Received: 17 January 2025 | Revised: 13 June 2025 | Accepted: 16 June 2025

DOI: 10.1002/pro.70212

RESEARCH ARTICLE

THE PROTEIN SOCIETY WILEY

HAIRpred: Prediction of human antibody interacting residues in an antigen from its primary structure

Ruchir Sahni[1,2] | **Nishant Kumar**[1] | **Gajendra P. S. Raghava**[1]

[1]Department of Computational Biology, Indraprastha Institute of Information Technology Delhi, New Delhi, India

[2]Indian Institute of Science Education and Research (IISER) Pune, Pune, India

Correspondence
Gajendra P. S. Raghava, Department of Computational Biology, Indraprastha Institute of Information Technology Delhi, Okhla Industrial Estate, Phase III (Near Govind Puri Metro Station), New Delhi 110020, India.
Email: raghava@iiitd.ac.in

Funding information
Department of Biotechnology, Ministry of Science and Technology, India, Grant/Award Number: BT/PR40158/BTIS/137/24/2021

Review Editor: Nir Ben-Tal

Abstract

In the past, several methods have been developed for predicting conformational B-cell epitopes in antigens that are not specific to any host. Our primary analysis of antibody–antigen complexes indicated a need to develop host-specific B-cell epitopes. In this study, we present a novel approach to predict conformational B-cell epitopes specific to human hosts by focusing on human antibody interacting residues in antigens. We trained, tested, and evaluated our models on 277 complexes of human antibody–antigen complexes. Initially, we employed machine learning models based on the one hot encoding sequence profile of antigens, achieving a maximum area under the receiver operating characteristic curve (AUROC) of 0.61. The performance of the model improved significantly with the AUROC increasing from 0.61 to 0.67 when evolutionary profiles were used instead of one hot encoding profile. Models developed using embeddings from fine-tuned protein language models reached an AUROC of 0.61. Additionally, models utilizing predicted surface relative solvent accessibility achieved an AUROC of 0.67. Our ensemble model, which combined relative surface accessibility with evolutionary profiles, achieved the highest precision with an AUROC of 0.72. All models in this study were trained using fivefold cross-validation on a training dataset and evaluated on an independent dataset not used for training or validation. Our method outperforms existing approaches on the independent dataset. Furthermore, we used the SHAP eXplainable AI (XAI) method to interpret the importance of elements in features contributing to the predictions made by our models. To support the scientific community, we have developed a standalone software and web server, HAIRpred, for predicting human antibody interacting residues in proteins (https://webs.iiitd.edu.in/raghava/hairpred/).

KEYWORDS
antibody–antigen interaction, antibody interacting residues, B-cell epitopes, machine learning, protein language models

1 | INTRODUCTION

Antibodies are specialized glycoproteins produced by the immune system, playing a vital role in identifying and neutralizing pathogens, as well as in establishing immunological memory. Their essential functions make antibodies crucial for maintaining human health and supporting immune responses. Given their wide-ranging applications in diagnostics, vaccines, immunotherapy, and therapeutics, the study of antibody activity and interactions is a fundamental aspect of immunological research. The interaction between an antibody and its corresponding antigen predominantly occurs through specific surface residues on the antigen, termed antibody interacting residues (Zeng et al. 2023). These residues are critical in defining conformational B-cell epitopes, as they represent the direct contact points between antibodies and their targets. Therefore, it is

essential to accurately identify these residues for predicting conformational B-cell epitopes, which play a significant role in immune recognition. Conventional experimental techniques for identifying antibody interacting regions, such as X-ray crystallography and alanine scanning mutagenesis, are labor-intensive, costly, and often unsuitable for large-scale analysis (Kozlova et al. 2018).

Recent progress in machine learning and artificial intelligence has led to the development of various computational models aimed at predicting antibody interacting residues or conformational B-cell epitopes. These computational models are based on experimental datasets like IEDB (Vita et al. 2025), PDB (Burley et al. 2019), and simulated datasets like Absolut! (Robert et al. 2022). Broadly, existing methods for predicting conformational B-cell epitopes can be categorized into two groups, namely structure-based and sequence-based methods. Following are examples of structure-based methods commonly used for predicting conformational B-cell epitopes in an antigen from its tertiary structure CEP, DiscoTope, PEPITO, Epitopia, SEPPA, SEMA, and Epitope3D (da Silva et al. 2022; Haste Andersen et al. 2006; Kulkarni-Kale et al. 2005; Rubinstein et al. 2009; Shashkova et al. 2022; Solihah et al. 2020; Sun et al. 2009; Sweredoski and Baldi 2008). One of the limitations of structure-based methods is that they need the tertiary structure of the antigen. In order to address this challenge, a number of sequence-based methods have been developed for predicting conformational B-cell epitopes in an antigen from its amino acid sequence. The following are commonly used sequence-based methods BepiPred-3.0, SEMA-2.0, CBTope, and CLBTope (Ansari and Raghava 2010; Clifford et al. 2022; Ivanisenko et al. 2024; Kumar et al. 2024). However, a notable limitation of these models is their inability to accommodate host-specific differences in antibody–antigen interactions. Antibodies are finely adapted to their respective host organisms that the epitopes recognized by human antibodies can differ significantly from those recognized by antibodies from other species, such as mice. Additionally, these models frequently overlook variations in immune responses, antibody structures, and post-translational modifications across different species (Almagro et al. 1998; Mestas and Hughes 2004). Such limitations raise concerns about the generalizability of these models across datasets obtained from diverse experimental settings. Recent research by Cia et al. (2023) has highlighted the shortcomings of existing methods by demonstrating their limited performance when tested against experimentally derived complexes from the Protein Data Bank (PDB) (Cia et al. 2023). We believe that the lack of host specificity resulted in the poor performance of the existing generalized models. As a result, focusing on human antibody interacting residues to predict epitopes that are specific to human antibodies is a crucial step toward enhancing the accuracy and relevance of immune response predictions.

In this study, we present HAIRpred, a novel computational tool designed to predict human antibody interacting regions in antigens with improved accuracy and interpretability. Our models were developed and validated using experimentally derived human antibody–antigen complexes, ensuring a focus on host specificity. We utilized a carefully curated set of sequence-derived features that have shown predictive efficacy in prior studies while ensuring biological relevance. This feature set integrates structural, sequential, and evolutionary information about the antigen that can be extracted from its amino acid sequence. The performance of HAIRpred was rigorously evaluated against existing methods using a diverse dataset of experimentally verified antibody–antigen complexes. Additionally, the SHAP method was used to interpret the models, which provided insight into the relative contribution of each element in the feature to the prediction outcomes.

2 | RESULTS

We have divided the result section into six categories: (i) importance of host specificity, (ii) data analysis, (iii) machine learning methods, (iv) benchmarking, (v) feature importance analysis, and (vi) web server implementation. The complete workflow of the study is illustrated in Figure 1, and the details of the following subsections can be found below.

