Computational tools for designing therapeutic molecules against virulent factors of pathogens

A THESIS SUBMITTED IN THE PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

By

Neelam Sharma
PhD16201

under the guidance of

Prof. Gajendra Pal Singh Raghava
Professor, IIIT-Delhi

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY DELHI

Department of Computational Biology
Indraprastha Institute of Information Technology, New Delhi

June 2022

Computational tools for designing therapeutic molecules against virulent factors of pathogens

A THESIS SUBMITTED IN THE PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

By

Neelam Sharma
PhD16201

under the guidance of

Prof. Gajendra Pal Singh Raghava
Professor, IIIT-Delhi

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY DELHI

Department of Computational Biology
Indraprastha Institute of Information Technology, New Delhi

June 2022

Certificate

This is to certify that the thesis titled "**Computational tools for designing therapeutic molecules against virulent factors of pathogens**" being submitted by **Neelam Sharma** 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.

June, 2022

Prof. Gajendra P.S. Raghava

Professor

Indraprastha Institute of Information Technology Delhi

New Delhi-110 020

Declaration

This is to be certified that the dissertation titled "**Computational tools for designing therapeutic molecules against virulent factors of pathogens"** being submitted by **Neelam Sharma** to the Indraprastha Institute of Information Technology Delhi, for the award of degree of Doctor of Philosophy, is a bonafide work carried out by me. This research work has been carried out under the supervision of **Prof. Gajendra P.S. Raghava.**

The study pertains to this dissertation has not been submitted in part or in full, to any other University or Institution for the award of any other degree.

Neelam Sharma

PhD Student

IIIT-Delhi

Acknowledgements

Transitioning from one field to another (experimental lab to *in-silico* lab) was a challenging for me. It would not have been achievable without the assistance and guidance of several people. Lot of people made a substantial contribution to making this challenge easier for me. I would, therefore, like to take this opportunity to extend my sincere and heartfelt thankfulness to all of them.

First and foremost, I take this opportunity to express my deepest gratitude to my supervisor **Prof. Gajendra P. S. Raghava**, for his valuable guidance, ideas and perceptions in each stride of my research work without which it would have been impossible to complete. I was enthused by his energetic, vibrant and commitment in all his works and was a great privilege and honour to do my work under his guidance. He motivated me to stay focused, enthusiastic and imaginative about the research. At the same time, he taught me that one must not be afraid of making mistakes and eventually should learn from them. Every discussion with him inspires me to learn more and contribute to the betterment of people around us as he always says that "one can only succeed best and quickest by helping others". Lastly, I want to say that I will be always indebted to him for showing me the right way to learn, progress, and stay motivated for living well.

I would like to thank my Doctoral Committee members Dr. K.Sriram and Dr. Vibhor Kumar for their valuable comments and inputs which made me to improve my research work. I am highly thankful to all the teaching and research staff in IIIT-Delhi, who gave me an exciting experience of learning something new every day. I want to say thanks to Ms. Priti Patel, Ms. Shipra Jain and Ms. Sheetu along with others in the administrative department for helping me with all the academia related work. I also want to thank all IIIT Delhi admin facility staff for keeping IIITD campus clean and one of the best place to live in Delhi. I would like to praise the FMS and mess staff for providing great home-like facilities and food at IIIT-Delhi. I am highly thankful to the IT helpdesk and other system administrator staff for the management of the computer facilities at IIIT-Delhi, which aided in the smooth conduct of the research work.

A special thanks to the funding agency **Department of Science and Technology (DST-INSPIRE)** for providing me with financial aid to support my research work.

I would also like to thank my yoga teacher Mr. Ajay Saxena for his yoga classes and encourage me to keep myself physically fit and mentally confident. I am extremely grateful Dr. Amita Puri and Mr. Khushpinder Sharma for their invaluable help, motivation, emotional support throughout my PhD journey. I would also like thank my spiritual teachers namely BK Shivani, BK Bharti, BK Suraj, BK Pragya, BK Sujata to motivate me through their angelic lives and make me a divine being with good management skills, positivity, emotionally independent and energetic through their spiritual guidance and meditation sessions.

I want to mention special thanks to my lab mates and colleagues. I am deeply grateful to my seniors, Dr. Leima, Dr. Akshara, Dr. Pawan, Dr. Sherry, Dr. Piyush, Dr. Salman, Dr. Chakit, Dr. Anjali Lathwal, Dr. Harpreet, Dr. Rajesh, Dr. Vinod, Dr. Lubna for their timely help and guidance, wherever required. They have been a constant source of learning in the lab. I also want to thanks my colleagues Shipra, Akanksha, Dashleen, Dilraj, Sumeet, Anjali Dhall, Ritu, Nishant, Shubham, Nisha, Anand, Sarita, Smriti and Neetesh for their support.

I am feeling profound happiness and pride in acknowledging the role of some great friends that were no less than the pillars of strength during my Ph.D. years. I got many great friends for a lifetime Dr. Leima, Dr. Kajal, Dr. Niharika, Dr. Tania, Nidhi, Shipra, Ramneek, Dipa and Gunjan for their invaluable friendship and emotional support. They have been a great source of happiness, encouragement, strength and support during my Ph.D. journey. I found a great supportive friend in Dr Leima, who inspired me to remain calm and focused during one of the most difficult phases of my Ph.D. She is the most positive, cheerful and joyful person and has helped me to bring out the best in me, even when I myself was doubtful about the same. Shipra has been a great support both in professional and personal life. We had lots of fruitful discussions on a variety of topics. I am immensely obliged to my best friends Ritu, Ankita, Shalini, Sangam and Kirti for being with me and supporting in ups and downs of life. Having friends like them is a blessing.

No words can ever be strong enough to explain the gratitude to my parents, Smt. Hemlata and Shri. Ashok Kumar Sharma for their constant support of inspiration, encouragement, invaluable assistance and giving me the liberty to choose what I preferred. I am really grateful to my siblings CA Neha Sharma and CA Nihal Sharma who are always standing by me in hard times and always been a constant support. They have encouraged, motivated, and tolerated me during the most complex and demanding times.

Last but not the least my dear Supreme Shiv Baba, although there are no worldly words to express the gratitude to the most special being of my life. I would like to dedicate my pure feelings to express thanks for giving me the companionship, powers, blessings, recharging, enlightenment and pure energy throughout my life. There is a special place in my soul for you. I dedicate this thesis to you.

Neelam Sharma

Abstract

Microbes are minute, unicellular, multicellular organisms, such as bacteria, algae, fungi, viruses and protozoans, that can be only visible through the microscope. These can be infectious as well as non-infectious in nature. The ability of the infectious agent to cause diseases in the host cell is known as pathogenicity, while the ability of the pathogen to infect or cause damage to the host tissues is determined by the virulence factors. According to Centers for Diseases Control and Prevention reports, various infectious diseases such as COVID-19, tuberculosis, measles, and influenza are responsible for causing morbidity in the human population. Microbial pathogens pose an alarming threat to the healthcare sector worldwide. These micro-organisms cause severe diseases that lead to high mortality and morbidity rate. The rise in re-emergence of life-threatening infectious diseases and increasing incidence of antibiotic-resistant strains of the pathogens poses a danger to the healthcare sector. In the past, several studies reported that many distinct pathogenic micro-organisms share the common mechanisms of causing the infections to the host cell. Virulence factors of these pathogens play an important role in host-pathogen interactions and disease-mechanism. These factors include invasion, colonization and damage to the host cell, which contribute to pathogenicity. Thus, virulence factors are major drug/vaccine targets for designing therapeutic molecules against these pathogens. Some pathogenic micro-organisms release several molecules which cause damage to the host cell, induce the infection and evoke the diseases. These molecules include toxins released by certain pathogens to cause toxicity and induce allergic reactions in the host cell. Advances in various technologies led to the explosive growth of experimentally verified proteomic data related to virulence factors, which is available in the form of repositories. Thus, the present thesis focuses on utilising the publicly available experimentally verified data to develop computational tools to identify the potential virulence factors and pathogenicity associated with pathogens, such as toxicity and allergenicity, and to design the therapeutic molecules against them.

Taking this into consideration, we aimed to develop *in-silico* models to explore, predict and identify the potential virulence factors of pathogens. We build a machine learning-based method named 'VirFacPred' to identify novel virulence factors to aid the clinicians and scientific community. The best performing model achieved the maximum area under the receiver operating characteristic curve 0.97 with Matthews correlation coefficient 0.77 on the dataset. The best machine learning models have been implemented in the web server, which

allows the prediction, designing, mapping and motif search for the virulent proteins of the pathogens.

To address the pathogenicity caused by the pathogenic organisms to the host cell, such as toxicity and allergy, we have developed “ToxinPred2” and “AlgPred 2.0” that will facilitate the identification of toxic and allergic proteins. Toxins are one of the major virulence factors that play a crucial role in damaging the host cell. We have developed a highly accurate method, ToxinPred2, for predicting toxins with better precision. We have integrated a hybrid method that combines three approaches, i.e., similarity-, motif-, and composition-based machine learning model, which achieved a maximum area under the receiver operating characteristic curve around 0.99 with Matthews correlation coefficient 0.91 on the dataset. We have provided the standalone version of the method, which can be freely accessed at GitHub. This is a general method developed for predicting the toxicity of proteins regardless of their source of origin. On the other hand, a method called “AlgPred 2.0” has been developed for identifying allergenic proteins with high accuracy that allows the prediction of allergens, designing of non-allergenic proteins, mapping of IgE epitope, motif search and BLAST search. The ensemble approach, i.e., similarity-, motif-, and composition-based machine learning model, has been used for predicting allergenic protein by combining prediction scores. The best model achieved maximum performance in terms of area under the receiver operating characteristic curve 0.98 with Matthews correlation coefficient 0.85 on the dataset.

Besides proteins and peptides, some chemical compounds are known to induce allergic reactions to the host cell, known as chemical allergy. A therapeutic molecule may cause side effects due to its allergic potential. A first attempt has been made to develop the method using machine learning techniques that can predict the allergenic potential of chemicals. To aid the scientific community, we developed a novel method named “ChAlPred” that allows to predict and design the chemicals with allergenic properties. Our fingerprint-based analysis suggests that certain chemical fingerprints such as PubChemFP129 and GraphFP1014 are abundant in allergic compounds. We have also identified the FDA-approved drugs causing allergic symptoms (e.g., Cefuroxime, Spironolactone, Tioconazole) using our best model incorporated in the web server.

In summary, attempts have been made to develop *in-silico* models that can be used to design directly/indirectly therapeutic molecules against disease-causing agents. To facilitate the scientific community, web-based services and standalone software have been developed.

List of Publications

Thesis Related Publications

1. **Sharma, N.**, Patiyal, S., Dhall, A., Pande, A., Arora, C., & Raghava, G. (2021). AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. Briefings in bioinformatics, *22*(4), bbaa294. https://doi.org/10.1093/bib/bbaa294.

2. **Sharma, N.**, Patiyal, S., Dhall, A., Devi, N. L., & Raghava, G. (2021). ChAlPred: A web server for prediction of allergenicity of chemical compounds. Computers in biology and medicine, 136, 104746. https://doi.org/10.1016/j.compbiomed.2021.104746.

3. **Sharma N.**, Devi NL, Jain S, Raghava GPS. ToxinPred2: An improved method for predicting toxicity of proteins. Briefings in bioinformatics, bbac174. https://doi.org/10.1093/bib/bbac174.

4. **Sharma N.**, Devi NL, Raghava GPS. VirFacPred: A method for the prediction of virulent factors of the pathogens. (*manuscript under process*)

Other Publications

5. Gupta, S., **Sharma, N.**, Naorem, L. D., Jain, S., & Raghava, G. (2022). Collection, compilation and analysis of bacterial vaccines. *Computers in biology and medicine*, *149*, 106030. Advance online publication. https://doi.org/10.1016/j.compbiomed.2022.106030 (*equal contribution*)

6. **Sharma, N.,** Naorem, L. D., Gupta, S., & Raghava, G. P. S. (2021). Computational resources in healthcare. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1437. https://doi.org/10.1002/widm.1437

7. Dhall A., Jain S., **Sharma N.**, Naorem, L.D., Kaur, D., Patiyal, S., & Raghava, G. (2021). In silico tools and databases for designing cancer immunotherapy. Advances in Protein Chemistry and Structural Biology doi.org/10.1016/bs.apcsb.2021.11.008

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

9. Dhall, A., Patiyal, S., **Sharma, N.**, Usmani, S. S., & Raghava, G. (2021). Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in

COVID-19. *Briefings in bioinformatics*, *22*(2), 936–945. https://doi.org/10.1093/bib/bbaa259

10. Agrawal, P., Bhagat, D., Mahalwal, M., **Sharma, N.**, & Raghava, G. (2021). AntiCP 2.0: an updated model for predicting anticancer peptides. *Briefings in bioinformatics*, *22*(3), bbaa153. https://doi.org/10.1093/bib/bbaa153

11. Mathur, D., Kaur, H., Dhall, A., **Sharma, N.**, & Raghava, G. (2021). SAPdb: A database of short peptides and the corresponding nanostructures formed by self-assembly. *Computers in biology and medicine*, *133*, 104391. https://doi.org/10.1016/j.compbiomed.2021.1043

12. Kumar, V., Patiyal, S., Dhall, A., **Sharma, N.**, & Raghava, G. (2021). B3Pred: A Random-Forest-Based Method for Predicting and Designing Blood-Brain Barrier Penetrating Peptides. *Pharmaceutics*, *13*(8), 1237. https://doi.org/10.3390/pharmaceutics13081237

13. Tarca, A. L., Pataki, B. Á., Romero, R., Sirota, M., Guan, Y., Kutum, R., Gomez-Lopez, N., Done, B., Bhatti, G., Yu, T., Andreoletti, G., Chaiworapongsa, T., **DREAM Preterm Birth Prediction Challenge Consortium**, Hassan, S. S., Hsu, C. D., Aghaeepour, N., Stolovitzky, G., Csabai, I., & Costello, J. C. (2021). Crowdsourcing assessment of maternal blood multi-omics for predicting gestational age and preterm birth. *Cell reports. Medicine*, *2*(6), 100323. https://doi.org/10.1016/j.xcrm.2021.100323

14. Gabor, A., Tognetti, M., Driessen, A., Tanevski, J., Guo, B., Cao, W., Shen, H., Yu, T., Chung, V., Single Cell Signaling in Breast Cancer **DREAM Consortium members**, Bodenmiller, B., & Saez-Rodriguez, J. (2021). Cell-to-cell and type-to-type heterogeneity of signaling networks: insights from the crowd. *Molecular systems biology*, *17*(10), e10402. https://doi.org/10.15252/msb.202110402

15. Patiyal, S., Kaur, D., Kaur, H., **Sharma, N.**, Dhall, A., Sahai, S., Agrawal, P., Maryam, L., Arora, C., & Raghava, G. (2020). A Web-Based Platform on Coronavirus Disease-19 to Maintain Predicted Diagnostic, Drug, and Vaccine Candidates. *Monoclonal antibodies in immunodiagnosis and immunotherapy*, *39*(6), 204–216. https://doi.org/10.1089/mab.2020.0035 (*equal contribution*)

16. Kaur, D., Patiyal, S., **Sharma, N.**, Usmani, S. S., & Raghava, G. (2019). PRRDB 2.0: a comprehensive database of pattern-recognition receptors and their ligands. *Database: the journal of biological databases and curation*, *2019*, baz076. https://doi.org/10.1093/database/baz076

17. Agrawal, P., Raghav, P. K., Bhalla, S., **Sharma, N.**, & Raghava, G. (2018). Overview of Free Software Developed for Designing Drugs Based on Protein-Small Molecules

Interaction. *Current topics in medicinal chemistry*, *18*(13), 1146–1167. https://doi.org/10.2174/1568026618666180816155131

18. 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. (2022). Computing wide range of protein/peptide features from their sequence and structure. *Journal of Computational Biology.*

Table of Contents

Abbreviations

AAC	Amino acid composition
Acc	Accuracy
AUC	Area Under the Receiver Operating Characteristic curve
BLAST	Basic Local Alignment Search Tool
Blastp	protein–protein Basic Local Alignment Search Tool
CD-HIT	Cluster Database at High Identity with Tolerance
COVID-19	Coronavirus Disease 2019
CV	Cross Validation
DPC	Dipeptide composition
DT	Decision Tree
E-value	Expect value
FDA	Food and Drug Administration
GNB	Gaussian Naive Bayes
HTML	HyperText Markup Language
IEDB	Immune Epitope Database
IgE	Immunoglobulin E
KNN	k-nearest neighbors
LightGBM	Light Gradient Boosting Machine
LPS	Lipopolysaccharides
LR	Logistic Regression
MAST	Motif Alignment & Search Tool
MCC	Matthews correlation coefficient
MEME	Multiple Em for Motif Elicitation
MERCI	Motif-EmeRging and with Classes-Identification
ML	Machine Learning
MLP	Multi-layer Perceptron
NCBI	National Center for Biotechnology Information
PSI-BLAST	Position-Specific Iterated Basic Local Alignment Search Tool
PSSM	Position Specific Scoring Matrix
QSAR	Quantitative Structure-Activity Relationship
RF	Random Forest

Sens	Sensitivity
Sklearn	Scikit-learn
Spec	Specificity
SVC	Support Vector Classifier
UniProtKB	UniProt Knowledgebase
VF	Virulence Factor
XGB	Extreme Gradient Boosting

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Micro-organisms are tiny, living, microscopic organisms that are omnipresent and cannot be seen by naked eyes. These organisms can be unicellular, multicellular, prokaryotes and eukaryotes, such as bacteria, algae, fungi, viruses and protozoans (InformedHealth.org, 2006). Based on the ability of these micro-organisms to cause harm to the host cell, they are classified as pathogens and non-pathogens. Pathogenic micro-organisms are capable of causing an infection, hence leading to disease to the host cell, whereas non-pathogenic micro-organisms are harmless, and do not cause any infection or disease (Piglowski, 2019). The competence of the infectious agent to cause diseases in the host cell is known as pathogenicity or infectivity. Whereas the ability of the pathogen to infect or cause damage to the host tissues is determined by the virulence factors (VFs). To induce the infection, the pathogenic microbes invade the host cells, multiply their growth and cause damage to the host organism. The capacity of the microbe to cause mild or severe, or acute infection depends on its relative pathogenicity (Shapiro-Ilan et al., 2005). Pathogenic microbes include *Bacillus anthracis, Influenza virus, Candida albicans, cyanobacteria, Plasmodium spp.*, whereas *Escherichia coli*, *Lactobacillus acidophilus*, *Brevibacterium linens* fall under the category of non-pathogenic microbes.

Infectious diseases caused by pathogenic micro-organisms are the leading cause of mortality worldwide. As per the report by the Centers for Diseases Control and Prevention (CDC), there are various infectious diseases, such as COVID-19, tuberculosis, measles, and influenza, which are responsible for causing morbidity in the human population (CDC, 2021), (ECDC, 2022). The rise in re-emergence of life-threatening infectious diseases and increasing incidence of antibiotic-resistant strains of the pathogens poses a danger to the healthcare sector. Many past studies state that although many distinct pathogenic micro-organisms are present in the environment, they share the common mechanisms of causing the infections to the host cell. For instance, different pathogenic bacteria have similar processes for attaching to the host membrane, penetrating it, causing damage, evading the host immune response, and establishing the infection. VFs play a very significant role in enhancing the pathogenicity of pathogens. Figure 1.1 depicts the various aspects needed by the pathogenic micro-organism to establish the infection using its various virulence factors and releasing certain responses such as toxicity and allergy (pathogenicity) to the host cell. The first step for causing the infection is binding the pathogen to the host outer membrane, for which adhesin proteins play a crucial role. After that, burrowing into the cell, also known as penetration or invasion inside the host cell, is mediated by invasion factors present in the pathogen and followed by colonising within the

host cell using the colonization factors. Some pathogenic micro-organisms use immune evasion strategies, and some suppress the immune response to escape from the host's defence mechanism and survive within the cell. Thus leading to pathogenic reactions such as exploiting the normal functioning by damaging the cell, causing toxicity and allergic responses. The mechanism of microbial pathogenesis is shown in Figure 1.1. To understand the pathogenicity of the micro-organism, it is necessary to have a more comprehensive understanding of various molecular mechanisms associated with the virulence of the pathogen because the ability of the pathogenic micro-organism is directly related to its virulence (Peterson, 1996).

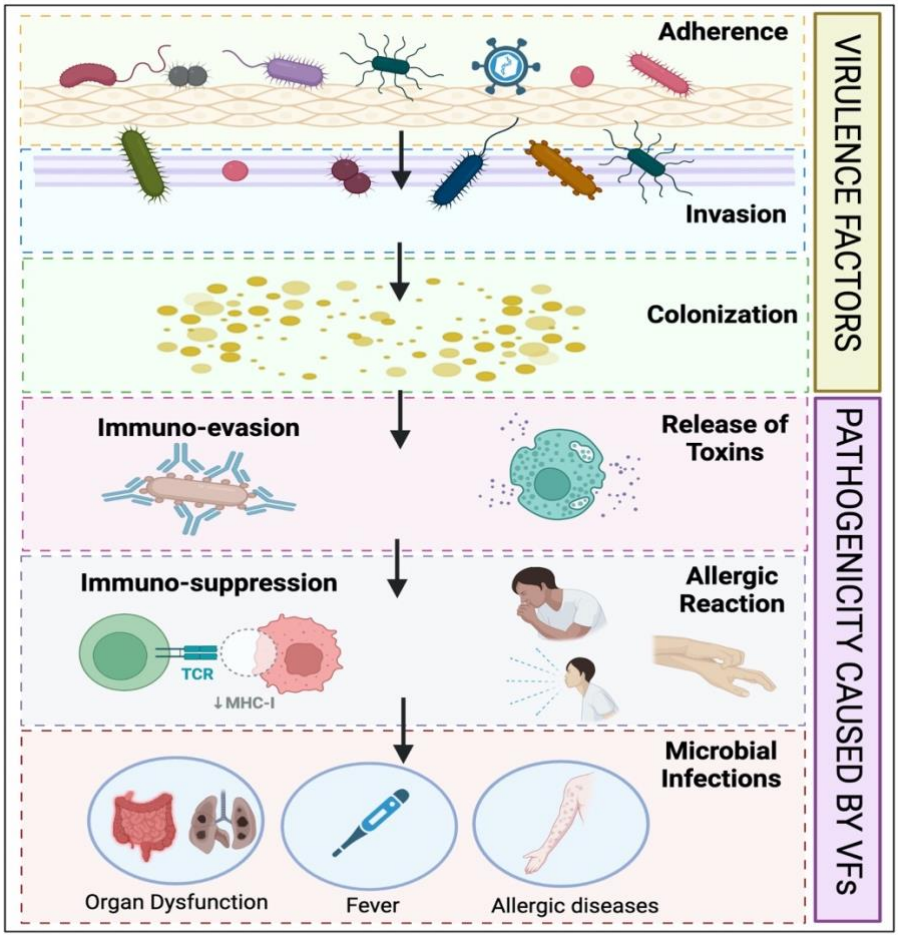

Figure 1.1: The mechanism of microbial pathogenesis

1.2 Virulence

There are millions of microbial species present, out of which very few are known to cause diseases (Blackwell, 2011). The micro-organisms that cause disease in humans and other organisms are pathogenic in nature. Interaction of these pathogenic microbes is the result of a parasitic relationship with the host, which results in causing the disease to the host organism

(Casadevall & Pirofski, 2000). Survival and growth of the micro-organism causing infection and damage to the host cell are the two processes required by the pathogens to cause pathogenicity (Peterson, 1996), (Balloux & van Dorp, 2017). Immune evasion factors are the essential aspect of causing infection and are required for the survival of the pathogens inside the host cells (Hilleman, 2004). To become infectious, the pathogen needs to cause damage to the host. Other than these, VFs play a significant role that can immediately cause damage to the host organism (Peterson, 1996), (Sharma, Dhasmana, et al., 2017). VFs are the molecules synthesized by micro-organisms, especially pathogens which include bacteria, viruses, protozoans and fungi (Peterson, 1996), (Shapiro-Ilan et al., 2005). These factors enhance the tendency of pathogens to cause diseases in different host organisms. They are classified into different groups based on their functions and mechanism of causing disease.

1.2.1 Adherence factors

Adhesins are the molecules involved in the attachment of the pathogen to the host cell, which is extensively found in the pathogenic microbes (Berne et al., 2015). Commensal microbes primarily reside within the mucus, at some distance from the anti-microbial peptide secreting epithelial cells. This helps the pathogen remain in intimate contact with the host cell, which is required to manipulate and invade it (Gallo & Hooper, 2012). *Streptococcus mutans* is a gram-positive bacterium which causes human dental caries. It uses Adhesin P1 protein to adhere to the tooth surface (Sullan et al., 2015). Viruses such as the Influenza virus adhere to the sialic acid present on the host's membrane of respiratory and intestinal cells using its spike protein hemagglutinin, infecting the host with influenza (Matrosovich et al., 2015), (Dou et al., 2018). *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEMP1) is an adhesin protein present in a protozoan *Plasmodium falciparum* that causes malaria in humans, which is used to adhere with endothelial cells (Senczuk et al., 2001). Hyphal wall protein 1 (HWP1) is a vital adhesin protein expressed by various *Candida* species (Abastabar et al., 2016).

1.2.2 Invasion factors

Invasion into the host cell and tissues are usually described as one of the defining activities of the pathogenic microbes (Ribet & Cossart, 2015). Some bacteria invade the host cell by distorting the outer membrane and triggering endocytosis (Dalle et al., 2010). After invading the cell, it releases certain enzymes and toxins. For example, *Staphylococcus aureus* uses enzymes like Hyaluronidase S (Ibberson et al., 2014) and *Bacillus anthracis* uses enzymes like Phospholipase C (Titball, 1993) to invade the host cell by disrupting its outer cell membrane

(Pomerantsev et al., 2003). Many pathogens invade by entering the bloodstream, releasing toxins in the bloodstream leading to sepsis and can even lead to pus formation (Minasyan, 2019).

1.2.3 Colonization factors (Growth in the host cell)

After invasion inside the host cells and tissues, the next step is colonising, which is mediated by colonisation factors. These factors help the pathogen to colonise within the host cell (Ribet & Cossart, 2015). *Streptococcus pneumoniae* colonises the respiratory tract and causes major respiratory and invasive diseases (Green et al., 2021).

1.2.4 Immuno-evasion

Immuno-evasion helps the pathogens to evade the host immune response by escaping from being recognised by the innate and adaptive immune response (Finlay & McFadden, 2006). Some classes of pathogenic species have developed various immune-evasion strategies, such as suppressing their recognition via immunogenic surface receptors present on the outer membrane of the host cell, i.e., pathogen-associated molecular patterns (PAMPs) (Mogensen, 2009). Certain bacteria like *Helicobacter pylori* and *Porphyromonas gingivalis* synthesize an altered form of lipopolysaccharides (LPS) which helps them not being recognised by the host immune system (Hornef et al., 2002). Few fungal pathogens coat them with capsules or even from the molecules derived from the host itself to escape the immune response (Latge & Beauvais, 2014), (van de Veerdonk et al., 2008). For example, capsular polysaccharides shield the pathogens from the host's defense mechanism (Chai et al., 2009).

1.2.5 Immunosuppression

Immunosuppression is the obstruction of host immune response by some effector proteins, which is the popular strategy in pathogenic microbes (Peterson, 1996). These effector proteins mainly interact with host molecules to manipulate the host cell in favor of the pathogen (Rajamuthiah & Mylonakis, 2014). Effector proteins may include invasins which trigger receptor-mediated host cytoskeleton rearrangements and induce endocytosis (Phan et al., 2007). Traversing the host membrane is very crucial for the action of the effector molecules (Pizarro-Cerda & Cossart, 2006).

1.2.6 Damage to the host cell

Toxins

Toxins are the virulence factors that exploit the normal functions of the host cell by constraining the essential processes to facilitate pathogenic infection (do Vale et al., 2016). To resist the host defense mechanism, some pathogens secrete various toxins that target innate immune cells, specifically neutrophils and macrophages, hence obliterating a key component of the host immune response (do Vale et al., 2016). Toxins can be of two types: endotoxins and exotoxins. The LPS present in the outer cell wall of the Gram-negative bacteria is a classical example of endotoxins. Lipid A component of LPS triggers the inflammatory response of the host immune system causing fever, severe changes in blood pressure, lethal shock, and multi-organ failure, which can even lead to death (Peterson, 1996).

Certain pathogenic bacterial species produce various types of protein toxins and enzymes, which fall under the category of exotoxins (Popoff, 2018). This includes a wide variety of toxins such as cytotoxins, neurotoxins, and enterotoxins. Exotoxins can be grouped into three categories: intracellular targeting, membrane disrupting, and superantigens (Peterson, 1996), (Spaulding et al., 2013).

Some examples of intracellular targeting exotoxins are diphtheria toxins produced by the gram-positive bacteria *Corynebacterium diphtheriae* (Murphy, 1996). Cholera toxin is an enterotoxin produced by the gram-negative bacterium *Vibrio cholerae* (Reidl & Klose, 2002). Botulinum toxin (also known as Botox) is a neurotoxin produced by the gram-positive bacterium *Clostridium botulinum* (Nigam & Nigam, 2010). Tetanus toxin is produced by the gram-positive bacterium *Clostridium tetani* (George et al., 2022). These toxins inhibit protein synthesis, hence killing the host cell.

Membrane disrupting exotoxins obstruct the outer cell membrane of the host cell by forming pores in the membrane (Los et al., 2013). For instance, leukocidins (lysis of white blood cells) (Spaan et al., 2017), hemolysins (lysis of red blood cells) (Dinges et al., 2000), streptolysins of *Streptococcus pyogenes* (Molloy et al., 2011), pneumolysin of *Streptococcus pneumoniae* (Hirst et al., 2004) and an alpha-toxin produced by *Staphylococcus aureus* are pore-forming exotoxins which cause the leakage of cytoplasmic contents leading to the cell death (Seilie & Bubeck Wardenburg, 2017). Superantigens are the class of exotoxins that elicit a strong immune response, hence activating the immune cells to excessively release the cytokines leading to a cytokine storm (Proft & Fraser, 2003), (Popugailo et al., 2019). One example of superantigen exotoxin is toxic shock syndrome caused due to the colonization of *Staphylococcus aureus* (Xu & McCormick, 2012).

Exoenzymes

Some pathogenic microbes can directly cause damage to the host cells by secreting factors like toxins, hydrolytic enzymes, and physical forces during host cell invasion or escape (Mayer et al., 2013). Secreted factors include hydrolytic enzymes like proteases or lipases and toxins in the form of small metabolites or peptides that poison the host cell and causes tissue damage (Scharf et al., 2014).

1.2.7 Siderophores

Siderophores are the small chemical compounds or secondary metabolites that help pathogens in the uptake of iron (chelation) or obtain nutrition from the host cell (Ahmed & Holmstrom, 2014). The secretion of siderophores by many pathogenic species is considered an important virulence factor (Peterson, 1996). Based on the chemical moiety involved in the iron chelation, there are four types of siderophores. These are catecholates (enterobactin), hydroxamates (ferrioxamine B, alcaligin), carboxylates (rhizobactin, rhizoferrin) and phenolates (pyochelin) (Kramer et al., 2020). Some of the examples of siderophores are enterochelin produced by *Escherichia coli* (Scholz & Greenberg, 2015), pyochelin from *Pseudomonas aeruginosa* (Ross-Gillespie et al., 2015), ornibactin from *Burkholderia cenocepacia* (Sathe et al., 2019) and vibrioferrin from *Vibrio* species (Cordero et al., 2012).

1.3 Pathogenicity caused by virulence factors

After the successful attachment, invasion, and multiplying within the host cell, the pathogenic micro-organisms secrete certain molecules which cause damage to the host cell, induce the infection and evoke the diseases. These molecules include toxins released by certain pathogens to cause toxicity and induce allergic reactions in the host cell.

1.3.1 Toxicity

Toxins are substances that have the potential to cause deleterious effects on living organisms. They are naturally present in plants or can be produced by animals (snakes, spiders, cone snails) and different types of microbes (such as bacteria and fungi to enhance their pathogenicity) (Clark et al., 2019). When toxins enter the body in any form, they can cause fatal illnesses or death. Generally, toxins can be categorized into endotoxins and exotoxins. Endotoxins such as LPS (Raetz & Whitfield, 2002) and exotoxins like tetanus toxin, botulinum toxin, and mycotoxin can cause immense inflammation (Pasechnik et al., 1992), (Nigam & Nigam, 2010). It induces the degranulation of mast cells and basophils as a part of the immune response,

causing the release of inflammatory mediators. These mediators include cytokines, interleukins, and histamine, which cause or lead to allergic symptoms like fever, skin rashes, septic shock, and other diseases (Mukai et al., 2018). Proteins and peptides are naturally occurring molecules that play various functions and processes in the body that are essential to sustain cellular mechanisms (Shaji & Patole, 2008). Their aberrant activity has been involved in several pathological conditions such as cancer, neurodegenerative disorders and diabetes (Bruno et al., 2013). Thus, using them as therapeutic agents is regarded as a promising way to fight against a variety of diseases. In recent years, they have the potential to revolutionize medical therapy and are the preferred choice over small molecules and antibodies due to their high target specificity, tissue penetration, high biological activity and inexpensive (Bruno et al., 2013). However, there are certain prime concerns in the development of proteins/peptide-based drug discovery, such as toxicity, immunogenicity and stability (Otvos & Wade, 2014). Due to this, the assessment of toxic properties of protein/peptide is of great necessity.

1.3.2 Allergenicity

Several studies have also shown the correlation between microbial colonization and allergic diseases. Allergy is an exaggerated immune response caused by foreign substances called allergens. Certain microbial species like *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Alternaria*, *Aspergillus* and *Influenza* virus are known to cause severe allergic inflammation (Nordengrun et al., 2018). Moreover, allergens like dust mites, pollens, and many others, induce the Type I hypersensitivity reactions which elicit IgE antibodies. This allergic reaction results in the release of inflammatory mediators, such as histamine, cytokines from mast cells and basophils (Masoli et al., 2004), which affects the population at a large scale, particularly in skin sensitization (Sutton & Gould, 1993), (Broadfield et al., 2002). Also, there is a wide range of molecules that can pose a threat as allergens, including biological molecules like proteins and peptides or some chemical compounds (Sharma et al., 2020), (Dang & Lawrence, 2014), (Goodman et al., 2005). Other than that, molecules like lipids (Del Moral & Martinez-Naves, 2017), carbohydrates (Commins & Platts-Mills, 2010), nucleic acid (mRNA vaccines) (Rubin, 2021) and some engineered nanoparticles (Alsaleh & Brown, 2020) can also stimulate some specific allergic reactions like asthma, food allergies and chronic kidney disease, respectively.

1.4 Designing of therapeutic molecules

Increasing reports of drug resistance are becoming more common, posing a severe threat to world healthcare. There is a critical need for new and more effective medicinal drugs to be discovered. Therapeutic agents have attracted much attention from researchers as a safe and effective alternative because of their high efficacies, low toxicity, strong cell penetration, and ease of production. Therapeutic molecules could be proteins, peptides and small chemicals used to treat several fatal illnesses. These molecules could act as potential drug and vaccine targets. Over the last few decades, significant progress has been made in the development of computational resources to accelerate the process of drug and vaccine discovery. Several properties of therapeutic molecules such as half-life, toxicity, allergenicity, haemolytic, antimicrobial, and cell-penetrating properties need to be evaluated before being released to the market.

In the past, various repositories and *in-silico* tools have been developed to maintain and predict the essential properties of the therapeutic molecules. CAMP (Waghu et al., 2014), AVPdb (Qureshi et al., 2014), SATPdb (Singh et al., 2016), CPPsite (Gautam et al., 2012), TumorHoPe (Kapoor et al., 2012) and Hemolytik (Gautam et al., 2014) are some of the databases that comprehend the information about different properties of therapeutic molecules. Computational tools like ToxinPred (Gupta et al., 2015), AlgPred (Saha & Raghava, 2006a), PlifePred (Mathur et al., 2018), AHTPin (Kumar et al., 2015) and HemoPI (Chaudhary et al., 2016) have been developed to predict the pharmacologically important properties of the molecules, such as toxicity, allergenicity, half-life, anti-hypertensive and haemolytic, respectively. These prediction methods aid not only in the development of peptides and chemical analogues with better physicochemical features but also in the screening of libraries for the desired therapeutic property. The most effective method for identifying novel therapeutic molecules is computational screening, followed by *in-vitro* and *in-vivo* studies.

