 induction is likely influenced by a combination of sequence features rather than a single dominant residue. These findings support the use of amino acid composition (AAC) as a baseline feature set in predictive modeling. Additionally, the plot suggests that further exploration of position-specific composition or higher-order features (e.g., dipeptide composition, motif analysis) may yield even greater discriminative power.

4.3 Dipeptide Composition (DPC) Analysis

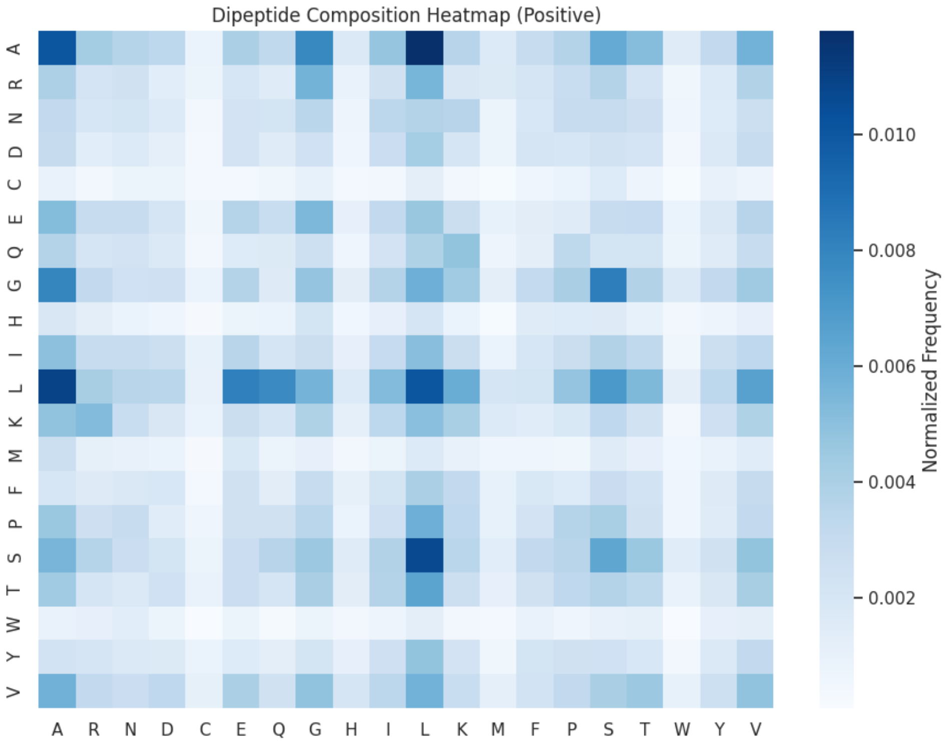


Figure 7: Dipeptide Composition of positive dataset

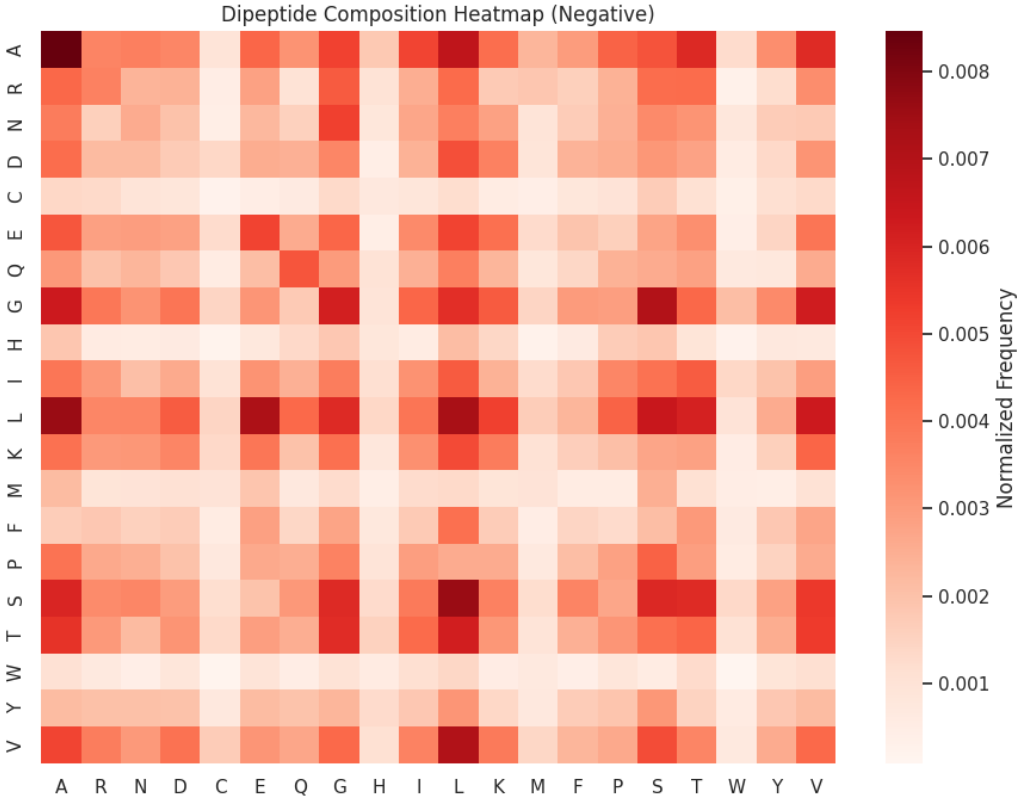


Figure 8: Dipeptide Composition of negative dataset

The dipeptide composition heatmaps for IL-2 inducing (positive) and non-inducing (negative) peptides reveal distinct patterns in the frequency of dipeptide pairs across the two classes. In the positive dataset, the distribution of dipeptide frequencies appears more uniform, with a few specific dipeptides showing higher enrichment, as indicated by darker blue shades. This suggests that certain dipeptide motifs may be preferentially associated with IL-2 induction, potentially reflecting sequence patterns important for T-cell activation and cytokine release. In contrast, the negative dataset exhibits a more heterogeneous distribution, with several dipeptides showing relatively higher frequencies (darker red), indicating that non-inducing peptides may be characterized by different or more repetitive dipeptide patterns. These compositional differences are significant for machine learning model development, as they highlight the potential of dipeptide composition features to discriminate between IL-2 inducers and non-inducers. Incorporating dipeptide frequencies as input variables can enhance the model’s ability to capture subtle sequence motifs and dependencies that are not apparent from amino acid composition alone. Biologically, the enrichment of specific dipeptides in the positive class may point to underlying structural or functional motifs required for effective immune signaling. Overall, this analysis underscores the value of dipeptide composition as an informative feature set, supporting both the interpretability and predictive power of computational models for IL-2 inducing peptide identification.