2.1 | Importance of host specificity

The hypothesis of the study is that the major limitation of existing generalized models is the failure to account for the interspecies variation in antibody–antigen interactions. It is therefore important to show the existence of the difference in antibody–antigen interactions, which was done by comparing the antibody interacting residues of different species. We compared the amino acid composition (AAC) for both human and Mus musculus interacting residues (Figure 2). The general proteome was also introduced in the comparison to highlight the characteristics of human and mouse epitopes. It was observed that the AAC is significantly different between the two species. For example, amino acids like cysteine (C), glutamine (Q), arginine (R), tryptophan (W), and tyrosine (Y) are preferred in antibody–antigen interaction in the case of humans but not in the case of Mus musculus. Similarly, residues like glycine (G), lysine (K), asparagine (N), and serine (S) are preferred in Mus musculus antibody interaction but not in human

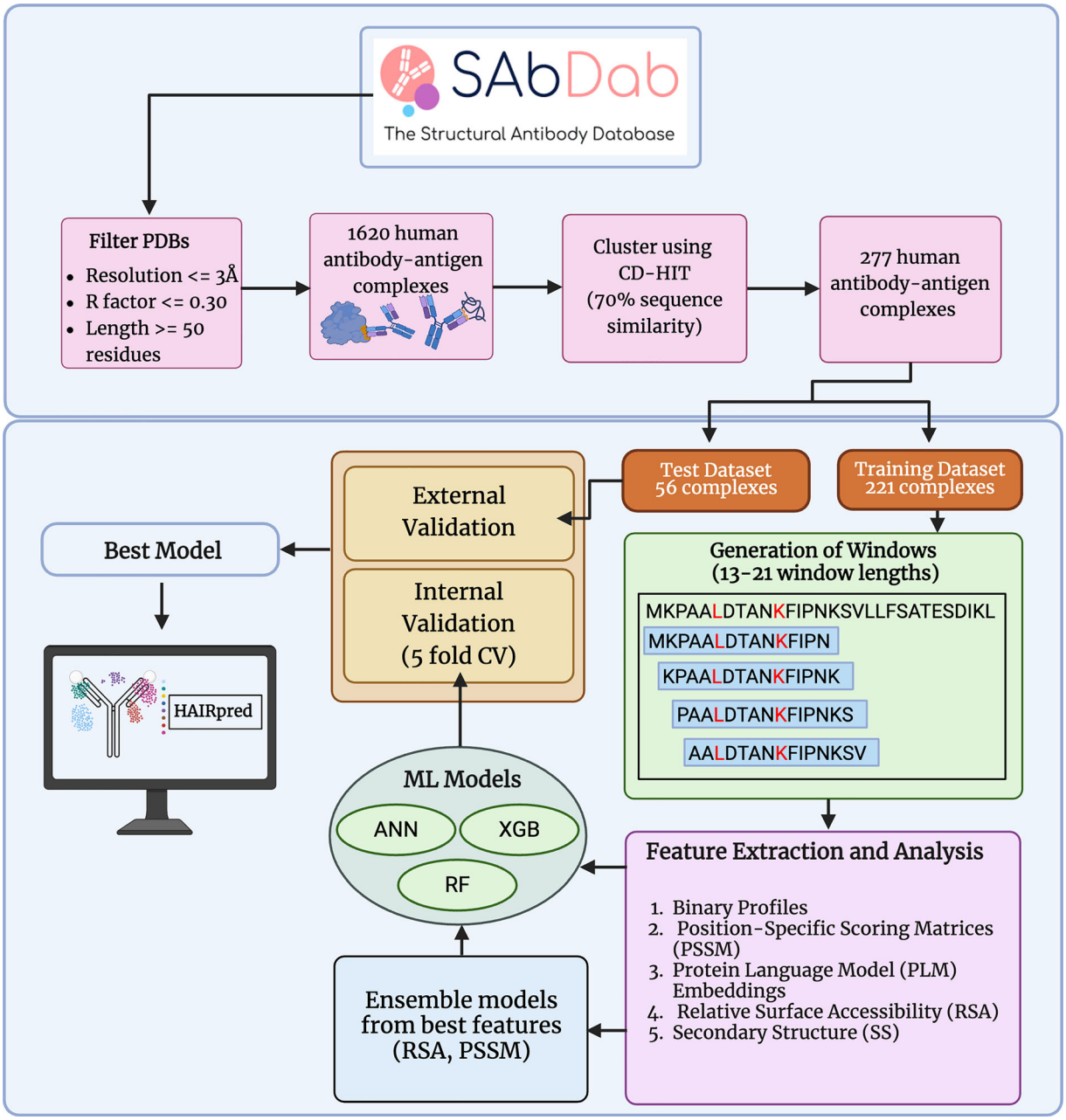


FIGURE 1 The complete workflow of the study.

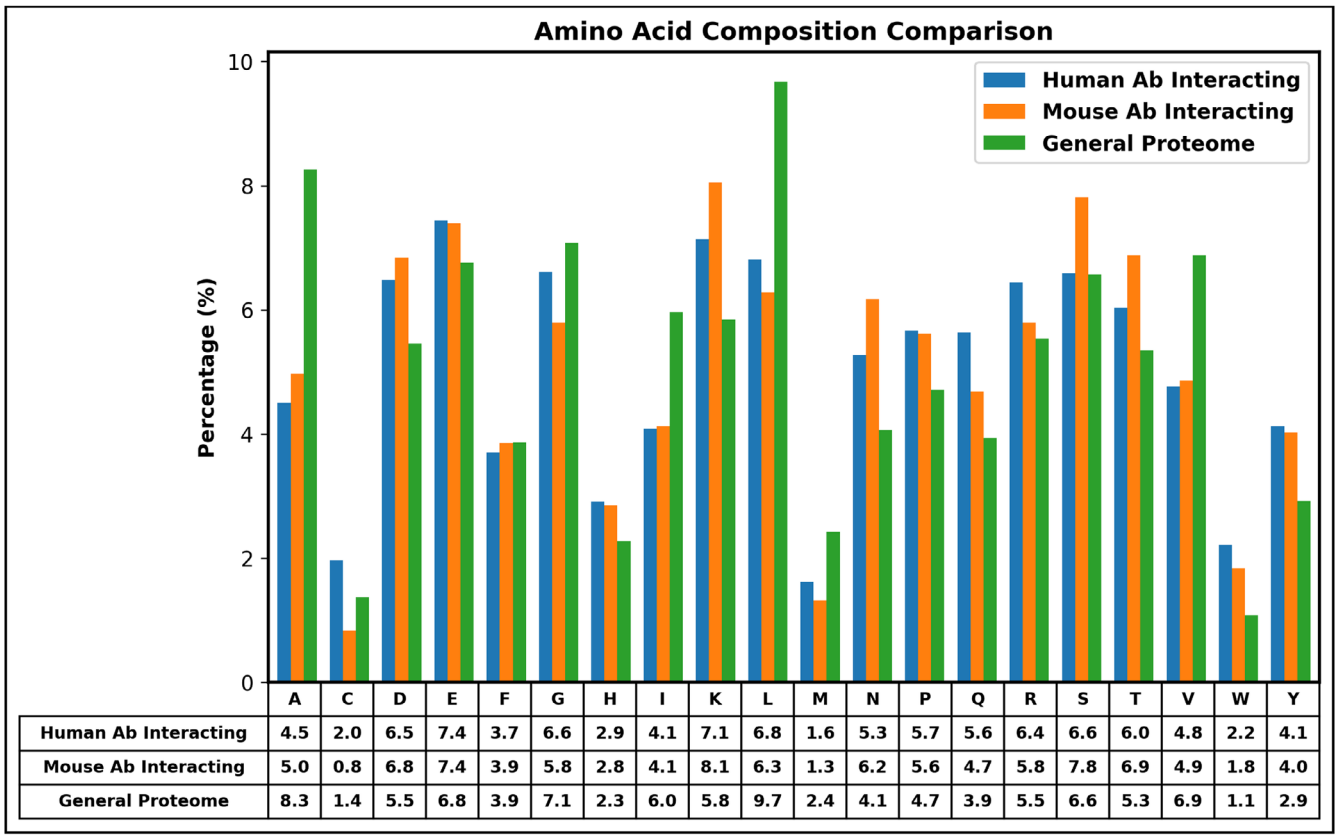
antibody interaction. These findings highlight the variations in antibody–antigen interactions between humans and mice and underscore the importance of host specificity, emphasizing the need for host-specific predictors.

2.2 | Data analysis

In data analysis, we analyze the human-specific antibody–antigen complexes, which would be used to train, test, and evaluate the models used in this study. We have performed the compositional analysis for the antibody interacting and non-interacting residues. Additionally, the Two Sample Logo was developed on the patterns generated from the antigen sequences.

2.2.1 | Analysis of human interacting residues

We compared the amino acid composition of antibody interacting residues and non-interacting residues (Figure 3) to understand the preference of residues. The *p*-values are generated using the chi-square test. The composition of residues aspartic acid (D), glutamic acid (E), histidine (H), lysine (K), proline (P), glutamine



	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 2 Comparison of amino acid composition of antigens that interact with human and Mus musculus antibodies. It also includes amino acid composition of general proteome to provide comparison with base.

