Prediction of Conformational B-cell Epitopes in Structure of Protein

by Megha

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

Indraprastha Institute of Information Technology Delhi
June, 2025

Prediction of Conformational B-cell Epitopes in Structure of Protein

by Megha

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 "**Prediction of Conformational B-cell Epitopes in Structure of Protein**" being submitted by Megha 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 in 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 "**Prediction of Conformational B-cell Epitopes in Structure of Protein**"

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.

Megha

Roll No.: MT23125

Abstract

The correct prediction of conformational B-cell epitopes is absolutely crucial to the rational design of subunit vaccines and antibody-based therapies. Contrary to the nature of linear epitopes, conformational epitopes are comprised of non-contiguous residues brought into close spatial proximity through the three-dimensional folding of the antigen. The process of identifying them has become a computationally challenging task. Initial models typically used only sequence-derived features and performed poorly. Researchers have begun to incorporate 3D structural information into prediction workflows to break these limitations. Several notable tools have been developed to exploit structural cues. DiscoTope (2006) employed spatial proximity and solvent accessibility; ElliPro (2008) utilized ellipsoid-based modeling and residue protrusion indices; SEPPA (2009, 2011) utilized neighborhood-preserving geometry; EpiPred (2013) combined docking and energetic modeling; and BEpro (previously PEPITO) utilized surface clustering and propensity scoring. The tools usually employ structural features like solvent accessibility, secondary structure, flexibility, and topological features, which are typically derived through DSSP, and utilize statistical or machine learning methods. In this work, we present a structure-based machine learning model for prediction of conformational B-cell epitopes from a benchmark dataset developed by Cia et al. (2023), consisting of high-resolution antibody–antigen complexes. We extracted a broad spectrum of structural features derived from DSSP, ranging from secondary structure, absolute and relative solvent accessibility (ACC, RSA) ,backbone torsion angles (phi and psi), to hydrogen bonding metrics. The residues were encoded with a sliding window to maintain local structural context. We compared several machine learning models, including Random Forest, Logistic Regression, LightGBM, and Gradient Boosting. The gradient boosting classifier yielded the best results, with an AUROC of 0.76 and MCC of 0.43 on the validation set. Analysis of feature importance unveiled torsion angles, solvent accessibility, and secondary structure as high contributors. This work stresses a successful application of structure-driven features for epitope prediction and offers a robust, transferable pipeline for subsequent immunoinformatics applications.

List of Figures

List of Tables

Contents

Chapter 1: INTRODUCTION

1.1 Background

The adaptive immune system relies heavily on the recognition of foreign antigens, a process facilitated by B-cell epitopes—specific regions on antigen surfaces that are recognized and bound by B-cell receptors or antibodies. These epitopes are central to initiating immune responses and have immense therapeutic relevance in vaccine development, antibody engineering, and diagnostics. B-cell epitopes are typically classified into two main categories: linear (continuous) epitopes, which consist of sequential amino acid residues, and conformational (discontinuous) epitopes, which arise from amino acids that are spatially close in the protein's 3D structure but not necessarily adjacent in the primary sequence [1].

The accurate identification of conformational B-cell epitopes is vital for advancing peptide-based vaccines and therapeutic antibodies. Traditional experimental techniques such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy have been employed to determine epitope regions[2] . While these methods are accurate, they are also laborious, costly, and time-intensive, making them less suitable for high-throughput screening.

To overcome these limitations, computational methods for epitope prediction have gained considerable attention. Early approaches primarily relied on sequence-derived features like amino acid composition, flexibility, and hydrophilicity[3]. However, these models often performed poorly when applied to conformational epitope prediction, largely due to the absence of 3D structural context. This gap has led to the emergence of structure-based predictive models, which integrate spatial, topological, and physicochemical descriptors obtained from high-resolution antigen structures [4].

Several structural attributes have been identified as influential in determining epitope likelihood:

- Secondary structure elements (helix, sheet, coil) which influence residue accessibility and stability.
- Relative Solvent Accessibility (RSA), as surface-exposed residues are more likely to interact with antibodies [5].

- Backbone torsion angles (ϕand ψ), which define local protein conformation and flexibility [[6]].
- Hydrogen bonding patterns, which help identify structurally constrained or flexible regions conducive to epitope formation.
- Atomic depth and residue mobility, which are associated with antigenicity and immune recognition [7].

With the increasing availability of structural data through databases like the Protein Data Bank (PDB) [8], machine learning (ML)-based models now have the potential to leverage this structural information for high-accuracy epitope prediction. Recent studies have shown that supervised learning algorithms, particularly ensemble-based methods such as Random Forests, Gradient Boosting, and LightGBM, perform well in capturing complex, non-linear relationships between structural features and epitope propensity [9,10].

In the present work, we explore the predictive potential of these structural features by applying a window-based encoding strategy to capture local structural environments around each residue. Using a combination of well-curated antigen–antibody complexes and advanced ML algorithms, this study aims to develop robust models for conformational B-cell epitope prediction, addressing a critical gap in computational immunology and contributing to the development of next-generation vaccine and antibody therapies.

1.2 Motivation of work

The development of effective vaccines and antibody-based therapeutics remains a global priority in combating infectious diseases. Central to this goal is the accurate identification of B-cell epitopes, which are specific regions on antigens recognized by B-cell receptors or antibodies. These epitopes serve as critical targets for eliciting an immune response, thereby playing a fundamental role in the design of subunit vaccines and antibody therapeutics [1]. However, the majority of experimentally confirmed B-cell epitopes are conformational in nature—formed by spatially proximate amino acid residues in a protein's tertiary structure, but non-contiguous in its primary sequence [2]. This intrinsic complexity poses a major challenge for epitope prediction.

Despite decades of research and the availability of computational tools, many existing models continue to exhibit suboptimal performance, particularly in predicting conformational epitopes. A primary reason is the heavy reliance on sequence-based features and training on small or outdated datasets, which do not adequately capture the spatial and physicochemical intricacies of epitope regions [3][5]. As revealed in a recent benchmark study by Cia et al. (2023), even widely-used tools trained on earlier data [11].

Adding to this challenge is the underutilization of structural information in many predictive frameworks. Critical features such as secondary structure elements, solvent accessibility (RSA), torsional angles (ϕ/ψ), and hydrogen bonding patterns are either ignored or insufficiently modeled, despite being known to influence antigenicity and antibody binding [4, 6, [7].

Thus, the motivation behind this work is threefold:

1. To retrain and refine epitope prediction models using up-to-date and well-curated datasets that better reflect real-world antigen–antibody interactions.
2. To investigate the predictive value of diverse structural features, including RSA, secondary structure, torsional angles, and hydrogen bonding, in distinguishing epitope from non-epitope residues.
3. To develop robust, window-based ML classifiers that integrate spatially local structural context and offer improved accuracy in conformational epitope prediction.
4. By focusing on the integration of structural features with modern supervised learning algorithms, this work aims to overcome the limitations of earlier sequence-based models and contribute a reliable, generalizable tool to support vaccine development, antibody design, and computational immunology research.

1.3 Objective

The major objective of this research is to develop a machine learning-based model for the accurate prediction of conformational B-cell epitopes by leveraging structural features extracted from experimentally resolved antigen–antibody complexes. This work aims to address the limitations of sequence-only prediction methods by incorporating rich 3D structural context into the model training process.

1.3.1 Specific Objectives

To achieve this goal, the study is guided by the following specific objectives:

- To curate and preprocess a high-quality dataset of antigen–antibody complexes from a benchmark set, ensuring accurate residue-level epitope annotations.
- To extract and encode structural features relevant to epitope prediction, including:
    - Secondary structure
    - Relative solvent accessibility (RSA)
    - Backbone torsion angles (ϕ and ψ)
    - Hydrogen bonding patterns
- To apply a window-based encoding strategy that captures local structural environments of each residue.
- To develop and evaluate multiple machine learning models, including ensemble methods, using both cross-validation and independent testing for robust performance assessment.
- To compare model performance across various structural feature combinations and determine the most informative descriptors for predicting conformational B-cell epitopes.
- To analyze and interpret feature importance in biological terms, offering insights into structural determinants of antibody binding.

1.4 Scope

This work centers on predicting conformational B-cell epitopes through a structure-based machine learning strategy. In contrast to the conventional sequence-based approach, this research utilizes a varied collection of structural features derived from DSSP, such as secondary structure, solvent accessibility (RSA, ACC), torsion angles (phi, psi), and hydrogen bonding patterns, to describe the local environment of antigen residues.

The scope encompasses:

Preprocessing and feature extraction of structural features from an experimentally solved antibody–antigen benchmark dataset. Window-based feature encoding to retain local context surrounding each residue. Training and evaluation of various machine learning models, ensemble classifiers, and deep learning architectures to compare predictive performance. Comparative analysis of the contributions of individual features and their relation with epitope regions .Determination of the best-performing model and feature set by metrics such as AUROC, MCC, and accuracy. This paper is not about deploying a software tool or web server but wants to add a feature-rich, structure-sensitive modeling pipeline that may facilitate future developments in epitope prediction, immunoinformatics, and rational vaccine design.

Chapter 2: LITERATURE REVIEW

2.1 Introduction

Conformational B-cell epitopes (CBEs) are specific surface-exposed regions of antigens that are recognized by antibodies. Accurate identification of these epitopes is vital in vaccine design, antibody production, and immunodiagnostics [1,2]. Traditionally, the identification of CBEs has relied on laborious and time-intensive experimental methods such as X-ray crystallography and cryo-electron microscopy [8]. However, with the advent of computational biology, machine learning (ML)-based predictive models have become a powerful alternative due to their speed, cost-effectiveness, and ability to analyze large datasets [3,4,7].

