Species-Specific Prediction of B Cell Epitopes

A Thesis

submitted to

Indian Institute of Science Education and Research Pune in partial fulfilment of the requirements for the BS-MS Dual Degree Programme

by

Ruchir Sahni

Indian Institute of Science Education and Research Pune

Dr. Homi Bhabha Road,

Pashan, Pune 411008, INDIA.

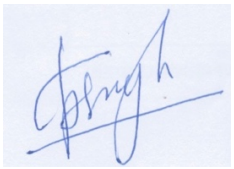
April, 2025

Supervisor: Prof. Gajendra Pal Singh Raghava (IIIT Delhi)

Expert: Prof. M.S. Madhusudhan (IISER Pune)

Certificate

This is to certify that this dissertation entitled **Species-Specific Prediction of B Cell Epitopes** towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Ruchir Sahni at Indraprastha Institute of Information Technology Delhi (IIIT-Delhi) under the supervision of Dr. Gajendra Pal Singh Raghava, Head and Professor, Department of Computational Biology, during the academic year 2024-2025.



Prof. Gajendra Pal Singh Raghava

Indraprastha Institute of Information Technology, Delhi

Committee Members:

Supervisor:
Prof. Gajendra Pal Singh Raghava
Head & Professor
Department of Computational Biology

IIIT Delhi

Expert:
Prof. M.S Madhusudhan
Professor
Department of Biology
and Data Science (Joint)
IISER Pun

Declaration

I hereby declare that the matter embodied in the report entitled ***Species-Specific Prediction of B Cell Epitopes*** are the results of the work carried out by me at the Department of Computational Biology, Indraprastha Institute of Information Technology, Delhi, under the supervision of Prof. Gajendra Pal Singh Raghava, and the same has not been submitted elsewhere for any other degree. Wherever others contribute, every effort is made to indicate this clearly, with due reference to the literature and acknowledgement of collaborative research and discussions.

Ruchir Sahni

Roll Number: 20201123

Table of Contents

1. List of Figures

2. List of Tables

3. List of Papers

1. Sahni, R., Kumar, N., & Raghava, G. P. (2024). HAIRpred: Prediction of human antibody interacting residues in an antigen from its primary structure. bioRxiv (Cold Spring Harbor Laboratory). https://doi.org/10.1101/2024.12.17.628825

4. Acknowledgments

I am deeply grateful to my supervisor, Professor G.P.S. Raghava, for his insightful guidance and unwavering support throughout the course of my thesis. His knowledge and feedback has been crucial in shaping my work. The extensive discussions with him regarding my project always led me to new ideas.

I would also like to acknowledge my PhD mentor, Nishant Kumar, for his constant encouragement and mentorship. His guidance has been a source of inspiration, and his feedback has helped me refine my research.

Additionally, I would like to acknowledge the entire Raghava Lab, which has provided me with a stimulating environment and the necessary guidance to conduct my research.

I would like to appreciate my family for their unwavering love and support. Their constant encouragement and patience have been instrumental in helping me overcome the challenges of my research journey. I would also like to extend my gratitude to my friends, who have been a source of comfort, motivation, and joy. Finally, I would like to specially acknowledge Akanksha K for her constant presence and support. It made my journey more enjoyable, and I cherish the memories we have shared.

5. Abstract

Antibody-Antigen interactions are fundamental in immunological research, influencing vaccines, diagnostics and therapeutic developments. However, existing computation models for predicting the antibody-interacting residues (epitopes) suffer from significant limitations, especially the lack of species specificity. These models are trained on data from diverse species and hence fail to account for the interspecies variation in antibody-antigen interactions.

To address these limitations, this study presents a novel framework for species-specific B-Cell epitope prediction. We first establish the existence of interspecies differences in the interactions by comparing the amino acid composition of human-specific and mouse-specific epitopes, highlighting the need for species-specific models. Next, we develop HAIRpred (Human Antibody Interacting Residue predictor), a second-generation human-specific ensemble model based on random forests. HAIRpred integrates information from position-specific scoring matrices (PSSM) and relative solvent accessibility (RSA), which were identified as key features determining antibody-antigen interaction. The model achieves the area under the receiver operating characteristic curve (AUROC) score of 0.72 on the independent test dataset which surpasses the performance of the current state-of-the-art models. Additionally, we use the SHAP explainable AI (XAI) approach to analyse the contribution of residue position in the pattern in the eventual prediction.

Furthermore, we recognize the need for mouse-specific prediction tools and we extend our methodology to develop a mouse-specific predictor. However, data scarcity is a significant challenge with only 290 available antibody-antigen complexes forming 94 clusters at 70% sequence identity threshold. To address this, we implement a cluster-based data splitting strategy, preventing overestimation of model performance.

Our findings highlight the need for species-specific models and set a new standard for epitope prediction tools. By addressing the limitations of previous studies, we anticipate that these models will be a valuable resource in immunological research.

6. Contributions

Contributor name	Contributor role
G.P.S Raghava, Nishant Kumar	Conceptualization Ideas
G.P.S Raghava, Nishant Kumar, Ruchir Sahni	Methodology
Ruchir Sahni	Software
G.P.S Raghava, Nishant Kumar, Ruchir Sahni	Validation
Ruchir Sahni	Formal analysis
Ruchir Sahni	Investigation
G.P.S Raghava	Resources
Ruchir Sahni	Data Curation
Ruchir Sahni	Writing - original draft preparation
Ruchir Sahni, Nishant Kumar, G.P.S Raghava	Writing - review and editing
Ruchir Sahni, Nishant Kumar, G.P.S Raghava	Visualization
G.P.S Raghava	Supervision
G.P.S Raghava	Project administration
G.P.S Raghava	Funding acquisition

Note: The majority of the work presented in this thesis has been previously described in the preprint titled *HAIRpred: Prediction of human antibody interacting residues in an antigen from its primary structure,* available at [https://doi.org/10.1101/2024.12.17.628825].

Chapter 1 : Introduction

1.1 Immunology

Immunology is the study of the immune system, which is an interconnected network of organs, cells, proteins and chemicals that protects the body from diseases. The immune system detects and responds to infection-causing pathogens, parasitic worms and cancer cells. The idea of disease immunity dates back to 5th century BC Greece, where Thucydides observed that individuals who recovered from the plague did not succumb to it again and described them as "immune" or "exempt" (Greenberg, 2003). The modern discipline of immunology arose in the 18th Century from the pioneering works of Edward Jenner, leading to the first demonstration of vaccination (Jenner, 2023).

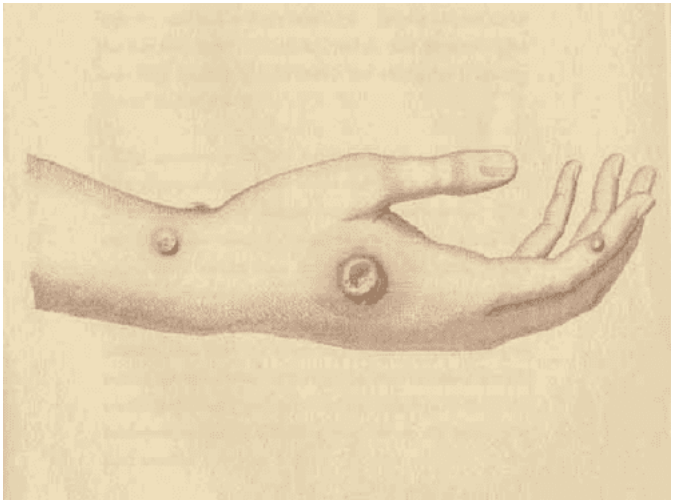


Figure 1: Drawing of Cowpox lesion in Edward Jenner's notebooks
Sourced from Greenberg, 2003

The immune system is generally divided into two main subcategories: the innate immune system and the adaptive immune system. The innate immune system serves as the body's first line of defence against pathogens. It works against a wide variety of targets and gives rise to an instantaneous response. However, it does not give rise to immunological memory and is not specific to any target. The innate immune system consists of anatomical barriers, cells which can target and kill pathogens, signalling molecules that influence the workings of the innate immune system, and the complement system. Among these, immune cells such as macrophages, neutrophils and dendritic cells express Pattern Recognition Receptors (PRRs) that detects conserved molecular features, which are called Pathogen-Associated Molecular Patterns (PAMPs), to target pathogens and initiate an effective immune response (Janeway, 2001).

In contrast, the adaptive immune system uses a target-specific approach to combat pathogens. The response involves specialised white blood cells that recognise the pathogen using specific receptors and eliminating it. The adaptive immune system generates a diverse repertoire of receptors from a limited set of genes to identify the pathogens. Additionally, the adaptive immunity is characterised by the ability to develop immunological memory, leading to quicker and more effective response upon repeat exposure. However unlike the innate immune system,

the adaptive immune response takes longer to develop. Adaptive immunity primarily comprises of B and T cells (Janeway, 2001).

All immune cells originate from a single precursor called the hematopoietic stem cell (HSCs). HSCs can differentiate to either myeloid or lymphoid progenitor cells. Myeloid progenitor cells give rise to the innate immune cells including granulocytes such as neutrophils, basophils and eosinophils as well as antigen presenting cells like macrophages and dendritic cells. Similarly, lymphoid progenitor cells predominantly differentiates into adaptive immune system cells like B and T cells (Janeway, 2001). This process is described in Figure 2. In this study, we focus on B-Cells and associated antibody-mediated immunity.

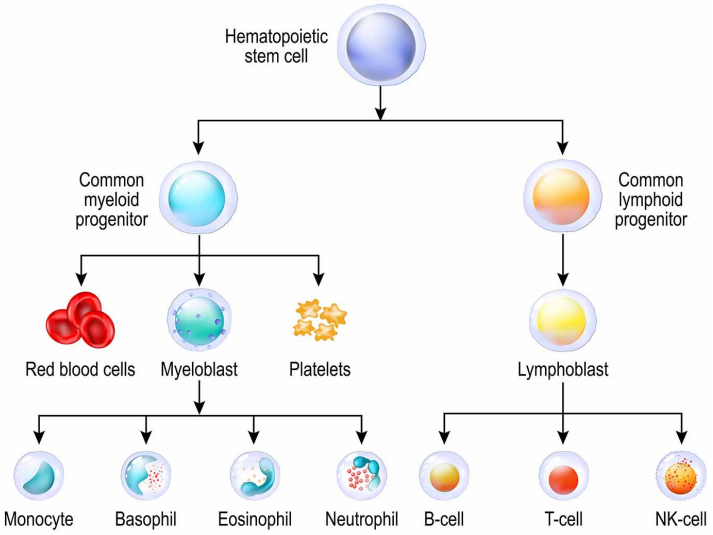


Figure 2: Process of Hematopoiesis
Sourced from Beckman Coulter

1.2 B cells and Antibodies

B lymphocyte is a type of white blood cell which develops and matures in the bone marrow and plays a crucial role in the adaptive immune system. Their primary role is to develop humoral immunity by secreting antibodies. B-Cells differentiates from hematopoietic stem cells through a process which includes stages of progenitor B-Cells, precursor B-Cells, immature B cells and mature B cells (LeBien and Tedder, 2008).

Mature B cells migrate to lymph organs, like the spleen and lymph nodes, where they encounter foreign particles. B cells gets activated from the simultaneous recognition of antigen from their B-cell receptors and signals from $CD4^+$ T helper cells. This activation leads to proliferation of B cells and the B cell differentiates into two types of effector cells : plasma cells and memory B cells. Memory B cells provide long-term immunity by developing immunological memory. Plasma cells, on the other hand, are the main effector cells responding to the pathogen by secreting large amount of antibodies (LeBien and Tedder, 2008).

Antibodies are Y-shaped glycoproteins that are produced by plasma cells, and they bind to extracellular pathogen associated molecules called antigens. They are tetramers, comprising of two identical chains which are called the heavy and light chains. The chains are connected together by disulphide bonds. Each antibody belongs to one of five classes (isotypes)—IgG, IgA, IgE, IgD, and IgM—determined by the structure of their heavy chains. Each antibody

class has a distinct receptor, cellular localization, and biological role (Chiu *et al.*, 2019). The structure of an antibody is described in Figure 3.

Each antibody has two key regions : Fc (fragment crystallisable) and Fab (fragment antigen-binding) region. The Fc region binds to other immune cells to mediate immune responses. It is also critical for activating the complement system which is a series of proteins that work together to destroy pathogens. On the other hand, the Fab region contains the variable domains that bind to antigens and the binding is crucial for neutralizing pathogens or marking them for elimination by other immune cells. The specificity of the binding is determined by its complementary-determining regions (CDRs), which binds to antigens through non-covalent forces including electrostatic interactions, van der Waals forces, hydrogen bonds, and hydrophobic interactions (Chiu *et al.*, 2019).

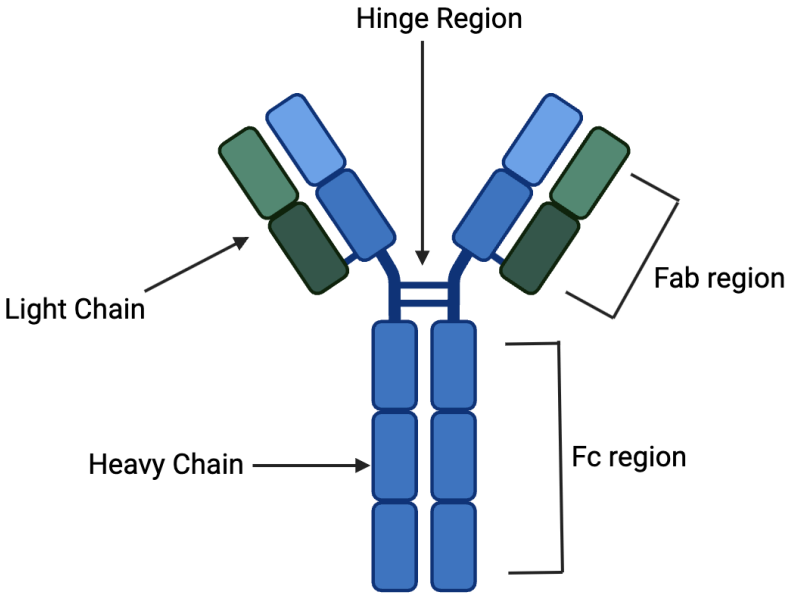


Figure 3: Structure of an Antibody

Antibodies have the ability to recognize and bind to a diverse range of antigens due to the diversity in antibody structures. Antibody diversity arises from four main mechanisms. First, the diversity results from V(D)J recombination, where random variable (V), diversity (D), and joining (J) gene segments are joined to create the variable domain of the antibody. The light chains are formed by recombing V and J gene segments whereas the heavy chains undergo a more complex rearrangement involving V, D, and J segments. Since multiple copies of these gene segments are present in the genome, different combinations result into a significant level of diversity(Janeway, 2001).

The second mechanism is called junctional diversity and it is introduced through the changes at the junctions where V, D, and J segments are joined, as a result of improper joining of the gene segments. There is addition and subtraction of nucleotides at the junctions due to the imprecise recombination process, resulting in diversity in the antibody structure. The third mechanism arises from the combinatorial pairing of light and heavy chains. Each immunoglobulin contains one heavy and light chain and the pairing of different heavy chains and light chains results in exponential increase in number of possible antibody structures (Janeway, 2001).

Lastly, somatic hypermutation introduces additional variation by introducing point mutations in the rearranged V-region genes. This process occurs after antigen exposure and it involves the enzyme Activation-Induced cytidine Deaminase (AID) deaminating the cytosine residues, which leads to mutations that can either increase or decrease the affinity for the antigen. Then, B cells producing antibodies with improved affinity is selected. Together, these four mechanisms result in a large repertoire of possible antibody structures which allows the antibodies to bind to a wide variety of antigens (Janeway, 2001).

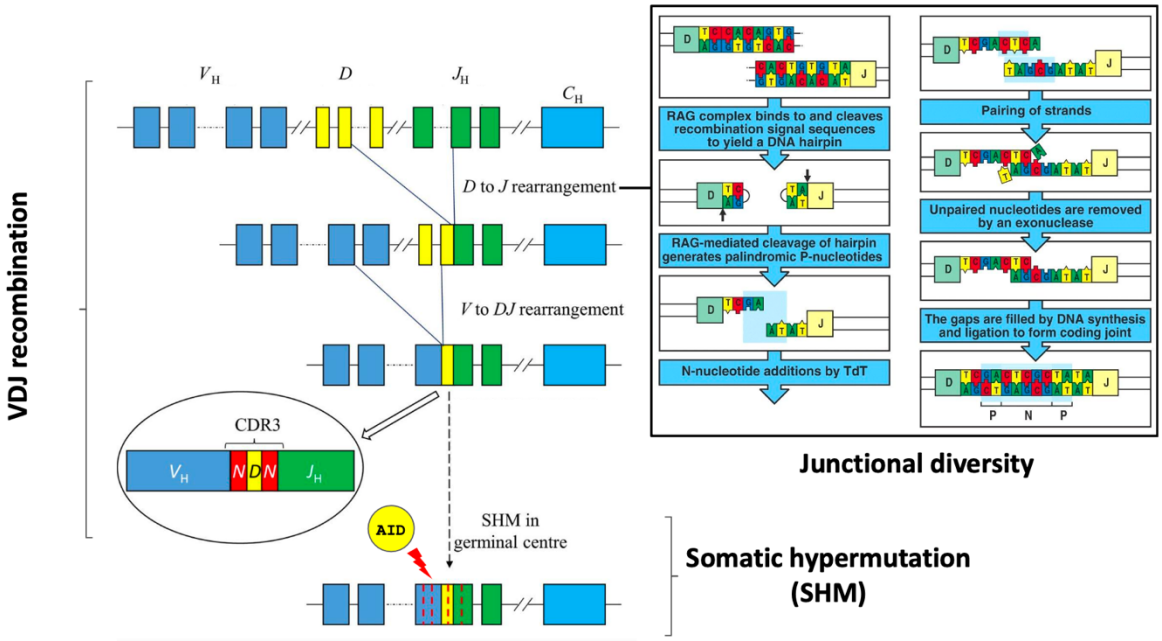


Figure 4: Mechanisms which lead to diversity in antibody structures
Sourced from Gabriel CIA BERIAIN, 2023

1.3 B Cells Epitopes

B cells play an important role in the adaptive immune response by recognizing and responding to pathogens. Upon encountering these pathogens, B cells secrete antibodies specific to these antigens. The antigen region which binds to the antibody is called the B cell epitope. These epitopes play an important role in the mediating the immune response as they are critical in the recognition of pathogens and the subsequent neutralization (Alberts, 2014).

