An Improved Method for predicting Conformational B-Cell Epitopes in a Protein sequence from its Primary Structure

by Anupma Pandey

Under the Supervision of Dr. G.P.S. Raghava

Indraprastha Institute of Information Technology Delhi
June, 2025

An Improved Method for predicting Conformational B-Cell Epitopes in a Protein sequence from its Primary Structure

by Anupma Pandey

Submitted in partial fulfillment of the requirements for the degree of Master of Technology

to

Indraprastha Institute of Information Technology Delhi

June, 2025

Certificate

This is to certify that the thesis titled "**An Improved Method for predicting Conformational B-Cell Epitopes in a Protein sequence from its Primary Structure**" being submitted by Anupma Pandey to the Indraprastha Institute of Information Technology Delhi for the award of the Master of Technology is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

June, 2025

G.P.S. Raghava

Department of Computational Biology

Indraprastha Institute of Information Technology Delhi

New Delhi 110020

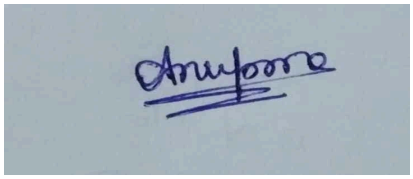
Acknowledgements

First and foremost, I want to express my gratitude to Dr. G. P. S. Raghava, Head of the Computational Biology Department at the Indraprastha Institute of Information Technology (IIIT) in New Delhi, for his support, encouragement, inspiration, and helpful suggestions. I want to thank him from the bottom of my heart for allowing me to work with him to finish my thesis, which was named "**An Improved Method for predicting Conformational B-Cell Epitopes in a Protein sequence from its Primary Structure**".

I want to extend my deep appreciation to him for all of her help with the project, including advice, encouragement, insightful remarks, and vast expertise.

I also want to express my gratitude to the non-teaching staff members for their wonderful assistance in whatever manner they could.

Last but not the least, this work would have never been a success without the constant inspiration of my parents, family members and my friends.



Anupma Pandey

Roll No.: MT23019

Abstract

The original CBTOPE method, introduced in 2009, was one of the first approaches for predicting conformational B-cell epitopes using only protein sequence information. While it has been widely adopted in the scientific community, recent benchmarking by Cia et al. (2023) revealed that many existing tools, including CBTOPE, exhibit suboptimal performance on rigorously curated and expanded datasets. This performance gap is largely attributed to limitations in earlier training data and feature representation. In response, we present CBTOPE2, an improved version of the original model, developed using modern machine learning techniques and trained on the high-quality dataset curated by Cia et al. (2023). The enhanced pipeline integrates multiple types of features, including binary profile, position-specific scoring matrices (PSSM), and relative solvent accessibility (RSA), to better capture evolutionary and structural signals relevant to epitope recognition. Initial models using binary features achieved a maximum AUC of 0.61 on the validation set. The addition of evolutionary features via PSSM improved performance to 0.67, while combining PSSM with RSA further boosted the AUC to 0.70. Multiple classifiers were evaluated, with Gradient Boosting trained on PSSM+RSA features yielding the best results: AUC = 0.70 and MCC = 0.28 on an independent validation set. To support practical application, CBTOPE2 is available as both a web server and a standalone Python package. The web platform allows users to submit antigen sequences in FASTA format and receive predictions of antibody-binding residues, while the standalone version (pip install cbtope2) supports offline analysis. The CBTOPE2 platform (https://webs.iiitd.edu.in/raghava/cbtope2) provides a significantly improved and accessible tool for conformational B-cell epitope prediction, designed to aid researchers in vaccine development, antibody engineering, and immunodiagnostics.

List of Figures

List of Tables

Contents

Chapter 1: INTRODUCTION

1.1 Background

Proteins are fundamental components of life, involved in nearly every biological process. Among their many roles, proteins serve as key elements of the immune response, where they function as antigens capable of triggering the production of antibodies. The specific regions of antigens that are recognized and bound by antibodies are known as B-cell epitopes, which are critical for the development of vaccines, immunotherapies, and diagnostic tools [1,2]. These epitopes are classified into two types: linear (continuous) and conformational (discontinuous). While linear epitopes consist of consecutive amino acid residues in the primary sequence, conformational epitopes are formed by amino acids that are brought into proximity in the 3D structure of the protein, despite being distant in the linear sequence [3].

The correct identification of B-cell epitopes is a building block in the development of successful subunit vaccines and antibody-based drugs [4]. Experimental techniques to map conformational B-cell epitopes, e.g., X-ray crystallography and NMR spectroscopy, are time-consuming, costly, and not always practicable [5]. Therefore, computational methods have become necessary tools for B-cell epitope prediction, especially the conformational ones that comprise about 90% of all B-cell epitopes[6] .

To overcome this challenge, various computational tools have been proposed during the last two decades. These approaches commonly make use of features extracted from amino acid sequences, evolutionary profiles, structural annotations, and physicochemical properties, and perform predictions using machine learning algorithms. Even with these endeavors, predicting conformational epitopes is still a challenging problem, primarily because of the inherent complexity of the underlying biological process and the lack of high-quality, experimentally verified datasets.

One of the initial software contributions to this area was CBTOPE, which our group created in 2009. CBTOPE was one of the first computer programs to predict conformational B-cell epitopes directly from a protein's amino acid sequence, without knowledge of its 3D structure. CBTOPE received significant visibility and has been extensively applied by the scientific community.

Nevertheless, the tool was only trained on small and comparatively aged datasets, which impacted its generalizability and performance on newer benchmark datasets [7].

Latest tests, for example, of Cai et al. (2023) [8], have shown that most currently available tools—like CBTOPE—fail when applied to new, well-annotated datasets. These results highlight the necessity of newer models trained on well-annotated data and using more informative features. Additionally, developments in machine learning, combined with availability of richer biological annotations and sequence-derived features, offer new prospects for enhancing predictive accuracy.

Therefore, this thesis seeks to fill a significant conformational B-cell epitope prediction gap by introducing a stable, new, and user-friendly tool—CBTOPE2—supported by rigorous validation and based on cutting-edge features and algorithms.

1.2 Motivation of Work

Infectious diseases and autoimmune disorders continue to pose a major challenge to global public health. A critical component in combating these diseases is the ability of the immune system to recognize and neutralize foreign pathogens. Among the most promising approaches in modern immunotherapy and vaccine development is the use of epitope-based subunit vaccines, which rely on identifying precise antigenic determinants—known as B-cell epitopes—capable of eliciting a targeted antibody response .

Within this domain, conformational B-cell epitopes represent the majority of antibody-recognized regions. Unlike linear epitopes that consist of continuous amino acid stretches, conformational epitopes are formed by residues that are spatially close in the 3D structure but scattered along the primary sequence. Predicting these epitopes remains a significant challenge, particularly in the absence of structural data, which is often unavailable for many antigens of interest.

Although early efforts—including our own previously developed tool, CBTOPE—have laid the foundation for sequence-based prediction of conformational B-cell epitopes, recent evaluations

reveal that their performance is often limited. These tools were developed using outdated and relatively small datasets, and have not benefited from advances in computational biology or from newly curated high-quality experimental datasets.

The motivation for developing CBTOPE2 stems from these limitations. The need for a robust, accurate, and updated sequence-based tool for predicting conformational B-cell epitopes has become increasingly urgent with the surge in emerging infectious diseases, vaccine-resistant pathogens, and the rising demand for personalized immunotherapies. In particular:

- Experimental methods for identifying epitopes are expensive, time-consuming, and impractical for high-throughput analysis.
- Most existing computational tools are structure-dependent, excluding a vast number of potential antigens whose 3D structures are unknown.
- There has been a lack of benchmarking on recently curated, unbiased datasets, leading to overestimation of model performance in real-world scenarios.
- Existing models often neglect the integration of multiple discriminative features such as evolutionary conservation and solvent accessibility, which have been shown to be key indicators of epitope regions.

To address these gaps, our work on CBTOPE2 seeks to build a next-generation epitope predictor that leverages advanced machine learning algorithms and integrates diverse biological features. By combining rigorous validation strategies with interpretability and usability, CBTOPE2 aims to empower researchers in epitope discovery, vaccine development, and immunodiagnostics—particularly in resource-limited settings where experimental facilities are inaccessible. In essence, the development of CBTOPE2 is motivated by the pressing need for accessible, accurate, and interpretable tools to accelerate epitope discovery in silico, enabling the design of safer and more effective vaccines and antibody therapies.

1.3 Objective

The primary objective of this research was to develop a robust and updated machine learning-based model for the prediction of conformational B-cell epitopes directly from protein sequences. Unlike linear epitopes, conformational epitopes are non-contiguous and require structural context for accurate identification. However, due to the scarcity of high-resolution structural data, there is an urgent need for sequence-based prediction tools that do not rely on 3D structure information.