Over the years, several bioinformatics tools and computational models have been developed to predict CBEs. These methods can be broadly classified into structure-based and sequence-based approaches. Structure-based tools such as Epitopia, SEPPA, and Discotope rely on three-dimensional (3D) structural information of antigen-antibody complexes, limiting their applicability to only proteins with resolved structures [6,8]. Despite these advances, recent benchmarking by Cia et al. (2023) has revealed that many of the existing tools, perform poorly when tested on large, high-quality datasets of antibody-antigen complexes [11]. One key limitation is that earlier models were trained on small and outdated datasets, leading to reduced generalizability [5]. Furthermore, many methods do not incorporate sufficient biological context, such as solvent accessibility or evolutionary conservation, which are crucial for accurate epitope prediction [4,5].

2.2 Feature Engineering

To improve the accuracy of conformational B-cell epitope prediction, various structure-derived features have been explored [4,5]. These include secondary structure elements, which influence the overall fold and local accessibility of residues; and relative solvent accessibility (RSA), which reflects the exposure of amino acids on the protein surface—a critical factor in antibody recognition [6,8]. Additional geometric descriptors such as absolute solvent accessibility (ACC)

and backbone torsion angles (phi and psi) provide further insight into residue orientation and flexibility within the 3D structure .[6,12]

Moreover, hydrogen bonding interactions, including NH→O and O→NH bond energies and relative indices (for both first and second-order hydrogen bond partners), serve as proxies for local structural stability [4,13]. Inter-residue contacts and binary profile features—capturing residue identity in a windowed format—enable the encoding of both spatial and sequential context [3,10,13]. These features collectively provide a detailed representation of the residue's local environment, which can be leveraged by machine learning models to distinguish epitope and non-epitope regions more effectively [7,10,13].

Recent studies indicate that combining these diverse structural descriptors enables models, particularly ensemble classifiers like gradient boosting, to capture complex patterns associated with epitope formation while maintaining robustness across variable protein topologies [10,11,14].

2.3 Use of Machine Learning Models

In this study, a wide range of machine learning and deep learning models were implemented for conformational B-cell epitope prediction. The models included:

- Logistic Regression, which serves as a baseline linear classifier;
- Random Forest and Decision Tree, known for their interpretability and ability to handle complex feature interactions;
- Gaussian Naïve Bayes, a probabilistic model assuming feature independence;
- Gradient Boosting, XGBoost, and LightGBM (LGBMClassifier), which are ensemble learning techniques optimized for high-dimensional and imbalanced biological data;
- Convolutional Neural Network (CNN) and Bi-directional LSTM (BiLSTM), used to capture spatial and sequential dependencies;
- SimpleRNN, a recurrent model for sequence encoding.

These models were chosen to explore both conventional and advanced techniques for pattern learning in structural feature space.

While simpler models like Logistic Regression and Naïve Bayes offer interpretability, ensemble methods such as Gradient Boosting, XGBoost, and LGBMClassifier have shown superior performance by effectively capturing nonlinear and high-order interactions among structural features. Deep learning models like CNN and BiLSTM were also evaluated, though their performance was comparatively limited on the available structural dataset.

Unlike other platforms such as PreTP-Stack, PEPred-Suite, and PPTPP, which often use synthetic peptide datasets or predicted properties, the models developed in this study were trained on experimentally validated epitope annotations derived from high-resolution antibody–antigen complex structures. All structural features—including secondary structure, solvent accessibility, torsion angles, and hydrogen bond metrics—were computed using the DSSP tool, ensuring consistency and reliability in annotation.

This integrated modeling approach allows for a more biologically grounded and realistic prediction framework, directly applicable to real-world immunoinformatics and vaccine design efforts.

2.4 Gap in Existing Literature

A common shortcoming of most existing models is their reliance on predicted or assumed therapeutic sequences rather than experimentally validated druggable proteins. This limits their utility in drug development pipelines where precise, biologically validated predictions are critical. Moreover, these methods often ignore key characteristics of real-world antigens and antibodies, such as post-translational modifications and structural flexibility, which influence epitope accessibility.

2.5 Summary

To address the limitations of existing sequence-based methods, this study explores a structure-informed machine learning approach for conformational B-cell epitope prediction. Using a high-quality benchmark dataset of antibody–antigen complexes curated by Cia et al. (2023), we extract a rich set of structural features, including secondary structure, relative and absolute solvent accessibility (RSA, ACC), backbone torsion angles (phi and psi), and hydrogen

bonding metrics, all computed using DSSP. Additionally, binary profile representations were incorporated to capture local sequence identity within structural contexts.

A range of machine learning and deep learning models—including Gradient Boosting, Random Forest, XGBoost, LGBM, CNN, BiLSTM, and SimpleRNN—were evaluated using a sliding window-based encoding strategy. Among these, ensemble tree-based models demonstrated the most reliable performance. This structure-aware framework significantly improves the model's ability to identify conformational epitopes, offering insights that are more biologically grounded and generalizable compared to earlier sequence-only models.

This work focuses on evaluating feature contributions and model performance, contributing to the growing body of research in structure-based immunoinformatics and providing a foundation for future tool development.

Chapter 3: METHODOLOGY

3.1 Preparation of Dataset

This work uses a benchmark dataset compiled by Cia et al. (2023) that is explicitly prepared for the assessment of conformational B-cell epitope prediction tools [11]. The dataset includes 286 non-redundant, high-resolution antibody–antigen complexes derived from the Protein Data Bank (PDB) . They were thoroughly screened on the basis of quality criteria such as resolution and completeness to guarantee homogeneity and reliability in subsequent analysis.

For every antigen structure, residue-level labels describe whether a residue in question is included in a conformational B-cell epitope. The labels were defined according to spatial proximity to antibody chains in the complex, with the final verdict (epitope or non-epitope) noted in the last column of the respective CSV files.

A total of 268 antigen structures were chosen from the entire benchmark for building models. Each structure was preprocessed to obtain local residue patterns, which acted as input examples for machine learning. Rather than sequence features alone, this work is concerned with structure-derived descriptors, such as: Secondary structure assignments, Relative and absolute solvent accessibility (RSA and ACC), Backbone torsion angles (phi and psi), Hydrogen bonding energies and relative indices (NH→O and O→NH), Residue-level binary profile in a windowed setting.

The data was split into two mutually exclusive subsets: 214 structures (80%) for training and 54 structures (20%) for independent validation, in such a way that no structure was common to both. These subsets were used reliably throughout the entire machine learning pipeline for training, cross-validation, and final performance assessment.

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	5H3S_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_1
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 Generating Pattern

In order to identify the local structural environment of each residue that is implicated in antigen–antibody interactions, sliding window-based pattern extraction was utilized. Overlapping residue patterns were generated with window lengths between 3 and 21 amino acids. Each pattern was centered on a target residue, which was given a binary label—positive if the central residue belonged to a conformational B-cell epitope (i.e., antibody-interacting), and negative otherwise. For residues that occur at or near the termini of protein sequences, padding using a dummy amino acid token was used to guarantee that all patterns had a standard length throughout the dataset. Such homogeneity enables the implementation of fixed-size input vectors within machine learning algorithms. This method has been extensively utilized in previous epitope prediction research and allows the models to leverage both the features of the central residue and also the structural context around it, which is essential for proper identification of epitope areas.

3.1.2 Class Distribution and Dataset Balancing

This work uses a benchmark dataset compiled by Cia et al. (2023) that is explicitly prepared for the assessment of conformational B-cell epitope prediction tools. The dataset includes 286 non-redundant, high-resolution antibody–antigen complexes derived from the Protein Data Bank (PDB). They were thoroughly screened on the basis of quality criteria such as resolution and completeness to guarantee homogeneity and reliability in subsequent analysis.

For every antigen structure, residue-level labels describe whether a residue in question is included in a conformational B-cell epitope. The labels were defined according to spatial proximity to antibody chains in the complex, with the final verdict (epitope or non-epitope) noted in the last column of the respective CSV files.

A total of 268 antigen structures were chosen from the entire benchmark for building models. Each structure was preprocessed to obtain local residue patterns, which acted as input examples for machine learning. Rather than sequence features alone, this work is concerned with structure-derived descriptors, such as: Secondary structure assignments, Relative and absolute solvent accessibility (RSA and backbone torsion angles (phi and psi), Hydrogen bonding energies and relative indices (NH→O and O→NH), Residue-level binary profile in a windowed setting.

The data was split into two mutually exclusive subsets: 214 structures (80%) for training and 54 structures (20%) for independent validation, in such a way that no structure was common to both. These subsets were used reliably throughout the entire machine learning pipeline for training, cross-validation, and final performance assessment.

3.2 Generation of Features

For training and assessing machine learning models to predict conformational B-cell epitopes, we used a feature generation approach based on the structural and physicochemical characteristics of protein residues. Each pattern based on residues was described with a set of features that reflect its local conformation, surface exposure, and bonding properties.The following features were extracted: Secondary structure labels,Relative and absolute solvent accessibility (RSA and ACC), Backbone torsion angles (phi and psi), Hydrogen bond contacts,

relative indices and energies of both NH→O and O→NH bonds (first and second order), Binary profile vectors, residue identity within a windowed environment.

These features were obtained from structural annotations produced by DSSP, which is done to ensure consistency and accuracy for all the patterns. By combining both spatial and physicochemical properties, the generated feature set offers a dense description of the structural environment of every residue, thus enabling improved discrimination of epitope from non-epitope regions by the model.

