3-23-2022

In silico Identification of Vaccine Candidates against Viral Infections

Prabin Baral
pbara006@fiu.edu

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

IN SILICO IDENTIFICATION OF VACCINE CANDIDATES AGAINST VIRAL INFECTIONS

A dissertation submitted in partial fulfillment of

the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PHYSICS

by

Prabin Baral

2022

To: Dean Michael R. Heithaus
College of Arts, Sciences and Education

This dissertation, written by Prabin Baral, and entitled *In silico* Identification of Vaccine Candidates Against Viral Infections, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

_____
Jin He

_____
Yuan Liu

_____
Prem P. Chapagain, Co-Major Professor

_____
Bernard S. Gerstman, Co-Major Professor

Date of Defense: March 23, 2022

The dissertation of Prabin Baral is approved.

_____
Dean Michael R. Heithaus
College of Arts, Sciences and Education

_____
Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University, 2022

DEDICATION

I dedicate this work to my parents.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisors Professor Bernard S. Gerstman and Professor Prem P. Chapagain for their invaluable guidance and motivation throughout the years of my Ph.D. program that provided me necessary training to pursue my research. I am very grateful to my dissertation committee members Professor Jin He and Professor Yuan Liu for their helpful suggestions and comments. I am thankful to the Department of Physics Faculty and Staff members for systematizing all the official tasks. I would like to acknowledge the past and present members of the Biophysics Research Group Elumalai, Rudra, Nisha, Lokman, Tej, Michael, and Hugo for the friendly lab environment as well as help in my research projects.

I am thankful to the COVID-Informatics research team at FIU for the research assistantship and collaboration: Professor Giri Narasimhan, Professor Kalai Mathee, Professor Ananda Mohan Mondal, Professor Jessica Siltberg-Liberles, Professor Trevor Cickovski as well as my peers Vitalii Stebliankin, Janelle Nunez Castilla, Masrur Sobhan and Christian Balbin. Special thanks to Ziyou Zhou from the Luna labs and the U.S. Army Contracting Command – Aberdeen Proving Ground – Natick Contracting Division for the collaboration and computational support during the epitopes predictions in the Marburgvirus project. Thanks to the CASTIC, as well as FIU HPC team for the computer support. I would like to thank the University Graduate School at Florida International University for providing me

with a Dissertation Year Fellowship Award that helped me concentrate on writing my dissertation as well as completing my research project.

Finally, I would like to thank my parents for their support and patience during my Ph.D. journey. Thanks to my beloved wife Lila for all the love, care, and understanding.

ABSTRACT OF THE DISSERTATION

IN SILICO IDENTIFICATION OF VACCINE CANDIDATES AGAINST VIRAL INFECTIONS

by

Prabin Baral

Florida International University, 2022

Miami, Florida

Professor Bernard S. Gerstman, Co-Major Professor

Professor Prem P. Chapagain, Co-Major Professor

There are many viral diseases without effective treatments or vaccines. A critical step in the immune system's fight against viruses involves an immunological protein molecule binding to a viral protein molecule. I investigate the atomic and molecular details of binding site recognition and binding interactions and dynamics for three important viruses.

Antigens are molecules, such as viral proteins, that are foreign to the human body and can generate an immune response such as the production of antibody proteins to attack the antigen. An epitope is the part of an antigen molecule that is the site for antibody binding. They are categorized as T-cell or B-cell epitopes based upon which type of immunological cell can bind to the epitope. The identification of epitopes is an essential step for the discovery and development of epitope-based vaccines. Experimental identification of epitopes involves

expensive and time-consuming steps and therefore *in silico* identification is a powerful tool to facilitate the identification of potential epitope candidates and can decrease the time and expense spent on validation experiments.

I employed several epitope computational prediction methods that are based upon the antigen protein's amino acid sequence and conformation for the glycoprotein of the Lassa virus as well as for different proteins of the Marburg virus. The predicted epitopes are further filtered based on a consensus approach that resulted in the identification of new epitopes that have not yet been tested experimentally. I performed molecular dynamics computational simulations on the most promising epitopes to determine atomic-level details of the epitope's interactions and dynamics.

In addition, I performed MD simulations to investigate the dynamics and antibody evasion behavior by the B.1.617.2 (delta) variant of SARS-CoV-2. I found that the receptor-binding β-loop-β motif in the spike protein adopts an altered conformation that causes binding difficulty for some of the neutralizing antibodies that were generated against the original coronavirus strain. This study reflects the possible mechanism for the immune evasion exhibited by the delta variant.

ABBREVIATIONS AND ACRONYMS

CHARMM	Chemistry at Harvard Molecular Mechanics
LJ	Lennard-Jones
MD	Molecular Dynamics
NAMD	Not Another Molecular Dynamics
PDB	Protein Data Bank
SASA	Solvent Accessible Surface Area
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
MHC	Major Histocompatibility Complex
HLA	Human Leukocyte Antigen
%	Percentage
ns	Nanoseconds

TABLE OF CONTENTS

LIST OF TABLES

TABLES PAGE

LIST OF FIGURES

FIGURES PAGE

1. INTRODUCTION

1.1 Viral Infections

There are many viral diseases without effective treatments or vaccines. These viruses can cause catastrophic epidemics such as the Lassa, Ebola, and Marburg viruses. Similarly, the recent coronavirus pandemic is of great concern as new variants are continuously emerging with decreased susceptibility to antibodies and vaccines that were developed for earlier strains. A critical step in the immune system fight against viruses involves an immunological protein molecule binding to a viral protein molecule. I investigate the atomic and molecular details of binding site recognition and binding interactions and dynamics for three important viruses.

Viruses are infectious agents that typically consist of genetic material (DNA or RNA) encapsulated in a protein coat. Viruses are able to multiply by using the machinery within the living cells of a host[1]. A person is susceptible to different viruses which can enter or transmit through various routes such as the mouth, nose, eyes, genitals, as well as wounds. Once Inside a cell, the viral DNA or RNA starts replicating and can make millions of copies of its genetic material and protein shell. These copies leave the infected host cell and then migrate to uninfected cells to repeat the process. The process of replication can lead to pathogenesis if the immune system fails to control the virus[2].

Viruses are dangerous for a variety of reasons. Unlike bacteria, fungi, and parasites[3], viruses are too small to be visible under a standard optical microscope. Viruses have different modes of transmission that could range from contaminated

surfaces, airborne, or contaminated body fluid. Worldwide, different viruses have caused a substantial amount of disease and death. Examples include respiratory viruses such as influenza virus, rhinovirus, adenovirus, coronavirus, etc., as well as the bodily fluid and blood-borne viruses such as Lassa, Ebola, Marburg, hepatitis virus and Human immunodeficiency virus (HIV)[4]. Currently, there is a lack of effective treatment methods against most of these infections. Thus, the identification of effective methods to contain such viruses is of extreme importance.

The development and deployment of therapeutic and vaccine candidates against such viral infections is of significant urgency as the twenty-first century has witnessed the emergence of different viruses that have already caused several epidemics or pandemic[5] including the ongoing COVID-19 pandemic. Vaccines are made by using different processes in which they might contain attenuated live viruses (eg. MMR vaccines, Varicella, Influenza, Rotavirus vaccines, etc), inactivated viruses as in Polio and Hepatitis A, inactivated toxins against bacterial diseases such as Diptheria, Tetanus, etc. Other vaccines against viruses such as Hepatitis B, Influenza, Pneumococcal, Meningococcal are developed by utilizing segments of pathogens which include subunits as well as conjugate vaccines[6,7] that pair a weak virus with a strong virus.

1.2 Epitopes

To create a vaccine, I focus on epitopes, which are parts of an antigen. Antigens are molecules, such as viral proteins, that are foreign to the human body and can generate an immune response such as the production of antibody proteins to

attack the antigen. An epitope is the part of an antigen molecule that is the site for antibody binding. They are categorized as T-cell or B-cell epitopes based upon which type of immunological cell can bind to the epitope. The B cell epitopes are surface segments of the antigen protein that are recognized by the immunological B-cells, whereas T-cell epitopes are peptides derived from antigens and recognized by immunological T-cells when the epitope is bound to a major histocompatibility complex molecule. A vaccine can contain either a weakened virus or an antigen molecule from the virus. It is the few amino acids of the antigen in the form of epitopes that interact with the immune system[8].

Recently, on the basis of computational approaches that utilize advanced algorithms and epitope databases, many new epitope-based vaccine candidates have been proposed[9,10]. These epitopes are classified into two types based on the immunological cells (lymphocytes) that the body generates to attack the antigen. The lymphocytes that mature in the thymus are called T-cell lymphocytes, and epitopes recognized by them are referred to as T-cell epitopes. The lymphocytes that mature in the bone marrow are B lymphocytes and the epitopes that they recognize are called B-cell epitopes[11]. There is further classification into two categories: linear and conformational epitopes. Linear epitopes are a contiguous segment of amino acids on an antigen, while conformational epitopes are amino acids on the antigen that are not contiguous in the primary peptide sequence but are brought together by the structural arrangement in a folded protein[12,13]. T-cells contain proteins, CD4 or CD8, that can bind to proteins on other cells. CD4 receptors bind to a cell's major histocompatibility complex molecule (MHC-II

molecule) and help in the activation of CD8 cells and B-cells, while CD8 receptors bind to MHC-I molecules, which then acts in the removal of infected cells and pathogens[14,15].

Since an epitope is only a small but unique segment of an antigen, an epitope-based vaccine has fewer side effects compared to conventional vaccines. For the immune system to be effective, the cell's MHC molecules must bind to an epitope on the antigen. Since MHC molecules are polymorphic with a range of conformations and amino acid variations, the efficacy of a vaccine will increase if we can identify epitopes that can bind to several different types of MHC molecules (promiscuous epitopes). Experimental identification of promiscuous epitopes involves many expensive and time-consuming steps, including the production of antibodies to map antigenic regions on a target protein, animal models, and determination of the crystal structure of antigen-antibody complexes using X-ray crystallography. To narrow down the set of possible epitopes, computational identification of epitopes is employed as a powerful and fast approach to facilitate the identification of potential epitope candidates that can decrease the number of validation experiments[16,17]. In addition to decreasing the experimental cost of searching for epitopes, computational investigations also provide molecular and atomic level details of the interactions and dynamics of epitope candidates that are not accessible with experimental methods. These computational investigations have accelerated epitope-based vaccine development for several viral infections and cancer[18,19].

1.3 Investigating epitopes of Lassa virus, Marburg virus, and Coronavirus

My work focuses on the discovery and investigation of T and B cell epitopes for the Lassa virus, Marburg virus epitopes, as well as the molecular and atomic dynamics that allow antibody evasion by the SARS-CoV-2 spike protein antigen. Biophysical and biochemical databases are used to identify potential viral protein epitopes. The interaction of identified epitopes with the different versions (alleles) of MHC proteins is then investigated by molecular docking as well as Molecular Dynamics (MD) simulations studies to determine the epitope promiscuity. The calculations from MD simulations study the relative stability of epitope-allele complexes as well as the identification of novel epitopes. In addition to the Lassa virus, I also study the dynamics of the SARS-CoV-2 spike protein by using MD simulations. The dynamics and flexibility of the spike protein receptor binding domain (RBD) of different variants are explored and binding of antibodies to the spike protein are also studied. This study reveal that mutations in an amino acid loop in the spike RBD is responsible for the immune escape mechanism of the delta variant and thereby increasing the number of COVID-19 cases[20].

2. VIRAL SYSTEMS INVESTIGATED

I investigated the molecular details of proteins encoded by three different viruses, Lassa, Marburg, and the B.1.617.2 delta variant of SARS-CoV-2. These three systems were chosen because they have caused current or recent epidemics, and the crystallographic structure of the proteins are known.

2.1 Lassa Virus Disease

Lassa virus (LASV), a member of the *Arenaviridae*[21], is an ambisense RNA virus that causes a severe hemorrhagic Lassa fever in humans. LASV is endemic, particularly in the West African countries of Sierra Leone, The Republic of Guinea, Nigeria, and Liberia[22,23]. The transmission of LASV to humans occurs through the urine or feces of infected Mastomys rats and the virus spreads human-to-human through direct contact with the blood, urine, feces, or other bodily secretions of an infected person. LASV can be fatal and no approved effective therapeutics are currently available. The development of therapeutics such as antibodies and vaccines for the treatment of LASV is therefore of significant urgency[24,25,26].

Of the four proteins that are encoded by the two RNA segments of the LASV genome, the glycoprotein (GP) is the only protein on the viral surface. GP results from the cleavage of a 75 kDa precursor polypeptide, GPC by signal peptidase and then further glycosylated and processed into GP1 and GP2[27]. GP1 is the receptor-binding subunit, and GP2 is the membrane-spanning fusion subunit[28,29,30]. The virion envelope protein spikes are composed of three heterotrimers, with each heterotrimer containing signal peptide, GP1, and

GP2[31,32], shown in Figure 2.1. A chalice-like GP trimer interacts with receptors on the cell surface, for example matriglycan, which mediates the entry of the virus into the host cell. In addition, the GP also interact with ERGIC-53 in the exocytic pathway, which helps to form infectious virions[33]. GP is considered to be a key factor for LASV growth, cell tropism, host range and pathogenicity, and as it is the only protein situated on the LASV virion surface, GP becomes a primary target for vaccine design[24].