Epitopes can be divided into two types : Linear epitopes and Conformational epitopes. Epitopes composed of a continuous sequence of adjacent residues are known as continuous or linear epitopes. However, crystallographic studies of antibody–antigen complexes have revealed that majority of the epitopes are discontinuous and they are formed by atoms from non-adjacent residues that come together on the protein's surface in its tertiary structure (Regenmortel, 1996a). These epitopes, which involve residues that are spaced out in the linear sequence but are spatially close in the folded protein, are known as conformational epitopes. The aim of this study is to identify these conformational B cell epitopes computationally.

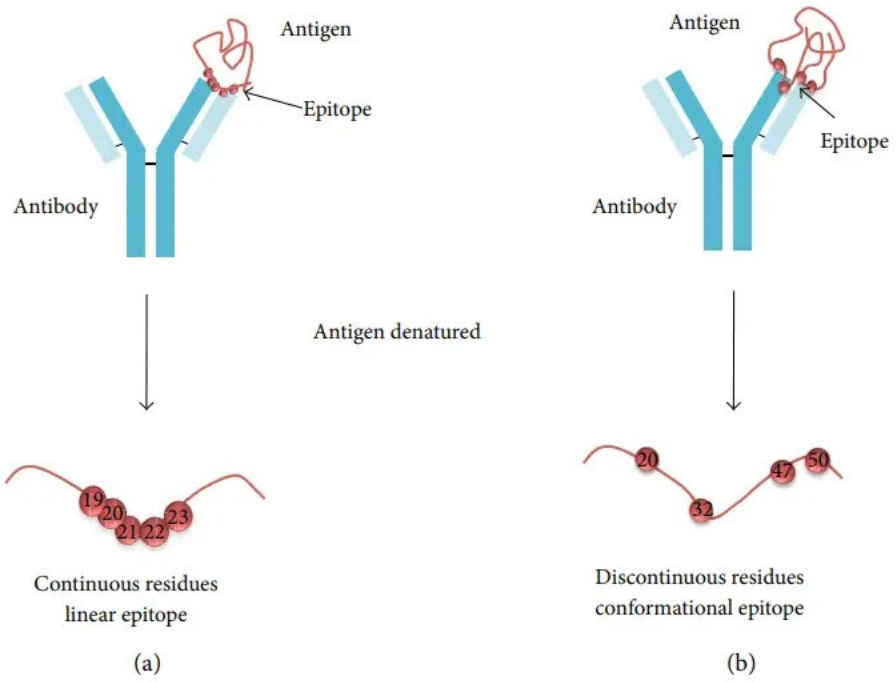


Figure 5: Two types of epitopes
Sourced from Tankeshwar (microbeonline.com)

The primary goal of B cell epitope identification is to design molecules that can substitute for an antigen in antibody production or detection. These molecules, which can be synthesized and expressed through cloned genes, are a safer and cost-effective alternative to live pathogens as they do not pose any infection risks (Ponomarenko and Van Regenmortel, 2009).

Antibodies and synthetic peptides derived from epitopes are widely used in diagnosing infectious, autoimmune, and allergic diseases. Diagnostic immunoassays use these peptides or antipeptide antibodies to detect disease markers like viral proteins and bacterial lipopolysaccharides (Ponomarenko and Van Regenmortel, 2009). In autoimmune research, predicting B cell epitopes in neurological autoimmune diseases like multiple sclerosis could help identify potential therapeutic targets (Talucci and Maric, 2024). Similarly, in allergy treatment, understanding IgE-binding epitopes helps in the development of hypoallergenic protein variants for immunotherapy (Dall'Antonia and Keller, 2019). Beyond diagnostics, epitope identification is crucial in monoclonal antibody production for therapeutic applications (Liew et al., 2021).

Another important application of epitope identification is synthetic vaccine design. Traditional vaccines which are usually derived from live-attenuated or inactivated pathogens (e.g., polio, measles) carry safety risks due to the potential reversion of virulence or adverse immune reactions. In comparison, synthetic peptides offer a safer, cost-effective alternative. Additionally, predicting immunogenic peptides that elicit neutralizing antibodies would lead to rational vaccine design (Ponomarenko and Van Regenmortel, 2009).

1.4 Identification of B Cell Epitopes

The identification of B cell epitopes is crucial for understanding vaccine design, immune responses, and therapeutic antibody development. These predictions can be made through two broad approaches: experimental and computational methods.

1.4.1 Experimental Methods of Identification of Epitopes

Experimental methods for identify B cell epitopes can be categorised into structural and functional approaches. These methods differ in the perspective of what is classified as an epitope. Structural methods, such as nucleic magnetic resonance (NMR), X-ray crystallography, and electron microscopy (EM), define structural epitopes as the collection of antigen atoms which are in direct contact with the antibody (Ponomarenko and Van Regenmortel, 2009). Ponomarenko and Bourne (2007) analysed a dataset comprising 59 epitopes from single-chain proteins bound to two-chain antibody fragments and estimated that structural epitopes consist of 10 to 22 residues, with an average of 16 residues. This estimation was performed by defining structural epitopes as antigen residues with at least one atom within the 4 Å radius of the antibody atom (Ponomarenko and Bourne, 2007).

Functional methods, on the other hand, are identified through functional assays that look at changes in antibody-antigen binding affinity after antigen modifications including site-directed mutagenesis (Benjamin and Perdue, 1996; Ponomarenko and Van Regenmortel, 2009). The functional epitopes tend to be a subset of structural epitopes – around 3 to 5 residues as these residues significantly affect the binding affinity of the antibody and the antigen.

It is important to note that the epitopes identified from the experimental methods may not always correspond exactly to the actual immunogenic epitope. This can be accounted for by the ability of the antibodies to recognize both the epitopes they were elicited against and the structurally similar epitopes. Therefore, antibodies generated against a native protein may exhibit cross-reactivity with linear peptides that only partly resemble the original epitope (Ponomarenko and Van Regenmortel, 2009). Additionally, protein preparations used may contain denatured molecules, with the observed cross-reactivity resulting from the antibodies specific to the denatured protein (Laver et al., 1990; Regenmortel, 1996b). Furthermore, the residues identified as functional residues may not actually be epitopic. Instead, they have been identified as functional residues as their substitution might induce conformational changes in the protein which will indirectly affect the antibody-antigen binding (Ponomarenko and Van Regenmortel, 2009).

While these experimental methods have been important in advancing epitope identification, they are often costly, labor-intensive, and unsuitable for large-scale studies. Conventional techniques, such as alanine scanning mutagenesis and X-ray crystallography, require significant resources and time which limits their application in high-throughput epitope mapping.

1.4.2 Computational Methods of Prediction of Epitopes

With advancements in computational power and the emergence of sophisticated algorithms, particularly in machine learning, computational prediction of epitopes have become more efficient and scalable. Machine learning models use large datasets of known antigen–antibody interactions to identify patterns and make accurate predictions. These models utilize features

which provides information about the antibody-antigen binding to make the predictions. By capturing complex relationships affecting the binding, computational approaches offer a promising alternative to accelerate antibody research and improve vaccine design and therapeutic antibody development.

The first epitope prediction algorithm was introduced by Hopp and Woods in 1981 to predict linear epitopes under the assumption that hydrophilic residues are primarily located on the protein surface (Hopp and Woods, 1981). Since then, different approaches have been developed for linear epitope prediction (Chou and Fasman, 1974; Kyte and Doolittle, 1982; Parker *et al.*, 1986; Larsen *et al.*, 2006; Saha and Raghava, 2006; Sweredoski and Baldi, 2008a; Wang *et al.*, 2011; Liu *et al.*, 2020; Bahai *et al.*, 2021; Collatz *et al.*, 2021a; Liu *et al.*, 2024). These tools utilize a diverse array of amino acid properties, such as hydrophilicity, solvent accessibility, secondary structure, and flexibility.

However, the prediction of conformational B cell epitopes is more complicated. The lack of high-quality structural data for many antigens makes it difficult to get accurate predictions. The current approaches can be categorized into two groups : sequence-based methods and structure-based methods. Structure-based methods utilize the three-dimensional information of the antigen to predict the antigenic residues which are most likely to interact with the antibody. Several structure-based methods have been created including, but not limited to, CEP, DiscoTope, PEPITO, Epitopia, SEPPA, CBTope, SEMA, CLBTope and Epitope3D (Kulkarni-Kale *et al.*, 2005; Haste Andersen *et al.*, 2006; Sweredoski and Baldi, 2008b; Rubinstein *et al.*, 2009; Sun *et al.*, 2009; Ansari and Raghava, 2010; Solihah *et al.*, 2020; da Silva *et al.*, 2021; Ivanisenko *et al.*, 2024; Kumar *et al.*, 2024). These methods analyse structural features like solvent accessibility, spatial proximity of residues and residue conservation to predict conformational epitopes. These approaches provide a more precise representation of conformation epitopes because they incorporate structural data. However, a significant limitation is their dependence on high quality data on the tertiary structure of antigens which are often not available. Experimental techniques like X-ray crystallography and cryo-electron microscopy are expensive and time-consuming and the computational structure prediction methods, such as AlphaFold, may not always provide sufficiently accurate 3-D structures.

To address this limitation, several sequence-based models have been created to predict conformational B cell epitopes directly from the primary sequence of an antigen. Prominent sequence-based methods are BepiPred-3.0, SEMA-2.0, Epidope, CBTope, and CLBTope (Ansari and Raghava, 2010; Collatz *et al.*, 2021b; Clifford *et al.*, 2022; Ivanisenko *et al.*, 2024; Kumar *et al.*, 2024). These models use features derived from the primary sequence which include amino acid composition, hydrophobicity, flexibility, and evolutionary conservation to infer conformational epitopes.

The recent more advanced prediction models also integrate protein language model embeddings which significantly improve the sequence-based predictions as the embeddings implicitly encode structural and evolutionary information of proteins. Protein Language Models (PLMs) are large language models trained on extensive protein sequence databases which enables them to learn the contextual representations of amino acids. The embeddings generated by these models capture structural information like the residue-level interactions, secondary structure, solvent accessibility and other tertiary structure information (Elnaggar *et al.*, 2022; Lin *et al.*, 2023). By incorporating these embeddings, the model receives a more accurate representation of the antigen and leads to better predictions.

1.4.3 Limitation of the Computational Prediction Models

Despite significant advancements in the prediction methodology, the computational models still have major limitations. Cia et al. (2023) conducted a comprehensive benchmarking study of 9 popular B cell epitope prediction models, including sequence-based and structure-based methods. It highlighted that many existing methods achieve low predictive performance, with some failing to outperform randomly generated surface residue patches (Figure 6). Additionally, they reported that there is a lack of consensus among different predictors—an antigen that is well-predicted by one method may perform poorly in another. The study incorporated multiple metrics to assess the performance of the models and all indicate that the existing models are underperforming.

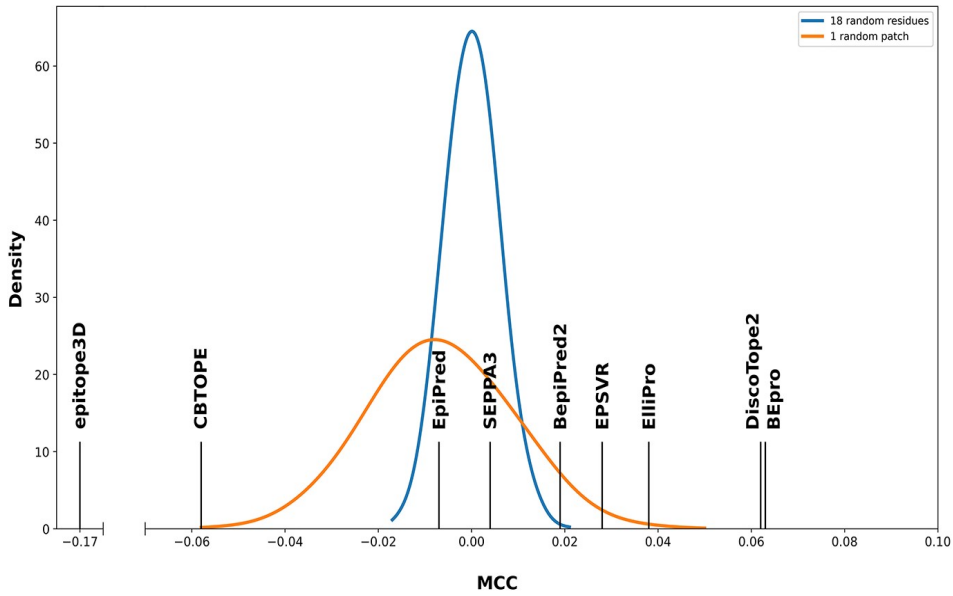


Figure 6: Comparison of state-of-art conformational B cell epitope prediction tools. Additionally, these tools are compared to bootstrap distributions of a random procedure that produces 1 random patch of 18 nearby surface residues (displayed in orange) or 18 random surface residues (displayed in blue) Sourced from Cia et al., 2023

The major drawback of the current computational models is the diverse nature of antigen–antibody interactions across different species and the lack of species specificity in these models. Antibodies are highly specific to their host organisms and the epitopes recognized by human antibodies can vary considerably from the ones targeted by antibodies from other species. The existing prediction tools are trained on datasets comprising antibodies from a broad range of species, including mice, rabbits, and other model organisms commonly used in experimental immunology. These models often do not take into consideration the interspecies variations in immune response, antibody repertoire, and post-translational modifications (Almagro et al., 1998; Mestas and Hughes, 2004). As a result, predictions generated from these generalized models may not accurately capture species-specific epitope recognition. This lack of species specificity is a significant challenge to the applicability of computational epitope prediction.

1.5 Creating a Species-Specific Prediction Model

This study addresses the challenge of species specificity in computational epitope prediction. It demonstrates the differences in antibody–antigen binding between humans and mice which

highlights the limitations of existing generalized models that fail to capture species-specific immune responses. To overcome this limitation, we developed sequence-based computational tools designed to predict human-specific and mouse-specific epitopes directly from the antigen sequence. To ensure host specificity, these computational tools were trained and evaluated using experimentally derived species-specific antibody–antigen complexes.

To develop these tools, machine learning models were created based on a carefully selected set of features derived from the antigen's primary sequence. The selection of these features was guided by their predictive effectiveness in previous studies and their biological significance. This feature set incorporates structural, sequential, and evolutionary information about the antigen which can be derived from its primary sequence. The machine learning models then underwent rigorous testing to identify key features affecting the interaction between the antibody and the antigen. The machine learning models based on these key determinants were subsequently integrated into an ensemble framework, forming the final predictive tool.

The performance of the species-specific epitope prediction tools were assessed in comparison to existing methods using the test dataset of experimentally derived antibody–antigen complexes. Furthermore, the SHAP (Shapley Additive Explanations) method was employed to interpret the predictions, offering insights into the relative contribution of each residue position in the pattern to the final outcomes. By prioritizing feature selection and model interpretability, these tools improves predictive accuracy as well as provides valuable insights into the mechanisms governing these interactions. Ultimately, this study provides a foundational framework for species-specific epitope prediction, enabling more accurate predictions that will facilitate advancements in immunotherapy, rational vaccine design, and the development of therapeutic antibodies.

Chapter 2: Results

The Results chapter has been divided into three sections : (i) Importance of Host Specificity, (ii) Development of human-specific B cell epitope predictor (HAIRpred), and (iii) Mouse-specific B cell epitope prediction.

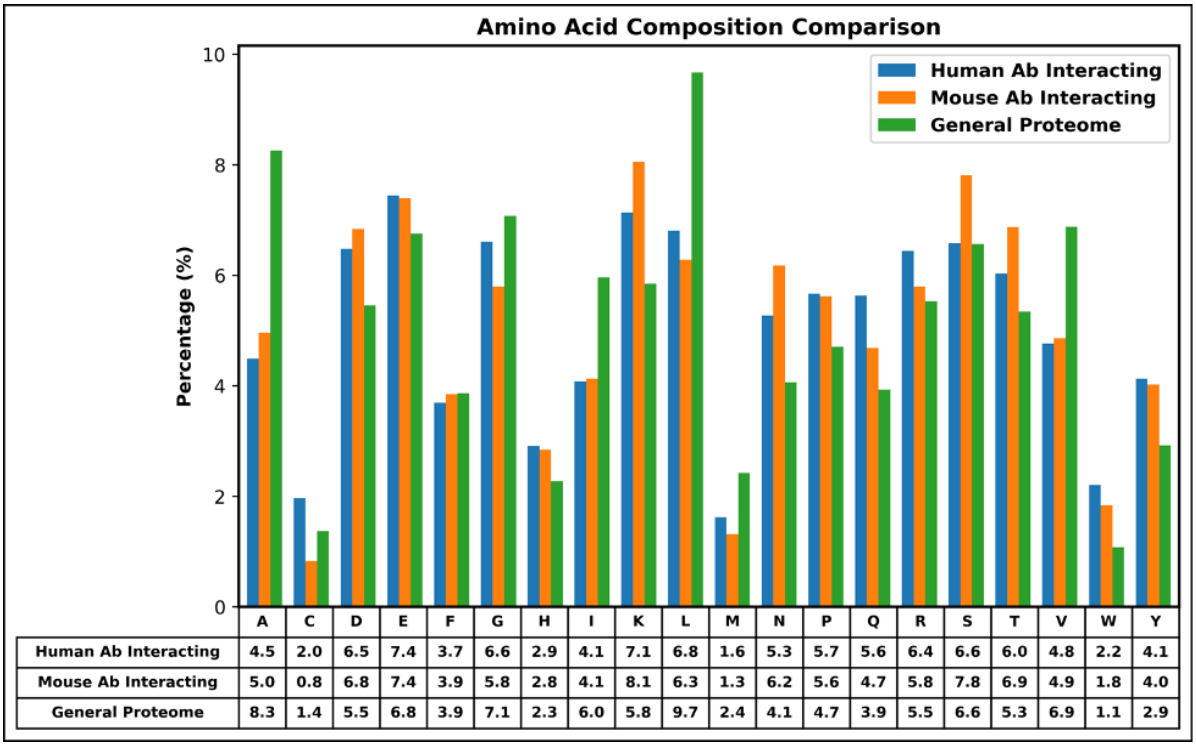
All the results are based on the data of antigen-antibody complexes derived from the SAbDab database (Dunbar *et al.*, 2014). The details of the data acquisition and further processing are provided in Section 4.1 under Materials and Methods.