3.2.1 Binary Profile

Binary profile features represent every amino acid in a sequence window with a 21-dimensional one-hot encoded vector. The vector encodes the identity of the amino acid, with one dimension equal to 1 (for the current amino acid) and all others as 0. Another dummy amino acid ('X') was added to encode for padding residues in shorter sequences. For a specified window size W, the output binary profile is a 21×W matrix, which is reshaped to a 1×(21×W) vector for model input [3,10]. This representation maintains the residue sequential order and composition, allowing machine learning models to capture patterns strictly on the basis of amino acid identity [7,10]. Binary profiles were created with Pfeature's BinaryProfile module, which has been extensively employed in tasks of protein function and epitope prediction [15].

3.2.3 Relative Solvent Accessibility (RSA)

Relative Solvent Accessibility (RSA) refers to a value of how much an amino acid residue is solvent-accessible in the tertiary structure of a protein [6]. In this paper, RSA values were calculated by using the DSSP (Define Secondary Structure of Proteins) program, which predicts solvent accessibility from experimentally derived PDB structures based on atomic coordinates [16]. DSSP operates by comparing the 3D structure of a protein and establishing hydrogen bonding patterns and solvent exposure for individual residues. It employs a specific rolling water sphere model to estimate the accessible surface area of every amino acid by modeling how much of each residue's surface is solvent-exposed. Secondary structure is assigned and the Absolute Solvent Accessibility (ASA) was obtained through tracing of atoms contacting solvent considering neighbor atoms' van der Waals radii. Normalized absolute ASA values extracted

from the DSSP program were obtained by dividing ASA by the highest accessible surface area, as defined by a standard reference scale, for each amino acid type in order to calculate the RSA values. For each residue in a windowed pattern, the RSA value was derived from the DSSP output, and only for the central residue in the pattern the value was retained. A single value that is the fractional surface exposure relative to the maximum possible accessibility for the type of residue was rounded to two places and appended to the feature vector for further modeling. RSA is a critical structural feature of epitope prediction because solvent-exposed residues are more likely to be involved in the interaction with the antibody, and buried residues are generally inaccessible [5,6]. The incorporation of RSA improves the model's ability to differentiate likely epitope regions from non-epitope regions based on surface accessibility [4,11].

3.2.4 ACC

Accessible Contact Capacity (ACC) is a structural descriptor, which measures the extent to which a residue is both in contact with neighboring residues and solvent-accessible [16]. In contrast to mere RSA, which measures only solvent exposure, ACC considers each residue's local packing environment within the 3D structure [6,17]. Increased ACC values generally reflect the fact that a residue is both exposed and loosely packed, features which are commonly found in surface-exposed areas such as B-cell epitopes [5,11,16]. In capturing this subtle interaction between solvent exposure and residue contact, ACC offers insightful clues for the differentiation of possible epitope sites from the protein interior. In this research, ACC was taken as a 7-dimensional vector for every residue window to facilitate more accurate identification of conformational epitopes.

3.2.5 PHI - PSI

Phi (ϕ) and Psi (ψ) angles are the two primary torsion angles which determine the backbone conformation of amino acid residues of a protein structure [6,17]. These are the angles of rotation about the N–Cα (phi) and Cα–C (psi) bonds, respectively, and play an important role in defining the local secondary structure, including α-helices and β-sheets [18]. Their ranges are limited by steric hindrance and tend to cluster in certain areas of the Ramachandran plot [[18]. Through the inclusion of phi and psi angles as features in this research, we seek to capture

crucial conformational traits that differentiate surface-exposed and flexible residues—features commonly referred to in B-cell epitopes [5,11,16]]. The use of these angular features enables the model to better perceive the spatial organization of residues in three dimensions.

3.2.6 The NH→O and O→NH Relative Index position

The NH→O and O→NH Relative Index positions are representations of hydrogen bonding interactions in the protein backbone[4,17,19] . The indices reflect the relative residue position involved in the hydrogen bond, offering insights into local and non-local interactions that are important for protein stability [6,18,19]. In particular, NH→O 1 and 2 RelIdx indicate the acceptor residue position with respect to the donor, whereas O→NH 1 and 2 RelIdx indicate the donor's position with respect to the acceptor. These relative indices are useful in depicting structural motifs, such as turns, helices, or sheets, that tend to affect the accessibility and flexibility of surface residues [5]. These features thus help the model understand the hydrogen bonding network of the protein more thoroughly, adding to a better discrimination capability between epitopes and non-epitopes [11,16,19].

3.2.7 Energy

NH→O and O→NH Energy features: These are measures that quantify the strength of hydrogen bonds formed by backbone atoms of the protein [19]. Their values, calculated strictly using geometrical and distance-based criteria, symbolize the energetic contribution of each hydrogen bond to the overall structure of the protein [4,19]. NH→O 1 and 2 Energy denotes the energy of bonds in which the backbone amide hydrogen is the donor, whereas O→NH 1 and 2 Energy is employed to describe bonds in which the carbonyl oxygen is an acceptor [18,19]. Stable hydrogen bonds (low values of energy) tend to stabilize secondary structures such as α-helices and β-sheets and affect the structural stiffness and surface exposure of residues [5,6,19]. These energy-based descriptions enable the model to detect subtle conformational signals, thus improving its capacity for distinguishing between interacting epitope residues and non-interacting ones [11,16] .

3.2.8 Secondary Structure (SS)

Secondary Structure (SS) refers to the local spatial arrangement of a protein's backbone atoms, typically categorized into helices (H), beta-sheets (E), and coils (C) [4,6]. These structural motifs are essential for understanding the functional and physical properties of proteins. In the context of B-cell epitope prediction, secondary structures provide valuable information about the conformational flexibility and accessibility of residues [5,7]. For instance, residues in coil regions are generally more exposed and flexible, making them more likely to participate in antigen-antibody interactions [2,11]. In this study, we used DSSP to assign the secondary structure of each residue and encoded this information as a 3-dimensional probability vector for helix, sheet, and coil [20][21]. These values were then included in the feature representation of each pattern to capture structural tendencies associated with epitope regions.

3.3 Exploratory Data Analysis

We also performed exploratory data analysis (EDA) to look at the distribution and important features of the dataset before developing models. EDA provided significant information on class imbalance, residue-specific patterns, and the biological significance of the features that were extracted.

3.3.1 Handling Dataset Imbalance

As demonstrated in Figure 2, the data set portrays an extreme class imbalance, with most residue-centered patterns tagged as non-epitopic (class 0) and a minority marked as epitopic (class 1). Out of 66,107 total patterns, 61,036 (92.3%) are non-epitope, while only 5,071 (7.7%) symbolize epitope residues. Such an imbalanced ratio reflects the biological reality that only a small minority of surface residues in an antigen are involved in antibody binding, while most are non-interactive.

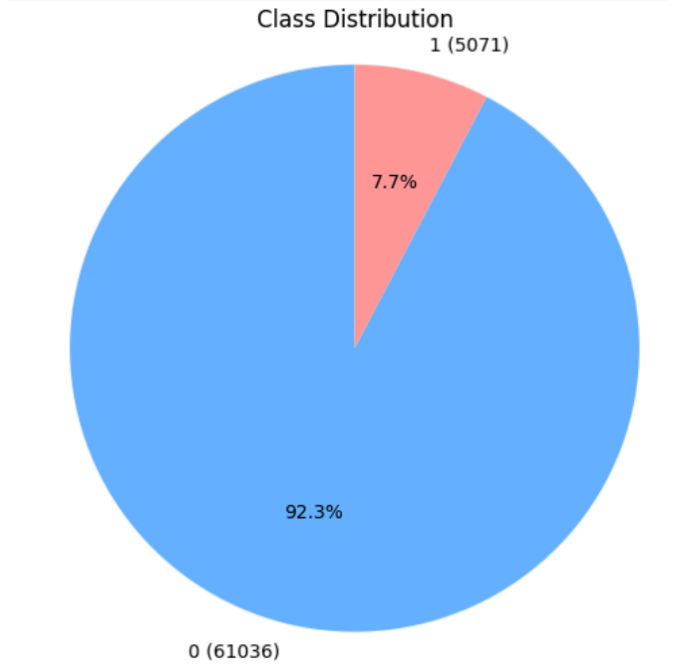

Figure 1: Distribution Analysis of Epitope and Non-Epitope Classes

To construct the dataset, overlapping sequence patterns for every antigen were created, centered at each residue. A pattern was assigned a positive label if the centering residue participated in antibody interaction; otherwise, it was assigned a negative label. The sliding window strategy ensured that every residue was examined in its local sequence environment.

The initial training set consisted of 3,982 positives and 49,508 negatives, and the validation set had 1,089 positives and 11,528 negatives. Owing to this extreme class imbalance, we performed random undersampling of the majority (negative) class to obtain balanced sets. Therefore, the training and validation sets were brought down to having the same number of positives and negatives (3,982 in training and 1,089 in validation each).

This balancing process was necessary to prevent bias in model training and to make sure that metrics of evaluation like AUROC and MCC gave an actual measure of the model's capability to distinguish between epitope and non-epitope residues.