By utilizing a newly curated and experimentally validated dataset from Cai et al. (2023), this work aims to overcome the limitations of previous models—including our own earlier tool, CBTOPE—which were trained on outdated and limited datasets [8]. This study leverages modern feature representations and state-of-the-art machine learning techniques to enhance predictive performance and reliability.

1.3.1 Overall Objective

The main goal of this project was to design a strong computational model for predicting conformational B-cell epitopes with recent, high-quality data and biologically informative features. The goal was to improve the predictive accuracy of the original CBTOPE method and facilitate the scientific community in epitope-based vaccine design. To facilitate practical utility, the top-performing model was released as a facile webserver and also as a standalone software program, so it can be used for wider applications in immunology research and drug development.

1.4 Scope

Prediction of conformational B-cell epitopes is important for the design of successful subunit vaccines and immunotherapies [7,8]. Due to their intricate structural nature, their identification based on sequence information alone is not easy. The goal of this work is to facilitate the prediction of such epitopes by combining various types of biological features, such as binary profiles, evolutionary data (PSSM), and relative solvent accessibility (RSA), in machine learning algorithms trained on high-quality, well-curated data.

In addition to model development, this research involves deploying the top-performing model via an intuitive webserver and standalone program. The software is intended to enable researchers to make predictions for possible epitopes and streamline vaccine design through offering a steady, sequence-based prediction platform.

Chapter 2: LITERATURE REVIEW

2.1 Introduction

Conformational B-cell epitopes (CBEs) are well-defined surface-exposed areas on antigens that are targeted by antibodies. Precise determination of these epitopes is crucial in vaccine development, antibody generation, and immunodiagnostics. The determination of CBEs has conventionally been based on tedious and time-consuming experimental techniques like X-ray crystallography and cryo-electron microscopy. Yet, with the introduction of computational biology, machine learning (ML)-based predictive models have emerged as a strong contender because they are fast, inexpensive, and capable of processing large data sets.

Various bioinformatics tools and computational models have been created to make predictions of CBEs over the years. These approaches can be placed in two broad categories: structure-based and sequence-based approaches. Structure-based methods like Epitopia, SEPPA, and Discotope depend on three-dimensional (3D) structural data of antigen-antibody complexes and are therefore applicable only to proteins with solved structures. Sequence-based tools like CBTOPE, BepiPred 2.0, and epitope3D provide a more generalized solution by predicting the epitopes directly from the amino acid sequence.

In spite of these developments, recent benchmarking by Cia et al. (2023) has shown that most of the current tools, such as CBTOPE, fail when applied to big, high-quality datasets of antibody-antigen complexes [8]. A major drawback is that older models were trained on small and old-fashioned datasets, which results in less generalizability. Also, most methods lack enough biological context, like solvent accessibility or evolutionary conservation, which are important for correct epitope prediction.

2.2 Advances in Feature Engineering

For enhancing predictive performance, several features have been investigated. Binary profile patterns encode the amino acid composition as a fixed-length vector, allowing machine learning algorithms to discover sequence patterns. PSSM (Position-Specific Scoring Matrix) profiles, in

turn, are derived from PSI-BLAST searches and incorporate evolutionary information by emphasizing conserved areas between homologous sequences. RSA (Relative Solvent Accessibility) offers structural information by telling us which residues are solvent exposed on the protein surface, a critical feature for antigen-antibody binding.

Recent research has shown that the combination of these attributes improves performance. For instance, combining PSSM and RSA improved the discrimination of epitope regions by incorporating both conservation and structural exposure into consideration. Gradient boosting techniques, for example, have proved to perform well in learning from high-dimensional feature spaces of arbitrary complexity.

2.3 Use of ML Models

Multiple machine learning classifiers have been tried on the problem, such as Logistic Regression, Random Forests, Naïve Bayes, LightGBM, and Gradient Boosting. Though they provide interpretability, simpler models like Logistic Regression and Naïve Bayes tend to sacrifice predictive power, while ensemble methods like Random Forest and Gradient Boosting tend to provide better predictive performance by retaining nonlinear relationships within the data.

PreTP-Stack, PEPred-Suite, and PPTPP are examples of ensemble-based systems that make use of more than one feature and classifier for better therapeutic peptide prediction. These models always work with synthetic or theoretical data. They hardly make use of real sequences from DrugBank or reported therapeutic proteins, hence constraining their applicability to actual drug discovery problems.

2.4 Gap in Existing Literature

One of the most frequent flaws in current models is their dependence on forecasted or hypothesized therapeutic pathways instead of experimentally confirmed druggable proteins. This makes them less effective in drug development pipelines where accurate, biologically confirmed predictions are essential. In addition, most such methods disregard essential features of real

antigens and antibodies like post-translational modifications and structural flexibility, which determine epitope accessibility.

2.5 Summary

To fill these gaps, our work introduces CBTOPE2, an enhanced sequence-based tool learned from high-quality antibody-antigen complexes filtered by Cia et al. (2023) [8]. By incorporating binary profiles, PSSM, and RSA, and employing powerful ML algorithms like Gradient Boosting, CBTOPE2 significantly outperforms previous tools in predictive performance. Additionally, its availability as a web server and free software renders it easily available to researchers, expediting epitope discovery and vaccine design.

Chapter 3: METHODOLOGY

3.1 Preparation of Dataset

In the research, we utilized a benchmark dataset that was compiled and published by Cia et al. (2023) and specifically developed for benchmarking conformational B-cell epitope prediction tools. The dataset consists of high-resolution, non-redundant structures of antibody-antigen complexes taken from the Protein Data Bank (PDB) [8]. 286 PDB structures were chosen to be included in the final benchmark dataset (final_benchmark_dataset.txt). These buildings were strictly filtered for quality and resolution for reliability and consistency in downstream predictive modeling.

Each antigen structure was associated with residue-level annotations indicating whether a residue was part of a conformational B-cell epitope or not. This labeling was derived from the interaction interface with corresponding antibody chains. The final label for each residue (epitope or non-epitope) appears in the last column of the CSV tables for each antigen.

From the full list of antigens, 268 structures were used to generate positive and negative sequence patterns for training and evaluation. The dataset was split into 80% training (214 structures) and 20% validation (54 structures) sets, ensuring that each structure was uniquely represented in only one subset. The training and validation partitions were used throughout the machine learning pipeline for training, testing, and final evaluation.

Table 1: List of PDB IDs used for the study

3ZKN_2	2IBZ_1	4OGY_2	3S35_1	5WK3_2	3U30_2	5D8J_1	6O3B_1
3SOB_1	6H3T_1	1V7M_1	5NUZ_1	4DKF_1	6B08_1	5E94_2	3RU8_1
4ZSO_2	4K94_1	6I8S_1	5VPL_1	5XAJ_3	4OII_2	6ID4_2	4AG4_1
1LK3_2	3L5W_1	4QWW_1	6IEB_1	2VH5_1	5LSP_2	2BDN_1	5IKC_2
3KLH_1	3U9P_1	1JRH_1	4MWF_2	6A78_1	3MXW_1	2VXQ_1	4JQI_1
5VLP_1	5DFV_1	3GRW_1	2Q8A_1	4R8W_1	3LH2_1	5B71_2	6IAP_2
1WEJ_1	4LU5_2	3L95_1	4NP4_2	1XIW_2	3LEV_1	6DKJ_2	3G04_1
4HLZ_1	2DD8_1	4DTG_1	4LMQ_1	5E8E_1	2AEQ_1	5GJS_1	4K2U_2
5B3J_1	6BGT_1	4AEI_3	5W5X_1	3HB3_1	2VXT_1	4U6V_2	5L0Q_1
4F3F_1	5O14_2	6CW2_1	3QA3_2	5TIH_1	4YWG_1	1ORS_1	4J4P_2
5D72_1	4I3S_1	5MEV_1	1OB1_2	4XP4_1	5VTA_2	3EJZ_1	4HT1_1
4XAK_1	2J6E_1	5H35_2	4KXZ_4	5TZ2_1	4DW2_1	6CMI_1	4D9Q_1
3O2D_1	6MUF_1	2ZCH_1	4OKV_1	5MVZ_2	1EGJ_1	4ETQ_1	2R56_1
3KS0_2	4OT1_1	4L5F_1	5CZX_2	5BO1_1	5O1R_2	5EBM_1	5X2Q_1
5BK1_2	3VG9_1	4QNP_1	3B9K_2	5W4L_1	4EDW_1	3NID_1	1GC1_1
5OB5_1	5X0T_2	4KI5_2	5YOY_2	5D93_1	6CXY_1	4CAD_4	6AOD_1
4RDQ_3	4HWB_1	4LIQ_1	3LD8_1	1DVF_2	2QQK_1	6OTC_1	1PG7_2
4KUC_2	5T6L_1	1UAC_1	6E62_1	4O9H_1	1OAK_1	5DHV_1	3MJ9_1
1NMB_1	6FXN_6	1NFD_2	3SKJ_1	3V6O_2	1AHW_1	4F37_1	3QWO_17
5A3I_2	5TRU_2	5VQM_2	4M5Z_1	4LVN_1	4JLR_2	3BN9_2	6CK9_1
3LIZ_1	6MLK_1	4IRZ_1	5KW9_1	6DDV_1	2QQN_1	6MI2_1	3R1G_1
4JRE_1	1KB5_1	3GI9_1	1FE8_1	2ARJ_1	5NGV_1	6AQ7_1	2JEL_1
4KHT_1	4CMH_1	6EWB_1	3EOA_2	4RGM_1	4YQX_2	3BT2_1	2XQY_2