These results have been uploaded in Biorxiv as a preprint (Sahni *et al.*, 2024) and the publication is currently under review.

2.1 Importance of Host Specificity

The objective of this study is to understand the drawbacks of the current B cell epitope prediction methods and develop better models overcoming these drawbacks. One of the major limitations of the current models is that antigen-antibody interactions differ across species and the existing generalized models cannot account for these differences. To observe and understand the difference in the interaction, experimentally-derived human-specific and mouse-specific antigen-antibody complexes were obtained, and the residues interacting with the antibody (epitopes) were identified. The details of this identification is provided in Section 4.1 and 4.2 under Materials and Methods.

We compared the Amino Acid Composition (AAC) for both human and mouse epitopes to observe the differences between the epitopes. The general proteome was also introduced in the comparison to highlight the characteristics of human and mouse epitopes. The results of the AAC comparison is displayed in Figure 7.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Human Ab Interacting	4.5	2.0	6.5	7.4	3.7	6.6	2.9	4.1	7.1	6.8	1.6	5.3	5.7	5.6	6.4	6.6	6.0	4.8	2.2	4.1
Mouse Ab Interacting	5.0	0.8	6.8	7.4	3.9	5.8	2.8	4.1	8.1	6.3	1.3	6.2	5.6	4.7	5.8	7.8	6.9	4.9	1.8	4.0
General Proteome	8.3	1.4	5.5	6.8	3.9	7.1	2.3	6.0	5.8	9.7	2.4	4.1	4.7	3.9	5.5	6.6	5.3	6.9	1.1	2.9

Figure 7: Comparison of AAC of human and mouse epitopes

Analysis of the AAC revealed notable differences between the epitopes from the different species. Specifically, amino acids such as Cysteine (C), Glutamine (Q), Arginine (R), Tryptophan (W), and Tyrosine (Y) were more prevalent in human antibody–antigen interactions, whereas Glycine (G), Lysine (K), Asparagine (N), and Serine (S) were favored in mouse interactions. These findings highlights the variations in immune interactions between humans and mice and demonstrates the necessity of species-specific predictors to ensure accurate epitope prediction.

2.2 Development of human-specific B cell epitope predictor (HAIRpred)

Given the observed differences in AAC between human and mouse antibody-interacting residues, it is evident that a species-specific predictor is needed to accurately capture species-specific epitope patterns. To address this need, this study develops HAIRpred - a human-specific B cell epitope predictor. HAIRpred uses a combination of machine learning techniques based on biologically informed features to predict B cell epitopes with high precision.

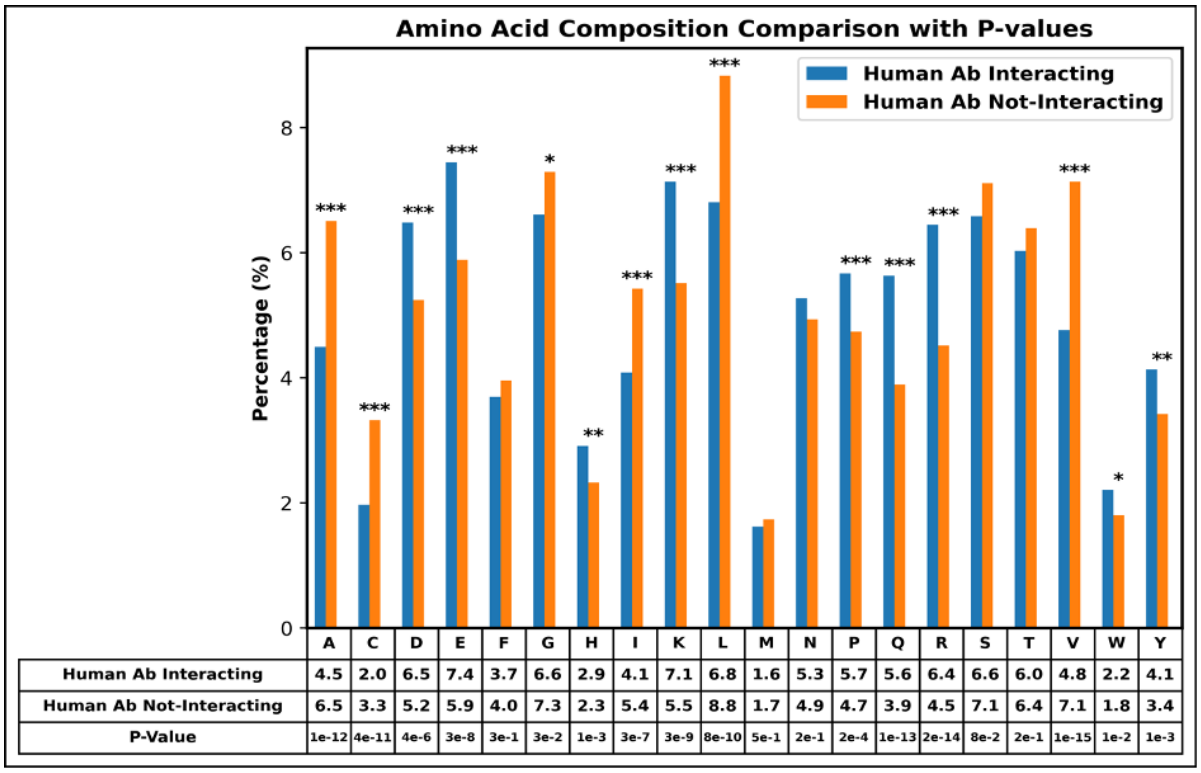
This section has been divided into five subsections : (i) Data analysis, (ii) Development of a human-specific predictor (iii) 2.2.3 Comparison between HAIRpred and existing methods, (iv) SHAP Analysis and (v) Web Server

2.2.1 Data Analysis

The objective of this part of the study was to create a human-specific B cell epitope predictor. Therefore, human-specific antigen-antibody complexes were obtained from the SAbDab database (Dunbar *et al.*, 2014). The complexes were filtered to get a quality dataset of 1620 human-specific antibody-antigen complexes. In order to avoid redundancy, the antigens from the complexes were clustered based on sequence identity using CD-HIT (Fu et al., 2012) with

a 70% sequence identity threshold. This resulted in a high quality non-redundant dataset of 277 antibody-antigen complexes. The antigen residues were then labelled as antibody-interacting and antibody not-interacting based on the algorithm outlined in Section 4.2 under Materials and Methods.

To investigate amino acid preferences in antibody interaction, the AAC of interacting and not-interacting residues were compared. The results are displayed in Figure 8. Statistical significance was assessed using the chi-square test. The analysis revealed a significantly higher abundance of Glutamic Acid (E), Glutamine (Q), Aspartic Acid (D), Histidine (H), Lysine (K), Proline (P), Tryptophan (W), Arginine (R), and Tyrosine (Y) in antibody-interacting regions. In contrast, Alanine (A), Leucine (L), Cysteine (C), Isoleucine (I), Glycine (G), and Valine (V) were significantly underrepresented in these regions. Our results indicate that antibody-interacting regions are enriched in polar, charged, and aromatic residues, whereas non-interacting regions are predominantly composed of mildly polar and hydrophobic residues.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Human Ab Interacting	4.5	2.0	6.5	7.4	3.7	6.6	2.9	4.1	7.1	6.8	1.6	5.3	5.7	5.6	6.4	6.6	6.0	4.8	2.2	4.1
Human Ab Not-Interacting	6.5	3.3	5.2	5.9	4.0	7.3	2.3	5.4	5.5	8.8	1.7	4.9	4.7	3.9	4.5	7.1	6.4	7.1	1.8	3.4
P-Value	1e-12	4e-11	4e-6	3e-8	3e-1	3e-2	1e-3	3e-7	3e-9	8e-10	5e-1	2e-1	2e-4	1e-13	2e-14	8e-2	2e-1	1e-15	1e-2	1e-3

Figure 8: Comparison of AAC of human-specific residues which interact with the antibody and the residues that do not interact with the antibody

Next, we wanted to understand the effect of the neighboring residues on the epitope determination. Previous investigations have demonstrated that the residue's interactions is affected by its surrounding environment (Kaur and Raghava, 2003, 2004; Ansari and Raghava, 2010). To account for this local influence, overlapping sequence patterns (or sliding windows) were generated for every residue within the antigen. The details of creation of patterns is given under Section 4.3 in Materials and Methods. We demonstrate in later results (under Results section 2.2.2.2) that patterns of window length 15 best captures the effect of the neighborhood residues on the epitope prediction.

We then wanted to observe the frequency of residues in the local environment around epitopic and non-epitopic residues. Therefore, we generated the two sample logo to learn which residues are preferred at a particular position in these patterns. The patterns were created for both antibody-interacting and not interacting residues and were used to create the two sample logo shown in Figure 9. The details of the construction of the two sample logo is given under Materials and Methods section 4.5.

The analysis of the two sample logo reveals a notable compositional difference between antibody-interacting and non-interacting residues. Specifically, Alanine (A) is enriched among non-interacting residues, whereas Glutamine (Q) and Cysteine (C) exhibit a higher abundance in the antibody-interacting residue group. Additionally, our analysis from the Amino Acid Composition (AAC) comparison are observed in the central residue of the pattern in the two sample logo. The enrichment of Alanine (A) in non-interacting residues suggests its hydrophobic role, whereas the higher abundance of Glutamine (Q) and Cysteine (C) in interacting residues indicates their involvement in hydrogen bonding and electrostatic interactions, leading to binding.

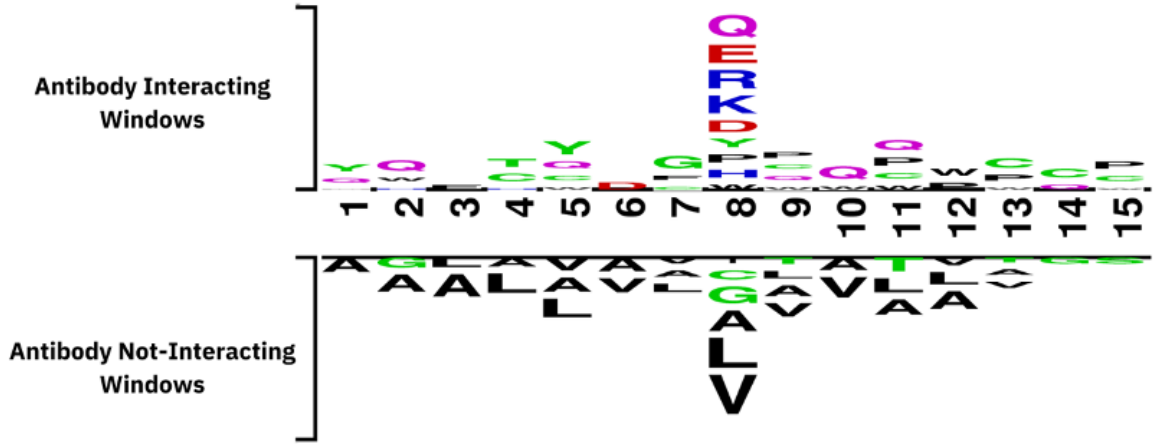


Figure 9: Two Sample Logo generated from the human-specific dataset. The antibody-interacting patterns were inputted as positive samples and the antibody not-interacting patterns which contained no epitopic residues were inputted as negative samples.

2.2.2 Development of a human-specific predictor

The previous results demonstrate the significance of species specificity and highlights the necessity for species-specific predictors. In response to this, we developed a human-specific B cell predictor, HAIRpred, which predicts conformational epitopes from the antigen sequence. **HAIRpred** (**H**uman **A**ntibody **I**nteracting **R**esidue **pred**ictor) is an ensemble machine learning model based on sequential, structural and evolutionary features derived from the antigen sequence (Sahni *et al.*, 2024). The workflow of the experiments leading to the development of HAIRpred is displayed in Figure 10.

The high-quality non-redundant dataset curated in section 2.2.1 consists of 277 antigens, ensuring that no two antigens are more than 70% similar. This dataset was used to develop the

human-specific predictor and was subsequently partitioned into training and independent test dataset in a 80:20 ratio. The training dataset contains 221 antigens , while the independent test set consists of 56 antigens.

Machine Learning Models

We developed machine learning models on our dataset to classify each residue within an input antigen sequence as either "antibody-interacting" or "non-antibody-interacting." Since most machine learning algorithms require numerical input data, it was necessary to generate features that capture antigen information in numerical vectors or matrices.

Here, a diverse set of features was created that contains sequential ,structural, and evolutionary information of antigen which can be derived from its primary sequence. We generated feature vectors on 'patterns' of different window lengths. This allowed us to also observe the effect of the neighbouring environment of each residue on the prediction. The details of 'pattern' creation is given under Section 4.3 under Materials and Methods. Feature vectors were derived from the training dataset and utilized to train the machine learning models. Each model was then evaluated for its performance using the independent test dataset.

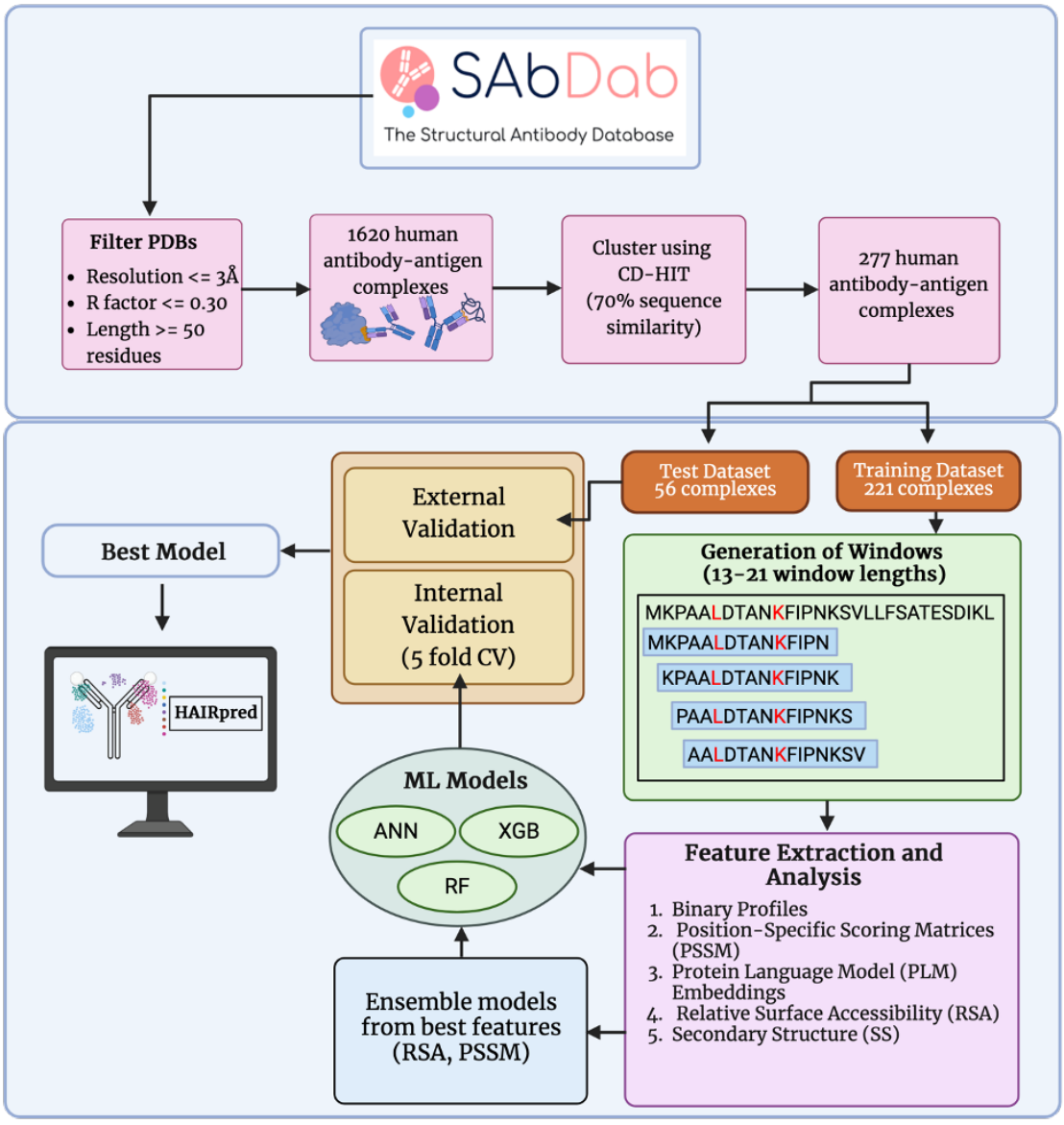


Figure 10: Workflow of the development of HAIRpred

ML models based on sequential and evolutionary features

To develop classification algorithms for antibody-interacting residues, features capturing both sequence and evolutionary information were generated for each antigen pattern. Two key feature representations used in this study were Binary Profiles and Position-Specific Scoring Matrices (PSSM). These features were designed to represent the structural and evolutionary characteristics of each residue respectively in a format suitable for machine learning models.

Binary Profiles were constructed by encoding each amino acid with a unique binary representation, resulting in a 20-length vector per residue. These were generated for each pattern of length N, resulting in feature vectors with dimensions N × 20. These features were utilized to train the models and its predictive performance is displayed in Table 1.

Position-Specific Scoring Matrices (PSSM) quantify the probability of each amino acid appearing at a given position, which is extracted from multiple sequence alignments. In this study, PSSM profiles were generated for each antigen in the dataset, segmented into patterns, and subsequently used to train machine learning models. The results of the performance of these models are reported in Table 1 and it is observed that our the PSSM profile based model obtain higher AUCROC scores than the binary profile based models. These findings are in agreement with prior studies, which have demonstrated the superior performance of PSSM-based models (Kaur and Raghava, 2003, 2004).