4.4 Sequence Logo Generation

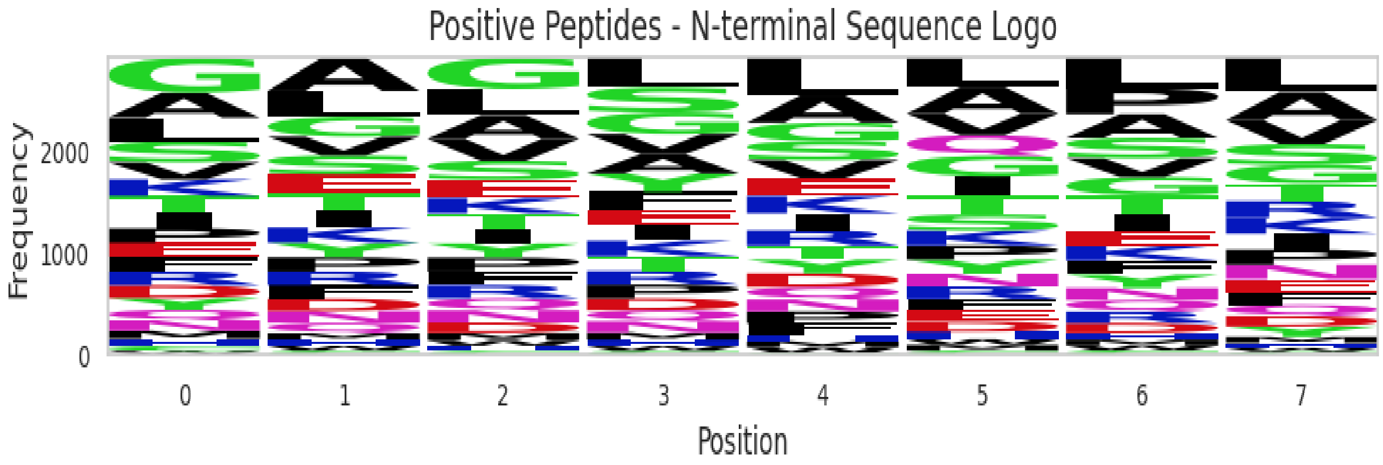


Figure 9: N-terminal sequence logo for positive peptides

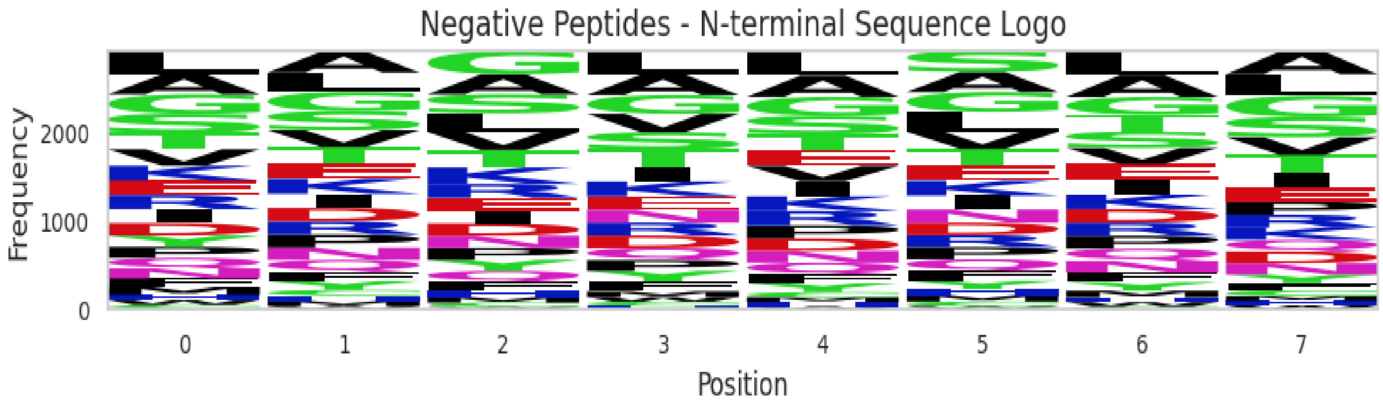


Figure 10: N-terminal sequence logo for negative peptides

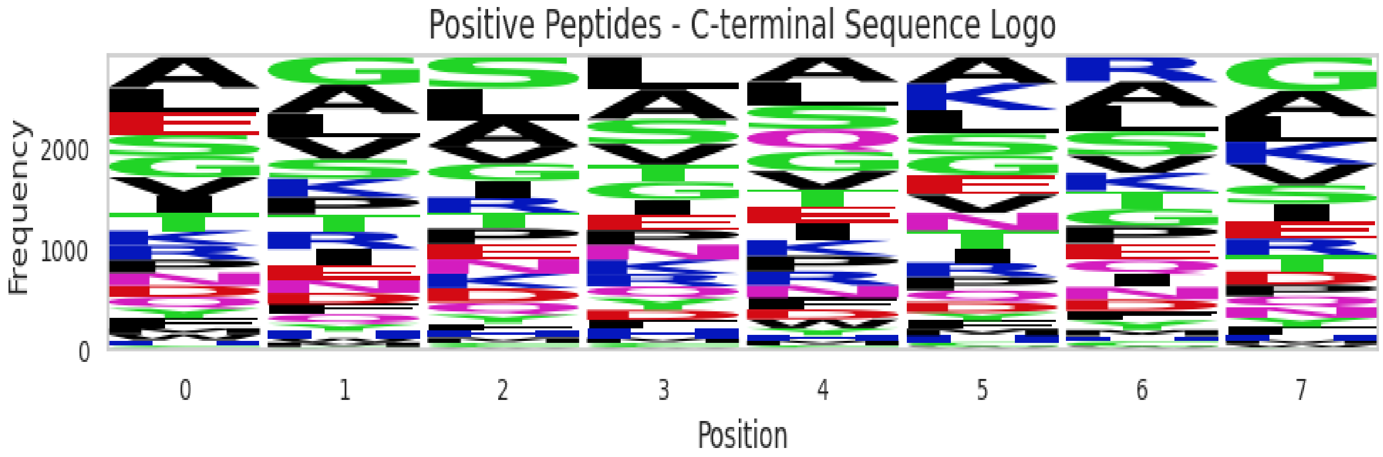


Figure 11: C-terminal sequence logo for negative peptides

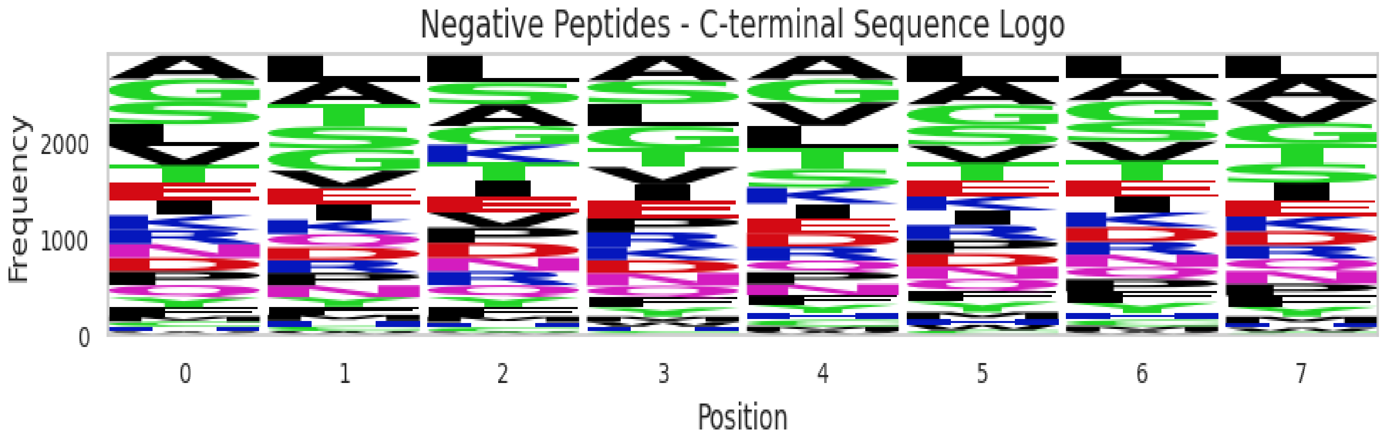


Figure 12: C-terminal sequence logo for negative peptides

The sequence logo analysis of both N-terminal and C-terminal regions for IL-2 inducing (positive) and non-inducing (negative) peptides provides valuable insights into position-specific amino acid preferences that can influence peptide function and immunogenicity. In the sequence logos for positive peptides, certain amino acids such as glycine (G), alanine (A), and leucine (L) are prominently enriched at specific positions, as indicated by their larger letter heights, suggesting a strong positional conservation and potential importance in IL-2 induction. In contrast, the logos for negative peptides display a more diverse and less conserved pattern, with no single amino acid dominating at most positions, reflecting higher sequence variability. This contrast implies that IL-2 inducing peptides may rely on specific sequence motifs or structural features at their termini, which could be critical for MHC binding or T-cell receptor recognition. The observed conservation in positive peptides, especially at the N-terminus, supports the hypothesis that certain residue arrangements are favored for biological activity, while the lack of such patterns in negatives may contribute to their non-inducing nature. From a machine learning perspective, these position-specific enrichment patterns provide a rationale for incorporating positional encoding or motif-based features into predictive models, as they capture information beyond global amino acid composition. Overall, the sequence logo analysis highlights the functional relevance of terminal motifs in IL-2 induction and underscores the potential of using such positional features to enhance the interpretability and accuracy of computational models for peptide immunogenicity prediction.