4BZ2_1	4UTA_1	4JZJ_2	6IW2_6	3WIH_2	1BGX_1	2H9G_2	6AL5_1
3HMX_1	5LBS_2	4ZFG_1	3LHP_1	1DVF_1	2WUB_1	3WFC_1	4FQJ_1
4XWG_1	4XWO_1	3VI4_2	4QHU_1	5EU7_1	2NYY_1	2OZ4_1	6CBV_1
1ZTX_1	6O39_1	5OCC_1	6AL0_1	5FB8_1	2B2X_1	5UTZ_1	4HCR_2
1TZH_1	5K59_2	5HDQ_1	5GGT_1	2J88_1	2XRA_1	5F3B_1	5B8C_1
4PLJ_2	5XAJ_2	1RJL_1	4LF3_1	5MHR_5	4Y5Y_2	5TL5_1	6MEI_1
1NSN_1	1FJ1_2	5JQ6_1	5Y11_1	5IES_1	1FSK_2	6CYF_1	5LDN_1
5W2B_1	4Z5R_3	6C9U_1	3KR3_1	1IAI_2	1E6J_1	6J5D_1	4QCI_2
4RRP_2	1H0D_1	5TH9_3	4LEO_1	2XQB_1	6APB_1	4I9W_2	5WI9_2
4UU9_2	1OAZ_2	1PKQ_1	4K3J_1	5K9K_1	1YJD_1	5WHK_1	6BPA_1
2YC1_1	5VEB_1	5HBV_1	5XWD_1				

3.1.1 Pattern Generation

To capture contextual information relevant to antigen-antibody interactions, we generated overlapping residue patterns of varying window lengths (7 to 31 amino acids) [9]. Each pattern was centered on a specific residue and assigned a binary label based on whether the central residue was antibody-interacting (positive) or not (negative). Terminal residues were padded with a dummy token to maintain uniform window sizes across the dataset. This approach aligns with previous studies in the domain and ensures that both local and flanking residue information contribute to model training.

3.1.2 Dataset Statistics and Balancing

As expected, non-interacting residues greatly outnumbered the interacting ones, introducing class imbalance. To mitigate this, we employed random undersampling of the negative class to match the number of positive samples. The final balanced training dataset included 3,978 positive and 3,978 negative patterns, while the validation dataset comprised 1,069 positive and 1,069 negative patterns, ensuring a 1:1 ratio of classes for supervised learning tasks.

This dataset was used to generate multiple feature representations—Binary Profile, PSSM (Position-Specific Scoring Matrix), and RSA (Relative Solvent Accessibility)—to train and benchmark various machine learning models. The curated PDB structures, pattern annotations, and supplementary data files (such as final_benchmark_dataset.txt and Antigen_zip) are publicly available via the GitHub repository maintained by Cia et al. and referenced in this study - https://github.com/3BioCompBio/BCellEpitope [8].

3.2 Feature Generation

To train and evaluate machine learning models for conformational B-cell epitope prediction, we employed a feature generation strategy that captures diverse biochemical and structural characteristics of protein residues. Specifically, we utilized three types of residue-level descriptors: Binary Profile, Position-Specific Scoring Matrix (PSSM), and Relative Solvent Accessibility (RSA). All features were generated using tools from the Pfeature suite, which offers a comprehensive platform for computing structural and compositional features of protein sequences.

3.2.1 Binary Profile

Binary profile features represent each amino acid in a given sequence window using a 21-dimensional one-hot encoded vector. This vector captures the identity of the amino acid, where one dimension is set to 1 (representing the current amino acid), and all others are set to 0. An additional dummy amino acid ('X') was included to represent padding residues in shorter sequences. For a given window size W, the resulting binary profile is a 21×W matrix, which is flattened into a 1×(21×W) vector for model input.

This representation retains the sequential order and composition of residues, enabling machine learning models to learn patterns based solely on amino acid identity. Binary profiles were generated using Pfeature's BinaryProfile module [10], which has been widely used in protein function and epitope prediction tasks [2,11].

3.2.2 Position-Specific Scoring Matrix (PSSM)

PSSM features encode evolutionary conservation information by capturing the likelihood of each amino acid substitution at every residue position in a protein sequence. PSSMs were generated by running three iterations of PSI-BLAST against the Swiss-Prot database. The resulting 20-dimensional PSSM vectors for each residue were then scaled using Pfeature's min-max normalization to bring the values into a consistent range[12–14] .

For each pattern of length W, the PSSM is represented as a 20×W matrix, which is flattened into a 1×(20×W) vector for downstream processing. This feature set is known to significantly enhance classification performance by providing insights into residue conservation across curated protein sequences [10,15].

3.2.3 Relative Solvent Accessibility (RSA)

RSA measures the solvent exposure of amino acids in a protein's tertiary structure. For the purpose of this study, SPOT-1D, a state-of-the-art deep learning method was employed to predict RSA values for estimating the structural properties directly from protein sequences [5],[16]. SPOT-1D first encodes the input FASTA sequence with a pre-trained bidirectional LSTM neural network, capturing both local and global contextual dependencies across the sequence. The model is trained on large datasets of proteins with known 3D structures and learns to predict all sorts of residue-level structural attributes including Relative Solvent Accessibility based on patterns in amino acid composition and evolutionary signals.

To calculate RSA, SPOT-1D assigns each amino acid in the input sequence to a scalar value representing relative accessibility to solvent, normalized against a reference maximum exposure value. We extracted the predicted RSA value for each residue in a window pattern but kept only the RSA of the center residue, as it best represents the local context of the pattern. We rounded this scalar value to two decimal places and appended this as a feature to the corresponding pattern.

RSA can act as a proxy for surface exposure, distinguishing buried from accessible residues to antibodies. RSA is an important structural descriptor in epitope prediction, since epitope residues

are solvent-exposed more frequently [17,18]. Incorporating SPOT-1D's RSA predictions enabled us to introduce structural context into our sequence-based model in a useful manner.

3.3 Exploratory Data Analysis

We performed exploratory data analysis (EDA) to understand the underlying distribution and characteristics of the dataset before model development. This step provided critical insights into class imbalance, residue-level patterns, and the biological relevance of the extracted features.

3.3.1 Class Imbalance

As illustrated in Figure 1, the dataset exhibits a significant class imbalance, with a vast majority of residues labeled as non-epitopic (class 0) and only a small fraction labeled as epitopic (class 1). Out of 66,107 total residue-centered patterns, 61,036 (92.3%) belong to the non-epitope class, while only 5,071 (7.7%) correspond to epitopic residues. This distribution reflects a biological reality — only a limited number of surface residues in an antigen are actually involved in antibody binding, while the majority remain non-interacting.

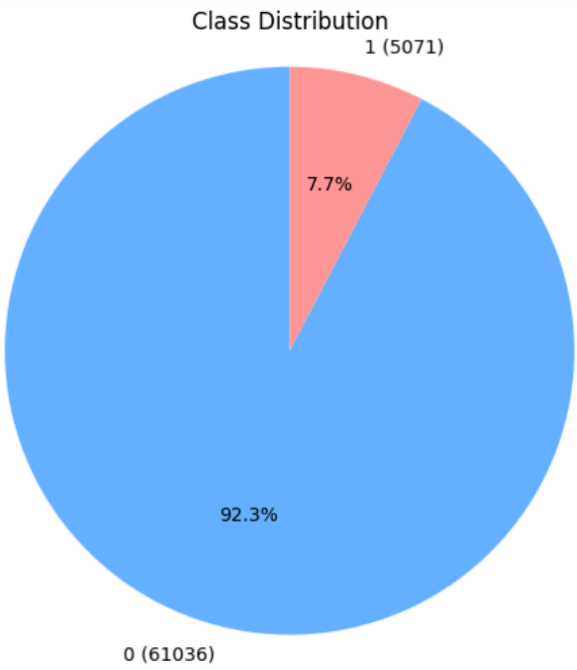


Figure 1: Distribution of Epitopic Vs Non-Epitopic classes in the Dataset

To construct the dataset, we generated overlapping patterns of each antigen, where each pattern is centered around a residue. A pattern was labeled positive if its central residue was found to interact with an antibody; otherwise, it was labeled negative. This approach ensured that all residue positions were contextually represented through sequence windows.