Table 1 : Evaluation of human-specific models using Binary and PSSM Profiles as features

Models	Performance on human-specific Independent Test Dataset				
	Sensitivity	Specificity	Accuracy	AUCROC	MCC
Window Length – 13					
Binary Profile					
Artificial Neural Networks	0.55	0.50	0.54	0.54	0.04
XGBoost	0.47	0.61	0.58	0.55	0.06
Random Forests	0.54	0.57	0.56	0.58	0.08
PSSM					
Artificial Neural Networks	0.65	0.51	0.53	0.61	0.12
XGBoost	0.66	0.56	0.58	0.66	0.17
Random Forests	0.70	0.53	0.56	0.67	0.17
Window Length – 15					
Binary Profile					
Artificial Neural Networks	0.50	0.56	0.55	0.55	0.05
XGBoost	0.58	0.52	0.54	0.57	0.08
Random Forests	0.59	0.51	0.53	0.57	0.08

PSSM					
Artificial Neural Networks	0.62	0.55	0.56	0.62	0.13
XGBoost	0.67	0.56	0.58	0.68	0.18
Random Forests	0.70	0.55	0.57	0.68	0.19
Window Length – 17					
Binary Profile					
Artificial Neural Networks	0.57	0.54	0.58	0.58	0.08
XGBoost	0.56	0.55	0.55	0.58	0.08
Random Forests	0.58	0.57	0.57	0.61	0.11
PSSM					
Artificial Neural Networks	0.62	0.54	0.55	0.62	0.12
XGBoost	0.66	0.57	0.58	0.67	0.18
Random Forests	0.67	0.55	0.57	0.67	0.17
Window Length – 19					
Binary Profile					
Artificial Neural Networks	0.55	0.52	0.52	0.54	0.05
XGBoost	0.55	0.54	0.54	0.56	0.07
Random Forests	0.56	0.57	0.56	0.59	0.10
PSSM					
Artificial Neural Networks	0.55	0.60	0.59	0.62	0.12
XGBoost	0.63	0.57	0.58	0.67	0.16
Random Forests	0.67	0.56	0.58	0.67	0.18
Window Length – 21					
Binary Profile					
Artificial Neural Networks	0.57	0.50	0.52	0.55	0.06
XGBoost	0.57	0.53	0.54	0.57	0.08
Random Forests	0.57	0.56	0.56	0.59	0.10
PSSM					
Artificial Neural Networks	0.58	0.58	0.58	0.62	0.12
XGBoost	0.66	0.58	0.60	0.67	0.19

Random Forests	0.68	0.57	0.59	0.68	0.19

ML models based on LLM embeddings and structural properties features

Next, embeddings were derived from Protein Language Models (PLMs) were used as features. Protein Language Models (PLMs) are transformer-based models that are trained on a large corpus of protein data and thereby, encapsulate contextual relationships between amino acid residues and encode evolutionary and functional information (Rao et al., 2020). Here, we used ESM2 (Lin et al., 2023) and Encoder-only ProtTrans (ProtT5-XL-UniRef50) (Elnaggar et al., 2022) embeddings of the patterns as features to train prediction machine learning models.

In addition to embeddings, PLMs can predict structural properties of proteins, including Secondary Structure (SS) and Relative Solvent Accessibility (RSA). RSA represents the proportion of a residue's surface exposed to the solvent which indicates its potential involvement in interactions. SS describes the local folding pattern of the protein chain, typically classified as alpha-helices, beta-sheets, or coils, which is provide insights into protein function and interactions. Here, RSA features were created using ESMFold and the SS features were derived from both ESMFold and ProtTrans and divided into patterns to train the machine learning models.

Similar to previous experiments, the training dataset which comprised of patterns was used to generate feature vectors which was then inputted to train the machine learning models. Each model was then evaluated for its performance using the independent test dataset. The performance metrics of these models are shown in Table 2. We observed that the models based of RSA feature vectors exhibit a stronger predictive performance than other features. RSA is observed to a good indicator for predicting B Cell epitopes.

Table 2 : Evaluation of human-specific models using PLM-Derived Embeddings, RSA, and SS as features

Models	**Performance on human-specific Independent Test Dataset**				
	Sensitivity	**Specificity**	**Accuracy**	**AUCROC**	**MCC**
Window Length – 13					
PLM Embeddings					
ESM-2					
Artificial Neural Networks	0.41	0.62	0.58	0.52	0.02
XGBoost	0.44	0.62	0.58	0.54	0.05
Random Forests	0.40	0.64	0.59	0.53	0.03
ProtT5					
Artificial Neural Networks	0.63	0.46	0.49	0.55	0.07
XGBoost	0.51	0.57	0.56	0.56	0.06

Random Forests	0.47	0.60	0.58	0.56	0.06
Relative Solvent Accessibility					
ESMFold					
Artificial Neural Networks	0.82	0.42	0.50	0.64	0.20
XGBoost	0.86	0.38	0.47	0.65	0.20
Random Forests	0.85	0.39	0.48	0.66	0.20
Secondary Structure					
ESMFold					
Artificial Neural Networks	0.57	0.49	0.50	0.54	0.04
XGBoost	0.54	0.55	0.55	0.56	0.07
Random Forests	0.55	0.54	0.54	0.56	0.07
ProtT5					
Artificial Neural Networks	0.68	0.40	0.45	0.56	0.07
XGBoost	0.61	0.50	0.52	0.57	0.09
Random Forests	0.59	0.51	0.53	0.57	0.08
Window Length – 15					
PLM Embeddings					
ESM-2					
Artificial Neural Networks	0.56	0.47	0.48	0.52	0.02
XGBoost	0.56	0.49	0.51	0.54	0.04
Random Forests	0.42	0.63	0.59	0.54	0.04
ProtT5					
Artificial Neural Networks	0.53	0.53	0.53	0.54	0.05
XGBoost	0.53	0.56	0.55	0.56	0.07
Random Forests	0.49	0.60	0.58	0.57	0.07
Relative Solvent Accessibility					
ESMFold					
Artificial Neural Networks	0.81	0.43	0.50	0.64	0.19
XGBoost	0.86	0.37	0.47	0.65	0.19
Random Forests	0.79	0.45	0.51	0.66	0.19

Secondary Structure					
ESMFold					
Artificial Neural Networks	0.52	0.54	0.53	0.54	0.04
XGBoost	0.55	0.52	0.52	0.56	0.05
Random Forests	0.57	0.53	0.54	0.57	0.08
ProtT5					
Artificial Neural Networks	0.61	0.47	0.50	0.56	0.07
XGBoost	0.60	0.50	0.52	0.57	0.08
Random Forests	0.61	0.49	0.51	0.56	0.08
Window Length – 17					
PLM Embeddings					
ESM-2					
Artificial Neural Networks	0.48	0.61	0.59	0.57	0.07
XGBoost	0.60	0.54	0.55	0.60	0.10
Random Forests	0.46	0.69	0.65	0.61	0.11
ProtT5					
Artificial Neural Networks	0.52	0.56	0.55	0.56	0.06
XGBoost	0.61	0.48	0.50	0.57	0.07
Random Forests	0.47	0.65	0.61	0.58	0.09
Relative Solvent Accessibility					
ESMFold					
Artificial Neural Networks	0.73	0.49	0.54	0.64	0.18
XGBoost	0.52	0.68	0.65	**0.66**	0.16
Random Forests	0.78	0.46	0.52	**0.67**	0.20
Secondary Structure					
ESMFold					
Artificial Neural Networks	0.54	0.53	0.53	0.55	0.05
XGBoost	0.16	0.89	0.75	0.57	0.06
Random Forests	0.22	0.85	0.73	0.59	0.08
ProtT5					

Artificial Neural Networks	0.55	0.52	0.53	0.55	0.06
XGBoost	0.49	0.60	0.58	0.57	0.07
Random Forests	0.52	0.58	0.57	0.57	0.08
Window Length – 19					
PLM Embeddings					
ESM-2					
Artificial Neural Networks	0.61	0.43	0.46	0.52	0.03
XGBoost	0.55	0.52	0.52	0.54	0.05
Random Forests	0.38	0.69	0.63	0.55	0.06
ProtT5					
Artificial Neural Networks	0.61	0.47	0.50	0.55	0.06
XGBoost	0.56	0.52	0.53	0.56	0.06
Random Forests	0.43	0.66	0.62	0.57	0.08
Relative Solvent Accessibility					
ESMFold					
Artificial Neural Networks	0.82	0.42	0.49	0.64	0.19
XGBoost	0.86	0.38	0.47	0.66	0.20
Random Forests	0.79	0.45	0.51	0.67	0.19
Secondary Structure					
ESMFold					
Artificial Neural Networks	0.65	0.41	0.45	0.55	0.04
XGBoost	0.56	0.54	0.54	0.57	0.08
Random Forests	0.56	0.55	0.55	0.58	0.09
ProtT5					
Artificial Neural Networks	0.53	0.56	0.56	0.56	0.07
XGBoost	0.58	0.51	0.52	0.57	0.07
Random Forests	0.59	0.52	0.53	0.57	0.08
Window Length – 21					
PLM Embeddings					
ESM-2					

Artificial Neural Networks	0.52	0.51	0.51	0.52	0.03
XGBoost	0.51	0.53	0.53	0.53	0.03
Random Forests	0.37	0.68	0.62	0.55	0.05
ProtT5					
Artificial Neural Networks	0.56	0.55	0.55	0.58	0.09
XGBoost	0.56	0.51	0.52	0.56	0.05
Random Forests	0.44	0.68	0.64	0.59	0.10
Relative Solvent Accessibility					
ESMFold					
Artificial Neural Networks	0.82	0.41	0.49	0.64	0.19
XGBoost	0.85	0.39	0.48	0.65	0.20
Random Forests	0.79	0.45	0.51	0.67	0.19
Secondary Structure					
ESMFold					
Artificial Neural Networks	0.57	0.51	0.52	0.56	0.07
XGBoost	0.53	0.55	0.55	0.56	0.06
Random Forests	0.55	0.56	0.56	0.58	0.09
ProtT5					
Artificial Neural Networks	0.50	0.59	0.57	0.57	0.07
XGBoost	0.58	0.53	0.54	0.58	0.08
Random Forests	0.60	0.53	0.54	0.58	0.09

2.2.2.1 Ensemble Models

We observed, from our experiments with different features, that models trained on PSSM profile and Relative Solvent Accessibility (RSA) outperform those based on other features. To integrate the strength of both features, we created an ensemble model that combines information from both the models. The ensemble model was developed using a straightforward approach: the predicted probabilities from individual models were averaged, and the final classification was based on these averaged probabilities. The performance metrics of ensemble models are displayed in Table 3.

Table 3 : Impact of varying window lengths on the human-specific Ensemble Model's performance on the independent test dataset

Models	Performance on human-specific Independent Test Dataset				
	Sensitivity	**Specificity**	**Accuracy**	**AUCROC**	**MCC**
Window Length – 13					
Artificial Neural Networks	0.66	0.57	0.59	0.67	0.18
XGBoost	0.72	0.54	0.57	0.70	0.20
Random Forests	0.79	0.48	0.54	0.71	0.22
Window Length – 15					
Artificial Neural Networks	0.65	0.57	0.58	0.66	0.17
XGBoost	0.73	0.53	0.57	0.71	0.21
Random Forests	0.78	0.51	0.56	**0.72**	0.23
Window Length – 17					
Artificial Neural Networks	0.65	0.58	0.59	0.67	0.18
XGBoost	0.71	0.56	0.59	0.70	0.21
Random Forests	0.76	0.52	0.57	0.71	0.22
Window Length – 19					
Artificial Neural Networks	0.65	0.57	0.58	0.66	0.17
XGBoost	0.71	0.56	0.59	0.70	0.21
Random Forests	0.76	0.53	0.57	0.71	0.23
Window Length – 21					
Artificial Neural Networks	0.59	0.66	0.65	0.67	0.20
XGBoost	0.72	0.56	0.59	0.71	0.22
Random Forests	0.75	0.53	0.57	0.71	0.22

It is observed that the Random Forest ensemble model based on Window Length 15 has the highest prediction performance with AUCROC score of 0.72 and MCC score of 0.23. The ROC (Receiver Operating Characteristic) curve is displayed in Figure 11. This ensemble model, consisting of Random Forest classifiers trained on Relative Solvent Accessibility (RSA) and Position-Specific Scoring Matrices (PSSM) with a window length of 15, demonstrated the highest predictive performance for identifying antibody-interacting regions within antigens. and is therefore named **HAIRpred** (**H**uman **A**ntibody **I**nteracting **R**esidue **pred**ictor).

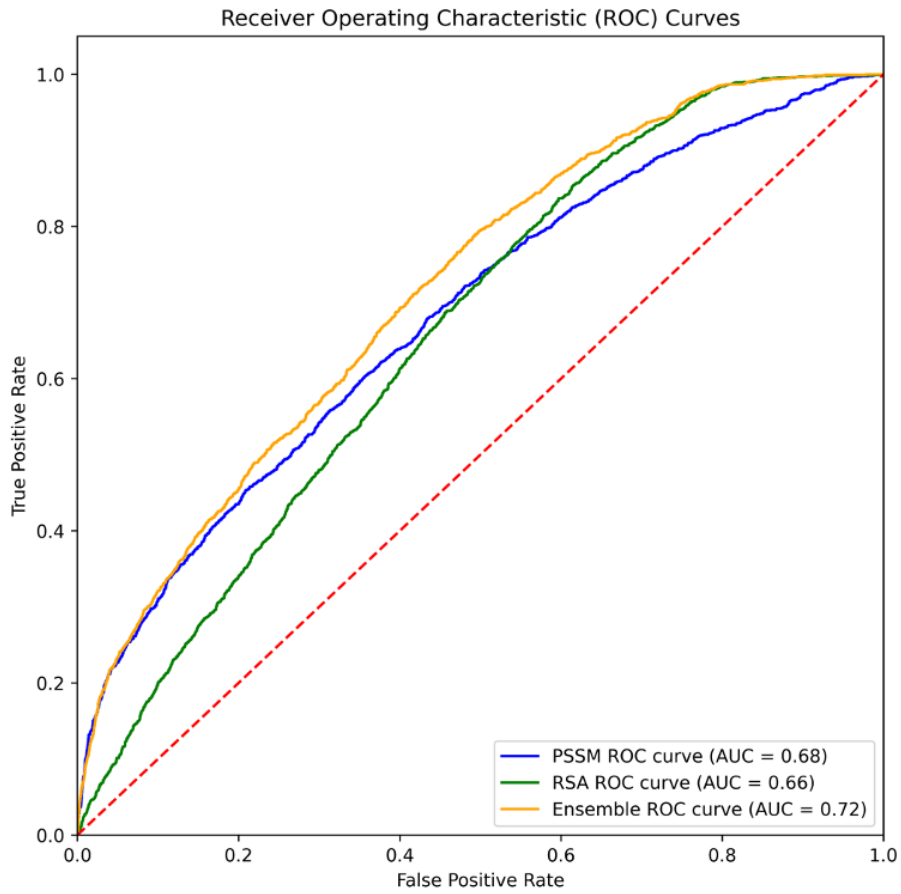


Figure 11: Receiver Operating Curve (ROC) curve of the best performing ensemble model along with ROC curves of the contributing features

2.2.3 Comparison between HAIRpred and existing methods

It is important for any study developing prediction tools to be benchmarked against all other existing tools. As we can tell, there is no existing method in literature which has been created to predict conformational B-cell epitopes for humans. Therefore, a direct comparison of HAIRpred with other prediction methods is not feasible.

However, several methods have been proposed in the literature for predicting conformational B-cell epitopes without species specificity. To establish a benchmark for our method, we evaluated the predictive performance of these generalized B-cell epitope prediction approaches on the independent human-specific test dataset used in this study. The methods considered for creating a benchmark included CBTOPE (Ansari and Raghava, 2010), Bepipred-2.0 (Jespersen et al., 2017), Bepipred-3.0 (Clifford et al., 2022), SEMA-1d (Ivanisenko et al., 2024), and Epidope (Collatz et al., 2021a). We display the results in Table 4. We observe that the highest performance of these generalized methods was the AUROC score of 0.62, while HAIRpred achieves the AUROC score of 0.72 on the independent test dataset.

In addition to this, we also wanted to understand how our method would perform on other hosts. Therefore, we used HAIRpred to predict the epitopic sites on the mouse-specific test dataset. HAIRpred performs worse on the mouse dataset with a lower AUCROC score of 0.68. This observation of models performing better on the train host's data further supports the necessity of species-specific models to predict antibody-interacting residues.