4.5 Motif Detection

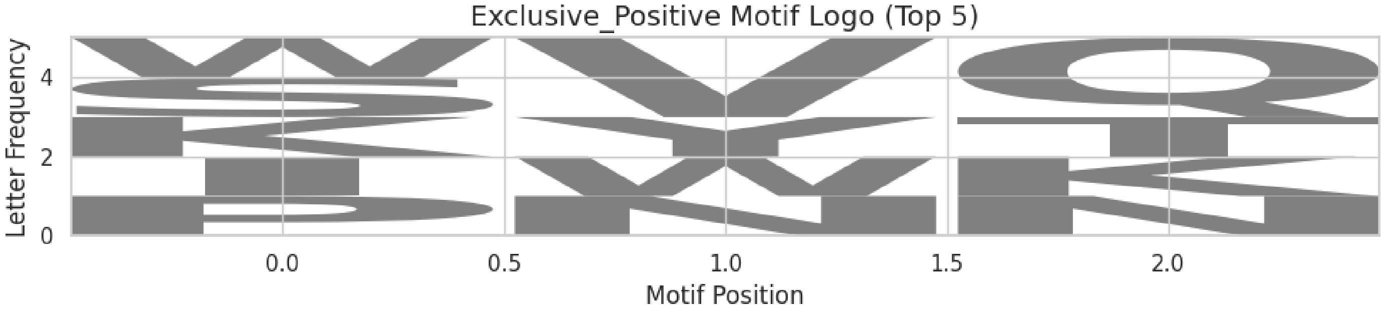


Figure 13: Top 5 Motif logo for positive peptides

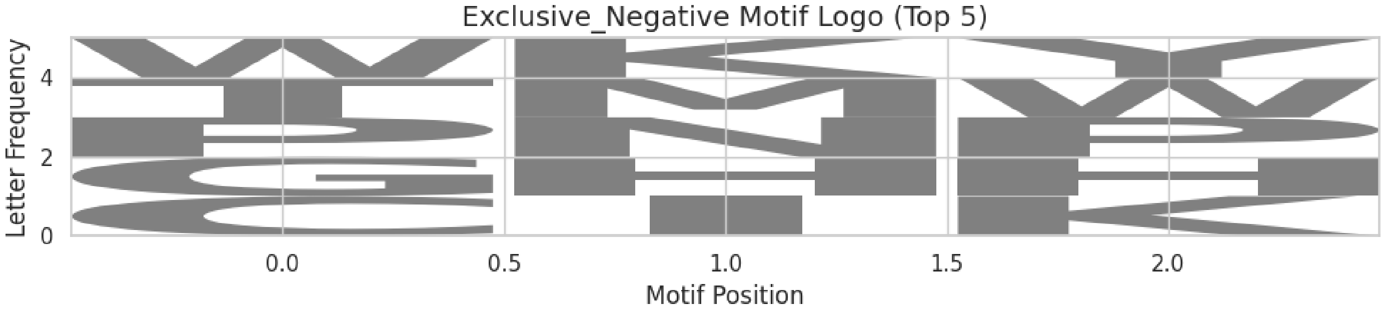


Figure 14: Top 5 Motif logo for negative peptides

The motif logo plots for the top five exclusive motifs in IL-2 inducing (positive) and non-inducing (negative) peptides provide important insights into the sequence patterns that distinguish these two classes. In the positive motif logo, certain amino acids—such as W, Y, and Q—are highly conserved at specific motif positions, as indicated by their larger letter heights. This strong positional conservation suggests that these residues may be critical for the biological activity associated with IL-2 induction, potentially facilitating specific interactions with MHC molecules or T-cell receptors. In contrast, the negative motif logo displays a different pattern, with residues like W, Y, and G appearing prominently but at different positions or with less conservation, reflecting alternative sequence motifs that may lack the capacity to trigger IL-2 release. The presence of distinct, highly conserved motifs in the positive set underscores the functional importance of particular short sequence patterns in driving immunogenicity. For machine learning applications, these motifs serve as highly informative features: their presence or absence can be encoded as binary or frequency-based variables, enhancing the model's ability to discriminate between inducers and non-inducers. Furthermore, the exclusivity of these motifs to either the positive or negative class reduces ambiguity and improves the interpretability of predictive models. Overall, the motif detection results highlight the biological and computational relevance of short, conserved sequence patterns in peptide immunogenicity, supporting their integration into feature engineering pipelines for robust IL-2 induction prediction.

4.6 Physiochemical Property Analysis

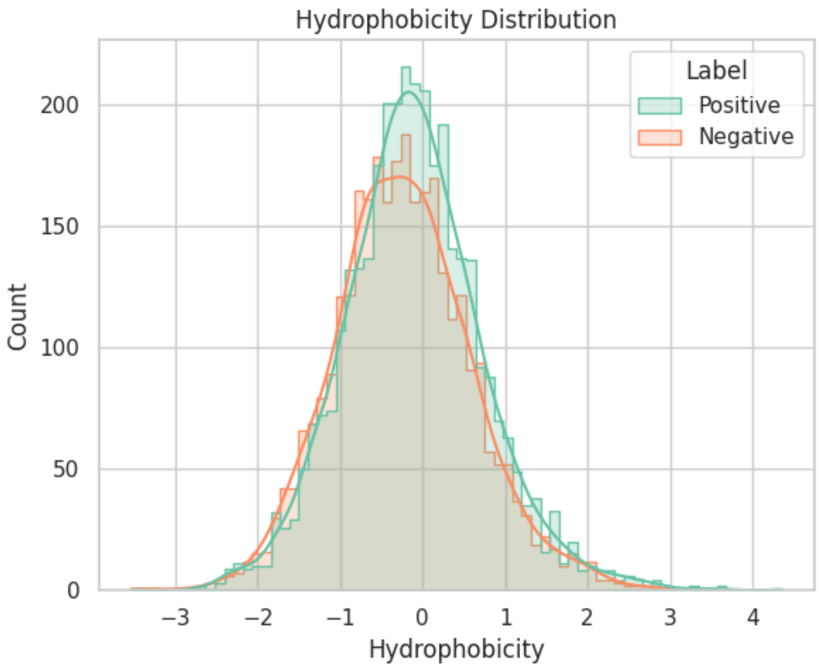


Figure 15: Hydrophobicity distribution of positive and negative peptides

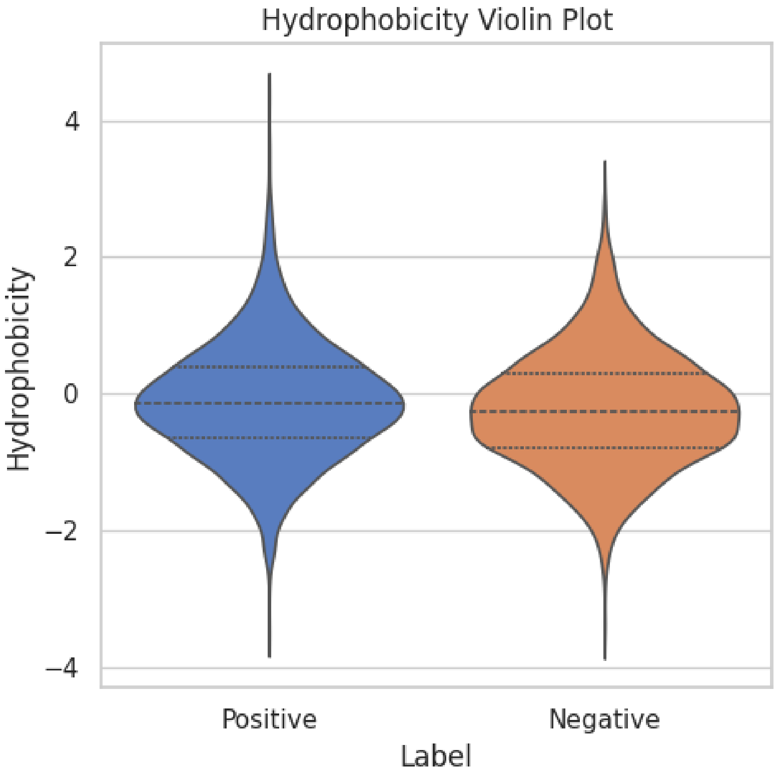


Figure 16: Hydrophobicity distribution of positive and negative peptides

The hydrophobicity distribution and violin plots reveal that both IL-2 inducing (positive) and non-inducing (negative) peptides exhibit a broadly similar, approximately normal distribution of hydrophobicity values, centered near zero. However, positive peptides show a slightly higher frequency of values around the mean and a marginally broader spread, indicating greater variability in hydrophobicity among inducers. The violin plot further suggests that while the central tendency is comparable between the two classes, positive peptides may include more extreme hydrophobic and hydrophilic sequences. These subtle differences imply that hydrophobicity, while not a sole distinguishing factor, could contribute as a supportive feature in machine learning models, potentially capturing nuanced physicochemical patterns relevant to IL-2 induction.