In the training set, we initially obtained 3,982 positive and 49,508 negative patterns. Likewise, the validation set contained 1,089 positive and 11,528 negative patterns. To address this highly skewed distribution, we applied random undersampling on the majority (negative) class. This resulted in balanced sets containing 3,982 positive and 3,982 negative patterns for training, and 1,089 positive and 1,089 negative patterns for validation.

This class balancing strategy was crucial to prevent machine learning models from being biased toward predicting only the majority class. It also ensured that performance metrics such as AUROC and MCC accurately reflected model effectiveness in distinguishing epitope residues from non-epitope residues.

3.3.2 Central Amino Acid Distribution in Epitopic Regions

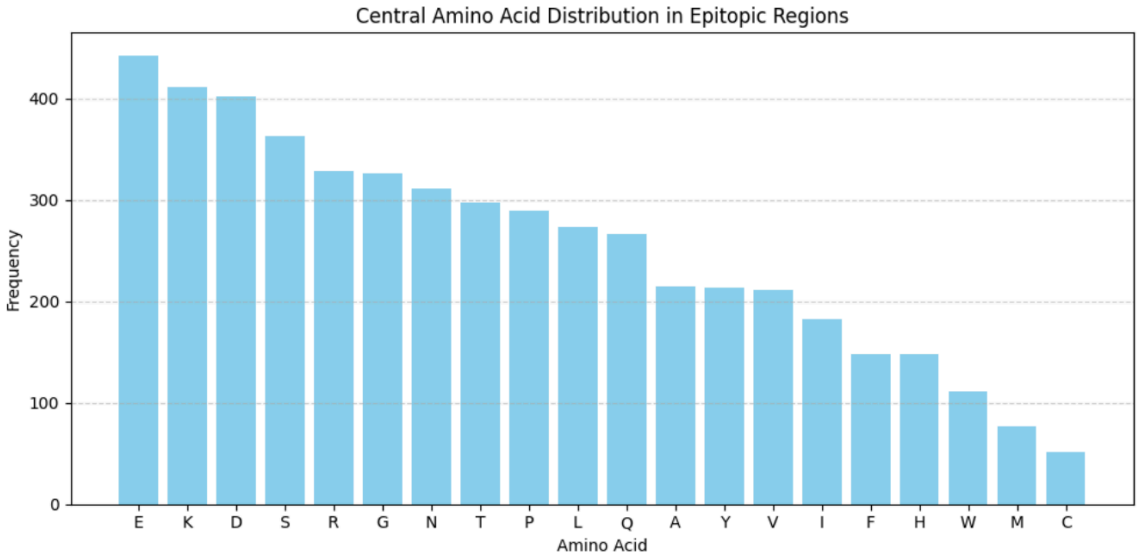


Figure 2: Frequency of Amino Acids in Epitopic Regions

Figure 2 presents the frequency distribution of amino acids that occur at the central position of patterns labeled as epitopic residues (i.e., Interacting = 1). Since each residue-centered pattern is

constructed using a fixed window size (W = 17), the amino acid at the center is considered most indicative of potential epitope characteristics.

From the bar plot, we observe that certain charged and polar residues such as Glutamic acid (E), Lysine (K), Aspartic acid (D), Serine (S), and Arginine (R) are highly enriched in epitopic positions. These residues are known for their hydrophilic nature, often being exposed on the protein surface, which is a desirable trait for epitope formation due to their accessibility to antibodies.

Conversely, hydrophobic residues like Tryptophan (W), Methionine (M), and Cysteine (C) are underrepresented at these positions. These residues are generally buried within the protein core and thus are less likely to participate in antigen-antibody interactions.

This distribution further supports the biological relevance of the dataset and highlights sequence-level preferences in epitope composition, which can be useful for designing more informed prediction models.

3.3.3 RSA Distribution for Epitopic vs Non-Epitopic Residues

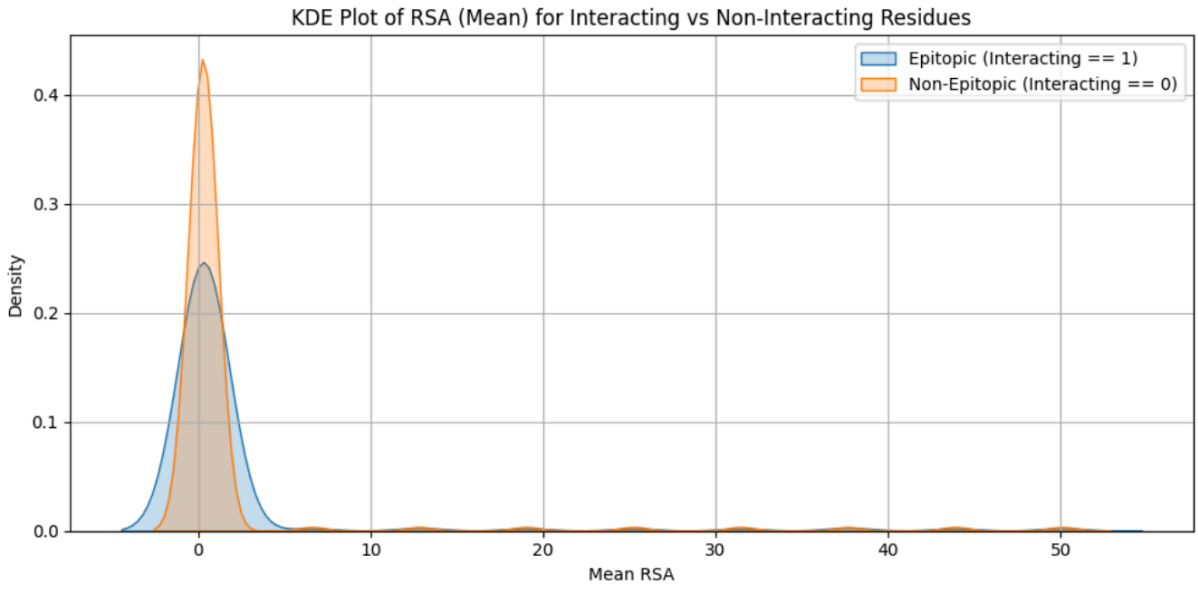


Figure 3: Comparison of RSA values of Epitopic & Non-Epitopic Residues

Figure 3 depicts the Kernel Density Estimate (KDE) plot comparing the distribution of Relative Solvent Accessibility (RSA) values for epitopic (Interacting = 1) and non-epitopic (Interacting = 0) residues. For each pattern, the mean RSA across the window was calculated, and distributions were plotted separately for the two classes.

As shown in the figure, epitopic residues (blue curve) tend to have slightly higher RSA values than non-epitopic residues (orange curve). This aligns with the biological understanding that epitopes are more likely to be located on solvent-exposed regions of the antigen, making them accessible for antibody binding. In contrast, non-epitopic residues are often more buried within the protein structure, leading to lower solvent accessibility.

Although the distributions overlap significantly, the shift in the density curve of the epitopic residues indicates a statistical trend toward higher exposure, suggesting that RSA is a discriminative and biologically meaningful feature for epitope prediction.

This observation supports the inclusion of RSA in our feature set for model development, as it captures valuable structural information that can enhance the prediction of conformational B-cell epitopes.

3.3.4 Mean PSSM Score Distribution in Epitopic vs. Non-Epitopic Residues

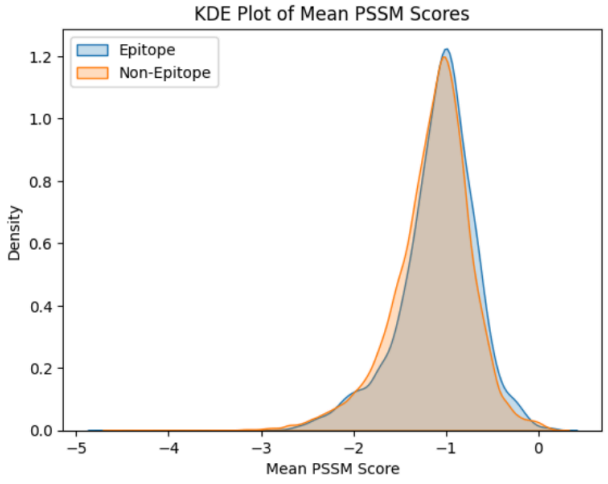


Figure 4: Comparison of Mean PSSM Scores for Epitopic & Non-Epitopic regions

To investigate the role of evolutionary conservation in epitope prediction, we plotted a Kernel Density Estimate (KDE) of the mean PSSM scores for residues labeled as epitopic (interacting) and non-epitopic (non-interacting). As shown in Figure 4, the distribution of average PSSM values is slightly shifted between the two classes.