1.5 Origin of proposal

Several attempts have been made in the past to identify the novel virulence factors of the pathogens, their mechanism of causing diseases, and how to use these factors as potential drug targets. Efforts have been made to study the specific virulence factors of the pathogens, such as SPAAN, a method developed to predict the adhesin proteins, which play a significant role in the virulence of the pathogens (Sachdeva et al., 2005). Some are developed for specific organisms like MAAP, a method for predicting adhesin proteins for malarial parasites (Ansari

et al., 2008). Likewise, limited attempts have been made to study the toxicity caused by the virulence factors of the pathogenic organisms. After the invasion of pathogenic micro-organisms into the host cell, microbes release certain toxins to damage the cell. Various methods are available that can be used to predict the toxicity caused by a certain group of organisms, such as BTXpred (Saha & Raghava, 2007a) and NTXpred (Saha & Raghava, 2007b) developed for classifying the toxins of bacterial origin and neurotoxins, respectively. Other methods like ClanTox (Naamati et al., 2009), SpiderP (Wong et al., 2013), and TOXIFY (Cole & Brewer, 2019) are developed to identify the toxins of certain animal origin. After releasing the toxins into the cells, it induces intense inflammation, hence causing allergic symptoms. Several studies have also shown the correlation between microbial colonization and allergic diseases, and some of the virulent factors are responsible for causing allergic reactions. Allergy can be caused by proteins, peptides, small chemical compounds and other molecules like lipids and carbohydrates. There are several tools and methods developed in the past for the prediction of allergy caused by proteins and peptides, but no attempt has been made to predict the allergy caused by small chemical compounds. Keeping all these limitations in mind, the present study provides the *in-silico* approach that can be used to identify the virulence factors against all the pathogenic micro-organisms. Also, it provides the tools with state-of-the-art techniques to predict the toxic and allergenic proteins along with the prediction of allergenicity caused by the small chemical compounds. Besides, it also includes the methods to design the therapeutic molecules against the virulence factors, toxicity and allergenicity caused by proteins, peptides and small molecules.

1.6 Objectives of the thesis

To overcome the above-discussed shortcomings, we have put an effort to study the virulence factors and their mechanism related to causing toxicity and allergenicity to the host cell. The present study primarily focuses on the identification of virulence factors of the pathogens, which is an important aspect to understand the host-pathogen interactions and disease-mechanism. Therefore, these molecular determinants of virulence can be used as promising targets for developing novel anti-virulence drugs. The first aim of this study is designing the drug targets against disease-causing pathogens based on virulence factors without killing or inhibiting environment-friendly bacteria. Secondly, to study the toxicity caused by different organisms, we have developed the tool named “ToxinPred2”, which can be used to predict the toxicity of the protein sequences. Apart from that, to study the mechanism of allergy caused by

proteins, we have developed the method “AlgPred 2.0”. In addition to proteins, allergy is also caused by chemical compounds, known as chemical allergy. So, to address this problem, we have developed a novel method to predict the allergenic potential of chemicals and employed in the form of a web server name “ChAlPred”. Based on this, the following objectives have been framed:

- A prediction method for the classification of virulent and non-virulent proteins
- An updated *in-silico* method for classifying toxins and designing non-toxic proteins
- An improved method for predicting allergenic proteins and mapping of IgE epitopes
- A computational method for the prediction of chemical allergens and designing of non-allergenic compounds

The outline of the thesis is depicted in Figure 1.2.

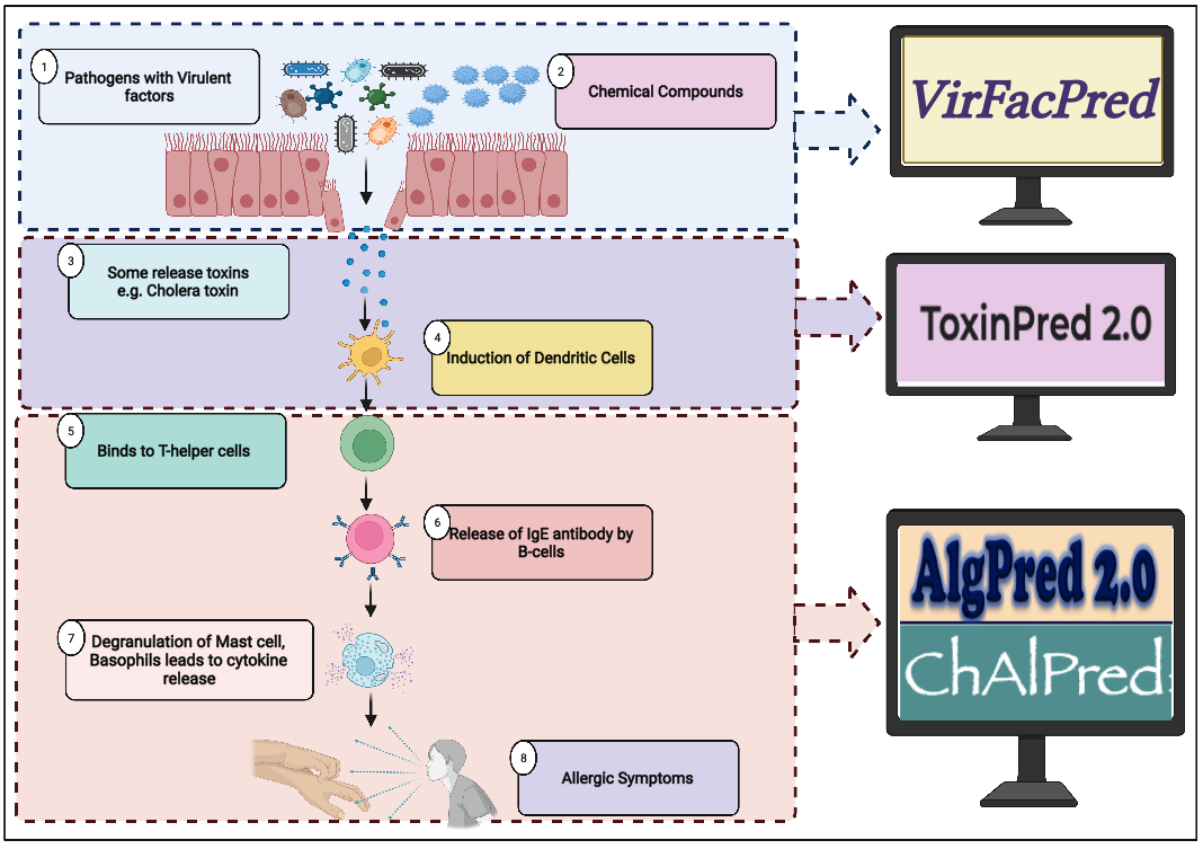

Figure 1.2: The overall workflow of the study

1.7 Organization of the chapters

The thesis is organized into seven chapters containing the information discussed below:

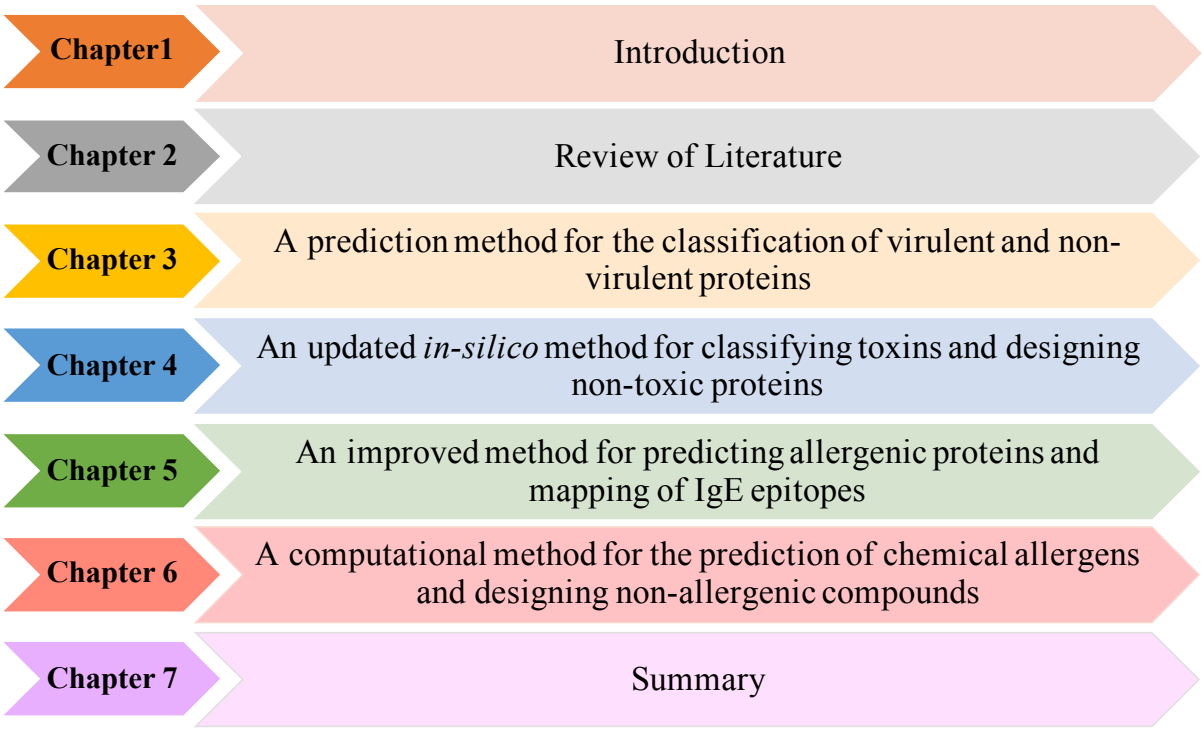

Figure 1.3: Organization of thesis and title of the chapters

Chapter 1 – This chapter introduces the virulence of the pathogens along with a description of all the associated factors. This is followed by a brief information of the toxicity caused due to these virulence factors. It also discusses the allergy caused by proteins and the small chemical compounds. At the end of the chapter, the organization and the overall objectives undertaken in the thesis are elaborated.

Chapter 2 – This chapter reviews the existing and published literature. The existing knowledge on the virulence factors of the pathogens, followed by toxicity and allergenicity caused by them, has been discussed. The emphasis has been laid on the limitations and challenges of the current methods. Overall, this chapter explains the motivation behind this study.

Chapter 3 – This chapter describes the first objective of the thesis, which is the development of machine learning-based model for the classification of virulent proteins of various pathogens. This study describes a method developed to classify the protein sequence as virulent

and non-virulent. A web server "VirFacPred" has also been developed that allows the prediction of virulent proteins along with similarity-based search using BLAST, motifs and designing of therapeutic proteins against them.

Chapter 4 – This chapter is the second objective of the thesis, which deals with the prediction of toxic proteins. It is a general method that can be used to classify toxic and non-toxic proteins irrespective of its source of origin. This chapter also explains the "ToxinPred2" web server that allows the prediction of toxins, designing non-toxic proteins, scanning for the motifs, and similarity-based search using the BLAST module.

Chapter 5 – This chapter explains the third objective of the thesis, which is an improved method developed for the prediction of allergenic proteins. We have built an *in-silico* tool for the classification of allergenic proteins and the mapping of IgE epitopes. A web server, "AlgPred 2.0", has also been developed that allows the prediction of allergens, designing, mapping of IgE epitopes, motifs and BLAST search.

Chapter 6 – This chapter corresponds to the fourth objective of the thesis. This describes the novel method "ChAlPred" developed for predicting chemical allergens and designing chemical analogs with desired allergenicity. To aid the research community, this method has been deployed as a smart-device compatible web server that allows to predict and design the chemicals with allergenic properties.

Chapter 7 – This chapter provides the overall summary of the work and concludes by providing a holistic view of the thesis and explains the contribution of the work in the area of immune-informatics.

Chapter 2

Review of Literature

2.1 Virulence factors as a drug/vaccine target

Infectious diseases caused by various pathogenic micro-organisms such as bacteria, viruses, fungi and protozoans are among the leading causes of death globally. In addition to the re-occurrence of fatal infectious diseases, there is an increment in the antimicrobial-resistant strains of the pathogenic organisms, leading to the emergence of new diseases and is an alarming threat to the healthcare sector and the well-being. Research in the past has stated the fact that there are several pathogenic species present in the environment, but the underlying mechanism of causing diseases is found to be common among them. For instance, many pathogenic microbial species share common mechanisms from establishing the contact to the host cell causing infection. Various bacteria, viruses, fungi and other pathogenic species have the same abilities to adhere, invade, colonize, damage the host cell and evade and suppress the host immune response (Weiss, 2002), (Poulin & Combes, 1999). Some of the examples of virulence factors that are possessed by most of the pathogens to infect the host cell are adhesins, invasins, colonizing factors, immune-evasion, immune-suppression, acquisition of the nutrients from the host cell, siderophores, toxins and enzymes which contribute to pathogenicity (Cross, 2008), (Casadevall & Pirofski, 2009). Thus, virulence factors are major drug/vaccine targets for designing therapeutic molecules against these pathogens. Advances in various technologies led to the explosive growth of experimentally verified proteomic data related to virulence factors, which is available in the form of repositories. Therefore, these molecules can also be used in vaccine formulations for priming the host immune system to generate the antibodies in advance so that it can neutralize the effect of pathogens after their encounter to the host cell (Casadevall & Pirofski, 2009) (Casadevall & Pirofski, 1999).

To become a potential drug/vaccine target, virulence factors of the pathogens should meet up several criteria, such as it should be crucial for the survival of the pathogen and do not have any similarity with respect to functions of the host. Also, the target should be ‘druggable’ so that the ligand molecule can alter its function and should be stable, with rare or no point of mutation. It is essential to keep these factors in mind while opting for a suitable drug/vaccine target for efficient anti-microbial drug/vaccine discovery (Silver, 2011). Anti-virulence approaches are proven to be highly compelling in the treatment of the infectious diseases caused by pathogenic species while reducing drug resistance. Drugs that target the virulence factors or mechanism associated with them, such as adhesion, invasion, damaging the host cell,

releasing toxins, causing allergy, suppress the pathogenesis without hampering the normal functioning of the host cell (Rasko & Sperandio, 2010).

2.2 Traditional approaches to identify virulence factors

With the advancements in technology, genomics and proteomics have transformed biological research. Sequencing the whole genome of the pathogenic strains has opened a window of opportunities to identify the putative virulence factors via sequence analysis (Russ & Lampel, 2005). These findings are used for the development of novel computational tools and methods. There are a few bioinformatics-based approaches which are commonly used for the identification of the virulence factors, such as:

2.2.1 Sequence similarity-based search: This is a standard approach which is used in sequence analysis. In this, the query sequence is searched against the sequences in the database of all the proteins with similar functions. This concept can be used to identify the orthologues of virulence factors with known functions. The popular algorithm widely used for sequence similarity is the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990). It provides a heuristic search against a query protein sequence. Virulence factors like adhesin proteins of micro-organisms like *Escherichia coli* (Pichel et al., 2000), *Mycoplasma pneumonia* (Nakane et al., 2011), *Rickettsial species* (Renesto et al., 2006), *Mycoplasma agalactiae* (Fleury et al., 2002) and *Leptospira interrogans* (Palaniappan et al., 2002) have been identified using the BLAST algorithm.

2.2.2 Motif-based search: The motif is a recurring pattern of amino acids or nucleotides that occur in protein or DNA, respectively (D'Haeseleer, 2006). In general, motifs can act as a unique signature for a protein sequence and can be used to detect proteins with similar functional roles. Hence, motifs can also be used to identify the virulence factors in various pathogens, as described in Table 2.1.

Table 2.1: List of motifs associated with adhesion proteins of various pathogens reported in previous studies

Motif	Description	Virulence factor	Pathogen	Reference
C-terminal GPI-motif	Glycosylphosphatidylinositol-modified (GPI) proteins linked to plasma membrane via pre- formed GPI anchor	Adhesion	Fungi (*Candida albicans*)	(De Groot et al., 2003), (Richard & Plaine, 2007)
HExxH motif containing metalloprotease adhesins	Zinc binding sequence motif His-Glu-Xaa-Xaa-His	Adhesion	Bacteria (*Treponema pallidum*)	(Houston et al., 2011)
RGD, SGxG motif	Arginine-glycine-aspartic acid (RGD) and glycosaminoglycan binding site (SGXG) motifs present in autotransporter family proteins pertactin (Prn)	Adhesion	Bacteria (*Bordetella pertussis*)	(Bokhari et al., 2012)
FxxN, GGA (I, L, V)	Tetrapeptide motifs FxxN and GGA (I, L, V) present in polymorphic membrane protein family (Pmp)	Adhesion	Bacteria (*Chlamydia pneumonia*)	(Molleken et al., 2010)

2.2.3 Signal sequence: It is also known as signal or leader peptide, which is a short peptide usually located at N-terminal and occasionally at C-terminal of the protein, which directs the protein towards the secretory pathway. A study conducted by Champion et al. shows that in *Mycobacterium tuberculosis*, virulence factor secretion is promoted by C-terminal signal peptide (Champion et al., 2006). Adhesins are the proteins attached to the membrane that normally have N-terminal signal peptide, which mediates its translocation to the endoplasmic reticulum (Lee & Schneewind, 2001). Hence, this information regarding N-terminal and C-terminal signal sequences can be used to identify the putative adhesin proteins. A computational tool like PrediSi is developed to predict the signal peptides (Hiller et al., 2004). Along with this, SignalP 6.0 recent method developed in 2022 predicts the signal peptides from the protein sequences (Teufel et al., 2022).

2.2.4 Transmembrane domains: These are the membrane-spanning protein domains which traverse across the integral membrane. These domains are predominantly alpha-helices which are packed closely to form a mesh-like structure and can also adopt different conformations. Some adhesin proteins that do not have a leader peptide sequence are known as ‘anchorless adhesins’. Such adhesin proteins have the transmembrane domain, and the prediction of these

domains could be used to identify these virulent proteins. Various methods such as TOPPER (Deng et al., 2013), TOPCONS (Tsirigos et al., 2015), TMHMM (Krogh et al., 2001), AllesTM (Honigschmid et al., 2020) are some of the popular tools which are used for the prediction of transmembrane proteins. These approaches can also be used to identify the putative virulence factors.

2.2.5 Protein domains: These are the conserved region of the protein chain that folds separately and are stable, functional unit (Phillips, 1966). The function of the whole protein sequence is determined by these conserved domains. Some of the virulence factors have well-annotated domains, and the presence of these domains can be used to predict the new virulence factors of the pathogens. There are various computational tools and databases which store the information related to these domains, such as the Conserved Domain Database (Marchler-Bauer et al., 2013), NCBI Conserved Domain Search (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) and InterPro (Blum et al., 2021). These resources can be used to predict the function of the query sequence based on the presence of known virulence factors associated with domains. Table 2.2 shows some of the known conserved domains of the virulence factors of pathogens.

Table 2.2: List of conserved domains in adhesin proteins of pathogens

Conserved Domain	Virulence Factor	Pathogen	Reference
PA14 domain	Toxins, enzymes, adhesins and signaling molecules	Bacteria	(Rigden et al., 2004)
YadA collagen-binding domain	Adhesin	Bacteria (*Yersinia enterocolitica*)	(Nummelin et al., 2004)
Fibrinogen-binding domain (CadF protein)	Adhesin	Bacteria (*Campylobacter jejuni*)	(Konkel et al., 2005)
ALS_N domain	Adhesin	Fungi (*Candida albicans*)	(Phan et al., 2007)
GLEYA domain	Adhesin	Fission yeasts (*Schizosaccharomyces pombe* and *Schizosaccharomyces japonicus*)	(Linder & Gustafsson, 2008)
Gingipain domain	Adhesin	Bacteria (*Porphyromonas gingivalis*)	(Li et al., 2010)

2.3 Identification of virulence factors

2.3.1 Available databases for virulence factors

There are various computational repositories and databases that have been developed for the identification of VFs of various pathogens. Some databases are pathogen-specific, some cover VFs of many pathogenic species and are tabulated in Table 2.3. Pathogens-specific VF repositories like virulence factor database (VFDB), developed by Chen et al., is a well-annotated and comprehensive database of known VFs in bacteria. It also stores the information of other genes and proteins that play a significant role in enhancing the virulence of the bacteria (Chen et al., 2005). VFanalyzer is an automatic pipeline provided by the VFDB group to identify known as well as potential VFs from the genomes of bacterial species (Liu et al., 2019). DFVF is a database of VFs of fungal pathogens known to cause infectious diseases in plant and animal hosts (Lu et al., 2012). Pathosystems Resource Integration Center (PATRIC) is a curation, integration and visualization of bacterial VFs (Snyder et al., 2007), (Davis et al., 2020). ProtVirDB is a database that holds the information of unique virulent proteins of protozoan species (Ramana & Gupta, 2009). Examples of resources that stores the information of VFs irrespective of the pathogenic species include, Victors a knowledgebase that contains the manually curated data regarding the VFs of various pathogenic species such as bacteria, virus, parasite and fungus (Sayers et al., 2019). MvirDB is a centralized microbial data warehouse that comprises all the information from different resources on a single platform. It includes information regarding VFs, toxin proteins, and antibiotic resistance genes from eight open-source databases for various microbial species (Zhou et al., 2007). Pathogen-host interactions database (PHI-base) catalogues pathogenicity, virulence, and effector genes from bacterial, protist and fungal pathogens that infect insects, plants, animals and human hosts (Urban et al., 2020). Apart from the above-discussed databases and repositories, UniProtKB and Swiss-Prot are the knowledgebases that are the central hub containing the information of proteins along with their function, description, structure, modifications with a more accurate and high level of annotation (UniProt, 2021), (Bairoch & Apweiler, 1997).

Table 2.3: List of databases developed for maintaining information regarding VFs of various pathogenic species

Name (Year)	Description (Weblink)	Working Status
MvirDB (2007)	Database of microbial virulent, toxin proteins and antibiotic resistance genes (http://mvirdb.llnl.gov)	No
ProtVirDB (2009)	Database of protozoan virulent proteins (http://bioinfo.icgeb.res.in/protvirdb/)	Yes
DFVF (2012)	Database of VFs for fungal species (http://sysbio.unl.edu/DFVF/)	Yes
VFDB (2019)	Database of VFs for bacterial species (http://www.mgc.ac.cn/VFs/)	Yes
PATRIC (2020)	Curation, integration and visualization of bacterial VFs (https://www.patricbrc.org)	Yes
PHI-base (2020)	Pathogen-host interaction database (http://www.phi-base.org)	Yes
UniProtKB/ Swiss-Prot (2021)	Database of protein sequence and functional information (https://www.uniprot.org)	Yes

2.3.2 Computational resources for virulence factors

Efforts have been made in the past to create data-driven techniques for predicting VFs of the pathogens that can be used as drug and vaccine targets. Some of the *in-silico* methods and tools that are developed for the identification of VFs are listed in Table 2.4. The majority of available tools are extensively specialized for identifying specific virulence factors of certain pathogens. For instance, VICMPred, proposed by Saha et al., is developed to predict the major functions of gram-negative bacterial proteins from their amino acid sequences (Saha & Raghava, 2006b). MAAP is a method for predicting the adhesins and adhesin-like proteins for malarial parasites (Ansari et al., 2008). Another pathogen-specific methods, such as FaaPred (Ramana & Gupta, 2010) and FungalRV (Chaudhuri et al., 2011), have been developed specifically to predict the adhesins and adhesin-like proteins for fungal species. Some methods have been developed to predict the virulent proteins and associated factors irrespective of their pathogenic species. SPAAN, a neural network-based method, has been developed to classify the pathogens' adhesins and adhesins-like proteins (Sachdeva et al., 2005). VirulentPred has been developed to predict the bacterial virulent protein sequences (Garg & Gupta, 2008). MP3 is a tool that predicts pathogenic proteins using genomic and metagenomic data (Gupta et al., 2014). PathoFact is a metagenomic data pipeline for predicting virulence factors and antibiotic resistance genes (de Nies et al., 2021).

Table 2.4: List of *in-silico* tools developed for predicting VFs of various pathogenic species

Name (Year)	Description (Weblink)	Virulence Factor	Pathogen
SPAAN (2005)	Classification of adhesins and adhesins-like proteins of the pathogens (ftp://203.195.151.45)	Adhesins	All pathogens
VICMPred (2006)	A method to predict the major functions of proteins of gram-negative bacteria (https://webs.iiitd.edu.in/raghava/vicmpred/)	All	Bacteria
MAAP (2007)	A method for the identification of adhesins of malarial parasite (https://maap.igib.res.in)	Adhesins	Malarial Parasite
VirulentPred (2008)	Prediction of the bacterial virulent protein (http://bioinfo.icgeb.res.in/virulent/)	All	Bacteria
FaaPred (2010)	A method for the identification of adhesins of fungal species (http://bioinfo.icgeb.res.in/faap/faap.html)	Adhesins	Fungi
FungalRV (2011)	Fungal adhesins prediction method (https://fungalrv.igib.res.in)	Adhesins	Fungi
MP3 (2014)	Prediction of pathogenic proteins using genomic and metagenomic data (http://metagenomics.iiserb.ac.in/mp3/)	All	Bacteria
PathoFact (2021)	Pipeline for the prediction of VFs, toxins and antimicrobial resistance genes (https://git-r3lab.uni.lu/laura.denies/PathoFact/)	All	Bacteria

2.4 Identification of toxicity

Toxins are one of the major virulence factors that play a crucial role in damaging the host cell. Some pathogenic microbes can directly cause damage to the host cells by secreting factors like toxins, hydrolytic enzymes, and physical forces during host cell invasion or escape (Mayer et al., 2013). Secreted factors include hydrolytic enzymes like proteases or lipases and toxins in the form of small metabolites or peptides that poison the host cell and cause tissue damage (Scharf et al., 2014). In general, toxins can be of two types: endotoxins and exotoxins. With the advent of highly accurate and cost-effective methods, the scientific community has adopted data-driven computational methods, such as databases and machine learning techniques, to curate the information regarding toxins and predict the toxicity of the molecules (Pérez Santín E, 2021).

2.4.1 Databases for proteins and chemical toxicity

Various databases have been created to assist the scientific community during the last few decades. These databases provide a variety of information regarding the toxicity of the

molecules, such as hazardous chemicals, newly synthesized chemical compounds, and proteins. Table 2.5 enlists the databases for the toxicity of the chemical compounds and proteins.

Table 2.5: List of repositories developed for maintaining information for the toxicity of the chemical compounds and proteins

Name (Year)	Description (Weblink)	Reference
Databases for toxic chemical compounds		
Distributed Structure Searchable Toxicity (DSSTox) (2002)	Database providing information about chemical structures as well as their toxicity (https://www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database)	(Richard & Williams, 2002)
SuperToxic (2008)	Comprehensive database of toxic compounds (http://bioinformatics.charite.de/supertoxic)	(Schmidt et al., 2009)
Toxin and Toxin Target Database (T3DB) or Toxic Exposome Database (2010)	Database that stores the detailed information of toxins along with its target data (http://www.t3db.ca)	(Lim et al., 2010)
Chemical Entities of Biological Interest (ChEBI) (2010)	Database focused on small chemical compounds (https://www.ebi.ac.uk/chebi/)	(de Matos et al., 2010)
Toxicity Reference Database (ToxRefDB) (2019)	Database containing the structural information of toxic compounds	(Watford et al., 2019)
Comparative Toxicogenomics Database (CTD) (2020)	Database containing the toxicological information for chemicals, genes, phenotypes and diseases (http://ctdbase.org/)	(Davis et al., 2021)
Tox 21 (2020)	Database cataloguing the information of around 8500 chemical compounds	-
RISCTOX	A comprehensive database on toxic and hazardous substances (https://risctox.istas.net/en/)	-
Exposure Forecaster Database (ExpoCastDB)	EPA's database for aggregating chemical exposure information	-
Databases for toxic proteins molecules		
ArachnoServer (2009)	Repository of protein toxins from spiders (http://www.arachnoserver.org)	(Wood et al., 2009)
DBETH (2012)	Database of bacterial exotoxins for human (http://www.hpppi.iicb.res.in/btox/)	(Chakraborty et al., 2012)

BtoxDB (2015)	A comprehensive database of protein structural data on toxin–antitoxin systems (http://www.gurupi.uft.edu.br/btoxdb)	(Barbosa et al., 2015)
TASmania (2019)	Database of bacterial toxin-antitoxin systems (https://shiny.bioinformatics.unibe.ch/apps/tasmania/)	(Akarsu et al., 2019)

2.4.2 Computational tools for proteins and chemical toxicity

Over the years, a plethora of research work has been published to study the toxicity of chemicals, namely DeepTox (Mayr A, 2016), ProTox-II (Banerjee et al., 2018) and eToxPred (Pu et al., 2019). The majority of available tools are extensively specialized for toxins of certain animal origins; for instance, the prediction methods named BTXpred (Saha & Raghava, 2007a) and NTXpred (Saha & Raghava, 2007b) were developed for the classification of bacterial toxins and neurotoxins, respectively. ClanTox, developed by Naamati et al., is a classifier of animal toxins from their primary protein sequences (Naamati et al., 2009). Another method, SpiderP, has been developed to predict the propeptide cleavage sites in spider toxins (Wong et al., 2013). Similarly, ToxClassifier was developed by Gacesa et al. to identify venom toxins (Gacesa R, 2016). Deep learning-based methods such as TOXIFY can be used to classify animal venom proteins from non-toxic proteins (Cole & Brewer, 2019), whereas ToxDL can be used to assess the protein toxicity of animal origin (Pan et al., 2021).

In 2013, Gupta et al. proposed a general method ToxinPred for predicting the toxicity of peptides and proteins irrespective of their source. This method is heavily used by the research community to predict the toxicity of proteins/peptides. It is an SVM-based method that utilizes several features like amino acid composition (AAC), dipeptide composition (DPC) and finding toxic motifs/ regions derived from the sequences (Gupta, Kapoor, et al., 2013). NNTox is a machine learning method to detect the toxicity of protein-based on several gene ontology annotations (Jain & Kihara, 2019). Recently, deep learning-based tools such as ATSE (Wei et al., 2021) and ToxIBLT (Wei et al., 2022) were developed by Wei et al. for the prediction of protein/peptide toxicity using structural, evolutionary and physicochemical properties of the sequences. In addition, a number of methods have been developed for predicting the specific type of toxicity, like hemo-toxicity. In Table 2.6, we provide a comprehensive list of toxicity prediction methods.

Table 2.6: List of computational tools developed for predicting wide range of toxicity of peptides, proteins and small molecules

Tools	Year	Description (Weblink)
Tools for predicting toxicity of chemical compounds		
ToxiPred	2016	Prediction of aqueous toxicity of small chemical molecules (https://webs.iiitd.edu.in/raghava/toxipred/)
HemoPI	2016	Prediction of hemolytic or hemotoxic nature of peptides (https://webs.iiitd.edu.in/raghava/hemopi/)
DeepTox	2016	Prediction of chemical compounds using deep learning (http://www.bioinf.jku.at/research/DeepTox/)
HemoPred	2017	Predicting the hemolytic activity of peptides (http://codes.bio/hemopred/)
ToxiM	2017	Prediction of toxicity of small molecules (http://metagenomics.iiserb.ac.in/ToxiM/)
CLC-Pred	2018	Prediction of the cytotoxicity of a chemical compound (http://way2drug.com/Cell-line/)
ProTox-II	2018	Prediction of toxicity of chemicals (http://tox.charite.de/protox_II)
eToxPred	2019	To predict the toxicity of drug candidates (https://github.com/pulimeng/etoxpred)
HLPpred	2020	Prediction of hemolytic peptides and its activity (http://thegleelab.org/HLPpred-Fuse)
Tools for predicting toxicity of proteins and peptides		
BTXpred	2007	Classification of bacterial toxins (exotoxins and endotoxins) (https://webs.iiitd.edu.in/raghava/btxpred/)
NTXpred	2007	Prediction of neurotoxins (https://webs.iiitd.edu.in/raghava/ntxpred/)
ClanTox	2009	Classification of animal toxins from their primary protein sequences (http://www.clantox.cs.huji.ac.il)
SpiderP	2013	Prediction of the propeptide cleavage sites in spider toxins (http://www.arachnoserver.org/spiderP.html)
ToxClassifier	2016	Prediction of venom toxins from other proteins (http://bioserv7.bioinfo.pbf.hr/ToxClassifier/)
TOXIFY	2019	Deep learning approach for the classification of animal venom proteins (https://www.github.com/tijeco/toxify)

NNTox	2019	Detection of protein toxicity based on gene ontology annotations (http://www.github.com/kiharalab/NNTox)
ToxDL	2020	Prediction of toxic proteins from animal species like snakes and spiders (http://www.csbio.sjtu.edu.cn/bioinf/ToxDL/)
ATSE	2021	Prediction of peptide toxicity with their structural and evolutionary information (http://server.malab.cn/ATSE)
ToxIBLT	2022	A deep learning approach for the prediction of peptide toxicity using information bottleneck and transfer learning (http://server.wei-group.net/ToxIBTL)

2.5 Identification of allergenicity

Several studies have also shown the correlation between microbial colonization and allergic diseases. Allergy is an exaggerated immune response caused by foreign substances called allergens. In some instances, virulent factors are responsible for allergy, which is the hypersensitivity of the immune system. In addition to protein, allergy is also caused by chemical compounds, known as chemical allergy. A therapeutic molecule may cause side effects due to its allergic potential. Based on this, there are various tools and data repositories which maintain the information regarding the allergenic proteins and chemical compounds.

2.5.1 Repositories for proteins and chemical allergens

A substantial number of allergenic proteins and chemicals have been found and characterised in the last few decades. As a result, various databases are developed to aggregate the available scattered information. The World Health Organization /International Union of Immunological Societies (WHO/IUIS) Allergen Nomenclature database maintains the systematic and unambiguous nomenclature for proteins that induce IgE-mediated allergies in humans. AllergenOnline is a repository that provides a peer-reviewed list of allergens (Goodman et al., 2016). Comprehensive Protein Allergen Resource (COMPARE) database stores the information of protein sequences of known allergens (van Ree et al., 2021). AllerBase is a knowledgebase that integrates the data of allergens, its sequences, and IgE epitope binding affinity on a single platform (Kadam et al., 2017). Structural Database of Allergenic Proteins (SDAP) incorporates the information of allergenic protein sequences along with other tools (Ivanciuc et al., 2003). The Immune Epitope Database (IEDB) is a well-known and widely used database that catalogs experimental data of B and T cell epitopes studied in different hosts with respect to various diseases like allergy, autoimmune and infectious diseases (Dhanda et al., 2019). The Database of Allergen Families (AllFam) provides the classification of the

allergens into the known protein families (Radauer et al., 2008). ChEMBL is a manually curated database of bioactive molecules with drug-like properties (Bento et al., 2014). DrugBank Online is a comprehensive, open-access database that contains information on drugs and drug targets (Wishart et al., 2018). Table 2.7 enlists the databases developed for the proteins and chemical allergens.

Table 2.7: List of databases developed for the proteins and chemical allergens

Name	Year	Weblink	Working Status
WHO/IUIS Allergen Nomenclature	2000	http://allergen.org	Yes
SDAP	2003	https://fermi.utmb.edu	Yes
AllFam	2007	https://www.meduniwien.ac.at/allfam/	Yes
ChEMBL	2014	https://www.ebi.ac.uk/chembl/	Yes
AllergenOnline	2016	http://www.allergenonline.org	Yes
COMPARE	2017	https://comparedatabase.org	Yes
AllerBase	2017	http://bioinfo.unipune.ac.in/AllerBase/Home.html	Yes
IEDB	2019	https://www.iedb.org	Yes
DrugBank Online	2018	https://go.drugbank.com	Yes

2.5.2 *In-silico* tools for proteins and chemical allergens

Several computational tools and methods have been developed for the prediction of allergenic proteins. However, limited efforts have been made to develop a method or tool to predict the allergenicity of chemicals causing the allergy. Below is a brief description of the methods developed for the prediction of allergenic proteins. A hybrid method AlgPred (Saha & Raghava, 2006a), was developed that combines the different approaches for predicting allergenic proteins. AllerTool is a method which combines a similarity-based approach for predicting allergenicity and allergic cross-reactivity in proteins (Zhang et al., 2007). AllerHunter was developed for predicting allergenic proteins, where models were developed using SVM-pairwise sequence similarity (Muh et al., 2009). AllerTOP and its updated version, AllerTOPv2, have been developed to classify the allergenic and non-allergenic proteins (Dimitrov et al., 2013), (Dimitrov, Bangov, et al., 2014). In the case of PREAL, models were developed to predict the allergenic protein using biochemical and physicochemical properties (Wang et al., 2013). AllergenFP is a method that incorporates descriptor-based fingerprints for developing prediction models (Dimitrov, Naneva, et al., 2014). AllerCatPro has been developed for predicting the allergenicity potential of a protein from its sequence and 3D

epitope mapping (Maurer-Stroh et al., 2019). Table 2.8 enlists the *in-silico* methods developed to predict the allergenic proteins.