(Q), arginine (R), tryptophan (W), and tyrosine (Y) is significantly higher in antibody interacting regions. Similarly, the residues alanine (A), cysteine (C), glycine (G), isoleucine (I), leucine (L), and valine (V) are significantly underrepresented in antibody interacting regions. The results indicate that the antibody interacting regions are characterized by a mix of charged, polar, and aromatic residues, whereas non-interacting regions are dominated by hydrophobic and mildly polar residues.

2.2.2 | Two Sample Logo

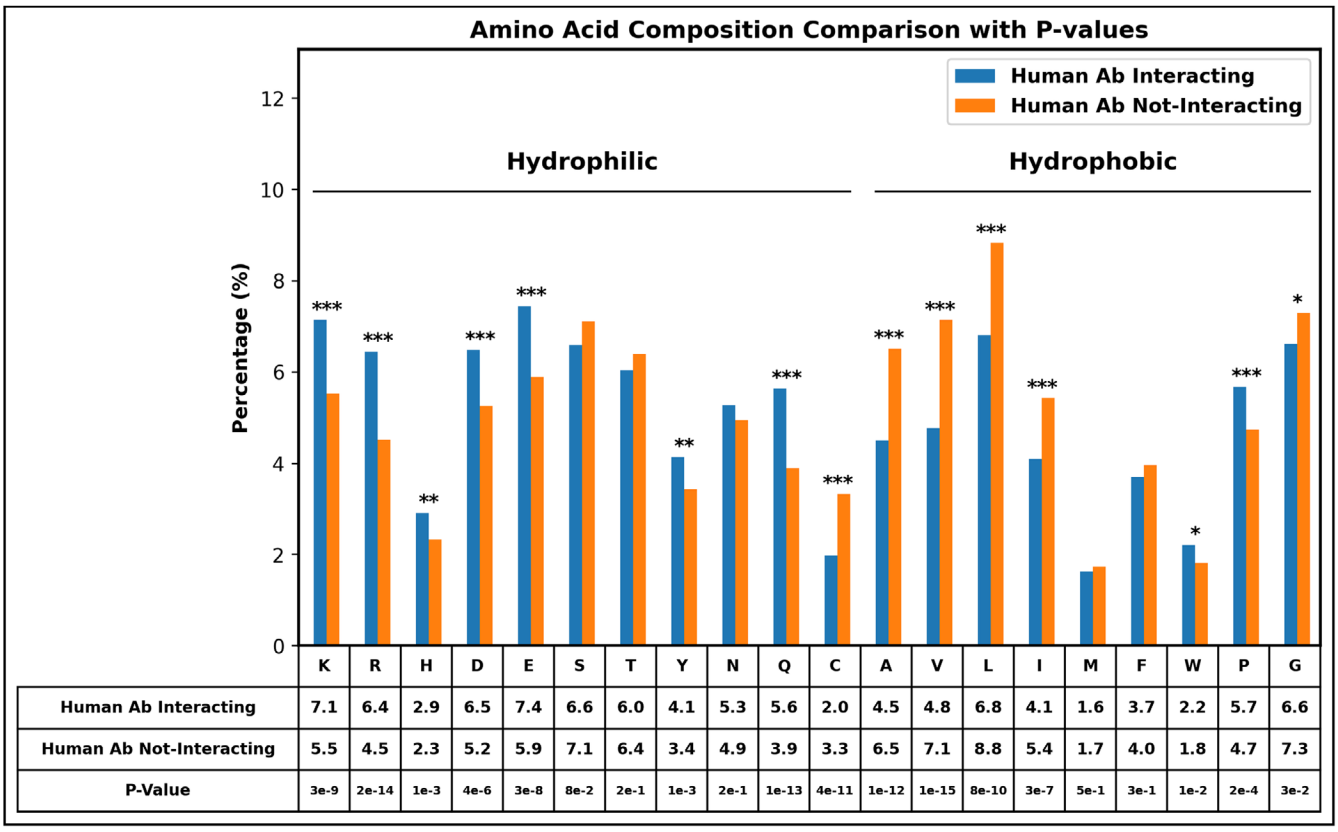
In this study, we have built the Two Sample Logo to understand the preference of a residue at a specific position in the human antibody interacting patterns. The two-sample logo is displayed in Figure 4. The central residue results in the Two Sample Logo are similar to the AAC performed in Figure 3. Alanine (A) is highly abundant in antibody non-interacting residues while in antibody interacting residues, glutamine (Q) and cysteine (C) are in abundance.

2.3 | Machine learning methods

In this section, machine learning (ML) models were developed to predict each residue in the input antigen sequence as either "antibody interacting" or "antibody non-interacting." As most of the machine learning techniques need inputs in fixed length numerical vectors to develop prediction models, we generate numerical features for the antibody interacting and antibody non-interacting patterns. The features, which capture antigen information, are generated for the patterns of length 17, as it has been shown in a number of studies in the past that a pattern/window length of 17 is most effective for predicting interacting residues (Chauhan et al. 2012; Panwar et al. 2013; Patiyal et al. 2023).

2.3.1 | Models developed using sequence and evolutionary features

In this study, we derived the one hot encoding and PSSM profiles for each pattern to extract sequence and evolutionary information of a pattern. These features were used to train the machine learning algorithms and the predictive performance for each model was then evaluated against the independent test dataset. As shown in Table 1, the random forest (RF) model, trained on one hot encoding profile feature, achieved the maximum performance with the area under the receiver operating characteristic (AUROC) score of 0.61 and the Matthews correlation coefficient



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

FIGURE 3 Percentage composition of antibody interacting residues and antibody non-interacting residues in antigen.

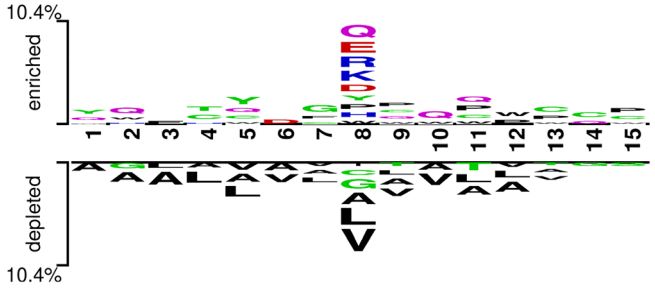


FIGURE 4 Two Sample Logo constructed from the interacting and non-interacting patterns in the human dataset.

(MCC) of 0.11. The performance of our models improved significantly when we used evolutionary information of patterns in the form of a PSSM profile. As shown in Table 1, our RF-based model obtained the highest AUROC of 0.67 with an MCC of 0.17. These results agree with previous studies where the PSSM-based models perform better than the sequence-based models (Kaur and Raghava 2003; Kaur and Raghava 2004).

2.3.2 | Models developed using PLM embeddings and structural features

In addition to sequence and evolutionary information, we also used embeddings generated by protein language models (PLMs). PLMs are transformer-based models that generate embeddings that capture contextual relationships between amino acids from large-scale protein data (Rao et al. 2020). In this study, we used PLM embeddings from ESM2 and ProtTrans as a feature vector. Additionally, PLMs were used to predict structural information, including relative solvent accessible surface area (RSA) and secondary structure (SS), which were used as feature vectors. Here, RSA features were created using ESMFold, and the SS features were derived from both ESMFold and ProtTrans. Similar to sequence-based features, each feature vector was generated from the training dataset on the patterns of length 17 and used to train the machine learning models. The predictive performance of each feature was then evaluated against the independent test dataset.

We observe that the PLM embeddings yielded the highest AUROC score of 0.61, which is lower than models utilizing evolutionary profiles. One of the possible reasons for the poor performance of the PLM model is that it extracts embeddings from a single sequence, whereas evolutionary information-based models use multiple sequences (homologs). Evolutionary information from PSSM captures more information than information in a single sequence. It is observed that the RSA feature gives the best performance, with the AUROC score reaching 0.67. The results are displayed in Table 2. Additional data is available in Table S1, Supporting Information.

TABLE 1 The performance of machine learning based models on independent dataset, developed using binary and PSSM profiles at window length 17.