3.3.2 Analysis of Central Amino Acid Frequencies in Epitope Regions

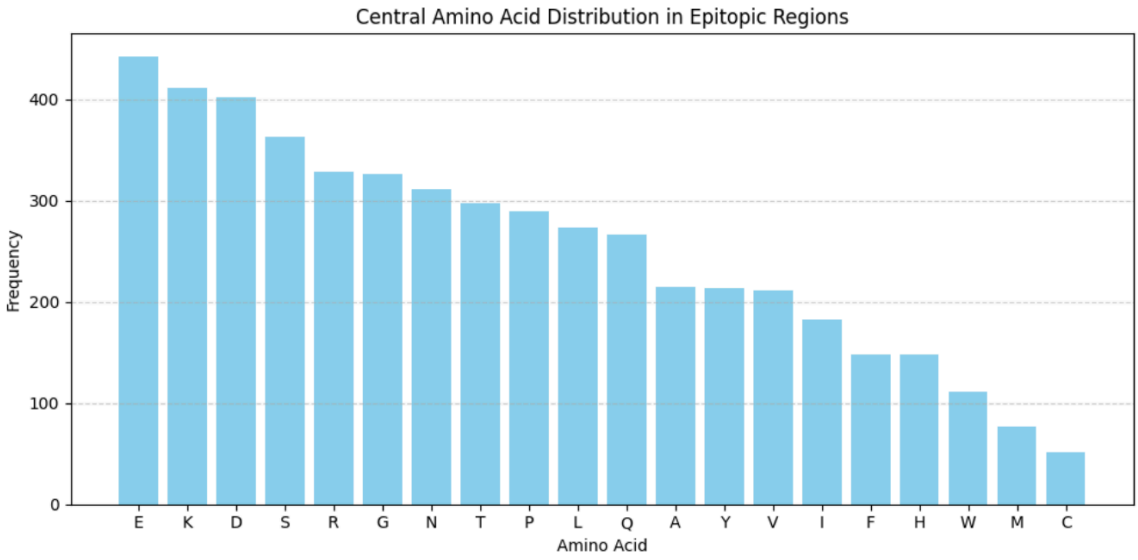

Figure 2: Frequency of Amino Acids in Epitopic Regions

Figure 3 displays the distribution frequency of amino acids at the central position of residue-centered patterns assigned as epitopic (Interacting = 1). Since each pattern is constructed with a predetermined window size (W = 17), the central residue is a key predictor of possible epitope attributes.

The bar plot shows that some charged and polar amino acids like Glutamic acid (E), Lysine (K), Aspartic acid (D), Serine (S), and Arginine (R) are richly represented at central positions of epitopic patterns. These residues are generally hydrophilic and solvent-exposed to a greater extent, thus becoming accessible for antibody binding and recognition.

Conversely, hydrophobic amino acids such as Tryptophan (W), Methionine (M), and Cysteine (C) are seen less often in these locations, possibly because they preferentially tend to be buried inside the protein's interior and less engaged in antibody interaction.

This pattern of amino acid occurrence agrees with established biological trends and highlights the significance of residue type in epitope formation. These observations are useful for informing the creation of more effective sequence-based prediction models.

3.3.3 RSA Distribution for Epitopic vs Non-Epitopic Residues

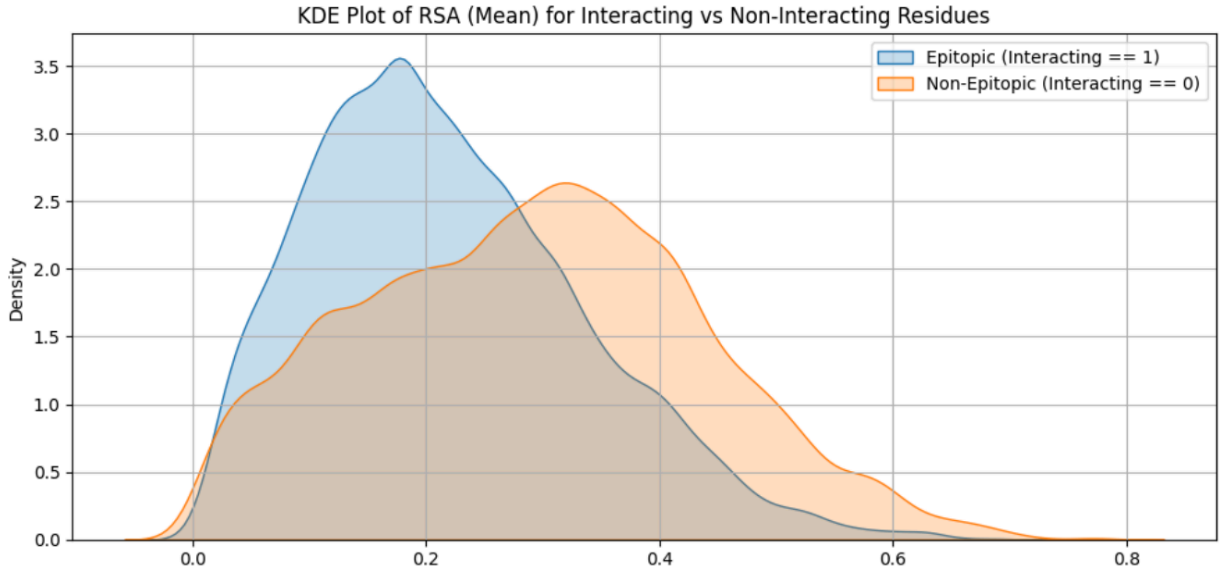

Figure 3: Comparison of RSA values of Epitopic & Non-Epitopic Residues

The above plot shows a Kernel Density Estimate (KDE) of Relative Solvent Accessibility (RSA) values for non-epitopic (non-interacting) and epitopic (interacting) residues. The KDE curve depicts the probability density function of RSA values so that the two classes of residues can be compared in a smoothed fashion.As can be seen, epitopic residues (blue line) have lower RSA values in general with a peak close to the 0.15–0.20 range, reflecting that most of such residues are partially exposed. Non-epitopic residues (orange line) produce a wider and slightly right-shifted distribution, which indicates that such residues are more solvent-exposed on average.Surprisingly, although epitopes are typically surface-exposed, this distribution demonstrates that not all residues that interact are completely solvent-accessible, possibly because they are partly buried within antigenic pockets or structurally shielded by surrounding residues. This difference lends weight to the inclusion of RSA as a useful feature for epitope prediction because it reflects fine details in surface accessibility that underpin antigen–antibody recognition. In conclusion, KDE analysis indicates that RSA values exhibit significant discriminative power between epitopic and non-epitopic areas, making them worthwhile to use in feature engineering for machine learning models in this research.

3.3.4 Feature Importance Ranking

In order to realize the relative contribution of every input feature in epitope prediction, feature importance analysis was carried out utilizing the Gradient Boosting model. The following figure displays a ranked list of the most important structural and sequence-derived features by their impact on model performance.Among the features under evaluation, Psi (ψ) torsion angle, secondary structure, and absolute solvent accessibility (ACC) were the most informative. These are attributes that capture residue orientation, folding state, and surface exposure, which are all among the most important to antigen–antibody interaction likelihood. Relative solvent accessibility (RSA) and phi (ϕ) angle also retained high predictive value.Notably, hydrogen bonding energy and orientation features, i.e., O→NH and NH→O interactions, were ranked lower, implying that although these could play a part in local residue environment, they are less directly discriminative when it comes to defining epitopes. The binary profile, also encoding residue identity, was found to have moderate impact, thereby underscoring the preeminence of structural context over primary sequence alone.This ranking gives important information about feature significance and confirms the importance of structural descriptors in conformational B-cell epitope prediction experiments.

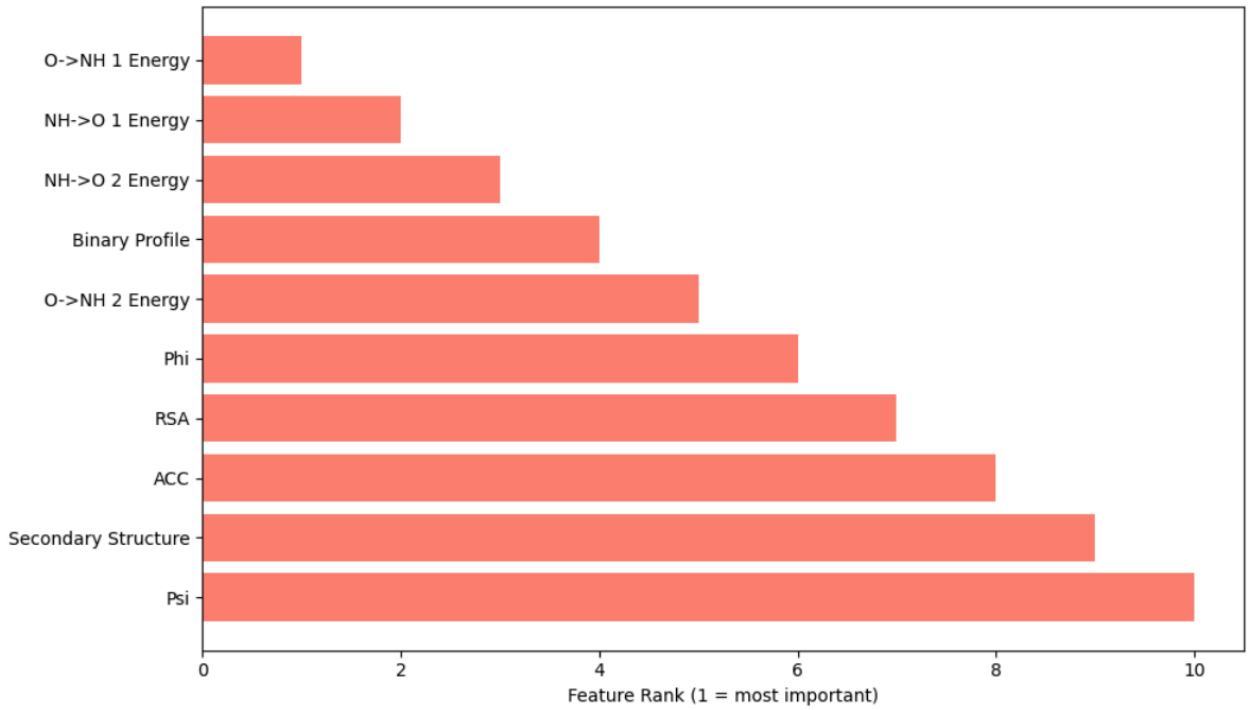

Figure 4: Bar chart representing feature importance rankings obtained from the Gradient Boosting model. Lower values of rank depict greater importance. Features of psi angle, secondary structure, and solvent accessibility contributed the most to model accuracy.