Table 4 : Comparison of HAIRpred and other existing methods on the independent test dataset

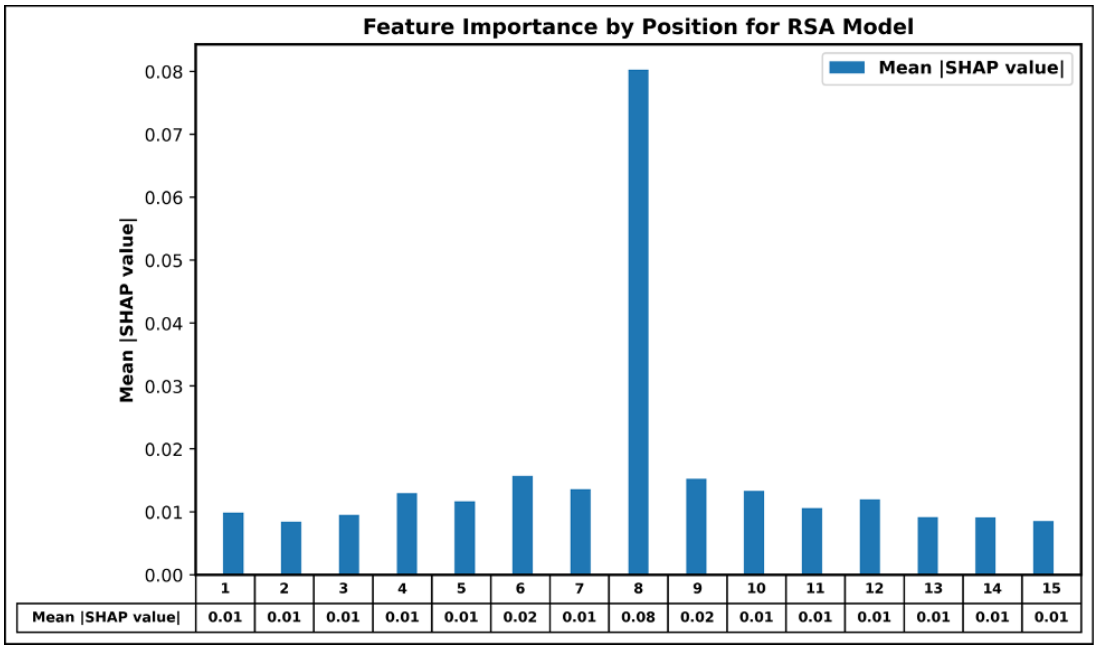
Model	Year	Sensitivity	Specificity	Accuracy	AUCROC	MCC
CBTOPE	2010	0.09	0.94	0.80	0.52	0.04
Bepipred - 2.0	2017	0.77	0.22	0.31	0.50	-0.01
Epidope	2021	0.03	0.96	0.81	0.50	-0.02
Bepipred - 3.0	2022	0.65	0.58	0.59	0.62	0.17
Sema - 1d	2024	0.65	0.60	0.61	0.62	0.18
HAIRpred on Mouse Dataset	-	0.71	0.54	0.56	0.68	0.17
HAIRpred	-	0.78	0.51	0.56	0.72	0.23

2.2.4 SHAP Analysis

We created a human-specific B-cell epitope predictor by developing an ensemble of two Random Forest models which are trained on Position-Specific Scoring Matrix (PSSM) and Relative Solvent Accessibility (RSA). To increase the interpretability of our model, we used the SHapley Additive exPlanations (SHAP) method to analyse our ensemble mode. SHAP is a game theory based approach to understand the model and it generates Shapley values for each position in a feature vector. These Shapley values indicate the importance of that position in the feature vector in the final prediction (Lundberg and Lee, 2017).

We calculates the Shapley Values for the two constituent random forests using the SHAP TreeExplainer from the python shap module. The Shapley values gave us the importance of the residue position in the pattern's feature vector. Figure 12 shows the Shapley Values for each residue position in the pattern for the RSA trained RF model and these values given an insight into each position's importance. The SHAP analysis revealed that the central position in the pattern's RSA feature vector has the highest influence in the final prediction. This is consistent with the expectation that the central residue is critical in the determination of its interaction with the antibody. Since RSA represents the extent of solvent exposure, the model likely prioritizes the central residue's solvent exposure to distinguish interacting and non-interacting residues.

The SHAP analysis for the PSSM-based model identified the most influential positions in the pattern's feature vector which are, in descending order, 8Q, 8E, 8R, 8K, 8H, 8N, 8C, 8D, 8A and 11A. These positions are here displayed with the number as the row number of the feature and the letter as the amino acid column in the PSSM feature matrix. Similar to the RSA model, the central position in the pattern's feature vector has a large imoact in the final prediction. Additionally, neighbouring residues also exhibit moderate Shapley values, indicating that their contributions provide contextual information that refines the model's decision-making.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Mean |SHAP value|	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.08	0.02	0.01	0.01	0.01	0.01	0.01	0.01

Figure 12: Comparison of the Shapley Values of the RSA model generated for each residue locus in the pattern

2.2.5 Web Server Implementation

HAIRpred was integrated into a webserver to improve accessibility and provide ease of us to the scientific research community. This platform is designed to facilitate easy and efficient prediction of epitopic residues in antigen sequences, making it accessible to researchers without requiring extensive bioinformatics expertise. The best-performing prediction models are hosted in the 'Predict' module. In this module, users upload or type in their antigen sequences in the FASTA format or just the sequence of amino acids in one-letter format. The server then processes these sequences using the prediction model to identify antibody-interacting sites within each submission. The webserver's 'Predict' module is displayed in Figure 13.

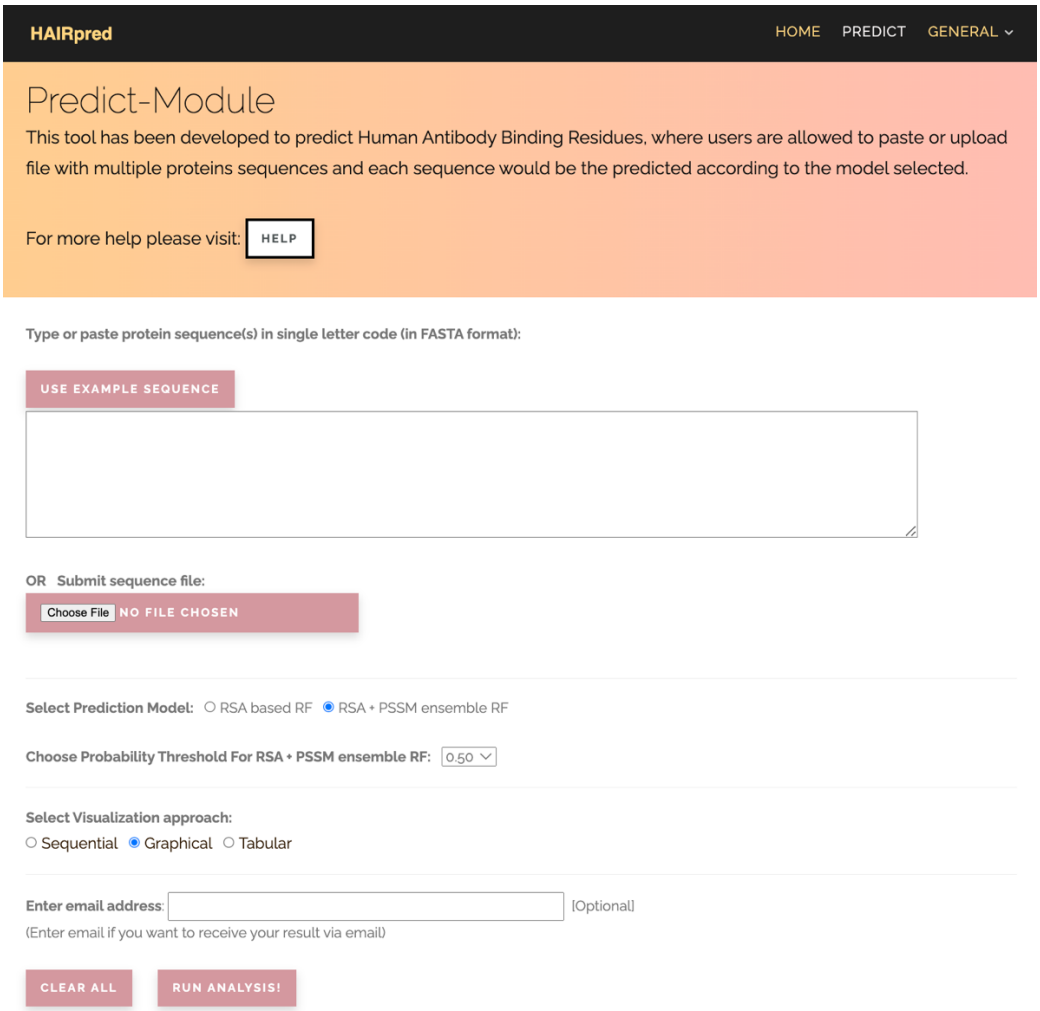


Figure 13: Predict Module of HAIRpred webserver

We designed the output of HAIRpred to be both informative and user-friendly and therefore the 'Predict' module offers three different visualisation approach : sequential, graphical and tabular approach. The sequential approach presents prediction results in a linear format where antibody-interacting residues are highlighted within the input antigen sequence by color-coding to distinguish between interacting and non-interacting sites which allows for easy tracking of interaction patterns. The graphical approach provides an intuitive, visual representation of interacting residues and provides an interactive approach to exploring binding patterns. The tabular approach organizes the results into a structured table and displays key details such as residue position, predicted HAIRpred score, and classification (interacting vs. non-interacting). The visualisation approaches are displayed in Figure 14.

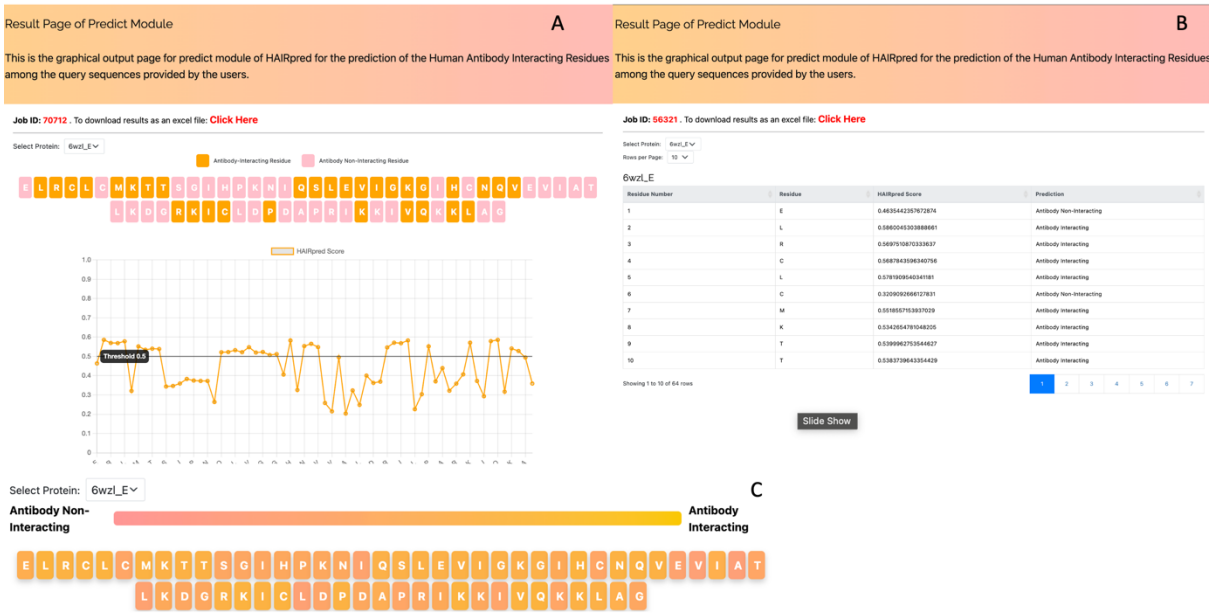


Figure 14: Visualisation Approaches of HAIRpred webserver - A) Graphical, B) Tabular and C) Sequential

The server also has the facility to download the train and the independent test data used to create the model. The server also has a 'Help' section to assist the users navigating the features of the platform effectively. The server has been created using HTML, PHP and Javascript to ensure a robust and user-friendly interface which will allow it to run across various devices including laptops, mobile phones and iPads.

To further enhance accessibility, we have created a standalone version of HAIRpred. This package enables researchers to generate predictions without requiring an internet connection. The standalone software is provided on the webserver and the GitHub page (https://github.com/raghavagps/HAIRpred) and can additionally be downloaded as a Python package via pip (pip install hairpred). The links of the python package and the standalone are also listed in the webserver. To provide an estimate of the speed of our standalone software, we performed a time analysis of the standalone software which is displayed in Table 5. We measured the time it took for the system to run different number of sequences of varying size. These experiments were performed on a system with i5 14th Generation with 32GB RAM.

This initiative ensures that researchers worldwide can use state-of-the-art prediction models without technical barriers which will help promote innovation in the immunological research field. The open-source web server is freely accessible at https://webs.iiitd.edu.in/raghava/hairpred/.

Table 5 : Time Analysis of the standalone software

		Number of Sequences				
		1	**10**	**25**	**50**	**100**
Size of Sequences (bp)	**60**	7.61s	35.80s	83.71s	161.90s	318.36s
	99	8.46s	43.68s	133.87s	199.88s	395.26s
	206	11.74s	74.45s	188.16s	357.88s	702.79s
	307	13.63s	91.71s	233.22s	456.11s	891.24s

2.3 Development of mouse-specific B cell epitope predictor

Mice have long served as indispensable models in preclinical research due to its genetic similarity to humans and short lifespan (Boguski, 2002). Murine models are the standard experimental models across a broad spectrum of biomedical fields. Despite their widespread use, existing computational frameworks for predicting antibody-interacting residues rely on generalized models that overlook the distinct immunological characteristics of murine systems. This limitation highlights the need for the development of a predictive tool specifically tailored to capture the unique antigen-antibody interaction patterns observed in mice.

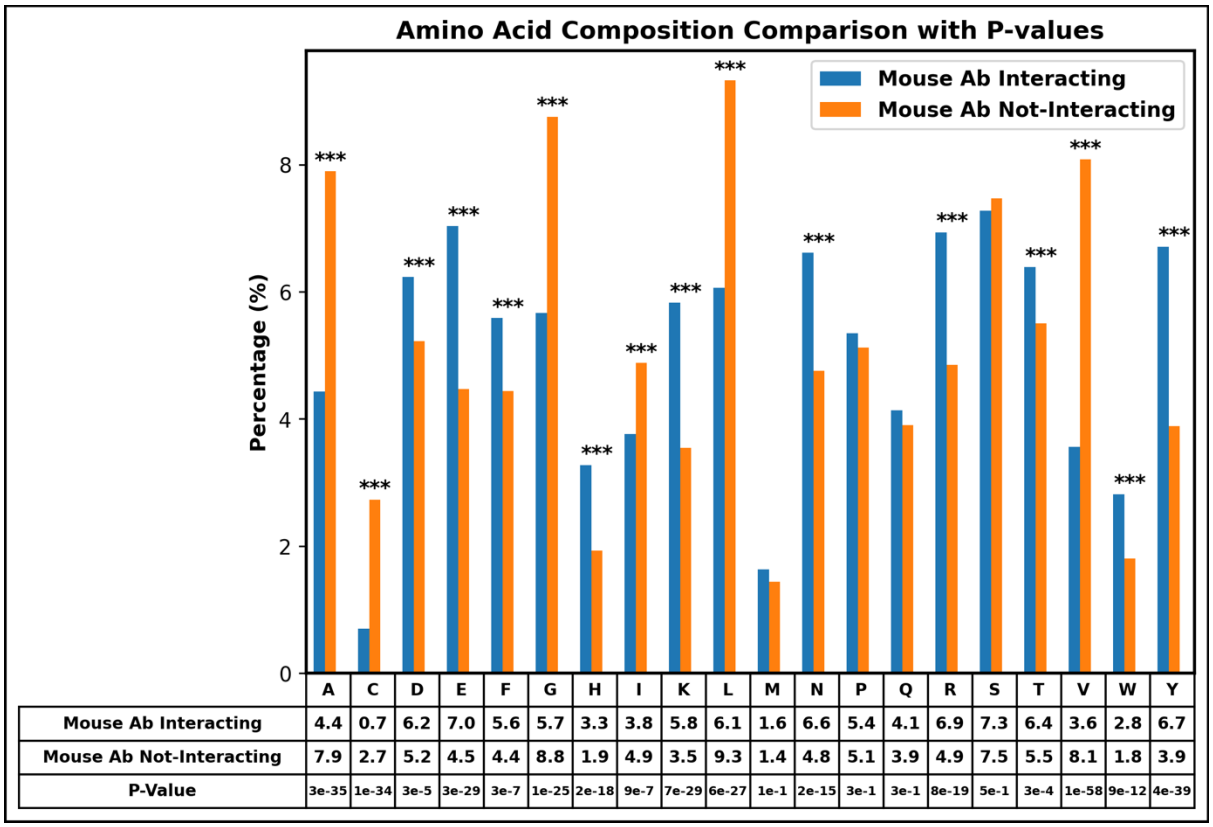
Such an approach is particularly important in studies which regularly depend on murine models for early-stage discovery and validation including vaccine development, infectious disease studies, cancer immunotherapy, and autoimmune disorder research. The ability to reliably predict the interaction sites within murine antigens offers the practical benefit of reducing unnecessary experimental iterations, thereby conserving both time and resources.

To support our experimental research community, this study develops a mouse-specific antibody interaction prediction tool. This tool is developed on a similar approach to the human-specific tool with strong focus on ML models based on sequential, structural and evolutionary features derived from the mouse antigen's sequence. This section has been divided into three subsections : (i) Data analysis, (ii) Development of a mouse-specific predictor (iii) Benchmarking

2.3.1 Data Analysis

Similar to the methodology of the development of HAIRpred, we downloaded mouse antibody-antigen complexes from the SAbdab database (Dunbar *et al.*, 2014). The complexes were similarly filtered to get a high quality dataset of 290 mouse-specific antibody-antigen complexes. This dataset has limited number of antibody-antigen complexes with only 94 non redundant antigen clusters after clustering antigens according to their sequence identity using CD-HIT (Fu et al., 2012) with a 70% sequence identity threshold. Therefore, instead of only using the representative antigen of a cluster, we use all antibody-antigen complexes in model development. However, to ensure that the model does not become biased or overfit due to high sequence similarity between antigens in the same cluster, we implement a cluster-based data splitting strategy which is described in Materials and Methods Section 4.6.