	Performance on independent dataset				
Models	**Sensitivity**	**Specificity**	**Accuracy**	**AUROC**	**MCC**
Binary profile					
ANN	0.57	0.54	0.58	0.58	0.08
XGB	0.56	0.55	0.55	0.58	0.08
RF	0.58	0.57	0.57	0.61	0.11
PSSM					
ANN	0.62	0.54	0.55	0.62	0.12
XGB	0.66	0.57	0.58	0.67	0.18
RF	0.67	0.55	0.57	0.67	0.17

Abbreviations: ANN, Artificial Neural Network; AUROC, area under the receiver operating characteristic; MCC, Matthew's correlation coefficient; RF, Random Forest; XGB, XGBoost.

TABLE 2 The performance of machine learning models developed using PLM generated embeddings, predicted secondary structure and relative surface accessibility.

	Performance on independent dataset				
Models	**Sensitivity**	**Specificity**	**Accuracy**	**AUROC**	**MCC**
PLM embeddings					
ESM-2					
ANN	0.48	0.61	0.59	0.57	0.07
XGB	0.60	0.54	0.55	0.60	0.10
RF	0.46	0.69	0.65	0.61	0.11
ProtT5					
ANN	0.52	0.56	0.55	0.56	0.06
XGB	0.61	0.48	0.50	0.57	0.07
RF	0.47	0.65	0.61	0.58	0.09
Relative solvent accessibility					
ESMFold					
ANN	0.73	0.49	0.54	0.64	0.18
XGB	0.52	0.68	0.65	0.66	0.16
RF	0.78	0.46	0.52	0.67	0.20
Secondary structure					
ESMFold					
ANN	0.54	0.53	0.53	0.55	0.05
XGB	0.16	0.89	0.75	0.57	0.06
RF	0.22	0.85	0.73	0.59	0.08
ProtT5					
ANN	0.55	0.52	0.53	0.55	0.06
XGB	0.49	0.60	0.58	0.57	0.07
RF	0.52	0.58	0.57	0.57	0.08

Abbreviations: ANN, Artificial Neural Network; AUROC, area under the receiver operating characteristic; MCC, Matthew's correlation coefficient; RF, Random Forest; XGB, XGBoost.

2.3.3 | Ensemble models

In our analysis, we observed that the models developed using PSSM profile and RSA performed better than other models. In order to utilize the strength of both these features, we developed an ensemble model that combines predictions from these two types of features. As shown in Table 3, we achieved the highest AUROC of 0.71 with an MCC of 0.22 at pattern length 17. So far, we used window/pattern length 17, as most

TABLE 3 The performance of the ensemble model using different window lengths on an independent dataset.

	Metrics evaluated on the independent test dataset				
Models	**Sensitivity**	**Specificity**	**Accuracy**	**AUROC**	**MCC**
Window length 13					
ANN	0.66	0.57	0.59	0.67	0.18
XGB	0.72	0.54	0.57	0.70	0.20
RF	0.79	0.48	0.54	0.71	0.22
Window length 15					
ANN	0.65	0.57	0.58	0.66	0.17
XGB	0.73	0.53	0.57	0.71	0.21
RF	0.78	0.51	0.56	0.72	0.23
Window length 17					
ANN	0.65	0.58	0.59	0.67	0.18
XGB	0.71	0.56	0.59	0.70	0.21
RF	0.76	0.52	0.57	0.71	0.22
Window length 19					
ANN	0.65	0.57	0.58	0.66	0.17
XGB	0.71	0.56	0.59	0.70	0.21
RF	0.76	0.53	0.57	0.71	0.23
Window length 21					
ANN	0.59	0.66	0.65	0.67	0.20
XGB	0.72	0.56	0.59	0.71	0.22
RF	0.75	0.53	0.57	0.71	0.22

Abbreviations: ANN, Artificial Neural Network; AUROC, area under the receiver operating characteristic; MCC, Matthew's correlation coefficient; RF, Random Forest; XGB, XGBoost.

of the previous studies used length 17. We also wanted to understand the effect of window length on the performance of our models. We developed models using pattern lengths 13–21 and computed the performance of our ensemble models. It was observed that our RF-based model performs best on pattern length 15 with AUROC 0.72 and MCC 0.23. Generally, a pattern length of 15 performs better than a pattern length of 17 by a marginal amount. Additional data is available in Table S2.

We observe that the Random Forest ensemble of classifiers trained on RSA and PSSM with a pattern length of 15 achieved the highest performance in predicting antibody interacting regions in the antigen. This model is termed as HAIRpred (Human Antibody Interacting Residue predictor). The performance of the ensemble model compared to its individual features is displayed in Figure 5.

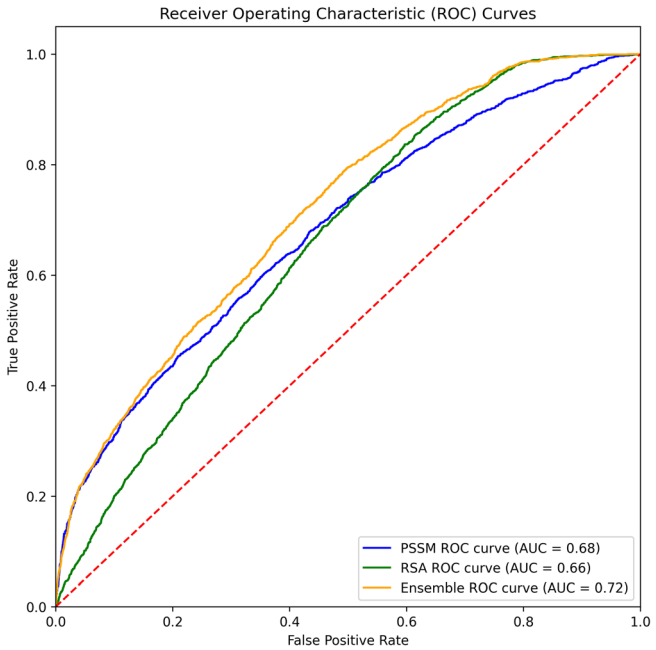


FIGURE 5 ROC curves of ensemble model (pattern length = 15) compared with individual features.

2.4 | Benchmarking

It is important for any study to benchmark its tools against existing state-of-the-art methods. To the best of our knowledge, no method has been developed so far for predicting conformational B-cell epitopes for human hosts or for predicting human antibody interacting residues. Thus, a direct comparison of our method with previous studies is not possible.

TABLE 4 The performance of our method HAIRpred and existing methods on independent dataset.

Model	Year	Sensitivity	Specificity	Accuracy	AUROC	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

Abbreviations: AUROC, area under the receiver operating characteristic; MCC, Matthew's correlation coefficient.

However, several methods have been developed in the literature for predicting conformational B-cell epitopes not specific to any host. Thus, we evaluate these general methods of B-cell epitope prediction on an independent dataset used in this study to benchmark these methods with the proposed HAIRpred method. The B-cell epitope prediction methods used for benchmarking include 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. 2021). As shown in Table 4, none of the existing methods achieved an AUROC of more than 0.62 on the independent dataset, whereas our methods achieved an AUROC of 0.72 on the same dataset. The results in Table 4 demonstrate the superior performance of HAIRpred compared to existing methods.

In addition to this, we also wanted to understand how our method would perform on data from other hosts. Therefore, we also tested HAIRpred on the Mus musculus dataset derived from experimental results through SAbDaB. It is observed that the performance of our methods decreases significantly on the Mus musculus dataset. These observations further support the development of host-specific methods for predicting conformational B-cell epitopes.