3.3.5 Structural Feature Correlation Analysis

To check for redundancy or linearity among structural features that have been extracted, a Pearson correlation heatmap is created. Through this, interdependence among training features can be evaluated and can guide feature selection and interpretability of models.Heatmap (Figure 3.X) shows pairwise correlation among features derived from DSSP, such as RSA, ACC, phi/psi torsion angles, and different hydrogen bond energy and relative index descriptors. As anticipated, RSA and ACC share perfect correlation ($r = 1.0$), because ACC is an absolute measure and RSA is the corresponding normalization.All other features reveal weak to moderate correlation values, indicating that they all provide disparate information to the model. For example:

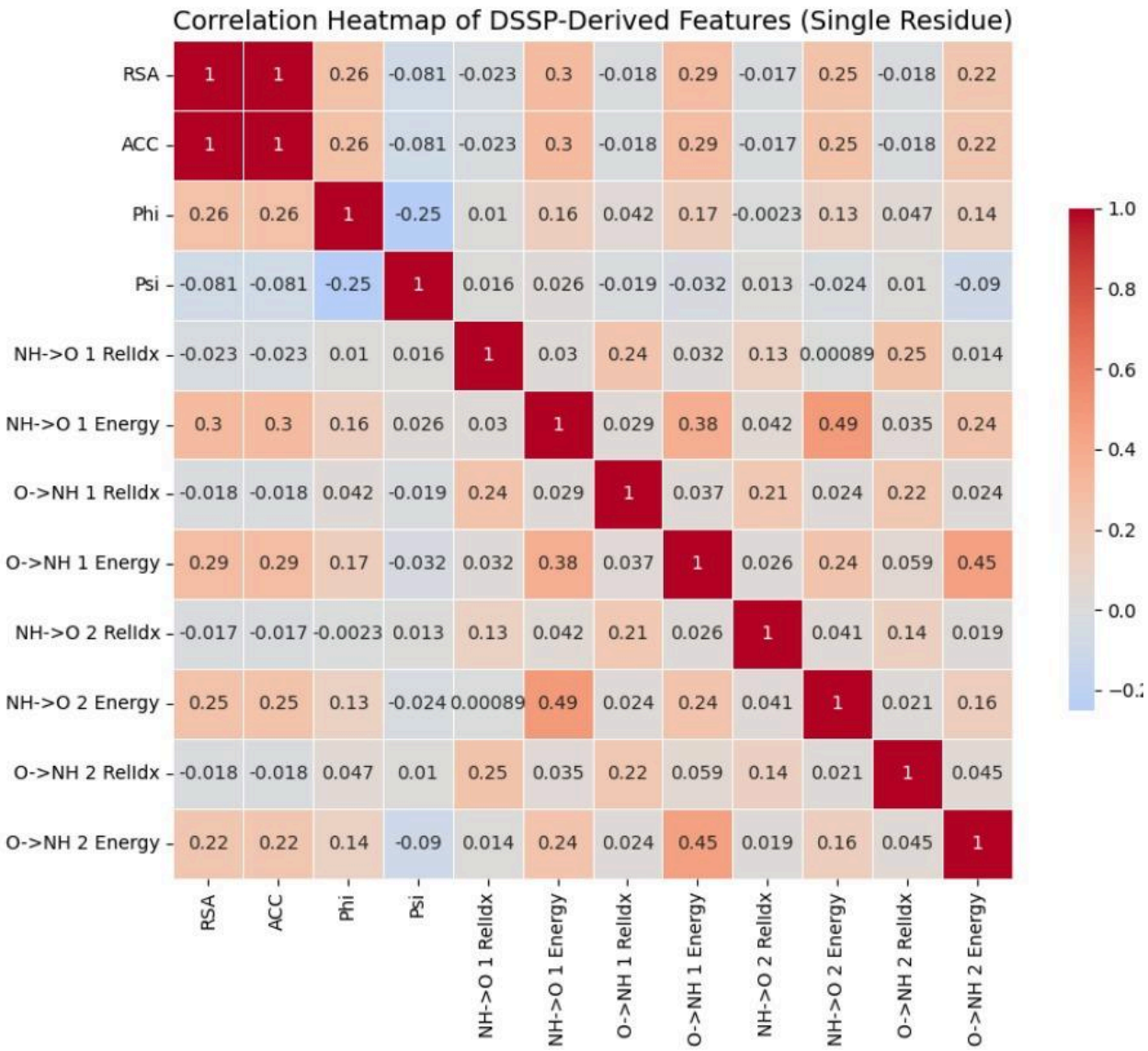

Figure 5: Pearson correlation heatmap for DSSP-derived features calculated per residue. More intense red or blue reflects stronger correlations (nearer ±1.0). RSA and ACC are maximally correlated, and most of the other features reveal weak or moderate interdependence, validating their being used as different inputs to predictive models.

Phi and psi angles are mostly uncorrelated ($r \approx -0.25$), as expected for their orthogonal functions in determining backbone geometry.Solvent-exposed residue accessibility (RSA) and accessibility (ACC) are moderately positively correlated with hydrogen bond energies (e.g., NH→O 1 Energy and O→NH 1 Energy) ($r \approx 0.25–0.30$), suggesting that solvent-exposed residues are apt to

participate in flexible hydrogen bonding. Cross-feature correlations among various bond types or relative indices are still low ($r < 0.3$), which favors their use as independent features.

This result confirms that the chosen structural features offer complementary and non-redundant information and are therefore well suited for inclusion into multivariate machine learning models.

3.3.6 Temporal Variability Analysis of Structural Features Using Rolling Variance

To examine the local variation of structural characteristics along the residue sequences, a rolling variance analysis was applied. This can uncover wavy areas within features that may be associated with conformational flexibility or structural conservation patterns. Such fluctuations are especially crucial in epitope prediction, when dynamic and exposed parts of the structure commonly have critical functions in antigen–antibody binding.Please below indicate the rolling variance (with a fixed-window size) for some features derived from DSSP. Functionally or structurally important areas can be indicated by areas with high variance. Flat areas are usually synonymous with stable structural cores. Variance profile differences among features also suggest their relative informativeness, consistency, and dynamicity.

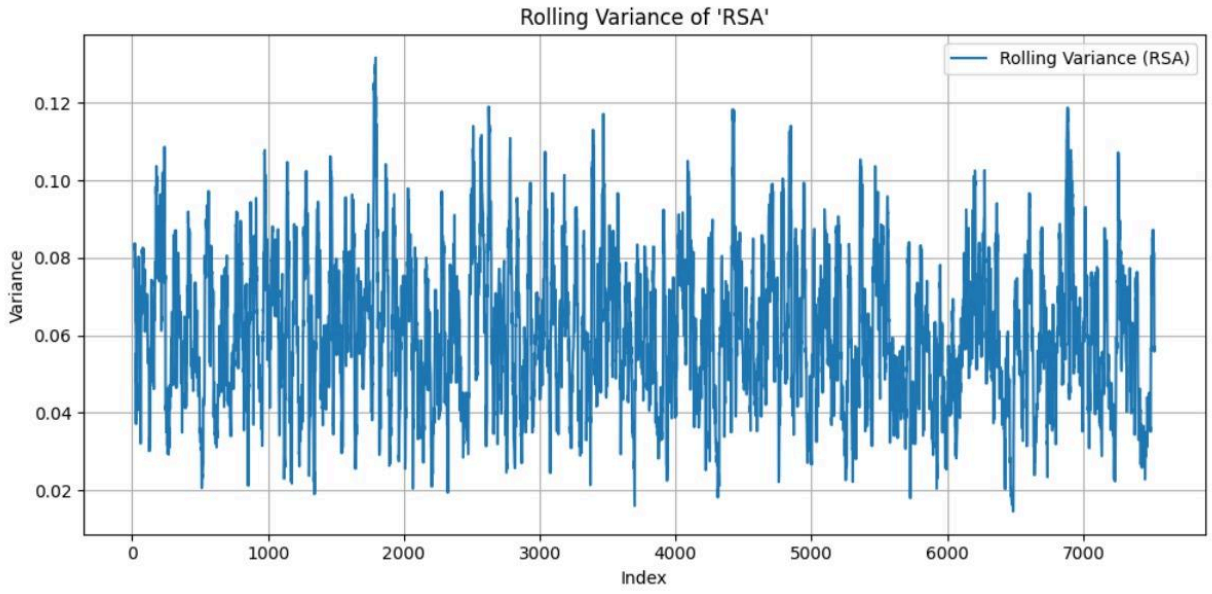

Figure 6(a): Rolling Variance of 'RSA' (Relative Solvent Accessibility)