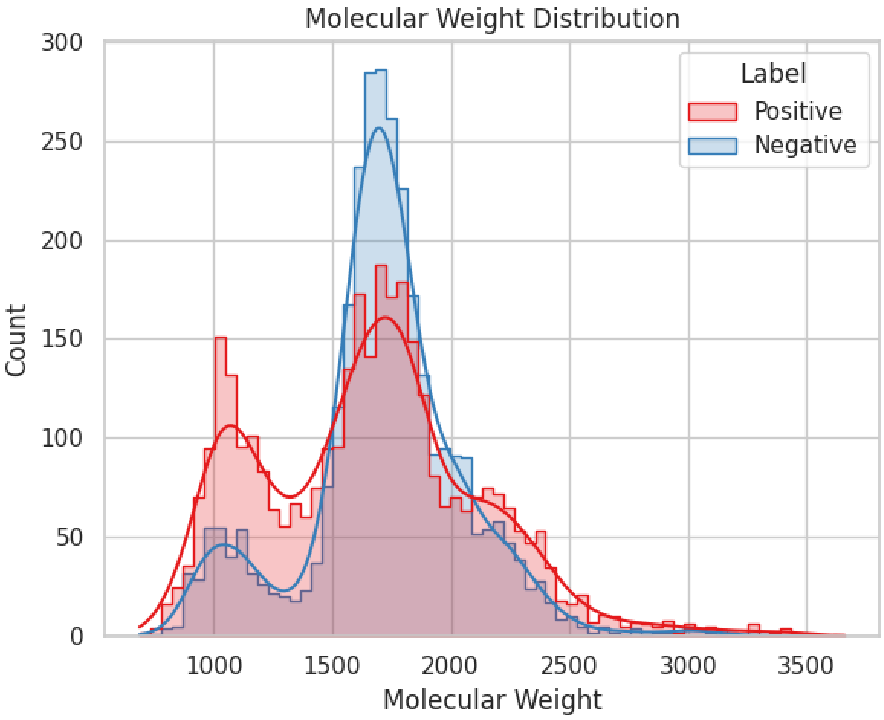


Figure 17: Molecular weight distribution of positive and negative peptides

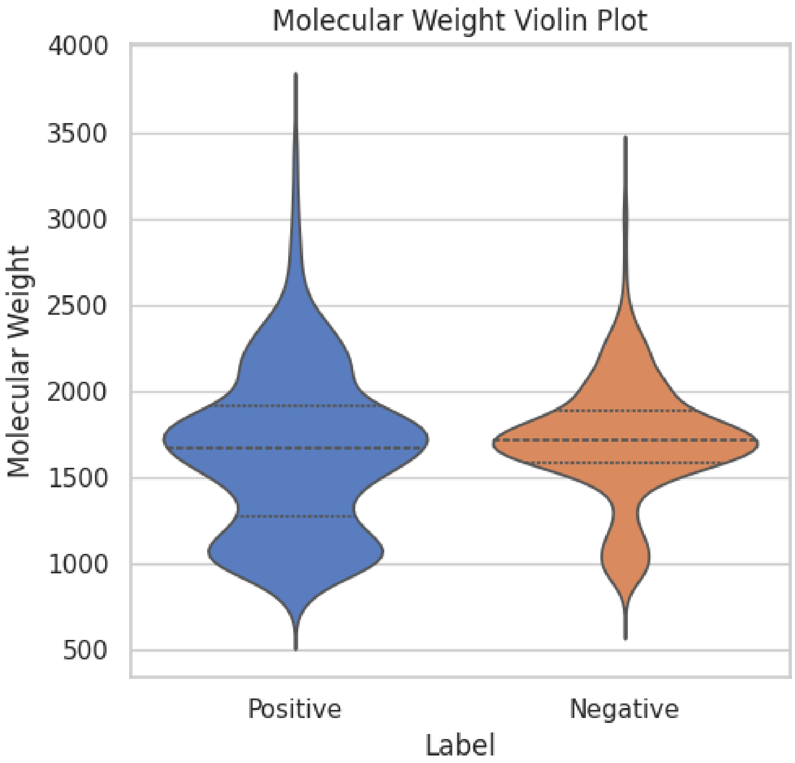


Figure 18: Molecular weight distribution of positive and negative peptides

The molecular weight distribution plots show that both IL-2 inducing (positive) and non-inducing (negative) peptides span a similar range, but positive peptides display a broader and more variable distribution, with multiple peaks and a higher frequency of both lower and higher molecular weights. In contrast, negative peptides are more tightly clustered around the central peak. The violin plot further confirms this greater diversity in molecular weight among inducers. These differences suggest that molecular weight, while not a sole discriminator, may contribute as a useful feature in machine learning models by capturing underlying sequence diversity relevant to IL-2 induction12.

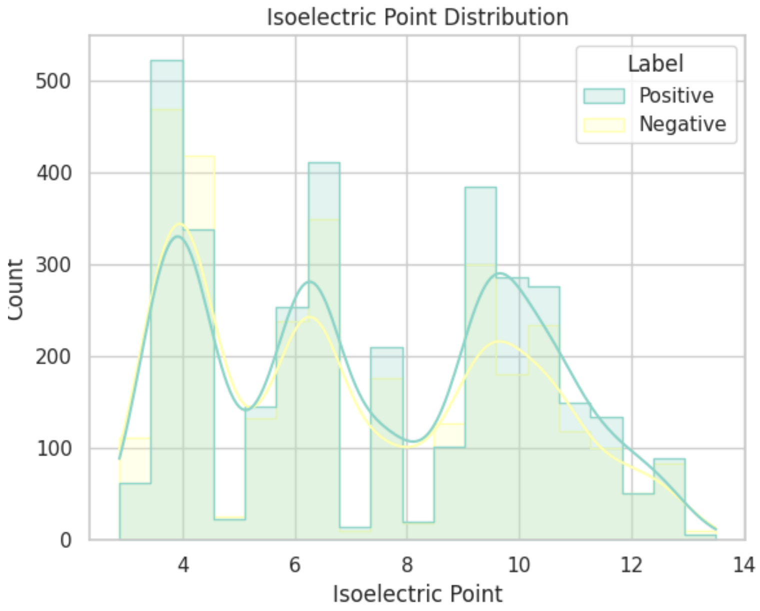


Figure 19: Isoelectric Point distribution of positive and negative peptides

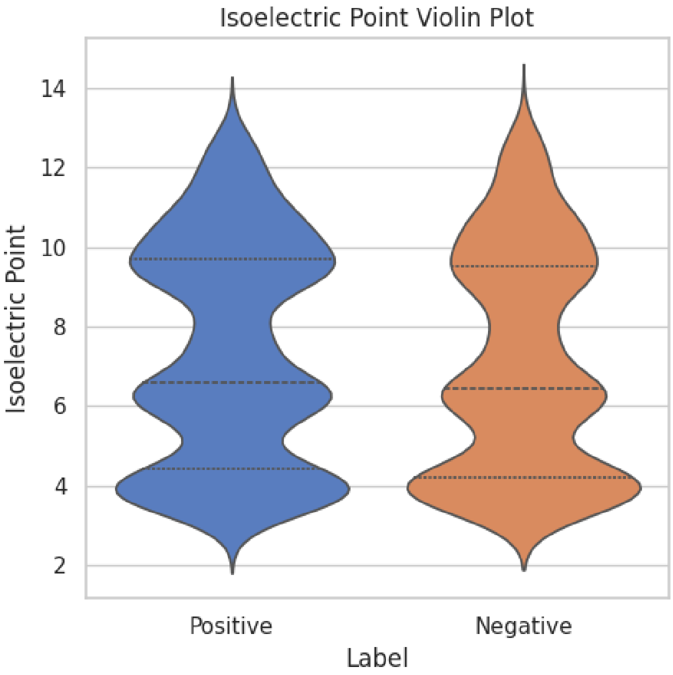


Figure 20: Isoelectric Point distribution of positive and negative peptides

The isoelectric point distribution and violin plots show that both IL-2 inducing and non-inducing peptides span a wide range of pI values, with multiple peaks observed in both classes. While the overall distribution patterns are similar, positive peptides display slightly higher frequencies at certain pI ranges and a marginally broader spread, suggesting greater diversity in their isoelectric points. These subtle differences indicate that isoelectric point could serve as a supportive feature for machine learning models, potentially capturing physicochemical nuances relevant to IL-2 induction.