Table 2.8: List of *in-silico* methods developed for the prediction of allergenic proteins

Name	Year	Weblink	Working Status
AlgPred	2006	https://webs.iiitd.edu.in/raghava/algpred/	Yes
AllerTool	2007	http://research.i2r.a-star.edu.sg/AllerTool/	No
AllerHunter	2009	http://tiger.dbs.nus.edu.sg/AllerHunter	No
AllerTOPv2	2013	http://www.ddg-pharmfac.net/AllerTOP/	Yes
PREAL	2013	http://gmobl.sjtu.edu.cn/PREAL/index.php	No
AllergenFP	2014	http://ddg-pharmfac.net/AllergenFP/	Yes
AllerCatPro	2019	https://allercatpro.bii.a-star.edu.sg	Yes

2.6 Conclusion

Virulence factors play an important role in enhancing the pathogenicity of the pathogenic micro-organisms. Multiple studies in the past have revealed that virulent proteins and other associated factors can be used as potential drug and vaccine candidates. *In-silico* tools and methods for the prediction of virulent, toxic and allergenic proteins irrespective of their sources are not available. Hence, proper identification of these factors involved in the virulence as well as pathogenicity, such as toxicity and allergenicity, is the need of an hour to develop proper therapeutics against the pathogens.

Chapter 3

Prediction of virulent proteins

3.1 Introduction

Virulence factors are the molecules synthesized by pathogenic micro-organisms. They have the tendency to cause diseases in different host organisms. They are of several types, such as adhesins, colonization factors, invasions, immune-evasion, immunosuppression, toxins, capsular polysaccharides and siderophores (Peterson, 1996), (Zheng et al., 2012), (Sharma, Dhasmana, et al., 2017). Growth is essential but not always required throughout the pathogen's life. Dormant phase, also known as transient non-replicative phase, can be beneficial for the pathogen, which helps in the persistence of microbe within the host organism (Rittershaus et al., 2013), (Fanning & Mitchell, 2012). It is the common tendency for moderate or no growth of the pathogen associated with resistance to antibiotics and biocides (Brown & Barker, 1999). To grow inside the human host, it is required by the pathogen to take up and metabolize the host-derived nutrients (Ene et al., 2014). Adhesins are the molecules that are involved in the attachment of the pathogen to the host cell, and these are extensively found on the pathogenic microbes. Commensal microbes primarily reside within the mucus, with some distance from the anti-microbial peptide secreting epithelial cells. This helps the pathogen to remain in intimate contact with the host cell, which is required to manipulate and invade it.

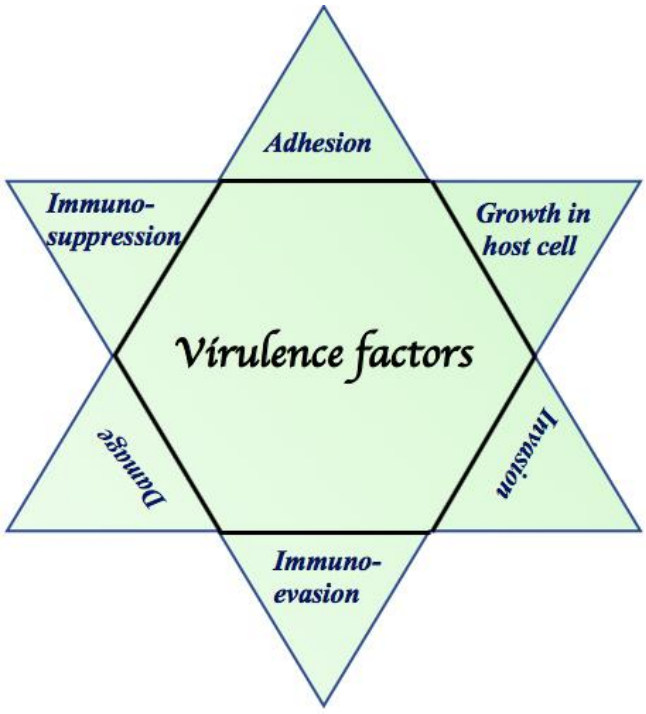

Figure 3.1: Major virulence factors involved in the pathogenesis

Invasion into the host cell and tissues are usually described as one of the defining activities of the pathogenic microbes (Ribet & Cossart, 2015). Some bacteria invade the host cell by

distorting the outer membrane and triggering endocytosis (Dalle et al., 2010). After invasion inside the host cells and tissues, the next step is colonizing, which is mediated by colonization factors. Immuno-evasion helps the pathogen to evade the host immune response by escaping from being recognized by immunogenic surface receptors present on the outer membrane of the host cell, i.e., PAMPs (Latge & Beauvais, 2014), (van de Veerdonk et al., 2008), (Lata & Raghava, 2008). For example, capsular polysaccharides shield the pathogens from the defense mechanism of the host (Chai et al., 2009), (Kaur et al., 2019). Immunosuppression is the obstruction of host immune response by some effector proteins, a popular strategy in pathogenic microbes. These effector proteins mainly interact with host molecules to manipulate the host cell in favor of the pathogen. Effector proteins may include invasins which trigger receptor-mediated host cytoskeleton rearrangements and induce endocytosis (Phan et al., 2007). Traversing the host membrane is very crucial for the action of the effector molecules. Some pathogenic microbes can directly cause damage to the host cells by secreting factors like toxins, hydrolytic enzymes, and physical forces during host cell invasion or escape (Mayer et al., 2013). Secreted factors include hydrolytic enzymes like proteases or lipases and toxins in the form of small metabolites or peptides that poison the host cell and causes tissue damage (Scharf et al., 2014). Virulence factors of the pathogens are depicted in Figure 3.1.

Over the years, limited attempts have been made for the prediction of virulent proteins of the pathogens. The majority of available tools are extensively specialized for certain virulence factors of a certain pathogen. To fill this gap, we have made an attempt to develop a highly accurate method for predicting virulent proteins. Most of the methods are developed on small datasets. To address this limitation, we have proposed a method named “VirFacPred” to classify the virulent and non-virulent protein sequences. Models developed in this study have been trained and evaluated on the latest dataset consisting of 7058 virulent sequences. In addition, several features have been integrated into VirFacPred, which enhance the performance of the model with high precision.

3.2 Materials & methods

3.2.1 Dataset compilation

The dataset used in this study is compiled from different databases, such as VFDB (Liu et al., 2019) and Victors (Sayers et al., 2019). Besides this, we also extracted data from UniProt release 2021_04 (released on 17 November 2021) (UniProt, 2021) using various other terms related to the virulence of the pathogens, such as virulence, adhesin, adhesion, adherence,

toxin, invasion, and capsule. It is challenging to obtain a negative dataset where experimentally validated data is not readily available. Thus, in this study, we carefully extracted and assigned non-virulent proteins from UniProt. We extracted 565 918 proteins using the query 'NOT virulence proteins AND reviewed:yes'; these proteins were assigned as a negative dataset. In this study, we have only taken the proteins which are reviewed and manually annotated. All the protein sequences comprising 'BJOUXZ', and non-virulent sequences similar to virulent sequences were removed. After all the pre-processing, 7058 sequences were referred to as positive dataset, and 462 844 sequences were called as negative dataset. After that, CD-HIT (Li & Godzik, 2006) with 40% sequence identity was applied to both datasets. It leads to a reduced number of sequences for positive and negative datasets. After applying CD-HIT, the positive dataset is reduced to 4714 sequences from 7058, whereas the negative dataset is reduced to 89310 sequences from 462 844. In the study, we have created two datasets: Main Dataset: It contains 7058 virulent and 7058 non-virulent proteins redundant sequences. Alternate Dataset: It contains 4714 virulent and 4714 non-virulent proteins. In case of main dataset, we have not remove redundant virulent proteins, the aim was to train our model on largest possible dataset. In case of alternate dataset, redundant virulent protein has been removed, the aim was to train our model on non-redundant dataset to follow standard practice in the field of classification. The construction of the datasets are shown in Figure 3.2

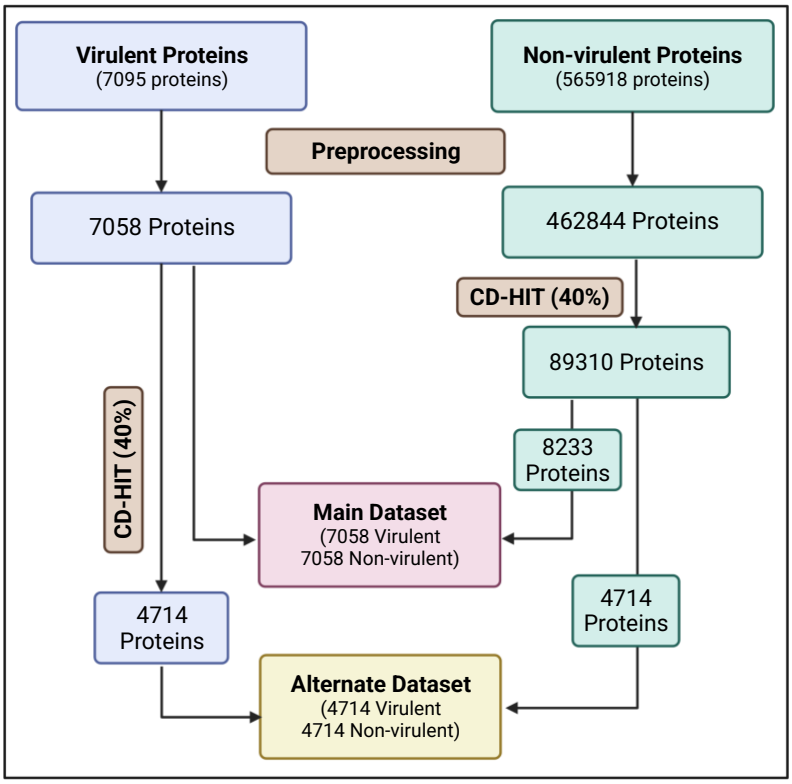

Figure 3.2: Creation of datasets followed in the study

3.2.2 Generation of protein features

We have generated more than 9000 features using the well-known feature extraction method, Pfeature (Pande A, 2019). We have employed a composition-based module of Pfeature to compute 9529 features which are further used for developing machine learning models. The detailed information of each feature, along with the length of the vector, is tabulated in Table 3.1 .

Table 3.1: List of all the computed features along with their vector length

Name of the Feature	Feature vector length
Amino acid composition (AAC)	20
Amphiphilic pseudo amino acid composition (APAAC)	23
Atom composition (ATC)	5
Bond composition (BTC)	4
Composition enhanced Transition Distribution (CTD)	189
Conjoint Triad Calculation (CTC)	343
Dipeptide composition (DPC)	400
Distance distribution of residue (DDOR)	20
Physicochemical Properties Composition (PCP)	30
Pseudo amino acid composition (PAAC)	21
Quasi-sequence order (QSO)	42
Residue Repeat Information (RRI)	20
Shannon Entropy of Physicochemical Property (SPC)	25
Shannon Entropy of Residues (SER)	20
Shannon-Entropy of Protein (SEP)	1
Tripeptide composition (TPC)	8000

3.2.3 BLAST for similarity search

BLAST is used to identify the virulent proteins based on the similarity of a protein with virulent and non-virulent sequences (Altschul et al., 1990). The BLAST search module was developed, where the query sequences were searched against the database of virulent and non-virulent proteins. The performance of the method was evaluated according to various E-value thresholds.

3.2.4 Motif analysis

The motifs corresponding to virulent proteins were extracted using Motif-EmeRging and with Classes-Identification (MERCI) program (Vens et al., 2011).

3.2.5 Feature selection and ranking

Past studies have shown that all the features generated are not important; hence it is necessary to select the relevant features from an extensive feature set. For this, the SVC-L1-based feature selection technique was used to fetch the significant features (Sharma et al., 2021). We have selected 132 features for the main dataset and 91 for the alternate dataset from 9529 features. Moreover, feature ranking was done using a decision tree-based algorithm, Light Gradient Boosting Machine (LightGBM), to rank the selected features (Sharma et al., 2021). The obtained top-ranked features were used to build the different machine learning prediction models in both datasets, respectively (Sharma et al., 2022).

3.2.6 Machine learning models

In the current study, different machine learning techniques have been used for the classification of virulent and non-virulent protein sequences. We used Logistic Regression (LR), k-nearest neighbors (KNN), Decision Tree (DT), Gaussian Naive Bayes (GNB), XGBoost (XGB), Support Vector Classifier (SVC), and Random Forest (RF) based techniques for the classification.

3.2.7 Evaluation parameters

We have also applied 5-fold cross-validation (CV) on 80% of training data for the internal training, testing and model evaluation. The performance of machine learning models was evaluated using the standard evaluation parameters such as sensitivity (Sens), specificity (Spec), accuracy (Acc), Matthews correlation coefficient (MCC) and area under the receiver operating characteristic (AUC).

3.2.8 Hybrid approach

We have also applied a hybrid approach to enhance the accuracy of the prediction model. For this, the following three techniques have been integrated: (i) similarity-based approach using BLAST, (ii) motif-based approach using MERCI and (iii) Machine learning-based techniques. An ensemble BLAST (top five hits) was initially used to classify the given protein sequence at E-value of $10^{-6.}$ The score of '+0.5' was allotted for virulent proteins, '-0.5' for non-virulent proteins and '0' for no prediction. Next, using MERCI based approach, the same protein sequence was classified. The score of '+0.5' was allotted if the motif was present and '0' if the motif was absent. Finally, for developing the hybrid approach, we have computed the overall score, which has been obtained by integrating the scores of these three methods. The protein

sequence is categorized as virulent and non-virulent based on the overall score at different threshold values. The overall workflow of the methodology is depicted in Figure 3.3.

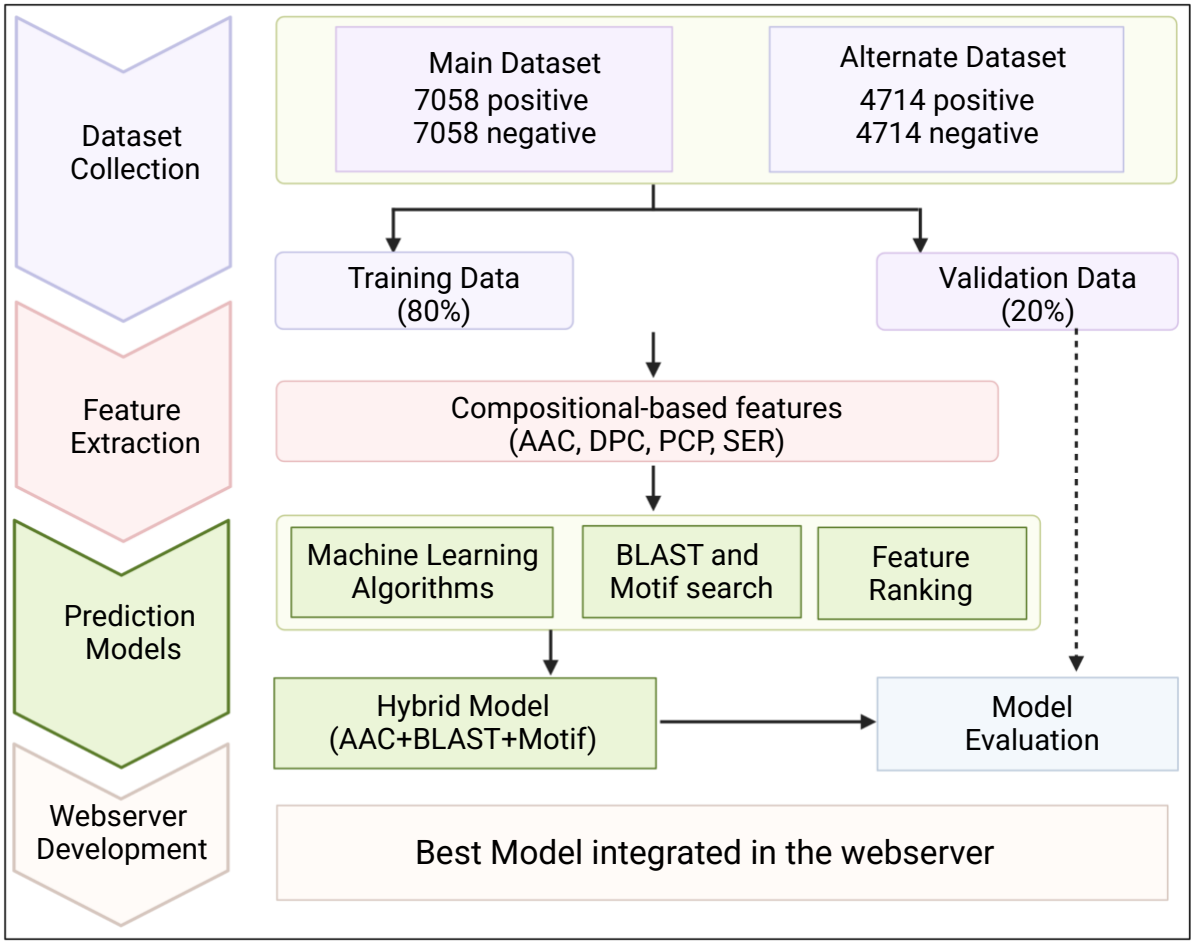

Figure 3.3: Complete methodology used for predicting virulent factors

3.3 Results

3.3.1 Compositional analysis

The amino acid composition (AAC) was computed for virulent and non-virulent proteins. We observed that the average AAC of residues such as alanine, glycine, isoleucine, and leucine are higher in the virulent sequences. In contrast, cysteine, glutamate, proline and serine are abundant in the non-virulent sequences. Figure 3.4 depicts the comparison of average AAC for virulent and non-virulent proteins.

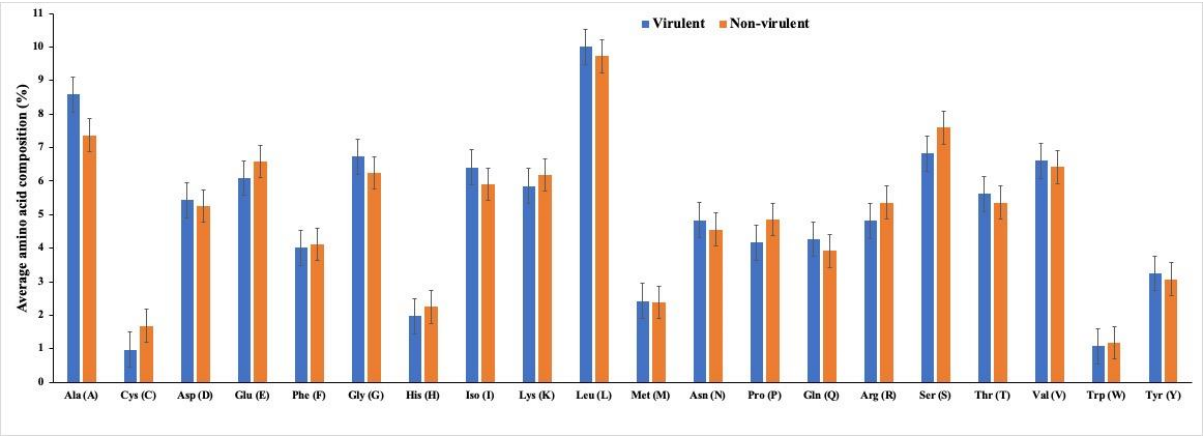

Figure 3.4: Shows amino acid composition of virulent and non-virulent proteins

3.3.2 Similarity search-based prediction

We have applied an ensemble of top five hits to reduce the false prediction. For the main dataset, the number of correct hits (sensitivity) increased from 17.25% to 21.62% for the training dataset and from 18.91% to 23.26% for the validation dataset with E-value ranging from 10^{-6} to 10^{-1}. This also increases the number of wrong hits (error), as shown in Table 3.2.

Table 3.2: The performance of BLAST-based search on main dataset

E-value	Training				Validation			
	Virulent		Non-virulent		Virulent		Non-virulent	
	Correct hits (Sens)	Wrong hits (error)	Correct hits (spec)	Wrong hits (error)	Correct hits (Sens)	Wrong hits (error)	Correct hits (spec)	Wrong hits (error)
10^{-6}	1948 (17.25%)	97 (0.86%)	486 (4.3%)	226 (2%)	534 (18.91%)	41 (1.45%)	159 (5.63%)	55 (1.95%)
10^{-5}	2033 (18%)	106 (0.94%)	539 (4.77%)	238 (2.11%)	554 (19.62%)	45 (1.59%)	173 (6.13%)	56 (1.98%)
10^{-4}	2128 (18.85%)	123 (1.09%)	588 (5.21%)	262 (2.32%)	578 (20.47%)	48 (1.7%)	186 (6.59%)	64 (2.27%)
10^{-3}	2213 (19.6%)	136 (1.2%)	650 (5.76%)	290 (2.57%)	599 (21.21%)	54 (1.91%)	206 (7.29%)	73 (2.58%)
10^{-2}	2322 (20.56%)	169 (1.5%)	738 (6.54%)	320 (2.83%)	623 (22.06%)	58 (2.05%)	233 (8.25%)	76 (2.69%)
10^{-1}	2441 (21.62%)	197 (1.74%)	859 (7.61%)	353 (3.13%)	657 (23.26%)	73 (2.58%)	267 (9.45%)	88 (3.12%)

3.3.3 Performance of prediction models

3.3.3.1 Composition-based features

The features including AAC of virulent and non-virulent were computed to develop several machine learning models. The RF-based model was observed to perform relatively well

compared to other models for both datasets. For main dataset, it achieved a maximum AUC of 0.83 and 0.84 on training and validation datasets, respectively. For the alternate dataset, model attained AUC of 0.77 and 0.78 on training and validation datasets, respectively (Table 3.3).

Table 3.3: The performance of ML-based models developed using amino acid composition

Main Dataset										
ML	**Training**					**Validation**				
	Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	63.49	64.14	63.82	0.69	0.28	61.67	65.76	63.70	0.71	0.28
RF	74.68	73.89	74.28	0.83	0.49	75.25	74.47	74.86	0.84	0.50
LR	67.46	68.30	67.88	0.74	0.36	67.37	68.69	68.02	0.74	0.36
XGB	72.36	72.53	72.44	0.80	0.45	72.15	73.97	73.05	0.81	0.46
KNN	70.49	72.03	71.26	0.80	0.43	71.03	72.90	71.96	0.80	0.44
GNB	66.87	66.57	66.72	0.72	0.33	68.14	67.83	67.99	0.73	0.36
SVC	74.38	74.56	74.47	0.82	0.49	73.84	75.39	74.61	0.83	0.49
Alternate Dataset										
DT	61.65	61.36	61.51	0.66	0.23	64.95	61.72	63.31	0.68	0.27
RF	70.14	69.88	70.01	0.77	0.41	70.54	69.25	69.88	0.78	0.40
LR	67.34	65.57	66.46	0.72	0.33	68.82	65.27	67.02	0.73	0.34
XGB	68.47	69.67	69.07	0.76	0.38	66.24	69.67	67.98	0.75	0.36
KNN	67.79	69.32	68.55	0.76	0.37	66.77	70.71	68.77	0.76	0.38
GNB	66.07	65.62	65.85	0.71	0.32	64.52	63.60	64.05	0.70	0.28
SVC	70.24	69.80	70.02	0.76	0.40	70.54	70.61	70.57	0.76	0.41

3.3.3.2 Models using selected features

The reduced features 132 (main dataset) and 91 (alternate dataset) were used to develop different classification models on both datasets. The performance of these models is illustrated in Table 3.4. It is clearly shown in the table that the RF-based model performed better for both datasets.

Table 3.4: The performance of ML-based models developed using selected features

Main Dataset										
ML	**Training**					**Validation**				
	Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	61.76	61.88	61.82	0.66	0.24	65.75	62.20	63.99	0.69	0.28
RF	74.02	76.17	75.10	0.84	0.50	71.82	77.60	74.68	0.84	0.50

LR	72.64	72.35	72.49	0.79	0.45	71.31	73.47	72.38	0.79	0.45
XGB	73.58	73.94	73.76	0.82	0.48	71.38	75.54	73.44	0.82	0.47
KNN	60.95	61.05	61.00	0.66	0.22	58.79	63.48	61.12	0.66	0.22
GNB	69.09	68.23	68.66	0.75	0.37	67.79	69.12	68.45	0.75	0.37
Alternate Dataset										
DT	59.83	59.37	59.60	0.64	0.19	58.39	59.00	58.70	0.63	0.17
RF	70.80	72.11	71.45	0.79	0.43	69.57	71.23	70.41	0.78	0.41
LR	71.56	70.30	70.94	0.77	0.42	70.32	69.67	69.99	0.76	0.41
XGB	71.22	70.22	70.72	0.78	0.41	67.74	69.04	68.40	0.76	0.37
KNN	55.21	58.38	56.79	0.60	0.14	53.76	56.28	55.04	0.57	0.11
GNB	67.76	67.83	67.79	0.74	0.36	66.02	71.03	68.56	0.74	0.37

3.3.3.3 Motif-based approach

The motifs such as 'LSSGLRI, KDDAAG, SAKDDA and SGLRIN' are solely found in virulent proteins. Composition-based models (AAC) built using different ML techniques were integrated with the MERCI approach. The performance of the combined approach (ML+MERCI) for both datasets is shown in Table 3.5.

Table 3.5: The performance of motif-based approach when combined with machine learning

Main Dataset										
ML	**Training**					**Validation**				
	Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
RF	76.17	72.83	74.50	0.83	0.49	76.58	73.32	74.97	0.84	0.50
SVC	69.43	78.62	74.04	0.82	0.48	69.27	79.46	74.33	0.83	0.49
XGB	75.30	70.19	72.74	0.80	0.46	75.25	71.04	73.16	0.81	0.46
KNN	64.78	77.00	70.90	0.80	0.42	65.68	77.89	71.74	0.80	0.44
LR	71.06	66.30	68.68	0.74	0.37	71.45	65.76	68.63	0.75	0.37
GNB	60.68	72.05	66.37	0.72	0.33	61.81	72.61	67.17	0.73	0.35
DT	65.81	62.20	64.00	0.69	0.28	65.82	63.91	64.87	0.71	0.30
Alternate Dataset										
RF	74.50	66.37	70.45	0.78	0.41	75.05	66.32	70.63	0.79	0.42
SVC	66.44	73.76	70.09	0.77	0.40	66.02	74.58	70.36	0.77	0.41
KNN	78.75	60.72	69.77	0.77	0.40	78.82	61.93	70.26	0.77	0.41
XGB	73.41	65.94	69.69	0.77	0.40	70.97	66.00	68.45	0.76	0.37
LR	69.82	64.66	67.25	0.73	0.35	71.51	64.12	67.76	0.74	0.36
GNB	71.32	61.60	66.47	0.72	0.33	70.65	59.52	65.01	0.71	0.30
DT	65.51	58.68	62.11	0.67	0.24	68.28	59.83	64.00	0.69	0.28

3.3.3.4 BLAST-based model

The similarity search approach BLAST and ML-based models were synergized to build an enhanced method. The BLAST search was initially implemented for a query sequence; if a BLAST hit was obtained, the query sequence was assigned as virulent and non-virulent based on the BLAST result. If no hit is obtained, then the composition-based model is utilized to predict the same sequence. Table 3.6 shows the performance of BLAST when combined with ML.

Table 3.6: The performance of BLAST when combined with machine learning

Main Dataset										
ML	**Training**					**Validation**				
	Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
KNN	85.73	88.21	86.97	0.96	0.74	87.48	88.59	88.03	0.96	0.76
RF	88.89	86.33	87.61	0.96	0.75	89.81	87.16	88.49	0.97	0.77
LR	83.71	89.36	86.54	0.95	0.73	84.74	90.01	87.36	0.95	0.75
SVC	85.88	90.13	88.01	0.95	0.76	87.06	90.87	88.95	0.96	0.78
XGB	85.08	88.95	87.02	0.95	0.74	87.20	89.8	88.49	0.95	0.77
GNB	81.83	86.42	84.13	0.93	0.68	82.63	86.73	84.67	0.94	0.69
DT	81.37	84.21	82.79	0.91	0.66	82.28	85.95	84.1	0.93	0.68
Alternate Dataset										
KNN	87.45	76.21	81.85	0.92	0.64	89.78	79.92	84.78	0.94	0.71
RF	84.17	79.94	82.06	0.92	0.64	86.77	82.22	84.46	0.94	0.69
SVC	79.71	85.55	82.62	0.92	0.65	82.15	88.61	85.42	0.94	0.71
LR	82.48	77.86	80.18	0.91	0.61	85.48	80.65	83.03	0.93	0.66
XGB	84.06	79.54	81.81	0.91	0.64	85.05	82.32	83.67	0.93	0.67
GNB	81.21	77.78	79.51	0.89	0.59	84.09	79.39	81.71	0.92	0.64
DT	80.26	74.53	77.41	0.87	0.55	84.19	78.14	81.12	0.91	0.62

3.3.3.5 Models using hybrid approach

In order to overcome the limitations of individual methods, different approaches have been integrated. These approaches were used to predict the virulent proteins with better performance. AAC-based ML model is combined with BLAST-based similarity and MERCI-based motif approaches. First, proteins were classified using ensemble BLAST at E-value of 10^{-6} followed by MERCI approach. The protein sequences left unpredicted by these two approaches are predicted by an ML-based model. The performance of the hybrid method is significantly improved, which is not achievable by using all these methods independently. The performance of the hybrid method is shown in Table 3.7. RF-based model performed best for

both the datasets on training and validation datasets. It achieved AUC of 0.96 and 0.97 (main dataset) and AUC of 0.92 and 0.94 (alternate dataset) on training and validation dataset.

Table 3.7: The performance of hybrid method combining machine learning, BLAST and MERCI

Main Dataset										
ML	**Training**					**Validation**				
	Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
KNN	85.73	88.22	86.98	0.96	0.74	87.55	88.59	88.07	0.96	0.76
RF	88.89	86.37	87.63	0.96	0.75	89.87	87.16	88.53	0.97	0.77
LR	83.73	89.64	86.69	0.95	0.74	84.81	90.23	87.51	0.95	0.75
SVC	85.88	90.15	88.02	0.95	0.76	87.13	90.87	88.99	0.96	0.78
XGB	85.08	89.01	87.04	0.95	0.74	87.27	89.8	88.53	0.95	0.77
GNB	81.85	86.42	84.14	0.93	0.68	82.71	86.73	84.71	0.94	0.69
DT	81.37	84.31	82.84	0.91	0.66	82.35	86.02	84.17	0.93	0.68
Alternate Dataset										
KNN	89.89	80.02	84.89	0.94	0.71	87.58	76.37	81.99	0.92	0.64
RF	84.28	80.02	82.15	0.92	0.64	86.88	82.32	84.57	0.94	0.69
SVC	79.97	85.63	82.79	0.92	0.66	82.37	88.71	85.58	0.94	0.71
LR	82.58	78.34	80.47	0.91	0.61	85.59	81.28	83.42	0.93	0.67
XGB	84.28	79.56	81.93	0.92	0.64	85.16	82.53	83.83	0.93	0.68
GNB	81.45	77.94	79.72	0.89	0.59	84.19	79.62	81.87	0.92	0.64
DT	80.51	74.77	77.65	0.87	0.55	84.41	78.45	81.39	0.91	0.63

3.3.3.6 Best models developed in the study

The performance of the best classification models developed using the different features is listed in Table 3.8 .

Table 3.8: List of the features along with the performance of the best machine learning algorithm

Method	**Features**	**ML**	**Training**					**Validation**				
			Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
Composition based features	AAC	RF	74.68	73.89	74.28	0.83	0.49	75.25	74.47	74.86	0.84	0.5
Feature Selection	132 features	RF	74.02	76.17	75.10	0.84	0.5	71.82	77.6	74.68	0.84	0.5
Motif based approach	Motifs + AAC	RF	76.17	72.83	74.50	0.83	0.49	76.58	73.32	74.97	0.84	0.5
BLAST based approach	BLAST + AAC	RF	88.89	86.33	87.61	0.96	0.75	89.81	87.16	88.49	0.97	0.77
Hybrid approach	BLAST + Motifs + AAC	RF	88.89	86.37	87.63	0.96	0.75	89.87	87.16	88.53	0.97	0.77

3.4 Web-based service

A freely accessible web server named "VirFacPred" (https://webs.iiitd.edu.in/raghava/virfacpred/) has been developed to predict the virulent proteins. The key modules such as (i) prediction, (ii) design, (iii) motif scan, (iv) BLAST search and (v) Download are integrated into the web server. Prediction: It permits the user to submit the single as well as multiple protein sequences in FASTA format. This module can efficiently classify virulent and non-virulent proteins based on our two best performing models, i.e., Model-1 (AAC-based RF approach) and Model-2 (hybrid approach). Design: It generates all the possible mutants corresponding to a protein by mutating a single residue at a time. Motif scan: It uses MERCI software to detect the motifs present in the virulent protein sequences. It also maps or scans the motifs in the query protein sequence given by the user and distinguishes them as virulent and non-virulent. BLAST search: It assists the user to perform a similarity-based search using BLAST against a virulent and non-virulent protein database. Download: Using this module, the user can download the python-based standalone package as well as the dataset used in the study.

3.5 Benchmarking with previous methods

It is necessary to benchmark the existing methods with our new proposed method VirFacPred. Our model outperformed the various baseline methods such as SPAAN, VICMPred, VirulentPred, and MP3. For instance, SPAAN is a neural network-based software specifically for predicting adhesins and adhesins-like proteins of the pathogens. It has attained a sensitivity of 89% and specificity of 100% on the dataset. The method named VICMPred, developed by Saha et al. is an SVM-based classification approach that classifies bacterial proteins into four different functional classes: virulence factors, information molecule, cellular process and metabolism molecules, and has achieved an accuracy of 70.75%. VirulentPred, another SVM-based method developed by Garg et al., predicts the virulent proteins of bacterial pathogens. It has attained AUC of 0.86 with MCC of 0.64. Our method VirFacPred is trained on a larger dataset containing 7058 virulent and non-virulent proteins. Our hybrid approach was implemented, combining AAC-based RF model, similarity-based search using BLAST, and motif search by MERCI. It achieved AUC of 0.97 and MCC of 0.77 with balanced sensitivity and specificity. The comparison of VirFacPred with existing methods is shown in Table 3.9.

Table 3.9: Comparison of proposed method VirFacPred with existing methods

Method	Type of method	Pathogen Type	Sens	Spec	Acc	AUC	MCC
VirFacPred	General	All pathogens	89.87	87.16	88.53	0.97	0.77
FungalRV	Specific	Fungus	NA	NA	NA	NA	0.87
FaaPred	Specific	Fungus	82.66	86.8	86.05	NA	0.61
VirulentPred	General	All pathogens	82.00	81.50	81.80	0.86	0.64
MAAP	Specific	Malarial parasite	NA	NA	NA	NA	0.80–0.904
VICMPred	Specific	Bacteria	NA	NA	70.75	NA	NA
SPAAN	General	All pathogens	89.00	100	NA	NA	NA

3.6 Discussion & conclusion

The human body has an environment that is needed by any pathogen to grow and survive. The attributes required for survival and growth in this particular environment are generally considered as virulence factors. These virulence factors play a significant role in favor of the pathogen both during the pathogenesis and growth phase. It protects the microbe from the mammalian body temperature and helps escape phagocytosis. In a broader sense, virulence factors are necessary for the survival of the pathogen. Direct action of the virulence factors into the host response leads to the disruption of the host homeostasis, leading to disease. Thus, there is a dire need to determine the virulence property of the proteins produced by different pathogens. In the present study, we have proposed a method named VirFacPred for predicting virulent and non-virulent proteins irrespective of their source. The similarity-based search using BLAST has been carried out to identify query sequences. If a query protein sequence is similar to a known protein, then the same function is assigned to that query protein. It has been observed from Table 3.2 that BLAST can identify some virulent proteins with the probability of correct prediction of more than 23%, with an extremely low error rate. The result suggests that BLAST produces a substantial number of no hits; thus, it cannot be used to predict the unknown protein that is not similar to known virulent and non-virulent proteins. In order to overcome this constraint, a hybrid model was developed using AAC-based ML model, BLAST and MERCI. The highest performance is attained using a hybrid model with maximum accuracy, as shown in Table 3.7. In this study, we created a web-based platform for users to classify virulent and non-virulent proteins. We have provided a freely accessible web server and a standalone package of “VirFacPred” to assist the scientific community working on protein therapeutics.