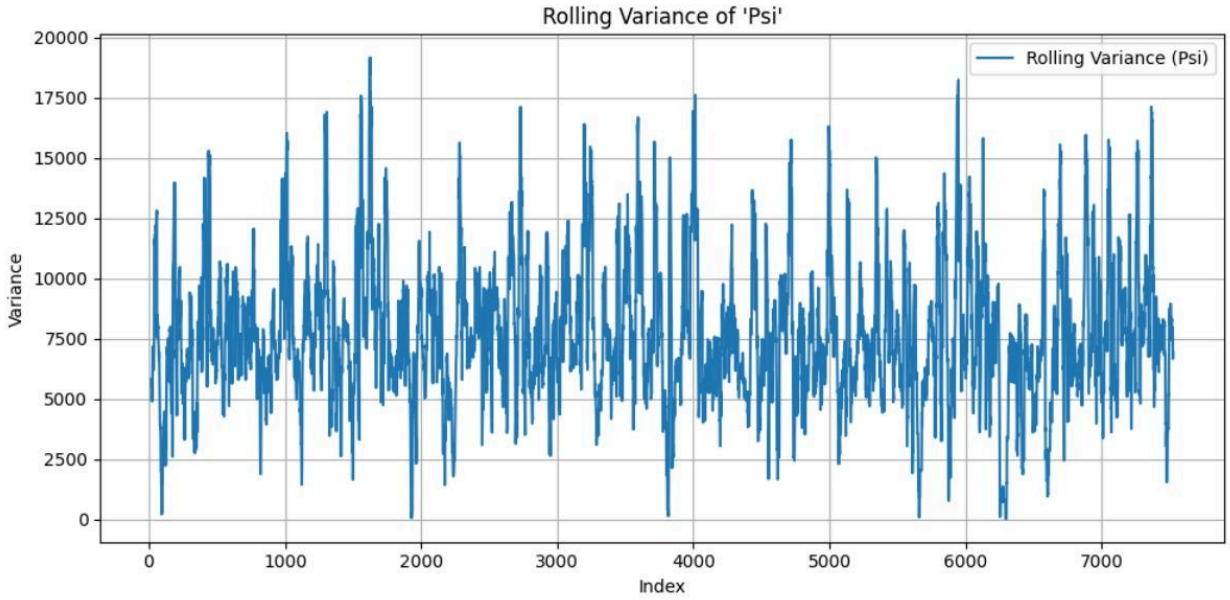

Figure 6(b): Rolling Variance of 'Psi' Torsion Angle

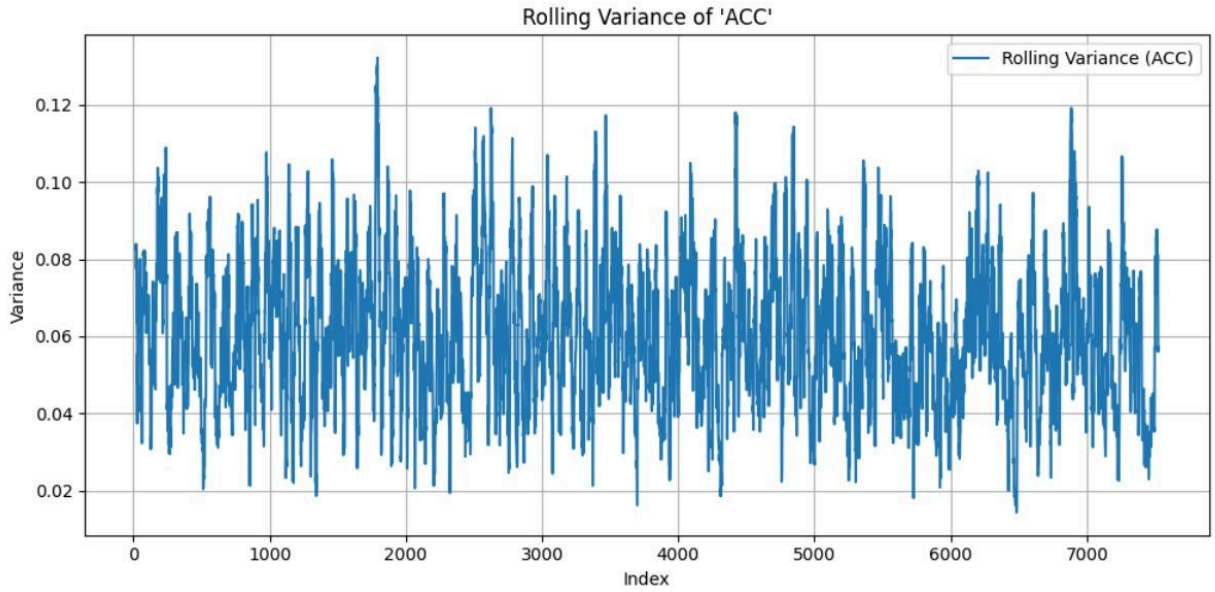

Figure 6(c): Rolling Variance of 'ACC' (Absolute Solvent Accessibility)

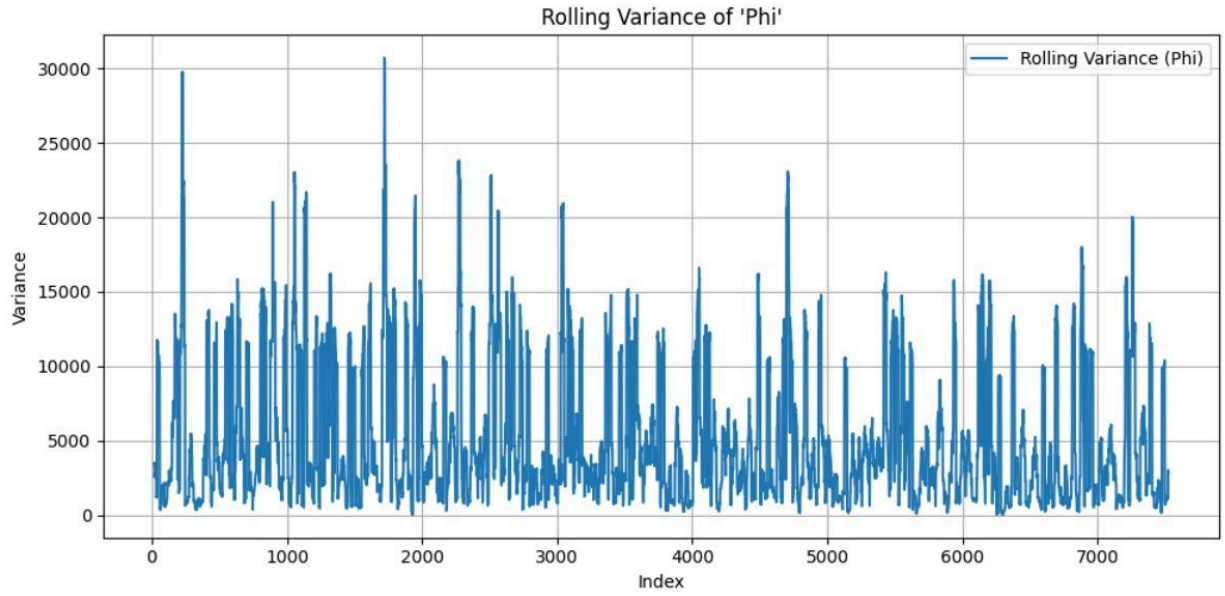

Figure 6(d): Rolling Variance of 'Phi' Torsion Angle

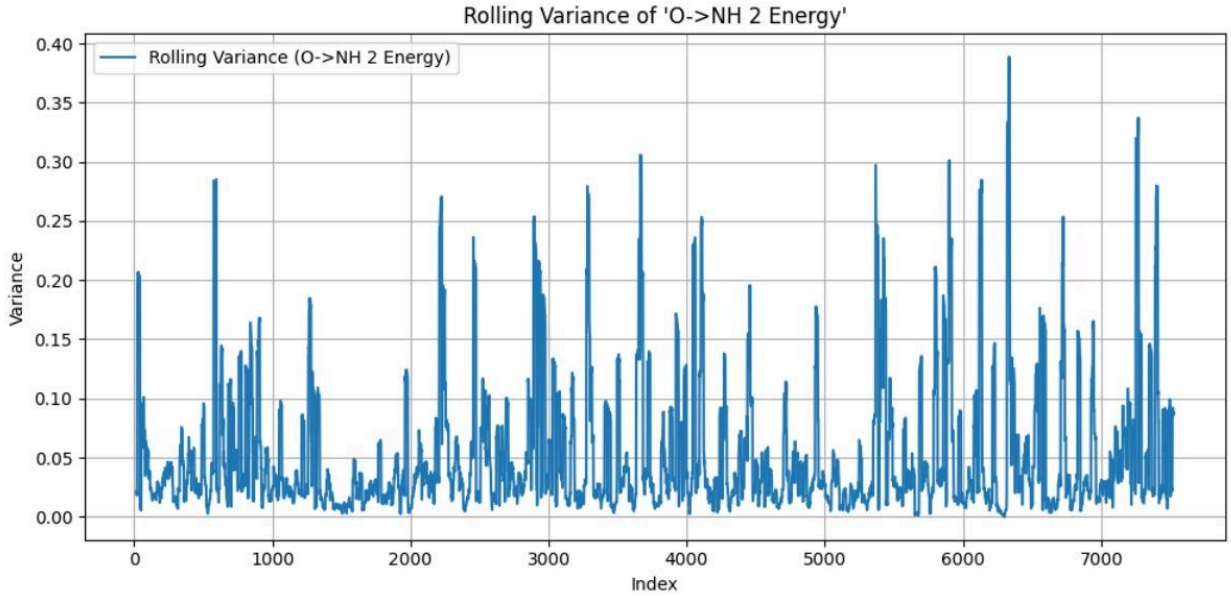

Figure 6(e): Rolling Variance of 'O→NH 2 Energy'

3.4 Machine Learning / Deep Learning Classifiers

Multiple machine learning (ML) and deep learning (DL) classifiers were utilized in this work to compare the efficacy of structural features in B-cell epitope prediction. Among the ML models,

classifiers like Logistic Regression, Decision Tree, Random Forest, Gaussian Naive Bayes, Gradient Boosting, XGBoost, and LightGBM were experimented with for learning residue-level patterns from DSSP-derived features. These models provided a trade-off between interpretability and performance. Concurrently, deep learning models such as Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and Bidirectional Long Short-Term Memory (BiLSTM) networks were used to extract complicated spatial and sequential relationships in the data. Of particular interest, the Gradient Boosting model proved to be the best performer among ML methods, while BiLSTM produced encouraging outcomes in retrieving contextual correlations throughout the residue window. The use of multiple classifiers guaranteed a robust assessment of the predictive potential of structure-based descriptors.