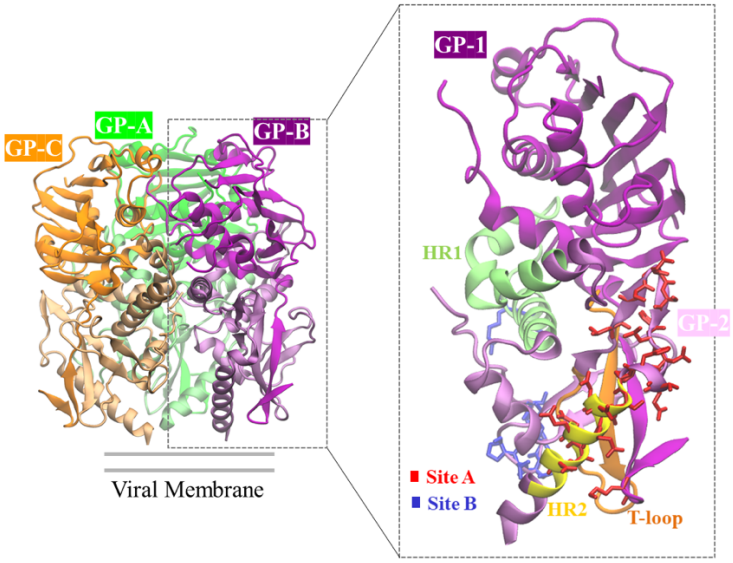


Figure 2.1 Structure of the LASV GP trimer consisting of the three GPs (GP-A, GP-B, GP-C). Each GP has a GP1 subunit and a GP2 subunit (zoomed view). Each monomer is colored differently in the GP trimer. In the zoomed view, the GP2 subunit is lightly shaded to differentiate from the GP1 subunit, and some of the antibody binding sites are highlighted (figure generated from the crystal structure of the LASV GP in the Protein Data Bank[34], PDB ID: 5VK2[24]).

The crystal structure of the trimeric LASV GP in complex with the 37.7H neutralizing antibody from a human survivor (PDB ID: 5VK2, Figure 2.1) has been determined, thereby providing insight into the structural basis for antibody design.

Analysis of the GP-37.7H antibody complex shows that the antibody simultaneously binds to two GP monomers at the base of the GP trimer. The binding involves four discontinuous regions of LASV GP: two in site A and two in site B. Site A contains residues 62 and 63 of the N-terminal loop of GP1 and residues 387 to 408 in the T-loop (residues 365-384) and HR2 (residues 400-412) regions of GP2. Site B contains residues 269 to 275 of the fusion peptide and residues 324 to 325 of HR1 (residues 311-355) of GP2[24,35]. Although the antibody predominantly binds to GP2, GP1 is required to maintain the proper prefusion conformation of GP2 for antibody binding[24].

2.2 Marburg Virus Disease

Filovirus infections in humans can cause severe hemorrhagic fever in humans with other symptoms such as vomiting, cough, diarrhea, jaundice, and result in high fatalities[36,37]. Of the different genera of Filovirus family, Marburgvirus (MARV) is the deadliest infectious agent and has reemerged a multiple times since its discovery in 1967[38]. Bat is considered the primary source of this disease and it is transmitted to humans through direct contact, droplets of body fluids from infected persons, or contact with equipment and other objects contaminated with infectious blood or tissues. However, currently there are currently no approved vaccines or post exposure treatment methods available[36,37]. Filoviruses are enveloped virions (Figure 2.2a), which express glycoprotein (GP), nucleoprotein (NP), VP35, VP40, VP30, VP24, and large polymerase protein (L). The GP is situated on the virions' surface and helps enter into the target cells, and is therefore considered a primary

target for antibodies and vaccines design[36,37,39]. A 3.6 Aº resolution crystal structure of trimeric MARV surface glycoprotein complexed with human antibody MR78 (PDB ID: 5UQY, Figure 2.2 b) has been determined[40]. The GP1 has the receptor-binding site (RBS) that is responsible for the host receptor binding[41,42]. The GP2 anchors the membrane and catalyzes membrane fusion and viral entry[43]. Nucleoprotein (NP) binds directly to the viral genome and plays an important role in the replication of the genome and nucleocapsid formation[44,45]. It is considered as a suitable target for vaccine development because of its abundance in Filovirus infected cells and its strong antigenicity[46,47]. The crystal structures of MARV NP in both apo and VP35-chaperoned form have been reported (PDB ID: 5F5M[48], Figure 2.2c). The NP apo form exists as hexamers while the VP35 chaperoned NP remains in monomeric state losing affinity to single stranded RNA[48]. VP35 is a multifunctional protein and participates in the assembly of nucleocapsid[49]. Similarly, VP40 is the most abundant matrix protein in the Filovirus and provides a link between the nucleocapsid structure and membrane[50] (Figure 2.2a). The VP40 interacts with negatively charged phospholipids within lipid bilayers to promote virus assembly and budding[49,51,52]. VP24 is a minor matrix protein and plays a key role in viral packaging and immune response inhibition. It is also involved in viral uncoating, formation of ribonucleoprotein complex together with NP and VP35 for nucleocapsid transport and genome packaging[52-55]. VP30 binds with NP and facilitate the formation of virus like particles (VLP's), which is essential for viral RNA synthesis[49]. The protein L is thought to play a role in the catalytic functions required for the viral transcription and replication[56]. These proteins could be

possible target for the discovery and development of vaccine candidates as they involve in various human immune systems[57,58].

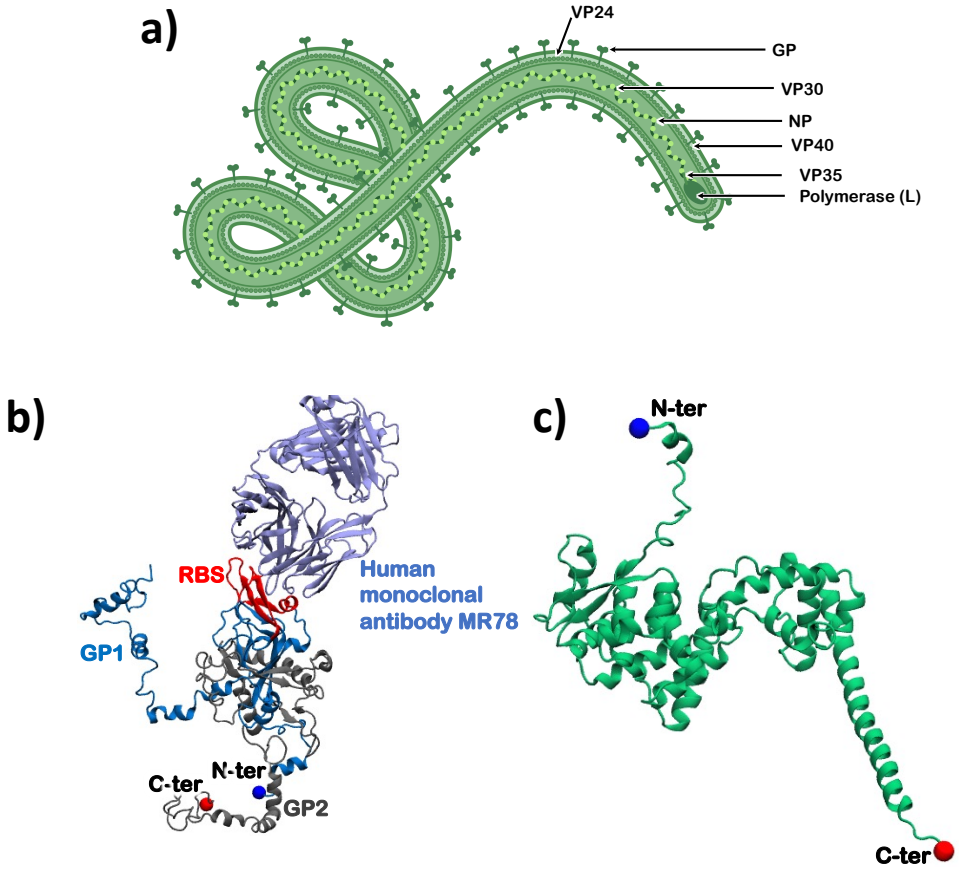


Figure 2.2 Filamentous Filovirus structure showing a) the organization of different proteins in the viral particle (made from BioRender). Crystal structures of the monomers of (b) the MARV GP in complex with human survivor antibody MR78 (PDB ID: 5UQY) and (c) the MARV NP (PDB ID: 5F5M). The N and C terminals in b-c are highlighted with blue and red spheres respectively.

2.3 SARS-CoV-2 Spike Protein

The highly contagious severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) is a form of severe acute respiratory syndrome coronavirus (SARS) that had an outbreak in China in 2003 and causes the COVID-19 disease in a number

of animal species including humans[59-61]. The first case of SARS-CoV-2 was detected in December 2019 in Wuhan, China, and a pandemic was declared in March 2020 due to its worldwide transmission. A person contracting this disease can have symptoms including fever, dry cough, headache, breathing difficulties, and pneumonia[60-62]. Several vaccines against the SARS-CoV-2 infections have been developed[63], with other vaccine candidates are in clinical trials[64]. The vaccines authorized for emergency use have been highly effective in reducing the number of cases and deaths[63,65]. In addition, therapeutic measures are also being pursued in parallel. These include identification of therapeutic small molecules, convalescent plasma, decoy receptors, and neutralizing antibodies[66-76]. Several studies have considered neutralizing antibodies (Abs) that can bind to the virus' receptor binding domain (RBD)[77-82] and have advantageous pharmacokinetics and the ability to be produced on a large scale[83]. The primary target for the vaccines or Abs is the receptor binding domain (RBD) of the spike protein, which is responsible for binding of the virus to the human ACE2 receptor on the host cell and facilitating viral entry into the human cells[75,84,85]. The spike protein in complex with ACE2 receptor is shown in Figure 2.3.

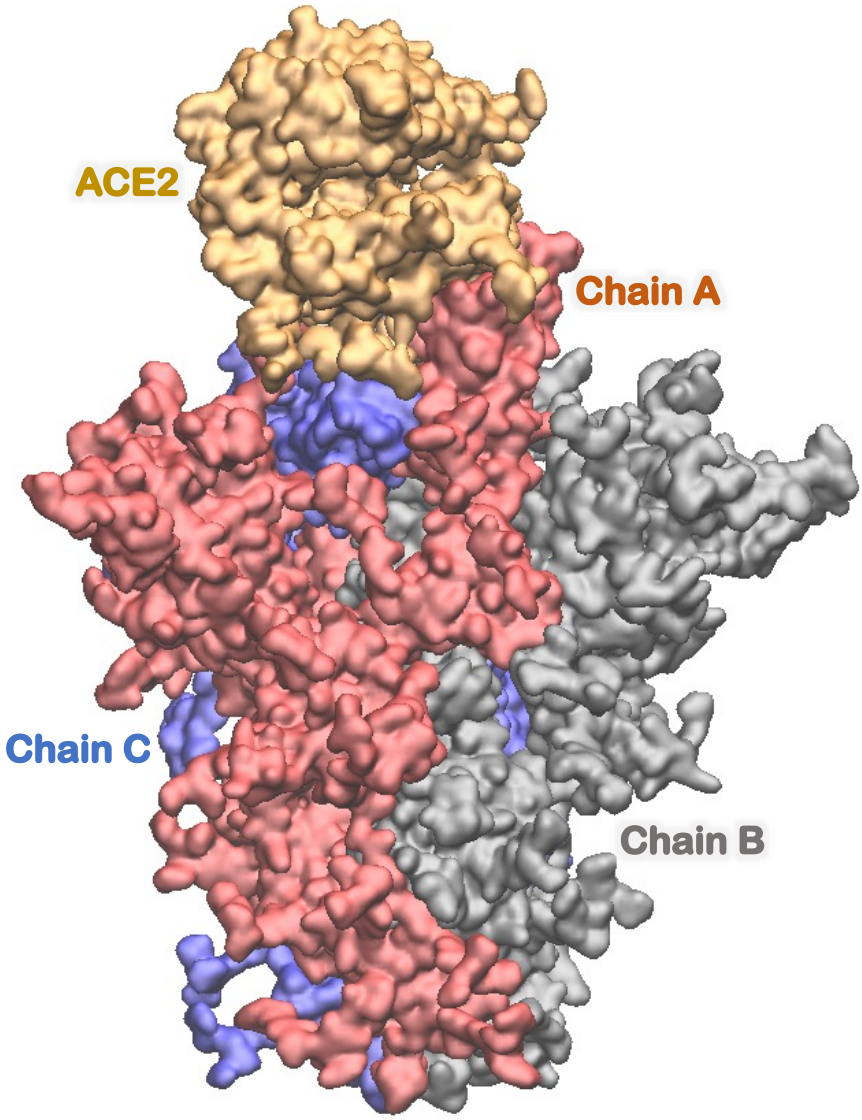


Figure 2.3 SARS-CoV-2 spike trimer in complex with ACE2 receptor. Each monomer in the spike protein is labelled as Chain A, Chain B and Chain C.

While vaccination efforts are ongoing worldwide, new genetic variants are emerging and spreading. Notably, the viruses that emerged in the United Kingdom (B.1.1.7 lineage, also known as alpha variant), in South Africa (B.1.351 lineage, also known as beta variant) and in India (B.1.617.2 lineage, also known as the delta variant) that are reported to be critical among several other variants by increasing the transmission of the virus and host immune evasion[86,87]. Although,

the current vaccines appear to be effective against many variants of concern[88,89], the lack of vaccination coverage worldwide has allowed the virus to spread and continue to evolve, decreasing the chances of quickly getting rid of this viral infection. In the spike protein alone, the B.1.1.7 variant has amino acid deletions at H69, V70, and Y144 and mutations N501Y, A570D, P681H, T716I, S982A, and D1118H, the South African variant has mutations L18F, D80A, D215G, R246I, K417N, E484K, N501Y, and A701V [90], while the B.1.617.2 variant has mutations T19R, G142D, Δ156-157, R158G, L452R, T478K, D614G, P681R, and D950N[91].

3. METHODS

3.1 Retrieval of the viral proteins

The amino acid sequences of the viral proteins for different strains investigated were obtained from the NIAID Virus Pathogen Database and Analysis Resource (ViPR) website[92]. ViPR often provides multiple sequences for the same protein in each viral strain. To choose the sequence to perform investigations, I used the Clustal Omega program[93] to align the different sequences and determine which sequence is the most highly conserved for a single strain. I used UniprotKB[94] to obtain information about the 3D structure of each strain sequence for each protein. The most conserved sequences were used as the principal sequence for epitope predictions.