2.5 | Feature importance analysis

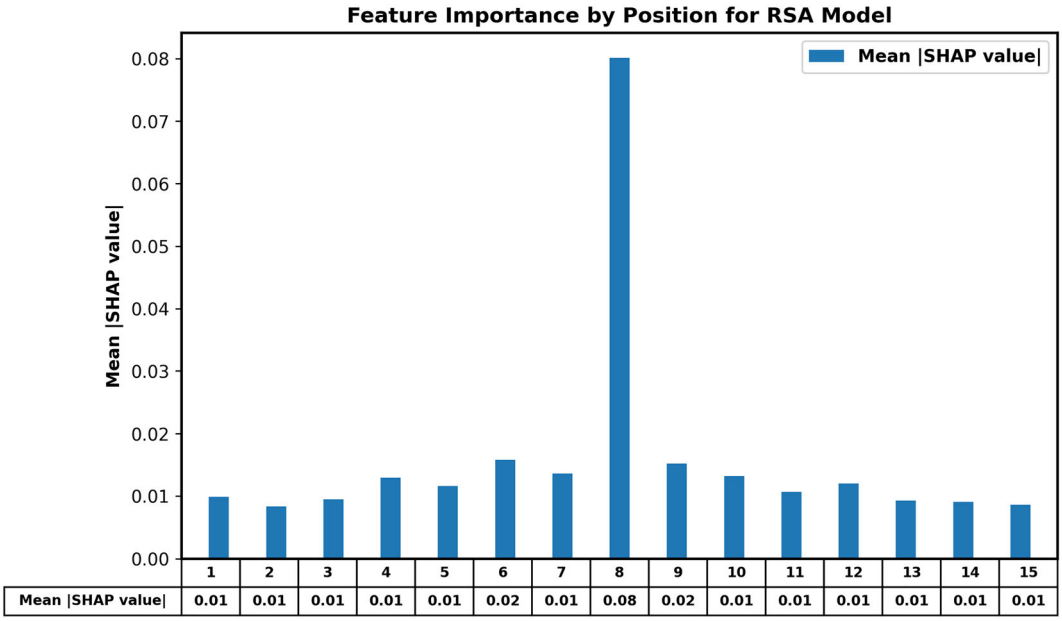
HAIRpred is an ensemble method that combines two Random Forest-based models, trained on RSA and PSSM on pattern length 15, respectively. To understand the decision process of these two individual models, Shapley values were calculated for each individual model using the Python shap package. The Shapley values indicate the importance of a particular position in the feature vector in the final prediction.

Figure 6 shows the feature importance by position on the pattern for the RSA trained model. It is observed that the RSA of the central residue in the pattern has the largest influence on the decision making process of the RSA-trained Random Forest model. Similarly, Shapley values were calculated for the PSSM-trained model, and the top features with the highest Shapley values are: 8Q, 8E, 8R, 8K, 8H, 8N, 8C, 8D, 8A, and 11A. Here, the number refers to the position of the residue in the pattern and represents the column of the PSSM matrix. It is similarly observed that the central residue in the pattern has the largest influence on the decision making process for the PSSM-trained model.

2.6 | Web server implementation

To serve the scientific community, we have developed a user-friendly web server that integrates the best-performing prediction models from our study. This web server provides an accessible platform for researchers to predict antibody interacting residues in antigen sequences. Users can submit antigen sequences in FASTA format, and the server uses the selected model to predict antibody-binding sites in each submitted antigen. The output is designed to be both extensive and user-friendly, offering results in three different formats: sequential, graphical, and tabular (see Figure 7). In addition to providing predictions, the server allows users to download the training and testing datasets used in this study. The server also has a "Help" section to assist the users in navigating the features of the platform effectively. The server has been implemented using a combination of HTML, JavaScript, and PHP scripts. The web server has robust functionality and compatibility across a wide range of devices, including laptops, Android smartphones, iPhones, and iPads. There is one limitation of the web server due to the computational restraints: the lack of a design module that could use the findings of HAIRpred to redesign the antigen: antibody complexes, similar to approaches done in previous studies (Padhi et al. 2021). We hope that the scientific community will use our findings to develop such modules.

We have also developed an open-source standalone version of HAIRpred to enhance its accessibility. This package offers researchers the flexibility to run predictions locally without the need for an active internet



	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 6 Shapley values calculated for each position of the pattern for the RSA model.

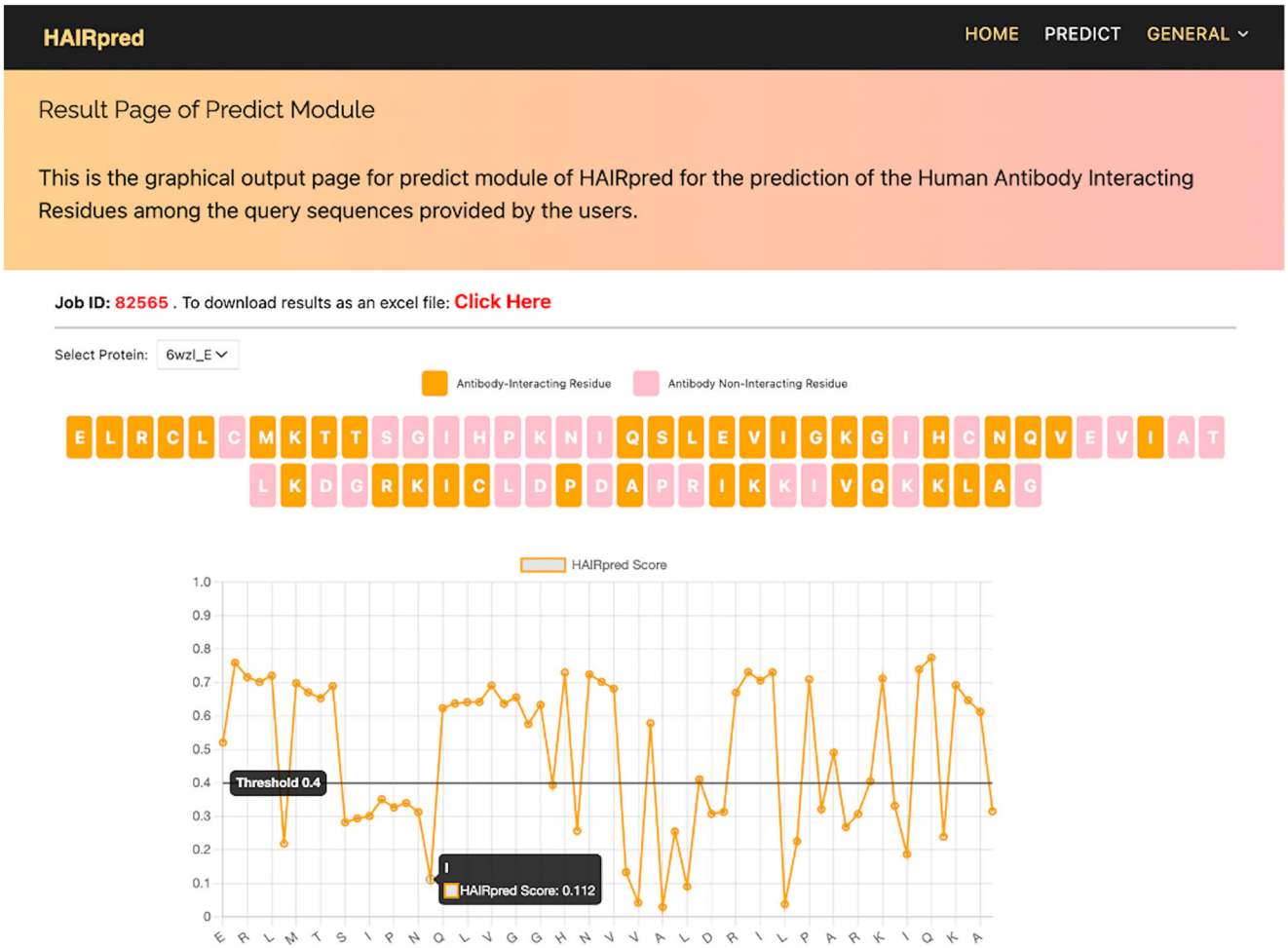


FIGURE 7 HAIRpred web server interface showing antibody interacting residue prediction results.

connection. The standalone version is available on GitHub (https://github.com/raghavagps/HAIRpred) and can also be installed as a python package, accessible through pip (pip install HAIRpred). The standalone software has been designed in such a way that small antigens (less than 400 residues) can be predicted on resource-constrained systems. However, for larger antigens, GPU support is recommended to ensure optimal performance and faster predictions. We have also performed a time analysis of the standalone software to provide an estimate of the speed of our software. We measured the time it took for the software to run different numbers of sequences of varying sizes on a system with i5 14th Generation processor with 32GB RAM, and the results are displayed in Table S3.