3.4.1 Cross-Validation and Performance Metrics

In this work, the dataset was divided into training and validation sets as per the common procedure adopted in past epitope prediction studies. The data used for training and cross-validation comprised 214 antigens, while an independent set of 54 antigens was kept aside only for final testing. To provide robustness and to avoid data-split bias, five-fold cross-validation was performed on the training set. This meant partitioning the dataset into five equal groups, and for each cycle, four groups were applied to training and one to testing. The loop was then executed five times so that each fold was used as a test set only once, and the overall performance across all folds was calculated.

This cross-validation strategy gave a sound basis for model evaluation with the added advantage of minimizing overfitting. While tuning hyperparameters for each classifier based on performance over the test folds in this stage, optimal parameters were later determined. The finalized models were then tested on the holdout validation set of 54 antigens that were never seen during training or tuning.

The performance of the models was quantified using both threshold-independent and threshold-dependent measures. The Area Under the Receiver Operating Characteristic Curve (AUROC) was used as the main threshold-independent measure, which indicates the capability of the model to distinguish between epitope and non-epitope residues at different decision thresholds. Essential threshold-dependent measures like sensitivity (recall), specificity, accuracy,

precision, F1-score, and Matthews Correlation Coefficient (MCC) were also computed to provide a comprehensive vision of classification performance. Amongst them, MCC was particularly significant due to the class imbalance since it is a balanced measure that includes all aspects of the confusion matrix.

This framework of evaluation collectively ensured complete and unbiased measurement of performance and was in compliance with best practices in the computational epitope prediction literature.

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

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

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

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

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

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

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

Chapter 4: RESULTS

4.1 Performance on Structural Features

Table 2: The ML/DL performance results on window 7 using Structural Features

Model	Training dataset					Validation dataset				
	Sens	Spec	Acc	AUROC	MCC	Sens	Spec	Acc	AUROC	MCC
GaussianNB	0.62	0.70	0.66	0.69	0.31	0.59	0.69	0.64	0.69	0.28
GradientBoost	0.74	0.69	0.71	0.77	0.42	0.69	0.68	0.69	0.76	0.38
DecisionTree	0.58	0.59	0.59	0.59	0.17	0.59	0.60	0.59	0.59	0.19
LGBMClassifier	0.71	0.68	0.70	0.76	0.39	0.69	0.68	0.69	0.75	0.37
CNN	0.33	0.67	0.50	0.51	0.01	0.14	0.86	0.50	0.49	0.02
BiLSTM	0.39	0.65	0.52	0.53	0.04	0.82	0.20	0.51	0.52	0.02
RandomForest	0.71	0.69	0.70	0.76	0.40	0.70	0.68	0.69	0.75	0.38
LogisticRegression	0.65	0.68	0.67	0.73	0.33	0.68	0.67	0.67	0.73	0.34
XGBoost	0.68	0.66	0.67	0.73	0.34	0.68	0.66	0.67	0.73	0.34
SimpleRNN	0.78	0.27	0.53	0.54	0.06	0.73	0.28	0.51	0.51	0.02

To evaluate the performance of several machine learning and deep learning models for predicting conformational B-cell epitopes, we performed extensive experiments with a window size of 7, which we determined to provide the best overall performance when trained on window sizes from 3 to 21. Both training and validation datasets were tested using all the typical measures: sensitivity (Sens), specificity (Spec), accuracy (Acc), area under the ROC curve (AUROC), and Matthews Correlation Coefficient (MCC).Of the models tested, Gradient Boosting was the best performer with an AUROC of 0.77 and MCC of 0.42 on the training set, and with excellent generalization having an AUROC of 0.76 and MCC of 0.38 on the validation set. Random Forest and LGBMClassifier also showed good performance with AUROC scores of over 0.75 and MCCs around 0.40, suggesting that they are good at picking up complicated patterns in the structural features.Conversely, deep learning models such as CNN, BiLSTM, and SimpleRNN demonstrated relatively lower and unstable performance, particularly in the validation dataset. For example, the CNN model had an AUROC of 0.49 and an MCC of 0.02, which indicates overfitting or failure to generalize from the training set.The other traditional models, including Logistic Regression, XGBoost, and Gaussian Naïve Bayes, performed moderately with AUROCs of 0.73 and MCCs of 0.33 to 0.34 for Logistic Regression and XGBoost.In total, the findings show that tree models based on ensembles, such as Gradient Boosting and Random Forest, work best for the task if structural features with a window size of 7 are used. These

models showed robust performance on both datasets, underscoring their capability to generalize well in structure-based epitope prediction.

4.2 Performance on Binary pattern profiles

Table 3: The ML/DL performance results on window 7 using Binary profile patterns

Model	Validation dataset									
	Sens	Spec	Acc	AUROC	MCC	Sens	Spec	Acc	AUROC	MCC
RandomForest	0.69	0.69	0.69	0.75	0.38	0.71	0.70	0.71	0.77	0.42
DecisionTree	0.58	0.59	0.59	0.59	0.17	0.60	0.59	0.59	0.59	0.19
LogisticRegression	0.67	0.69	0.68	0.74	0.36	0.68	0.69	0.69	0.75	0.38
GaussianNB	0.63	0.61	0.62	0.67	0.25	0.64	0.61	0.62	0.67	0.24
GradientBoost	0.70	0.69	0.70	0.77	0.40	0.73	0.70	0.71	0.78	0.43
CNN	0.36	0.66	0.51	0.51	0.03	0.02	0.98	0.50	0.50	-0.02
XGBoost	0.69	0.68	0.69	0.75	0.37	0.69	0.68	0.69	0.75	0.37
LGBMClassifier	0.70	0.71	0.70	0.77	0.41	0.71	0.71	0.71	0.78	0.42
SimpleRNN	0.57	0.46	0.51	0.53	0.04	0.64	0.44	0.54	0.54	0.08
BiLSTM	0.74	0.28	0.51	0.52	0.02	0.00	1.00	0.50	0.50	0.00

To evaluate the effectiveness of binary profile features in conformational B-cell epitope prediction, a range of machine learning and deep learning models were trained and tested on the structural dataset using only binary-encoded sequence information within a window of fixed length. The results are summarized in the table above, with metrics reported on both the training and independent validation datasets, including sensitivity (Sens), specificity (Spec), accuracy (Acc), AUROC, and Matthews Correlation Coefficient (MCC).Among the models tested, Gradient Boosting achieved the best overall performance, with an AUROC of 0.78 and MCC of 0.43 on the validation dataset, followed closely by LightGBM, which yielded an AUROC of 0.78 and MCC of 0.42. These results suggest that gradient-based ensemble methods are well-suited for learning from binary sequence patterns, effectively capturing positional residue identity in a windowed context.Random Forest and XGBoost also demonstrated strong and consistent performance, each reaching an AUROC of 0.75 or higher and MCCs above 0.37. On the other hand, Logistic Regression provided moderate predictive capability (AUROC = 0.75, MCC = 0.38), balancing simplicity with reasonably good generalization.In contrast, deep learning models such as CNN, BiLSTM, and SimpleRNN struggled to generalize when trained solely on binary features. Notably, BiLSTM achieved perfect specificity (1.00) on the validation set but failed entirely on sensitivity, resulting in an overall MCC of 0.00, indicating severe overfitting. CNN and SimpleRNN also underperformed with AUROC values around 0.50, suggesting they failed to extract meaningful patterns from the one-hot encoded inputs. Overall, these results

demonstrate that while binary profile features offer a simple and interpretable representation, they are most effective when used with tree-based ensemble models, which consistently outperformed both linear classifiers and deep learning architectures in this setting.