To analyse amino acid preferences in antibody-interacting versus non-interacting regions, we calculated and compared the AAC of the interacting and not-interacting residues. The results, presented in Figure 15, were statistically evaluated using the chi-square test. The analysis indicates that residues such as Aspartic Acid (D), Phenylalanine (F), Glutamic Acid (E), Histidine (H), Tryptophan (W), Methionine (M), and Tyrosine (Y) are significantly more abundant in the antibody-interacting residues. In contrast, Alanine (A), Cysteine (C), Isoleucine (I), Glycine (G), and Valine (V) are significantly enriched in non-interacting residues. These findings suggest that antibody-interacting sites are predominantly characterized by polar, charged, and aromatic residues, while non-interacting regions are enriched in small, hydrophobic, and aliphatic residues.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Mouse Ab Interacting	4.4	0.7	6.2	7.0	5.6	5.7	3.3	3.8	5.8	6.1	1.6	6.6	5.4	4.1	6.9	7.3	6.4	3.6	2.8	6.7
Mouse Ab Not-Interacting	7.9	2.7	5.2	4.5	4.4	8.8	1.9	4.9	3.5	9.3	1.4	4.8	5.1	3.9	4.9	7.5	5.5	8.1	1.8	3.9
P-Value	3e-35	1e-34	3e-5	3e-29	3e-7	1e-25	2e-18	9e-7	7e-29	6e-27	1e-1	2e-15	3e-1	3e-1	8e-19	5e-1	3e-4	1e-58	9e-12	4e-39

Figure 15: Comparison of AAC in mouse-specific data of residues that interact with the antibody with residues that do not interact with the antibody

Next, we wanted to see how mouse-specific B Cell epitopes differ from human-specific epitopes. We had shown in Section 2.1 that human and mouse epitopes have different amino acid composition (AAC) (Figure 7). It was revealed that the human-specific epitopes prefer amino acids like Cysteine (C), Arginine (R), Glutamine (Q), Tryptophan (W), and Tyrosine (Y) while mouse-specific epitopes prefer Glycine (G), Asparagine (N), Lysine (K), and Serine (S).

Furthermore, We wanted to observe if the local environment around epitopic and non-epitopic residues for mouse-specific data differs from the human data. Therefore, we created the two sample logo to understand which residues are preferred at a particular position for the mouse data for pattern length 15. The two sample logo generated is shown in Figure 16.

We observe that in the antibody-interacting windows (above the axis), the central residue, is predominantly occupied by charged and polar residues, most notably lysine (K), arginine (R),

glutamic acid (E), and tyrosine (Y). These residues bring a combination of positive and negative charges, as well as aromatic and polar side chains, enabling a rich potential for electrostatic interactions, hydrogen bonding, and π-π stacking. In the neighbouring residues, we see an enrichment of small polar residues such as serine (S) and threonine (T), along with glutamine (Q) and asparagine (N). These neighbouring positions appear to form a supportive, flexible environment and the glycine (G) in nearby positions further enhances local flexibility. On the other hand, the non-interacting windows (below the axis) are dominated across the sequence by hydrophobic and aliphatic residues, particularly alanine (A), leucine (L) and valine (V). These residues favor the buried, core-like regions of the antigen.

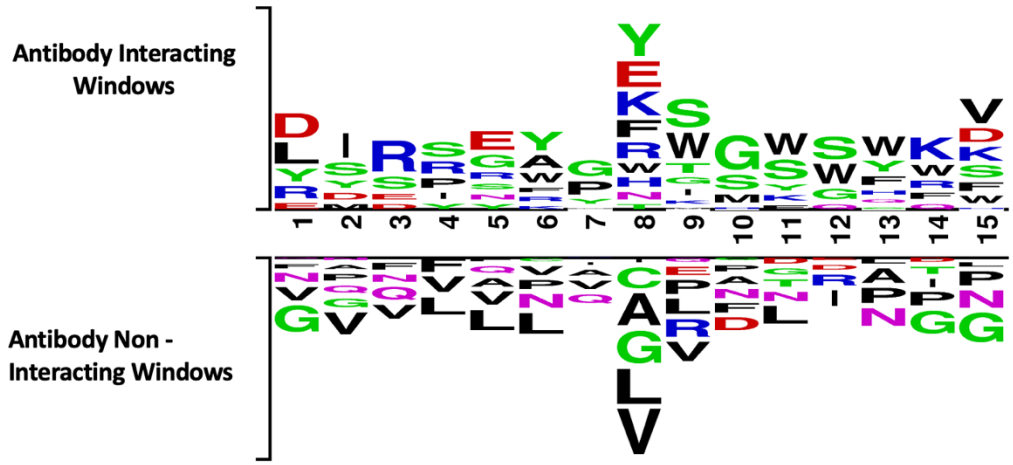


Figure 16: Comparison of AAC of mouse-specific residues which interact with the antibody and the residues that do not interact with the antibody

The mouse-specific two sample logo differs from the human-specific logo in its complexity and the diversity of residue representation across the sequence window. While the human-specific logo is dominated by a strong, localized enrichment of charged and polar residues at the central position, the mouse-specific logo displays a more distributed pattern, with a broader variety of residues, including polar (serine, tyrosine), charged (lysine, glutamic acid), and aromatic (phenylalanine, tryptophan) residues spread across multiple neighbouring positions. Additionally, the non -interacting pattern in the human-specific logo is overwhelmingly hydrophobic and centered, whereas in the mouse-specific logo, hydrophobic residues are consistently present throughout the non-interacting regions. These results reinforce the hypothesis that there exists species-specific difference in the interaction of antibody and antigens and therefore, highlights the need for a mouse-specific predictor.

2.3.2 Development of a mouse-specific predictor

There is an important need for a mouse-specific B Cell epitope predictor. In response to this need, we developed machine learning models that predict mouse-specific conformational B Cell epitopes from the antigen's primary sequence. These models are based on features that generate sequential, structural and evolutionary information from the antigen's primary sequence.

The dataset used for training and testing these models are experimentally derived mouse-specific antibody-antigen complexes downloaded from the SAbDab database (Dunbar *et al.*, 2014). As mentioned previously, the filtered dataset comprises of only 290 mouse-specific antibody-antigen complexes with just 94 clusters at 70% sequence identity. Therefore, instead of limiting the dataset to only the representative antigen from each cluster (which would further reduce the already small dataset), we include all antibody-antigen complexes from each cluster during model development. However, to ensure that the model does not become biased or overfit due high sequence similarity among the antigens, we implement a cluster-based data splitting strategy. In this approach, we assign entire clusters exclusively to either the training or testing set, ensuring that no antigen cluster (sharing ≥70% sequence identity) is split across both. This prevents highly similar antigens from appearing in both sets, which would lead to overly optimistic performance metrics and compromise the ability of the model to generalise. The resulting training and independent testing dataset comprises of 232 and 58 mouse-specific experimentally derived antibody-antigen complexes.

Furthermore, our five-fold cross-validation is also performed following this cluster-based splitting approach. During cross validation, the training dataset is split into disjoin subsets (folds) and in each iteration, complexes from one fold is used for validation while the other folds are used for training the model. Here, we ensure that no antigen cluster is shared between training and validation folds, thereby preventing the mixing of highly similar antigens across the folds. This ensures a more realistic and reliable evaluation of the model's generalizability and its performance on unseen and diverse antigen sequences.

2.3.2.1 Machine Learning Models

Similar to the development methodology of HAIRpred, we create machine learning models trained on our dataset which can classify each residue as either "antibody-interacting" or "non-antibody-interacting." These machine learning models are based on features which encode information about the antigen. In this study, we trained our models on features similar to ones we used for HAIRpred and they encode sequential, structural and evolutionary information of an antigen from its sequence. These features are calculated on ‘patterns’ of different window lengths to understand the effect of the local environment. Then, these features are used to train the machine learning models. The training is performed using five-fold cross validation which follows cluster-based data splitting strategy. Each model was then evaluated for its performance using the independent test dataset.

ML models based on sequential and evolutionary features

Similar to the human-specific models, we developed binary classification models to identify mouse-specific antibody-interacting residues. First, we developed machine learning models on features capturing sequential as well as evolutionary information from the antigen sequence. These features include Binary Profiles and Position-Specific Scoring Matrices (PSSM).

Binary Profiles were constructed by encoding each amino acid with a unique binary representation, resulting in a 20-length vector per residue. These were created for each pattern of length N, resulting in feature vectors with dimensions $N \times 20$. These features were utilized to train the Random Forest (RF), XGBoost (XGB), and Artificial Neural Network (ANN) models and its predictive performance are displayed in Table 6. It is observed that the models trained on binary profile has the highest performance at AUROC at 0.56 and MCC at 0.06.

Position-Specific Scoring Matrices (PSSM) quantify the probability of each amino acid appearing at a given position, which is extracted from multiple sequence alignments. In this study, PSSM profiles were generated for each antigen in the dataset, segmented into patterns, and subsequently used to train machine learning models. The results of the performance of these models are reported in Table 6 and it is observed that our the PSSM profile based model obtain higher AUCROC scores, with AUROC of 0.56 and MCC of 0.06, than the binary profile based models. These findings are in agreement with the prior literature, which have demonstrated the superior performance of PSSM-based models (Kaur and Raghava, 2003, 2004).

Table 6 : Evaluation of mouse-specific models using Binary and PSSM Profiles as features

Models	Performance on human-specific Independent Test Dataset				
	Sensitivity	Specificity	Accuracy	AUCROC	MCC
Window Length – 13					
Binary Profile					
Artificial Neural Networks	0.52	0.55	0.52	0.55	0.04
XGBoost	0.48	0.61	0.59	0.55	0.05
Random Forests	0.44	0.64	0.62	0.56	0.05
PSSM					
Artificial Neural Networks	0.53	0.49	0.49	0.52	0.01
XGBoost	0.62	0.47	0.49	0.57	0.06
Random Forests	0.74	0.40	0.43	0.59	0.09
Window Length – 15					
Binary Profile					
Artificial Neural Networks	0.49	0.57	0.56	0.54	0.04
XGBoost	0.44	0.62	0.60	0.55	0.04
Random Forests	0.42	0.65	0.62	0.56	0.04
PSSM					
Artificial Neural Networks	0.21	0.84	0.78	0.54	0.05
XGBoost	0.63	0.47	0.49	0.58	0.06
Random Forests	0.73	0.41	0.44	0.58	0.09
Window Length – 17					
Binary Profile					

Artificial Neural Networks	0.52	0.53	0.53	0.53	0.03
XGBoost	0.42	0.63	0.61	0.55	0.03
Random Forests	0.41	0.68	0.65	0.56	0.06
PSSM					
Artificial Neural Networks	0.52	0.52	0.52	0.54	0.03
XGBoost	0.63	0.48	0.50	0.58	0.07
Random Forests	0.74	0.41	0.45	0.59	0.10
Window Length – 19					
Binary Profile					
Artificial Neural Networks	0.55	0.50	0.50	0.54	0.03
XGBoost	0.43	0.62	0.60	0.55	0.03
Random Forests	0.44	0.63	0.61	0.56	0.05
PSSM					
Artificial Neural Networks	0.48	0.52	0.51	0.50	0.00
XGBoost	0.59	0.48	0.49	0.56	0.04
Random Forests	0.74	0.40	0.44	0.59	0.09
Window Length – 21					
Binary Profile					
Artificial Neural Networks	0.55	0.53	0.53	0.55	0.05
XGBoost	0.45	0.62	0.60	0.55	0.04
Random Forests	0.39	0.68	0.65	0.56	0.04
PSSM					
Artificial Neural Networks	0.56	0.51	0.51	0.55	0.04
XGBoost	0.65	0.47	0.49	0.58	0.07
Random Forests	0.75	0.41	0.44	0.60	0.10

ML models based on LLM embeddings and structural properties features

Next, embeddings were derived from Protein Language Models (PLMs) were used as features. Similar to the human-specific models, we used embeddings generated from ESM2

and Encoder-only ProtTrans as features. We generated embeddings for the patterns and used the embeddings to train the machine learning models. Then, we evaluated these models on the mouse independent test dataset and the results are displayed in Table 6. The highest performance of the models generated on LLM embeddings are from ProtT5 embeddings with AUROC of 0.57 and MCC of 0.06.

Furthermore, we wanted to use structural information to develop the machine learning models. To achieve this, we used PLMs to generate Relative Solvent Accessibility (RSA) and Secondary Structure (SS). RSA was generated by predicting structure using ESMFold and then implementing the DSSP algorithm. Similarly, SS was generated using ESMFold. Additionally, we predicted the SS from ProtTrans as well. Similar to previous experiments, the training dataset which comprised of patterns was used to generate feature vectors which was then inputted to train the machine learning models. Each model was then evaluated for its performance using the independent test dataset. The performance metrics of these models are shown in Table 7. It is observed that the RSA based models perform significantly better than other features with maximum AUROC of 0.67.

Table 7 : Evaluation of mouse-specific models using PLM-Derived Embeddings, RSA and SS as features

Models	Performance on human-specific Independent Test Dataset				
	Sensitivity	**Specificity**	**Accuracy**	**AUCROC**	**MCC**
Window Length – 13					
PLM Embeddings					
ESM-2					
Artificial Neural Networks	0.49	0.52	0.51	0.51	0.01
XGBoost	0.61	0.39	0.41	0.52	0.00
Random Forests	0.75	0.27	0.32	0.51	0.01
ProtT5					
Artificial Neural Networks	0.56	0.50	0.50	0.55	0.03
XGBoost	0.67	0.43	0.45	0.57	0.06
Random Forests	0.77	0.29	0.35	0.55	0.04
Relative Solvent Accessibility					
ESMFold					
Artificial Neural Networks	0.72	0.54	0.56	0.67	0.16
XGBoost	0.63	0.57	0.58	0.65	0.13
Random Forests	0.64	0.57	0.58	0.65	0.14

Secondary Structure					
ESMFold					
Artificial Neural Networks	0.59	0.46	0.47	0.55	0.03
XGBoost	0.38	0.68	0.64	0.53	0.04
Random Forests	0.52	0.52	0.52	0.53	0.02
ProtT5					
Artificial Neural Networks	0.49	0.54	0.53	0.53	0.02
XGBoost	0.62	0.40	0.43	0.53	0.01
Random Forests	0.60	0.43	0.45	0.52	0.02
Window Length – 15					
PLM Embeddings					
ESM-2					
Artificial Neural Networks	0.61	0.44	0.46	0.52	0.03
XGBoost	0.62	0.38	0.40	0.50	-0.01
Random Forests	0.72	0.30	0.34	0.51	0.01
ProtT5					
Artificial Neural Networks	0.53	0.52	0.52	0.54	0.03
XGBoost	0.63	0.43	0.45	0.54	0.04
Random Forests	0.80	0.28	0.33	0.55	0.05
Relative Solvent Accessibility					
ESMFold					
Artificial Neural Networks	0.72	0.54	0.56	0.67	0.16
XGBoost	0.66	0.55	0.57	0.65	0.14
Random Forests	0.63	0.58	0.59	0.65	0.13
Secondary Structure					
ESMFold					
Artificial Neural Networks	0.60	0.44	0.46	0.54	0.02
XGBoost	0.56	0.48	0.49	0.54	0.02
Random Forests	0.50	0.53	0.52	0.53	0.02

ProtT5					
Artificial Neural Networks	0.50	0.54	0.53	0.54	0.03
XGBoost	0.62	0.41	0.44	0.52	0.02
Random Forests	0.61	0.42	0.44	0.52	0.02
Window Length – 17					
PLM Embeddings					
ESM-2					
Artificial Neural Networks	0.46	0.52	0.51	0.49	-0.01
XGBoost	0.56	0.41	0.43	0.49	-0.02
Random Forests	0.69	0.30	0.34	0.50	-0.01
ProtT5					
Artificial Neural Networks	0.49	0.53	0.53	0.53	0.02
XGBoost	0.64	0.42	0.45	0.55	0.04
Random Forests	0.82	0.23	0.29	0.54	0.04
Relative Solvent Accessibility					
ESMFold					
Artificial Neural Networks	0.69	0.56	0.57	0.67	0.16
XGBoost	0.62	0.57	0.58	0.64	0.12
Random Forests	0.64	0.58	0.59	0.65	0.14
Secondary Structure					
ESMFold					
Artificial Neural Networks	0.67	0.39	0.43	0.55	0.04
XGBoost	0.56	0.49	0.50	0.54	0.03
Random Forests	0.52	0.54	0.53	0.54	0.03
ProtT5					
Artificial Neural Networks	0.47	0.55	0.55	0.54	0.02
XGBoost	0.59	0.43	0.45	0.52	0.01
Random Forests	0.57	0.43	0.45	0.50	0.00

Window Length – 19					
PLM Embeddings					
ESM-2					
Artificial Neural Networks	0.45	0.58	0.56	0.50	0.02
XGBoost	0.59	0.42	0.43	0.50	0.00
Random Forests	0.72	0.26	0.31	0.50	-0.01
ProtT5					
Artificial Neural Networks	0.51	0.59	0.58	0.56	0.06
XGBoost	0.63	0.42	0.44	0.54	0.03
Random Forests	0.84	0.19	0.26	0.54	0.03
Relative Solvent Accessibility					
ESMFold					
Artificial Neural Networks	0.69	0.55	0.56	0.67	0.15
XGBoost	0.61	0.58	0.58	0.64	0.12
Random Forests	0.61	0.59	0.59	0.65	0.13
Secondary Structure					
ESMFold					
Artificial Neural Networks	0.49	0.57	0.57	0.55	0.04
XGBoost	0.52	0.53	0.53	0.54	0.03
Random Forests	0.50	0.54	0.54	0.53	0.02
ProtT5					
Artificial Neural Networks	0.60	0.42	0.44	0.54	0.02
XGBoost	0.58	0.44	0.45	0.51	0.01
Random Forests	0.56	0.44	0.45	0.49	0.00
Window Length – 21					
PLM Embeddings					
ESM-2					
Artificial Neural Networks	0.46	0.60	0.58	0.53	0.04
XGBoost	0.65	0.42	0.44	0.53	0.04

Random Forests	0.71	0.31	0.35	0.52	0.02
ProtT5					
Artificial Neural Networks	0.57	0.54	0.54	0.57	0.06
XGBoost	0.45	0.61	0.59	0.54	0.04
Random Forests	0.84	0.19	0.26	0.54	0.03
Relative Solvent Accessibility					
ESMFold					
Artificial Neural Networks	0.68	0.56	0.58	0.67	0.16
XGBoost	0.64	0.56	0.57	0.65	0.13
Random Forests	0.61	0.60	0.60	0.65	0.13
Secondary Structure					
ESMFold					
Artificial Neural Networks	0.51	0.59	0.58	0.56	0.06
XGBoost	0.51	0.50	0.50	0.52	0.01
Random Forests	0.49	0.56	0.55	0.53	0.03
ProtT5					
Artificial Neural Networks	0.38	0.66	0.63	0.54	0.02
XGBoost	0.58	0.44	0.45	0.51	0.01
Random Forests	0.54	0.43	0.45	0.49	-0.02

2.3.3 Benchmarking

Similar to our approach to HAIRpred, the models created for the mouse-specific data must be compared with existing prediction tools in the literature. As we can tell, there is no existing method in literature which has been specifically created to predict conformation B-cell epitopes for mouse. Therefore, we cannot directly compare our models with other tools.