Both the web server and standalone package aim to democratize access to advanced computational tools, fostering advancements in antigen–antibody interaction research. The open-source web server can be accessed at https://webs.iiitd.edu.in/raghava/hairpred/, providing a seamless and reliable resource for the scientific community.

3 | DISCUSSION AND CONCLUSION

Antibodies are vital glycoproteins in the adaptive immune system, designed to bind to foreign particles through specific regions called epitopes (Zeng et al. 2023). Identifying these antibody-binding regions is crucial for advancing immunological research and enabling the development of vaccines, therapeutics, and diagnostic tools (Correia et al. 2014; Csepregi et al. 2020; Norman et al. 2020). However, existing predictive models face several challenges, including limited host specificity and an inability to account for species-specific variations in immune responses and antibody structures (da Silva et al. 2022; Haste Andersen et al. 2006; Kulkarni-Kale et al. 2005; Rubinstein et al. 2009; Shashkova et al. 2022; Solihah et al. 2020; Sun et al. 2009; Sweredoski and Baldi 2008). These shortcomings often lead to suboptimal performance when applied to experimental datasets (Cia et al. 2023). Furthermore, many models lack interpretability, offering little understanding of the factors influencing antibody–antigen interactions. To address these limitations, our study focused exclusively on human antibody–antigen complexes derived from experimental data. By narrowing the scope to human systems, we sought to overcome the drawbacks of generalized models that neglect species-specific differences. We curated a comprehensive feature set that includes structural, sequential, and evolutionary characteristics of antigens. Through rigorous testing using data from the Protein Data Bank and the SAbDab database, we identified key features such as RSA and PSSM as the most important determinants of antibody–antigen interactions.

Additionally, we investigated the influence of local antigenic environments by analyzing sequence patterns of varying lengths. This analysis resulted in the development of the Human Antibody Interacting Residue Predictor (HAIRpred), an ensemble-based random forest model designed to predict antibody interacting regions in antigens with high precision. HAIRpred represents a second-generation, host-specific tool that prioritizes human antibody interactions over generalized, cross-species predictions. The model demonstrated a remarkable AUROC of 0.72 on an independent experimental dataset, outperforming existing state-of-the-art models. It also demonstrated superior performance in identifying human antibody interacting residues compared to Mus musculus-derived antibodies, underscoring the importance of host specificity. Insights from SHAP (SHapley Additive exPlanations) analysis highlighted that central residues in antigenic patterns are pivotal in predicting interactions, enhancing the model's interpretability.

While these results mark significant progress, this study acknowledges one limitation: the sequence similarity threshold of 70% used for data clustering. This threshold was necessary to ensure an adequate amount of training data. However, as more experimental data becomes available, lowering this threshold to 40% will allow a more in-depth exploration of antigen–antibody interactions across diverse antigen structures, further improving the model's generalizability and performance. In conclusion, HAIRpred offers a major advancement in antibody–antigen interaction prediction by adopting a host-specific, human-centric approach. Its robust performance on human data makes it an invaluable resource for immunological research, with potential applications in vaccine design, therapeutic development, and diagnostics. By addressing the limitations of earlier models, HAIRpred sets a new standard for second-generation, host-specific prediction tools. To ensure accessibility, we have made both a user-friendly web server, PyPI package, and a standalone version of HAIRpred available at https://webs.iiitd.edu.in/raghava/hairpred/.

4 | MATERIALS AND METHODS

4.1 | Creation of datasets

We obtained antibody–antigen complexes for the human hosts from the SAdDab database (Dunbar et al. 2014). Antigens or proteins having a resolution less than 3.0 Å and an R-factor less than 0.30 were selected for further processing. All antigens with less than 50 residues were removed from the dataset. This resulted in a set of 1620 human antibody–antigen complexes. In a similar way, we obtained 290 Mus musculus antibody–antigen complexes from the SAbDaB database. In order to remove redundancy from antigens recognized by human antibodies, we clustered these antigens using CD-HIT (Fu et al. 2012) with a 70% sequence similarity cut-off. Our dataset contains 277 high-quality

non-redundant antigens, where no two antigens have more than 70% similarity with each other. This dataset was further divided into training and independent datasets (80:20 ratio), where the training dataset contained 221 and the independent dataset contained 56 antigens. In order to assign antibody interacting residues in these antigens, we examined changes in the RSA of these residues with and without antibody binding, similar to the algorithm performed by Cia et al. (2023). A residue is assigned antibody interacting residues if it undergoes a change in RSA of at least 5% upon binding with an antibody ($RSA_{unbound} - RSA_{bound} \geq 5\%$) (Cia et al. 2023).

4.2 | Creation of overlapping patterns

Numerous studies have shown that the residue's function is influenced by its local environment (Ansari and Raghava 2010; Kaur and Raghava 2003; Kaur and Raghava 2004). To capture the local environment of each residue, we created overlapping patterns (or windows) for each residue in the antigen, with window sizes ranging from 13 to 21. To generate a pattern for the terminal residues, we added a dummy variable "*X*" at both ends of the antigen for the residues not covered in the pattern. Each pattern's corresponding label (antibody interacting/non-interacting) is the label of the central residue. This approach has been commonly used in multiple studies (Bhasin and Raghava 2005; Kumar et al. 2008).

Patterns were created for both the training dataset and the independent test dataset. The patterns generated for the training dataset were imbalanced, that is, the non-interacting patterns are much larger in number. This imbalance can introduce bias in the model, causing it to favor the majority class (non-interacting patterns). To avoid any bias due to the imbalance, the balanced training dataset was made, which consisted of all antibody interacting patterns and an equal number of randomly selected non-interacting patterns.

4.3 | Compositional analysis

Amino acid composition is the frequency of each of the 20 amino acids within a peptide or protein sequence. It is represented as a feature vector with 20 elements, where each component corresponds to the fraction of a particular amino acid within the sequence (Pande et al. 2023). It can be calculated by Equation (1),

$$AAC_j = \frac{P_j}{Q}, \quad (1)$$

where AAC_j is an amino acid composition of residue type j and P_j and Q are the number of residues of type j and the length of the sequence, respectively.

In Figure 2, the AAC of human and mouse epitopes are compared to understand their differences. The labeled data of both humans and mice were filtered to only get the antibody interacting residues (epitopes), and the AAC was calculated using Equation (1). The general proteome AAC was 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) (Walker 2005).

In Figure 3, the AAC of human antibody interacting residues and non-interacting residues were compared. The labeled human-specific antigen data were classified as per the labels, and AAC was calculated as per Equation (1). *p*-values were calculated using the chi-square test to understand the significance of the difference in AAC of each residue, and it was calculated using the scipy package (Virtanen et al. 2020). The label for significance is (*), and the number of (*) is calculated using thresholds with (*) for a *p*-value below 0.05, (**) for a *p*-value below 0.01, and (***) for a *p*-value below 0.001.

4.4 | Sequence logo

To show sequence conservation in antibody interacting and non-interacting patterns, we generated the sequence logo using a web-based application called "Two Sample Logo." This tool calculates the significance of each residue at each position and compares distributions between positive and negative samples. Residues are grouped based on enrichment or depletion in the positive sample, and the graphical output can display statistically significant residues with symbol sizes proportional to the difference between the groups. The *p*-value is calculated using a *t* test (Vacic et al. 2006). To generate the Two Sample Logo, we inputted all the antibody interacting patterns in the positive sample. To get a better understanding, we only considered those antibody non-interacting patterns in which no residue was antibody interacting. This was done to ensure the elimination of potential bias from the antibody interacting residues. From these patterns, we then randomly selected the negative sample to have the total number of patterns to be equal in both positive and negative sample.