3.7 Limitation of the study

In this study, we have proposed a highly accurate method for predicting virulent proteins. However, this is a general method developed for predicting the virulent proteins irrespective of the specific pathogen. We have considered all the pathogens and have tried to develop a generalised method on latest dataset with high precision.

Chapter 4

Identification of toxins and designing of non-toxic proteins

4.1 Introduction

Proteins and peptides are naturally occurring molecules that play various functions and processes in the body that are essential to sustain cellular mechanisms (Shaji & Patole, 2008). Their aberrant activity has been involved in several pathological conditions such as cancer, neurodegenerative disorders and diabetes. Thus, using them as therapeutic agents is a promising way to fight against various diseases. In recent years, they have the potential to revolutionize medical therapy as well as a preferred choice over small molecules and antibodies due to their high target specificity, tissue penetration, high biological activity and inexpensive (Bruno et al., 2013). However, there are certain prime concerns in the development of protein/peptide-based drug discovery, such as toxicity, immunogenicity and stability (Otvos & Wade, 2014). Due to this, the assessment of toxic properties of proteins/peptides is of great necessity.

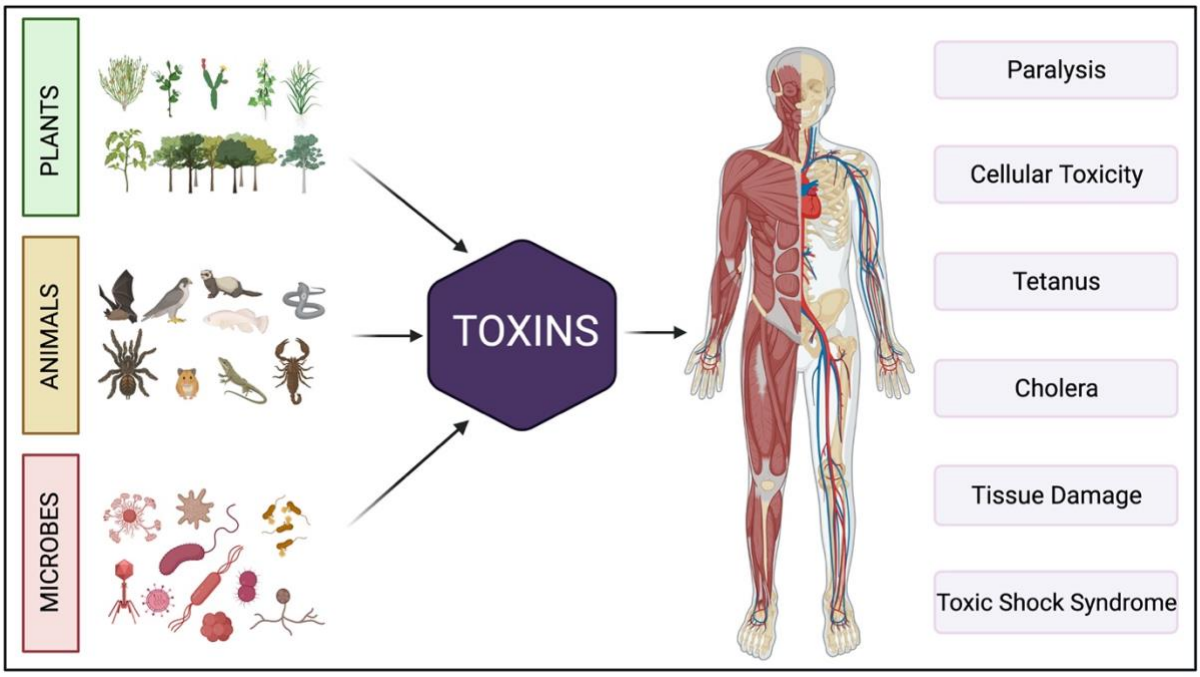

Figure 4.1: Source of toxins and their effects on humans (10.1093/bib/bbac174)

Toxins are naturally occurring poisonous substances that have the ability to cause harm to other organisms. It can be of plant, animal origin or released by several types of microbes (Clark et al., 2019). It can cause deadly diseases or death when it enters the body. A variety of toxins from certain animals can lead to several lethal effects, such as scorpion venom can overstimulate neuronal signalling leading to paralysis (Petricevich, 2010). Further, snake venom could be neurotoxic, causing neuromuscular paralysis as well as haemotoxic damaging

the circulatory system and leading to acute tissue damage (Casewell et al., 2020), (Slagboom et al., 2017). The various effects of toxins from different sources are depicted in Figure 4.1. Conventional experimental techniques are used to evaluate the toxicity of unknown proteins, peptides and chemical compounds. However, these approaches are laborious, cost-intensive and involve animal testing for in-vivo assessment. These impediments lead to an inclination towards the applicability of in-silico techniques (Sharma N, 2021). With the advent of highly accurate and cost-effective methods, the scientific community has adopted data-driven computational methods, such as machine learning techniques, to predict the toxicity of molecules (Pérez Santín E, 2021). In this study, we have attempted to develop a highly accurate method for predicting the toxicity of large proteins, which will complement our previous method ToxinPred. Although, ToxinPred is highly accurate and used widely by the scientific community, but there are several constraints that necessitate improvement. ToxinPred was trained on 1805 toxic peptides where the maximum length was 35 amino acids. Thus, ToxinPred is suitable only for peptides or small peptides of length up to 50 amino acids but not suitable for large proteins. To address these limitations, we have proposed the updated method named "ToxinPred2" to classify the toxic and non-toxic protein sequences, which is trained and evaluated on large proteins/toxins. Models developed in this study have been trained and evaluated on the latest dataset consisting of 8233 toxic sequences. In addition, several features have been integrated into ToxinPred2, which enhance the performance of the model with high precision.

4.2 Materials & methods

4.2.1 Dataset collection

The dataset was retrieved from UniProt release 2021_03 (released on 2 June 2021) (UniProt, 2021) using different keywords for obtaining toxic and non-toxic proteins/peptides. We extracted 9940 toxic proteins using the keyword 'toxin AND reviewed:yes'. All protein sequences comprising 'BJOUXZ', less than 35 amino acids and non-toxic sequences similar to toxic sequences were discarded. Ultimately, we obtained 8233 toxic sequences, referred to as a positive dataset. The compilation of experimentally validated or well-annotated toxic peptides is possible, whereas it is challenging to obtain non-toxic peptides. Therefore, we have extracted the negative dataset from Swiss-Prot (Bairoch & Apweiler, 2000) using keywords 'NOT toxin NOT allergen AND reviewed: yes' and obtained 554 145 proteins. In the present study, we have considered proteins that are reviewed and manually curated. From this data, we

have discarded the sequences with length less than 35 amino acids and non-standard characters. Hence, we proceeded with 460 257 non-toxic sequences as a negative dataset. The creation of datasets is depicted in Figure 4.2.

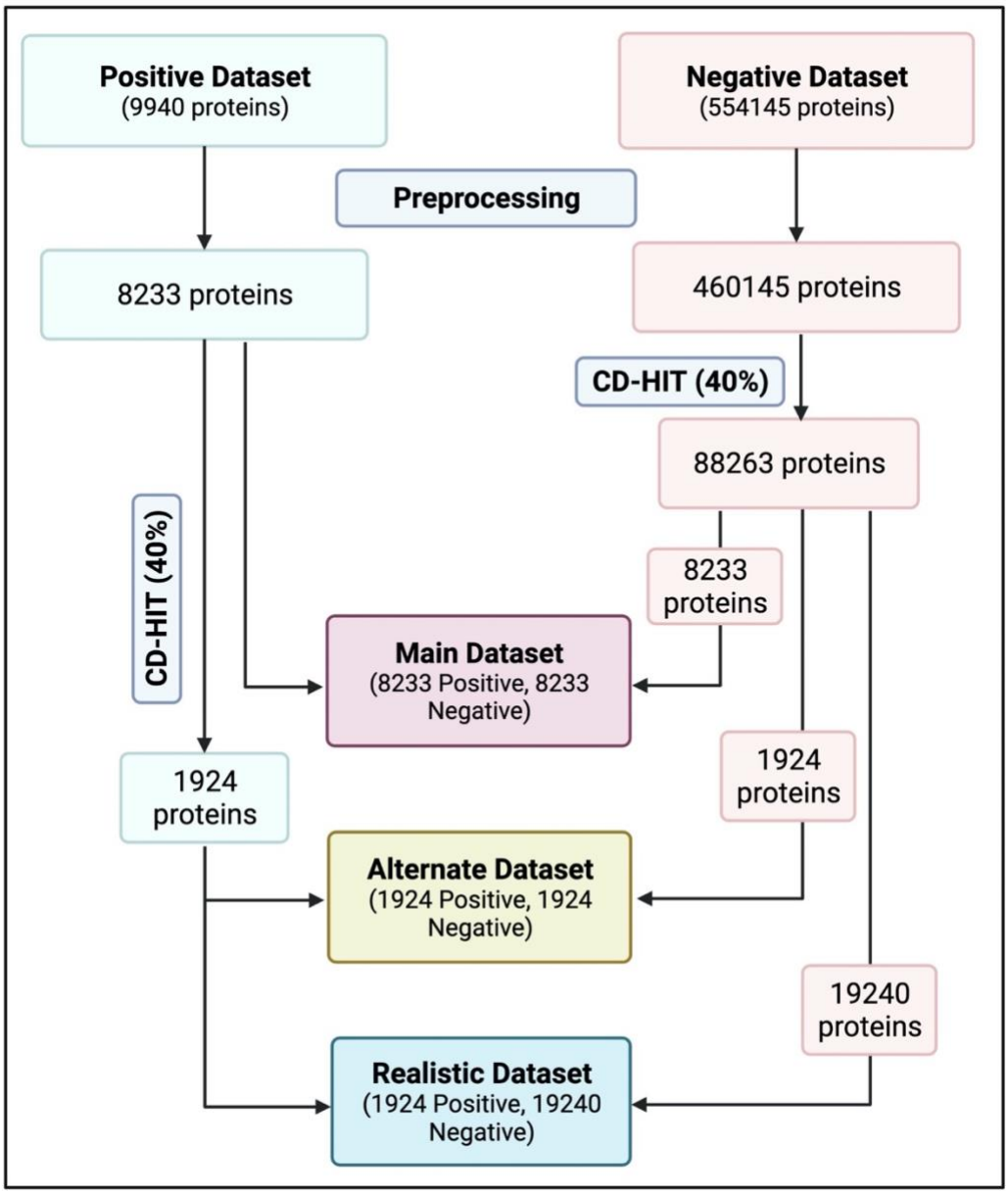

Figure 4.2: Compilation of datasets for developing toxin prediction method (10.1093/bib/bbac174)

After that, CD-HIT software (Li & Godzik, 2006) was applied to both datasets at 40% sequence identity. It leads to a reduced number of sequences for positive and negative datasets. After applying CD-HIT, the positive dataset is reduced to 1924 sequences from 8233, whereas the negative dataset is reduced to 88263 sequences from 460 257. We have created three datasets based on the number of toxic and non-toxic protein sequences, as described below:

(a) Main dataset: This dataset contains 8233 toxic (obtained after pre-processing of positive data) and 8233 non-toxic (randomly selected from 88263 negative data obtained after CD-HIT) protein sequences.
(b) Alternate dataset: This dataset contains 1924 toxic (obtained after applying CD-HIT on 8233 positive data) and 1924 non-toxic (randomly selected from 88263 sequences obtained after CD-HIT) non-redundant protein sequences. No two proteins have more than 40% sequence similarity in this dataset.
(c) Realistic dataset (Ten times Negative Dataset): This dataset consists of 1924 toxic and 19240 non-toxic protein sequences. These toxic sequences are same as those used in the alternate dataset, where no two proteins have more than 40% similarity. The non-toxic protein sequences were randomly selected from non-redundant 88263 non-toxic sequences obtained after applying CD-HIT.

4.2.2 BLAST-based similarity search

In this study, we have used BLAST to identify toxins based on the similarity of a protein sequence with toxic and non-toxic sequences (Altschul et al., 1990). The similarity-based search module was created in which the query sequences were searched against the database of toxins and non-toxins. The performance of the method was assessed based on the various E-value cutoffs.

4.2.3 Scanning of motifs

The toxic proteins were searched for the motifs using MERCI tool, a program to locate motifs in any sequence (Vens et al., 2011). Motif analysis provides the information related to recurring patterns present in the toxic sequences.

4.2.4 Feature generation

We have used a standalone tool, Pfeature, to generate a wide range of features such as composition and evolutionary information-based features (Pande A, 2019). Using a composition-based feature module of Pfeature, a vector of 9163 features was calculated against each sequence for all three datasets. To extract the evolutionary information for a given protein, Position-specific scoring matrix (PSSM) composition was calculated using Position-Specific Iterated BLAST (PSI-BLAST) (Altschul et al., 1997).

4.2.5 Feature selection and ranking

It has been shown in previous studies that all the features are not significant. Thus, selecting the relevant features from a larger set of features is a major challenge. In this study, we have used the SVC-L1-based feature selection technique to select the significant features from the high-dimensional feature set (Sharma et al., 2021). Using this method, we have listed important features for all three datasets from the pool of 9163 features. Out of that, 129 features were selected for the main dataset, 32 for the alternate dataset and 52 for the realistic dataset. Further, a feature-selector tool was utilized for ranking the top key features. It uses a decision tree-based algorithm called the LGBM to rank the feature frequently used to split the data across all trees (Sharma et al., 2021). The obtained top-ranked features were used to build the different machine learning prediction models in all three datasets.

4.2.6 Machine learning techniques

Several machine learning techniques have been used to discriminate toxic from non-toxic proteins. RF, LR, GNB, DT, KNN, XGB, and SVC were implemented to develop the classification models.

4.2.7 Performance evaluation parameters

These classifiers were optimized using various hyper-parameters, and the best results were included. We have also applied a 5-fold CV on 80% of training data for the internal training, testing and model evaluation. The performance of machine learning models was evaluated using the standard evaluation parameters such as sensitivity, specificity, accuracy, MCC and AUC. The complete workflow of ToxinPred2 is depicted in Figure 4.3.

4.2.8 Combined approach

In this study, we have also implemented a combined approach to enhance the prediction of the model. The hybrid approach is the weighted scoring method, in which the score is computed by integrating three different methods (i) similarity-based approach using BLAST, (ii) motif-based approach using MERCI and (iii) ML-based technique. First, the given protein sequence was classified using BLAST at E-value of 10^{-6}. We assigned the weight of '+0.5' for the positive predictions (toxic proteins), '−0.5' for negative predictions (non-toxic proteins) and '0' for no hits. Second, the same protein sequence was classified using MERCI. We assigned the score of '+0.5' if the motifs were found and '0' if the motifs were not found. In the case of a combined approach, scores obtained from three methods (i.e., BLAST, MERCI and ML

scores) were combined to compute the overall score. Based on the overall score at different threshold values, the protein sequence is categorized as toxic and non-toxic. This hybrid approach has been extensively employed in several studies (Sharma et al., 2020), (Gupta, Kapoor, et al., 2013).

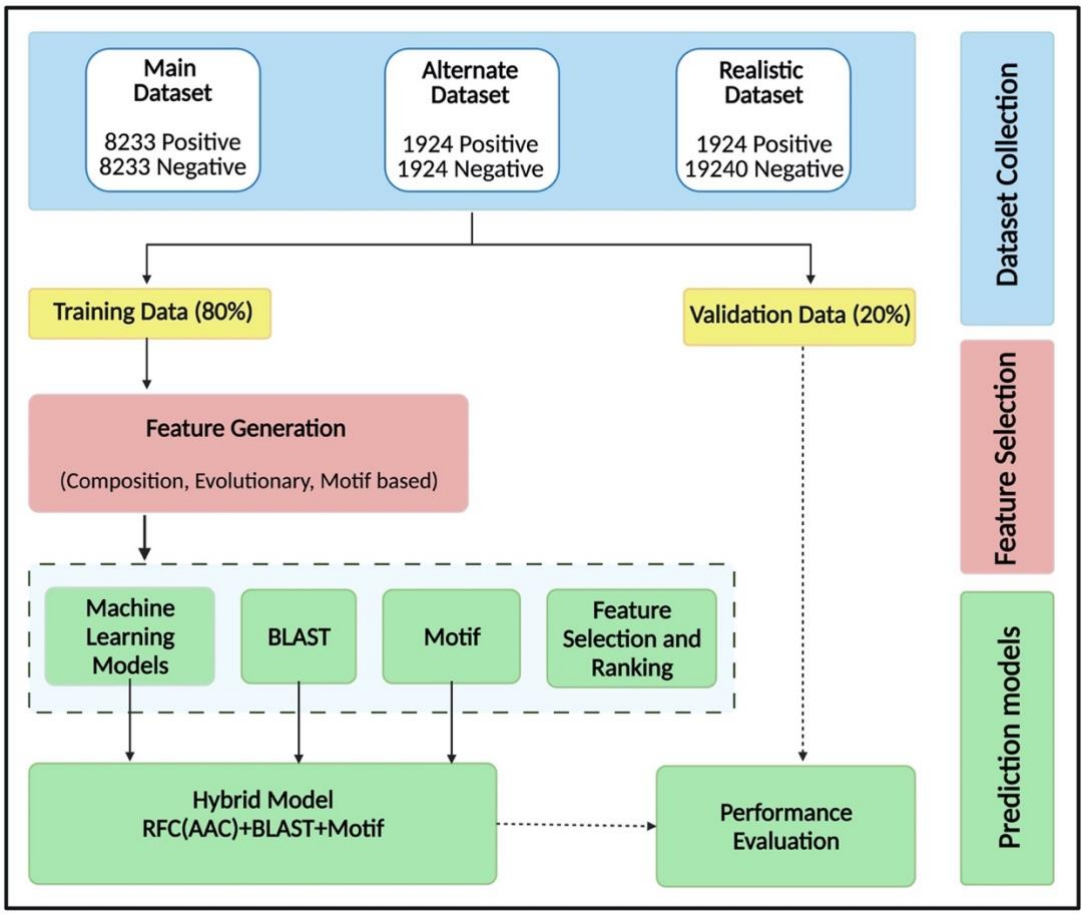

Figure 4.3: Flowchart depicting the overall architecture of ToxinPred2 (10.1093/bib/bbac174)

4.3 Results

4.3.1 Compositional analysis

In the study, AAC for both toxic and non-toxic proteins was computed. We found that the average AAC of amino acid residues such as cysteine, glycine, lysine, and tryptophan are abundant in the toxic sequences, whereas alanine, glutamate, isoleucine, leucine and serine are higher in the non-toxic sequences. Also, we have compared the average AAC between the peptides and proteins of ToxinPred and ToxinPred2, respectively. It was observed that peptides of ToxinPred are exceptionally rich in cysteine and proline. In contrast, the proteins of

ToxinPred2 are rich in lysine and valine. The comparison of average AAC between ToxinPred and ToxinPred2 is depicted in Figure 4.4.

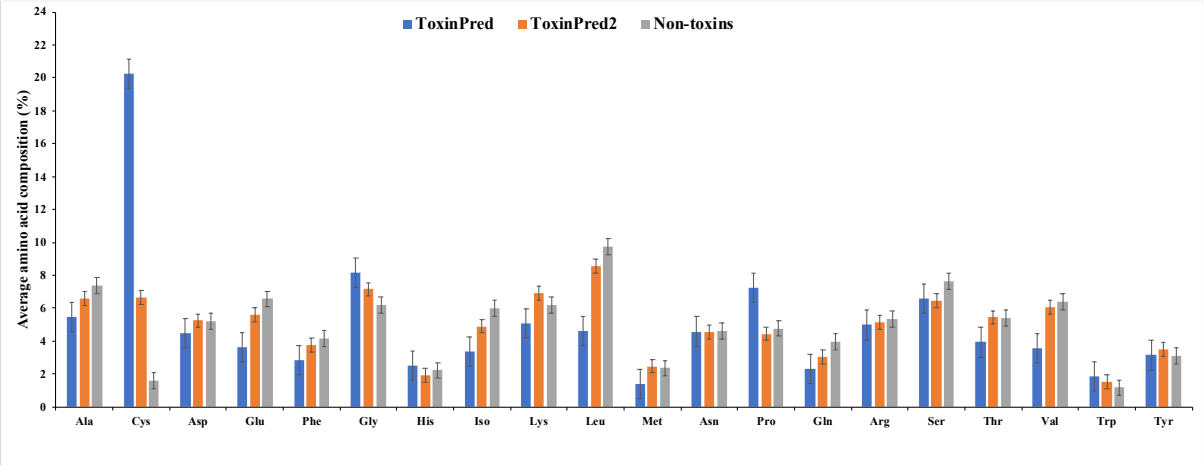

Figure 4.4: Shows amino acid composition of toxic peptides in ToxinPred, toxins in ToxinPred2 and non-toxins (10.1093/bib/bbac174)

It has been shown in the literature that toxic peptides/proteins are rich in cysteine amino-acid. A subunit of pertussis toxin is rich in cysteine residue which forms disulfide bonds and is responsible for binding to the specific cell surface receptors (Burns et al., 1989). Another study by Zhu et al., has shown that if amino acids (cysteine or lysine) are accumulated in high concentrations, the protein may be toxic to the other organisms (Zhu et al., 2004) (Kishor et al., 2020). Eisenhut et al., have stated that higher concentration of glycine is very toxic and can lead to inhibition or even death in bacteria, plants and humans (Eisenhut et al., 2007). Figure 4.4 clearly shows that peptides as well as proteins which are rich in cysteine, glycine and tryptophan amino acid are toxic in nature, and this information can be used as a feature that can be employed for distinguishing non-toxic proteins.

4.3.2 BLAST-based analysis

We have implemented an ensemble of top five hits to reduce the false prediction. For the main dataset, the number of correct hits (sensitivity) increased from 35% to 38.79% for the training dataset and from 36.44% to 40.23% for the validation dataset, with E-value ranging from 10^{-6} to 10^{-1}. This also leads to an increase in the number of wrong hits (error), as shown in Table 4.1.

Table 4.1: The performance of BLAST-based model on main dataset

E-value	Training				Validation			
	Toxins		Non-toxins		Toxins		Non-toxins	
	Correct hits (Sens)	Wrong hits (error)	Correct hits (Spec)	Wrong hits (error)	Correct hits (Sens)	Wrong hits (error)	Correct hits (Spec)	Wrong hits (error)
10^{-6}	4610 (35%)	68 (0.52%)	610 (4.63%)	137 (1.04%)	1201 (36.44%)	16 (0.49%)	182 (5.52%)	42 (1.27%)
10^{-5}	4681 (35.54%)	71 (0.54%)	662 (5.03%)	146 (1.11%)	1224 (37.14%)	17 (0.52%)	198 (6.01%)	45 (1.37%)
10^{-4}	4762 (36.16%)	77 (0.58%)	735 (5.58%)	157 (1.19%)	1251 (37.96%)	19 (0.58%)	212 (6.43%)	51 (1.55%)
10^{-3}	4869 (36.97%)	87 (0.66%)	812 (6.17%)	174 (1.32%)	1271 (38.56%)	24 (0.73%)	248 (7.52%)	55 (1.67%)
10^{-2}	4976 (37.78%)	102 (0.77%)	900 (6.83%)	192 (1.46%)	1293 (39.23%)	28 (0.85%)	271 (8.22%)	61 (1.85%)
10^{-1}	5109 (38.79%)	124 (0.94%)	1034 (7.85%)	214 (1.62%)	1326 (40.23%)	35 (1.06%)	319 (9.68%)	72 (2.18%)

4.3.3 Performance of ML-based models

4.3.3.1 Models using composition

The features including AAC of toxins and non-toxins were computed to develop several machine learning models. For the main dataset, it was observed that RF-based models performed quite well when compared to other models and achieved a maximum AUC of 0.93 and 0.92 on training and validation datasets, respectively. For the alternate dataset, RF-based model attained AUC of 0.76 and 0.75 on training and validation dataset, respectively. In the case of a realistic dataset, it was found that the model based on XGB obtained AUC of 0.75 and 0.74 for training and validation dataset, respectively (Table 4.2).

Table 4.2: The performance of machine learning-based techniques developed using amino acid composition

Main Dataset										
ML	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
DT	77.19	77.90	77.55	0.85	0.55	75.73	80.52	78.13	0.85	0.56
RF	86.59	86.15	86.37	0.93	0.73	86.47	84.10	85.29	0.92	0.71
LR	75.54	75.11	75.32	0.84	0.51	75.61	73.30	74.45	0.84	0.49
XGB	83.55	83.98	83.77	0.91	0.68	82.34	82.77	82.56	0.91	0.65
KNN	82.60	81.72	82.16	0.91	0.64	82.40	82.16	82.28	0.90	0.65

GNB	75.22	74.49	74.85	0.83	0.50	74.76	73.60	74.18	0.82	0.48
SVC	84.39	84.63	84.51	0.92	0.69	83.62	82.16	82.89	0.91	0.66
Alternate Dataset										
DT	62.14	59.61	60.88	0.66	0.22	58.85	55.47	57.16	0.64	0.14
RF	68.64	68.96	68.80	0.76	0.38	64.32	67.97	66.15	0.75	0.32
LR	63.05	62.14	62.60	0.70	0.25	63.02	62.76	62.89	0.69	0.26
XGB	67.53	69.29	68.41	0.75	0.37	66.93	65.89	66.41	0.73	0.33
KNN	65.33	66.17	65.75	0.73	0.32	61.20	63.54	62.37	0.70	0.25
GNB	61.88	61.30	61.59	0.69	0.23	59.64	65.10	62.37	0.69	0.25
SVC	65.02	64.55	64.77	0.72	0.30	63.02	66.41	64.71	0.71	0.29
Realistic Dataset										
DT	60.84	58.97	59.14	0.65	0.12	71.62	47.25	49.46	0.65	0.11
RF	71.04	68.71	68.92	0.78	0.24	67.97	68.06	68.05	0.77	0.22
LR	62.60	63.82	63.71	0.71	0.16	63.54	62.89	62.95	0.69	0.16
XGB	67.14	68.04	67.95	0.75	0.21	68.75	66.50	66.71	0.75	0.21
KNN	67.86	65.72	65.92	0.74	0.21	65.89	62.99	63.26	0.72	0.17
GNB	63.77	60.92	61.17	0.70	0.14	63.28	61.33	61.51	0.70	0.14
SVC	77.86	51.29	53.71	0.74	0.17	83.85	40.88	44.78	0.73	0.15

4.3.3.2 PSSM-based models

The PSSM profiles based on evolutionary information were also generated for protein sequences and used to develop ML-based models. We found that XGB achieved AUC of 0.94 on training and 0.93 on validation for the main dataset. Further, for an alternate dataset, RF-based model attained AUC of 0.80 on training and 0.79 on the validation dataset. For the realistic dataset, XGB performed better and obtained the maximum AUC of 0.80 on training and 0.79 on validation (Table 4.3).

Table 4.3: The performance of machine learning-based techniques developed using PSSM profiles

Main Dataset										
ML	**Training**					**Validation**				
	Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	75.41	75.85	75.63	0.82	0.51	75.00	75.12	75.06	0.82	0.50
RF	86.52	86.65	86.58	0.94	0.73	85.19	85.80	85.50	0.93	0.71
LR	82.17	81.85	82.01	0.90	0.64	81.13	80.58	80.86	0.90	0.62
XGB	86.96	86.62	86.79	0.94	0.74	85.98	86.17	86.07	0.93	0.72
KNN	84.18	84.19	84.18	0.92	0.68	82.65	83.92	83.28	0.92	0.67

GNB	73.06	74.82	73.94	0.82	0.48	72.03	74.03	73.03	0.81	0.46
SVC	79.73	79.20	79.46	0.88	0.59	78.58	80.22	79.40	0.87	0.59
Alternate Dataset										
DT	61.88	64.35	63.12	0.68	0.26	56.77	63.28	60.03	0.66	0.20
RF	71.69	71.95	71.82	0.80	0.44	72.14	70.83	71.48	0.79	0.43
LR	68.64	68.83	68.73	0.76	0.38	69.27	70.83	70.05	0.76	0.40
XGB	69.03	70.00	69.51	0.78	0.39	68.75	69.79	69.27	0.77	0.39
KNN	69.48	70.13	69.81	0.76	0.40	67.97	70.05	69.01	0.76	0.38
GNB	62.27	62.34	62.31	0.65	0.25	62.50	65.37	63.93	0.66	0.28
SVC	65.91	66.43	66.17	0.72	0.32	66.93	69.53	68.23	0.73	0.37
Realistic Dataset										
DT	63.96	62.01	62.18	0.68	0.15	57.81	63.36	62.85	0.64	0.13
RF	71.23	75.36	74.99	0.81	0.30	69.01	76.51	75.83	0.81	0.29
LR	68.90	69.04	69.02	0.77	0.23	69.27	69.80	69.75	0.77	0.24
XGB	71.88	70.42	70.54	0.80	0.26	71.62	71.23	71.27	0.79	0.26
KNN	71.04	68.23	68.49	0.77	0.24	72.92	65.93	66.56	0.77	0.23
GNB	61.95	61.95	61.95	0.65	0.14	62.50	61.95	62.00	0.64	0.14
SVC	67.21	64.66	64.90	0.74	0.19	63.54	65.39	65.22	0.72	0.17

4.3.3.3 Selected features

The reduced features 129 (main dataset), 32 (alternate dataset) and 52 (realistic dataset) were used to develop different classification models on all the three datasets. The performance of these models is illustrated in Table 4.4. It is clearly shown in the table that RF-based model performed better in all three datasets.

Table 4.4: The performance of machine learning-based techniques developed using selected features

Main Dataset										
ML	**Training**					**Validation**				
	Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	76.55	76.46	76.51	0.84	0.53	75.85	75.61	75.73	0.82	0.52
RF	86.27	86.29	86.28	0.94	0.73	85.50	86.35	85.92	0.93	0.72
LR	80.64	80.50	80.57	0.89	0.61	79.19	82.16	80.67	0.88	0.61
XGB	86.21	85.83	86.02	0.93	0.72	84.41	86.77	85.59	0.93	0.71
KNN	78.53	79.04	78.79	0.86	0.58	76.88	79.43	78.16	0.86	0.56
GNB	67.46	85.95	76.71	0.82	0.54	66.14	85.68	75.91	0.81	0.53
Alternate Dataset										

DT	63.25	61.30	62.27	0.68	0.25	67.19	48.96	58.07	0.64	0.16
RF	70.71	70.58	70.65	0.79	0.41	70.31	69.79	70.05	0.77	0.40
LR	66.88	66.36	66.62	0.74	0.33	66.67	66.93	66.80	0.73	0.34
XGB	68.70	67.53	68.12	0.76	0.36	68.75	71.62	70.18	0.76	0.40
KNN	65.97	65.65	65.81	0.70	0.32	62.51	65.10	63.80	0.70	0.28
GNB	64.22	64.29	64.25	0.71	0.29	64.84	64.84	64.84	0.71	0.30
Realistic Dataset										
DT	63.31	60.89	61.11	0.69	0.14	54.17	66.29	65.19	0.66	0.12
RF	73.44	68.98	69.39	0.80	0.26	71.35	68.66	68.90	0.79	0.24
LR	67.01	67.80	67.72	0.75	0.21	64.58	66.79	66.59	0.74	0.19
XGB	66.75	71.12	70.72	0.77	0.23	67.19	70.66	70.35	0.77	0.23
KNN	63.64	66.89	66.59	0.70	0.18	57.81	65.41	64.72	0.67	0.14
GNB	50.07	82.69	79.72	0.71	0.23	47.92	81.91	78.83	0.70	0.21

4.3.3.4 Motif-based models

The motifs such as 'GCYCG, MKTLL, TLLLTL and LLLTLV' are solely found in toxic proteins. Composition-based models (AAC) built using different ML techniques were integrated with the MERCI approach. The performance of the combined approach (ML+MERCI) for all three datasets is shown in Table 4.5.

Table 4.5: The performance of motif-based approach when combined with machine learning techniques

Main Dataset										
ML	**Training**					**Validation**				
	Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	75.46	80.90	78.18	0.85	0.56	74.15	83.13	78.64	0.85	0.58
GNB	72.12	79.56	75.84	0.83	0.52	72.76	79.37	76.06	0.82	0.52
KNN	81.72	83.40	82.56	0.91	0.65	81.74	84.16	82.95	0.90	0.66
LR	74.02	78.76	76.39	0.84	0.53	74.21	77.00	75.61	0.84	0.51
RF	84.60	89.36	86.98	0.94	0.74	84.59	87.93	86.26	0.93	0.73
SVC	85.24	82.20	83.72	0.92	0.68	84.59	80.40	82.49	0.91	0.65
XGB	83.28	84.53	83.90	0.92	0.68	81.80	83.50	82.65	0.91	0.65
Alternate Dataset										
DT	69.55	52.53	61.04	0.69	0.22	66.15	51.30	58.72	0.67	0.18
GNB	60.91	66.49	63.70	0.72	0.27	60.16	69.01	64.58	0.72	0.29
KNN	58.57	75.39	66.98	0.74	0.35	55.21	73.70	64.45	0.72	0.29

LR	57.08	71.69	64.38	0.72	0.29	57.55	74.48	66.02	0.72	0.33
RF	64.55	76.49	70.52	0.78	0.41	59.90	76.82	68.36	0.77	0.37
SVC	57.86	75.58	66.72	0.75	0.34	56.25	80.21	68.23	0.74	0.38
XGB	62.08	78.18	70.13	0.77	0.41	61.46	73.96	67.71	0.76	0.36
Realistic Dataset										
DT	66.88	52.03	53.38	0.67	0.11	72.66	47.38	49.67	0.67	0.12
GNB	47.53	84.65	81.27	0.71	0.24	44.53	84.17	80.58	0.71	0.21
KNN	56.10	79.54	77.41	0.75	0.24	55.21	76.53	74.60	0.73	0.21
LR	55.26	73.40	71.75	0.72	0.18	55.21	72.30	70.75	0.71	0.17
RF	60.65	81.63	79.72	0.79	0.29	56.51	82.10	79.77	0.77	0.27
SVC	35.52	95.86	90.37	0.75	0.35	34.12	95.92	90.31	0.74	0.34
XGB	55.20	82.56	80.07	0.76	0.27	54.69	81.94	79.47	0.76	0.26

4.3.3.5 BLAST-based models

To build an enhanced method, the similarity search approach BLAST and ML-based models were synergized. The BLAST search was initially implemented for a query sequence; if a BLAST hit was obtained, the query sequence was assigned as toxin and non-toxin based on the BLAST result. If there is no hit obtained, then the composition-based model is utilized to predict the same sequence. Table 4.6 shows the performance of BLAST when combined with machine learning techniques.

Table 4.6: The performance of BLAST when combined with machine learning techniques

Main Dataset										
ML	**Training**					**Validation**				
	Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	88.63	91.77	90.20	0.97	0.81	88.53	90.59	89.56	0.97	0.79
GNB	94.08	91.19	92.63	0.96	0.85	94.91	91.57	93.23	0.96	0.87
KNN	93.53	95.22	94.37	0.98	0.89	94.05	95.57	94.81	0.98	0.91
LR	93.65	92.44	93.04	0.98	0.86	94.72	92.61	93.66	0.98	0.87
RF	93.85	96.14	95.02	0.98	0.91	94.36	95.75	95.05	0.99	0.92
SVC	91.69	91.47	91.58	0.98	0.83	91.14	90.78	90.96	0.98	0.82
XGB	94.73	93.35	94.04	0.98	0.88	95.02	93.14	94.08	0.98	0.88
Alternate Dataset										
DT	73.90	77.92	75.91	0.83	0.52	72.41	77.61	75.01	0.83	0.51
GNB	74.87	78.83	76.85	0.84	0.54	74.74	80.21	77.47	0.85	0.55
KNN	73.70	82.53	78.12	0.88	0.56	72.41	81.25	76.82	0.88	0.54
LR	75.91	77.21	76.56	0.87	0.53	75.01	78.91	76.95	0.87	0.54
RF	79.61	80.45	80.03	0.89	0.61	75.78	82.03	78.91	0.89	0.58

SVC	76.75	79.42	78.08	0.88	0.56	75.01	82.81	78.91	0.88	0.58
XGB	78.57	82.27	80.42	0.89	0.61	76.31	80.73	78.52	0.89	0.57
Realistic Dataset										
DT	83.25	77.20	77.75	0.91	0.39	87.24	78.31	79.11	0.92	0.42
GNB	78.44	89.22	88.24	0.92	0.52	77.60	89.5	88.42	0.93	0.52
KNN	77.08	90.81	89.56	0.94	0.54	78.65	90.67	89.58	0.95	0.55
LR	81.11	86.64	86.14	0.94	0.49	82.03	87.32	86.84	0.94	0.51
RF	83.12	90.11	89.47	0.95	0.57	81.77	91.19	90.34	0.96	0.58
SVC	69.87	97.08	94.61	0.94	0.67	71.88	97.14	94.85	0.95	0.69
XGB	80.39	90.61	89.68	0.94	0.56	80.47	91.40	90.41	0.95	0.58

4.3.3.6 Models using combined approach

Ultimately, multiple approaches were integrated in order to overcome the limitations of individual methods. These approaches were developed to detect the toxins with better precision. A composition-based model is combined with BLAST- and MERCI-based approaches. Initially, proteins were classified using ensemble BLAST at E-value of 10^{-6}, led by MERCI approach. The ML-based model then predicts the protein sequences not predicted by these two approaches. The combined approach significantly enhanced the coverage and accuracy, which is not feasible by using all these methods individually. The performance of the combined approach has improved by integrating all these methods, as shown in Table 4.7. RF-based model performs best for all three datasets on training and validation datasets. It achieved AUC of 0.98 and 0.99 (main dataset), AUC of 0.90 and 0.90 (alternate dataset) and AUC of 0.95 and 0.96 (realistic dataset) on training and validation dataset.