3.2 Prediction of T-cell epitopes and their antigenicity and allergenicity

Sequence-based MHC-I T-cell epitope predictions for the selected proteins were carried out using three different servers, ProPred-I[95], CTLPred[96] and NetCTL1.2[97]. Sequence-based MHC-II T-cell epitope predictions were made by using three independent servers: ProPred[98], NetMHCII2.3[99] and EpiTop[100] separately. ProPred uses a quantitative matrix[98] approach and NetMHCII2.3 uses ANN[99], while EpiTOP 3.0 uses Quantitative Structure–Activity Relationship models (QSAR)[101] to predict the MHC-II T-cell epitopes.

3.3 Prediction of B-cell epitopes

Servers such as BepiPred2.0[102], BCPREDSs[103] and BcePred[104] are implemented for the sequence-based B-cell epitope predictions. These servers predict epitopes based on physico-chemical properties of amino acids, and these servers accept the primary sequence as an input. The BepiPred predicts the epitopes based on a random forest algorithm trained on epitopes annotated from antibody-antigen structures. BCPREDS predicts epitopes by using SVM combined with a different kernel method, including string kernels, radial basis kernels, and subsequence kernels. The BcePred locates B-cell epitopes using four physicochemical properties like hydrophilicity, polarity, exposed surface and beta-turns[104]. Similarly, the structure-based B-cell epitope prediction were carried out using different independent servers: ElliPro[105], Epitopia[106] and DiscoTope[107]. These servers predict epitope regions based on the geometrical and solvent surface-accessibility of a protein structure, and these servers accept the 3D structure of a protein as an input. ElliPro predicts linear and conformational epitopes by incorporating the antigenicity, solvent accessibility, and flexibility of protein structures[105]. Epitopia uses a machine learning algorithm to analyze the antigenic features on protein structure and predicts the probable conformational epitope regions[106]. DiscoTope uses amino acid statistics, spatial information, and surface accessibility on the protein 3D structure to predict residue-by- residue conformational epitopes[107]. The consensus epitopes from both sequence and structure-based predictions were selected as potential epitopes for novelty analysis.

The antigenicity scores of all the predicted epitopes were predicted by VaxiJen 2.0[108] while the allergenicity assessment was made by using AllerTOP V.2 server[109].

3.4 Docking-based virtual screening

For the MHC-I epitopes, the IEDB provided 27 viable MHC protein variants (alleles) that appear most frequently in the global human population[110]. The structures of all 27 alleles were minimized and MD simulations were performed for 100 ns for each allele. The MD simulations were performed using the conditions described later. 100 different conformations, separated by an interval of 1 ns were generated from the MD trajectories of each allele, resulting in 2,700 structures for the 27 alleles. The MHC-I epitopes that have antigenicity greater than a threshold value of 0.5 and IC_{50} values less than 500 nM were selected for further investigations of their binding affinities to alleles. The structure of each epitope and all 2700 allele structures were prepared for docking using AutodockTools version 1.5.6[111]. Autodock Vina 1.1.2[112] was used for epitope-allele docking with a grid space that covered the entire allele and each epitope was docked with all 2,700 allele structures generated from the earlier MD simulations. This allowed us to determine the strongest binding affinity for each epitope to any of the 2,700 allele structures.

3.5 All Atom Molecular Dynamics (AAMD) Computational Simulations

I used molecular dynamics (MD) computer simulations in most of my Ph.D. projects. MD simulations investigates the evolution of system and its conformational changes concerning the structural transformations. The first MD simulation of protein dates back to the late 1970s[113] and by now, it is used in various fields including the drug discovery[114]. In this process, the atoms in the biomolecules are treated as a system of classical particles and the energy function is used to characterize the interaction between them. The overall energy of the system is calculated from the force field and the equation representing the overall energy of the system is given in equation (3.5.1).

$$U_{\text{total}} = \sum_{\text{bonds}} K_b\left(r - r_0\right)^2 + \sum_{\text{angles}} K_\theta\left(\theta - \theta_0\right)^2 + \sum_{\text{dihedrals}} K_\varphi\left[1 + \cos\left(n\varphi - \varphi_0\right)\right] + \sum_{\substack{\text{van der Waals} \\ i,j \text{ pairs}}} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + \sum_{\substack{\text{electrostatic} \\ i,j \text{ pairs}}} \frac{q_i q_j}{\varepsilon r_{ij}} \tag{3.5.1}$$

The different interaction parameters are represented in Figure 3.1. A virtual spring is used as a model to represent the bonds between adjacent atoms. The bonds, angles and the dihedrals represent the bonded interactions while the Van der Waal's energy defined by Lenard Jones 6-12 potential and the pairwise electrostatic interactions given by Coulomb's law are non-bonded interactions. A cutoff distance is chosen by the user to reduce the computational time by applying the switching function to smoothly cut off the non-bonded interactions. The MD simulations uses a general cutoff distance of 12-14 Å.

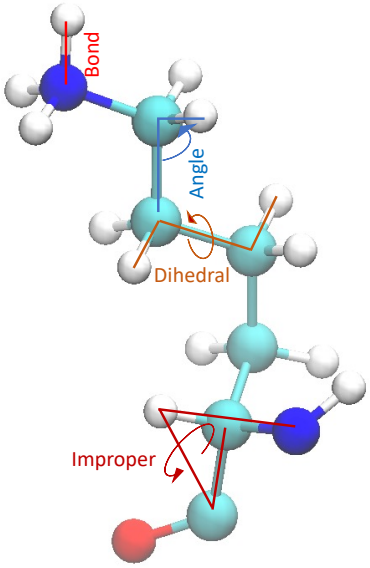


Figure 3.1 Illustration of bonded interactions of a molecule.

In MD simulations protocol, the initial setting of a biomolecule is assigned from the positions of atoms obtained from the three-dimensional crystallography, NMR, cryo-EM, or homology modeling of the structure. A certain temperature is assigned to these particles which then sets their initial velocity by using Maxwell Boltzmann distribution of velocities. The position and velocities of the atoms are then updated and saved after a user defined short time step. As such, the position and velocity of particles at every time steps are saved as time evolution data which forms the MD trajectory. This MD trajectory is utilized to investigate the thermodynamic and kinetic properties of complex biomolecules by using various methods and tools of statistical physics. Thus obtained results can be then validated with the experimental observations.

The large-scale structural changes occur in the biomolecules in the timescales that ranges from milliseconds, seconds or hours. However, the MD time steps are of the order of a femtosecond. As such, it is impossible to observe the significant conformational changes by using conventional MD. These days, the advancement in Graphical Processing Units (GPUs) have significantly enhanced the computational power and a few research groups have been able to mark millisecond scale of simulation with their most advanced computational facilities. The current study employed MD to understand the dynamics as well as protein-protein interactions.

All atom MD simulations were performed in High Performance Computing (HPC) and GPU resources by using the GPU version of NAMD 2.14[115] employing the Charmm36m force field[116,117] in explicit solvent. All systems were solvated in TIP3P water molecules in 0.15 M salt concentration using CHARMM-GUI[118,119]. All the structures were equilibrated for at least 1 ns with a timestep of 2 fs after a short minimization for 10,000 steps. The production runs were performed under constant pressure of 1 atm, controlled by a Nose−Hoover Langevin piston[120] with a piston period of 50 fs and a decay of 25 fs to control the pressure. All simulations were performed at a temperature of 303 K and controlled by Langevin temperature coupling with a damping coefficient of 1/ps. The Particle Mesh Ewald (PME)[121] method was used for long-range electrostatic interactions with periodic boundary conditions and all covalent bonds with hydrogen atoms were constrained by Shake algorithm[122]. The hydrogen bonds were calculated with the cutoff distance and

angle of 3.5 Å and 30^0 respectively. Visualization and analysis of the trajectories were done with Visual Molecular Dynamics (VMD)[123].

4. IN-SILICO IDENTIFICATION OF THE VACCINE CANDIDATE EPITOPES AGAINST THE LASSA VIRUS HEMORRHAGIC FEVER

This chapter was published in 2020: P. Baral, E. Pavadai, B.S. Gerstman, and P.P. Chapagain, *In silico* identification of the vaccine candidate epitopes against the Lassa virus hemorrhagic fever, Sci Rep 10, 7667 (2020).

In this study, I have identified and characterized T and B-cell epitopes for the LASV GP using different sequence and structure-based computational epitope prediction methods. I then selected potential B and T-cell epitopes for the LASV GP based on a consensus approach, and the novelty of the epitopes was examined with the Immune Epitope Database (IEDB) tools. Subsequently, I identified strongly binding alleles to the MHC-I T-cell epitopes and modeled the allele structures and performed docking to understand the interaction between alleles and epitopes. I further investigated the stability and dynamics of the epitope-allele complexes using molecular dynamics simulations. Analyses of root-mean square deviations, hydrogen bond, interaction energy, and solvent accessibility showed that epitope-allele complexes are stable, indicating that the epitopes strongly bind to the alleles. The identified B and T-cell epitopes of LASV GP in the study can be useful for the development of effective vaccines against Lassa hemorrhagic fever.

4.1 Prediction of T-cell Epitopes

4.1.1 MHC-I T-cell Epitopes

MHC-I T-cell epitope prediction with the *LASV GP sequence* was performed using three different methods separately: ProPred-1, CTLPred, and NetCTL1.2, and the results are shown in Table 4.1.1.

Table 4.1.1 MHC class I epitopes prediction.

Propred-1	CTLPred	NetCTL 1.2
IEEVMNIVLI	TFFQEVPHV	TTSLYKGVY
VMNIVLIALSVLAVLKGL	EVPHVIEEV	CTKNNSHHY
VMNIVLIAL	VPHVIEEVM	ITEMLQKEY
NIVLIALSV	VIEEVMNIV	VLAVLKGLY
IALSVLAVL	LIALSVLAV	LSIPNFNQY
SVLAVLKGL	AVLKGLYNF	NTSIINHKF
FATCGLVGLVTFL	FATCGLVGL	MTSYQYLII
FATCGLVGL	VYELQTLEL	LSQRTRDIY
GLVGLVTFL	KNNSHHYIM	NWDCIMTSY
NETGLELTL	IINHKFCNL	FSRPSPIGY
IINHKFCNL	ALMSIISTF	IMCIPYCNY
LSDAHKKNL	MSIISTFHL	LRDIMCIPY
SRPSPIGYL	QYNLSHSYA	QADNMITEM
RRGTFTWTL	GVLQTFMRM	GRSCTTSLY
LIEAELKCF	MAWGGSYIA	MRMAWGGSY
	AWGGSYIAL	YCNYSKYWY
	FSRPSPIGY	NMETLNMTM
	SQRTRDIYI	VQYNLSHSY
	RTRDIYISR	
	RRGTFTWTL	
	RWMLIEAEL	
	ELKCFGNTA	
	MLRLFDFNK	

FNKQAIQRL
NALINDQLI
IPYCNYSKY

The epitopes listed by at least two of the methods are listed in Table 4.1.2 along with their binding affinity (IC_{50}), antigenicity, and allele. Among these four consensus epitopes, the nanomer E1 epitope FATCGLVGL shows the lowest average IC_{50} value of 34 nM against the A1 allele as predicted by the IEDB, and it has also a reasonable antigenicity score of 1.65. This was followed by the E3 epitope FSRPSPIGY, which has an average IC_{50} value of 88 nM against the A3 allele, and also has a better antigenicity score of 2.50 compared to the FATCGLVGL epitope. Interestingly, the E4 epitope RRGTFTWTL is predicted by all three methods though its IC_{50} and antigenicity scores are not as good as the other epitopes (Table 4.1.2). All four of these consensus epitopes were docked to the alleles and I performed the MD simulations to investigate the stability and dynamics of the allele-epitope complex as discussed later.

Table 4.1.2. Consensus prediction of the MHC-I T-cell epitopes.

Epitope	Sequence	Interval	Prediction method			Antigenicity	IC_{50} (nM)		Allele
			ProPred-1	CTLPred	NetCTL		ANN	SMM	
E1	FATCGLVGL	38-46	✓	✓	×	1.65	11.91	55.79	HLA-A*02:06 (A1)
E2	IINHKFCNL	112-120	✓	✓	×	1.23	101.6	214.8	HLA-A*02:03 (A2)
E3	FSRPSPIGY	233-241	×	✓	✓	2.50	81.63	94.1	HLA-B*35:01 (A3)
E4	**RRGTFTWTL**	**258-266**	**✓**	**✓**	**✓**	**1.04**	**109.6**	**727.7**	**HLA-A*32:01(A4)**

The epitope predicted by all three methods is highlighted in boldface.

4.1.2 MHC-II T-cell Epitopes

MHC-II T-cell epitope prediction with the *LASV GP sequence* was performed using three different methods separately: ProPred, NetMHCII 2.3, and EpiTOP 3.0, and the results are shown in Table 4.1.3. ProPred uses a quantitative matrix[98] approach and NetMHCII2.3 uses ANN[99], while EpiTOP 3.0 uses Quantitative Structure–Activity Relationship models (QSAR)[101] to predict the MHC-II T-cell epitopes.

Table 4.1.3 MHC class II epitopes prediction (The core epitopes predicted by NetMHCII2.3 are highlighted in bold).