4.5 | Curation of feature set

Most of the existing machine learning techniques need input in the form of a numerical vector or matrices. Thus, we need to generate features that would represent the interacting and non-interacting patterns of antigens in the form of a numerical vector. In this study, we curated a wide range of features that capture structural, sequential, and evolutionary information about antigens from their sequence. The curated set of features includes one hot encoding profile, PLM embeddings, PSSM, predicted RSA, and SS.

One hot encoding profile was generated by assigning binary values to each amino acid. This results in a 20-length vector for each residue in which only one position corresponding to that amino acid is set to 1, while all other positions are set to 0. The amino acid order used here is ['A','R','N','D','C','Q','E','G','H','I','L','K','M','F','P','S','T','W','Y','V']. For example, arginine (A) was represented as [1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]. Additionally, to account for the patterns of the terminal residues, the dummy variable *X* is encoded as 20-length zero vector.

One hot encoding profile was created for each pattern by concatenating the vectors of all residues in the pattern, resulting in the dimension of the feature vector as $N \times 20$, where *N* is the pattern length (Ansari and Raghava 2010).

PSSM are matrices containing information about the probability of amino acids occurring in each position. It is derived from a multiple sequence alignment and it represents the evolutionary history of the protein. Here, we generated PSSM profiles of a sequence using Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) (Altschul et al. 1997) and searching against the Swiss Prot database (Bairoch and Apweiler 1999). The search was performed with an *E*-value threshold of 0.1 and a word size of 3 for protein sequences in three iterations. The alignment was conducted using BLOSUM62 as the scoring matrix with gap penalties of values of 11 (gap opening) and 1 (gap extension). We generated PSSM profiles for each antigen in the complex and then divided them into patterns. The dummy variable *X* is assigned a 20-length zero vector while creating the feature vector. The resulting feature vector is a matrix of dimensions $N \times 20$, where *N* is the length of the pattern.

PLMs are transformer-based models which generate embeddings that capture contextual relationships between amino acids, reflecting evolutionary and functional information. Here we use ESM2 (esm2_t6_8-M_UR50D) (Lin et al. 2023) and Encoder only ProtTrans (ProtT5-XL-UniRef50) (Elnaggar et al. 2022) to generate embeddings for the patterns. The patterns are inputted into the models and the last layer embeddings are extracted as a feature vector.

RSA gives a measure of the extent of exposure of a residue in the 3D structure. To predict RSA from sequence, the 3D structure was predicted using ESMFold (Lin et al. 2023), a PLM-based protein structure prediction model. The predicted structures were then processed using the DSSP algorithm (Kabsch and Sander 1983) to get RSA values for each residue. These predicted RSA values were then divided into patterns corresponding to their respective residues. The dummy variable *X* is assigned the RSA of value 0. The resulting feature vector is a vector of length *N*, where *N* is the length of the pattern. We conducted a comparative analysis between the RSA values predicted using this approach and those calculated directly from the PDB structure to evaluate the reliability of the predicted RSA. The results indicate a Pearson correlation coefficient of 0.75 between the two sets of values. This substantial correlation provides strong evidence supporting the accuracy of the RSA predictions generated using ESM, thereby justifying their application in this context. We also calculated the results of our models using the actual RSA to see our comparison with predicted RSA. These results are displayed in Table S5.

SS provides information about the local structure of the protein backbone. Secondary structure was predicted in two ways: (1) from ProtT5-XL-U50 embeddings (8-class SS) (Elnaggar et al. 2022) and (2) from the DSSP algorithm applied to the 3D structure predicted by ESMFold (3-class SS) (Lin et al. 2023). The predicted secondary structures were then divided into patterns corresponding to their respective residues. The dummy variable *X* is assigned the secondary structure of a coil. The secondary structure is then encoded using a label encoder, resulting in a vector of length *N*, where *N* is the length of the pattern.

4.6 | Machine learning and neural network models

The machine learning algorithms were implemented using scikit-learn (Pedregosa et al. 2012) to develop good predictive models. We focused on the Random Forest and XGBoost algorithms due to their strong classification performance and ability to handle variable datasets. The hyperparameters of the models were optimized using GridSearchCV from the scikit-learn package. In addition to traditional machine learning algorithms, we implemented neural network models using PyTorch (Paszke et al. 2019). Neural networks are highly effective for capturing complex patterns in data due to their multilayered architecture. The neural networks were optimized using backpropagation and the Adam optimizer. Cross-entropy loss was used as the loss function to measure prediction error and guide model optimization. Ensemble models were also created to combine information from different features. We have used a simple method for creating ensemble models, which involves averaging the probabilities generated by the individual models. The averaged probabilities are then used to decide the predicted label.

4.7 | Training and evaluation of models

To optimize our classification models, we used the five-fold cross-validation technique. In this approach, the dataset is randomly divided into five equal-sized "folds." The model is trained on four of these folds and evaluated on the remaining fold, with this process

repeated five times. The results from each fold are then averaged to provide an estimate of the model's performance. This method helps optimize the model's hyperparameters to maximize AUROC. Although different folds were used for training and testing, some degree of overoptimization cannot be fully ruled out with five-fold cross-validation. Therefore, we evaluated our final models on an independent dataset that was not used for hyperparameter optimization.

4.7.1 | Evaluation parameters

In this study, the models are performing binary classification on each window to label each residue as antibody interacting/non-interacting. The performance of the models is evaluated based on the standard evaluation parameters. The evaluation parameters can be divided into threshold-dependent and threshold-independent parameters. The threshold-independent parameter used in this study is the AUROC, and the threshold-dependent parameters are Sensitivity, Specificity, Accuracy, and MCC. These parameters are implemented using the scikit-learn package,

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}, \tag{3}$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}, \tag{4}$$

$$\text{MCC} = \frac{(\text{TP}*\text{TN})-(\text{FP}*\text{FN})}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}. \tag{5}$$

Here TN, TP, FN, and FP stand for true negative, true positive, false negative, and false positive, respectively.

4.8 | Explainable machine learning using SHAP

Understanding the decision making process of a given machine learning model is crucial for model reliability, especially in complex biological applications. In this study, we used SHapley Additive exPlanations (SHAP), a post-hoc interpretability method based on Shapley Values and cooperative game theory, to analyze machine learning models (Lundberg and Lee 2017) and generate 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. The Python SHAP package was used for SHAP TreeExplainer to interpret the results of the individual models of the HAIRpred's Random Forest ensemble.

4.9 | Web server

To aid the scientific community, we designed a web server called "HAIRpred" to predict human antibody interacting residues in the antigen (https://webs.iiitd.edu.in/raghava/hairpred/). The user-friendly front end was developed using HTML, Javascript, and PHP scripts. In addition to the web-based platform, we developed a standalone version of HAIRpred and a pip package.

AUTHOR CONTRIBUTIONS

Ruchir Sahni: Investigation; validation; software; formal analysis; data curation; writing – original draft; writing – review and editing; methodology; visualization. **Nishant Kumar:** Methodology; validation; writing – review and editing; writing – original draft; software; formal analysis; visualization. **Gajendra P. S. Raghava:** Conceptualization; investigation; funding acquisition; writing – original draft; writing – review and editing; validation; methodology; software; project administration; data curation; supervision; visualization; resources; formal analysis.

ACKNOWLEDGMENTS

The current work has been supported by the Department of Biotechnology (DBT) grant BT/PR40158/BTIS/137/24/2021. Authors are thankful to the University Grants Commission (UGC) and DST-Inspire (KVPY), for fellowships and financial support, and the Department of Computational Biology, IIITD New Delhi for infrastructure and facilities. We would like to acknowledge that figures were created using BioRender.com.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

All the datasets used in this study are available at the "HAIRpred" web server, at https://webs.iiitd.edu.in/raghava/hairpred/.