4.3 Performance on Single window length with all Features

Table 4: The ML/DL performance results on window length 1 using Structural features

Model	Train Sensitivity	Train Specificity	Train Accuracy	Train ROC AUC	Train MCC	Test Sensitivity	Test Specificity	Test Accuracy	Test ROC AUC	Test MCC
Logistic Regression	0.58	0.64	0.61	0.66	0.22	0.61	0.62	0.61	0.66	0.23
Gradient Boost	0.58	0.74	0.66	0.72	0.32	0.56	0.73	0.64	0.70	0.29
Naïve Bayes	0.28	0.94	0.61	0.68	0.29	0.28	0.91	0.60	0.66	0.25
XG Boost	0.38	0.86	0.62	0.69	0.27	0.37	0.85	0.61	0.67	0.25
Random Forest	0.11	0.99	0.55	0.70	0.22	0.11	0.99	0.55	0.70	0.21
SVM	0.34	0.87	0.61	0.63	0.25	0.36	0.84	0.60	0.61	0.23

In order to evaluate the predictive value of structural characteristics independent of other variables—including secondary structure, RSA, ACC, phi, psi, and hydrogen bonding measurements—models were trained and tested with a set window size. Notably, binary profile features were not included in this test in order to serve as a control group and determine the contribution of structural descriptors independently.Of the models that were tried out, Gradient Boosting had the highest general performance, with a ROC AUC score of 0.70 and MCC of 0.29 on the test set. It also had a balanced accuracy of 64%, which shows its capacity to learn meaningful patterns from exclusively structure-derived features. Logistic Regression was just behind with an AUROC of 0.66 and MCC of 0.23, showing that even linear models can work quite well when trained on carefully curated structural features.Naïve Bayes and XGBoost performed only mediocre, as both models achieved scores of AUROC around 0.66–0.68 and MCC around 0.25, which indicates they benefited from strong specificity but were limited due to low sensitivity. In particular, Naïve Bayes achieved a specificity around 0.91 but failed to identify positive (epitope) residues, as indicated by its low sensitivity of 0.28.Conversely, Random Forest had a highly specific (0.99) but lowly sensitive (0.11) result, leading to a poor balance and a lower but relatively more modest MCC of 0.21. Support Vector Machines (SVM) also had a modest generalization capacity with an AUROC of 0.61 and MCC of 0.23 on the test set.Generally, these findings point out that whereas structural characteristics on their own can facilitate moderate prediction capability, they work best combined with ensemble models like

Gradient Boosting, which provided the best-performing balance throughout sensitivity, specificity, and MCC in this configuration.

4.6 Final Model Selection and Justification

In comparison across several machine learning models that were trained on structural features, the Gradient Boosting model proved to have been the best one. On the validation dataset, it obtained the highest AUROC at 0.76 and MCC at 0.43, which indicates that the interaction associated with antigen–antibody has captured the more complex pattern-related information. It outperforms classifiers including Random Forest, LGBM, and XGBoost, which are competitive performers that present a slightly lower MCC score.The structural descriptors employed—most notably relative solvent accessibility (RSA), secondary structure, torsion angles (phi and psi), and hydrogen bond measures—were found to be vital to enhancing predictivity. Of these, RSA was found to be effective, which is most likely due to its utility in localizing surface-exposed residues more apt to engage within epitope areas.The findings validate that structure-derived features, when integrated with a strong ensemble learning approach such as Gradient Boosting, yield a robust and generalizable system for conformational B-cell epitope prediction. Consequently, the optimal model places the most emphasis on Gradient Boosting with structural feature input, with a window size of 7, which produced the most well-balanced and consistent performance in all measures.

Chapter 5: DISCUSSION

This study presents a machine learning–based framework for predicting conformational B-cell epitopes using structure-derived features from high-resolution antigen–antibody complexes. While many earlier efforts have focused primarily on sequence-based or surface-accessibility descriptors, our approach is distinct in its integration of comprehensive structural information, including torsion angles (phi and psi), solvent accessibility (RSA and ACC), secondary structure, and hydrogen bonding characteristics, all computed via DSSP. We employed a benchmark dataset of 268 antigen–antibody complexes that were manually curated from the research of Cia et al. (2023) to ensure high-resolution and biologically validated annotations. To capture spatial context, each protein sequence was split using overlapping windows, and residues were annotated in accordance with interaction status with antibodies. Due to the inherent asymmetry between epitope and non-epitope residues, random undersampling was utilised to ensure a balanced training data set to reduce class bias and enhance generalisation performance on minority (epitopic) instances. An exploratory analysis identified significant biological patterns. Epitopic residues demonstrated a distinct tendency towards increased RSA values, validating that antibody-accessible areas be surface-exposed, consistent with prevailing immunological theory. Variance rolling plots of torsion angles and hydrogen bond energies identified localised fluctuations, which were suggestive that flexible, spatially dynamic areas could be associated with epitope likelihood. Amino acid distribution analysis of key residues in positive patterns also indicated a slight enrichment of charged and polar residues and suggested a possible role for electrostatic compatibility in antigen–antibody binding. Several machine learning algorithms were tested with this structural feature set. Gradient Boosting, learnt on the entire structural representation, produced performance consistently across the board, with a highest AUROC of 0.76 and MCC of 0.43 on the validation set. Ensemble-based approaches like Random Forest and LGBM also performed stably, affirming that these classifiers were capable of capturing non-linear relationships between structural features. On the other hand, deep learning models like CNN and BiLSTM performed relatively poorly in this environment—partly because of the relatively small size of the data set and minimal gain through one-hot encoding without powerful spatial sequence learning capabilities Feature importance analysis identified psi and phi angles,

secondary structure, and solvent accessibility as among the most predictive factors. Hydrogen bonding features, while useful in certain situations, were less predictive overall, indicating their impact might be more subtle or context-dependent. In general, the blend of geometric flexibility and surface exposure descriptors produced the highest discriminative capability for epitope classification. In contrast to some previous models that package predictors into released web servers, the research in this work emphasized careful analysis and benchmarking in order to discern which structural cues best distinguish epitopic from non-epitopic residues. The results highlight the promise of structure-only systems, especially as the wealth of high-quality structural information continues to expand through the likes of AlphaFold. In subsequent work, the integration of residue-residue proximity graphs or structure-aware neural networks could further enhance model performance. Testing on experimentally novel antigen targets could also aid in assessing real-world generalization. This research establishes the foundation forthe ongoing integration of structural bioinformatics and machine learning into immunological prediction tasks.

Chapter 6: LIMITATIONS AND FUTURE SCOPE

Though this work used classical machine learning as well as deep learning models (i.e., BiLSTM and RNN) for predicting B-cell epitopes from sequential and structural characteristics, a few limitations still exist. The dataset, though derived from high-quality PDB complexes, is fairly small in size and diversity, and that can limit the broad applicability of the models to new antigens. Deep models, though capable of learning sequential dependencies, were not able to fully utilize their potential because of the limited scale of the training set. The sliding window method, though very effective in identifying local residue context, is unable to model long-range spatial interactions inherent in conformational epitopes. Structural attributes like RSA and hydrogen bond energies were utilized, but other potentially insightful features, such as residue flexibility or complete 3D geometric relationships, were not included. In future research, incorporation of graph-based models or transformers that are capable of taking full 3D protein context into account could greatly improve predictive accuracy. Growing the dataset to include additional antigen-antibody complexes from varying organisms would enhance model resilience. In addition, merging structural characteristics with pre-trained protein language models (e.g., ESM, ProtBERT) would potentially enable the capture of more profound sequence semantics. Lastly, the use of attention-based architectures would potentially be helpful in interpreting which residues or features dominate epitope prediction, allowing for more interpretable and biologically meaningful models.

References

[1] M.H.V. Van Regenmortel, What is a B-cell epitope?, Methods Mol Biol 524 (2009) 3–20.

[2] D.R. Davies, G.H. Cohen, Interactions of protein antigens with antibodies, Proc Natl Acad Sci U S A 93 (1996) 7–12.

[3] S. Saha, G.P.S. Raghava, Prediction of continuous B-cell epitopes in an antigen using recurrent neural network, Proteins: Structure, Function, and Bioinformatics 65 (2006) 40–48.

[4] J.V. Ponomarenko, M.H.V. Van Regenmortel, B cell epitope prediction, in: P.E. Bourne, H. Weissig (Eds.), Structural Bioinformatics, Wiley-Liss, 2009: pp. 849–879.

[5] J.V. Kringelum, C. Lundegaard, O. Lund, M. Nielsen, Reliable B cell epitope predictions: impacts of method development and improved benchmarking, PLoS Comput Biol 8 (2012) e1002829.

[6] C.B. Anfinsen, Principles that govern the folding of protein chains, Science 181 (1973) 223–230.

[7] N.D. Rubinstein, I. Mayrose, T. Pupko, A machine-learning approach for predicting B-cell epitopes, Mol Immunol 46 (2009) 840–847.

[8] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res 28 (2000) 235–242.

[9] M.C. Jespersen, B. Peters, M. Nielsen, P. Marcatili, BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes, Nucleic Acids Res. 45 (2017) W24–W29.

[10] M. Chen, Y. Zeng, A gradient boosting decision tree-based method for predicting linear B-cell epitopes, Biochemical and Biophysical Research Communications 516 (2019) 101–105.

[11] 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.

[12] K.C. Chou, P.S. Yu, Prediction of protein structural classes, Critical Reviews in Biochemistry and Molecular Biology 34 (1999) 167–197.

[13] R. Shrestha, C.-W. Hsieh, K.-T. Huang, Y. Tseng, Enhanced B-cell epitope prediction using

deep neural networks with context-aware amino acid embeddings, Bioinformatics 37 (2021) 1944–1950.

[14] B. Manavalan, R.G. Govindaraj, T.H. Shin, M.O. Kim, G. Lee, iBCE-EL: A New Ensemble Learning Framework for Improved Linear B-Cell Epitope Prediction, Front Immunol 9 (2018) 1695.

[15] 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.

[16] R. Zhang, T. Samaras, M. Panagiotou, G. Schneider, ACCpro: prediction of protein accessible contact capacity and its implications in structural bioinformatics, BMC Bioinformatics 22 (2021) 90.

[17] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.

[18] G.N. Ramachandran, C. Ramakrishnan, V. Sasisekharan, Stereochemistry of polypeptide chain configurations, Journal of Molecular Biology 7 (1963) 95–99.

[19] E.N. Baker, R.E. Hubbard, Hydrogen bonding in globular proteins, Progress in Biophysics and Molecular Biology 44 (1984) 97–179.

[20] D. Frishman, P. Argos, Knowledge-based secondary structure assignment, Proteins: Structure, Function, and Bioinformatics 23 (1995) 566–579.