ProPred	NetMHCII2.3	EpiTOP3.0
MGQIVTFFQ	IALSV**LAVLKGLYN**F	MGQIVTFFQ
FFQEVPHVI	ALSV**LAVLKGLYN**FA	FQEVPHVIE
VMNIVLIALSVL	LSV**LAVLKGLYN**FAT	IEEVMNIVL
MNIVLIALSVLAV	SV**LAVLKGLYN**FATC	VLIALSVLA
IVLIALSVL	V**LAVLKGLYN**FATCG	LIALSVLAV
LKGLYNFATCG	**LAVLKGLYN**FATCGL	LSVLAVLKG
LVGLVTFLLLCGRSC	DCIMTS**YQYLIIQNT**	LAVLKGLYN
VYELQTLEL	CIMTS**YQYLIIQNT**T	LKGLYNFAT
LNMTMPLSC	IMTS**YQYLIIQNT**TW	FATCGLVGL
LELTLTNTSII	TS**YQYLIIQNT**TWED	LCGRSCTTS
INHKFCNLS	RPSPIG**YLGLLSQR**T	LYKGVYELQ
MSIISTFHLSI	PSPIG**YLGLLSQRT**R	YKGVYELQT
MSIISTFHL	SPIG**YLGLLSQRT**RD	VYELQTLEL
FNQYEAMSC	PIG**YLGLLSQRT**RDI	YELQTLELN
VQYNLSHSY	IG**YLGLLSQRT**RDIY	LQTLELNME
ISVQYNLSH	G**YLGLLSQRT**RDIYI	LELNMETLN
YNLSHSYAG	MSIIST**FHLSIPNFN**	LNMETLNMT
YAGDAANHC	SIIST**FHLSIPNFN**Q	LNMTMPLSC
MRMAWGGSY	IIST**FHLSIPNFN**QY	IMVGNETGL

ProPred	NetMHCII2.3	EpiTOP3.0
LQTFMRMAW	IST**FHLSIPNFN**QYE	MVGNETGLE
MTSYQYLII	ST**FHLSIPNFN**QYEA	VGNETGLEL
YQYLIIQNT	T**FHLSIPNFN**QYEAM	LELTLTNTS
LIIQNTTWE	GVLQTF**MRMAWGGSY**	LTNTSIINH
IGYLGLLSQ	VLQTF**MRMAWGGSY**I	IINHKFCNL
LLSQRTRDI	LQTF**MRMAWGGSY**IA	INHKFCNLS
IYISRRRRG	QTF**MRMAWGGSY**IAL	LSDAHKKNL
WMLIEAELK	TF**MRMAWGGSY**IALD	LYDHALMSI
LKCFGNTAV	F**MRMAWGGSY**IALDS	LMSIISTFH
LRLFDFNKQ	LSVLAV**LKGLYNFAT**	MSIISTFHL
IQRLKAPAQ	SVLAV**LKGLYNFAT**C	IISTFHLSI
IQLINKAVN	VLAV**LKGLYNFAT**CG	FHLSIPNFN
LINDQLIMK	LAV**LKGLYNFAT**CGL	FNQYEAMSC
LIMKNHLRDIM	AV**LKGLYNFAT**CGLV	MSCDFNGGK
LRDIMCIPY	V**LKGLYNFAT**CGLVG	FNGGKISVQ
YCNYSKYWY	VGLVT**FLLLCGRSC**T	ISVQYNLSH
WYLNHTTTGR	GLVT**FLLLCGRSC**TT	LSHSYAGDA
LVSNGSYLN	LVT**FLLLCGRSC**TTS	VLQTFMRMA
	VT**FLLLCGRSC**TTSL	FMRMAWGGS
	NNSHH**YIMVGNETG**L	MRMAWGGSY
	NSHH**YIMVGNETG**LE	WGGSYIALD
	SHH**YIMVGNETG**LEL	LDSGCGNWD
	HH**YIMVGNETG**LELT	WDCIMTSYQ
	HKKNL**YDHALMSII**S	LIIQNTTWE
	KKNL**YDHALMSII**ST	WEDHCQFSR
	KNL**YDHALMSII**STF	FSRPSPIGY
	NL**YDHALMSII**STFH	IGYLGLLSQ
	L**YDHALMSII**STFHL	LLSQRTRDI
	GVLQT**FMRMAWGGS**Y	LSQRTRDIY
	VLQT**FMRMAWGGS**YI	YISRRRRGT
	LQT**FMRMAWGGS**YIA	ISRRRRGTF
	QT**FMRMAWGGS**YIAL	WMLIEAELK
	T**FMRMAWGGS**YIALD	MLIEAELKC

ProPred	NetMHCII2.3	EpiTOP3.0
	AQT**SIQLINKAV**NAL	LIEAELKCF
	QT**SIQLINKAV**NALI	IEAELKCFG
	T**SIQLINKAV**NALIN	FGNTAVAKC
	QTSIQL**INKAVNALI**	FCDMLRLFD
	TSIQL**INKAVNALI**N	FDFNKQAIQ
	SIQL**INKAVNALI**ND	FNKQAIQRL
	IQL**INKAVNALI**NDQ	LKAPAQTSI
	QL**INKAVNALI**NDQL	IQLINKAVN
	SQRTRD**IYISRRRRG**	LINKAVNAL
	QRTRD**IYISRRRRG**T	LINDQLIMK
	RTRD**IYISRRRRG**TF	LIMKNHLRD
	TRD**IYISRRRRG**TFT	IMKNHLRDI
	RD**IYISRRRRG**TFTW	LRDIMCIPY
	D**IYISRRRRG**TFTWT	WYLNHTTTG
	IYISRRRRGTFTWTL	LNHTTTGRT
	NLYDHA**LMSIISTFH**	WLVSNGSYL
	LYDHA**LMSIISTFH**L	LVSNGSYLN
	YDHA**LMSIISTFH**LS	FSDDIEQQA
	DHA**LMSIISTFH**LSI	ITEMLQKEY
	HA**LMSIISTFH**LSIP	MLQKEYMER
	MTSYQ**YLIIQNTTW**E	
	TSYQ**YLIIQNTTW**ED	
	SYQ**YLIIQNTTW**EDH	
	YQ**YLIIQNTTW**EDHC	
	Q**YLIIQNTTW**EDHCQ	
	LIALS**VLAVLKGLY**N	
	IALS**VLAVLKGLY**NF	
	ALS**VLAVLKGLY**NFA	
	LS**VLAVLKGLY**NFAT	
	VANGV**LQTFMRMAW**G	
	ANGV**LQTFMRMAW**GG	
	NGV**LQTFMRMAW**GGS	
	GV**LQTFMRMAW**GGSY	

ProPred	NetMHCII2.3	EpiTOP3.0
	V**LQTFMRMAW**GGSYI	
	DFNGG**KISVQYNLS**H	
	FNGG**KISVQYNLS**HS	
	NGG**KISVQYNLS**HSY	
	GG**KISVQYNLS**HSYA	
	G**KISVQYNLS**HSYAG	
	NETGLE**LTLTNTSII**	
	ETGLE**LTLTNTSII**N	
	TGLE**LTLTNTSII**NH	
	GLE**LTLTNTSII**NHK	
	LE**LTLTNTSII**NHKF	
	LYDHAL**MSIISTFHL**	
	YDHAL**MSIISTFHL**S	
	DHAL**MSIISTFHL**SI	
	HAL**MSIISTFHL**SIP	
	AL**MSIISTFHL**SIPN	
	L**MSIISTFHL**SIPNF	

The epitopes that were predicted by at least two methods are listed in Table 4.1.4. Among these consensus MHC-II T-cell epitope predictions, the E9 and E13 epitopes were predicted by all three methods and have a reasonable antigenicity score of 0.7, indicating that these two epitopes can be potential candidates for the design of MHC-II T-cell based vaccines. ProPred and EpiTOP 3.0 predict most epitopes as nanomers whereas NetMHCII 2.3 predicts varying lengths of epitopes (Table 4.1.4). Interestingly, the 15-mer epitopes predicted by NetMHCII have the consensus core nanomer epitopes, suggesting that the core region is responsible for strong binding of the epitope into the MHC-II binding site[124,125,126].

Table 4.1.4 Prediction of the MHC-II T-cell epitopes.

Epitope	Sequence	Interval	Prediction Method			Antigenicity
			ProPred	NetMHCII	EpiTOP	
E5	MGQIVTFFQ	1-9	✓	×	✓	-0.1820
E6	VYELQTLEL	65-73	✓	×	✓	0.8600
E7	LNMTMPLSC	78-86	✓	×	✓	0.9390
E8	INHKFCNLS	113-121	✓	×	✓	1.5060
E9	***MSIISTFHL***	**134-142**	**✓**	**✓**	**✓**	**0.7080**
	LYDHAL**MSIISTFHL**	128-142	×	✓	×	0.2896
	YDHAL**MSIISTFHL**S	129-143	×	✓	×	0.4907
	DHAL**MSIISTFHL**SI	130-144	×	✓	×	0.4809
	HAL**MSIISTFHL**SIP	131-145	×	✓	×	0.1949
	AL**MSIISTFHL**SIPN	132-146	×	✓	×	0.2066
	L**MSIISTFHL**SIPNF	133-147	×	✓	×	0.2428
E10	FNQYEAMSC	147-155	✓	×	✓	0.5520
E11	ISVQYNLSH	162-170	✓	×	✓	1.1310
E12	LQTFMRMAW	188-196	✓	✓	×	0.2620
	VANGV**LQTFMRMAW**G	183-197	×	✓	×	0.1328
	ANGV**LQTFMRMAW**GG	184-198	×	✓	×	0.1683
	NGV**LQTFMRMAW**GGS	185-199	×	✓	×	0.0579
	GV**LQTFMRMAW**GGSY	186-200	×	✓	×	0.1572
	V**LQTFMRMAW**GGSYI	187-201	×	✓	×	0.1895
E13	***MRMAWGGSY***	**192-200**	**✓**	**✓**	**✓**	**0.7630**
	GVLQTF**MRMAWGGSY**	186-200	×	✓	×	0.1572
	VLQTF**MRMAWGGSY**I	187-201	×	✓	×	0.1895
	LQTF**MRMAWGGSY**IA	188-202	×	✓	×	0.1902
	QTF**MRMAWGGSY**IAL	189-203	×	✓	×	0.3470
	TF**MRMAWGGSY**IALD	190-204	×	✓	×	0.4434
	F**MRMAWGGSY**IALDS	191-205	×	✓	×	0.3543
E14	YQYLIIQNT	217-225	✓	✓	×	0.4720
	DCIMTS**YQYLIIQNT**	211-225	×	✓	×	0.6600
	CIMTS**YQYLIIQNT**T	212-226	×	✓	×	0.7075
	IMTS**YQYLIIQNT**TW	213-227	×	✓	×	0.6029
	TS**YQYLIIQNT**TWED	215-229	×	✓	×	0.6874

Epitope	Sequence	Interval	Prediction Method			Antigenicity
			ProPred	NetMHCII	EpiTOP	
E15	LIIQNTTWE	220-228	✓	×	✓	0.9100
E16	IGYLGLLSQ	239-247	✓	×	✓	1.5300
E17	LLSQRTRDI	244-252	✓	×	✓	1.7310
E18	IYISRRRRG	252-260	✓	✓	×	1.5560
	SQRTRD**IYISRRRRG**	246-260	×	✓	×	1.6434
	QRTRD**IYISRRRRG**T	247-261	×	✓	×	1.5276
	RTRD**IYISRRRRG**TF	248-262	×	✓	×	1.7213
	TRD**IYISRRRRG**TFT	249-263	×	✓	×	1.4112
	RD**IYISRRRRG**TFTW	250-264	×	✓	×	1.5207
	D**IYISRRRRG**TFTWT	251-265	×	✓	×	1.4261
	IYISRRRRGTFTWTL	252-266	×	✓	×	1.2680
E19	WMLIEAELK	283-291	✓	×	✓	1.3250
E20	IQLINKAVN	334-342	✓	×	✓	0.7710
E21	LINDQLIMK	344-352	✓	×	✓	-0.0481
E22	LRDIMCIPY	355-363	✓	×	✓	1.0590
E23	LVSNGSYLN	387-395	✓	×	✓	0.3450

The epitopes predicted by all three methods are highlighted in boldface with Italic font. The consensus core regions highlighted in boldface are in the epitopes predicted by NetMHCII 2.3.

4.2 Prediction of B-cell epitopes

4.2.1 Sequence based epitopes

In addition to the T-cell epitope predictions, I also predicted the linear B-cell epitopes for the LASV GP using sequence-based methods BepiPred 2.0[102], BCPREDS[103], and BcePred[104]. The BepiPred predicts the epitopes based on a random forest algorithm trained on epitopes annotated from antibody-antigen structures. BCPREDS predicts epitopes by using SVM combined with a different kernel method, including string kernels, radial basis kernels, and subsequence kernels. The BcePred locates B-cell epitopes using four physicochemical properties like hydrophilicity, polarity, exposed surface and beta-turns[104]. The epitope E30 containing 10 residues was predicted by all three of these sequence methods (Table 4.2.1) but with a negative antigenicity score.

4.2.2 Structure based epitopes

I also performed structure-based B-cell epitope prediction using three representative structural and geometrical properties-based methods: ElliPro, Epitopia and DiscoTope. For this, the experimental 3D structure LASV GP (PDB ID: 5VK2) with the modeled missing residues was used. ElliPro predicts linear and conformational epitopes by incorporating the antigenicity, solvent accessibility, and flexibility of protein structures[105]. Epitopia uses a machine learning algorithm to analyze the antigenic features on protein structure and predicts the probable conformational epitope regions[106]. DiscoTope uses amino acid statistics, spatial

information, and surface accessibility on the protein 3D structure to predict residue-by- residue conformational epitopes[107]. The E24, E29, E32 and E33 structure-based epitopes in Table 4.2.1 are especially interesting as potential candidates as they were predicted by all three methods. In Table 4.2.1, I also ranked each epitope based upon how many of the sequence and structure-based methods predicted each epitope, which do not always correlate with the highest antigenicity scores of E24, E26, E28, E29 and E31.

Table 4.2.1 Prediction of the B-cell epitopes. The epitopes predicted by either all three sequence- or structure-based methods are highlighted by boldface. Conformational epitopes chosen by all three structure-based methods are indicated in italics.