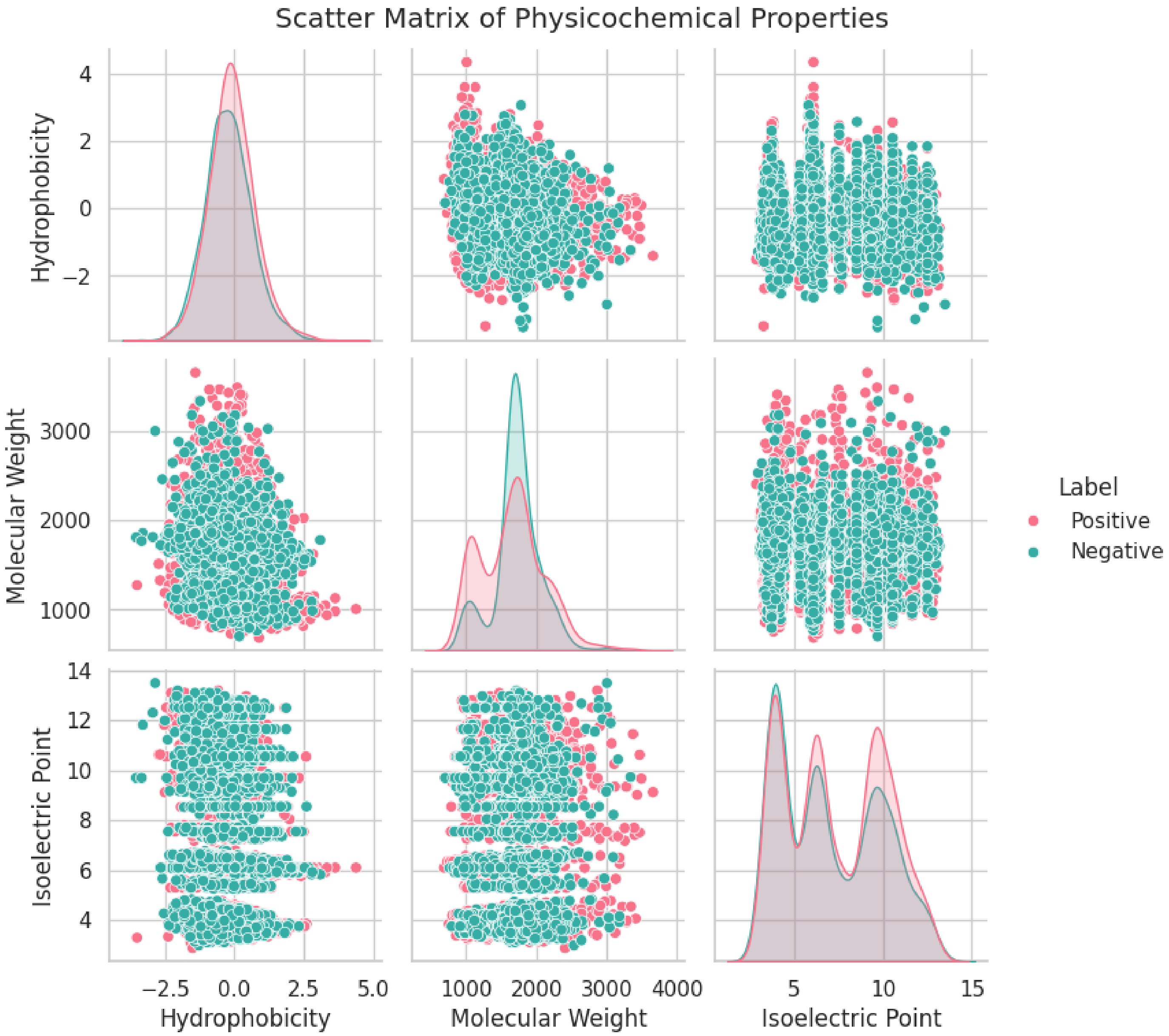


Figure 21: Scatter Matric of different Physiochemical properties of positive and negative peptides

The scatter matrix of physicochemical properties illustrates the relationships among hydrophobicity, molecular weight, and isoelectric point for IL-2 inducing (positive) and non-inducing (negative) peptides. The diagonal plots show that both classes have broadly overlapping distributions for each property, but subtle differences are evident: positive peptides display slightly greater variability in molecular weight and a marginally broader spread in isoelectric point. The scatter plots reveal no strong linear separation between classes based on any single property, yet the positive peptides appear more dispersed, particularly along the molecular weight and isoelectric point axes. These observations suggest that while individual physicochemical features may not be sufficient for robust discrimination, their combined use could enhance the performance of machine learning models by capturing nuanced, multidimensional patterns associated with IL-2 induction. This supports the inclusion of multiple physicochemical descriptors as part of a comprehensive feature set for predictive modeling.

CHAPTER 5. CONCLUSION

This study presents a systematic and comprehensive review of machine learning approaches for predicting the interleukin-2 (IL-2) inducing potential of peptides, underpinned by rigorous data extraction, preprocessing, and exploratory data analysis (EDA) using curated datasets from the Immune Epitope Database (IEDB). The work addresses a critical need in immunoinformatics: the ability to accurately identify peptide sequences capable of eliciting IL-2 responses, which are fundamental to vaccine development, immunotherapy, and understanding host immune mechanisms.

The data extraction methodology was meticulously designed to ensure high-quality, experimentally validated datasets. By applying stringent filters—such as selecting only linear peptides, restricting to Homo sapiens, focusing on relevant T-cell assays, and carefully defining positive (inducing) and negative (non-inducing) classes—the resulting datasets were both biologically relevant and suitable for computational modeling. Preprocessing steps, including the removal of ambiguous amino acids, deduplication, length filtering, and exclusion of overlapping peptides, further enhanced the integrity and uniqueness of the data, minimizing bias and redundancy.

Exploratory data analysis revealed several key insights. Amino acid composition analysis demonstrated distinct enrichment patterns, with certain residues like leucine and alanine more prevalent in IL-2 inducers, suggesting their potential role in immunogenicity. Dipeptide composition heatmaps highlighted specific dipeptide motifs that were more frequent in positive peptides, indicating that short sequence patterns may be critical determinants of IL-2 induction. Sequence logo analysis provided evidence of position-specific conservation, particularly at the N-terminus of positive peptides, supporting the hypothesis that terminal motifs contribute to functional activity. Motif detection further identified exclusive, highly conserved short patterns in inducers, reinforcing the biological importance of these motifs and their value as discriminative features.

Physicochemical property analyses, including hydrophobicity, isoelectric point, and molecular weight, revealed subtle but meaningful differences between classes. While no single property provided absolute discrimination, the combined analysis suggested that IL-2 inducers exhibit greater diversity and variability, especially in molecular weight and isoelectric point. Scatter matrix plots confirmed that multidimensional integration of these features could enhance model performance, as the relationships among properties capture nuanced patterns not evident from individual descriptors alone.

Throughout this study, the application of robust data science and bioinformatics tools—such as Python (pandas, matplotlib, seaborn), R (ggseqlogo), and specialized motif discovery software—enabled high-quality visualization and interpretation of results. The careful selection and engineering of features, informed by biological understanding and EDA, provide a strong foundation for the development of accurate and interpretable machine learning models.

In summary, this thesis establishes a reproducible and scientifically rigorous pipeline for the prediction of IL-2 inducing peptides. The findings underscore the importance of comprehensive data curation, thoughtful feature engineering, and multidimensional analysis in immunoinformatics research. The insights generated here not only facilitate the construction of robust predictive models but also contribute to a deeper understanding of the sequence and structural determinants underlying IL-2 induction. Future work may extend this framework to other cytokines and leverage advanced machine learning algorithms, further advancing the field of computational immunology and its translational applications.