The KDE plot illustrates that epitopic residues (blue curve) tend to have marginally higher average PSSM scores than non-epitopic residues (orange curve). This suggests that epitopic residues may be more evolutionarily conserved, possibly due to their functional importance in antibody recognition. Although the overall difference is subtle, the trend aligns with the biological expectation that critical regions, such as antibody-binding sites, are subject to evolutionary pressure.

This observation reinforces the utility of PSSM features in distinguishing epitopic regions, as conservation patterns captured through multiple sequence alignments reflect underlying functional constraints relevant to B-cell epitope localization.

3.3.5 PSSM Heatmap Analysis of Epitopic vs. Non-Epitopic Regions

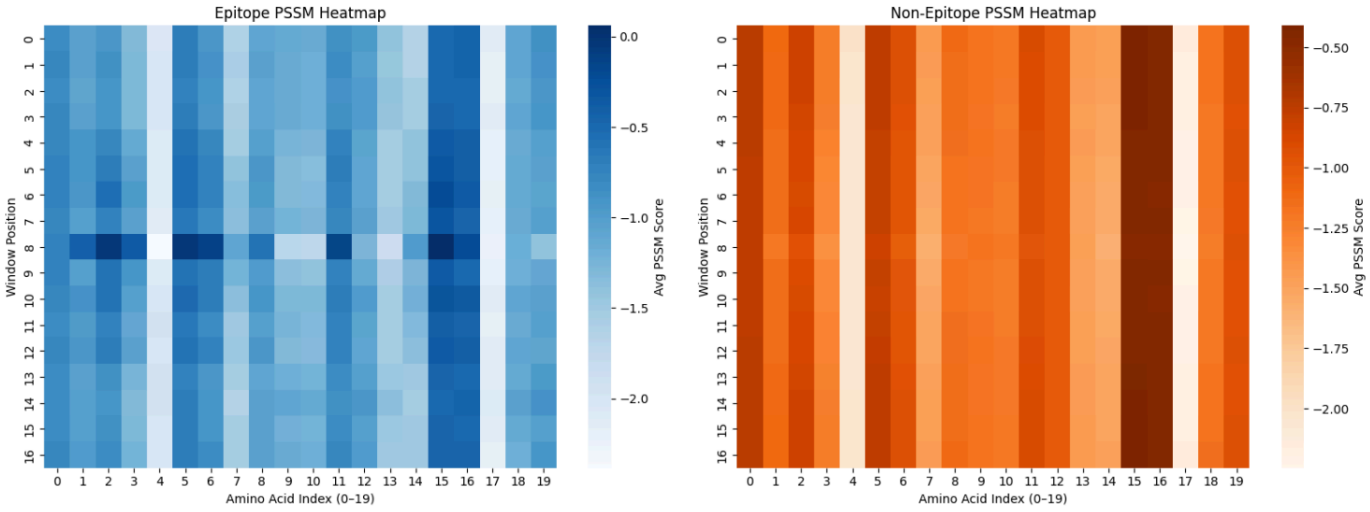


Figure 5: Heatmaps showing average PSSM scores across window positions for epitope (left) and non-epitope (right) residues, highlighting greater conservation in epitope regions.

To further understand the differences in conservation patterns between epitopic and non-epitopic regions, we visualized the Position-Specific Scoring Matrix (PSSM) data using heatmaps. Each heatmap in Figure 5 shows the average PSSM scores across all samples, with rows representing positions in the sliding window (length 17) and columns representing amino acid indices (0–19).

The left heatmap corresponds to epitopic regions, while the right heatmap represents non-epitopic regions.

Key observations:

- In the epitopic heatmap, there is more variation in PSSM scores across positions and amino acids, especially around the central window positions, indicating stronger conservation signals.
- In contrast, the non-epitopic regions show a more uniform pattern, with lower average PSSM values and fewer position-specific peaks.
- Certain amino acid indices exhibit consistently higher scores in epitopic regions, potentially indicating preferred or conserved residues involved in antibody binding.

This analysis supports the hypothesis that epitopic regions are more evolutionarily conserved, particularly around the central residues of the sliding window. The distinctive conservation patterns captured by the PSSM in these regions make them valuable features for distinguishing B-cell epitopes from non-epitopic residues.

3.4 Machine Learning Classifiers

To assess the performance of various feature sets in predicting B-cell epitopes, several machine learning models were applied to the refined dataset, each one chosen for its specific strengths. Random Forest (RF) [19] was also included as an ensemble learner, using multiple decision trees to improve predictive power and avoid overfitting. Gradient Boosting (GB) was also utilized for its iterative methodology in error correction done by earlier models to allow for better accuracy when dealing with complex data sets [19]. To balance model complexity, Naive Bayes (NB) was added for its speed and efficiency, especially to deal with large data

sets, even with the assumption of independence of features [20]. Besides, LightGBM (LGBM) was also used for its effectiveness in processing high-dimensional feature large datasets with high predictive accuracy using its one-side sampling and exclusive feature bundling methods based on gradients. Lastly, Logistic Regression (LR) was used for its interpretability as well as its good performance in binary classification problems, particularly where data is linearly separable [21]. Each model's unique strengths enabled a thorough test of the dataset, determining the impact of different sets of features on predictive accuracy.

3.4.1 Cross-Validation and Performance Metrics

In this study, the dataset was divided into training and validation subsets following the standard protocol used in previous studies. Specifically, we used 214 antigens for model training and internal evaluation, and a separate independent validation set of 54 antigens was reserved for final testing. The training dataset was further subjected to five-fold cross-validation to ensure reliable and unbiased evaluation of machine learning models[22] . In this technique, the data is partitioned into five equal folds, where in each iteration, four folds are used for training and the remaining one for testing. This cycle is repeated five times, such that every fold is used once for evaluation. The average performance across all folds is then computed.

This methodology allows models to be evaluated thoroughly while minimizing the bias introduced by a particular data split. Importantly, hyperparameter optimization for each classifier was conducted during cross-validation by monitoring performance on the test folds [23]. After finalizing model parameters, performance was assessed on the independent validation dataset comprising 54 antigens, which was not involved in training or internal testing.

Model performance was measured using both threshold-independent and threshold-dependent evaluation metrics. Area Under the Receiver Operating Characteristic Curve (AUROC) was used as the main threshold-independent metric to evaluate the model’s ability to distinguish between epitope and non-epitope residues without bias to a specific decision threshold [24].

In addition, threshold-dependent metrics such as sensitivity (recall), specificity, accuracy, precision, F1-score, and Matthews Correlation Coefficient (MCC) were calculated to provide a comprehensive performance overview. Among them, MCC is particularly valuable in

imbalanced datasets, as it takes into account true and false positives and negatives to provide a balanced assessment [25,26]. The average metrics from five-fold cross-validation, along with the results on the independent validation set, are reported in the Results section.

This strategy ensures that the evaluation remains rigorous, consistent, and in line with established practices in protein epitope prediction research[3,27] .

$$MCC = \frac{(T_P * T_N) - (F_P * F_N)}{\sqrt{(T_P + F_P)(T_P + F_N)(T_N + F_P)(T_N + F_N)}}$$

$$Specificity = \frac{T_N}{T_N + F_P}$$

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}$$

$$Sensitivity = \frac{T_P}{T_P + F_N}$$

$$F1 - Score = \frac{2T_P}{2T_P + F_P + F_N}$$

$$Precision = \frac{T_P}{T_P + F_P}$$

Where T_P, T_N, F_P, and F_N stand for true positive, true negative, false positive, and false negative, respectively.

3.4.2 Window Selection

In the context of epitope prediction, window size refers to the number of consecutive amino acid residues considered when forming feature patterns around a central residue. This central residue is the one ultimately labeled as epitopic or non-epitopic based on structural data. Selecting an optimal window size is critical, as it determines the amount of local sequence and structural context captured for each residue, which can directly impact model performance.

In this study, we experimented with sliding window sizes ranging from 7 to 31, increasing by increments of 2 (i.e., 7, 9, 11, ..., 31). This range was chosen to balance computational efficiency with biological relevance. Smaller window sizes may miss important contextual signals, whereas overly large windows can introduce noise and redundancy.