However, several methods have been proposed in the literature for predicting conformational B-cell epitopes without species specificity. In this study, we systematically assess the performance of these general B-cell epitope prediction tools on the independent, mouse-specific dataset to establish a benchmark for our proposed method. The predictive models employed for benchmarking include CBTOPE (Ansari and Raghava, 2010), Bepipred-2.0 (Jespersen et al., 2017), Bepipred-3.0 (Clifford et al., 2022) and SEMA-1d (Ivanisenko et al., 2024We show the performance of these tools in Table 7.

Table 8 : Comparison of existing methods on the mouse-specific independent test dataset

Model	Year	Sensitivity	Specificity	Accuracy	AUCROC	MCC
CBTOPE	2010	0.19	0.92	0.84	0.54	0.12
Bepipred - 2.0	2017	0.75	0.28	0.33	0.52	0.02
Bepipred - 3.0	2022	0.65	0.62	0.62	0.63	0.17
Sema - 1d	2024	0.71	0.63	0.63	0.67	0.21

There are still a few elements left to be performed in this study. Ensemble models need to be created to generate the model that gives the best predictive performance. These ensemble models must be then compared to these benchmarks. Following this, a feature importance analysis will be conducted using SHAP to improve the interpretability of the model. Finally, a web server will be implemented to facilitate accessibility and support the scientific community.

Chapter 3: Discussion

Antibodies are versatile glycoproteins which are secreted by the immune system that play a pivotal role in recognizing and removing pathogens, and generating immunological memory. This makes antibodies indispensable for human health and supporting immune defence mechanisms. Their extensive applications in diagnostics, vaccine development, immunotherapy, and therapeutic interventions underscores their critical importance in advancing modern medical science (Janeway, 2001). Therefore, elucidating antibody-antigen interactions remains a central objective in immunology.

Antibodies bind to specific regions of short sequences or conformational structures called the antibody-interacting regions or epitopes. Precise identification of these epitopes is crucial for advancing our understanding of immune interactions as well as the rational design of vaccines and the development of monoclonal antibodies exhibiting high affinity.

Traditional techniques for recognising epitopic regions, such as X-ray crystallography and alanine-scanning mutagenesis, are resource-intensive, time-consuming, and impractical for multiple antigens. The advancements in computational power and the emergence of sophisticated algorithms, particularly in machine learning, have facilitated the development of computational B-Cell epitope prediction tools including BepiPred-3.0 (Clifford *et al.*, 2022), SEMA-2.0 (Ivanisenko *et al.*, 2024), CLBTope (Kumar *et al.*, 2024), and CBTope (Ansari and Raghava, 2010). However, the current prediction models suffer from the challenges of lack of species specificity and overlooking the differences in immune responses and antibody structure across species. Most models are trained on datasets comprising antigenic information from diverse species, including mice, rabbits, and other commonly used experimental organisms. Consequently, these models fail to account for interspecies variations in antibody-antigen interactions. This leads to the lack of generalizability of these models which is shown by Cia *et al.*, 2023. They showed that all the current existing tools for B cell epitope prediction show poor performance.

In this study, we aimed to address these limitations and develop species-specific B cell epitope prediction tools. First, we showed that there is a difference between human and mouse epitopes by comparing their Amino Acid Composition (AAC). This result indicates that there is an interspecies variation in the antibody-antigen interactions between human and mice. It highlights the importance and the need for the development of species-specific predictors.

To address these challenges, this study focuses on the development of species-specific B-Cell epitope prediction tools. We design, train and test our models on species-specific antibody-antigen complexes. We start initially with developing a human-specific B cell epitope predictor. We created a high quality non-redundant dataset of human-specific antibody-antigen complexes. Then, we used this non-redundant dataset to generate features which encapsulated information about the antigen. The features were selected which have demonstrated predictive power in previous studies while maintaining biological relevance. We then analysed each feature's predictive performance and identified key features Relative Solvent Accessible surface area (RSA) and Position-Specific Scoring Matrix (PSSM) as the features which resulted in the highest predictive performance. In this analysis, we also compared the predictive performance across features of different window lengths to understand the influence of the neighbouring residues.

Through these experiments, the features with the best predictive performance were selected to create ensemble models. The analysis of the ensemble models helped us identify the best performing human-specific model which is an ensemble model of random forests which are trained on RSA and PSSM features derived from patterns of window length 15. This model is classified as Human Antibody Interacting Residue predictor (HAIRpred) and it is the first model in the literature, as far as we know, which is a species-specific B-Cell epitope prediction tool. Therefore, we consider it to be a second-generation of B-Cell epitope prediction tools. This model shows high prediction performance of AUROC of 0.72 on the independent test dataset and outperforms all the other models found in literature. Furthermore, we tried to improve the interpretability of the model by performing SHAP analysis on the individual models. The analysis revealed that the central position in the patterns, generated from the antigen sequence, plays a critical role in determining the final prediction.

There is, however, a limitation in our methodology in the development of HAIRpred. During the data processing stage to get the high quality non-redundant dataset, we clustered the antigens for redundancy with a 70% sequence similiarity threshold. This high threshold raises the possibility of bias or overfitting due to the fairly similar antigens used for training the models. The necessity of keeping a high threshold was to have to enough training data for the machine learning models. We hope that as more experimentally derived antibody-antigen complexes become available, we would be able to reduce this threshold to 40% and this would improve the performance and generalizability of our model.

In addition to the human-specific model, there is also a demand for mouse-specific models due to their widespread use in experimental studies. The existing tools overlook the characteristics of the murine immune system resulting in unnecessary delays in experimental research. To address this demand, we focus on the development of a mouse-specific prediction tool. We use a similar methodology as the human specific model involving analysing models based on features capturing sequential, structural and evolutionary information of the antigen. However, the challenge of limited data is even more stark here. The filtered dataset comprises of only 290 antibody-antigen complexes which after clustering at even 70% sequence identity threshold result in only 94 clusters. The development of machine learning models at such less amounts of data is challenging. We addressed the challenge by adopting a cluster-based data splitting strategy. In this approach, we include all antibody-antigen complexes from each cluster during model development, but we ensure that entire clusters are exclusively assigned to either the training or testing set. Additionally, during the five-fold cross validation process in the training of the models, it is ensured that all complexes in a cluster belong to either the training fold or the validation fold. This strategy ensures that highly similar antigens dont appear in both training and testing, which would lead to overly optimistic performance metrics and compromise the ability of the model to generalise.

In conclusion, these species-specific models offer a major advancement in the prediction of antibody-antigen interactions. Their robust performance makes them an important resource in immunological research, with applications in diagnostics, therapeutic development, and vaccine design. Our models overcome the drawbacks of previous models and we aim that they set a new standard for second-generation, species-specific prediction models. We hope that this study leads to the development of species-specific models in other areas of immunological research.

Chapter 4: Materials and Methods

4.1 Creation of Datasets

The datasets used in this study are experimentally-derived human (Homo Sapiens) and mouse (Mus Musculus) antibody-antigen complexes downloaded from the SAbDab database (Dunbar *et al.*, 2014) in June 2024. SAbDab (Structural antibody database) is an online database which contains and curates all publicly available annotated antibody structures. The database is updated weekly and it has different filters including species type, resolution, antibody-antigen affinity etc. These filters were used to create high quality datasets which would be used for training and testing machine learning models. The antibody-antigen complexes were filtered for only antigens which are proteins with more than 50 residues, and the resolution of the complex is less than 3.0 Å with the R-factor being less than 0.30. This resulted in high quality datasets comprising of 1620 human antibody-antigen complexes and 290 mouse antibody-antigen complexes.

Following this, the redundant antigens were removed to eliminate potential bias. Both human and mouse were clustered according to their sequence identity using CD-HIT (Fu et al., 2012) with a 70% sequence identity threshold. CD-HIT (Cluster Database at High Identity with Tolerance) is a clustering program which takes fasta format as input and generates clusters of sequences above a similarity threshold and also generates a set of non-redundant representative sequences as output. The program is aimed at removing sequence redundancy.

The human antigens were clustered using CD-HIT with 70% sequence identity threshold. This resulted in the generation of 277 antigen clusters with its non-redundant representative sequences. These representative sequences were then used for further analysis and therefore a high quality non-redundant dataset of 277 antibody-antigen complexes were generated.

The mouse antibody-antigen complexes, however, were limited and resulted in only 94 non redundant clusters after clustering antigens according to their sequence identity using CD-HIT (Fu et al., 2012) with a 70% sequence identity threshold. The non-redundant representative antigens alone would not be enough to train and test the machine learning models. Hence, instead of only using the representative antigen of a cluster, we use all antibody-antigen complexes in model development. This, however, leads to a concern of bias and overfitting of the model. To avoid potential bias, we employ cluster-based data splitting strategy while developing the model which has been described later in this chapter.

4.2 Labelling of Residues

The antibody-antigen complexes derived from SAbDab were processed to identify the antigen residues that interacted with the antibody. The labelling (1 for antibody interacting residues and 0 for antibody not interacting residues) algorithm used in this study is similar to Cia *et al.*, 2023. The algorithm examines the change in relative solvent accessible surface area (RSA) of these residues with and without antibody binding. A residue is classified as an antibody-interacting residue if it experiences a change in Relative Solvent Accessibility (RSA) of at least 5% upon binding with an antibody. All residues except antibody interacting residues are labelled as antibody not interacting residues.

$$RSAunbound - RSAbound \geq 5\%$$

Equation 1: Condition for antibody interacting residues

4.3 Generation of Patterns

Previous investigations have demonstrated that the residue's interactions are affected by its surrounding environment (Kaur and Raghava, 2003, 2004; Ansari and Raghava, 2010). We wanted to understand how the local environment affects the residue's interaction with the antibody. Additionally, we wanted to incorporate this effect into our prediction models. To account for this local influence, overlapping sequence patterns (or sliding windows) were generated for every residue within the antigen. The length of the generated patterns ranged from 13 to 21 residues. For each residue, a sliding window of a particular size was created with the residue being in the centre and this is termed as the pattern for that particular residue. This process is demonstrated in Figure 17. For the terminal residues, we used a dummy variable 'X' for the residues not covered in the pattern. The label (antibody interaction or antibody not-interacting) for each pattern is the corresponding label of the central residue. This methodology of capturing the local environment has been commonly used in previous studies (Bhasin and Raghava, 2005; Kumar *et al.*, 2008).

For this study, we created the patterns according to the forementioned methodology for both the human and mouse data.

Figure 17: Generation of Patterns of different window lengths

4.4 AAC Analysis

Amino Acid Composition (AAC) quantifies the relative abundance of each amino acid within a given sequence by computing the frequency of occurrence of each residue type. It is determined by normalizing the count of each amino acid to the total number of residues in the sequence. It is calculated by Equation 2, where AACi represents the amino acid composition

of residue type i; Pi denotes the number of residues of type i and Q corresponds to the total length of the sequence.

$$AAC_i\ (\%) = \frac{P_i}{Q} * 100$$

Equation 2: Calculation for AAC.

In Section 2.1, the AAC of human and mouse epitopes are compared to understand the differences between them. The labelled data of both human and mouse were filtered to only get the antibody-interacting residues (epitopes) and the AAC was calculated using Equation 2. The general proteome AAC were the AAC in the UniProtKB/Swiss-Prot data bank which are derived from the Release notes for UniProtKB/Swiss-Prot release 2013_04 - April 2013 (ed. JM Walker, 2005). The AAC were then compared by plotting it on a graph using matplotlib in python.

In Section 2.2.1, the AAC of human antibody interacting residues and not-interacting residues were compared. The labelled human-specific antigen data were classified as per the labels and AAC was calculated as per Equation 2. The AAC were compared by plotting it on a graph using matplotlib in python. P values were calculated using chi-square test to understand the significance of the difference in AAC of each residue and it was calculated using the scipy package in python. The label for significance is (*) and the number of (*) is calculated using thresholds with (*) for p-value below 0.05, (**) for p-value below 0.01 and (***) for p-value below 0.001. A similar procedure was performed for mouse-specific antibody-antigen complexes in Section 2.3.1.

4.5 Two Sample Logo

To investigate enrichment of residues within antibody-interacting and non-interacting patterns, we analyzed residue enrichment at specific positions using the "Two Sample Logo" web-based tool. This application allows comparative visualization of amino acid distributions by evaluating statistically significant differences between positive and negative sequence samples. The tool calculates the abundance of all amino acids at each position in the pattern and perform statistical tests to quantify the enrichment or depletion of the amino acids. The resulting graphical output represents these differences, with symbol sizes proportional to the magnitude of statistical significance. The statistical significance is determined using the t-test.

We generated the “Two Sample Logo” for both human and mouse datasets. Antibody interacting windows were designated as the positive samples. For the negative samples, we considered only those not-interacting patterns where no residue was epitopic. This was done to ensure elimination of potential bias from the antibody interacting residues. To mitigate potential bias , we ensured an equal number of positive and negative samples by randomly selecting non-interacting patterns to match the total count of positive samples.

4.6 Splitting of data into train and test

To train the machine learning models and test them, we need to divide them into training and independent test dataset. In this study, we divide the data into train and test dataset along the ration 80:20 respectively. The human-specific dataset comprised of high-quality non-redundant

277 antigens which were then divided into the training dataset comprising of 221 antigens and the independent test dataset comprising of 56 antigens.

However, the mouse dataset is limited with only 290 high quality antigens , which, after redundancy reduction through clustering, results in only 94 non-redundant representative antigens. Training a robust and generalizable model with such a small dataset poses a significant challenge as machine learning models usually require a substantial amount of diverse training data to avoid overfitting and to ensure reliable performance on unseen data.

To address this limitation, we adopt a cluster-based data splitting strategy. Instead of only using the representative antigen of a cluster, all the mouse-specific antibody-antigen complexes are used in model development. This approach allows us to increases the effective dataset size but it could lead to potential bias. To ensure the model does not become biased or overfit due to the high sequence similarity among the antigens, entire clusters are exclusively assigned to either the training or testing set. This restricts highly similar antigens from appearing in both sets, which would lead to overly optimistic performance metrics and compromise the ability of the model to generalise. By enforcing this cluster-wise split, we ensure that the model learns meaningful sequence patterns rather than memorizing highly similar sequences. The resulting training and independent testing dataset comprises of 232 and 58 mouse-specific experimentally derived antibody-antigen complexes.

Additionally, the training dataset has the not-interacting patterns in much larger numbers than the interacting patterns. This leads to an imbalanced training dataset. This imbalance can introduce bias in the model, skewing it towards the more abundant class (non-interacting patterns) while underrepresenting the minority class (interacting patterns). This might result in the model achieving high accuracy overall but performing poorly in correctly identifying interacting residues. To avoid this, we construct a balanced training dataset by including all antibody-interacting patterns and randomly selecting an equal number of non-interacting patterns. This ensures that the model is exposed to an equal representation of both classes which allows it to learn meaningful patterns for predicting interacting residues. This balancing strategy is applied to both the human and mouse datasets.

4.7 Generation of Features

Machine learning models require input in the form of numerical vectors. We developed these inputs as features which would capture information about the antigen. In our study, we focused on features which have been shown good predictive performance in the literature and also have theoretical relevance. A comprehensive set of features was curated to encapsulate diverse information about the antigens. The selected features are listed below.

4.7.1 Binary Profiles

Binary Profiles are commonly used technique to capture sequential information (Ansari and Raghava, 2010). Binary Profiles were generated by encoding each amino acid in the pattern as a binary vector of length 20. This encoding assigns a one-hot representation to each of the 20 standard amino acids i.e each amino acid is represented by a vector in which only one position corresponding to that amino acid is set to 1, while all other positions are set to 0. The amino acids are ordered as follows:

['A','R','N','D','C','Q','E','G','H','I','L','K','M','F','P','S','T','W','Y','V'].

For example, the amino acid Alanine (A) is encoded as a 20-length vector: [1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]. Similarly, other amino acids have a 1 at their respective positions in the 20-length vector. Additionally, to account for the patterns of the terminal residues, the dummy variable X is encoded as 20 length zero vector.

Binary Profiles were generated for each sequence pattern by concatenating the one-hot encoded vectors for all residues within that pattern. If a pattern consists of N residues, the feature vector has a dimension of N×20, where N represents the length of the pattern and 20 is the size of the binary vector for each residue.

4.7.2 Position-Specific Scoring Matrices (PSSM)

Position-Specific Scoring Matrices (PSSM) quantify the probability of each amino acid appearing at a given position based on homologous seqyences, which is extracted from multiple sequence alignments. PSSM profiles are widely used in bioinformatics for sequence analysis, functional annotation and evolutionary information (Kumar *et al.*, 2008; Kha *et al.*, 2025).

In this study, we generated PSSM profiles for each antigen sequence using the Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) tool (Altschul *et al.*, 1997) and searching against the Swiss Prot database (Bairoch and Apweiler, 1999). The PSI-BLAST search was conducted with an E-value threshold of 0.1 to filter out statistically insignificant alignments. A word size of 3 was used for protein sequences, and the search was performed for three iterations to refine the alignment and detect even the remote homologs. The sequence alignment was carried out using the BLOSUM62 scoring matrix, with a gap opening penalty of 11 and a gap extension penalty of 1.

Once the PSSM profiles were generated, we extracted the evolutionary information for each sequence and divided it into patterns. The dummy variable X was assigned a 20 length zero vector. For each antigen sequence, the final feature vector is structured as a matrix of dimensions N×20, where N represents the length of the pattern (number of residues), and 20 corresponds to the length of the PSSM vector for each residue.

4.7.3 PLM Embeddings

Protein Language Models (PLM) are natural language processing models which are designed to generate high-dimensional embeddings that encode the relations between amino acids in a protein sequence. These embeddings encode rich biological information including evolutionary and functional properties of protein sequences which makes them a valuable feature for biological models. Additionally, unlike traditional sequence-based features, PLM embeddings capture not only sequence information but also structural properties of proteins. These models learn implicit representations that correlate with secondary and tertiary structures of proteins. This structural awareness makes them valuable to study protein-protein interactions.