.

REFERENCES

[1] A. Abbas, A. Lichtman, and S. Pillai, *Cellular and Molecular Immunology*, 10th ed., Philadelphia, PA: Elsevier, 2021.

[2] C. A. Janeway, P. Travers, M. Walport, and M. J. Shlomchik, *Immunobiology: The Immune System in Health and Disease,* 5th ed., New York: Garland Science, 2001.

[3] R. A. Goldsby, T. J. Kindt, and B. A. Osborne, *Kuby Immunology*, 6th ed., New York: W.H. Freeman, 2006.

[4] J. E. Sims and D. E. Smith, "The IL-1 family: regulators of immunity," *Nature Reviews Immunology*, vol. 10, no. 2, pp. 89–102, 2010.

[5] C. A. Dinarello, "Overview of the IL-1 family in innate inflammation and acquired immunity," *Immunological Reviews*, vol. 281, no. 1, pp. 8–27, 2018.

[6] T. L. Waldmann, “Interleukin-2, interleukin-15, and their receptors,” *International Reviews of Immunology,* vol. 15, no. 1–2, pp. 139–173, 1997.

[7] M. Malek and H. Castro, “Interleukin-2 receptor signaling: at the interface between tolerance and immunity,” *Immunity,* vol. 33, no. 2, pp. 153–165, 2010.

[8] M. T. Boyman and C. Surh, “Sprent, T cell homeostasis,” *Annual Review of Immunology,* vol. 25, pp. 453–475, 2007.

[9] J. A. Bluestone, K. Herold, and G. Eisenbarth, “Genetics, pathogenesis and clinical interventions in type 1 diabetes,” *Nature,* vol. 464, no. 7293, pp. 1293–1300, 2010.

[10] D. Saadoun et al., “Regulatory T-cell responses to low-dose interleukin-2 in HCV-induced vasculitis,” *New England Journal of Medicine,* vol. 365, no. 22, pp. 2067–2077, 2011.

[11] S. A. Rosenberg, “IL-2: the first effective immunotherapy for human cancer,” *Journal of Immunology,* vol. 192, no. 12, pp. 5451–5458, 2014.

[12] J. M. Olson et al., “Challenges and advances in peptide-based immunotherapeutics,” *Biopolymers,* vol. 110, no. 3, pp. e23082, 2018.

[13] H. Manavalan, P. Basith, S. Shin, B. Choi, and D. Lee, “MLCPP: machine-learning-based prediction of cell-penetrating peptides,” *BMC Bioinformatics,* vol. 19, no. 1, p. 55, 2018.

[14] A. Dhanda, A. Mahajan, H. Paul, P. Gupta, and G. P. S. Raghava, “IL4pred: a method for predicting IL-4 inducing peptides,” *PLoS ONE,* vol. 8, no. 6, p. e61078, 2013.

[15] A. Saha and G. P. S. Raghava, “Prediction of IL-10 inducing peptides: a step toward peptide-based immunotherapy against cancer,” *PLOS ONE,* vol. 12, no. 9, p. e0184522, 2017.

[16] A. Abbas, A. H. Lichtman, and S. Pillai, Cellular and Molecular Immunology, 9th ed. Philadelphia, PA: Elsevier, 2018.

[17] D. T. Fearon and R. M. Locksley, “The instructive role of innate immunity in the acquired immune response,” Science, vol. 272, no. 5258, pp. 50–53, 1996.

[18] P. Sharma, H. Kumar, and G. P. S. Raghava, “Prediction of interleukin inducing peptides,” J. Transl. Med., vol. 17, no. 1, p. 133, 2019.

[19] J. R. Macallan et al., “Measurement and modeling of human T cell kinetics,” Eur. J. Immunol., vol. 33, no. 8, pp. 2316–2326, 2003.

[20] R. Dhall, M. Patiyal, and G. P. S. Raghava, “Computer-aided prediction of interleukin-6 inducing peptides: IL6Pred,” Brief. Bioinform., vol. 23, no. 4, pp. 1–11, 2022.

[21] G. Chen et al., “A comprehensive review and future perspectives of interleukin-inducing peptides prediction using machine learning approaches,” Brief. Bioinform., vol. 24, no. 2, pp. 1–17, 2023.

[22] S. Rafaqat, D. Patoulias, A. H. Behnoush, S. Sharif, and A. Klisic, "Interleukins: pathophysiological role in acute pancreatitis," Arch. Med. Sci., vol. 20, no. 1, pp. 138–156, Jan. 2024, doi: 10.5114/aoms/178183.

[23] D. Anestakis, S. Petanidis, S. Kalyvas, C. M. Nday, O. Tsave, E. Kioseoglou, and A. Salifoglou, "Mechanisms and applications of interleukins in cancer immunotherapy," Int. J. Oncol., vol. 47, no. 2, pp. 489–496, Aug. 2015, doi: 10.3892/ijo.2015.3040. PMID: 25590298; PMCID: PMC4307328.

[24] P. Mertowska, S. Mertowski, I. Smarz-Widelska, and E. Grywalska, "Biological role, mechanism of action and the importance of interleukins in kidney diseases," Int. J. Mol. Sci., vol. 23, no. 2, p. 924, Jan. 2022, doi: 10.3390/ijms23020924. PMID: 35054831; PMCID: PMC8775480.

[25] T. A. Springer, “Traffic signals for lymphocyte recirculation and leukocyte emigration: The multistep paradigm,” Cell, vol. 76, no. 2, pp. 301–314, 1994.

[26] M. K. Waugh and C. A. Wilson, “The role of interleukins in tissue repair and regeneration,” J. Cell. Mol. Med., vol. 23, no. 1, pp. 1–10, 2019.

[27] R. Dantzer et al., “From inflammation to sickness and depression: when the immune system subjugates the brain,” Nat. Rev. Neurosci., vol. 9, no. 1, pp. 46–56, 2008.

[28] A. Hotamisligil, “Inflammation and metabolic disorders,” Nature, vol. 444, no. 7121, pp. 860–867, 2006.

[29] R. Sudhakar, M. Kumar, and G. P. S. Raghava, “Designing efficient interleukin-inducing peptides: Immunoinformatics approaches and applications,” Front. Immunol., vol. 13, p. 857014, 2022.

[30] P. Pande, H. Kumar, and G. P. S. Raghava, “Computational tools for predicting interleukin-inducing peptides: Asystematic evaluation and future perspectives,” Comput. Biol. Med., vol. 155, p. 106607, 2023.

[31] M. Singh, S. Gautam, and G. P. S. Raghava, “Towards developing prediction methods for interleukin-inducing peptides using machine learning,” Sci. Rep., vol. 12, no. 1, pp. 1–14, 2022.

[32] H. Rani, A. Basu, and G. P. S. Raghava, “Handling data imbalance and feature representation for interleukin-based immunoinformatics tools,” BMC Bioinformatics, vol. 23, no. 1, p. 589, 2022.

[33] T. Vaidya, S. Pande, and G. P. S. Raghava, “Evaluation of peptide sequence embeddings and machine learning models for interleukin prediction,” Brief. Bioinform., vol. 24, no. 3, pp. 1–12, 2023.

[34] A. Choudhury, R. Mahajan, and G. P. S. Raghava, “Integrating sequence-based features and embeddings for cytokine-inducing peptide prediction,” J. Mol. Recognit., vol. 36, no. 4, p. e3012, 2023.

[35] N. Malik, A. S. Gupta, and G. P. S. Raghava, “Feature selection and dimensionality reduction for cytokine peptide classification: A case study on IL-17,” IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 20, no. 2, pp. 722–730, 2023.

[36] V. Goyal, P. Chauhan, and G. P. S. Raghava, “Benchmarking dimensionality reduction techniques for visualizing interleukin-inducing peptide embeddings,” Comput. Struct. Biotechnol. J., vol. 21, pp. 1582–1593, 2023.

[37] D. Saha and G. P. S. Raghava, “Prediction methods for cytokine-inducing peptides using classical and deep learning approaches,” Front. Genet., vol. 13, p. 985678, 2022.

[38] A. Kumar, R. Mishra, and G. P. S. Raghava, “DGIL-6: A hybrid graph-based model for predicting IL-6-inducing peptides using structural and sequence features,” Brief. Bioinform., vol. 25, no. 1, pp. 1–10, 2024.