For each window size, overlapping patterns were generated across the entire protein sequence, ensuring that each residue appeared as the central (target) residue in one pattern. These patterns were then encoded using binary profile, PSSM, and RSA features, either individually or in combination.

Each machine learning model was trained and evaluated separately for every window size using 5-fold cross-validation on the training dataset and independently validated on a separate test set. The performance was measured using AUROC, MCC, and other evaluation metrics. Based on the results, window sizes in the range of 17 to 23 consistently offered the best trade-off between performance and feature resolution, particularly for models trained using PSSM and RSA combinations.

The findings reinforce that capturing a sufficient local context is crucial for accurately modeling the spatial and physicochemical cues associated with antibody binding regions. Thus, window size optimization played a key role in fine-tuning the model performance.

3.5 Complete Workflow

Figure 6 presents the overall workflow adopted for developing the CBTOPE 2 prediction models. The process begins with the acquisition of 268 antigen–antibody complex structures

from the Protein Data Bank, as reported in Cia et al. For each antigen, overlapping residue-centered patterns were generated using a sliding window of sizes ranging from 7 to 31. To handle the inherent class imbalance—since only a small fraction of residues are antibody-interacting—each protein's negative patterns were randomly downsampled to match the number of positive patterns.

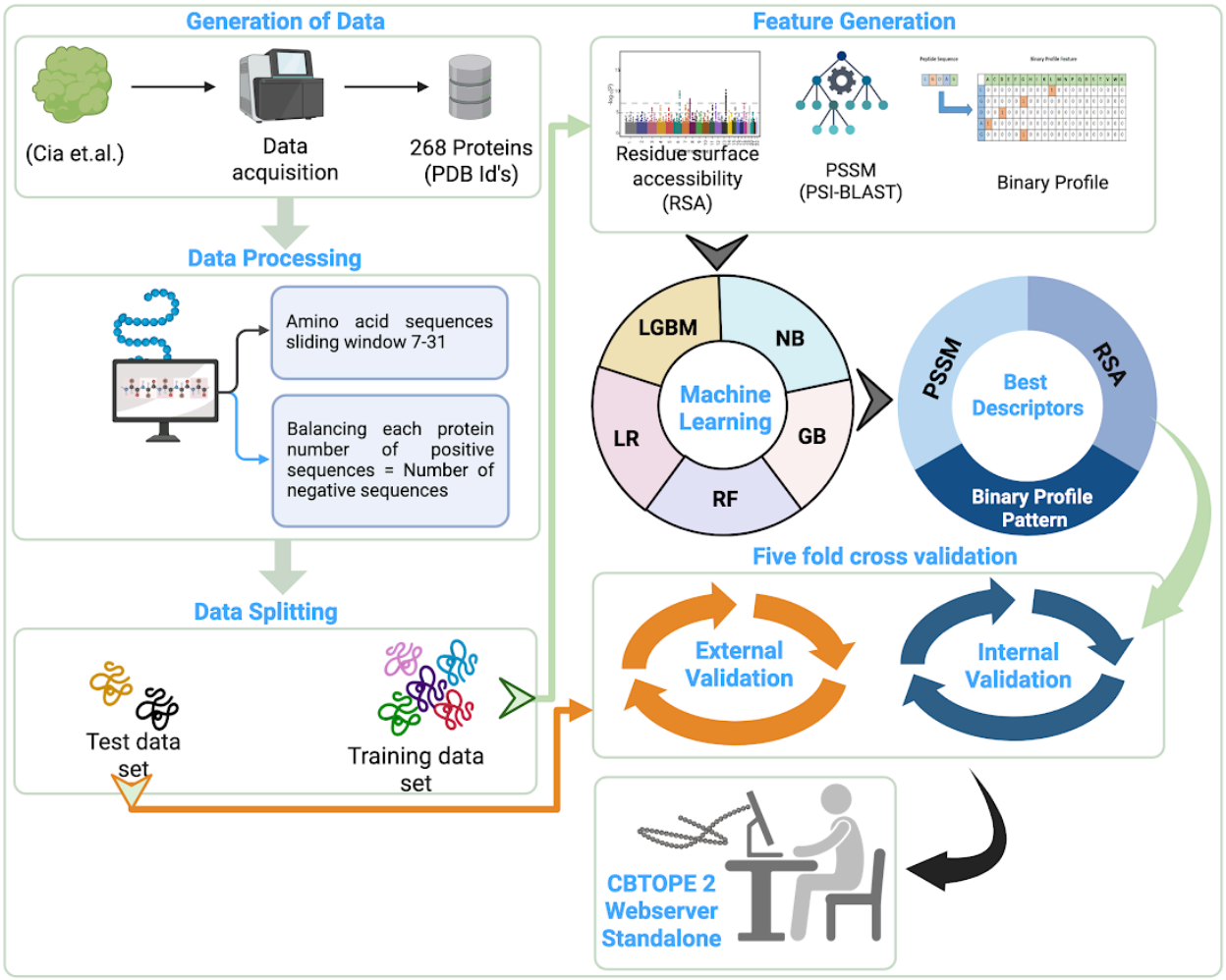


Figure 6: The complete workflow of the study

Next, multiple features were generated for each pattern: Relative Surface Accessibility (RSA) from SPOT-1D, evolutionary conservation through PSSM from PSI-BLAST (Swiss-Prot

database), and amino acid identity via binary profile encoding. The data was split into training (214 antigens) and external validation (54 antigens) sets. The training data was used in five-fold cross-validation to build and optimize machine learning models. Various classifiers were evaluated, including Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), Naive Bayes (NB), and Decision Tree (DT), using different feature combinations. The best-performing models were then validated on the independent test set and integrated into a standalone tool, CBTOPE 2, for epitope prediction.

Chapter 4: RESULTS

4.1 Binary pattern profiles

We developed machine learning models using binary profile patterns (BPP), where each protein/peptide sequence is represented as a 21×W matrix, with W denoting the window size. Models were trained and evaluated across multiple window sizes ranging from 7 to 31. Among the classifiers tested, the LightGBM (LGBM) model achieved the best overall performance at a pattern length of 17, with the highest AUROC of 0.58 and MCC of 0.14 on the validation dataset.

Other classifiers showed comparable but slightly lower performance. The Naive Bayes (NB) and Gradient Boost (GB) models both achieved an AUROC of 0.57 and an MCC of 0.11. The Logistic Regression (LR) model also reached an AUROC of 0.57, but with a lower MCC of 0.10, while the Random Forest (RF) model showed the least performance with an AUROC of 0.56 and MCC of 0.08.

Table 2: The ML performance results on window 17 using Binary profile patterns

Model	**Training dataset**					**Validation dataset**				
	Sens	**Spec**	**Acc**	**AUROC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUROC**	**MCC**
LR	0.58	0.57	0.58	0.60	0.15	0.54	0.56	0.55	0.57	0.10
RF	0.60	0.54	0.57	0.59	0.13	0.56	0.52	0.54	0.56	0.08
NB	0.58	0.57	0.57	0.60	0.15	0.56	0.55	0.56	0.57	0.11
LGBM	0.59	0.57	0.58	0.60	0.16	0.58	0.56	0.57	0.58	0.14
GB	0.63	0.53	0.58	0.60	0.16	0.60	0.51	0.56	0.57	0.11

LR, Logistic Regression; RF, Random Forest; NB, Naive Bayes; LGBM, Light Gradient Boosting Machine; GB, Gradient Boost; Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; AUROC, Area Under the Receiver Operating Characteristic; MCC, Matthews Correlation Coefficient

4.2 Performance on PSSM profiles

In addition to the binary profile, we also used the PSSM (Position-Specific Scoring Matrix) profile for developing prediction models. PSSM profiles, generated using PSI-BLAST, capture evolutionary conservation, making them highly effective for biological sequence classification tasks. A variety of machine learning classifiers were evaluated using PSSM-based features.

Table 3: The ML performance results on window 17 using PSSM matrix

Model	**Training dataset**					**Validation dataset**				
	Sens	**Spec**	**Acc**	**AUROC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUROC**	**MCC**
LR	0.65	0.56	0.61	0.65	0.21	0.64	0.55	0.59	0.63	0.18
RF	0.68	0.56	0.62	0.67	0.24	0.63	0.53	0.58	0.63	0.16
NB	0.71	0.51	0.61	0.64	0.22	0.67	0.53	0.60	0.62	0.20
LGBM	0.65	0.58	0.61	0.66	0.23	0.59	0.54	0.56	0.60	0.12
GB	0.69	0.55	0.62	0.66	0.24	0.65	0.52	0.58	0.62	0.17

LR, Logistic Regression; RF, Random Forest; NB, Naive Bayes; LGBM, Light Gradient Boosting Machine; GB, Gradient Boost; Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; AUROC, Area Under the Receiver Operating Characteristic; MCC, Matthews Correlation Coefficient

Among the tested models, Logistic Regression (LR) achieved the highest AUROC of 0.63 on the validation dataset, along with an MCC of 0.18, indicating it as the best-performing model overall. The Naive Bayes (NB) classifier also performed competitively with an AUROC of 0.62 and MCC of 0.20. Random Forest (RF) and Gradient Boost (GB) followed closely, both reaching AUROC scores of 0.63 and 0.62, with MCC values of 0.16 and 0.17, respectively. The LightGBM (LGBM) model showed slightly lower performance, achieving an AUROC of 0.60 and MCC of 0.12.