Epitope			Sequence based			Structure based				
	Sequence	Interval	BepiPre	BCPRED	BcePred	ElliPro	Epitopia	DiscoTop	Rank	Antigenicity
E24	***LSDAHKKNLYD***	**120-130**	✓	×	✓	✓	✓	✓	**5/6**	**0.74**
E25	PNFNQYEA	145-152	✓	×	✓	✓	✓	×	4/6	0.46
E26	DFNGGKI	156-162	×	✓	×	✓	✓	×	3/6	0.73
E27	LSHSYAGDAANH	168-182	✓	×	×	✓	✓	×	3/6	0.08
E28	LDSGCGNWDCIM	203-219	×	✓	×	✓	✓	×	3/6	1.08
E29	***ISRRRRGT***	**254-261**	×	×	✓	✓	✓	✓	**4/6**	**1.25**
E30	**SDSEGKDTPG**	**267-276**	✓	✓	✓	✓	✓	×	**5/6**	**-0.07**
E31	NHTTTGRT	373-380	×	✓	✓	✓	✓	×	4/6	0.99
E32	***ETHFSDDIE***	**396-404**	✓	×	✓	✓	✓	✓	**5/6**	**0.50**
E33	***MLQKEYMERQ***	**414-423**	×	✓	✓	✓	✓	✓	**5/6**	**-0.14**

Robinson et al.[35] have recently reported the cloning of many human monoclonal antibodies derived from memory B cells of Lassa fever survivors in West Africa. These antibodies specifically bind to both GP1 and GP2 epitopes of LASV. The comparison of our predicted B-cell epitopes with those epitopes shows that there are five consensus epitopes that share similarity with Robinson et al. (Table 4.2.2), and another five epitopes that do not share similarity, indicating that our consensus epitope prediction strategy has identified new epitopes.

Table 4.2.2 Comparison of B-cell epitopes (this work) with the Robinson et al.'s[35] B-cell epitopes. The common sequence between these two epitopes is highlighted in red color. Here, LASV-I, LASV-II and LASV-III represent different lineages from Nigeria while LASV-IV represents those from Sierra Leone[35]. The blue colors represent the amino acids differing in lineages.

Epitope	Sequence	Interval	Sequence (ref. Robinson)
E24	LSDAHKKNLYD		NLSDAHKKNLYDHALM (LASV-I, II, IV)
			NLSDAHKKNLYDHTLM (LASV-III)
E30	SDSEGKDTPG	267-276 (270-278)	EGKDTPGGY (LASV IV)
			EGNETPGGY (LASV I-III)
E31	NHTTTGRT	373-380 (369-373)	YWYLN (LASV I-IV)
E32	ETHFSDDIE	396-404 (401-415)	DDIEQQADNMITEML (LASV I, III, IV)
			DDIEQQADNMITELL (LASV II)
E33	MLQKEYMERQ		DDIEQQADNMITEML (LASV I, III, IV)
			DDIEQQADNMITELL (LASV II)

4.3 Epitope surface mapping

For efficacy of vaccines, the epitopes should be located on an accessible region of the protein so that the epitope will be able to bind with antibodies[127]. This is

especially important for the six epitopes that I list in the Tables above that do not share any part of their sequence with known epitopes: E1, E4, E18, E22, E27, E29. In Figure 4.1, I highlight the positions of these epitopes on LASV GP. I also highlight the positions of E2 and E3 because the four MHC-I T-cell epitopes have IC_{50} information readily available. Figure 4.1 shows that the E1, E2, E3, E4, E18, E22 and E27 epitopes are well located on the exposed regions and thus can interact well with the alleles.

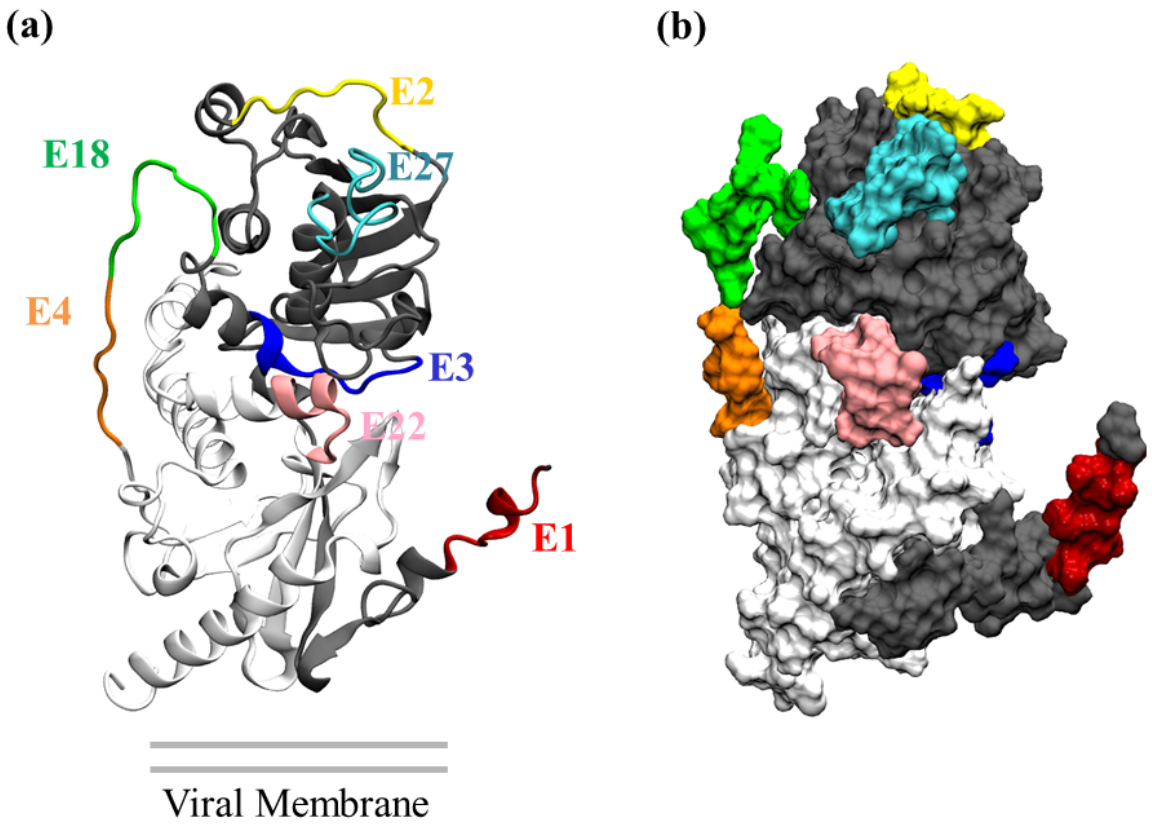


Figure 4.1 Mapping of some representative epitopes are highlighted on the LASV GP. Mapping of: (a) secondary structural elements, (b) surface accessibility. The location of the epitopes on the GP suggests that they are on the solvent exposed region, indicating promiscuity as they have easy access to alleles.

4.4 MHC-I T-cell Allele and epitope modeling and docking

Swiss-Model identified the 1.61 Å resolution crystal structure of the HLA class I antigen (PDB ID: 6EI2) as the best template for constructing models. The sequence identity between A4 and the template was 92%. The best model was then selected based on multiple validation methods, including GMQE (Global Model Quality Estimation) and QMEAN. The GMQE and QMEAN values[128,129] of the model are 0.75, and 0.6, respectively. In addition to these analyses, Ramachandran plots and ERRAT were also used for the model validation. Analysis of Ramachandran plots[130] of the model shows 99.6% of residues are either in favored or in allowed regions (Figure 4.2), indicating that backbone torsion angles of these models are acceptable. The ERRAT overall quality factor[131] score was computed as 99, which is greater than the normally accepted score range for a high quality model of 50. These analyses show that the model is within a high-quality range and can be used for further analysis.

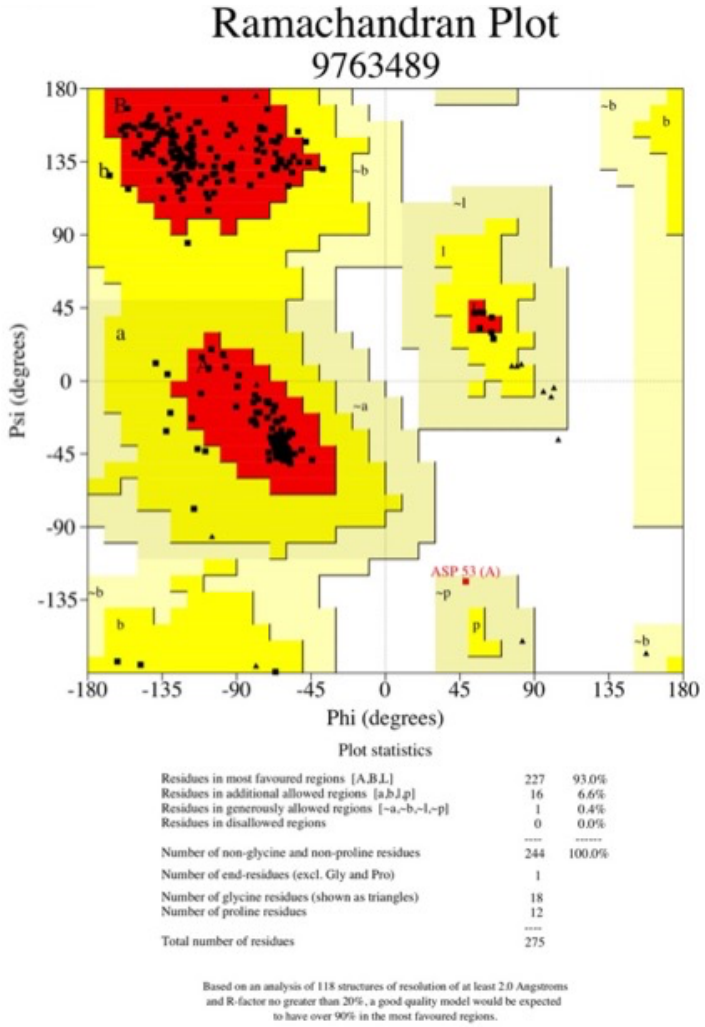


Figure 4.2 Ramachandran plot of the A4 model.

Docking of the four consensus MHC-I epitopes (Table 4.1.2) was performed using Autodock Vina, which enabled the docking of epitopes obtained from the sequence-based MHC-1 T-cell prediction into the promising allele structures. The Autodock Vina docking protocol has been previously demonstrated to successfully dock epitopes into allele structures[112]. However, I validated the capability of the docking protocol before docking the epitopes by redocking the epitopes into the allele crystal structure (PDB ID:3OX8) to see whether the crystal bound

conformation of the peptide could be reproduced or not. The docked allele-epitope complex showed the same residue-epitope interactions observed in the epitope bound crystal structure, indicating that the Autodock Vina docking protocol was capable of reproducing the experimentally observed binding mode of the epitope. I applied Autodock Vina to each of the four MHC-I allele-epitope complexes. Autodock Vina found that the highest ranked docking structure had the following binding affinities: -5.5 kcal/mol for A1::E1 -5.0 kcal/mol for A2::E2, -6.8 kcal/mol for A3::E3, and -6.0 for A4::E4. These epitopes-alleles docking complexes are shown in Figure 4.3.

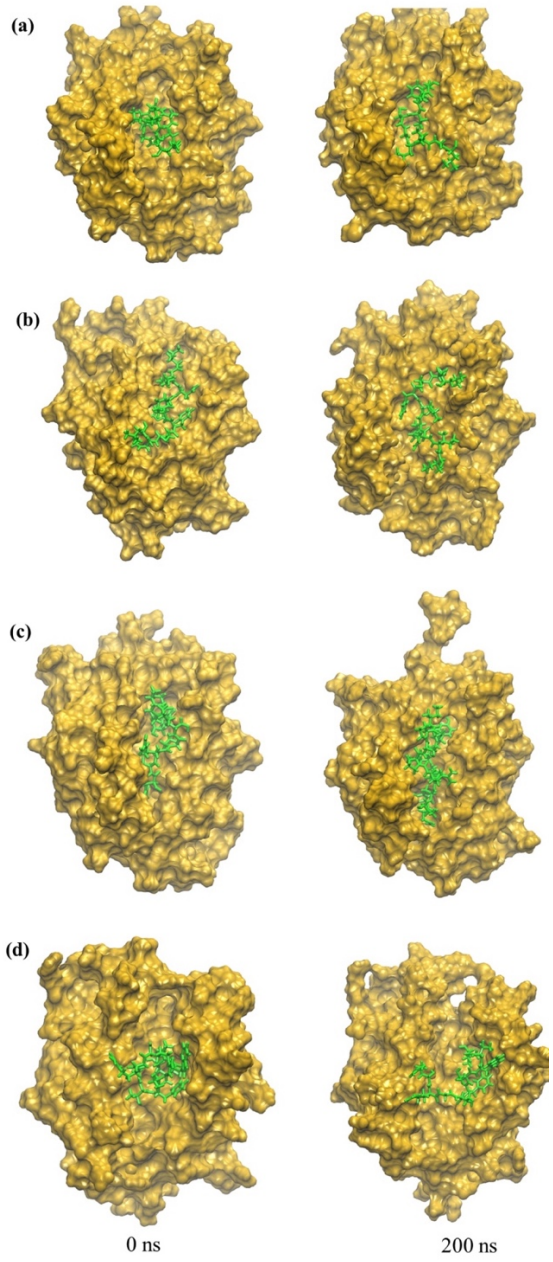


Figure 4.3 Snapshots of allele-epitope complexes. (a) A1::E1, (b) A2::E2, (c) A3::E3, and (d) A4::E4 at the beginning and end of the MD simulations: t=0 (minimized structure), t=200 ns. Allele is gold and epitope is green.