In this study, we used two state-of-the-art PLMs to generate embeddings for sequence patterns: ESM2 (Lin *et al.*, 2023) (esm2_t6_8M_UR50D) and Encoder-only ProtTrans (Elnaggar *et al.*, 2022) (ProtT5-XL-UniRef50). ESM2 is a model from the Evolutionary Scale Modeling (ESM) family and it captures residue-level relationships by pretraining on large-scale protein

databases. ProtT5-XL-UniRef50, part of the ProtTrans family, encodes protein sequences with high sensitivity to biophysical and functional features.

To generate embeddings, these models were downloaded from HuggingFace. The sequence patterns were inputted into these models and the last-layer embeddings were extracted as the feature vector. These embeddings provide a numerical representation of each sequence pattern and contains evolutionary as well as structural information.

4.7.4 Predicted Relative Solvent Accessibility

Relative Solvent Accessibility is a measure of the extent to which an amino acid residue is exposed to the surrounding environment in a protein's three-dimensional structure. This measure is important for understanding protein-protein interactions and functional sites, as exposed residues are more likely to participate in binding and interactions.

To predict RSA values directly from sequence, we first generated the 3D structure of each protein using ESMFold, a PLM-based protein structure prediction model (Lin et al., 2023). The predicted structures were then processed using the DSSP algorithm (Kabsch and Sander, 1983) which computes RSA value for every residue of the antigen. The RSA values generated for the antigen residues were then distributed into patterns. Additionally, to account for the patterns of the terminal residues, the dummy variable X was assigned the RSA value of 0. The final feature vector is an N length vector, where N represents the length of the sequence pattern.

4.7.5 Predicted Secondary Structure

The secondary structure of the protein is the local spatial conformation of the backbone of the protein. It provides an insight into the local folding patterns of the protein backbone and it plays an important role in determining the stability and function of the protein. It has been used in this study as a feature vector.

We predict the secondary structure using two different approaches. First, ProtT5-XL-U50 embeddings were utilized to predict 8-class SS, transferring the structural knowledge learned from large-scale protein datasets. Second, the DSSP algorithm was applied to 3D structures predicted by ESMFold, generating a 3-class SS representation (helix, strand, and coil). This allows us to integrate secondary structure knowledge through both sequence-based and structure-based predictions. The SS values generated were then divided into patterns. Additionally, to account for the patterns of the terminal residues, the dummy variable X is given the secondary structure of a coil. The final SS labels were then encoded using a label encoder which transformed the categorical SS into a numerical format. This resulted in a vector length corresponding to the length of the sequence pattern.

4.8 Machine Learning Models

The machine learning models used in this study were implemented using scikit-learn (Pedregosa et al., 2018), a widely used library in python to develop predictive models. We focused on two popular machine learning models : the Random Forest (RF) algorithm and the Extreme Gradient Boosting (XGBoost) algorithm, both known for their ability in classification and their ability to parse through complex datasets. Random Forest is an algorithm which creates and aggregates multiple decision trees to reduce overfitting and improve accuracy.

XGBoost, on the other hand, is an optimized gradient boosting framework designed for speed, performance, and scalability.

To optimize these model performance, hyperparameter tuning was performed using GridSearchCV from the scikit-learn package. GridSearchCV systematically explores different hyperparameter combinations and selects the combination that gives the best classification results based on cross-validation. This ensured that both Random Forest and XGBoost models were fine-tuned to maximize predictive accuracy.

Beyond conventional machine learning algorithms, we employed neural network models implemented in PyTorch (Paszke et al., 2019). The models were trained using backpropagation, with the optimization process facilitated by the Adam optimizer. Cross-entropy loss was used as the loss function to measure prediction error and guide model optimization.

4.9 Training and Evaluation of the Models

To ensure robust and unbiased models, we trained our human-specific models using the five-fold cross-validation technique. The train dataset was randomly divided into five equal-sized "folds", where four folds were used for training while the remaining fold was the validation set. This process was repeated five times with each fold serving as the validation set once. The results from each fold were then averaged to provide an estimate of the model's performance. This approach allows us to optimize the model's hyperparameters to improve the predictive performance.

In the mouse-specific models, a cluster-based splitting strategy was adopted during five-fold cross-validation to account for the sequence redundancy. Here, we ensure that no antigen cluster is shared between training and validation folds, thereby preventing the mixing of highly similar antigens across the folds. This ensures a more realistic and reliable evaluation of the model's generalizability and its performance on unseen and diverse antigen sequences.

The possibility of overoptimization cannot be entirely excluded when employing five-fold cross-validation as the hyperparameter tuning is performed on the same dataset. To address this, we evaluated our models on the independent test dataset which was not used during model training and hyperparameter optimization. This final assessment confirms the model's ability to generalize to unseen data.

4.10 Evaluation Metrics

In this study, the machine learning models were designed to perform binary classification on each sequence pattern and assigning labels to residues as antibody-interacting or non-interacting. To assess the predictive performance of these models, we used a set of evaluation metrics which can be broadly categorized into threshold-independent and threshold-dependent metrics.

For the threshold-independent evaluation, we used the Area Under the Receiver Operating Characteristic Curve (AUROC) score. The Receiver Operating Characteristic (ROC) curve is a graphical representation that plots the True Positive Rate against the False Positive Rate and the AUROC score is the area under the ROC curve. AUROC score measures the model's ability to discriminate between interacting and non-interacting residues irrespective of the thresholds, providing a robust assessment of overall model performance. Therefore, in this study, our key evaluation metric is AUROC score.

Other metrics we use in this study are threshold-dependent metrics. These metrics are Sensitivity, Specificity, Accuracy and Matthew's correlation coefficient (MCC). The formulas are listed in Equation 3.

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

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

Equation 3: Threshold Dependent Evaluation Metrics

Here TP, TN, FP, and FN represents true positive, true negative, false positive, and false negative respectively. These metrics are implemented using the scikit-learn package.

4.11 Ensemble Models

To further improve the predictive performance, ensemble models were created by combining results from different models which are trained on features. We employed a simple ensemble averaging strategy, where the predicted probabilities from individual models were averaged. The final classification decision was made based on the averaged probability. This enables the model to leverage the strengths of different classifiers. This ensemble approach helped us improve robustness and generalization of the models, leading to more reliable predictions.

4.12 Interpretability using SHAP

Gaining insight into the reasoning process of machine learning models is essential for ensuring their reliability. To increase the interpretability of our model, we used the SHapley Additive exPlanations (SHAP) method to analyse our ensemble mode. SHAP is a game theory based approach to understand the model and it generates Shapley values for each position in a feature vector. These Shapley values indicate the importance of that position in the feature vector in the final prediction (Lundberg and Lee, 2017).

To achieve interpretability in our models, we utilized SHAP TreeExplainer from the Python shap package to analyze the individual models within HAIRpred's Random Forest ensemble. The Shapley values provide a measure of each feature's importance and helps us understand how and to what extent different features contribute to the final prediction. SHAP allows us to to evaluate which residues most influence the model's classification decisions.

4.13 Web-Server

HAIRpred was integrated into a webserver to improve accessibility and provide ease of us to the scientific research community (https://webs.iiitd.edu.in/raghava/hairpred/). This platform is designed to facilitate easy and efficient prediction of epitopic residues in antigen sequences, making it accessible to researchers without requiring extensive bioinformatics expertise. The webserver was designed using HTML, PHP and Javascript. In addition to the web-based platform, we developed a standalone version of HAIRpred to improve accessibility for

researchers working in offline environments or requiring customized analyses. The standalone version is implemented entirely in Python and is available as an open-source repository on Github (https://github.com/raghavagps/HAIRpred). Furthermore, to simplify the integration of HAIRpred into machine learning pipelines, we have released it as a Python package, which can be installed via pip: pip install hairpred.

7. References

1. Alberts, B (2014). Molecular biology of the cell, Boca Raton, FL: CRC Press.
2. Almagro, JC, Hernández, I, Ramírez, MC, and Vargas-Madrazo, E (1998). Structural differences between the repertoires of mouse and human germline genes and their evolutionary implications. Immunogenetics 47, 355–363.
3. Altschul, SF, Madden, TL, Schäffer, AA, Zhang, J, Zhang, Z, Miller, W, and Lipman, DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389–3402.
4. Ansari, HR, and Raghava, GP (2010). Identification of conformational B-cell Epitopes in an antigen from its primary sequence. Immunome Res 6, 6.
5. Bahai, A, Asgari, E, Mofrad, MRK, Kloetgen, A, and McHardy, AC (2021). EpitopeVec: linear epitope prediction using deep protein sequence embeddings. Bioinformatics 37, 4517–4525.
6. Bairoch, A, and Apweiler, R (1999). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Res 27, 49–54.
7. Beckman Coulter Blood Cells.
8. Benjamin, DC, and Perdue, SS (1996). Site-Directed Mutagenesis in Epitope Mapping. Methods 9, 508–515.
9. Bhasin, M, and Raghava, GPS (2005). Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences. Nucleic Acids Res 33, W202-207.
10. Boguski, MS (2002). The mouse that roared. Nature 420, 515–516.
11. Chiu, ML, Goulet, DR, Teplyakov, A, and Gilliland, GL (2019). Antibody Structure and Function: The Basis for Engineering Therapeutics. Antibodies (Basel) 8.
12. Chou, PY, and Fasman, GD (1974). Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. Biochemistry 13, 211–222.
13. Cia, G, Pucci, F, and Rooman, M (2023). Critical review of conformational B-cell epitope prediction methods. Brief Bioinform 24, bbac567.
14. Clifford, JN, Høie, MH, Deleuran, S, Peters, B, Nielsen, M, and Marcatili, P (2022). BepiPred-3.0: Improved B-cell epitope prediction using protein language models. Protein Sci 31, e4497.
15. Collatz, M, Mock, F, Barth, E, Hölzer, M, Sachse, K, and Marz, M (2021a). EpiDope: a deep neural network for linear B-cell epitope prediction. Bioinformatics 37, 448–455.
16. Collatz, M, Mock, F, Barth, E, Hölzer, M, Sachse, K, and Marz, M (2021b). EpiDope: a deep neural network for linear B-cell epitope prediction. Bioinformatics 37, 448–455.
17. Dall'Antonia, F, and Keller, W (2019). SPADE web service for prediction of allergen IgE epitopes. Nucleic Acids Research 47, W496–W501.
18. Dunbar, J, Krawczyk, K, Leem, J, Baker, T, Fuchs, A, Georges, G, Shi, J, and Deane, CM (2014). SAbDab: the structural antibody database. Nucleic Acids Res 42, D1140-1146.
19. Elnaggar, A, Heinzinger, M, Dallago, C, Rehawi, G, Wang, Y, Jones, L, Gibbs, T, Feher, T, Angerer, C, Steinegger, M, *et al.* (2022). ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Trans Pattern Anal Mach Intell 44, 7112–7127.

20. Fu, L, Niu, B, Zhu, Z, Wu, S, and Li, W (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152.
21. Gabriel CIA BERIAIN (2023). Development of B-cell epitope prediction pipelines. Application to chronic lymphocytic leukemia and implications for disease origin. Ecole Polytechnique de Bruxelles.
22. Greenberg, SJ (2003). A Concise History of Immunology.
23. Haste Andersen, P, Nielsen, M, and Lund, O (2006). Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Science 15, 2558–2567.
24. Hopp, TP, and Woods, KR (1981). Prediction of protein antigenic determinants from amino acid sequences. Proceedings of the National Academy of Sciences 78, 3824–3828.
25. Ivanisenko, NV, Shashkova, TI, Shevtsov, A, Sindeeva, M, Umerenkov, D, and Kardymon, O (2024). SEMA 2.0: web-platform for B-cell conformational epitopes prediction using artificial intelligence. Nucleic Acids Res 52, W533–W539.
26. Janeway, CA (2001). Immunobiology: the immune system in health and disease ; [animated CD-ROM inside], New York, NY: Garland Publ. [u.a.].
27. Jenner, E (2023). An inquiry into the causes and effects of the variolae vaccinae: a disease discovered in some of the western counties of England, particularly Gloucestershire, and known by the name of the cow pox. In: Scientific and Medical Knowledge Production, 1796-1918, Routledge, 40–50.
28. Kabsch, W, and Sander, C (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637.
29. Kaur, H, and Raghava, GPS (2003). A neural-network based method for prediction of γ-turns in proteins from multiple sequence alignment. Protein Science 12, 923–929.
30. Kaur, H, and Raghava, GPS (2004). Prediction of α-turns in proteins using PSI-BLAST profiles and secondary structure information. Proteins: Structure, Function, and Bioinformatics 55, 83–90.
31. Kha, QH, Nguyen, HPL, and Le, NQK (2025). A deep learning and PSSM profile approach for accurate SNARE protein prediction. Methods Mol Biol 2887, 79–89.
32. Kulkarni-Kale, U, Bhosle, S, and Kolaskar, AS (2005). CEP: a conformational epitope prediction server. Nucleic Acids Research 33, W168–W171.
33. Kumar, M, Gromiha, MM, and Raghava, GPS (2008). Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins 71, 189–194.
34. Kumar, N, Tripathi, S, Sharma, N, Patiyal, S, Devi, NL, and Raghava, GPS (2024). A method for predicting linear and conformational B-cell epitopes in an antigen from its primary sequence. Comput Biol Med 170, 108083.
35. Kyte, J, and Doolittle, RF (1982). A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology 157, 105–132.
36. Larsen, JEP, Lund, O, and Nielsen, M (2006). Improved method for predicting linear B-cell epitopes. Immunome Res 2, 2.
37. Laver, WG, Air, GM, Webster, RG, and Smith-Gill, SJ (1990). Epitopes on protein antigens: Misconceptions and realities. Cell 61, 553–556.
38. LeBien, TW, and Tedder, TF (2008). B lymphocytes: how they develop and function. Blood 112, 1570–1580.

39. Liew, OW, Ling, SSM, Lilyanna, S, Zhou, Y, Wang, P, Chong, JPC, Ng, YX, Lim, AES, Leong, ERY, Lin, Q, *et al.* (2021). Epitope-directed monoclonal antibody production using a mixed antigen cocktail facilitates antibody characterization and validation. Communications Biology 4, 441.
40. Lin, Z, Akin, H, Rao, R, Hie, B, Zhu, Z, Lu, W, Smetanin, N, Verkuil, R, Kabeli, O, Shmueli, Y, *et al.* (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130.
41. Liu, F, Yuan, C, Chen, H, and Yang, F (2024). Prediction of linear B-cell epitopes based on protein sequence features and BERT embeddings. Sci Rep 14, 2464.
42. Liu, T, Shi, K, and Li, W (2020). Deep learning methods improve linear B-cell epitope prediction. BioData Mining 13, 1.
43. Lundberg, S, and Lee, S-I (2017). A Unified Approach to Interpreting Model Predictions.
44. Mestas, J, and Hughes, CCW (2004). Of mice and not men: differences between mouse and human immunology. J Immunol 172, 2731–2738.
45. Parker, JMR, Guo, D, and Hodges, RS (1986). New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and x-ray-derived accessible sites. Biochemistry 25, 5425–5432.
46. Paszke, A, Gross, S, Massa, F, Lerer, A, Bradbury, J, Chanan, G, Killeen, T, Lin, Z, Gimelshein, N, Antiga, L, *et al.* (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library.
47. Pedregosa, F, Varoquaux, G, Gramfort, A, Michel, V, Thirion, B, Grisel, O, Blondel, M, Müller, A, Nothman, J, Louppe, G, *et al.* (2018). Scikit-learn: Machine Learning in Python.
48. Ponomarenko, J, and Van Regenmortel, M (2009). B cell epitope prediction. Structural Bioinformatics.
49. Ponomarenko, JV, and Bourne, PE (2007). Antibody-protein interactions: benchmark datasets and prediction tools evaluation. BMC Structural Biology 7, 64.
50. Regenmortel, MHVV (1996a). Mapping Epitope Structure and Activity: From One-Dimensional Prediction to Four-Dimensional Description of Antigenic Specificity. Methods 9, 465–472.
51. Regenmortel, MHVV (1996b). Mapping Epitope Structure and Activity: From One-Dimensional Prediction to Four-Dimensional Description of Antigenic Specificity. Methods 9, 465–472.
52. Rubinstein, ND, Mayrose, I, Martz, E, and Pupko, T (2009). Epitopia: a web-server for predicting B-cell epitopes. BMC Bioinformatics 10, 287.
53. Saha, S, and Raghava, GPS (2006). Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins: Structure, Function, and Bioinformatics 65, 40–48.
54. Sahni, R, Kumar, N, and Raghava, GPS (2024). HAIRpred: Prediction of human antibody interacting residues in an antigen from its primary structure.
55. da Silva, BM, Myung, Y, Ascher, DB, and Pires, DEV (2021). epitope3D: a machine learning method for conformational B-cell epitope prediction. Briefings in Bioinformatics 23, bbab423.
56. Solihah, B, Azhari, A, and Musdholifah, A (2020). Enhancement of conformational B-cell epitope prediction using CluSMOTE. PeerJ Comput Sci 6, e275.

57. Sun, J, Wu, D, Xu, T, Wang, X, Xu, X, Tao, L, Li, YX, and Cao, ZW (2009). SEPPA: a computational server for spatial epitope prediction of protein antigens. Nucleic Acids Research 37, W612–W616.
58. Sweredoski, MJ, and Baldi, P (2008a). COBEpro: a novel system for predicting continuous B-cell epitopes. Protein Engineering, Design and Selection 22, 113–120.
59. Sweredoski, MJ, and Baldi, P (2008b). PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. Bioinformatics 24, 1459–1460.
60. Talucci, I, and Maric, HM (2024). Epitope landscape in autoimmune neurological disease and beyond. Trends in Pharmacological Sciences 45, 768–780.
61. Tankeshwar, A Epitopes: Types, Function, Epitope Spreading | Microbe Online — microbeonline.com.
62. Walker, JM (2005). The proteomics protocols handbook, Totowa, NJ: Humana Press.
63. Wang, Y, Wu, W, Negre, NN, White, KP, Li, C, and Shah, PK (2011). Determinants of antigenicity and specificity in immune response for protein sequences. BMC Bioinformatics 12, 251.