Table 4.7: The performance of combined method integrating machine learning, BLAST and MERCI techniques

Main Dataset										
ML	**Training**					**Validation**				
	Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
DT	88.64	91.81	90.23	0.97	0.80	88.59	90.59	89.59	0.97	0.79
GNB	94.08	91.19	92.63	0.96	0.85	94.91	91.57	93.23	0.96	0.87
KNN	93.53	95.22	94.37	0.98	0.89	94.05	95.57	94.81	0.98	0.90
LR	92.26	95.99	94.12	0.98	0.88	93.33	95.81	94.57	0.98	0.89
RF	92.95	97.65	95.31	0.98	0.91	93.69	97.39	95.54	0.99	0.91
SVC	91.69	91.63	91.66	0.98	0.83	91.14	90.96	91.05	0.98	0.82
XGB	94.73	93.35	94.04	0.98	0.88	95.02	93.14	94.08	0.98	0.88
Alternate Dataset										
DT	74.42	78.31	76.36	0.84	0.53	73.71	78.12	75.91	0.85	0.52

GNB	75.39	79.16	77.27	0.85	0.55	76.04	80.73	78.39	0.86	0.57
KNN	73.96	82.66	78.31	0.88	0.57	73.96	81.51	77.73	0.89	0.56
LR	76.11	77.34	76.72	0.87	0.53	76.04	79.43	77.73	0.88	0.56
RF	79.94	80.52	80.23	0.90	0.61	76.82	82.55	79.69	0.90	0.59
SVC	77.08	79.55	78.31	0.89	0.57	76.30	83.59	79.95	0.89	0.61
XGB	78.90	82.53	80.71	0.89	0.61	77.08	81.25	79.17	0.90	0.58
Realistic Dataset										
DT	83.25	77.44	77.97	0.91	0.39	87.24	78.56	79.35	0.92	0.42
GNB	78.44	89.24	88.26	0.93	0.52	77.60	89.51	88.42	0.93	0.52
KNN	77.08	90.92	89.66	0.94	0.54	78.65	90.77	89.67	0.95	0.55
LR	81.11	86.73	86.22	0.94	0.49	82.03	87.37	86.89	0.94	0.51
RF	83.12	90.16	89.52	0.95	0.57	81.77	91.22	90.36	0.96	0.58
SVC	69.94	97.11	94.63	0.95	0.67	71.88	97.17	94.87	0.95	0.69
XGB	80.45	90.68	89.75	0.94	0.56	80.47	91.42	90.43	0.95	0.58

4.3.3.7 Best models of the study

The performance of the best classification models developed using the different features is listed in Table 4.8 .

Table 4.8: The list of the features used in the study with the best performing ML models

Method	**Features**	**ML**	**Training**					**Validation**				
			Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
Composition based features	AAC	RF	86.59	86.15	86.37	0.93	0.73	86.47	84.11	85.29	0.92	0.71
PSSM based features	PSSM profiles	XGB	86.96	86.62	86.79	0.94	0.74	85.98	86.17	86.07	0.93	0.72
Feature Selection	132 features	RF	86.27	86.29	86.28	0.94	0.73	85.51	86.35	85.92	0.93	0.72
Motif based approach	Motifs + AAC	RF	84.61	89.36	86.98	0.94	0.74	84.59	87.93	86.26	0.93	0.73
BLAST based approach	BLAST + AAC	RF	93.85	96.14	95.02	0.98	0.91	94.36	95.75	95.05	0.99	0.92
Hybrid approach	BLAST + Motifs + AAC	RF	92.95	97.65	95.31	0.98	0.91	93.69	97.39	95.54	0.99	0.91

4.4 Web server and standalone software package

A web server, ToxinPred2 (https://webs.iiitd.edu.in/raghava/toxinpred2/), has been developed for predicting toxic proteins. We have executed our two best performing models, i.e., Model-1 (AAC-based RF approach) and Model-2 (hybrid approach). Both the models are trained on

the main dataset for predicting toxins. The major modules such as (i) prediction, (ii) design, (iii) motif scan, (iv) BLAST search and (v) Download are integrated into the web server. The 'prediction module' permits the user to submit the single and multiple protein sequences in FASTA format. This module can efficiently classify toxic and non-toxic proteins. The 'design module' is developed for generating all possible mutants of a protein by mutating a single residue at a time. The developed models were then used for predicting each mutated sequence as toxic or non-toxic protein. Using our webserver, user can design the non-toxic proteins and which can be used for further studies. The 'motif scan module' uses MERCI software to identify the motifs exclusively present in the toxic protein sequences. It also maps or scans the motifs in the query protein sequence given by the user and distinguishes them as toxin and non-toxin. The 'BLAST search module' aids the user to carry out a similarity-based search using BLAST against toxins and non-toxins database. The web server is built with a responsive HTML template and browser compatibility for various operating systems. To facilitate the users to predict toxins, we have also developed a python based standalone package of ToxinPred2, which can be accessed from 'Download' module of the web server.

4.5 Comparison with other methods

It is important to compare the performance of the proposed method with existing methods to justify the development of the new method. We have shown the comparison of the performance of a proposed method, ToxinPred2 and other existing methods as reported in the literature in Table 4.8. ToxinPred2 outperforms other existing methods under the heading 'Performance of existing methods reported in the literature'. In order to provide an unbiased comparison, we have computed the performance of ToxinPred2 on the validation dataset used in existing methods. As shown in Table 4.9, under the heading 'Performance of ToxinPred2 on validation dataset of existing methods', ToxinPred2 obtained AUC 0.96, 0.99 and 0.99 for protein datasets used in ToxClassifer, TOXIFY, and ToxDL, respectively. Our proposed method achieves AUC 0.94 on peptide datasets used in ToxinPred and ATSE, which are lower than their original performance. It is because ToxinPred2 is developed/trained for proteins not for peptides. We also attempted to evaluate the performance of existing methods on validation data (main dataset) of ToxinPred2. Since the dataset has larger protein sequences, the peptide toxicity prediction methods (ToxinPred, ToxIBLT and ATSE) cannot be implemented. Moreover, we could not predict the toxicity of proteins in the validation dataset of ToxinPred2 using ToxDL and ToxClassifier due to the limitations of their web services (one is non-functional, and

another allows a maximum of 10 sequences per submission for prediction). We used a standalone version of TOXIFY to predict the toxicity of proteins in validation dataset of ToxinPred2. It has predicted the toxicity of only 2617 out of 3296 protein sequences, which have length up to 500 amino acids. As shown in Table 4.9, under the heading 'Performance of existing methods on validation dataset of ToxinPred2', it achieved AUC 0.88, which is lower than the performance reported by ToxinPred2 on the same dataset. This comparison demonstrates the importance of the newly proposed method in the field of toxicity prediction.

Table 4.9: Comparison of proposed method ToxinPred2 with existing methods

Method	Type of Dataset used	Sensitivity	Specificity	Accuracy	AUC	MCC
ToxinPred2	All types of toxins (Proteins)	93.69	97.39	95.54	0.99	0.91
Performance of existing methods reported in the literature						
ToxinPred	All types of toxins (Peptides)	93.80	94.85	94.50	0.98	0.88
ToxClassifier	Animal venom toxins (Proteins)	96.70	99.80	99.70	NA	0.89
TOXIFY	Animal venom toxins (Proteins)	96.00	76.00	86.00	NA	0.74
ToxDL	Animal toxins (Proteins)	NA	NA	NA	0.98	0.79
ATSE	Toxic peptides	96.50	94.00	95.20	0.97	0.90
Performance of ToxinPred2 on validation dataset of existing methods						
ToxinPred	Toxic peptides	97.73	45.73	63.12	0.94	0.44
ToxClassifier	Toxic proteins	97.15	77.54	87.38	0.96	0.76
TOXIFY	Toxic proteins	96.48	92.71	94.59	0.99	0.89
ToxDL	Toxic proteins	100	88.81	89.74	0.99	0.63
ATSE	Toxic peptides	96.65	58.21	76.03	0.94	0.58
Performance of existing methods on validation dataset of ToxinPred2						
TOXIFY	Toxic proteins	68.94	97.94	81.85	0.88	0.68

4.6 Discussion & conclusion

One of the major challenges in the field of protein/peptide-based therapeutics is to identify toxic regions in a protein. There is a dire need to determine the toxic potential of newly synthesized proteins. Experimental techniques for determining toxicity proteins are costly and time-consuming. Thus, there is a need to develop computer-aided techniques for predicting the toxicity of proteins/peptides with high precision. In order to facilitate the scientific community, our group developed a method, ToxinPred, for predicting and designing toxic peptides. It is heavily used by the scientific community in the field of therapeutic peptides. This tool has been developed mainly for peptides as models have been trained on peptides having length up to 35 amino acids. In order to complement ToxinPred, we proposed a new method, ToxinPred2, for predicting the toxicity of proteins. In the present study, three datasets were created: main,

alternate and realistic datasets curated from Swiss-Prot. The main dataset consists of 8233 toxic and non-toxic proteins, alternate dataset contains 1924 non-redundant toxic and non-toxic proteins. Realistic dataset was generated to create realistic conditions in which negative data is multiple folds than positive data. Thus, 1924 toxic and 19240 non-toxic proteins were used in realistic dataset.

Various features for the protein sequences were computed using Pfeature tool. The relevant features were further selected and ranked using the SVC-L1 and feature-selector tool, respectively. Our compositional analysis exhibited that cysteine, glycine, lysine and tryptophan are dominant in toxic proteins in comparison to non-toxic proteins. It is noteworthy that the composition-based features are among the top selected features. This suggests that these features can be used to distinguish between toxic and non-toxic proteins. Furthermore, we have implemented the BLAST, a widely used tool to annotate any query protein sequence. If the query protein sequence shows high similarity with a known protein function, it designates the same function to the query protein. As shown in Table 4.1, BLAST has correctly identified some toxins with a probability of correct prediction of more than 40%, with a very low error rate. Thereby, it can be inferred that BLAST is generating a large number of no hits; hence it fails when the unknown protein has no similarity with toxins and non-toxins. Combined model was developed using ML models (AAC), BLAST and MERCI to overcome this limitation. We have achieved the highest performance with balanced sensitivity and specificity and higher accuracy, as shown in Table 4.7.

In the present study, we have provided a comprehensive platform where users can classify toxic and non-toxic proteins/peptides. To facilitate the scientific community and promote widespread usage of the proposed prediction method, we have provided a freely accessible web server and a standalone package of ToxinPred2. In the web server, we have incorporated the best performing model for correctly predicting the toxins and non-toxins. However, one of the limitations of our method is that it can classify toxins and non-toxins regardless of their source of origin. We hope that the researchers will extensively use our prediction method for designing improved and accurate protein/peptide-based therapeutics against various diseases.

4.7 Limitation of the study

In this study, we have made an attempt to develop the user friendly method that can be used to classify the toxic and non-toxic proteins. However, it is general method it can classify toxins and non-toxins regardless of their source of origin.

Chapter 5

Prediction of allergens and designing of non-allergens

5.1 Introduction

Allergy is the abnormal behaviour of the immune system against foreign substances called allergens. It involves a series of many reactions, which trigger various symptoms like allergic asthma, rhinitis, skin reactions, and difficulty in breathing that can lead to death. The rise in the occurrence of allergic diseases in the last few years has not only enhanced the costs of treatment but also adversely affected the quality of life of a large population (Obermeyer & Ferreira, 2005). Allergens like dust mites, pollens, and many others induce Type I hypersensitive reactions, which elicit IgE antibodies. This allergic reaction results in the release of inflammatory mediators, such as histamine, cytokines from mast cells and basophils (Masoli et al., 2004), which affects the population at large scale, particularly skin sensitization (Sutton & Gould, 1993), (Broadfield et al., 2002).

Sensitization is the first encounter of allergen, which develops the hypersensitivity, while the second encounter of the same allergen leads to the effector response. Type I hypersensitivity is mediated by immunoglobulin E (IgE), which is produced to act against allergens. Allergens induce the Type I hypersensitivity reaction, which sets off the production of allergen-specific IgE epitopes. These epitopes bind to the mast cell and basophils, known as the sensitization of mast cells and basophils. Re-exposure of the allergens to sensitized mast cells and basophils (which are already coated with IgE antibodies) leads to the degranulation and release of mediators and inflammatory molecules like histamine, leukotriene, etc., which leads from a mild allergic reaction to sudden death from anaphylactic shock (Mak TW, 2014). Overall processing of allergen, activation of IgE antibodies and release of histamine is shown in Figure 5.1. The guideline issued by Food and Agriculture Organization fails to identify allergens with high precision due to a large number of false-positive predictions (FAO/WHO, 2001) (FAO/WHO, 2003). Earlier methods developed before 2005 can be classified in the following categories; i) similarity search, ii) supervised learning-based models, and iii) motif-based approaches.

In 2006, a hybrid method AlgPred (Saha & Raghava, 2006a), was developed that combines the following approaches for predicting allergenic proteins; i) SVM-based model, ii) mapping of IgE epitopes, iii) MEME/MAST motifs (Bailey & Elkan, 1994), (Bailey & Gribskov, 1998), and iv) BLAST-based similarity search (Camacho et al., 2009). This method combines the power of different approaches, and it outperformed all methods developed before 2006. Following is a brief description of methods developed in the last 14 years. AllerTool is an SVM-based method developed in 2007; it combines a similarity-based approach for predicting

allergenicity and allergic cross-reactivity in proteins (Zhang et al., 2007). AllerHunter was developed in 2009 on 1356 allergenic proteins, where models were developed using SVM-pairwise sequence similarity (Muh et al., 2009). In 2013, AllerTOP was developed on 2210 allergens, and its updated version, AllerTOPv2, has been developed on 2427 allergens (Dimitrov et al., 2013), (Dimitrov, Bangov, et al., 2014). In the case of PREAL, an SVM-based model was developed on the 1176 allergenic protein using biochemical and physicochemical properties (Wang et al., 2013). In 2014, AllergenFP was developed on a dataset of 2427 allergens that incorporates descriptor-based fingerprints for developing prediction models (Dimitrov, Naneva, et al., 2014). Recently, AllerCatPro has been developed on 4180 allergens for predicting the allergenicity potential of a protein from its sequence and 3D epitope mapping (Maurer-Stroh et al., 2019).

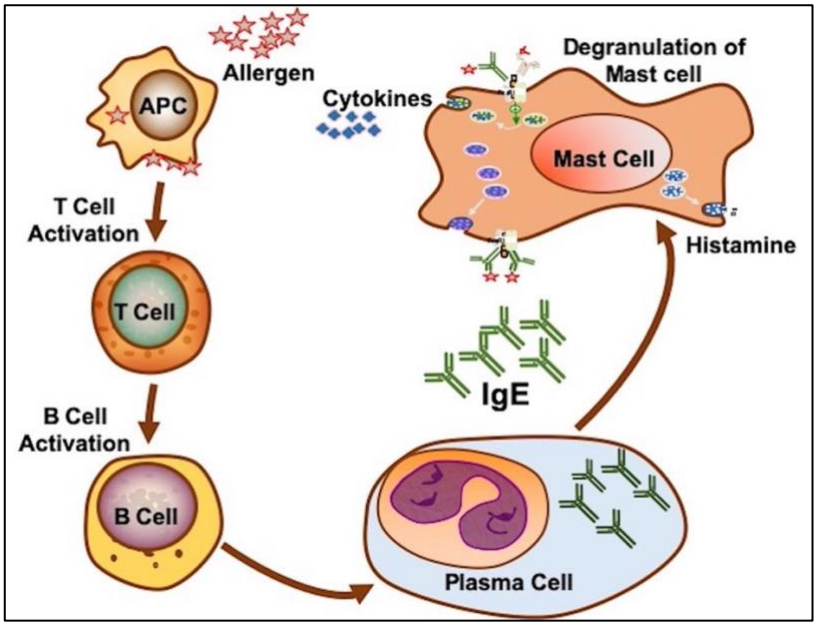

Figure 5.1: Mechanism showing processing of allergen, activation of IgE antibodies and release of histamine (10.1093/bib/bbaa294)

These allergen prediction methods are heavily used by the scientific community, particularly by experimental researchers in designing proteins with desired allergenicity. These methods have their limitations that include; i) most of the methods have been developed on small datasets, ii) redundant proteins in the dataset, iii) no mapping of IgE epitopes, iv) motif information not incorporated. In order to complement existing methods in serving the research community, in this study, we made a systematic attempt to improve our method AlgPred (Saha & Raghava, 2006a).

In this study, we proposed the prediction method to classify the allergenic and non-allergenic protein sequences, AlgPred 2.0, an improved version of AlgPred developed in 2006. AlgPred combined four different methods to predict the allergens. In contrast, in AlgPred 2.0, we have incorporated a number of new features to improve the performance of the method using the state of art techniques. In addition to the large dataset used for training our models, we have incorporated the following features. AlgPred 2.0 allows users to map 10451 experimentally validated IgE epitopes on a protein sequence instead of 178 epitopes in AlgPred. In this study, we have also used evolutionary information for building *in-silico* models where the evolutionary information was derived from PSSM profiles generated using PSI-BLAST. We also introduced a new approach for using BLAST called ensemble of BLAST hits, where the prediction of allergens is based on the top five hits. In the previous method, AlgPred, we had used only SVM using AAC, whereas, in AlgPred 2.0, we used several machine learning techniques, including RF, SVM, KNN, MLP, and DT. In order to map IgE epitopes on proteins, we used BLAST and MERCI software in AlgPred 2.0. In summary, AlgPred 2.0 is a hybrid method that combines most of the existing approaches to identify allergens with high accuracy.

5.2 Materials & methods

5.2.1 Compilation of dataset

The dataset used in this study was compiled from various databases and repositories, namely, COMPARE (2018 allergens) (https://comparedatabase.org), Allergen Online (2078 allergens) (Goodman et al., 2016), AlgPred (Saha & Raghava, 2006a), AllerTop (2427 allergens and 2427 non-allergens) (Dimitrov et al., 2013) and Swiss-Prot (1078 allergens with the query 'allergen AND reviewed:yes' (UniProt, 2021). All proteins containing non-standard characters (i.e. BJOUXZ) or less than 50 amino acids or non-allergen sequences similar to allergen sequences were removed. Finally, we got 10 075 allergen sequences, which we called a positive dataset. For obtaining a negative dataset, we extracted 545 820 proteins using the query 'NOT allergen NOT cancer NOT allergenic AND reviewed: yes'; these proteins were assigned as non-allergens. We got 533 719 non-allergenic sequences after removing sequences having less than 50 amino acids and containing the non-standard characters. We randomly pick up 10 075 non-allergenic sequences from 533 719 non-allergenic sequences. Finally, we got a dataset that contains 10 075 allergenic and 10 075 non-allergenic sequences.

5.2.2 Creation of non-redundant dataset

One should remove the redundancy among proteins in a dataset to develop a robust method. In the past, researchers have created non-redundant datasets at different levels of similarity from 30% to 100%. One of the major reasons to create a non-redundant dataset is to remove similar sequences among proteins in training and testing datasets. Unfortunately, the removal of redundant sequences also reduces the size of the dataset. In a previous study, we introduced the concept of data partitioning to create non-redundant training and testing datasets without reducing the size of the dataset (Saha & Raghava, 2006a). In this study, we also used the same approach to partition data in training and testing datasets, where no protein in the training set had more than 40% similarity with any protein in the test dataset. First, clusters were created using CD-HIT (Li & Godzik, 2006) software at a 40% sequence similarity for the positive dataset (allergens) as well as for the negative dataset (non-allergens). Second, clusters obtained for both allergens and non-allergens datasets were divided into 80% training data and 20% validation data. The clusters in training data (both for positive and negative data) were further fractionated into five sets such that all proteins of a given cluster are kept in one set, and sequences in one set do not have any similarity with sequences of other sets. It results in five positive sets and five negative sets. The 20% validation set also consists of positive and negative clusters that are not present in the training data.

5.2.3 Dataset of IgE epitopes

The epitopes were obtained from various sources that include IEDB (15 046) (Vita et al., 2019), AllerBase (863) (Kadam et al., 2017), and IgPred (2341) (Gupta, Ansari, et al., 2013). The non-IgE epitopes were obtained from IEDB (381 396) and IgPred (35 219). Finally, we got 10 451 IgE epitopes and 307 866 non-IgE epitopes after removing redundant epitopes and epitopes having less than five and more than 50 amino acids.

5.2.4 BLAST for similarity search

The similarity-based search module was developed using the blastp suite of BLAST+ version 2.7.1 (Camacho et al., 2009), where the query sequences were hit against the database of allergens and non-allergens. This study used two strategies to identify allergens: (i) top hit of BLAST and (ii) ensemble of top five hits of BLAST.

5.2.5 Motif scanning

The motif is a recurring pattern of amino acids or nucleotides occurring in protein or DNA. In this study, we used MEME/MAST and MERCI software to find out the motifs from experimentally validated IgE epitopes (Bailey & Elkan, 1994), (Vens et al., 2011).

5.2.6 Protein features

Residue information of the protein was used in the form of AAC and DPC for developing ML models. The web server 'Pfeature' was used for this purpose (Pande A, 2019). Evolutionary information of the protein in the form of PSSM profiles was also computed.

5.2.7 Machine learning models

Different classification models such as RF, SVM, DT, KNN and MLP were implemented using sklearn package from python. GridSearchCV was used for the optimization of hyper-parameters. Protein features such as AAC and PSSM-400 were used as fixed-length vectors for training and testing models. A 5-fold CV was used to evaluate the models using different performance measures. The complete architecture of AlgPred 2.0 is shown in Figure 5.2.

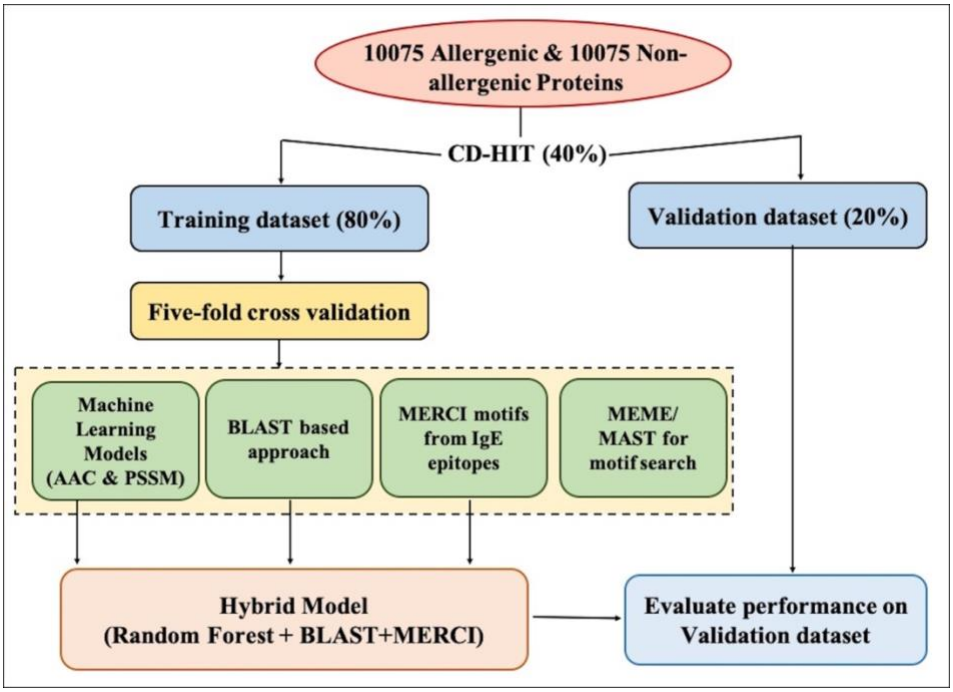

Figure 5.2: Flowchart shows the overall architecture of AlgPred 2.0 (10.1093/bib/bbaa294)

5.2.8 Evaluation of performance

A 5-fold CV was used to evaluate the performance of the prediction models developed in this study. A training set was formed by combining four negative and four positive sets, while the corresponding test set was formed by combining the remaining positive and negative sets. This method is repeated five times to ensure that the combination of a positive set and a negative set is used as a test set only once. These five training and testing sets were used for developing learning-based prediction models. Models were then evaluated by performing predictions on the unseen validation set. Threshold independent parameters such as sensitivity, specificity, accuracy, and MCC; and dependent parameter viz. AUC value, was used to evaluate the performance of the models.

5.2.9 Hybrid approach for classification

In order to improve the accuracy of classifying allergenic and non-allergenic proteins, we implemented a hybrid approach as used in AlgPred and other state-of-the-art methods. In AlgPred 2.0, we have also applied a hybrid approach. Here, the following three techniques have been combined: (i) similarity-based approach using BLAST, (ii) motif-based approach using MERCI and (iii) ML-based technique. First, the given protein sequence was classified using BLAST at E-value of 10^{-6}. We assigned the score of '+0.5' for the correct positive predictions (allergenic proteins), '-0.5' for correct negative predictions (non-allergenic proteins) and '0' for no hits. Second, the same protein sequence was classified using MERCI. We assigned the score of '+0.5' if the motifs were found and '0' if the motifs were not found. In the case of a hybrid approach, scores obtained from three methods (i.e., BLAST, MERCI and ML scores) were combined to compute the overall score. This overall score of the hybrid approach was used for assigning the protein as allergenic and non-allergenic protein at different thresholds.

5.3 Results

5.3.1 Prediction based on similarity

BLAST is the widely used software for similarity search. Hence, we have implemented the BLAST to segregate the allergens and non-allergens. We have implemented a 5-fold CV to avoid the biasness, in which the proteins in the one set, i.e., test set, were searched against the five datasets using the BLAST at various E-value cut-offs. This process is repeated five times

so that each set gets the chance to be the test set once to cover all the proteins in the dataset. As exhibited in Table 5.1, the number of correctly predicted allergens (sensitivity) increased from 57.3% to 63.68% and 43.57% to 47.00% for the training dataset and validation dataset, respectively, with E-value from 10^{-6} to 10^{-1}. The sensitivity is directly proportional to the error (% of non-allergens) with the increase of E-values. A similar trend is followed by the non-allergens; the specificity varied from 11.32% to 19.26% and 13.10% to 22.28% in training and validation dataset, respectively, with an increase in the error from 1.33% to 1.89% and 1.19% and 1.74%, for E-value from 10^{-6} to 10^{-1}. The maximum overall accuracies achieved using BLAST method are 41.47% and 34.64% for training and validation dataset, respectively, due to the significant number of no hits. The overall performance of BLAST is too poor due to a large number of no hits; it means BLAST alone cannot be used for predicting allergenic proteins, as represented in Table 5.1.

Table 5.1: Shows the results of similarity-based search developed using top five hits of BLAST

E-value	Training				Validation			
	Allergens		Non-allergens		Allergens		Non-allergens	
	Correct hits (Sens)	Wrong hits (Error)	Correct hits (Spec)	Wrong hits (Error)	Correct hits (Sens)	Wrong hits (Error)	Correct hits (Spec)	Wrong hits Error
10^{-6}	4618 (57.3%)	164 (2.03%)	912 (11.32%)	107 (1.33%)	878 (43.57%)	48 (2.38%)	264 (13.1%)	24 (1.19%)
10^{-5}	4665 (57.88%)	174 (2.16%)	1001 (12.42%)	111 (1.38%)	883 (43.82%)	48 (2.38%)	284 (14.09%)	24 (1.19%)
10^{-4}	4772 (59.21%)	189 (2.34%)	1093 (13.56%)	120 (1.49%)	887 (44.02%)	50 (2.48%)	311 (15.43%)	24 (1.19%)
10^{-3}	4940 (61.29%)	216 (2.68%)	1201 (14.9%)	127 (1.58%)	899 (44.62%)	52 (2.58%)	342 (16.97%)	24 (1.19%)
10^{-2}	5056 (62.73%)	264 (3.28%)	1349 (16.74%)	135 (1.67%)	913 (45.31%)	55 (2.73%)	383 (19.01%)	28 (1.39%)
10^{-1}	5133 (63.68%)	291 (3.61%)	1552 (19.26%)	152 (1.89%)	947 (47%)	70 (3.47%)	449 (22.28%)	35 (1.74%)

5.3.2 Mapping of IgE epitopes

It is known that a protein containing an IgE epitope is an allergen, as IgE epitopes are responsible for allergenicity. It has been observed in the case of AlgPred that only a few allergen sequences can be mapped on IgE epitopes. Thus in this study, we used a similarity-based approach for searching IgE epitopes in a protein sequence. As described in materials and methods, we used BLAST to search a protein against a database of IgE epitopes. As shown in Figure 5.3, the sensitivity increased from 55.3% to 72.99%, with E-value from 10^{-6} to 10^{-1}, implying that the allergens have similarities with IgE epitopes. Interestingly, this technique

has a low rate of false-positive or error (i.e., 0.03% to 0.66%). This technique can be used to assign allergens based on the BLAST hit of a protein against IgE epitopes. We integrated this technique into our web server to facilitate users to hit their protein against the database of IgE epitopes. In addition to BLAST-based mapping, we also used MERCI software to map IgE epitopes on a protein. In this case, IgE specific motifs that are exclusively found in the IgE epitope were discovered using MERCI software. Finally, we search these IgE specific motifs in a query protein. Though this technique was only able to identify 1% of allergens, but the false prediction was nearly negligible.

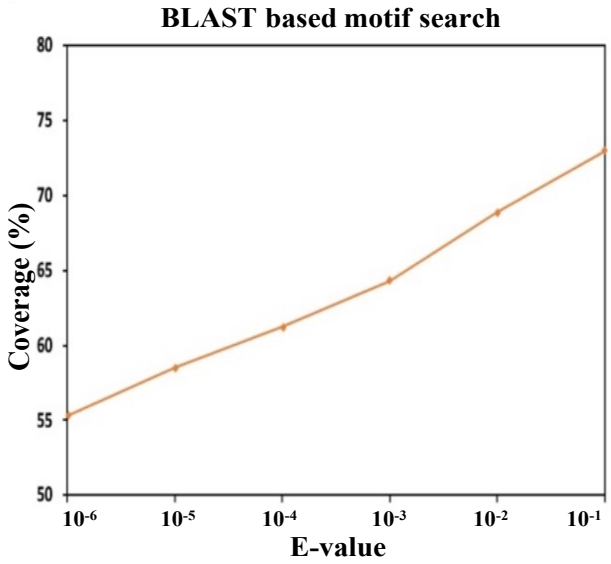

Figure 5.3: Shows the performance of BLAST with change in E-value
(x-axis shows E-value; y-axis shows the percentage of coverage)

5.3.3 Motif-based prediction

MEME/MAST

We identified motifs using MEME software from proteins in the training set. Then MAST module was used to search for matches to a set of motifs in the test set. This process is repeated five times to obtain the performance of MEME/MAST on the training dataset. As shown in Figure 5.4, sensitivity increases from 21.64% to 41.89% on the training dataset and from 10.52% to 36.97% on the validation dataset, respectively, at E-value ranging from 0.001 to 100. Although the sensitivity increased with an increase in E-value, the percentage of the wrong assignment of non-allergens to allergens also increased from 2.07% to 17.95% on the training dataset and 1.49% to 19.9% on the validation dataset, respectively. This depicts that motif-based approach alone is insufficient to discriminate between allergens and non-allergens.

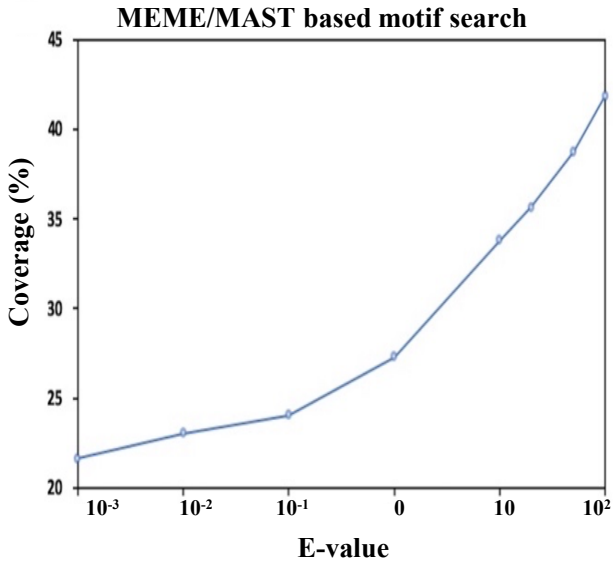

Figure 5.4: Shows the performance of MEME/MAST with change in E-value (x-axis shows E-value; y-axis shows the percentage of coverage)

MERCI

We used MERCI suite to identify the motifs/patterns present in allergens and non-allergens. Here, we have extracted motifs present exclusively in experimentally validated IgE epitopes and searched for these motifs in allergens and non-allergens protein sequences. Some motifs exclusively present in IgE epitopes and allergens are 'QQQFPQQ, FPQQQF, PQQQFP, and PYPQQ' etc. Out of 10 075 allergens proteins, 105 sequences were found to have the IgE motifs, and out of 10 075 non-allergens proteins, only 2 sequences have shown the motifs similar to IgE motifs.

5.3.4 Composition-based models

Firstly, we compute the AAC of allergen and non-allergen proteins. Several ML models (such as RF, SVM, KNN, MLP, and DT) implemented to achieve maximum performance. To achieve maximum accuracy with nearly equal sensitivity and specificity, we optimize several ML models by tuning the different parameters. RF model achieved maximum AUC 0.93 and 0.92 with balanced sensitivity and specificity on the training and validation dataset. The SVM-based model achieves reasonable performance compared to other ML models with AUC on training (0.89) and validation (0.90) dataset, as represented in Table 5.2.

Table 5.2: Shows the results of ML-based models developed using amino acid composition

ML	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
RF	89.16	82.7	85.93	0.93	0.72	84.52	83.62	84.07	0.92	0.68
SVM	85.89	79.9	82.9	0.89	0.66	87.79	77.92	82.85	0.9	0.66
KNN	85.57	79.26	82.41	0.9	0.65	85.11	79.95	82.53	0.9	0.65
MLP	81.81	80.79	81.3	0.89	0.63	85.76	76.67	81.22	0.89	0.63
DT	73.75	78.31	76.03	0.82	0.52	54.24	82.78	68.51	0.74	0.39

5.3.5 PSSM-based models

In order to compute evolutionary information of protein sequences, we generate PSSM profiles to develop several machine learning models based on positional and composition information of the allergen and non-allergen proteins. RF performs best among various ML models with balanced sensitivity and specificity. RF-based model attained maximum performance with AUC of 0.94, MCC of 0.76, and accuracy of 87.74% on the training and AUC of 0.92, MCC of 0.67, and accuracy of 83.33% on the validation dataset. The comprehensive performance of other machine learning models is represented in Table 5.3.