4.5 Dynamics of the allele-epitope complex

4.5.1 Root mean square deviation of allele, epitope and allele-epitope complex

In order to investigate the dynamics and stability of the four MHC-I allele-epitope complexes, I performed 200 ns all-atom, explicit solvent MD simulations. To quantitatively understand the stability of the allele-epitope complex, I calculated the root mean square deviations (RMSD) of the backbone atoms of the allele-epitope complexes as a function of simulation time as shown in Figure 4.4. Figure 4.4 also includes curves of the RMSD of the backbone atoms of just the allele, and separately, just the backbone atoms of the epitope. All alleles have an RMSD compared to their initial structures of approximately 2 Å, whereas the allele-epitope complexes have a bit higher RMSD of approximately 2.5 Å, indicating that the epitopes make the complexes more flexible. Interestingly, in the case of A3::E3, the allele and the complex show almost the same RMSD, suggesting that the complex is especially stable. To pinpoint why the complexes, show a higher RMSD, I further computed the RMSD of only the backbone atoms of the epitope in each the complex. Figure 4 shows that the initial configuration of epitopes E1 and E4 is compact, and that both of these epitopes rearrange their configuration in the binding site and elongate during the 200 ns MD simulation. This elongated configuration is consistent with the investigations of Antunes et al.[132] on MHC-I epitopes.

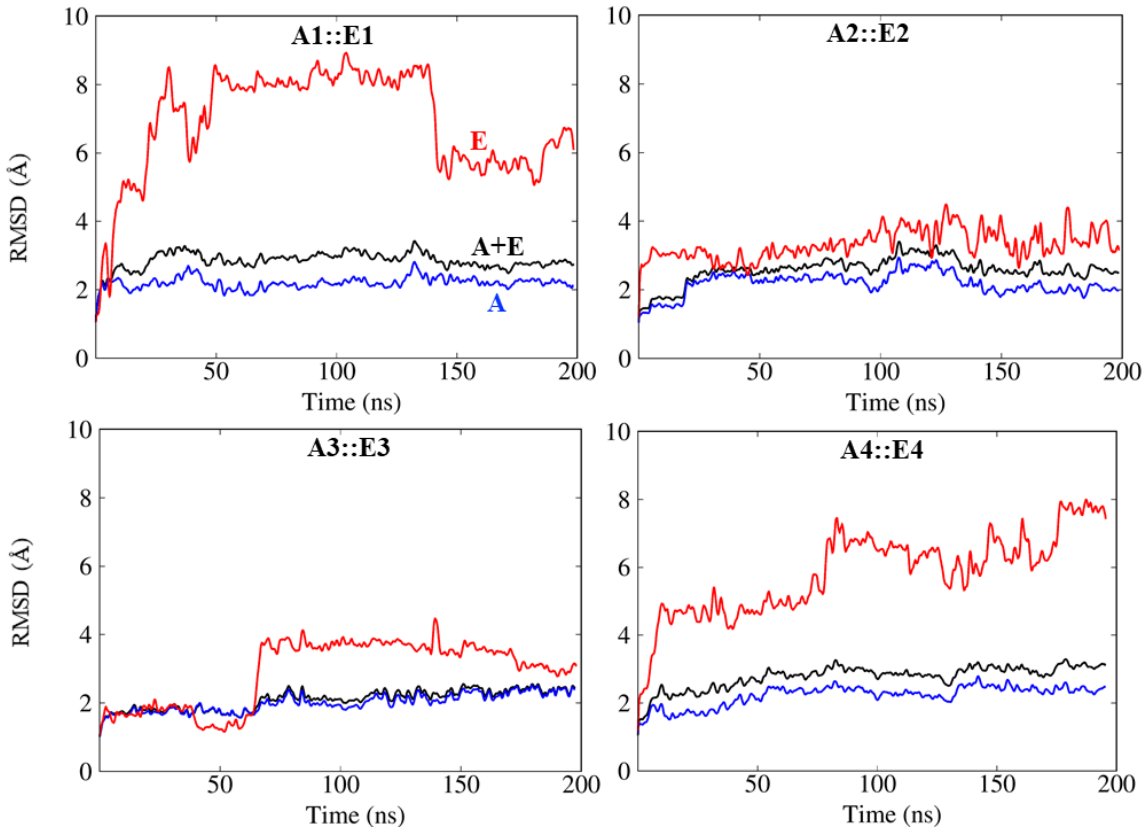


Figure 4.4 Root-mean-squared deviations (RMSD) calculated for the backbone atoms of allele (A), epitope (E) and complex (A+E) from MD simulations of MHC-I allele-epitope complexes.

4.5.2 Hydrogen bonds and interaction energy

Since the interactions between protein and epitope peptide are mostly influenced by non-covalent interactions, I computed the number of hydrogen bonds and the interaction energy between the allele and epitope as a function of the MD simulation time. The hydrogen bond was calculated between the protein interface atoms with a distance cut-off of 3.5 Å and angle cut-off of 30° between the donor and acceptor heavy atoms. As shown in Figure 4.5, the number of H-bonds fluctuates during the MD simulations for all the complexes. The A4 complex has

the largest number of H-bonds. Table 4.5.1 shows that during the last 50 ns of the MD simulation trajectory, the A4 complex averages 2.5 H-bonds.

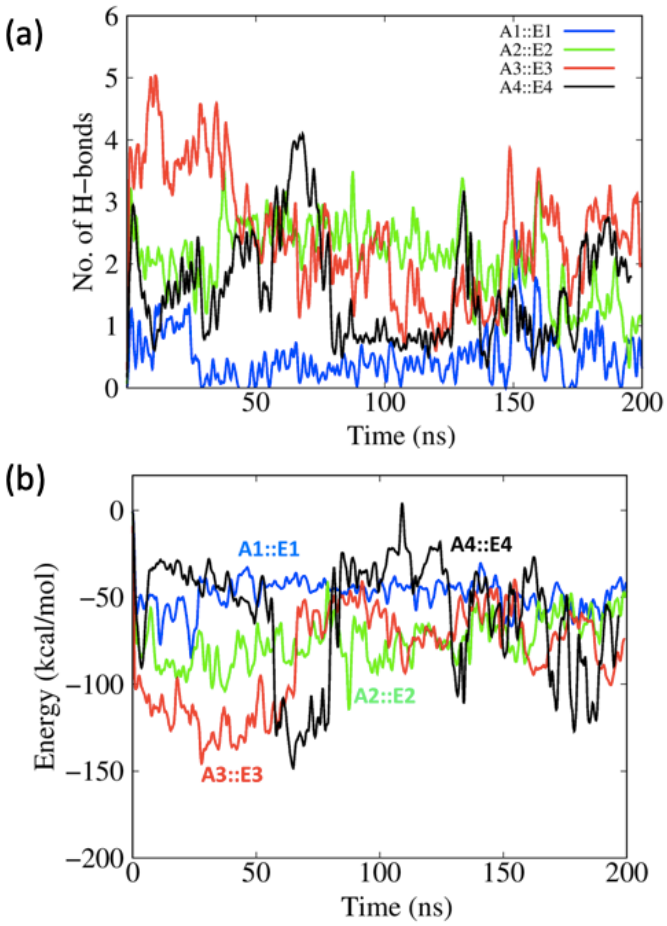


Figure 4.5 (a) The number of allele-epitope intermolecular hydrogen bonds as a function of MD simulation time. (b) Interaction energy calculated between allele and epitopes as a function of simulation time.

Figure 4.5 shows the interaction energy (electrostatic interaction + van der Waals contacts) throughout the entire MD simulation and Table 4.5.1 lists the average over the last 50 ns. The A3::E3 and A4::E4 display relatively stronger interaction energies than the A1:E1 and A2::E2 complexes. The comparison of RMSD,

hydrogen bond, and interaction energy information indicates that the E3 epitope is an especially promising epitope candidate.

Table 4.5.1 Allele–epitope interaction parameters calculated by averaging over the last 50 ns of the MD simulation trajectory. The best interaction is highlighted in bold.

Complex	Interaction	No. of H-bonds
A1::E1	-53.53 ± 7.40	0.64 ± 0.54
A2::E2	-64.54 ± 10.88	1.49 ± 0.63
A3::E3	**-74.85 ± 14.94**	**2.48 ± 0.50**
A4::E4	-73.23 ± 27.07	1.51 ± 0.67

4.6 Novelty analysis

The novelty of the four MHC I T-cell epitopes in Table 4.1.2, the nineteen MHC II T-cell epitopes in Table 4.1.4, and the ten B-cell epitopes in Table 4.2.1 identified in this study were analyzed using IEDB[110]. The IEDB database contains the epitopes that are annotated based on scientific literature. The IEDB showed that the E1, E4, E18, E22, E27, E29 epitopes, which bind to solvent exposed regions on the protein (Figure 4.1), have not been previously reported as LASV epitopes or vaccine candidates). In addition, this analysis further indicates that 24 other epitopes (E2, E3, E5, E6, E7, E8, E10, E11, E12, E14, E15, E16, E17, E19, E20, E23, E24, E25, E26, E28, E30, E31, E32, E33) have partial segments of their sequence reported as subsets of other epitopes, whereas E9, E13, E21 are exact match to previously reported sequences. For these epitopes, a comparison

showing the overlap between the predicted epitopes in this study and previously known epitopes documented in IEDB is given in Table 4.6.1. Therefore, the novelty results confirm that thirty epitopes have not been previously reported as LASV epitopes, suggesting that their therapeutic potentials in designing vaccines against LASV can be further explored.

Table 4.6.1 New Epitopes and the peptides in which they were reported in IEDB. The common sequences are highlighted in red.

Epitope	Sequence (this work)	IEDB Reported peptides	References
E2	IINHKFCNL	SIINHKFCNL	110,133
		SIINHKFCNLSDAHK	110
		TLTNTSIINHKFCNL	110
E3	FSRPSPIGY	CQFSRPSPIGYLGLL	110
E5	MGQIVTFFQ	MGQIVTFFQEVPHVI	110
E6	VYELQTLEL	LYKGVYELQTLELNM	110
E7	LNMTMPLSC	ETLNMTMPLSCTKNN	110
E8	INHKFCNLS	SIINHKFCNLSDAHK	110
E9	MSIISTFHL	LMSIISTFHL	110,134
		MSIISTFHL	110
		HALMSIISTFHLSIP	110
E10	FNQYEAMSC	HLSIPNFNQYEAMSC	110
		NFNQYEAMSCDFNGG	
E11	ISVQYNLSH	DFNGGKISVQYNLSH	110
		KISVQYNLSHSYAVD	
		KISVQYNLSH	
E12	LQTFMRMAW	GVLQTFMRMAWGGSY	110
E13	MRMAWGGSY	MRMAWGGSY	110,135
		FMRMAWGGSY	110,135
		FMRMAWGGSYIALDS	134
		GVLQTFMRMAWGGSY	134
E14	YQYLIIQNT	MRMAWGGSYI	110,135

Epitope	Sequence (this work)	IEDB Reported peptides	References
		SYQYLIIQNTTWEDH	
E15	LIIQNTTWE	SYQYLIIQNTTWEDH	110
E16	IGYLGLLSQ	PSPIGYLGLLSQRTR	110,136
E17	LLSQRTRDI	YLGLLSQRTRDIYIS	110,136
E19	WMLIEAELK	RWMLIEAELKCFG	110,137
		TRWMLIEAELKCFGN	110
		PGGYCLTRWMLIEAELKCF	110
		RWMLIEAELKCFGNTAVAKC	110,137
E20	IQLINKAVN	QMSIQLINKAVNALI	110
E21	LINDQLIMK	LINDQLIMK	110
		VNALINDQLIMKNHL	
E23	LVSNGSYLN	SLPKCWLVSNGSYLN	110
		WLVSNGSYLNETHFS	
E24	LSDAHKKNLYD	KFCNLSDAHKKNLYD	110
		SDAHKKNLY	
E25	PNFNQYEA	HLSIPNFNQYEAMSC	110
E26	DFNGGKI	DFNGGKISVQYNLSH	110
		EAMSCDFNGGKISVQ	
E28	LDSGCGNWDCIMT	DCIMTSYQY	110
E32	ETHFSDDIE	LNETHFSDDIEQQ	110,137
		ETHFSDDIEQQADNM	110
		GSYLNETHFSDDIEQ	
E33	MLQKEYMERQ	MLQKEYMER	110
		ITEMLQKEYMERQGK	

5. IN SILICO SCREENING AND MOLECULAR DYNAMICS INVESTIGATIONS OF MARBURGVIRUS EPITOPE-ALLELE COMPLEXES

In this study, I have used a consensus computational approach that provides higher confidence in the selection of epitopes as discussed in our recent work[138], to identify and characterize T- and B- cell epitopes for the MARV genome. I performed epitope prediction analyses for a complete MARV virion and then performed additional studies for the GP and NP epitopes as they are primary targets for MARV vaccine design. I first predicted and then selected the potential T- and B-cell epitopes based on their antigenic properties using the consensus approach. The novelty of the selected consensus epitopes was further examined with the Immune Epitope Database (IEDB) tools and literature. Several of the epitopes that I identified are novel. Variations in the predicted epitopes among the Lake Victoria, Angola, Musoke, and Ravn strains were analyzed. I screened epitope-allele complexes based upon the pharmaceutically important IC_{50} values. The MHC-I epitope-allele complexes for GP and NP with favorably low IC_{50} values were investigated using molecular dynamics computations to determine molecular details of the epitope-allele complex. To determine the strength of binding affinity to alleles, predicted GP and NP MHC-I T-cell epitopes with low IC_{50} were computationally docked to multiple conformations of a large array of MHC-I alleles. This study provides guidance for further experimental investigations and validation of potential epitopes for the design and development of MARV vaccine candidates.

The GP and NP epitope predictions given below (E1-E68) are provided in the following order: MHC-I T-cell, MHC-II T-cell, B-cell. For epitope comparison, the GP sequences of Lake Victoria MARV (UniprotKB ID: Q71VM1), Angola (UniprotKB ID: Q1PD50, strain Angola/2005), Musoke (UniprotKB ID: P35253, strain Kenya/Musoke/1980), and Ravn-1987 (UniprotKB ID: Q1PDC7, strain Kenya/Ravn/1987) were obtained from the UniprotKB[94] database. The multiple sequence alignment of MARV GP amino acid sequences resulted in the Lake Victoria being identified as a highly conserved strain, and I selected this as the principal strain for epitope predictions. The 3D structure of the Lake Victoria, Musoke and Angola MARV GP and NP proteins were not available in the PDB. I modeled these structures, as well as the missing residues of Ravn MARV (PDB: 6BP2 for GP and 5F5M for NP) with the use of Modeler.