These results highlight the predictive strength of PSSM features alone, with Logistic Regression demonstrating the best balance of sensitivity, specificity, and overall discriminative power. They also suggest that further improvements may be possible by

incorporating additional structural features, such as RSA, to complement evolutionary information.

4.3 Performance on PSSM with BPP

We also developed models using a combination of PSSM and binary profile features. As shown in Table 3, integrating these two feature types results in improved performance compared to using either profile individually. Among the classifiers evaluated, the Random Forest (RF) model achieved the best performance on the validation dataset, with the highest AUROC of 0.63 and MCC of 0.18, indicating strong discriminative capability.

Other models also delivered competitive results. The Gradient Boost (GB) classifier reached an AUROC of 0.62 and MCC of 0.17, followed by LightGBM (LGBM) with an AUROC of 0.62 and MCC of 0.16. Naive Bayes (NB) and Logistic Regression (LR) obtained AUROC values of 0.60 and 0.59, and MCC scores of 0.15 and 0.11, respectively.

These findings suggest that combining evolutionary (PSSM) and compositional (binary profile) features enhances the predictive power for conformational B-cell epitope identification.

Table 4: The ML performance results on window 17 using the PSSM matrix with BPP

Model	**Training dataset**					**Validation dataset**				
	Sens	**Spec**	**Acc**	**AUROC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUROC**	**MCC**
LR	0.60	0.58	0.59	0.63	0.18	0.58	0.53	0.56	0.59	0.11
RF	0.67	0.55	0.61	0.66	0.22	0.64	0.54	0.59	0.63	0.18
NB	0.65	0.54	0.60	0.63	0.19	0.60	0.55	0.58	0.60	0.15
LGBM	0.65	0.57	0.61	0.65	0.21	0.61	0.56	0.58	0.62	0.16
GB	0.69	0.55	0.62	0.66	0.24	0.64	0.52	0.58	0.62	0.17

LR, Logistic Regression; RF, Random Forest; NB, Naive Bayes; LGBM, Light Gradient Boosting Machine; GB, Gradient Boost; Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; AUROC, Area Under the Receiver Operating Characteristic; MCC, Matthews Correlation Coefficient;

4.4 Performance on PSSM with RSA

We also developed models using a combination of PSSM and Relative Surface Accessibility (RSA) profiles to predict conformational B-cell epitopes. As shown in Table 4, incorporating RSA alongside evolutionary information from PSSM improved predictive performance. The Random Forest (RF) model stood out among the classifiers, achieving the highest AUROC of 0.64 and an MCC of 0.18 on the validation dataset, highlighting its strong capability in distinguishing interacting from non-interacting residues.

Other classifiers also demonstrated competitive performance. Logistic Regression (LR) attained an AUROC of 0.60 and an MCC of 0.14, while LightGBM (LGBM) achieved an AUROC of 0.62 and MCC of 0.17. The Naive Bayes (NB) model showed moderate performance with an AUROC of 0.61 and an MCC of 0.18, whereas Gradient Boost (GB) recorded an AUROC of 0.62 and an MCC of 0.19, placing it alongside the best-performing models.

Table 5: The ML performance results on window 17 using PSSM matrix with RSA features

Model	**Training dataset**					**Validation dataset**				
	Sens	**Spec**	**Acc**	**AUROC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUROC**	**MCC**
LR	0.64	0.59	0.62	0.66	0.23	0.58	0.55	0.57	0.60	0.14
RF	0.67	0.56	0.61	0.66	0.23	0.64	0.54	0.59	0.64	0.18
NB	0.73	0.47	0.60	0.64	0.21	0.65	0.53	0.59	0.61	0.18
LGBM	0.65	0.58	0.62	0.66	0.23	0.63	0.55	0.59	0.62	0.17
GB	0.70	0.55	0.63	0.67	0.25	0.65	0.53	0.59	0.62	0.19

LR, Logistic Regression; RF, Random Forest; NB, Naive Bayes; LGBM, Light Gradient Boosting Machine; GB, Gradient Boost; Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; AUROC, Area Under the Receiver Operating Characteristic; MCC, Matthews Correlation Coefficient;

These results support the notion that RSA adds valuable structural context to evolutionary profiles, enhancing model performance for epitope prediction. In this study, RSA values were predicted using SPOT-1D, a deep learning framework trained on sequence-to-structure mappings. It estimates RSA directly from amino acid sequences using a bidirectional LSTM

architecture, enabling residue-level solvent accessibility predictions without requiring explicit 3D structures. This method gave us effective and precise RSA estimations, which were then incorporated into our pattern-based feature representations for enhanced prediction accuracy.

4.5 Performance on PSSM with BPP and RSA

Similarly, we also combined PSSM and RSA with the Binary Profile to develop prediction models [36]. As shown in Table 5, incorporating all three features—PSSM, RSA, and Binary Profile—resulted in performance comparable to using only PSSM and RSA. The model Random Forest (RF) remained the top performer, achieving an AUROC of 0.63 and an MCC of 0.18 on the validation dataset, reflecting its ability to effectively leverage complementary biological information. Gradient Boost (GB) also demonstrated strong performance, with an AUROC of 0.62 and MCC of 0.18, followed closely by LightGBM (LGBM) with an AUROC of 0.62 and MCC of 0.17. Logistic Regression (LR) and Naive Bayes (NB) produced modest results, with AUROC scores of 0.60 each and MCC values of 0.12 and 0.16, respectively. These results suggest that while the inclusion of Binary Profile alongside RSA and PSSM contributes marginally to overall performance, the structural (RSA) and evolutionary (PSSM) features remain the most influential for epitope prediction.

Table 6: The ML performance results on window 17 using PSSM matrix with RSA and Binary profile

	Training dataset					**Validation dataset**				
Model	**Sens**	**Spec**	**Acc**	**AUROC**	**MCC**	**Sens**	**Spec**	**Acc**	**AUROC**	**MCC**
LR	0.61	0.59	0.60	0.65	0.20	0.58	0.54	0.56	0.60	0.12
RF	0.67	0.57	0.62	0.67	0.24	0.63	0.55	0.59	0.63	0.18
NB	0.68	0.52	0.60	0.64	0.21	0.62	0.54	0.58	0.60	0.16
LGBM	0.65	0.58	0.61	0.66	0.23	0.61	0.56	0.58	0.62	0.17
GB	0.70	0.55	0.62	0.67	0.25	0.65	0.52	0.59	0.62	0.18

LR, Logistic Regression; RF, Random Forest; NB, Naive Bayes; LGBM, Light Gradient Boosting Machine; GB, Gradient Boost; Sens, Sensitivity; Spec, Specificity; Acc, Accuracy; AUROC, Area Under the Receiver Operating Characteristic; MCC, Matthews Correlation Coefficient;

4.6 Final Model

Based on the evaluation of different feature combinations and machine learning models, RF trained on PSSM and RSA emerged as the best-performing model. It achieved the highest AUROC of 0.64 and MCC of 0.18 demonstrating its effectiveness in capturing key structural and evolutionary patterns in antigen-antibody interactions. While the Binary Profile feature had minimal impact, RSA significantly enhanced model performance by providing insights into residue solvent exposure. The results confirm that structural accessibility (RSA) and evolutionary conservation (PSSM) are the most informative features for B-cell epitope prediction. Thus, the final model prioritizes PSSM with RSA with RF for optimal predictive performance.

Design and implementation of a web server

To facilitate the scientific community, we have also created an accessible and user-friendly web server, CBTOPE2, where the top-performing predictive models of this work are hosted. As a result of a comprehensive assessment across various feature sets and machine learning methods, the deployed model utilized utilizes the Random Forest (RF) classifier trained on a blend of Position-Specific Scoring Matrix (PSSM) and Relative Surface Accessibility (RSA) features. The best-performing model in validation attained an AUROC of 0.64 and MCC of 0.18, capturing the key structural and evolutionary relationships underlying antigen–antibody interactions efficiently. The RSA feature, containing residue solvent exposure information, was particularly important for improving prediction accuracy, while the Binary Profile feature contributed minimally. These outcomes validate that the union of structural accessibility and evolutionary conservation—implied via RSA and PSSM, respectively—are the most informative for B-cell epitope prediction.