Table 5.3: Shows the results of ML-based models developed using PSSM profiles

PSSM	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
RF	85.79	89.68	87.74	0.94	0.76	78.56	88.09	83.33	0.92	0.67
KNN	84.23	88.11	86.17	0.93	0.72	79.06	88.39	83.72	0.91	0.68
MLP	82.84	86.74	84.79	0.93	0.70	77.52	89.93	83.72	0.91	0.68
SVM	82.72	84.89	83.80	0.91	0.68	83.77	84.37	84.07	0.90	0.68
DT	59.03	84.27	71.65	0.73	0.45	46.80	86.75	66.77	0.68	0.37

5.3.6 ML+Motif-based models

MEME/MAST approach was combined with different ML approaches. As shown in Table 5.4, RF-based model performs best among various ML techniques. It achieves AUC (0.93 and 0.92), MCC (0.72 and 0.68), and accuracy of (85.99% and 84.22%) on training and validation dataset, which is quite higher compared to other models. The input feature combines the probability scores generated by machine learning after computing AAC and scores of MEME/MAST. Similarly, MERCI approach was also combined with ML approach. Further, several ML models were implemented by taking probability scores after computing AAC and

MERCI as input features. Results have shown that the RF-based model outperforms other models; we achieve maximum accuracy, AUC and MCC of 86%, 0.93, and 0.72; 84%, 0.93, and 0.69 on training and validation dataset. The performance of both approaches is shown in Table 5.4.

Table 5.4: Shows the results of motif-based approach when combined with ML

MAST+ML										
ML	**Training**					**Validation**				
	Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
RF	89.34	82.64	85.99	0.93	0.72	84.81	83.62	84.22	0.92	0.68
KNN	79.73	85.48	82.61	0.90	0.65	78.81	86.05	82.43	0.90	0.65
SVM	81.40	80.26	80.83	0.89	0.62	87.94	77.92	82.93	0.90	0.66
MLP	78.96	80.09	79.52	0.88	0.59	86.05	76.67	81.36	0.90	0.63
DT	72.88	73.2	73.04	0.80	0.46	66.45	67.39	66.92	0.75	0.34
MERCI+ML										
RF	88.96	82.64	85.80	0.93	0.72	85.51	83.62	84.57	0.93	0.69
KNN	85.57	79.26	82.41	0.91	0.65	85.81	79.95	82.88	0.90	0.66
SVM	81.14	80.26	80.7	0.89	0.61	88.49	77.92	83.2	0.91	0.67
MLP	78.41	80.09	79.25	0.87	0.59	86.75	76.67	81.71	0.90	0.64
DT	69.52	73.2	71.36	0.77	0.43	65.66	67.39	66.53	0.74	0.33

5.3.7 ML+BLAST-based models

In order to combine the power of the similarity search approach BLAST and machine learning-based models, we developed a method using these approaches. Firstly, the BLAST search was performed for a query sequence; if we got a BLAST hit, we assigned the query sequence based on the BLAST result. A composition-based model is used to predict allergenic and non-allergenic proteins if there is no hit. The performance of our RF-based model improved significantly from AUC 0.93 to 0.99 on the training dataset and AUC 0.92 to 0.98 on the validation dataset. We also combine the BLAST search with machine learning-based models developed using the PSSM profile. Table 5.5 shows the performance of different machine learning classifiers corresponding to the combination of BLAST+AAC and BLAST+PSSM for training and validation dataset.

Table 5.5: The performance of BLAST-based approach when combined with AAC and PSSM profiles

BLAST+AAC										
ML	**Training**					**Validation**				
	Sens	**Spec**	**ACC**	**AUC**	**MCC**	**Sens**	**Spec**	**ACC**	**AUC**	**MCC**
RF	93.10	95.36	94.23	0.99	0.88	88.44	95.09	91.76	0.98	0.84
SVM	91.55	85.22	88.39	0.96	0.77	95.83	83.97	89.90	0.97	0.80
KNN	94.47	90.45	92.46	0.98	0.85	93.35	91.46	92.41	0.97	0.85
MLP	88.86	92.84	90.85	0.96	0.82	88.19	92.51	90.35	0.96	0.81
DT	85.94	88.54	87.24	0.93	0.75	84.07	88.93	86.50	0.93	0.73
BLAST+PSSM										
RF	93.59	94.29	93.94	0.99	0.88	87.79	93.25	90.52	0.97	0.81
KNN	93.28	92.47	92.87	0.98	0.86	87.05	93.25	90.15	0.97	0.8
MLP	93.78	93.25	93.52	0.98	0.87	87.54	94.99	91.27	0.97	0.83
SVM	86.95	89.54	88.24	0.94	0.77	88.19	93.5	90.84	0.97	0.82
DT	80.05	90.38	85.22	0.9	0.71	78.56	91.36	84.96	0.91	0.71

5.3.8 Hybrid model

We develop a hybrid model intending to improve the performance of the allergen prediction method. In the hybrid approach, we combine two or more methods to overcome the limitation of independent models. Here, we have combined a composition-based model with BLAST and MERCI-based approaches. In order to integrate all three approaches, proteins were first classified using BLAST at E-value of 10^{-6} followed by MERCI and the proteins not classified by either method were predicted using machine learning. The hybrid method improved the coverage, which was previously missing using all the methods separately. As shown in Table 5.6, the performance of the hybrid method improved when all the methods were combined. The best performing model was RF-based model with an accuracy of 94.23%, AUC of 0.99 and MCC of 0.88 on training dataset and accuracy of 92.26%, AUC of 0.98 and MCC of 0.85 on validation dataset, which is relatively high as compared to the other existing models.

Table 5.6: The performance of hybrid method combining ML using amino acid composition, BLAST and MERCI

ML	**Training dataset**					**Validation dataset**				
	Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
RF	93.01±0.02	95.36±0.03	94.23±0.02	0.99±0.01	0.88±0.04	89.43	95.09	92.26	0.98	0.85
KNN	94.47±0.02	90.45±0.01	92.46±0.02	0.98±0.01	0.85±0.03	94.04	91.46	92.75	0.97	0.86
SVM	91.55±0.04	85.22±0.05	88.39±0.06	0.97±0.01	0.77±0.09	96.53	83.97	90.25	0.97	0.81

MLP	90.33±0.03	90.22±0.03	90.28±0.03	0.96±0.03	0.81±0.05	90.82	89.33	90.07	0.96	0.81
DT	86.43±0.05	88.54±0.06	87.48±0.06	0.93±0.02	0.75±0.09	85.06	88.93	87.01	0.93	0.74

5.3.9 Best models developed in the study

The performance of the best classification models developed using the different features is listed in Table 5.7 .

Table 5.7: List of the features used to develop AlgPred 2.0 along with the best performing ML models

Method	Features	ML	Training					Validation				
			Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
Composition based features	AAC	RF	89.16	82.7	85.93	0.93	0.72	84.52	83.62	84.07	0.92	0.68
PSSM based features	PSSM profiles	RF	85.79	89.68	87.74	0.94	0.76	78.56	88.09	83.33	0.92	0.67
MEME/MAST	Motifs + AAC	RF	89.34	82.64	85.99	0.93	0.72	84.81	83.62	84.22	0.92	0.68
MERCI	Motifs + AAC	RF	88.96	82.64	85.8	0.93	0.72	85.51	83.62	84.57	0.93	0.69
BLAST based approach	BLAST + AAC	RF	93.1	95.36	94.23	0.99	0.88	88.44	95.09	91.76	0.98	0.84
BLAST based approach	BLAST + PSSM	RF	93.59	94.29	93.94	0.99	0.88	87.79	93.25	90.52	0.97	0.81
Hybrid approach	BLAST + Motifs + AAC	RF	93.01	95.36	94.23	0.99	0.88	89.43	95.09	92.26	0.98	0.85

5.4 Comparison with existing methods

It is important to benchmark existing methods with our method. Unfortunately, most of the methods have been developed using different datasets; thus, it is unfair to compare their performance with each other. AllerCatPro was developed on 4,180 unique allergenic proteins, but the dataset contains the proteins which have 70% or more sequence identity with each other (Maurer-Stroh et al., 2019). AllerTOP v2 was developed using 2427 allergens and 2427 non-allergens; in this method, the dataset used for building the model was redundant (Dimitrov, Bangov, et al., 2014). The methods mentioned above have not been evaluated on the independent or external dataset. Recently, a state-of-the-art method, AllerHunter, has been developed, which uses both internal (5-fold) and external CV. In this study, they used 1356 allergens and 13449 non-allergens for developing prediction models (Muh et al., 2009). AllerHunter achieved a maximum MCC of 0.738 on an external dataset that contains 129 allergens and 1314 non-allergens. Our hybrid approach achieved a maximum MCC of 0.88 on training and 0.85 on validation dataset. We have also validated our method on the independent

dataset. The dataset consists of 297 positive proteins added recently in COMPARE and Swiss-Prot (not used in our study while training, testing and validation). Out of 297 positive proteins, 280 were correctly predicted by our method using the hybrid approach. Thus, our method also achieved high performance (accuracy of 94.28%) on an independent dataset that indicates the reliability of AlgPred 2.0. The comparison of AlgPred 2.0 with other existing methods is shown in Table 5.8.

Table 5.8: Comparison of AlgPred 2.0 with existing methods

Methods	Dataset	Sens	Spec	Acc	AUC	MCC	Web server Working
AlgPred 2.0	20150	93.1	95.36	94.23	0.99±0.01	0.88	Yes
AlgPred	1278	88.87	81.86	85.02	NA	0.7053	Yes
AllerCatPro	4180	100	67.00	84.00	NA	NA	Yes
AllerTOPv2	4854	86.70	90.70	88.70	NA	0.775	Yes
AllerTOP	4420	87.60	78.00	82.80	NA	0.671	No
AllerHunter	14805	83.70	96.40	95.30	0.928±0.004	0.738	No
AllerTool	1274	86.00	86.00	NA	0.90	NA	No
AllergenFP	4854	86.80	89.10	87.90	NA	0.759	Yes

5.5 Web server & standalone software

A web server AlgPred 2.0 (https://webs.iiitd.edu.in/raghava/algpred2/), has been developed for predicting allergenic proteins. It integrates five major modules; i) Prediction, ii) IgE epitope mapping, iii) Design, iv) Motif scan, and v) BLAST search. The 'Prediction module' allows users to submit the protein sequences in FASTA format to predict the allergenic and non-allergenic proteins. In this module, the hybrid approach and RF-based model using AAC have been integrated. The 'IgE epitope mapping module' facilitates the users to map the IgE epitope on a query protein sequence. The 'Design module' is developed for generating all possible mutants of a protein by mutating a single residue at a time. The 'Motif scan module' allows to scan or map motifs in the protein sequence given by the user. To derive the motifs, it uses two software, MEME/MAST and MERCI. The 'BLAST search module' facilitates the users to perform a similarity-based search using BLAST against allergen and non-allergen database and IgE epitopes database. The web server has been designed using a responsive HTML template and browser compatibility for different OS systems. To facilitate the users to predict allergens, we developed a standalone version of AlgPred 2.0, which is available at https://github.com/raghavagps/algpred2.

5.6 Discussion & conclusion

For the last five decades, there has been a rise in the prevalence of allergic diseases worldwide. These diseases include allergic rhinitis, drug allergy, food allergy, skin allergy and insect allergy, amongst many others. While the modern lifestyle and industrialization have been deemed as a cause of these diseases, there is still a lack of measures to curb this crucial issue. To this endeavour, several *in-silico* methods/techniques have been developed in the past that could be used to assess the allergenicity of proteins. Each method has its own merits and limitations. In the present study, we have developed an updated method of our previous methods called AlgPred, which combines a wide range of approaches that include SVM-based models, BLAST and mapping of IgE epitopes. One of the major limitations of AlgPred is that it was trained on limited data (i.e., 578 allergens, 700 non-allergens and 183 IgE epitopes) due to lack of data. In the last 14 years, a number of allergens and IgE epitopes have been discovered. Thus, there is a need to update AlgPred using recent advances in the field of immunology. In the present study, we have developed models using 10 075 allergens and 10 075 non-allergens. In addition, 10 451 IgE epitopes were used to identify antigenic regions in proteins.

The study also shows that similarity-based search, i.e., BLAST, and motif prediction-based models performed poorly when used individually due to a large number of no hits, as shown in Table 5.1 and Figure 5.3. To overcome this limitation, we developed machine learning models using various classifiers. Composition-based descriptors and evolutionary information based descriptors were employed as features since the allergenicity property has been widely accredited to the protein sequence. We developed hybrid models to combine the power of similarity search-based technique and ML-based models. As shown in Table 5.6, we got the highest performance on both training and testing datasets. Apart from the significant prediction accuracy, we also highlighted the distinction of our current method over the previously developed AlgPred in various contexts. The utilization of a more extensive dataset, MERCI-based prediction model, voting-based BLAST similarity search, evolutionary information as protein descriptors, and use of a wide range of ML classifiers are a few significant improvements. To facilitate the scientific community and promote extensive public usage of the proposed prediction method, we have also provided a free web server, AlgPred 2.0 (https://webs.iiitd.edu.in/raghava/algpred2/) as well as standalone version is provided at (https://github.com/raghavagps/algpred2). We believe our method would aid in more accurate

recognition of allergenic proteins and thereby bring a significant improvement in the field of allergy research and therapy.

5.7 Limitation of the study

In this study, we proposed the prediction method AlgPred 2.0 to classify the allergenic and non-allergenic protein sequences. This method that has been trained on the largest dataset and also combines most of the existing approaches to identify allergens with high accuracy. The dataset in this study has been collected from various databases consisting different sources, such as microbes, fish, plants and animals,. So, the limitation of the method is that it can only predict the protein as allergenic without specifying its source and type of allergy it can cause.

Chapter 6

Identification of chemical allergens and designing non-allergenic compounds

6.1 Introduction

Allergy is an inappropriate reaction of the immune response when it misidentifies a harmless foreign substance as a threat (Sharma et al., 2020), (Dimitrov, Bangov, et al., 2014), (Zhang et al., 2007). These foreign substances are known as allergens, triggering various allergic reactions and leading to various allergic diseases. Different types of aeroallergens (e.g., pollens, spores, dust mites), food allergens (e.g., eggs, peanuts, tree nuts, genetically modified foods), and chemical allergens in personal care products (e.g., fragrances in the skin and hair care products, dyes, creams) (Maurer-Stroh et al., 2019), (Sharma et al., 2020) can lead to allergic symptoms such as allergic asthma, rhinitis, skin reactions and anaphylaxis. It involves a series of allergic reactions from mild symptoms like itchy skin, rashes, facial swelling, irritation of the eyes leading to watery eyes and nose to severe symptoms like shortness of breath, lack of consciousness, weak pulse, nausea, vomiting, which can even lead to death if untreated (U.S. FDA, 2020), (Mak TW, 2014).

There is a wide variety of molecules that can pose a threat as allergens, including biological molecules like proteins and peptides (Dang & Lawrence, 2014), (Goodman et al., 2016) or some chemical compounds (Kimber et al., 2010). Other than that, molecules like lipids (Del Moral & Martinez-Naves, 2017), carbohydrates (Commins & Platts-Mills, 2010), nucleic acid (mRNA vaccines) (Rubin, 2021) and some engineered nanoparticles (Alsaleh & Brown, 2020) can also stimulate some specific allergic reactions like asthma, food allergies and chronic kidney disease, respectively.

In day-to-day life, the human body is exposed to innumerable chemical substances, including natural or organic compounds, pharmaceuticals, cosmetic products (such as makeup, soaps, perfumes, lotions, hair dyes etc.), various other chemicals (such as preservatives in food, metals in the jewellery) (Sharma, Srivastava, et al., 2017). Multiple new chemical entities are also introduced every year for designing new drugs or other purposes (Banerjee et al., 2016). Many of these chemical products are known to provoke allergic reactions, causing skin sensitization in some people, which results in skin or contact dermatitis. Some may cause the sensitization of the respiratory tract leading to occupational asthma, which can be lethal (Kimber et al., 2011). The allergic reaction caused by small chemical compounds is developed in two phases, i.e., (i) Sensitization or Induction and (ii) Elicitation. The mechanism of allergy caused by chemical allergens is depicted in Figure 6.1.

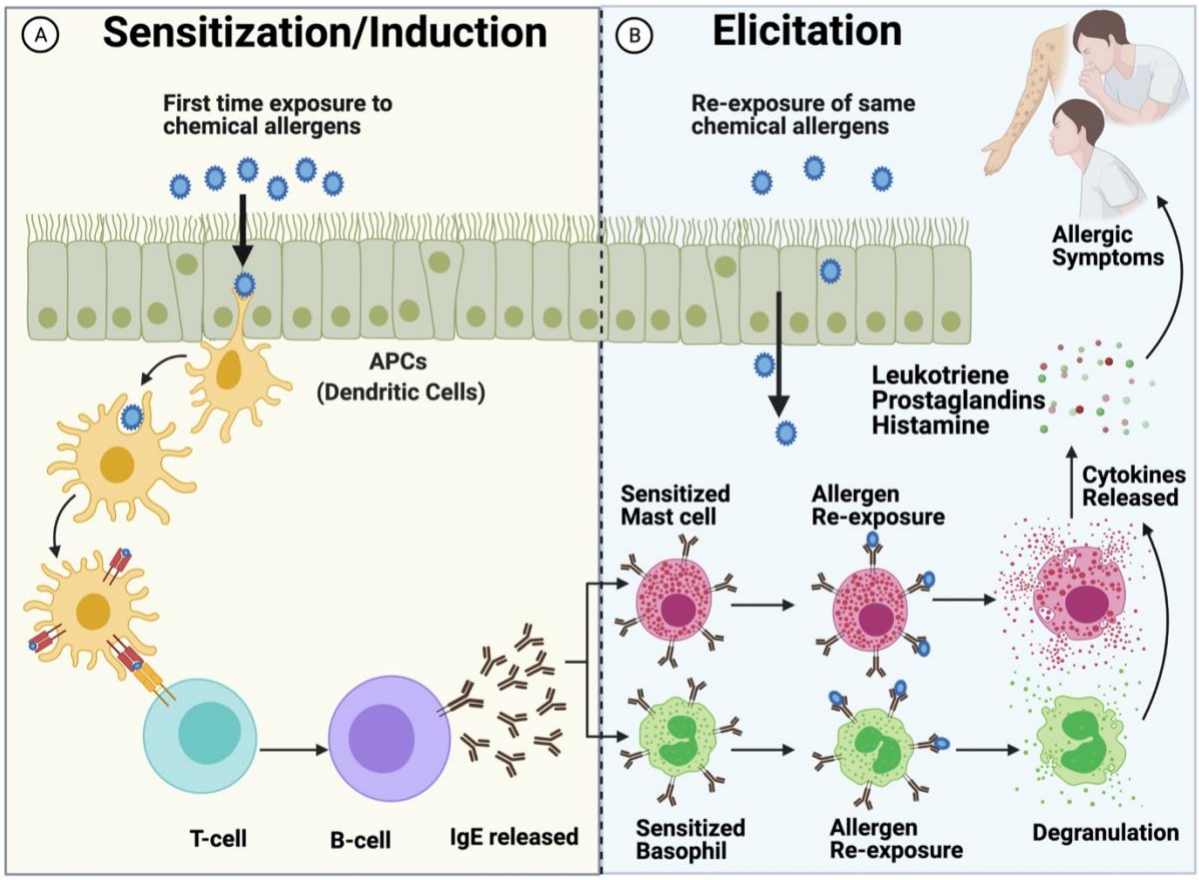

Figure 6.1: The mechanism of the allergy caused by chemical allergens (10.1016/j.compbiomed.2021.104746)

The protein/peptide allergen and chemical toxicity prediction tools and databases are widely used by the scientific community in designing the drug molecules with desired properties. However, limited efforts have been made to develop a method or tool to predict the allergenicity of chemicals causing allergy. In order to help the researchers to study the allergenicity of the chemical molecules, it is the first time we have made an attempt to develop a computational method named ChAlPred (Chemical Allergen Prediction) for predicting the chemical allergens. To create the dataset, we extracted the information of allergenic and non-allergenic chemicals from different resources and applied various machine learning approaches to develop the classification models. Our best models have been integrated into the web server, which can be freely accessible at https://webs.iiitd.edu.in/raghava/chalpred/.

6.2 Materials & methods

6.2.1 Dataset collection

In this study, we have collected allergenic and non-allergenic chemical compounds from IEDB (Vita et al., 2019) and Chemical Entities of Biological Interest (ChEBI) database (Hastings et

al., 2016). We obtain a total of 519 unique chemical compounds having allergenic properties from IEDB and ChEBI. On the other hand, we have taken 2211 non-allergenic chemical compounds with a filter of non-peptidic; No IgE; No histamine; No hypersensitivity; No allergy; No cancer collected from IEDB database. The chemical compounds with allergenic properties were considered as a positive dataset (allergens), and the compounds having non-allergenic properties were taken as a negative dataset (non-allergens).

6.2.2 Generation of descriptors

The chemical descriptors/features of allergen and non-allergen chemical compounds were computed using PaDEL software (Yap, 2011). It can compute a number of molecular descriptors, such as 2D, 3D and different types of fingerprints for a single chemical compound. It has computed 729 2D descriptors, 431 3D descriptors, and 16092 binary fingerprint-based (FP) descriptors for the 403 allergen and 1074 non-allergen chemical compounds. These 2D, 3D, and FP descriptor files were further used to develop different machine learning models.

6.2.3 Feature selection

In this study, we have used PaDEL software to compute the 2D, 3D and Fingerprint based features for the chemical compounds. It has computed 729 2D descriptors, 431 3D descriptors, and 16092 binary fingerprint-based descriptors for the 403 allergen and 1074 non-allergen chemical compounds. As the number of features computed is very large, so we have used various feature selection techniques to select the significant set of features. We used the variance threshold-based, correlation-based, and SVC-L1-based feature selection techniques. First, low-variance features were removed using VarianceThreshold feature selection method from the sklearn package (Pedregosa F, 2011), to remove those features with small value changes. Second, correlation-based feature selection method was used to remove the features having a correlation >=0.6. To further reduce the vector size, we have applied SVC-L1 method. Finally, we get the most important feature set, i.e., 14 descriptors out of 34 descriptors for 2D, 6 out of 8 descriptors for 3D and 22 FP descriptors out of 957.

6.2.4 Machine learning techniques

In the current study, different machine learning techniques have been used to classify allergen and non-allergen chemical compounds. We used LR, KNN, DT, GNB, XGB, SVC, and RF-based techniques for the classification.

6.2.5 Criteria for evaluating performance

We have also applied a 5-fold CV on 80% of training data for the internal training, testing and model evaluation. The performance of machine learning models was evaluated using the standard evaluation parameters such as sensitivity, specificity, accuracy, MCC and AUC. Figure 6.2 shows the comprehensive framework of the study.

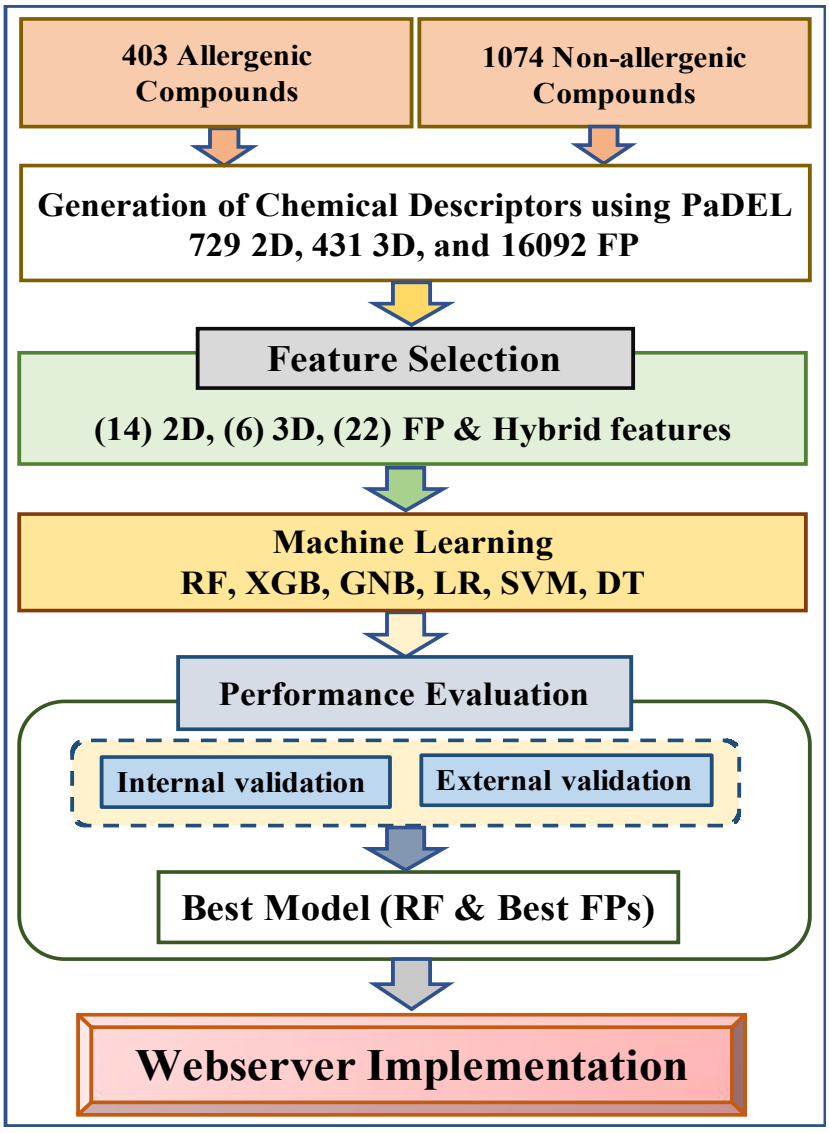

Figure 6.2: Shows the overall methodology used for developing method for chemical allergen prediction (10.1016/j.compbiomed.2021.104746)

6.3 Results

6.3.1 Performance of machine learning-based models

After computing various 2D, 3D and FP descriptors of the allergen and non-allergen chemical compounds, we have used these features to develop the ML models.

6.3.1.1 Models using 2D descriptors

The models developed using these ML techniques were optimized by tuning different parameters. For 14 (2D) descriptors, it was observed that the model based on the XGB algorithm performed better than other classifiers and achieved maximum AUC of 0.89 and 0.89 on the training and validation datasets, respectively (Table 6.1).

Table 6.1: The performance of machine learning-based models developed using 14 (2D) descriptors

ML	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
XGB	81.68	82.10	81.99	0.89	0.59	80.25	76.64	77.63	0.89	0.52
KNN	81.37	81.99	81.82	0.89	0.59	81.48	79.91	80.34	0.88	0.57
RF	81.99	81.29	81.48	0.89	0.59	83.95	81.31	82.03	0.90	0.60
LR	80.12	81.05	80.80	0.88	0.57	81.48	77.57	78.64	0.87	0.54
DT	79.50	79.65	79.61	0.85	0.55	67.90	76.62	74.24	0.80	0.41
GNB	78.57	78.36	78.42	0.86	0.52	81.48	77.57	78.64	0.86	0.54
SVC	78.26	77.78	77.91	0.87	0.52	85.18	78.04	80.00	0.88	0.58

6.3.1.2 Models using 3D descriptors

For 6 (3D) descriptors, a model based on the RF algorithm performed better than others and achieved a maximum AUC of 0.88 and 0.85 on the training and validation datasets, respectively. The result for 3D features is shown in Table 6.2.

Table 6.2: The performance of machine learning-based models developed using 6 (3D) descriptors

ML	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
RF	79.14	78.69	78.81	0.88	0.53	75.32	81.19	79.66	0.85	0.53
KNN	77.61	77.87	77.78	0.85	0.51	68.83	80.73	77.63	0.83	0.47
XGB	76.67	76.11	76.27	0.86	0.48	77.92	79.36	78.98	0.85	0.52

SVC	73.32	72.25	72.54	0.81	0.41	62.34	73.85	70.85	0.77	0.33
LR	68.41	71.31	70.51	0.73	0.36	70.13	72.48	71.86	0.76	0.38
GNB	68.41	70.61	70.00	0.75	0.36	64.93	73.39	71.17	0.75	0.35
DT	69.33	68.38	68.64	0.76	0.34	71.43	60.55	63.39	0.72	0.28

6.3.1.3 Models using FP descriptors

The machine learning model developed for 22 (FP) descriptors, a model based on the RF algorithm outperformed other classifiers and achieved maximum AUC of 0.92 and 0.92 on the training and validation datasets, respectively. The result for FP features is shown in Table 6.3.

Table 6.3: The performance of machine learning-based models developed using 22 (FP) descriptors

ML	**Training**					**Validation**				
	Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
RF	85.06	85.11	85.09	0.92	0.66	86.67	85.52	85.81	0.92	0.67
XGB	85.37	85.11	85.18	0.92	0.66	85.33	85.52	85.47	0.89	0.66
LR	83.84	83.82	83.83	0.91	0.64	81.33	81.45	81.42	0.86	0.58
SVC	83.54	83.01	83.15	0.91	0.62	82.67	80.54	81.08	0.86	0.58
KNN	82.93	83.12	83.06	0.91	0.62	85.33	80.54	81.76	0.87	0.60
GNB	79.57	79.37	79.42	0.88	0.55	70.67	81.45	78.72	0.83	0.45
DT	79.88	78.90	79.17	0.86	0.54	77.33	76.92	77.03	0.83	0.45

6.3.1.4 Models using hybrid features

Another ML-based model was developed by combining all three types of descriptors, i.e., 2D, 3D and FP. A total of 42 features were used for machine learning. It was shown that the RF-based model had achieved a maximum AUC of 0.94 and 0.93 on the training and validation datasets, respectively. As there were only 6 (3D) descriptors, we have excluded them and have developed the model with only 36 features (14 (2D) and 22 (FP) descriptors). The obtained results show that there was no significant change in the performance of the model. The RF-based model has achieved an AUC of 0.94 on the training dataset and 0.93 on the validation dataset. The results of machine learning models combining all three descriptors and excluding 3D descriptors are shown in Table 6.4.

Table 6.4: The performance of machine learning-based hybrid models developed after combining all descriptors

42 (2D+3D+FP) Descriptors										
ML	**Training**					**Validation**				
	Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
RF	85.63	86.00	85.90	0.94	0.68	87.95	82.55	84.07	0.93	0.66
SVC	84.06	84.01	84.03	0.91	0.64	93.98	78.77	83.05	0.92	0.66
KNN	84.06	83.66	83.77	0.92	0.63	80.72	81.60	81.36	0.92	0.58
XGB	83.75	83.78	83.77	0.92	0.63	85.54	79.72	81.36	0.92	0.60
LR	83.75	83.08	83.26	0.91	0.62	85.54	82.08	83.05	0.89	0.63
GNB	84.06	79.70	80.88	0.89	0.59	73.49	83.02	80.34	0.87	0.54
DT	79.06	78.76	78.85	0.87	0.53	81.93	80.66	81.02	0.88	0.58
36 (2D+FP) Descriptors										
RF	87.5	87.28	87.34	0.94	0.71	84.34	83.02	83.39	0.93	0.63
XGB	85.01	84.95	84.96	0.93	0.66	81.93	81.13	81.36	0.91	0.59
LR	84.06	84.48	84.37	0.91	0.64	84.34	83.96	84.07	0.90	0.64
KNN	83.44	84.13	83.94	0.93	0.63	81.93	82.08	82.03	0.91	0.60
SVC	84.06	83.08	83.35	0.90	0.63	86.75	81.60	83.05	0.91	0.63
GNB	84.06	79.00	80.37	0.88	0.58	77.11	83.96	82.03	0.88	0.58
DT	80.02	79.00	79.27	0.86	0.54	86.75	76.89	79.66	0.88	0.58

6.3.1.5 Best models of the study

The performance of the best classification models developed using the different features is listed in Table 6.5.

Table 6.5: List of the features used to develop ChAlPred along with the best performing ML models

Method	**Features**	**ML**	**Training**					**Validation**				
			Sens	**Spec**	**Acc**	**AUC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUC**	**MCC**
Chemical descriptors	14 (2D) descriptors	XGB	81.68	82.1	81.99	0.89	0.59	80.25	76.64	77.63	0.89	0.52
Chemical descriptors	6 (3D) descriptors	RF	79.14	78.69	78.81	0.88	0.53	75.32	81.19	79.66	0.85	0.53
Chemical descriptors	22 (FP) descriptors	RF	85.06	85.11	85.09	0.92	0.66	86.67	85.52	85.81	0.92	0.67
Feature selection	42 (2D+3D+FP) descriptors	RF	85.63	86	85.9	0.94	0.68	87.95	82.55	84.07	0.93	0.66
Hybrid approach	36 (2D+FP) descriptors	RF	87.5	87.28	87.34	0.94	0.71	84.34	83.02	83.39	0.93	0.63

6.3.1.6 Fingerprints-based analysis

In order to understand the importance of each FP in classifying allergens and non-allergens, we have computed the prediction ability of each FP. We used our in-house scripts to check the discrimination ability of fingerprint-based descriptors calculated by PaDEL. We ranked the FPs according to their probabilities for correctly classifying the chemical as allergen and non-allergen. Based on ranking, we identified the most important 20 FPs. Ten FPs are highly present in allergens and were called positive FPs, whereas other 10 which are highly present in non-allergens were called negative FPs. Figure 6.3 depicts the frequency of top 10 positive and 10 negative FPs in allergens and non-allergens. These 10 positive FPs are highly abundant in allergens but negligible in non-allergens. Similarly, 10 negative FPs are highly abundant in non-allergens but negligible in allergens.

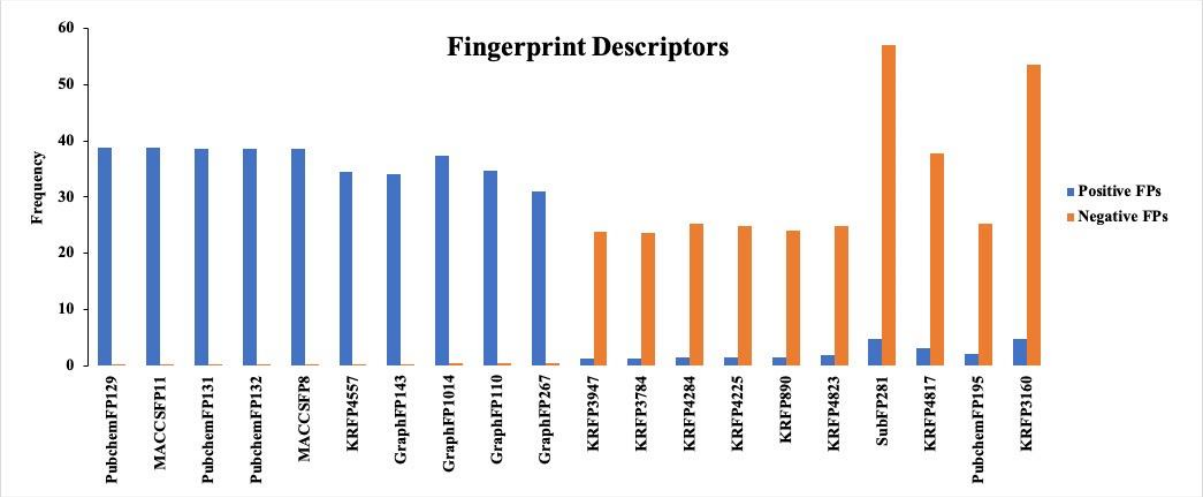

Figure 6.3: Shows the frequency of top 10 positive/ negative fingerprints in allergens and non-allergens (10.1016/j.compbiomed.2021.104746)

The description of top 10 positive and 10 negative FPs in allergens and non-allergens is presented in Table 6.6.