Sequence alignment of MARV NP sequences showed that the Lake Victoria and Angola strains (UniprotKB ID: Q1PD53*)* had identical amino acid sequences that were highly conserved. This was selected as the principal sequence for epitope predictions for NP. The NP sequences from Musoke (UniprotKB ID: P27588, strain Kenya/Musoke/1980) and Ravn-1987 (UniprotKB ID: Q1PDD0, strain Kenya/Ravn/1987) were also considered for sequence-based epitope predictions.

I also performed sequence alignment using UniprotKB to determine highly conserved sequences for the other MARV proteins for epitope predictions. The highest conserved sequence for each protein are: VP35 (UniProtKB: Q1PD52),

VP40 (UniProtKB: Q1PD51), VP30 (UniProtKB: Q1PD56), VP24 (UniProtKB: Q1PD48) and L (UniProtKB: Q1PD54).

5.1 Prediction of MHC-I T-cell epitopes and epitope-allele complexes for MARV GP and NP

MHC-I T-cell epitope predictions for the dominant strains of the MARV GP and NP were performed by using two independent servers separately: ProPred1 and CTLPred. I then examined the same region in the other strains to determine if there were variations in the epitope amino acid sequence. GP and NP epitopes predicted by either method that have at least one strain with antigenicity value greater than 0.5 and IC_{50} values less than 500 nM are listed in Table 5.1.1 along with their IC_{50}, allele, antigenicity, strongest binding affinity to any allele, and allergenicity. Among the 11 GP epitopes, six are allergenic, three are non-allergenic, and two epitopes have both allergenic and non-allergenic strains. As shown in Table 5.1.1, the E11 GP epitope, HAIDFLLTR, has the lowest IC_{50} value of 5.54 nM (IC_{50} value of 6.37 nM for HAIDFLLAR, Angola strain) and has a reasonably high value of antigenicity of 1.031 (0.8133 for Angola). The allergenicity assessment shows that both of these E11 epitope strains are allergenic in nature. The non-allergenic epitopes E5, E6 (Ravn), and E8 have relatively low IC_{50} values. As listed in Table 5.1.1, E20, E21 and E22 are three NP epitopes with the lowest IC_{50} (less than 11 nM) and, they have non-allergenic nature, indicating that they could be good epitope candidates. Table 5.1.1 also shows that the Ravn strains of GP epitopes E1 and

E6 have a greatly reduced IC_{50} compared to the Lake Victoria strain while maintaining the non-allergenic nature of the epitope.

5.1.1 Virtual screening to identify GP and NP MHC-I epitope-allele complexes

In order to identify the alleles that strongly bind to the selected MHC-I MARV GP and NP epitopes, I performed a docking-based virtual screening of all the GP and NP epitopes listed in Table 5.1.1 against a total of 27 MHC-I alleles that have been frequently found in humans[110]. For each of the 27 alleles, I ran a 100 ns MD simulation. For each allele, I obtained the conformation at 1 ns interval to obtain 100 different conformations. This provided a multiconformer allele database consisting of 2,700 unique conformations of the 27 alleles. The generation of multiple conformations of alleles by the use of MD simulations takes advantage of protein flexibility to help obtain reliable binding complexes of epitope and allele. Docking-based virtual screening was performed using the Autodock-Vina method, which enabled the docking of the selected GP and NP MHC-I epitopes to each of the 2,700 allele conformations. The reliability of the Autodock-Vina docking protocol was validated prior to docking the epitopes as mentioned in our previous work[138].

Table 5.1.1.Prediction of MHC-I T-cell epitopes and allele complexes for GP (GP1 and GP2) and NP proteins. The Lake Victoria (L) strain is usually the most conserved sequence and therefore used as the principal sequence for each epitope. Many epitopes have identical predicted sequences in the four Marburg strains. For epitopes that have slightly different predicted sequences in different Marburg strains, I provide the different sequences with strains abbreviated as Angola(A), Musoke(M), Ravn(R). In addition, the epitopes that were further investigated with epitope-allele MD simulations are underlined.

Protein	Epitope	Sequence	Interval	Antigenicity	Allergenic	Affinity kcal/mol	IC_{50} (nM)	Allele
GP1	E1	QGIKTLPIL(L)	14-22	0.659	N	-7.9	5240	HLA-A*30:01
		QGVKTLPIL(A)		0.667	N	-7.6	6601	HLA-A*30:01
		QGTKNLPIL(M)		0.585	A	-7.8	9502	HLA-B*08:01
		QSIKTLPVL(R)		0.546	N	-7.9	318	HLA-B*58:01
	E2	KVADSPLEA	58-66	0.583	A	-9.8	101	HLA-A*02:03
	E3	LEASKRWAF	64-72	0.839	A	-10	118	HLA-B*44:02
	E4	GQNPHAQGI	127-135	0.552	A	-9.9	342	HLA-A*02:06
	E5	NIAAMIVNK	164-172	0.626	N	-8.8	18	HLA-A*68:01
	E6	TTAPNMTNG(L)	404-412	0.772	N	-8.1	754	HLA-A*68:02
		TTAPNTTNE(M)		0.633	N	-8.4	956	HLA-A*68:02
		TTVPNTTNK(A)		0.73	A	-9.3	18	HLA-A*68:01
		MTTSDITSK(R)		1.185	N	-7.6	19	HLA-A*68:01

GP2	E7	REGDMFPFL(L,A,M)	440-448	1.215	A	-9.8	26	HLA-B*40:01
		KEGDIFPFL(R)		0.985	A	-9.8	24	HLA-B*40:01
	E8	QEDDLAAGL	522-530	1.218	N	-9.4	14	HLA-B*40:01
	E9	AAGLSWIPF	527-535	1.474	N	-10.9	266	HLA-A*23:01
	E10	NLVCRLRRL	554-562	0.863	A	-9.1	115	HLA-A*02:03
	E11	HAIDFLLTR(L,M,R)	589-597	1.031	A	-9.9	5.54	HLA-A*68:01
		HAIDFLLAR(A)		0.813	A	-10.5	6.37	HLA-A*68:01
NP	E12	MDLHSLLEL	1-9	0.553	A	-8	476.8	HLA-A*02:06
	E13	TAPHVRNKK	14-22	1.708	A	-8.2	131	HLA-A*68:01
	E14	HVRNKKVIL	17-25	1.082	N	-7.6	34	HLA-A*30:01
	E15	YLRDAGYEF	80-88	0.532	N	-8.9	30.84	HLA-B*15:01
	E16	EPHYSPLIL	106-114	1.385	N	-8.9	213	HLA-B*53:01
	E17	VLIHQGVNL	195-203	0.503	A	-8.3	122.4	HLA-A*02:03
	E18	LIHQGVNLV	196-204	0.638	N	-8.1	151.4	HLA-A*02:03
	E19	LEHGLYPQL	290-298	0.641	A	-9.7	17.7	HLA-B*40:01
	E20	GLYPQLSAI	293-301	0.951	N	-8.4	3	HLA-A*02:03
	E21	YPQLSAIAL	295-303	1.23	N	-8.7	10	HLA-B*35:01
	E22	TEITHSQTL	366-374	0.671	N	-8.1	10.23	HLA-B*40:01
	E23	TVQARPINR(L,A)	422-430	1.262	A	-7.7	33.5	HLA-A*31:01
		TVQARSINR(R)	422-430	0.969	A	-7.7	28.15	HLA-A*31:01

		TVQARPMNR(M)	422-430	1.417	N	-7.7	35.75	HLA-A*31:01
	E24	SFVDLNDPF	455-463	0.824	A	-9	161.3	HLA-A*23:01

For each epitope in Table 5.1.1, I list the allele that had the lowest IC_{50}. I also list the strongest binding energy to any allele for each epitope. These range from −7.6 to −10.9 kcal/mol. Interestingly, most epitopes bind to a similar binding site of the alleles (discussed later). Multiple epitopes prefer to bind to the same allele, for example E2, E10, E17, E18 and E20 epitopes share binding to the HLA-A*02:03 allele, and E7 E8, E19 and E22 epitopes share binding to the HLA-B*40:01 allele (Table 5.1.1)[139,140]. However, some epitopes prefer to bind to a specific allele that does not bind to other epitopes. For example, E3 binds to HLA-B*44:02, E6 binds to HLA-A*68:02, E15 binds to HLA-B*15:01, E16 binds to HLA-B*53:01. (Table 5.1.1).

5.1.2 Promising GP and NP MHC-I epitope-allele complexes

Of the listed GP epitopes, E9 shows the strongest allele binding affinity of -10.9 kcal/mol and has a non-allergenic nature (Table 5.1.1). Of the NP epitopes, E19 has the strongest allele binding affinity of -9.7 kcal/mol but has an allergenic nature (Table 5.1.1). The non-allergenic NP epitopes E15 and E16 both have a reasonably strong allele binding affinity of -8.9 kcal/mol with antigenicity of 0.5322 and 1.3854 respectively. The IC_{50} comparison of E15 and E16 shows that E15 could be a better candidate (IC_{50} value 30.84) compared to E16 (IC_{50} value 213). This result indicates that the E9 and E15 epitopes can be promising candidates against MARV.

Pharmaceutically, the most interesting epitopes are non-allergenic with low IC_{50}. To provide more molecular details, I chose the four GP epitope-allele complexes and the four NP complexes with the best combination of non-allergenicity and lowest IC_{50} values for all-atom MD investigations. For GP, these are: E5:HLA-A*68:01, E6: HLA-A*68:01 (Ravn), E8:HLA-B*40:01 and E9:HLA-A*23:01 complexes, and for NP: E15:HLA-B*15:01, E20:HLA-A*02:03, E21:HLA-B*35:01 and E22:HLA-B*40:01 complexes. The MD results are discussed in detail in section 5.5.

5.2 Prediction of MHC-II T-cell epitopes for MARV GP and NP

MHC-II T-cell epitope prediction with the MARV GP and NP sequences were carried out using three different independent methods: EpiTOP 3.0, NetMHCII 2.3, and ProPred. ProPred uses a quantitative matrix approach[98], NetMHCII2.3 uses an artificial neural network (ANN) machine learning[99] algorithm, while EpiTOP 3.0 uses proteochemometrics such as Quantitative Structure–Activity Relationship models (QSAR)[100] to predict the MHC-II T-cell epitopes. The consensus epitopes predicted by all three methods are listed in Table 5.2.1.

Among these consensus MHC-II T-cell epitope predictions that are non-allergenic, the E32 and E33 for GP, and E39 for NP have the highest antigenicity score and are good candidates for experimental validation and design of MHC-II T-cell based vaccines. Intriguingly for the E32 and E33 GP consensus epitopes listed in Table 5.2.1, the closely related MARV strains (Lake Victoria, Angola, Musoke, Ravn) have comparable antigenicity and allergenicity properties.

Table 5.2.1. Prediction of consensus MHC-II T-cell epitopes for GP and NP. The Lake Victoria (L) strain is usually the most conserved sequence and therefore used as the principal sequence for each epitope. Many epitopes have identical predicted sequences in the four Marburg strains. For epitopes that have slightly different predicted sequences in different Marburg strains, I provide the different sequences with strains abbreviated as Angola(A), Musoke(M), Ravn(R).

Protein	Epitopes	Sequence	Interval	Antigenicity	Allergenicity
GP					
	E25	LILIQGIKT(L)	10-18	0.4537	A
		LILIQGVKT(A)		0.6736	A
		LILIQGTKN(M)		0.8576	A
		LILIQSIKT(R)		0.5091	N
	E26	LIQGIKTLP(L)	12-20	0.2459	N
		LIQGVKTLP(A)		0.3951	N
		LIQGTKNLP(M)		0.3715	N
		LIQSIKTLP(R)		0.2351	N
	E27	LLLDPPTNI(L,A,M)	105-113	-0.5339	N
		LLLDPPSNI(R)		-0.4772	N
	E28	FLYDRIAST(L,A,M)	144-152	0.2317	N
		FLYDRVAST(R)		0.3004	N
	E29	MIVNKTVHK(L,A,M)	168-176	-0.0092	A
		MIVNKTVHR(R)		-0.0067	A
	E30	IVNKTVHKM(L,A,M)	169-177	-0.051	A
		IVNKTVHRM(R)		-0.0598	A
	E31	YRHMNLTST	186-194	1.8501	A
	E32	VYFRKKRSI(L)	429-437	1.3083	N
		VYFRRKRNI(A)		1.5966	N
		VYFRRKRSI(M)		1.4115	N
		IYFRKKRSI(R)		1.526	N
	E33	FRKKRSILW(L)	431-439	1.2182	N
		FRRKRNILW(A)		1.6415	N

Protein	Epitopes	Sequence	Interval	Antigenicity	Allergenicity
		FRRKRSILW(M)		1.2783	N
		FRKKRSIFW(R)		1.4661	N
	E34	LVCRLRRLA	555-563	0.72	A
NP					
	E35	ILALKTLES	113-121	0.7663	A
	E36	IGLFLSFCS	129-137	-0.0675	N
	E37	FILQKTDSG	235-243	0.7589	A
	E38	FKQALSNLA	262-270	-0.3282	N
	E39	FARVLNLSG	278-286	0.0884	N
	E40	ITHSQTLAV	368-376	0.5329	A

5.3 Prediction of B-cell Epitopes

In addition to the MHC-I and MHC-II T-cell epitope predictions, I performed the sequence and structure-based B-cell epitope predictions for GP and NP of different MARV strains. BepiPred 2.0 and BcePred servers were separately employed to predict sequence-based B-cell epitopes. The BepiPred uses a random forest algorithm that is trained on epitopes' annotated antibody-antigen structures, while BcePred predicts the B-cell epitopes using four physicochemical properties: hydrophilicity, polarity, exposed surface, and beta turns. Table 5.3.1 lists all GP and NP epitopes predicted by either server. The E44-E51 GP epitopes and the E55-E68 NP epitopes were predicted by both servers. As with T-cell epitopes, the Lake Victoria strain of the B-cell epitopes had the most conserved sequence and is listed as the principal sequence for each epitope.