ORCID

Ruchir Sahni https://orcid.org/0000-0002-9771-5496
Nishant Kumar https://orcid.org/0000-0001-7781-9602
Gajendra P. S. Raghava https://orcid.org/0000-0002-8902-2876

REFERENCES

Almagro JC, Hernández I, Ramírez MC, Vargas-Madrazo E. Structural differences between the repertoires of mouse and human

germline genes and their evolutionary implications. Immunogenetics. 1998;47:355–63.
Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25: 3389–402.
Ansari HR, Raghava GP. Identification of conformational B-cell epitopes in an antigen from its primary sequence. Immunome Res. 2010;6:6.
Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Res. 1999;27:49–54.
Bhasin M, Raghava GPS. Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences. Nucleic Acids Res. 2005; 33:W202–7.
Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res. 2019;47:D464–74.
Chauhan JS, Bhat AH, Raghava GPS, Rao A. GlycoPP: a webserver for prediction of N- and O-glycosites in prokaryotic protein sequences. PLoS One. 2012;7:e40155.
Cia G, Pucci F, Rooman M. Critical review of conformational B-cell epitope prediction methods. Brief Bioinf. 2023;24:bbac567. https://doi.org/10.1093/bib/bbac567
Clifford JN, Høie MH, Deleuran S, Peters B, Nielsen M, Marcatili P. BepiPred-3.0: improved B-cell epitope prediction using protein language models. Protein Sci. 2022;31:e4497.
Collatz M, Mock F, Barth E, Hölzer M, Sachse K, Marz M. EpiDope: a deep neural network for linear B-cell epitope prediction. Bioinformatics. 2021;37:448–55.
Correia BE, Bates JT, Loomis RJ, Baneyx G, Carrico C, Jardine JG, et al. Proof of principle for epitope-focused vaccine design. Nature. 2014;507:201–6.
Csepregi L, Ehling RA, Wagner B, Reddy ST. Immune literacy: reading, writing, and editing adaptive immunity. iScience. 2020;23: 101519.
da Silva BM, Myung Y, Ascher DB, Pires DEV. epitope3D: a machine learning method for conformational B-cell epitope prediction. Brief Bioinf. 2022;23:bbab423. https://doi.org/10.1093/bib/bbab423
Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, et al. SAbDab: the structural antibody database. Nucleic Acids Res. 2014;42:D1140–6.
Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell. 2022;44:7112–27.
Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28: 3150–2.
Haste Andersen P, Nielsen M, Lund O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci. 2006;15:2558–67.
Ivanisenko NV, Shashkova TI, Shevtsov A, Sindeeva M, Umerenkov D, Kardymon O. SEMA 2.0: web-platform for B-cell conformational epitopes prediction using artificial intelligence. Nucleic Acids Res. 2024;52:W533–9.
Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res. 2017;45:W24–9.
Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22:2577–637.
Kaur H, Raghava GPS. A neural-network based method for prediction of gamma-turns in proteins from multiple sequence alignment. Protein Sci. 2003;12:923–9.
Kaur H, Raghava GPS. Prediction of alpha-turns in proteins using PSI-BLAST profiles and secondary structure information. Proteins. 2004;55:83–90.
Kozlova EEG, Cerf L, Schneider FS, Viart BT, NGuyen C, Steiner BT, et al. Computational B-cell epitope identification and production of neutralizing murine antibodies against Atroxlysin-I. Sci Rep. 2018;8:14904.
Kulkarni-Kale U, Bhosle S, Kolaskar AS. CEP: a conformational epitope prediction server. Nucleic Acids Res. 2005;33:W168–71.
Kumar M, Gromiha MM, Raghava GPS. Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins. 2008; 71:189–94.
Kumar N, Tripathi S, Sharma N, Patiyal S, Devi NL, Raghava GPS. A method for predicting linear and conformational B-cell epitopes in an antigen from its primary sequence. Comput Biol Med. 2024;170:108083.
Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379:1123–30.
Lundberg S, Lee S-I. A unified approach to interpreting model predictions. 2017. arXiv [cs.AI] [Internet]. Available from: https://doi.org/10.48550/ARXIV.1705.07874
Mestas J, Hughes CCW. Of mice and not men: differences between mouse and human immunology. J Immunol. 2004; 172:2731–8.
Norman RA, Ambrosetti F, Bonvin AMJJ, Colwell LJ, Kelm S, Kumar S, et al. Computational approaches to therapeutic antibody design: established methods and emerging trends. Brief Bioinf. 2020;21:1549–67.
Padhi AK, Kumar A, Haruna K-I, Sato H, Tamura H, Nagatoishi S, et al. An integrated computational pipeline for designing high-affinity nanobodies with expanded genetic codes. Brief Bioinf. 2021;22:bbab338. https://doi.org/10.1093/bib/bbab338
Pande A, Patiyal S, Lathwal A, Arora C, Kaur D, Dhall A, et al. Pfeature: a tool for computing wide range of protein features and building prediction models. J Comput Biol. 2023;30:204–22.
Panwar B, Gupta S, Raghava GPS. Prediction of vitamin interacting residues in a vitamin binding protein using evolutionary information. BMC Bioinf. 2013;14:44.
Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. 2019. arXiv [cs.LG] [Internet]. Available from: https://doi.org/10.48550/ARXIV.1912.01703
Patiyal S, Dhall A, Bajaj K, Sahu H, Raghava GPS. Prediction of RNA-interacting residues in a protein using CNN and evolutionary profile. Brief Bioinf. 2023;24:bbac538. https://doi.org/10.1093/bib/bbac538
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. 2012. arXiv [csLG] [Internet]. Available from: https://doi.org/10.48550/ARXIV.1201.0490
Rao R, Meier J, Sercu T, Ovchinnikov S, Rives A. Transformer protein language models are unsupervised structure learners. 2020. bioRxiv [Internet]. Available from: https://doi.org/10.1101/2020.12.15.422761
Robert PA, Akbar R, Frank R, Pavlović M, Widrich M, Snapkov I, et al. Unconstrained generation of synthetic antibody-antigen structures to guide machine learning methodology for antibody specificity prediction. Nat Comput Sci. 2022;2:845–65.
Rubinstein ND, Mayrose I, Martz E, Pupko T. Epitopia: a web-server for predicting B-cell epitopes. BMC Bioinf. 2009;10:287.
Shashkova TI, Umerenkov D, Salnikov M, Strashnov PV, Konstantinova AV, Lebed I, et al. SEMA: antigen B-cell conformational epitope prediction using deep transfer learning. Front Immunol. 2022;13:960985.
Solihah B, Azhari A, Musdholifah A. Enhancement of conformational B-cell epitope prediction using CluSMOTE. PeerJ Comput Sci. 2020;6:e275.

Sun J, Wu D, Xu T, Wang X, Xu X, Tao L, et al. SEPPA: a computational server for spatial epitope prediction of protein antigens. Nucleic Acids Res. 2009;37:W612–6.

Sweredoski MJ, Baldi P. PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. Bioinformatics. 2008;24:1459–60.

Vacic V, Iakoucheva LM, Radivojac P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics. 2006;22:1536–7.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–72.

Vita R, Blazeska N, Marrama D, IEDB Curation Team Members, Duesing S, Bennett J, et al. The immune epitope database (IEDB): 2024 update. Nucleic Acids Res. 2025;53(D1):D436–43.

Walker JM. The proteomics protocols handbook. Totowa, NJ: Humana Press; 2005.

Zeng X, Bai G, Sun C, Ma B. Recent progress in antibody epitope prediction. Antibodies (Basel). 2023;12:52. https://doi.org/10.3390/antib12030052

AUTHOR BIOGRAPHIES

Ruchir Sahni is currently studying as an integrated BS-MS student at Indian Institute of Science Education and Research (IISER) Pune, India. He is currently working as an Intern on Project position at Department of Computational Biology, Indraprastha Institute of Information Technology (IIIT), New Delhi, India.

Nishant Kumar is currently working as PhD in Computational Biology from Department of Computational Biology, Indraprastha Institute of Information Technology (IIIT), New Delhi, India.

Gajendra P. S. Raghava is currently working as Professor and Head of Department of Computational Biology, Indraprastha Institute of Information Technology (IIIT), New Delhi, India.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Sahni R, Kumar N, Raghava GPS. HAIRpred: Prediction of human antibody interacting residues in an antigen from its primary structure. Protein Science. 2025; 34(8):e70212. https://doi.org/10.1002/pro.70212