Table 6.6: Description of top 10 positive and negative fingerprints in allergens and non-allergens

Finger prints	Fingerprint Name	Frequency	Descriptor Class	Description
Positive FP	PubchemFP129	38.82%	Pubchem FP	>= 1 any ring size 4
	MACCSFP11	38.82%	MACCS FP	MACCS keys
	PubchemFP131	38.51%	Pubchem FP	>= 1 saturated or aromatic nitrogen-containing ring size 4
	PubchemFP132	38.51%	Pubchem FP	>= 1 saturated or aromatic heteroatom-containing ring size 4
	MACCSFP8	38.51%	MACCS FP	MACCS keys

	KRFP4557	34.47%	Klekota-Roth FP	O=CNCCNC=O
	GraphFP143	34.16%	CDK graph only FP	Specialized version of the Fingerprinter which does not take bond orders into account
	GraphFP1014	37.27%	CDK graph only FP	Specialized version of the Fingerprinter which does not take bond orders into account
	GraphFP110	34.78%	CDK graph only FP	Specialized version of the Fingerprinter which does not take bond orders into account
	GraphFP267	31.06%	CDK graph only FP	Specialized version of the Fingerprinter which does not take bond orders into account
Negative FP	SubFP281	57.04%	Substructure FP	[OX2;$([r5]1@C@C@C(O)@C1),$([r6]1@C@C@C(O)@C(O)@C1)]
	KRFP3160	53.67%	Klekota-Roth FP	C1CCOCC1
	KRFP4817	37.83%	Klekota-Roth FP	OCC1OCC(O)C(O)C1O
	KRFP4284	25.26%	Klekota-Roth FP	NCC(O)CO
	PubchemFP195	25.26%	Pubchem FP	>= 3 saturated or aromatic heteroatom-containing ring size 6
	KRFP4225	24.91%	Klekota-Roth FP	NC(CO)CO
	KRFP4823	24.91%	Klekota-Roth FP	OCCCNC=O
	KRFP890	24.10%	Klekota-Roth FP	[!#1][NH]C(=O)[CH3]
	KRFP3947	23.86%	Klekota-Roth FP	CNCC(O)CO
	KRFP3784	23.52%	Klekota-Roth FP	CCNCC(O)CO

6.4 Case study

To identify the FDA-approved drugs that can cause allergic reactions, we have downloaded a total of 2675 FDA drug molecules from the DrugBank Database (Wishart et al., 2018). Out of 2675, we have only considered 1102 drugs that are approved. From 1102 drug molecules, the 2D structures were available only for 842 drugs. Finally, we have the structures of 842 FDA-approved drug molecules, which were used to identify which drug molecules could be allergenic and non-allergenic. We have used the hybrid model of the Predict module on the "ChAlPred" web server. The prediction was made using the default parameters. The hybrid model has predicted 114 drug molecules to be allergenic. Several studies also support our findings that some of these drugs can cause allergy in the patient when administered. We have identified 20 drug molecules that are used to cure some diseases but also tend to cause allergic symptoms. Table 6.7 depicts the information of the drug molecules which cause some allergic reactions.

Table 6.7: FDA-approved drug molecules predicted by our server (ChAlPred) causing allergic symptoms

Drug Bank ID	FDA-Approved Drugs	Prediction	Allergic symptoms
DB01112	Cefuroxime	Allergen	Anaphylactic reaction (Del Villar-Guerra et al., 2016)
DB00859	Penicillamine	Allergen	Skin allergy (Zhu et al., 2020)
DB01007	Tioconazole	Allergen	Contact hypersensitivity (Heikkila et al., 1996)
DB06209	Prasugrel	Allergen	Hypersensitivity skin reaction (Kim et al., 2014)
DB01330	Cefotetan	Allergen	Cefotetan-induced anaphylaxis (Nam et al., 2015)

6.5 Web server interface

We have developed a user-friendly web server named ChAlPred for the prediction of chemicals as allergens and non-allergens. In this server, we have provided the three modules: (i) Predict, (ii) Draw and (iii) Analog design module. The Predict module allows the user to submit the chemical compounds in different formats, such as SMILE, SDF and MOL formats, to predict whether the chemical could be allergenic or non-allergenic. The Draw module allows the user to draw or modify a molecule in an interactive way using Ketcher (Life Science open Source, 2021) and submit the molecule to predict whether the modified compounds will be allergenic or not. The Analog design module can be used to generate analogs based upon a combination of a given scaffold, building blocks and linkers. The server subsequently predicts the generated analogs as allergenic or non-allergenic. The web server has been designed using a responsive HTML template and browser compatibility for different OS systems.

6.6 Discussion & conclusion

One of the major challenges in the field of drug discovery is the side effect or adverse reactions of drugs. In the past, a number of drugs have already been withdrawn from the market due to their adverse effects. A wide range of toxicities are responsible for the side effects of drugs; it may be cytotoxicity, immuno-toxicity, hemo-toxicity, liver toxicity or allergenicity (Yang et al., 2018). Identification of toxicity is a costly, time-consuming and tedious task. Thus, there is a need to predict these toxicities using *in-silico* methods. Numerous tools have been developed to estimate the toxicity of the chemicals using different methodologies, such as The Toxicity Estimation Software Tool (TEST). It uses Quantitative Structure-Activity Relationships (QSAR) to estimate the toxicity of chemicals (U.S. EPA, 2020). VegaQSAR,

Toxtree (Patlewicz et al., 2008), and PreADMET are the other tools based on the QSAR model for toxicity prediction of the chemical molecules. ML-based tools such as ToxiM, developed by Sharma et al., predict the toxicity and toxicity-related properties of small chemical molecules using ML approaches (Sharma, Srivastava, et al., 2017), ProTox-II (Banerjee et al., 2018).

In contrast, no tool has been developed for predicting the allergenicity of chemicals. In this work, we have collected chemical compounds with their well-defined molecular descriptors utilizing publicly available databases such as IEDB and ChEBI. The data yielded several descriptors, which were reduced using various feature selection methods. We sorted the most important feature set, i.e., 14 for 2D, 6 for 3D and 22 for FP descriptors. Based on these selected features (14 2D and 22 FP), we have successfully employed several ML approaches and found that RF attained the highest AUC of 0.94 and 0.93 in the training as well as validation dataset. In addition, fingerprints-based analysis suggests that two positive FPs, i.e., PubChemFP129 (Extended Smallest Set of Smallest Rings (ESSSR) ring set >= 1 any ring size 4) and GraphFP1014 are highly present in allergenic chemical compounds, and three negative FPs, i.e., Klekota-Roth fingerprints (KRFP890 (!#1NHC(=O)CH3, KRFP3160 (C1CCOCC1)) and Substructure fingerprint (SubFP281 OX2;$(r51@C@C@C(O)@C1),$(r61@C@C@C(O)@C(O)@C1)) are abundant in non-allergenic chemical compounds. FDA-approved drug analysis has shown that few drugs used to treat certain diseases are also causing allergies as a side effect. Literature evidence has shown that the administration of FDA-approved drugs such as Cefuroxime (Del Villar-Guerra et al., 2016), Spironolactone (Ghislain et al., 2004), Penicillamine (Zhu et al., 2020) can cause allergic reactions like skin allergies, anaphylactic reactions, hypersensitivity. For instance, a case report has shown that 60 year old patient was experiencing an anaphylactic reaction after being given the antibiotic cefuroxime (Gu et al., 2019). Another report by Kinsara has shown that Spironolactone, a potassium sparing diuretic, was given to a patient diagnosed with idiopathic cardiomyopathy, and he developed macular rashes on both the arms (Kinsara AJ, 2018). A clinical study by Zhu et al. reported that the patients with Wilson disease were given D-penicillamine medication at first, but later they developed neurological symptoms as well as allergies (Zhu et al., 2020).

We can see that these medications can cause a variety of allergic reactions in patients, some of which can be fatal. To prevent these problems, there is a dire need to predict the allergenicity of chemical compounds before using them for treatment purposes. Eventually, we built a freely available web server, namely ChAlPred, for predicting allergenic and non-allergenic chemical

compounds using ML techniques based on their 2D, 3D and FP molecular descriptors. We hope that this study will be helpful in the future for designing drug molecules with no allergenic properties.

6.7 Limitation of the study

Various methods have been developed to predict the allergenicity of the proteins, but there is no method to study the allergenicity of the chemical molecules. So, in this study, we have made an attempt to develop a novel computational method named ChAlPred (Chemical Allergen Prediction) for predicting chemical allergens. Being the first of its kind, the limitation of the study is the small dataset. We have used 403 allergenic and 1074 non-allergenic chemical compounds obtained from IEDB database. There is a need for sufficient data size to develop an accurate and reliable method. In this study, a systematic attempt has been made to develop the best possible models in the present scenario.

Chapter 7

Summary

Microbial infections are one of the leading causes of high mortality and morbidity throughout the globe. Virulence factors associated with the pathogens play a vital role in establishing the interaction with the host as well as causing the disease to the host. Proteins related to the virulence factors are often intended as drug and vaccine targets for designing therapeutic molecules against pathogenic micro-organisms. However, increasing antibiotics and drug resistance poses a major challenge in the development of novel drugs. To combat this issue, there is an urgent need to accelerate the discovery of new antibiotics, identification of novel therapeutic targets and pharmacological compounds with unique mechanisms of action. Advances in various technologies led to the explosive growth of experimentally verified proteomic data related to virulence factors, which is available in the form of repositories. Computational approaches can be developed using the available data and can be utilised in prior detection of putative drug and vaccine targets to save time as well as money. Thus overall work focuses in developing computational tools for designing the therapeutic molecules against the virulent factors of pathogens.

Chapter 2 provides a glimpse of the traditional and conventional approaches used to identify the virulence factors and how these factors can be used as potential drug and vaccine candidates. In this chapter, we have discussed the computational resources, repositories, knowledgebase and *in-silico* tools that have been developed till now to store the information regarding the virulence factors. Apart from that, tools and databases developed for retrieving the information about toxins, allergy caused by proteins and chemicals have also been discussed.

Chapter 3 deals with the development of a web resource for the identification of putative virulence factors of various pathogens. The identification of virulence-associated factors is of utmost importance and is of great immunological interest. These factors can be used as potential drug and vaccine candidates to treat various microbial infections. Thus, to aid the clinicians and scientific community, we have developed a machine learning-based method named “VirFacPred” for the identification of novel virulence factors. The method has been trained, tested and evaluated on two datasets curated from the recent release of the Swiss-Prot. The main dataset contains 7058 positive and 7058 negative sequences, whereas the alternate dataset consists of 4714 virulent and 4714 non-virulent sequences. To provide unbiased evaluation, we performed internal validation on 80% of the data and external validation on the remaining 20% of data. More than 9000 features were generated for both datasets, and a feature

selection technique was also implemented to compute the best feature set. Firstly, a similarity-based search using BLAST was performed against the dataset, and virulent factors were predicted based on the level of similarity with known sequences. Secondly, MERCI-based motif search was implemented to identify the motifs which are exclusively present in virulent proteins. Thirdly, several prediction models have been developed using a wide range of machine learning techniques. Finally, a hybrid method that combines all three approaches has been developed to attain the maximum performance of the model with balanced sensitivity and specificity. Our best model achieved the maximum AUC around 0.97 with MCC 0.77 on the validation dataset. Moreover, we developed models on the alternate dataset as well. The best machine learning models have been implemented in the web server named "VirFacPred" (https://webs.iiitd.edu.in/raghava/virfacpred/), which allows the prediction, mapping, motif search for the virulent proteins of the pathogens as well as designing non-virulent proteins.

Chapter 4 mainly discusses about toxins, which are one of the major virulence factors that play a crucial role in damaging the host cell. Proteins/peptides have shown to be promising therapeutic agents for a variety of diseases. However, toxicity is one of the obstacles in protein/peptide-based therapy. In order to address this problem, a highly accurate method, ToxinPred2, has been developed for predicting toxins with high precision. This is an update of ToxinPred developed mainly for predicting the toxicity of peptides and small proteins. The method has been trained, tested and evaluated on three datasets curated from the recent release of the Swiss-Prot. To provide unbiased evaluation, we performed internal validation on 80% of the data and external validation on the remaining 20% of data. We have implemented the following techniques for predicting protein toxicity; (i) BLAST-based similarity, (ii) MERCI-based motif search and (iii) Prediction models. Similarity and motif-based techniques achieved a high probability of correct prediction with poor sensitivity/coverage, whereas models based on machine-learning techniques achieved balance sensitivity and specificity with reasonably high accuracy. Finally, we developed a hybrid method that combined all three approaches and achieved a maximum AUC around 0.99 with MCC 0.91 on the validation dataset. In addition, we developed models on alternate and realistic datasets. The best machine learning models have been implemented in the web server named "ToxinPred2", which is available at https://webs.iiitd.edu.in/raghava/toxinpred2/ and a standalone version at https://github.com/raghavagps/toxinpred2. This is a general method developed for predicting the toxicity of proteins regardless of their source of origin.

Chapter 5 focuses on the virulent factors which are responsible for allergy, which is the hypersensitivity of the immune system. To address this problem, a method called AlgPred 2.0 has been developed for identifying allergenic proteins with high accuracy. AlgPred 2.0 is a web server developed for predicting allergenic proteins and allergenic regions in a protein. It is an updated version of AlgPred developed in 2006. The dataset used for training, testing and validation consist of 10 075 allergens and 10 075 non-allergens. In addition, 10 451 experimentally validated IgE epitopes were used to identify antigenic regions in a protein. All models were trained on 80% of data called training dataset, and the performance of models was evaluated using 5-fold CV technique. The performance of the final model trained on the training dataset was evaluated on 20% of data called validation dataset; no two proteins in any two sets have more than 40% similarity. First, a BLAST search was performed against the dataset, and allergens were predicted based on the level of similarity with known allergens. Second, IgE epitopes obtained from the IEDB database were searched in the dataset to predict allergens based on their presence in a protein. Third, motif-based approaches like multiple EM for motif elicitation/motif alignment and search tools have been used to predict allergens. Fourth, allergen prediction models have been developed using a wide range of machine learning techniques. Finally, the ensemble approach has been used for predicting allergenic protein by combining prediction scores of different approaches. Our best model achieved maximum performance in terms of AUC 0.98 with MCC 0.85 on the validation dataset. A web server AlgPred 2.0, has been developed that allows the prediction of allergens, designing of non-allergenic proteins, mapping of IgE epitope, motif and BLAST search (https://webs.iiitd.edu.in/raghava/algpred2/), and a standalone software package is available at (https://github.com/raghavagps/algpred2).

Chapter 6 describes that in addition to protein and peptides, allergy is also caused by chemical compounds, known as chemical allergy. A therapeutic molecule may cause side effects due to its allergic potential. In the past, various methods have been generated for predicting the allergenicity of proteins and peptides. In contrast, there is no method that can predict the allergenic potential of chemicals. In order to overcome this issue, a novel method ChAlPred has been developed for predicting chemical allergens and designing chemical analogs with desired allergenicity. In this study, we have used 403 allergenic and 1074 non-allergenic chemical compounds obtained from the IEDB database. The PaDEL software was used to compute the molecular descriptors of the chemical compounds to develop different prediction models. All the models were trained and tested on the 80% training data and evaluated on the

20% validation data using the 2D, 3D and FP descriptors. We have developed different prediction models using several machine learning approaches. It was observed that the RF-based model developed using hybrid descriptors performed the best and achieved the maximum accuracy of 83.39% and AUC of 0.93 on validation dataset. The fingerprint analysis of the dataset indicates that certain chemical fingerprints are more abundant in allergens that include PubChemFP129 and GraphFP1014. We have also predicted the allergenicity potential of FDA-approved drugs using our best model and identified the drugs causing allergic symptoms (e.g., Cefuroxime, Spironolactone, Tioconazole). Our results agreed with the allergenicity of these drugs reported in the literature. In summary, attempts have been made to develop *in-silico* models that can be used to design directly/indirectly therapeutic molecules against disease-causing agents. To facilitate the scientific community, we developed a smart-device compatible web server ChAlPred (https://webs.iiitd.edu.in/raghava/chalpred/) that allows to predict and design the chemicals with allergenic properties.

The work presented in the thesis addresses various computational tools and methods developed for the identification of virulence factors as well as other factors associated with them. Computationally identified virulence factors, toxic proteins, allergenic proteins and chemical allergens can be utilised for developing drugs and vaccines against various infections. We anticipate that our findings will aid the clinicians, researchers, scientific community, and general public to understand the mechanism and develop better therapeutic approaches.

Bibliography

Abastabar, M., Hosseinpoor, S., Hedayati, M. T., Shokohi, T., Valadan, R., Mirhendi, H., Mohammadi, R., Aghili, S. R., Rahimi, N., Aslani, N., Haghani, I., & Gholami, S. (2016). Hyphal wall protein 1 gene: A potential marker for the identification of different Candida species and phylogenetic analysis. *Curr Med Mycol*, *2*(4), 1-8. https://doi.org/10.18869/acadpub.cmm.2.4.1

Ahmed, E., & Holmstrom, S. J. (2014). Siderophores in environmental research: roles and applications. *Microb Biotechnol*, *7*(3), 196-208. https://doi.org/10.1111/1751-7915.12117

Akarsu, H., Bordes, P., Mansour, M., Bigot, D. J., Genevaux, P., & Falquet, L. (2019). TASmania: A bacterial Toxin-Antitoxin Systems database. *PLoS Comput Biol*, *15*(4), e1006946. https://doi.org/10.1371/journal.pcbi.1006946

Alsaleh, N. B., & Brown, J. M. (2020). Engineered Nanomaterials and Type I Allergic Hypersensitivity Reactions. *Front Immunol*, *11*, 222. https://doi.org/10.3389/fimmu.2020.00222

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, *215*(3), 403-410. https://doi.org/10.1016/S0022-2836(05)80360-2

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, *25*(17), 3389-3402. https://doi.org/10.1093/nar/25.17.3389

Ansari, F. A., Kumar, N., Bala Subramanyam, M., Gnanamani, M., & Ramachandran, S. (2008). MAAP: malarial adhesins and adhesin-like proteins predictor. *Proteins*, *70*(3), 659-666. https://doi.org/10.1002/prot.21568

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, *2*, 28-36. https://www.ncbi.nlm.nih.gov/pubmed/7584402

Bailey, T. L., & Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, *14*(1), 48-54. https://doi.org/10.1093/bioinformatics/14.1.48

Bairoch, A., & Apweiler, R. (1997). The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res*, *25*(1), 31-36. https://doi.org/10.1093/nar/25.1.31

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

Balloux, F., & van Dorp, L. (2017). Q&A: What are pathogens, and what have they done to and for us? *BMC Biol*, *15*(1), 91. https://doi.org/10.1186/s12915-017-0433-z

Banerjee, P., Eckert, A. O., Schrey, A. K., & Preissner, R. (2018). ProTox-II: a webserver for the prediction of toxicity of chemicals. *Nucleic Acids Res*, *46*(W1), W257-W263. https://doi.org/10.1093/nar/gky318

Banerjee, P., Siramshetty, V. B., Drwal, M. N., & Preissner, R. (2016). Computational methods for prediction of in vitro effects of new chemical structures. *J Cheminform*, *8*, 51. https://doi.org/10.1186/s13321-016-0162-2

Barbosa, L. C., Garrido, S. S., & Marchetto, R. (2015). BtoxDB: a comprehensive database of protein structural data on toxin-antitoxin systems. *Comput Biol Med*, *58*, 146-153. https://doi.org/10.1016/j.compbiomed.2015.01.010

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Kruger, F. A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., &

Overington, J. P. (2014). The ChEMBL bioactivity database: an update. *Nucleic Acids Res*, *42*(Database issue), D1083-1090. https://doi.org/10.1093/nar/gkt1031

Berne, C., Ducret, A., Hardy, G. G., & Brun, Y. V. (2015). Adhesins Involved in Attachment to Abiotic Surfaces by Gram-Negative Bacteria. *Microbiol Spectr*, *3*(4). https://doi.org/10.1128/microbiolspec.MB-0018-2015

Blackwell, M. (2011). The fungi: 1, 2, 3 ... 5.1 million species? *Am J Bot*, *98*(3), 426-438. https://doi.org/10.3732/ajb.1000298

Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G. A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D. H., Letunic, I., Marchler-Bauer, A., . . . Finn, R. D. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res*, *49*(D1), D344-D354. https://doi.org/10.1093/nar/gkaa977

Bokhari, H., Bilal, I., & Zafar, S. (2012). BapC autotransporter protein of Bordetella pertussis is an adhesion factor. *J Basic Microbiol*, *52*(4), 390-396. https://doi.org/10.1002/jobm.201100188

Broadfield, E., McKeever, T. M., Scrivener, S., Venn, A., Lewis, S. A., & Britton, J. (2002). Increase in the prevalence of allergen skin sensitization in successive birth cohorts. *J Allergy Clin Immunol*, *109*(6), 969-974. https://doi.org/10.1067/mai.2002.124772

Brown, M. R., & Barker, J. (1999). Unexplored reservoirs of pathogenic bacteria: protozoa and biofilms. *Trends Microbiol*, *7*(1), 46-50. https://doi.org/10.1016/s0966-842x(98)01425-5

Bruno, B. J., Miller, G. D., & Lim, C. S. (2013). Basics and recent advances in peptide and protein drug delivery. *Ther Deliv*, *4*(11), 1443-1467. https://doi.org/10.4155/tde.13.104

Burns, D. L., & Manclark, C. R. (1989). Role of cysteine 41 of the A subunit of pertussis toxin. *The Journal of biological chemistry*, *264*(1), 564–568.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Casadevall, A., & Pirofski, L. A. (1999). Host-pathogen interactions: redefining the basic concepts of virulence and pathogenicity. *Infect Immun*, *67*(8), 3703-3713. https://doi.org/10.1128/IAI.67.8.3703-3713.1999

Casadevall, A., & Pirofski, L. A. (2000). Host-pathogen interactions: basic concepts of microbial commensalism, colonization, infection, and disease. *Infect Immun*, *68*(12), 6511-6518. https://doi.org/10.1128/IAI.68.12.6511-6518.2000

Casadevall, A., & Pirofski, L. A. (2009). Virulence factors and their mechanisms of action: the view from a damage-response framework. *J Water Health*, *7 Suppl 1*, S2-S18. https://doi.org/10.2166/wh.2009.036

Casewell, N. R., Jackson, T. N. W., Laustsen, A. H., & Sunagar, K. (2020). Causes and Consequences of Snake Venom Variation. *Trends Pharmacol Sci*, *41*(8), 570-581. https://doi.org/10.1016/j.tips.2020.05.006

CDC. (2021, September 13, 2021). *Infectious Disease*. Retrieved 09 June from https://www.cdc.gov/nchs/fastats/infectious-disease.htm

Chai, L. Y., Netea, M. G., Vonk, A. G., & Kullberg, B. J. (2009). Fungal strategies for overcoming host innate immune response. *Med Mycol*, *47*(3), 227-236. https://doi.org/10.1080/13693780802209082

Chakraborty, A., Ghosh, S., Chowdhary, G., Maulik, U., & Chakrabarti, S. (2012). DBETH: a Database of Bacterial Exotoxins for Human. *Nucleic Acids Res*, *40*(Database issue), D615-620. https://doi.org/10.1093/nar/gkr942

Champion, P. A., Stanley, S. A., Champion, M. M., Brown, E. J., & Cox, J. S. (2006). C-terminal signal sequence promotes virulence factor secretion in Mycobacterium tuberculosis. *Science*, *313*(5793), 1632-1636. https://doi.org/10.1126/science.1131167

Chaudhary, K., Kumar, R., Singh, S., Tuknait, A., Gautam, A., Mathur, D., Anand, P., Varshney, G. C., & Raghava, G. P. (2016). A Web Server and Mobile App for Computing Hemolytic Potency of Peptides. *Sci Rep*, *6*, 22843. https://doi.org/10.1038/srep22843

Chaudhuri, R., Ansari, F. A., Raghunandanan, M. V., & Ramachandran, S. (2011). FungalRV: adhesin prediction and immunoinformatics portal for human fungal pathogens. *BMC genomics*, *12*, 192. https://doi.org/10.1186/1471-2164-12-192

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., & Jin, Q. (2005). VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res*, *33*(Database issue), D325-328. https://doi.org/10.1093/nar/gki008

Clark, G. C., Casewell, N. R., Elliott, C. T., Harvey, A. L., Jamieson, A. G., Strong, P. N., & Turner, A. D. (2019). Friends or Foes? Emerging Impacts of Biological Toxins. *Trends Biochem Sci*, *44*(4), 365-379. https://doi.org/10.1016/j.tibs.2018.12.004

Cole, T. J., & Brewer, M. S. (2019). TOXIFY: a deep learning approach to classify animal venom proteins. *PeerJ*, *7*, e7200. https://doi.org/10.7717/peerj.7200

Commins, S. P., & Platts-Mills, T. A. (2010). Allergenicity of carbohydrates and their role in anaphylactic events. *Curr Allergy Asthma Rep*, *10*(1), 29-33. https://doi.org/10.1007/s11882-009-0079-1

Cordero, O. X., Ventouras, L. A., DeLong, E. F., & Polz, M. F. (2012). Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. *Proc Natl Acad Sci U S A*, *109*(49), 20059-20064. https://doi.org/10.1073/pnas.1213344109

Cross, A. S. (2008). What is a virulence factor? *Crit Care*, *12*(6), 196. https://doi.org/10.1186/cc7127

D'Haeseleer, P. (2006). What are DNA sequence motifs? *Nat Biotechnol*, *24*(4), 423-425. https://doi.org/10.1038/nbt0406-423

Dalle, F., Wachtler, B., L'Ollivier, C., Holland, G., Bannert, N., Wilson, D., Labruere, C., Bonnin, A., & Hube, B. (2010). Cellular interactions of Candida albicans with human oral epithelial cells and enterocytes. *Cell Microbiol*, *12*(2), 248-271. https://doi.org/10.1111/j.1462-5822.2009.01394.x

Dang, H. X., & Lawrence, C. B. (2014). Allerdictor: fast allergen prediction using text classification techniques. *Bioinformatics*, *30*(8), 1120-1128. https://doi.org/10.1093/bioinformatics/btu004

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., Wiegers, J., Wiegers, T. C., & Mattingly, C. J. (2021). Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Res*, *49*(D1), D1138-D1143. https://doi.org/10.1093/nar/gkaa891

Davis, J. J., Wattam, A. R., Aziz, R. K., Brettin, T., Butler, R., Butler, R. M., Chlenski, P., Conrad, N., Dickerman, A., Dietrich, E. M., Gabbard, J. L., Gerdes, S., Guard, A., Kenyon, R. W., Machi, D., Mao, C., Murphy-Olson, D., Nguyen, M., Nordberg, E. K., . . . Stevens, R. (2020). The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res*, *48*(D1), D606-D612. https://doi.org/10.1093/nar/gkz943

De Groot, P. W., Hellingwerf, K. J., & Klis, F. M. (2003). Genome-wide identification of fungal GPI proteins. *Yeast*, *20*(9), 781-796. https://doi.org/10.1002/yea.1007

de Matos, P., Alcantara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., & Steinbeck, C. (2010). Chemical Entities of Biological Interest: an update. *Nucleic Acids Res*, *38*(Database issue), D249-254. https://doi.org/10.1093/nar/gkp886

de Nies, L., Lopes, S., Busi, S. B., Galata, V., Heintz-Buschart, A., Laczny, C. C., May, P., & Wilmes, P. (2021). PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. *Microbiome*, *9*(1), 49. https://doi.org/10.1186/s40168-020-00993-9

Del Moral, M. G., & Martinez-Naves, E. (2017). The Role of Lipids in Development of Allergic Responses. *Immune Netw*, *17*(3), 133-143. https://doi.org/10.4110/in.2017.17.3.133

Del Villar-Guerra, P., Moreno Vicente-Arche, B., Castrillo Bustamante, S., & Santana Rodriguez, C. (2016). Anaphylactic reaction due to cefuroxime axetil: A rare cause of anaphylaxis. *Int J Immunopathol Pharmacol*, *29*(4), 731-733. https://doi.org/10.1177/0394632016664529

Deng, X., Liu, Q., Hu, Y., & Deng, Y. (2013). TOPPER: topology prediction of transmembrane protein based on evidential reasoning. *ScientificWorldJournal*, *2013*, 123731. https://doi.org/10.1155/2013/123731

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

Dimitrov, I., Bangov, I., Flower, D. R., & Doytchinova, I. (2014). AllerTOP v.2--a server for in silico prediction of allergens. *J Mol Model*, *20*(6), 2278. https://doi.org/10.1007/s00894-014-2278-5

Dimitrov, I., Flower, D. R., & Doytchinova, I. (2013). AllerTOP--a server for in silico prediction of allergens. *BMC bioinformatics*, *14 Suppl 6*, S4. https://doi.org/10.1186/1471-2105-14-S6-S4

Dimitrov, I., Naneva, L., Doytchinova, I., & Bangov, I. (2014). AllergenFP: allergenicity prediction by descriptor fingerprints. *Bioinformatics*, *30*(6), 846-851. https://doi.org/10.1093/bioinformatics/btt619

Dinges, M. M., Orwin, P. M., & Schlievert, P. M. (2000). Exotoxins of Staphylococcus aureus. *Clin Microbiol Rev*, *13*(1), 16-34, table of contents. https://doi.org/10.1128/CMR.13.1.16

do Vale, A., Cabanes, D., & Sousa, S. (2016). Bacterial Toxins as Pathogen Weapons Against Phagocytes. *Front Microbiol*, *7*, 42. https://doi.org/10.3389/fmicb.2016.00042

Dou, D., Revol, R., Ostbye, H., Wang, H., & Daniels, R. (2018). Influenza A Virus Cell Entry, Replication, Virion Assembly and Movement. *Front Immunol*, *9*, 1581. https://doi.org/10.3389/fimmu.2018.01581

ECDC. (2022, 24 Mar 2022). *Tuberculosis remains one of the deadliest infectious diseases worldwide, warns new report* https://www.ecdc.europa.eu/en/news-events/tuberculosis-remains-one-deadliest-infectious-diseases-worldwide-warns-new-report

Eisenhut, M., Bauwe, H., & Hagemann, M. (2007). Glycine accumulation is toxic for the cyanobacterium Synechocystis sp. strain PCC 6803, but can be compensated by supplementation with magnesium ions. *FEMS microbiology letters*, *277*(2), 232–237. https://doi.org/10.1111/j.1574-6968.2007.00960.x

Ene, I. V., Brunke, S., Brown, A. J., & Hube, B. (2014). Metabolism in fungal pathogenesis. *Cold Spring Harb Perspect Med*, *4*(12), a019695. https://doi.org/10.1101/cshperspect.a019695

Fanning, S., & Mitchell, A. P. (2012). Fungal biofilms. *PLoS Pathog*, *8*(4), e1002585. https://doi.org/10.1371/journal.ppat.1002585

FAO/WHO. (2001). *Evaluation of allergenicity of genetically modified foods.Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from*

Biotechnology.
http://www.fao.org/fileadmin/templates/agns/pdf/topics/ec_jan2001.pdf

FAO/WHO. (2003). *Joint FAO/WHO Food Standards Programme Codex Alimentarius Commission. Report of the Fourth Session of the Codex Ad Hoc Intergovernmental Task Force on Foods Derived from Biotechnology,* . http://www.fao.org/fileadmin/user_upload/gmfp/resources/al0334ae.pdf

Finlay, B. B., & McFadden, G. (2006). Anti-immunology: evasion of the host immune system by bacterial and viral pathogens. *Cell*, *124*(4), 767-782. https://doi.org/10.1016/j.cell.2006.01.034

Fleury, B., Bergonier, D., Berthelot, X., Peterhans, E., Frey, J., & Vilei, E. M. (2002). Characterization of P40, a cytadhesin of Mycoplasma agalactiae. *Infect Immun*, *70*(10), 5612-5621. https://doi.org/10.1128/IAI.70.10.5612-5621.2002

Gacesa R, B. D., Long PF. (2016). Machine learning can differentiate venom toxins from other proteins having non-toxic physiological functions. *PeerJ Computer Science*, *2:e90*. https://doi.org/https://doi.org/10.7717/peerj-cs.90

Gallo, R. L., & Hooper, L. V. (2012). Epithelial antimicrobial defence of the skin and intestine. *Nat Rev Immunol*, *12*(7), 503-516. https://doi.org/10.1038/nri3228

Garg, A., & Gupta, D. (2008). VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC bioinformatics*, *9*, 62. https://doi.org/10.1186/1471-2105-9-62

Gautam, A., Chaudhary, K., Singh, S., Joshi, A., Anand, P., Tuknait, A., Mathur, D., Varshney, G. C., & Raghava, G. P. (2014). Hemolytik: a database of experimentally determined hemolytic and non-hemolytic peptides. *Nucleic Acids Res*, *42*(Database issue), D444-449. https://doi.org/10.1093/nar/gkt1008

Gautam, A., Singh, H., Tyagi, A., Chaudhary, K., Kumar, R., Kapoor, P., & Raghava, G. P. (2012). CPPsite: a curated database of cell penetrating peptides. *Database (Oxford)*, *2012*, bas015. https://doi.org/10.1093/database/bas015

George, E. K., De Jesus, O., & Vivekanandan, R. (2022). Clostridium Tetani. In *StatPearls*. https://www.ncbi.nlm.nih.gov/pubmed/29494091

Ghislain, P. D., Bodarwe, A. D., Vanderdonckt, O., Tennstedt, D., Marot, L., & Lachapelle, J. M. (2004). Drug-induced eosinophilia and multisystemic failure with positive patch-test reaction to spironolactone: DRESS syndrome. *Acta Derm Venereol*, *84*(1), 65-68. https://doi.org/10.1080/00015550310005915

Goodman, R. E., Ebisawa, M., Ferreira, F., Sampson, H. A., van Ree, R., Vieths, S., Baumert, J. L., Bohle, B., Lalithambika, S., Wise, J., & Taylor, S. L. (2016). AllergenOnline: A peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reactivity. *Mol Nutr Food Res*, *60*(5), 1183-1198. https://doi.org/10.1002/mnfr.201500769

Goodman, R. E., Hefle, S. L., Taylor, S. L., & van Ree, R. (2005). Assessing genetically modified crops to minimize the risk of increased food allergy: a review. *Int Arch Allergy Immunol*, *137*(2), 153-166. https://doi.org/10.1159/000086314

Green, A. E., Howarth, D., Chaguza, C., Echlin, H., Langendonk, R. F., Munro, C., Barton, T. E., Hinton, J. C. D., Bentley, S. D., Rosch, J. W., & Neill, D. R. (2021). Pneumococcal Colonization and Virulence Factors Identified Via Experimental Evolution in Infection Models. *Mol Biol Evol*, *38*(6), 2209-2226. https://doi.org/10.1093/molbev/msab018

Gu, J., Liu, S., & Zhi, Y. (2019). Cefuroxime-induced anaphylaxis with prominent central nervous system manifestations: A case report. *J Int Med Res*, *47*(2), 1010-1014. https://doi.org/10.1177/0300060518814118

Gupta, A., Kapil, R., Dhakan, D. B., & Sharma, V. K. (2014). MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. *PLoS ONE*, *9*(4), e93907. https://doi.org/10.1371/journal.pone.0093907

Gupta, S., Ansari, H. R., Gautam, A., Open Source Drug Discovery, C., & Raghava, G. P. (2013). Identification of B-cell epitopes in an antigen for inducing specific class of antibodies. *Biol Direct*, *8*, 27. https://doi.org/10.1186/1745-6150-8-27

Gupta, S., Kapoor, P., Chaudhary, K., Gautam, A., Kumar, R., Open Source Drug Discovery, C., & Raghava, G. P. (2013). In silico approach for predicting toxicity of peptides and proteins. *PLoS ONE*, *8*(9), e73957. https://doi.org/10.1371/journal.pone.0073957

Gupta, S., Kapoor, P., Chaudhary, K., Gautam, A., Kumar, R., & Raghava, G. P. (2015). Peptide toxicity prediction. *Methods Mol Biol*, *1268*, 143-157. https://doi.org/10.1007/978-1-4939-2285-7_7

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., & Steinbeck, C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res*, *44*(D1), D1214-1219. https://doi.org/10.1093/nar/gkv1031

Heikkila, H., Stubb, S., & Reitamo, S. (1996). A study of 72 patients with contact allergy to tioconazole. *Br J Dermatol*, *134*(4), 678-680. https://doi.org/10.1111/j.1365-2133.1996.tb06969.x

Hilleman, M. R. (2004). Strategies and mechanisms for host and pathogen survival in acute and persistent viral infections. *Proc Natl Acad Sci U S A*, *101 Suppl 2*, 14560-14566. https://doi.org/10.1073/pnas.0404758101

Hiller, K., Grote, A., Scheer, M., Munch, R., & Jahn, D. (2004). PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res*, *32*(Web Server issue), W375-379. https://doi.org/10.1093/nar/gkh378

Hirst, R. A., Kadioglu, A., O'Callaghan, C., & Andrew, P. W. (2004). The role of pneumolysin in pneumococcal pneumonia and meningitis. *Clin Exp Immunol*, *138*(2), 195-201. https://doi.org/10.1111/j.1365-2249.2004.02611.x

Honigschmid, P., Breimann, S., Weigl, M., & Frishman, D. (2020). AllesTM: predicting multiple structural features of transmembrane proteins. *BMC bioinformatics*, *21*(1), 242. https://doi.org/10.1186/s12859-020-03581-8

Hornef, M. W., Wick, M. J., Rhen, M., & Normark, S. (2002). Bacterial strategies for overcoming host innate and adaptive immune responses. *Nat Immunol*, *3*(11), 1033-1040. https://doi.org/10.1038/ni1102-1033

Houston, S., Hof, R., Francescutti, T., Hawkes, A., Boulanger, M. J., & Cameron, C. E. (2011). Bifunctional role of the Treponema pallidum extracellular matrix binding adhesin Tp0751. *Infect Immun*, *79*(3), 1386-1398. https://doi.org/10.1128/IAI.01083-10

Ibberson, C. B., Jones, C. L., Singh, S., Wise, M. C., Hart, M. E., Zurawski, D. V., & Horswill, A. R. (2014). Staphylococcus aureus hyaluronidase is a CodY-regulated virulence factor. *Infect Immun*, *82*(10), 4253-4264. https://doi.org/10.1128/IAI.01710-14

InformedHealth.org. (2006). *What are microbes?* Institute for Quality and Efficiency in Health Care (IQWiG), Cologne, Germany. https://www.ncbi.nlm.nih.gov/books/NBK279387/

Ivanciuc, O., Schein, C. H., & Braun, W. (2003). SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res*, *31*(1), 359-362. https://doi.org/10.1093/nar/gkg010

Jain, A., & Kihara, D. (2019). NNTox: Gene Ontology-Based Protein Toxicity Prediction Using Neural Network. *Sci Rep*, *9*(1), 17923. https://doi.org/10.1038/s41598-019-54405-6

Kadam, K., Karbhal, R., Jayaraman, V. K., Sawant, S., & Kulkarni-Kale, U. (2017). AllerBase: a comprehensive allergen knowledgebase. *Database (Oxford)*, *2017*. https://doi.org/10.1093/database/bax066

Kapoor, P., Singh, H., Gautam, A., Chaudhary, K., Kumar, R., & Raghava, G. P. (2012). TumorHoPe: a database of tumor homing peptides. *PLoS ONE*, *7*(4), e35187. https://doi.org/10.1371/journal.pone.0035187

Kaur, D., Patiyal, S., Sharma, N., Usmani, S. S., & Raghava, G. P. S. (2019). PRRDB 2.0: a comprehensive database of pattern-recognition receptors and their ligands. *Database (Oxford)*, *2019*. https://doi.org/10.1093/database/baz076

Kinsara AJ (2018). Spironolactone - Induced Rash: A Case Report and Review. *Journal of Clinical Cardiology and Diagnostics*, *Volume 1*(Issue 2), 1-2.