The web server enables users to enter antigen sequences in FASTA format to provide residue-level predictions for likely epitope areas. Clearly displayed results indicate interacting residues in the antigen structure. Users are also able to download the full datasets incorporated into model construction, both training and validation sets, to enable additional analysis and reproducibility.

The server is implemented using HTML, JavaScript, and PHP, and is fully compatible across devices including desktops, laptops, smartphones, and tablets. To further improve accessibility and usability, a standalone version of CBTOPE2 is also provided. It can be downloaded from GitHub at https://github.com/raghavagps/cbtope2 and installed using PyPI with the command pip install cbtope2. This offline version ensures that researchers can perform predictions without requiring an active internet connection.

CBTOPE2 web server is available free of cost at https://webs.iiitd.edu.in/raghava/cbtope2/, and it can be used as a trusted resource for scientific communities involved in immunoinformatics and vaccine research.

Chapter 5: DISCUSSION

This study introduces CBTOPE2, a machine learning–based framework designed to predict conformational B-cell epitopes from antigen protein sequences. While several past efforts have focused on predicting linear B-cell epitopes or using surface features alone, our work is distinct in integrating structural and evolutionary descriptors—specifically, Relative Surface Accessibility (RSA) and Position-Specific Scoring Matrix (PSSM)—to improve the identification of conformational epitope residues.

We employed a systematic pipeline, beginning with a benchmark dataset of 268 high-resolution antigen–antibody complex structures curated from prior literature. Each antigen sequence was fragmented using overlapping sliding windows of lengths ranging from 7 to 31 residues, and labels were assigned based on whether the central residue in each window was involved in antigen–antibody interaction. However, only a small fraction of residues in protein structures form actual epitopes, leading to a significantly imbalanced dataset. To address this, we applied undersampling of negative samples to balance the number of interacting and non-interacting patterns during training, which helped mitigate model bias and enhanced learning on the minority class.

Extensive exploratory data analysis revealed several biological trends. For instance, the central residues of epitopic patterns were found to be dominated by charged and polar amino acids such as glutamic acid (E), lysine (K), and aspartic acid (D), suggesting that electrostatic interactions may play a key role in antigen recognition. Furthermore, our analysis of RSA distributions confirmed that epitopic residues tend to be more solvent-exposed than non-epitopic residues, in line with immunological expectations. PSSM-based analyses also demonstrated higher conservation scores in epitope regions, supporting the hypothesis that functionally important residues are more conserved.

Several machine learning classifiers were trained and evaluated across multiple feature sets and window sizes using five-fold cross-validation. Among the models, Random Forest (RF) trained on a combination of PSSM and RSA consistently emerged as the best performer, achieving a

maximum AUROC of 0.64 and MCC of 0.18 on the independent validation set. This combination outperformed models built on binary profile features alone or other classifiers like Naive Bayes, Logistic Regression, and Gradient Boost. The results underscore the complementary value of RSA and PSSM—capturing structural exposure and evolutionary conservation respectively—for robust epitope prediction. On the other hand, binary profile features showed limited improvement when added, indicating that their contribution to epitope characterization is less prominent in the presence of stronger descriptors.

To facilitate broader adoption, we deployed the final model as a user-friendly web server and standalone package. The server allows users to submit antigen sequences in FASTA format and receive residue-level epitope predictions. By providing both web and offline modes, along with access to training datasets, CBTOPE2 supports reproducibility, transparency, and usability for immunologists and vaccine developers.

Overall, our approach highlights the effectiveness of integrating structure-informed and sequence-derived features, coupled with rigorous model validation, to enhance B-cell epitope prediction. CBTOPE2 can serve as a valuable resource in the rational design of peptide-based vaccines and antibody therapies. Future extensions may explore graph-based neural networks and 3D structure-aware architectures to further push the boundaries of epitope prediction.

References

[1] W. Zhang, Y. Xiong, M. Zhao, H. Zou, X. Ye, J. Liu, Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature, BMC Bioinformatics 12 (2011) 341.

[2] I. Sela-Culang, V. Kunik, Y. Ofran, The structural basis of antibody-antigen recognition, Front Immunol 4 (2013) 302.

[3] N.D. Rubinstein, I. Mayrose, D. Halperin, D. Yekutieli, J.M. Gershoni, T. Pupko, Computational characterization of B-cell epitopes, Mol Immunol 45 (2008) 3477–3489.

[4] J.L. Pellequer, E. Westhof, M.H. Van Regenmortel, Predicting location of continuous epitopes in proteins from their primary structures, Methods Enzymol 203 (1991) 176–201.

[5] Van Regenmortel MHV, Mapping Epitope Structure and Activity: From One-Dimensional Prediction to Four-Dimensional Description of Antigenic Specificity, Methods 9 (1996) 465–472.

[6] Website, (n.d.). Website, (n.d.). Website, (n.d.). L. Breiman, Mach. Learn. 45 (2001) 5–32. https://doi.org/10.1023/a:1010933404324.

[7] H.R. Ansari, G.P. Raghava, Identification of conformational B-cell Epitopes in an antigen from its primary sequence, Immunome Res 6 (2010) 6.

[8] G. Cia, F. Pucci, M. Rooman, Critical review of conformational B-cell epitope prediction methods, Brief Bioinform 24 (2023). https://doi.org/10.1093/bib/bbac567.

[9] Website, (n.d.). Ansari, H.R., Raghava, G.P. Identification of conformational B-cell Epitopes in an antigen from its primary sequence. Immunome Res 6, 6 (2010). https://doi.org/10.1186/1745-7580-6-6.

[10] A. Pande, P. Gupta, S. Awasthi, G.P.S. Raghava, Pfeature: A Tool for Computing Wide Range of Protein Features from Sequence and Structure, bioRxiv (2019). https://doi.org/10.1101/599126.

[11] C.-Y.J. Peng, K.L. Lee, G.M. Ingersoll, An Introduction to Logistic Regression Analysis and Reporting, The Journal of Educational Research 96 (2002) 3–14.

[12] G.I. Webb, Na"ive Bayes, in: C. Sammut, G.I. Webb (Eds.), Encyclopedia of Machine Learning, Springer US, Boston, MA, 2011: pp. 713–714.

[13] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman,

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Research 25 (1997) 3389–3402.

[14] T.U. Consortium, UniProt: the universal protein knowledgebase in 2023, Nucleic Acids Research 51 (2023) D523–D531.

[15] A. Qureshi, A. Sacan, GPCRTree: An SVM-based method for GPCR classification, Computers in Biology and Medicine 43 (2013) 1104–1113.

[16] J. Hanson, K. Paliwal, T. Litfin, Y. Yang, Y. Zhou, SPOT-1D: accurate sequence-based prediction of protein secondary structure and solvent accessibility using deep neural networks, Bioinformatics 35 (2019) 4049–4056.

[17] S. Saha, G.P.S. Raghava, Prediction of Continuous B-cell Epitopes Using Physico-Chemical Properties and Machine Learning Methods, International Journal of Immunogenetics 34 (2007) 401–405.

[18] S. Liang, D. Zheng, C. Zhang, D.M. Standley, Fast and accurate prediction of protein solvent accessibility using a two-stage support vector regression approach, BMC Bioinformatics 10 (2009) 261.

[19] Website, (n.d.). Website, (n.d.). L. Breiman, Mach. Learn. 45 (2001) 5–32. https://doi.org/10.1023/a:1010933404324.

[20] Website, (n.d.). Website, (n.d.). G.I. Webb, Naïve Bayes, in: Encyclopedia of Machine Learning, Springer US, Boston, MA, 2011: pp. 713–714. https://doi.org/10.1007/978-0-387-30164-8_576.

[21] Website, (n.d.). Website, (n.d.). C.-Y.J. Peng, K.L. Lee, G.M. Ingersoll, An introduction to logistic regression analysis and reporting, J. Educ. Res. 96 (2002) 3–14. https://doi.org/10.1080/00220670209598786.

[22] R. Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), Morgan Kaufmann Publishers, 1995: pp. 1137–1143.

[23] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, Journal of Machine Learning Research 13 (2012) 281–305.

[24] T. Fawcett, An Introduction to ROC Analysis, Pattern Recognition Letters 27 (2006) 861–874.

[25] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over

F1 score and accuracy in binary classification evaluation, BMC Genomics 21 (2020) 6.

[26] D.M.W. Powers, Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation, Journal of Machine Learning Technologies 2 (2011) 37–63.

[27] L. Zhao, L. Wong, W.W.B. Goh, Evaluation of Computational Tools for B-cell Epitope Prediction: Opportunities, Challenges, and Future Directions, Briefings in Bioinformatics 20 (2019) 1127–1138.