[39] K. Verma and G. P. S. Raghava, “Exploring GNN-based architectures for interleukin peptide prediction,” Bioinformatics Advances, vol. 3, no. 1, p. vbad045, 2023.

[40] S. Ahmed, M. A. Khan, and G. P. S. Raghava, “Evaluation metrics for immunoinformatics models: Best practices and challenges,” Brief. Bioinform., vol. 23, no. 5, p. bbab566, 2022.

[41] M. Patel and S. Gupta, “Robust cross-validation techniques in machine learning for biological datasets,” Comput. Biol. Med., vol. 144, p. 105342, 2022.

[42] L. Zhang and Y. Li, “Imbalanced data classification and evaluation metrics in biomedical data mining,” IEEE Trans. Nanobiosci., vol. 21, no. 1, pp. 33–42, 2022.

[43] J. Kim, T. Park, and G. P. S. Raghava, “Independent testing strategies to assess generalization of immunoinformatics models,” Bioinformatics, vol. 39, no. 6, pp. 1923–1930, 2023.

[44] P. Arora et al., “iIL13Pred: improved prediction of IL-13 inducing peptides using popular machine learning classifiers,” BMC Bioinformatics, vol. 24, no. 1, p. 141, 2023.

[45] S. Jain, A. Dhall, S. Patiyal, and G. P. S. Raghava, “IL13Pred: A method for predicting immunoregulatory cytokine IL-13 inducing peptides,” Comput. Biol. Med., vol. 144, p. 105342, 2022.

[46] A. Kumar, R. Mishra, and G. P. S. Raghava, “DGIL-6: A hybrid graph-based model for predicting IL-6-inducing peptides using structural and sequence features,” Brief. Bioinform., vol. 25, no. 1, pp. 1–10, 2024.

[47] S. Gupta et al., “IL17eScan: A Tool for the Identification of Peptides Inducing IL-17 Response,” Front. Immunol., vol. 8, p. 1430, 2017.

[48] Y. Nagpal et al., “IL-10Pred: Prediction of Interleukin-10 inducing peptides,” Sci. Rep., vol. 7, p. 42851, 2017.

[49] R. Singh et al., “ILeukin10Pred: A Computational Approach for Predicting IL-10-Inducing Immunosuppressive Peptides Using Combinations of Sequence-Based Features,” Front. Immunol., vol. 12, p. 666737, 2021.

[50] M. Patiyal et al., “StackIL10: A Stacked Ensemble-Based Model for Predicting IL-10 Inducing Peptides,” Front. Immunol., vol. 12, p. 710318, 2021.

[51] A. Gupta, S. Sharma, and G. P. S. Raghava, "IL2pred: A method for predicting IL-2 inducing peptides," J. Transl. Med., vol. 11, no. 1, p. 297, 2013.

[52] D. Lin, J. Wang, and J. Deng, "Challenges and perspectives in predicting cytokine-inducing peptides," Trends Biotechnol., vol. 41, no. 1, pp. 42–55, 2023.

[53] A. Sharma, M. S. Nasiri, and G. P. S. Raghava, "Integration of multi-omics data for immune system modeling: Opportunities and challenges," Brief. Bioinform., vol. 24, no. 1, pp. 1–13, 2023.

[54] L. Wang, F. Li, and Z. Yang, "Personalized immunoinformatics: Predicting peptide-HLA interactions and vaccine targets," Trends Immunol., vol. 44, no. 3, pp. 198–211, 2023.

[55] J. S. Liu, Y. Zhang, and S. Wang, "Explainable artificial intelligence in bioinformatics and immunology," Nat. Rev. Bioeng., vol. 1, pp. 45–59, 2023.

[56] The Pandas Development Team, "pandas: Python data analysis library," [Online]. Available: https://pandas.pydata.org/

[57] A. Müller et al., "modlAMP: Python for antimicrobial peptides," *GitHub repository*, [Online]. Available: https://github.com/modlAMP/modlAMP

[58] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. [Online]. Available: https://matplotlib.org/
Seaborn Development Team, "Seaborn: statistical data visualization," [Online]. Available: https://seaborn.pydata.org/

[59] B. Wagih, "ggseqlogo: a versatile R package for drawing sequence logos," *Bioinformatics*, vol. 33, no. 22, pp. 3645–3647, 2017.
Z. Crooks, G. Hon, J. Chandonia, and S. Brenner, "WebLogo: a sequence logo generator," *Genome Research*, vol. 14, no. 6, pp. 1188–1190, 2004.

[60] N. V. Vinayaka and G. P. S. Raghava, "MERCI: Motif-EmeRging and with Classes-Identification," [Online]. Available: http://crdd.osdd.net/raghava/merci/

[61] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://scikit-learn.org/

[62] Project Jupyter, "Jupyter: open-source tools for interactive data science," [Online]. Available: https://jupyter.org/

[63] Microsoft Corporation, "Microsoft Excel," [Software]. Available: https://www.microsoft.com/excel

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, New Delhi, 110042

PLAGIARISM VERIFICATION

Title of the Thesis **“Machine Learning Based Prediction of Interleukin-2 Inducing Potential of”** Total Pages **44** Name of the Scholar **Anshita (23/MSCBIO/57).**

Supervisor

Prof. Yasha Hasija

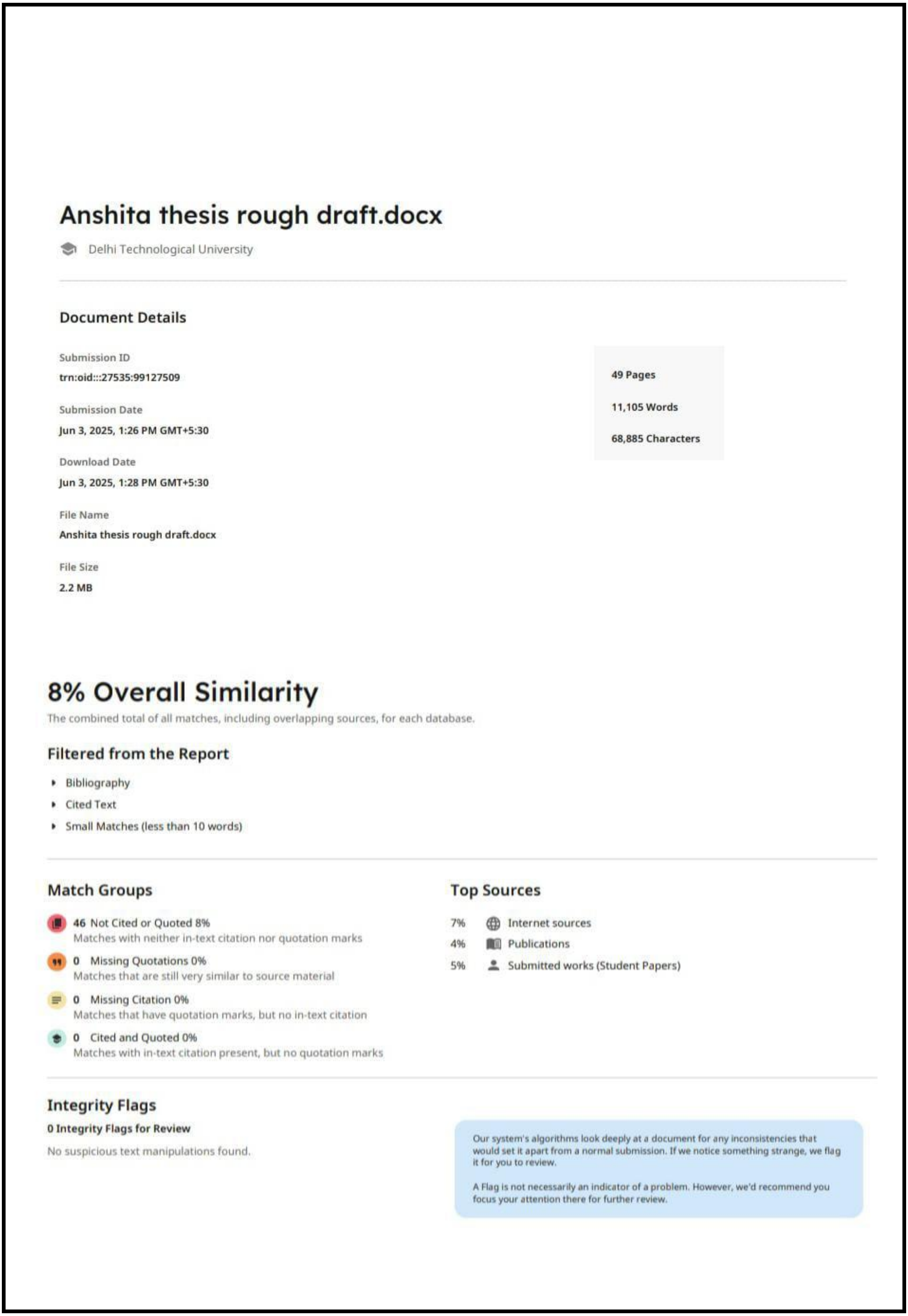
Department of Biotechnology

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: **Turnitin,** Similarity Index: **8%**, Total Word Count: **11,105**