Kim, S. H., Park, S. D., Baek, Y. S., Lee, S. Y., Shin, S. H., Woo, S. I., Kim, D. H., & Kwan, J. (2014). Prasugrel-induced hypersensitivity skin reaction. *Korean Circ J*, *44*(5), 355-357. https://doi.org/10.4070/kcj.2014.44.5.355

Kimber, I., Basketter, D. A., & Dearman, R. J. (2010). Chemical allergens--what are the issues? *Toxicology*, *268*(3), 139-142. https://doi.org/10.1016/j.tox.2009.07.015

Kimber, I., Basketter, D. A., Gerberick, G. F., Ryan, C. A., & Dearman, R. J. (2011). Chemical allergy: translating biology into hazard characterization. *Toxicol Sci*, *120 Suppl 1*, S238-268. https://doi.org/10.1093/toxsci/kfq346

Kishor, P., Suravajhala, R., Rajasheker, G., Marka, N., Shridhar, K. K., Dhulala, D., Scinthia, K. P., Divya, K., Doma, M., Edupuganti, S., Suravajhala, P., & Polavarapu, R. (2020). Lysine, Lysine-Rich, Serine, and Serine-Rich Proteins: Link Between Metabolism, Development, and Abiotic Stress Tolerance and the Role of ncRNAs in Their Regulation. *Frontiers in plant science*, *11*, 546213. https://doi.org/10.3389/fpls.2020.546213

Konkel, M. E., Christensen, J. E., Keech, A. M., Monteville, M. R., Klena, J. D., & Garvis, S. G. (2005). Identification of a fibronectin-binding domain within the Campylobacter jejuni CadF protein. *Mol Microbiol*, *57*(4), 1022-1035. https://doi.org/10.1111/j.1365-2958.2005.04744.x

Kramer, J., Ozkaya, O., & Kummerli, R. (2020). Bacterial siderophores in community and host interactions. *Nat Rev Microbiol*, *18*(3), 152-163. https://doi.org/10.1038/s41579-019-0284-4

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, *305*(3), 567-580. https://doi.org/10.1006/jmbi.2000.4315

Kumar, R., Chaudhary, K., Singh Chauhan, J., Nagpal, G., Kumar, R., Sharma, M., & Raghava, G. P. (2015). An in silico platform for predicting, screening and designing of antihypertensive peptides. *Sci Rep*, *5*, 12512. https://doi.org/10.1038/srep12512

Lata, S., & Raghava, G. P. (2008). PRRDB: a comprehensive database of pattern-recognition receptors and their ligands. *BMC genomics*, *9*, 180. https://doi.org/10.1186/1471-2164-9-180

Latge, J. P., & Beauvais, A. (2014). Functional duality of the cell wall. *Curr Opin Microbiol*, *20*, 111-117. https://doi.org/10.1016/j.mib.2014.05.009

Lee, V. T., & Schneewind, O. (2001). Protein secretion and the pathogenesis of bacterial infections. *Genes Dev*, *15*(14), 1725-1752. https://doi.org/10.1101/gad.896801

Li, N., Yun, P., Nadkarni, M. A., Ghadikolaee, N. B., Nguyen, K. A., Lee, M., Hunter, N., & Collyer, C. A. (2010). Structure determination and analysis of a haemolytic gingipain adhesin domain from Porphyromonas gingivalis. *Mol Microbiol*, *76*(4), 861-873. https://doi.org/10.1111/j.1365-2958.2010.07123.x

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

Life Sciences Open Source, Ketcher 2.0 (2021). https://lifescience.opensource.epam.com/ketcher/index.html#ketcher-2-0

Lim, E., Pon, A., Djoumbou, Y., Knox, C., Shrivastava, S., Guo, A. C., Neveu, V., & Wishart, D. S. (2010). T3DB: a comprehensively annotated database of common toxins and their targets. *Nucleic Acids Res*, *38*(Database issue), D781-786. https://doi.org/10.1093/nar/gkp934

Linder, T., & Gustafsson, C. M. (2008). Molecular phylogenetics of ascomycotal adhesins--a novel family of putative cell-surface adhesive proteins in fission yeasts. *Fungal Genet Biol*, *45*(4), 485-497. https://doi.org/10.1016/j.fgb.2007.08.002

Liu, B., Zheng, D., Jin, Q., Chen, L., & Yang, J. (2019). VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res*, *47*(D1), D687-D692. https://doi.org/10.1093/nar/gky1080

Los, F. C., Randis, T. M., Aroian, R. V., & Ratner, A. J. (2013). Role of pore-forming toxins in bacterial infectious diseases. *Microbiol Mol Biol Rev*, *77*(2), 173-207. https://doi.org/10.1128/MMBR.00052-12

Lu, T., Yao, B., & Zhang, C. (2012). DFVF: database of fungal virulence factors. *Database (Oxford)*, *2012*, bas032. https://doi.org/10.1093/database/bas032

Mak TW, S. M., Jett BD. (2014). *Immune hypersensitivity. In: Primer to the Immune Response.* https://www.elsevier.com/books/primer-to-the-immune-response/mak/978-0-12-385245-8

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M. K., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lanczycki, C. J., Lu, F., Lu, S., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Zhang, D., & Bryant, S. H. (2013). CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res*, *41*(Database issue), D348-352. https://doi.org/10.1093/nar/gks1243

Masoli, M., Fabian, D., Holt, S., Beasley, R., & Global Initiative for Asthma, P. (2004). The global burden of asthma: executive summary of the GINA Dissemination Committee report. *Allergy*, *59*(5), 469-478. https://doi.org/10.1111/j.1398-9995.2004.00526.x

Mathur, D., Singh, S., Mehta, A., Agrawal, P., & Raghava, G. P. S. (2018). In silico approaches for predicting the half-life of natural and modified peptides in blood. *PLoS ONE*, *13*(6), e0196829. https://doi.org/10.1371/journal.pone.0196829

Matrosovich, M., Herrler, G., & Klenk, H. D. (2015). Sialic Acid Receptors of Viruses. *Top Curr Chem*, *367*, 1-28. https://doi.org/10.1007/128_2013_466

Maurer-Stroh, S., Krutz, N. L., Kern, P. S., Gunalan, V., Nguyen, M. N., Limviphuvadh, V., Eisenhaber, F., & Gerberick, G. F. (2019). AllerCatPro-prediction of protein allergenicity potential from the protein sequence. *Bioinformatics*, *35*(17), 3020-3027. https://doi.org/10.1093/bioinformatics/btz029

Mayer, F. L., Wilson, D., & Hube, B. (2013). Candida albicans pathogenicity mechanisms. *Virulence*, *4*(2), 119-128. https://doi.org/10.4161/viru.22913

Mayr A, K. G., Unterthiner T and Hochreiter S. (2016). DeepTox: Toxicity Prediction using Deep Learning. *Front. Environ. Sci.*, *3*(80). https://doi.org/10.3389/fenvs.2015.00080

Minasyan, H. (2019). Sepsis: mechanisms of bacterial injury to the patient. *Scand J Trauma Resusc Emerg Med*, *27*(1), 19. https://doi.org/10.1186/s13049-019-0596-4

Mogensen, T. H. (2009). Pathogen recognition and inflammatory signaling in innate immune defenses. *Clin Microbiol Rev*, *22*(2), 240-273, Table of Contents. https://doi.org/10.1128/CMR.00046-08

Molleken, K., Schmidt, E., & Hegemann, J. H. (2010). Members of the Pmp protein family of Chlamydia pneumoniae mediate adhesion to human cells via short repetitive peptide motifs. *Mol Microbiol*, *78*(4), 1004-1017. https://doi.org/10.1111/j.1365-2958.2010.07386.x

Molloy, E. M., Cotter, P. D., Hill, C., Mitchell, D. A., & Ross, R. P. (2011). Streptolysin S-like virulence factors: the continuing sagA. *Nat Rev Microbiol*, *9*(9), 670-681. https://doi.org/10.1038/nrmicro2624

Muh, H. C., Tong, J. C., & Tammi, M. T. (2009). AllerHunter: a SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins. *PLoS ONE*, *4*(6), e5861. https://doi.org/10.1371/journal.pone.0005861

Mukai, K., Tsai, M., Saito, H., & Galli, S. J. (2018). Mast cells as sources of cytokines, chemokines, and growth factors. *Immunol Rev*, *282*(1), 121-150. https://doi.org/10.1111/imr.12634

Murphy, J. R. (1996). Corynebacterium Diphtheriae. In th & S. Baron (Eds.), *Medical Microbiology*. https://www.ncbi.nlm.nih.gov/pubmed/21413281

Naamati, G., Askenazi, M., & Linial, M. (2009). ClanTox: a classifier of short animal toxins. *Nucleic Acids Res*, *37*(Web Server issue), W363-368. https://doi.org/10.1093/nar/gkp299

Nakane, D., Adan-Kubo, J., Kenri, T., & Miyata, M. (2011). Isolation and characterization of P1 adhesin, a leg protein of the gliding bacterium Mycoplasma pneumoniae. *J Bacteriol*, *193*(3), 715-722. https://doi.org/10.1128/JB.00796-10

Nam, Y. H., Hwang, E. K., Ban, G. Y., Jin, H. J., Yoo, H. S., Shin, Y. S., Ye, Y. M., Nahm, D. H., Park, H. S., & Lee, S. K. (2015). Immunologic evaluation of patients with cefotetan-induced anaphylaxis. *Allergy Asthma Immunol Res*, *7*(3), 301-303. https://doi.org/10.4168/aair.2015.7.3.301

Nigam, P. K., & Nigam, A. (2010). Botulinum toxin. *Indian J Dermatol*, *55*(1), 8-14. https://doi.org/10.4103/0019-5154.60343

Nordengrun, M., Michalik, S., Volker, U., Broker, B. M., & Gomez-Gascon, L. (2018). The quest for bacterial allergens. *Int J Med Microbiol*, *308*(6), 738-750. https://doi.org/10.1016/j.ijmm.2018.04.003

Nummelin, H., Merckel, M. C., Leo, J. C., Lankinen, H., Skurnik, M., & Goldman, A. (2004). The Yersinia adhesin YadA collagen-binding domain structure is a novel left-handed parallel beta-roll. *EMBO J*, *23*(4), 701-711. https://doi.org/10.1038/sj.emboj.7600100

Obermeyer, G., & Ferreira, F. (2005). Can we predict or avoid the allergenic potential of genetically modified organisms? *Int Arch Allergy Immunol*, *137*(2), 151-152. https://doi.org/10.1159/000086313

Otvos, L., Jr., & Wade, J. D. (2014). Current challenges in peptide-based drug discovery. *Front Chem*, *2*, 62. https://doi.org/10.3389/fchem.2014.00062

Palaniappan, R. U., Chang, Y. F., Jusuf, S. S., Artiushin, S., Timoney, J. F., McDonough, S. P., Barr, S. C., Divers, T. J., Simpson, K. W., McDonough, P. L., & Mohammed, H. O. (2002). Cloning and molecular characterization of an immunogenic LigA protein of Leptospira interrogans. *Infect Immun*, *70*(11), 5924-5930. https://doi.org/10.1128/IAI.70.11.5924-5930.2002

Pan, X., Zuallaert, J., Wang, X., Shen, H. B., Campos, E. P., Marushchak, D. O., & De Neve, W. (2021). ToxDL: deep learning using primary structure and domain embeddings for assessing protein toxicity. *Bioinformatics*, *36*(21), 5159-5168. https://doi.org/10.1093/bioinformatics/btaa656

Pande A, P. S., Lathwal A, et al. (2019). Computing wide range of protein/peptide features from their sequence and structure. *bioRxiv*.

Pasechnik, V. A., Shone, C. C., & Hambleton, P. (1992). Purification of bacterial exotoxins. The case of botulinum, tetanus, anthrax, pertussis and cholera toxins. *Bioseparation*, *3*(5), 267-283. https://www.ncbi.nlm.nih.gov/pubmed/1369426

Patlewicz, G., Jeliazkova, N., Safford, R. J., Worth, A. P., & Aleksiev, B. (2008). An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. *SAR QSAR Environ Res*, *19*(5-6), 495-524. https://doi.org/10.1080/10629360802083871

Pedregosa F, V. G., Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830. http://jmlr.org/papers/v12/pedregosa11a.html

Pérez Santín E, R. g. S. R., González García M, et al. . (2021). Toxicity prediction based on artificial intelligence: A multidisciplinary overview. Overview. *WIREs Comput Mol Sci. 2021;e1516.*, *11*(5). https://doi.org/ https://doi.org/10.1002/wcms.1516

Peterson, J. W. (1996). Bacterial Pathogenesis. In th & S. Baron (Eds.), *Medical Microbiology*. https://www.ncbi.nlm.nih.gov/pubmed/21413346

Petricevich, V. L. (2010). Scorpion venom and the inflammatory response. *Mediators Inflamm*, *2010*, 903295. https://doi.org/10.1155/2010/903295

Phan, Q. T., Myers, C. L., Fu, Y., Sheppard, D. C., Yeaman, M. R., Welch, W. H., Ibrahim, A. S., Edwards, J. E., Jr., & Filler, S. G. (2007). Als3 is a Candida albicans invasin that binds to cadherins and induces endocytosis by host cells. *PLoS Biol*, *5*(3), e64. https://doi.org/10.1371/journal.pbio.0050064

Phillips, D. C. (1966). The three-dimensional structure of an enzyme molecule. *Sci Am*, *215*(5), 78-90. https://doi.org/10.1038/scientificamerican1166-78

Pichel, M., Binsztein, N., & Viboud, G. (2000). CS22, a novel human enterotoxigenic Escherichia coli adhesin, is related to CS15. *Infect Immun*, *68*(6), 3280-3285. https://doi.org/10.1128/IAI.68.6.3280-3285.2000

Piglowski, M. (2019). Pathogenic and Non-Pathogenic Microorganisms in the Rapid Alert System for Food and Feed. *Int J Environ Res Public Health*, *16*(3). https://doi.org/10.3390/ijerph16030477

Pizarro-Cerda, J., & Cossart, P. (2006). Bacterial adhesion and entry into host cells. *Cell*, *124*(4), 715-727. https://doi.org/10.1016/j.cell.2006.02.012

Pomerantsev, A. P., Kalnin, K. V., Osorio, M., & Leppla, S. H. (2003). Phosphatidylcholine-specific phospholipase C and sphingomyelinase activities in bacteria of the Bacillus cereus group. *Infect Immun*, *71*(11), 6591-6606. https://doi.org/10.1128/IAI.71.11.6591-6606.2003

Popoff, M. R. (2018). "Bacterial Toxins" Section in the Journal Toxins: A Fantastic Multidisciplinary Interplay between Bacterial Pathogenicity Mechanisms, Physiological Processes, Genomic Evolution, and Subsequent Development of Identification Methods, Efficient Treatment, and Prevention of Toxigenic Bacteria. *Toxins (Basel)*, *10*(1). https://doi.org/10.3390/toxins10010044

Popugailo, A., Rotfogel, Z., Supper, E., Hillman, D., & Kaempfer, R. (2019). Staphylococcal and Streptococcal Superantigens Trigger B7/CD28 Costimulatory Receptor Engagement to Hyperinduce Inflammatory Cytokines. *Front Immunol*, *10*, 942. https://doi.org/10.3389/fimmu.2019.00942

Poulin, R., & Combes, C. (1999). The concept of virulence: interpretations and implications. *Parasitol Today*, *15*(12), 474-475. https://doi.org/10.1016/s0169-4758(99)01554-9

Proft, T., & Fraser, J. D. (2003). Bacterial superantigens. *Clin Exp Immunol*, *133*(3), 299-306. https://doi.org/10.1046/j.1365-2249.2003.02203.x

Pu, L., Naderi, M., Liu, T., Wu, H. C., Mukhopadhyay, S., & Brylinski, M. (2019). eToxPred: a machine learning-based approach to estimate the toxicity of drug candidates. *BMC Pharmacol Toxicol*, *20*(1), 2. https://doi.org/10.1186/s40360-018-0282-6

Qureshi, A., Thakur, N., Tandon, H., & Kumar, M. (2014). AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic Acids Res*, *42*(Database issue), D1147-1153. https://doi.org/10.1093/nar/gkt1191

Radauer, C., Bublin, M., Wagner, S., Mari, A., & Breiteneder, H. (2008). Allergens are distributed into few protein families and possess a restricted number of biochemical functions. *J Allergy Clin Immunol*, *121*(4), 847-852 e847. https://doi.org/10.1016/j.jaci.2008.01.025

Raetz, C. R., & Whitfield, C. (2002). Lipopolysaccharide endotoxins. *Annu Rev Biochem*, *71*, 635-700. https://doi.org/10.1146/annurev.biochem.71.110601.135414

Rajamuthiah, R., & Mylonakis, E. (2014). Effector triggered immunity. *Virulence*, *5*(7), 697-702. https://doi.org/10.4161/viru.29091

Ramana, J., & Gupta, D. (2009). ProtVirDB: a database of protozoan virulent proteins. *Bioinformatics*, *25*(12), 1568-1569. https://doi.org/10.1093/bioinformatics/btp258

Ramana, J., & Gupta, D. (2010). FaaPred: a SVM-based prediction method for fungal adhesins and adhesin-like proteins. *PLoS ONE*, *5*(3), e9695. https://doi.org/10.1371/journal.pone.0009695

Rasko, D. A., & Sperandio, V. (2010). Anti-virulence strategies to combat bacteria-mediated disease. *Nat Rev Drug Discov*, *9*(2), 117-128. https://doi.org/10.1038/nrd3013

Reidl, J., & Klose, K. E. (2002). Vibrio cholerae and cholera: out of the water and into the host. *FEMS Microbiol Rev*, *26*(2), 125-139. https://doi.org/10.1111/j.1574-6976.2002.tb00605.x

Renesto, P., Samson, L., Ogata, H., Azza, S., Fourquet, P., Gorvel, J. P., Heinzen, R. A., & Raoult, D. (2006). Identification of two putative rickettsial adhesins by proteomic analysis. *Res Microbiol*, *157*(7), 605-612. https://doi.org/10.1016/j.resmic.2006.02.002

Ribet, D., & Cossart, P. (2015). How bacterial pathogens colonize their hosts and invade deeper tissues. *Microbes Infect*, *17*(3), 173-183. https://doi.org/10.1016/j.micinf.2015.01.004

Richard, A. M., & Williams, C. R. (2002). Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. *Mutat Res*, *499*(1), 27-52. https://doi.org/10.1016/s0027-5107(01)00289-5

Richard, M. L., & Plaine, A. (2007). Comprehensive analysis of glycosylphosphatidylinositol-anchored proteins in Candida albicans. *Eukaryot Cell*, *6*(2), 119-133. https://doi.org/10.1128/EC.00297-06

Rigden, D. J., Mello, L. V., & Galperin, M. Y. (2004). The PA14 domain, a conserved all-beta domain in bacterial toxins, enzymes, adhesins and signaling molecules. *Trends Biochem Sci*, *29*(7), 335-339. https://doi.org/10.1016/j.tibs.2004.05.002

Rittershaus, E. S., Baek, S. H., & Sassetti, C. M. (2013). The normalcy of dormancy: common themes in microbial quiescence. *Cell Host Microbe*, *13*(6), 643-651. https://doi.org/10.1016/j.chom.2013.05.012

Ross-Gillespie, A., Dumas, Z., & Kummerli, R. (2015). Evolutionary dynamics of interlinked public goods traits: an experimental study of siderophore production in Pseudomonas aeruginosa. *J Evol Biol*, *28*(1), 29-39. https://doi.org/10.1111/jeb.12559

Rubin, R. (2021). Allergic Reactions to mRNA Vaccines. *JAMA*, *325*(20), 2038. https://doi.org/10.1001/jama.2021.6941

Russ, A. P., & Lampel, S. (2005). The druggable genome: an update. *Drug Discov Today*, *10*(23-24), 1607-1610. https://doi.org/10.1016/S1359-6446(05)03666-4

Sachdeva, G., Kumar, K., Jain, P., & Ramachandran, S. (2005). SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. *Bioinformatics*, *21*(4), 483-491. https://doi.org/10.1093/bioinformatics/bti028

Saha, S., & Raghava, G. P. (2006a). AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res*, *34*(Web Server issue), W202-209. https://doi.org/10.1093/nar/gkl343

Saha, S., & Raghava, G. P. (2006b). VICMpred: an SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition. *Genomics Proteomics Bioinformatics*, *4*(1), 42-47. https://doi.org/10.1016/S1672-0229(06)60015-6

Saha, S., & Raghava, G. P. (2007a). BTXpred: prediction of bacterial toxins. *In Silico Biol*, *7*(4-5), 405-412. https://www.ncbi.nlm.nih.gov/pubmed/18391233

Saha, S., & Raghava, G. P. (2007b). Prediction of neurotoxins based on their function and source. *In Silico Biol*, *7*(4-5), 369-387. https://www.ncbi.nlm.nih.gov/pubmed/18391230

Sathe, S., Mathew, A., Agnoli, K., Eberl, L., & Kummerli, R. (2019). Genetic architecture constrains exploitation of siderophore cooperation in the bacterium Burkholderia cenocepacia. *Evol Lett*, *3*(6), 610-622. https://doi.org/10.1002/evl3.144

Sayers, S., Li, L., Ong, E., Deng, S., Fu, G., Lin, Y., Yang, B., Zhang, S., Fa, Z., Zhao, B., Xiang, Z., Li, Y., Zhao, X. M., Olszewski, M. A., Chen, L., & He, Y. (2019). Victors: a web-based knowledge base of virulence factors in human and animal pathogens. *Nucleic Acids Res*, *47*(D1), D693-D700. https://doi.org/10.1093/nar/gky999

Scharf, D. H., Heinekamp, T., & Brakhage, A. A. (2014). Human and plant fungal pathogens: the role of secondary metabolites. *PLoS Pathog*, *10*(1), e1003859. https://doi.org/10.1371/journal.ppat.1003859

Schmidt, U., Struck, S., Gruening, B., Hossbach, J., Jaeger, I. S., Parol, R., Lindequist, U., Teuscher, E., & Preissner, R. (2009). SuperToxic: a comprehensive database of toxic compounds. *Nucleic Acids Res*, *37*(Database issue), D295-299. https://doi.org/10.1093/nar/gkn850

Scholz, R. L., & Greenberg, E. P. (2015). Sociality in Escherichia coli: Enterochelin Is a Private Good at Low Cell Density and Can Be Shared at High Cell Density. *J Bacteriol*, *197*(13), 2122-2128. https://doi.org/10.1128/JB.02596-14

Seilie, E. S., & Bubeck Wardenburg, J. (2017). Staphylococcus aureus pore-forming toxins: The interface of pathogen and host complexity. *Semin Cell Dev Biol*, *72*, 101-116. https://doi.org/10.1016/j.semcdb.2017.04.003

Senczuk, A. M., Reeder, J. C., Kosmala, M. M., & Ho, M. (2001). Plasmodium falciparum erythrocyte membrane protein 1 functions as a ligand for P-selectin. *Blood*, *98*(10), 3132-3135. https://doi.org/10.1182/blood.v98.10.3132

Shaji, J., & Patole, V. (2008). Protein and Peptide drug delivery: oral approaches. *Indian J Pharm Sci*, *70*(3), 269-277. https://doi.org/10.4103/0250-474X.42967

Shapiro-Ilan, D. I., Fuxa, J. R., Lacey, L. A., Onstad, D. W., & Kaya, H. K. (2005). Definitions of pathogenicity and virulence in invertebrate pathology. *J Invertebr Pathol*, *88*(1), 1-7. https://doi.org/10.1016/j.jip.2004.10.003

Sharma, A. K., Dhasmana, N., Dubey, N., Kumar, N., Gangwal, A., Gupta, M., & Singh, Y. (2017). Bacterial Virulence Factors: Secreted for Survival. *Indian J Microbiol*, *57*(1), 1-10. https://doi.org/10.1007/s12088-016-0625-1

Sharma, A. K., Srivastava, G. N., Roy, A., & Sharma, V. K. (2017). ToxiM: A Toxicity Prediction Tool for Small Molecules Developed Using Machine Learning and Chemoinformatics Approaches. *Front Pharmacol*, *8*, 880. https://doi.org/10.3389/fphar.2017.00880

Sharma, N., Naorem, L. D., Jain, S., & Raghava, G. P. S. (2022). ToxinPred2: an improved method for predicting toxicity of proteins. *Brief Bioinform*. https://doi.org/10.1093/bib/bbac174

Sharma N, N. L., Gupta S, et al. . (2021). Computational resources in healthcare. *WIREs Data Min. Knowl Discov*, *e1437*. https://doi.org/https://doi.org/10.1002/widm.1437

Sharma, N., Patiyal, S., Dhall, A., Devi, N. L., & Raghava, G. P. S. (2021). ChAlPred: A web server for prediction of allergenicity of chemical compounds. *Comput Biol Med*, *136*, 104746. https://doi.org/10.1016/j.compbiomed.2021.104746

Sharma, N., Patiyal, S., Dhall, A., Pande, A., Arora, C., & Raghava, G. P. S. (2020). AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. *Brief Bioinform*. https://doi.org/10.1093/bib/bbaa294

Silver, L. L. (2011). Challenges of antibacterial discovery. *Clin Microbiol Rev*, *24*(1), 71-109. https://doi.org/10.1128/CMR.00030-10

Singh, S., Chaudhary, K., Dhanda, S. K., Bhalla, S., Usmani, S. S., Gautam, A., Tuknait, A., Agrawal, P., Mathur, D., & Raghava, G. P. (2016). SATPdb: a database of structurally annotated therapeutic peptides. *Nucleic Acids Res*, *44*(D1), D1119-1126. https://doi.org/10.1093/nar/gkv1114

Slagboom, J., Kool, J., Harrison, R. A., & Casewell, N. R. (2017). Haemotoxic snake venoms: their functional activity, impact on snakebite victims and pharmaceutical promise. *Br J Haematol*, *177*(6), 947-959. https://doi.org/10.1111/bjh.14591

Snyder, E. E., Kampanya, N., Lu, J., Nordberg, E. K., Karur, H. R., Shukla, M., Soneja, J., Tian, Y., Xue, T., Yoo, H., Zhang, F., Dharmanolla, C., Dongre, N. V., Gillespie, J. J., Hamelius, J., Hance, M., Huntington, K. I., Jukneliene, D., Koziski, J., . . . Sobral, B. W. (2007). PATRIC: the VBI PathoSystems Resource Integration Center. *Nucleic Acids Res*, *35*(Database issue), D401-406. https://doi.org/10.1093/nar/gkl858

Spaan, A. N., van Strijp, J. A. G., & Torres, V. J. (2017). Leukocidins: staphylococcal bi-component pore-forming toxins find their receptors. *Nat Rev Microbiol*, *15*(7), 435-447. https://doi.org/10.1038/nrmicro.2017.27

Spaulding, A. R., Salgado-Pabon, W., Kohler, P. L., Horswill, A. R., Leung, D. Y., & Schlievert, P. M. (2013). Staphylococcal and streptococcal superantigen exotoxins. *Clin Microbiol Rev*, *26*(3), 422-447. https://doi.org/10.1128/CMR.00104-12

Sullan, R. M., Li, J. K., Crowley, P. J., Brady, L. J., & Dufrene, Y. F. (2015). Binding forces of Streptococcus mutans P1 adhesin. *ACS Nano*, *9*(2), 1448-1460. https://doi.org/10.1021/nn5058886

Sutton, B. J., & Gould, H. J. (1993). The human IgE network. *Nature*, *366*(6454), 421-428. https://doi.org/10.1038/366421a0

Tang J, A. S., Liu H. (2014). Feature selection for classification: a review. *Data Classif Algorithms Appl*, *37*, 1871-1874.

Teufel, F., Almagro Armenteros, J. J., Johansen, A. R., Gislason, M. H., Pihl, S. I., Tsirigos, K. D., Winther, O., Brunak, S., von Heijne, G., & Nielsen, H. (2022). SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat Biotechnol*. https://doi.org/10.1038/s41587-021-01156-3

Titball, R. W. (1993). Bacterial phospholipases C. *Microbiol Rev*, *57*(2), 347-366. https://doi.org/10.1128/mr.57.2.347-366.1993

Tsirigos, K. D., Peters, C., Shu, N., Kall, L., & Elofsson, A. (2015). The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res*, *43*(W1), W401-407. https://doi.org/10.1093/nar/gkv485

UniProt, C. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*, *49*(D1), D480-D489. https://doi.org/10.1093/nar/gkaa1100

Urban, M., Cuzick, A., Seager, J., Wood, V., Rutherford, K., Venkatesh, S. Y., De Silva, N., Martinez, M. C., Pedro, H., Yates, A. D., Hassani-Pak, K., & Hammond-Kosack, K. E. (2020). PHI-base: the pathogen-host interactions database. *Nucleic Acids Res*, *48*(D1), D613-D620. https://doi.org/10.1093/nar/gkz904

U. S. EPA (2020). User's Guide for T.E.S.T. (version 5.1) (Toxicity Estimation Software Tool): A Program to Estimate Toxicity from Molecular Structure. In O. Chemical Characterization and Exposure Division Cincinnati (Ed.).

U. S. FDA (2020). *Allergens in cosmetics*. https://www.fda.gov/cosmetics/cosmetic-ingredients/allergens-cosmetics.

van de Veerdonk, F. L., Kullberg, B. J., van der Meer, J. W., Gow, N. A., & Netea, M. G. (2008). Host-microbe interactions: innate pattern recognition of fungal pathogens. *Curr Opin Microbiol*, *11*(4), 305-312. https://doi.org/10.1016/j.mib.2008.06.002

van Ree, R., Sapiter Ballerda, D., Berin, M. C., Beuf, L., Chang, A., Gadermaier, G., Guevera, P. A., Hoffmann-Sommergruber, K., Islamovic, E., Koski, L., Kough, J., Ladics, G. S., McClain, S., McKillop, K. A., Mitchell-Ryan, S., Narrod, C. A., Pereira Mouries, L., Pettit, S., Poulsen, L. K., . . . Bowman, C. (2021). The COMPARE Database: A Public Resource for Allergen Identification, Adapted for Continuous Improvement. *Front Allergy*, *2*, 700533. https://doi.org/10.3389/falgy.2021.700533

Vens, C., Rosso, M. N., & Danchin, E. G. (2011). Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics*, *27*(9), 1231-1238. https://doi.org/10.1093/bioinformatics/btr110

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., & Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res*, *47*(D1), D339-D343. https://doi.org/10.1093/nar/gky1006

Waghu, F. H., Gopi, L., Barai, R. S., Ramteke, P., Nizami, B., & Idicula-Thomas, S. (2014). CAMP: Collection of sequences and structures of antimicrobial peptides. *Nucleic Acids Res*, *42*(Database issue), D1154-1158. https://doi.org/10.1093/nar/gkt1157

Wang, J., Zhang, D., & Li, J. (2013). PREAL: prediction of allergenic protein by maximum Relevance Minimum Redundancy (mRMR) feature selection. *BMC Syst Biol*, *7 Suppl 5*, S9. https://doi.org/10.1186/1752-0509-7-S5-S9

Watford, S., Ly Pham, L., Wignall, J., Shin, R., Martin, M. T., & Friedman, K. P. (2019). ToxRefDB version 2.0: Improved utility for predictive and retrospective toxicology analyses. *Reprod Toxicol*, *89*, 145-158. https://doi.org/10.1016/j.reprotox.2019.07.012

Wei, L., Ye, X., Sakurai, T., Mu, Z., & Wei, L. (2022). ToxIBTL: prediction of peptide toxicity based on information bottleneck and transfer learning. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btac006

Wei, L., Ye, X., Xue, Y., Sakurai, T., & Wei, L. (2021). ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. *Brief Bioinform*, *22*(5). https://doi.org/10.1093/bib/bbab041

Weiss, R. A. (2002). Virulence and pathogenesis. *Trends Microbiol*, *10*(7), 314-317. https://doi.org/10.1016/s0966-842x(02)02391-0

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., . . . Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*, *46*(D1), D1074-D1082. https://doi.org/10.1093/nar/gkx1037

Wong, E. S., Hardy, M. C., Wood, D., Bailey, T., & King, G. F. (2013). SVM-based prediction of propeptide cleavage sites in spider toxins identifies toxin innovation in an Australian tarantula. *PLoS ONE*, *8*(7), e66279. https://doi.org/10.1371/journal.pone.0066279

Wood, D. L., Miljenovic, T., Cai, S., Raven, R. J., Kaas, Q., Escoubas, P., Herzig, V., Wilson, D., & King, G. F. (2009). ArachnoServer: a database of protein toxins from spiders. *BMC genomics*, *10*, 375. https://doi.org/10.1186/1471-2164-10-375

Xu, S. X., & McCormick, J. K. (2012). Staphylococcal superantigens in colonization and disease. *Front Cell Infect Microbiol*, *2*, 52. https://doi.org/10.3389/fcimb.2012.00052

Yang, H., Sun, L., Li, W., Liu, G., & Tang, Y. (2018). In Silico Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts. *Front Chem*, *6*, 30. https://doi.org/10.3389/fchem.2018.00030

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

Zhang, Z. H., Koh, J. L., Zhang, G. L., Choo, K. H., Tammi, M. T., & Tong, J. C. (2007). AllerTool: a web server for predicting allergenicity and allergic cross-reactivity in proteins. *Bioinformatics*, *23*(4), 504-506. https://doi.org/10.1093/bioinformatics/btl621

Zheng, L. L., Li, Y. X., Ding, J., Guo, X. K., Feng, K. Y., Wang, Y. J., Hu, L. L., Cai, Y. D., Hao, P., & Chou, K. C. (2012). A comparison of computational methods for identifying virulence factors. *PLoS ONE*, *7*(8), e42517. https://doi.org/10.1371/journal.pone.0042517

Zhou, C. E., Smith, J., Lam, M., Zemla, A., Dyer, M. D., & Slezak, T. (2007). MvirDB--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res*, *35*(Database issue), D391-394. https://doi.org/10.1093/nar/gkl791

Zhu, X. Q., Li, L. Y., Yang, W. M., & Wang, Y. (2020). Combined dimercaptosuccinic acid and zinc treatment in neurological Wilson's disease patients with penicillamine-induced allergy or early neurological deterioration. *Biosci Rep*, *40*(8). https://doi.org/10.1042/BSR20200654

Zhu, X., & Galili, G. (2004). Lysine metabolism is concurrently regulated by synthesis and catabolism in both reproductive and vegetative tissues. *Plant physiology*, *135*(1), 129–136. https://doi.org/10.1104/pp.103.037168