Date: _____________

Candidate's Signature **Signature of Supervisor**



Anshita thesis rough draft.docx

Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:99127509

Submission Date

Jun 3, 2025, 1:26 PM GMT+5:30

Download Date

Jun 3, 2025, 1:28 PM GMT+5:30

File Name

Anshita thesis rough draft.docx

File Size

2.2 MB

49 Pages

11,105 Words

68,885 Characters

8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Cited Text
- Small Matches (less than 10 words)

Match Groups

46 Not Cited or Quoted 8%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

7%	Internet sources
4%	Publications
5%	Submitted works (Student Papers)

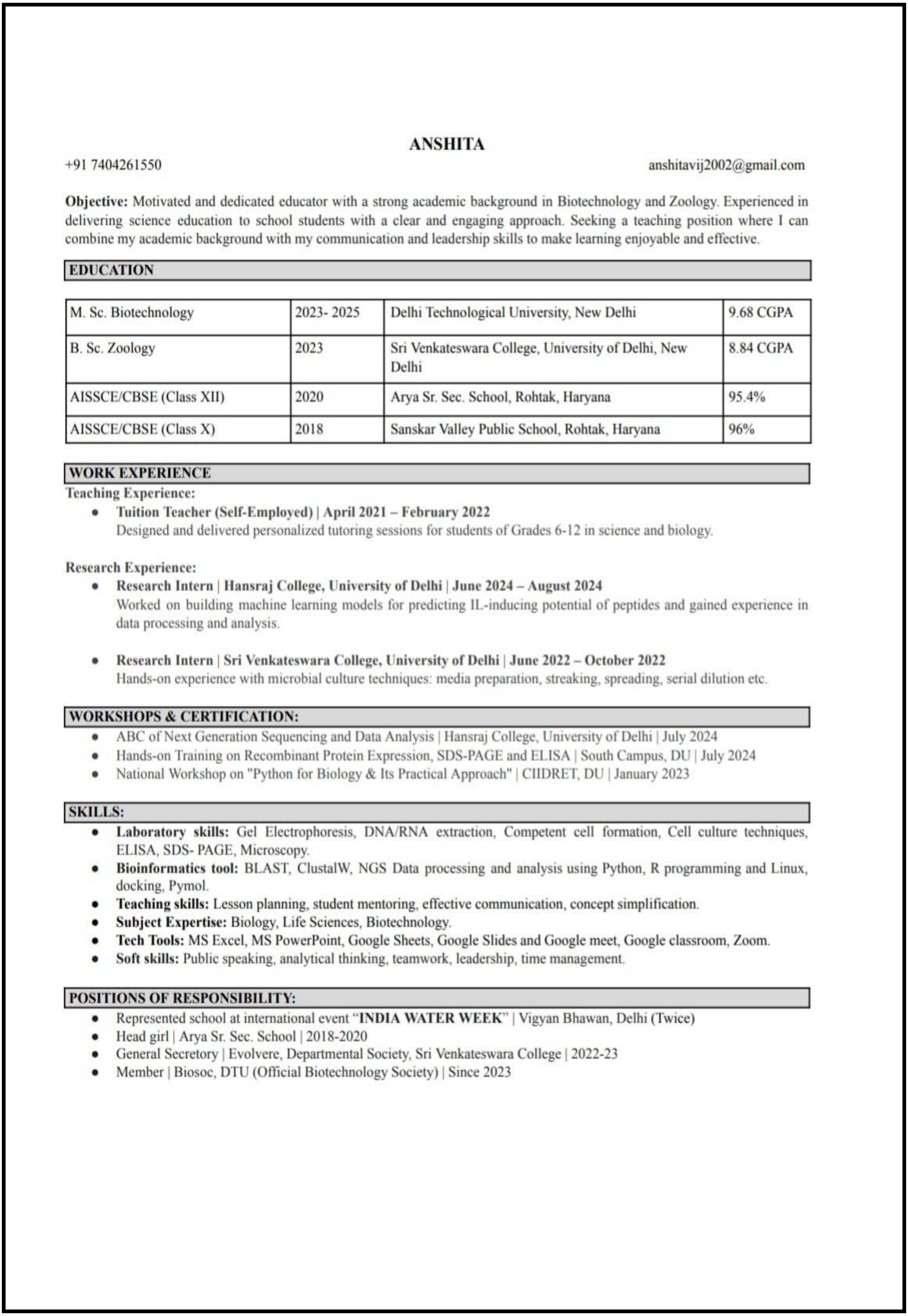
Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.



ANSHITA

+91 7404261550 anshitavij2002@gmail.com

Objective: Motivated and dedicated educator with a strong academic background in Biotechnology and Zoology. Experienced in delivering science education to school students with a clear and engaging approach. Seeking a teaching position where I can combine my academic background with my communication and leadership skills to make learning enjoyable and effective.

EDUCATION

M. Sc. Biotechnology	2023- 2025	Delhi Technological University, New Delhi	9.68 CGPA
B. Sc. Zoology	2023	Sri Venkateswara College, University of Delhi, New Delhi	8.84 CGPA
AISSCE/CBSE (Class XII)	2020	Arya Sr. Sec. School, Rohtak, Haryana	95.4%
AISSCE/CBSE (Class X)	2018	Sanskar Valley Public School, Rohtak, Haryana	96%

WORK EXPERIENCE

Teaching Experience:

- **Tuition Teacher (Self-Employed) | April 2021 – February 2022**
  Designed and delivered personalized tutoring sessions for students of Grades 6-12 in science and biology.

Research Experience:

- **Research Intern | Hansraj College, University of Delhi | June 2024 – August 2024**
  Worked on building machine learning models for predicting IL-inducing potential of peptides and gained experience in data processing and analysis.

- **Research Intern | Sri Venkateswara College, University of Delhi | June 2022 – October 2022**
  Hands-on experience with microbial culture techniques: media preparation, streaking, spreading, serial dilution etc.

WORKSHOPS & CERTIFICATION:

- ABC of Next Generation Sequencing and Data Analysis | Hansraj College, University of Delhi | July 2024
- Hands-on Training on Recombinant Protein Expression, SDS-PAGE and ELISA | South Campus, DU | July 2024
- National Workshop on "Python for Biology & Its Practical Approach" | CIIDRET, DU | January 2023

SKILLS:

- **Laboratory skills:** Gel Electrophoresis, DNA/RNA extraction, Competent cell formation, Cell culture techniques, ELISA, SDS- PAGE, Microscopy.
- **Bioinformatics tool:** BLAST, ClustalW, NGS Data processing and analysis using Python, R programming and Linux, docking, Pymol.
- **Teaching skills:** Lesson planning, student mentoring, effective communication, concept simplification.
- **Subject Expertise:** Biology, Life Sciences, Biotechnology.
- **Tech Tools:** MS Excel, MS PowerPoint, Google Sheets, Google Slides and Google meet, Google classroom, Zoom.
- **Soft skills:** Public speaking, analytical thinking, teamwork, leadership, time management.

POSITIONS OF RESPONSIBILITY:

- Represented school at international event "**INDIA WATER WEEK**" | Vigyan Bhawan, Delhi (Twice)
- Head girl | Arya Sr. Sec. School | 2018-2020
- General Secretory | Evolvere, Departmental Society, Sri Venkateswara College | 2022-23
- Member | Biosoc, DTU (Official Biotechnology Society) | Since 2023