The structure-based B-cell epitope predictions were performed using structural properties-based servers, Ellipro and DiscoTope. For this, 3D models for

the GP and NP of the Lake Victoria, Angola, Musoke and Ravn strains were modelled using the structures available in the PDB (GP: ID 6BP2, NP: ID 5F5M) as templates. The Ellipro algorithm predicts epitopes by incorporating the antigenicity, solvent accessibility, and flexibility of protein structures[141], while DiscoTope predicts residue-by-residue conformational epitopes by using amino acid statistics, spatial information, and surface accessibility[142]. The GP epitopes E42, E43, E48, E49, E51, E54 and the NP epitope E55 (Table 5.3.1) were predicted by both structure-based methods.

Table 5.3.1 shows that the B-cell epitopes E48, E49, E51, E55 were predicted by all of the sequence and structure-based prediction methods and thus were top ranked compared to other epitopes. This indicates that these can be promising epitope candidates for experimental validation and design of B-cell epitopes for MARV. Unlike the T-cell GP epitopes, none of the B-cell GP epitopes had identical sequences for all four strains except for E53.

Table 5.3.1 Prediction of consensus B-cell epitopes for GP and NP. The Lake Victoria (L) strain is usually the most conserved sequence and therefore used as the principal sequence for each epitope. For epitopes that have slightly different predicted sequences in different Marburg strains, I provide the different sequences with strains abbreviated as Angola(A), Musoke(M), Ravn(R). Boldface entries are epitopes that were predicted by all four methods.

Protein	Epitopes	Sequence	Interval	Length	Sequence-based		Structure-based		Rank	Antigenicity	Allergenicity
					BepiPre	BcePre	Ellipro	Discotop			
GP											
	E41	TGVPPKNVEYTEGEEAKTCY(L,M,R)	74-93	20	×	✓	✓	×	2/4	1.0498	N
		AGVPPKNVEYTEGEEAKTCY(A)								1.0603	N
	E42	TNIRDYPK(L,A,M)	111-	8	×	✓	✓	✓	3/4	0.819	A
		SNIRDYPK(R)								0.6469	A
	E43	PKCKTIHHIQGQNPHAQGIA(L,A,M)	117-	20	×	✓	✓	✓	3/4	0.6897	N
		PKCKTVHHIQGQNPHAQGIA(R)								0.6843	N
	E44	STPQQEGNNTDHSQG(L)	303-	15	✓	✓	×	×	2/4	0.6339	A
		STPQQGGNNTNHSQG(A)								0.4733	A
		STPQQGGNNTNHSQD(M)								0.5236	A
		STSQHEQNSTNPSRH(R)								0.6862	A
	E45	AQPSMPPHNTTAISTNNTSK(L)	329-	20	✓	✓	×	×	2/4	0.5846	A
		AQPSMPPHNTTTISTNNTSK(A,M)								0.5562	A
		TQPATLLNNTNTTPTYNTLK(R)								0.5	N

E46	TTTAPNMTNGHLTSPSPTPN(L)	403-	20	✓	✓	×	×	2/4	0.6614	N
	TTTVPNTTNKYSTSPSPTPN(A)								0.5967	N
	ATTAPNTTNEHFTSPPPTPS(M)								0.5568	N
	IMTTSDITSKHPTNSSPDSS(R)								0.6895	A
E47	NMTNGHLTSPSPTPNPT (L)	408-	17	✓	✓	×	×	2/4	0.6929	N
	NTTNKYSTSPSPTPNST (A)								0.5847	N
	NTTNEHFTSPPPTPSST (M)								0.5348	N
	DITSKHPTNSSPDSSPT(R)								0.8631	N
E48	**TIFDESSSSGASAEEDQHASP**(L,A,M)	**466-**	**21**	✓	✓	✓	✓	**4/4**	**0.4934**	**N**
	TIFDESPSFNTSTNEEQHTPP(R)								0.3299	A
E49	**FDESSSSGASAEEDQHASPN**(L,A,M)	**468-**	**20**	✓	✓	✓	✓	**4/4**	**0.5673**	**N**
	FDESPSFNTSTNEEQHTPPN(R)								0.1712	N
E50	PNINENTAYSGENENDCDAE(L,M)	496-	20	✓	✓	✓	×	3/4	0.4469	N
	PKVNENTAHSGENENDCDAE(A)								0.3503	A
	PDKNGDTAYSGENENDCDAE(R)								0.4963	N
E51	**TAYSGENENDC**(L,M,R)	**502-**	**11**	✓	✓	✓	✓	**4/4**	**0.7604**	**N**
	TAHSGENENDC(A)								0.695	N
E52	GPGIEGLYTAGLIKNQNNLV(L,A,R)	537-	20	×	×	✓	×	1/4	0.5119	N
	GPGIEGLYTAVLIKNQNNLV(M)								0.6288	N
E53	CKVLGPDCCIG	602-	11	×	✓	✓	×	2/4	0.1338	A
E54	QIDQIKKDEQKEGTGWGLGG(L,A,M)	623-	20	×	✓	✓	✓	3/4	0.1721	N
	QIDKIRKDEQKEETGWGLGG(R)								0.2547	N

NP	**E55**	**MDLHSLLELGTKPTAPHV**	**1-18**	**18**	✓	✓	✓	✓	**4/4**	**0.8425**	**N**
	E56	KNADATRFLDVIPNEPHYSPL(A,M)	92-112	21	✓	✓	✓	×	3/4	0.5486	N
		KNPDATRFLEVIPNEPHYSPL(R)								0.3473	N
	E57	KTLESTESQRG	117-	11	✓	✓	✓	×	3/4	0.8644	A
	E58	DSGVTLHPLVRTSKVKNEVAS(A,M)	241-	21	✓	✓	×	✓	3/4	0.4838	A
		DSGVALHPLVRTSKVKNEVAS(R)								0.4448	A
	E59	VLSQKREKLARL	376-	12	✓	✓	×	✓	3/4	0.5416	A
	E60	IENNIVEDQGFKQSQNR (A,M)	391-	17	✓	✓	✓	×	3/4	-0.2684	N
		IENNIAEDQGFKQSQNQ (R)								-0.1168	N
	E61	QSQDIDNSQGKQEDESTNLIKKP(A)	496-	23	✓	✓	✓	×	3/4	0.4557	N
		QSQDLNNSQGKQEDESTNPIKKQ(M)								0.6731	A
		QTQDLDISQKKQGNESTDPARKQ(R)								1.1635	A
	E62	ESIDQPGSDNEQGVDLPPPPL(A,M)	542-	21	✓	✓	✓	×	3/4	0.2922	N
		ESDDQPGSDNEQGVDLPPPPL(R)								0.6641	N
	E63	PPLYAQEKRQDPIQHP	560-	16	✓	✓	✓	×	3/4	0.822	A
	E64	DPIQHPAVSSQDPFGSIGDVNGDILEPI(A)	570-	28	✓	✓	✓	×	3/4	0.1342	A
		DPIQHPAANPQDPFGSIGDVNGDILEPI(M)								-0.0206	N
		DPIQHPAVSSQDPFGSIGDVDGDILEPI(R)								0.1103	A
	E65	RSPSSPSAPQ	598-	10	✓	✓	✓	×	3/4	0.1853	N
	E66	EDTRAREAYELSPDFTN (A)	608-	17	✓	✓	✓	×	3/4	0.9784	N
		EDTRMREAYELSPDFTN (M)								0.9384	A

	EDTRMGEAYELSPDFTS (R)								1.1065	N
E67	YEDNQQNWPQR(A,R)	625-	11	✓	✓	✓	×	3/4	0.2003	N
	DEDNQQNWPQR(M)								0.1383	N
E68	VVTKKGRTFLYPNDLL	636-	16	✓	✓	✓	×	3/4	0.6377	A

5.4 GP and NP epitope surface mapping

The ability of antibodies to access epitopes is a crucial property for vaccine candidates[143]. For Figure 5.1, I used VMD to display some of the four GP and four NP epitopes that I chose for MD simulations, as described in the Methods section. I display the most visually clear structures in the PDB: GP PDB 5UQY, and NP PDB:5F5M. In addition, Figure 5.1 also displays other epitopes that I found to be novel, as described later.

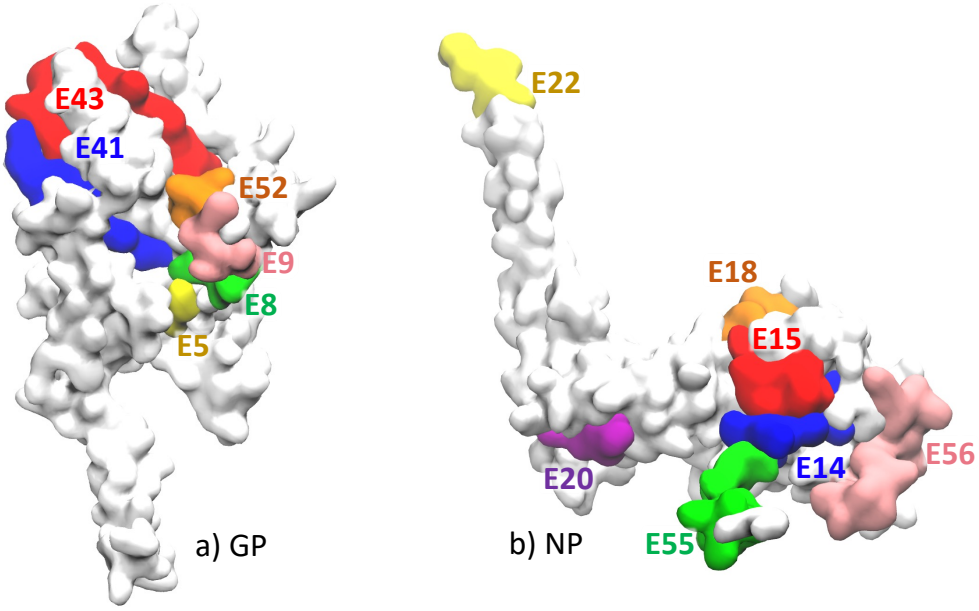


Figure 5.1: Surface mapping of some non-allergenic and good antigenicity epitopes chosen for MD simulations as well as some novel epitopes, for a) GP and b) NP structures.

In Table 5.4.1, for each epitope displayed in Figure 5.1 I list the solvent accessible surface area per amino acid (SASA/AA) in the epitope. E14 is a novel epitope and is especially interesting as a potential epitope candidate because it has the largest SASA per amino acid, is non-allergenic (Table 5.1.1), has low IC_{50} and mid-range values of antigenicity and binding affinity.

Table 5.4.1. For each epitope displayed in Figure 5.1, I list the solvent accessible surface area per amino acid (SASA/AA) in the epitope.

Epitope	Residues	SASA($Å^2$)	SASA($Å^2$) per amino acid
E5	164-172	1340.85	148.98
E8	522-530	1131.28	125.70
E9	527-535	1083.39	120.38
E14	17-25	1554.14	172.68
E15	80-88	1154.02	128.22
E18	196-204	1136.49	126.28
E20	293-301	1095.62	121.74
E22	366-374	1388.71	154.30
E41	74-93	2716.32	135.82
E43	117-136	2744.47	137.22
E52	537-556	2227.43	111.37
E55	1-18	2417.74	134.32
E56	92-112	2397.96	114.19

5.5 Dynamics of the MHC-I allele-epitope complex

In order to investigate the dynamics and stability of the selected consensus non-allergenic epitopes that have low IC_{50} values with specific alleles, I performed 200 ns all atom, explicit solvent MD simulations for four GP and four NP MHC-I T-cell epitope-allele complexes with the lowest IC_{50} and non-allergenic epitopes. These are GP epitope-allele complexes: E5-HLA-A*68:01(A8), E6-HLA-A*68:01(A8), E8-HLA-B*40:01(A26), E9-HLA-A*23:01(A18), and NP epitope-allele complexes: E15-HLA-B*15:01(A11), E20-HLA-A*02:03(A3), E21-HLA-B*35:01(A12), E22-

HLA-B*40:01(A26). A list of alleles and their PDB codes is provided in Table 5.5.1. The epitope-allele docking complexes are shown in Figure 5.2.

Table 5.5.1: List of the 27 alleles associated with the MHC-I T-cell epitopes. The PDB code is given for the first 17 alleles. For the last 10 alleles, M indicates that the allele structure was modelled using Swiss-Model.

No.	Alleles	PDB ID	No.	Alleles	PDB ID
A1	HLA-A*0101	3BO8	A15	HLA-B*5101	1E27
A2	HLA-A*0201	5ENW	A16	HLA-B*5701	5T6W
A3	HLA-A*0203	3OX8	A17	HLA-B*5801	5IM7
A4	HLA-A*0206	3OXR	A18	HLA-A*2301	M
A5	HLA-A*0301	3RL1	A19	HLA-A*2601	M
A6	HLA-A*1101	1X7Q	A20	HLA-A*3001	M
A7	HLA-A*2402	3VXN	A21	HLA-A*3002	M
A8	HLA-A*6801	4HWZ	A22	HLA-A*3101	M
A9	HLA-B*0702	5EO0	A23	HLA-A*3201	M
A10	HLA-B*0801	3X14	A24	HLA-A*3301	M
A11	HLA-B*1501	1XR8	A25	HLA-A*6802	M
A12	HLA-B*3501	2CIK	A26	HLA-B*4001	M
A13	HLA-B*4402	1M6O	A27	HLA-B*5301	M
A14	HLA-B*4403	1N2R			

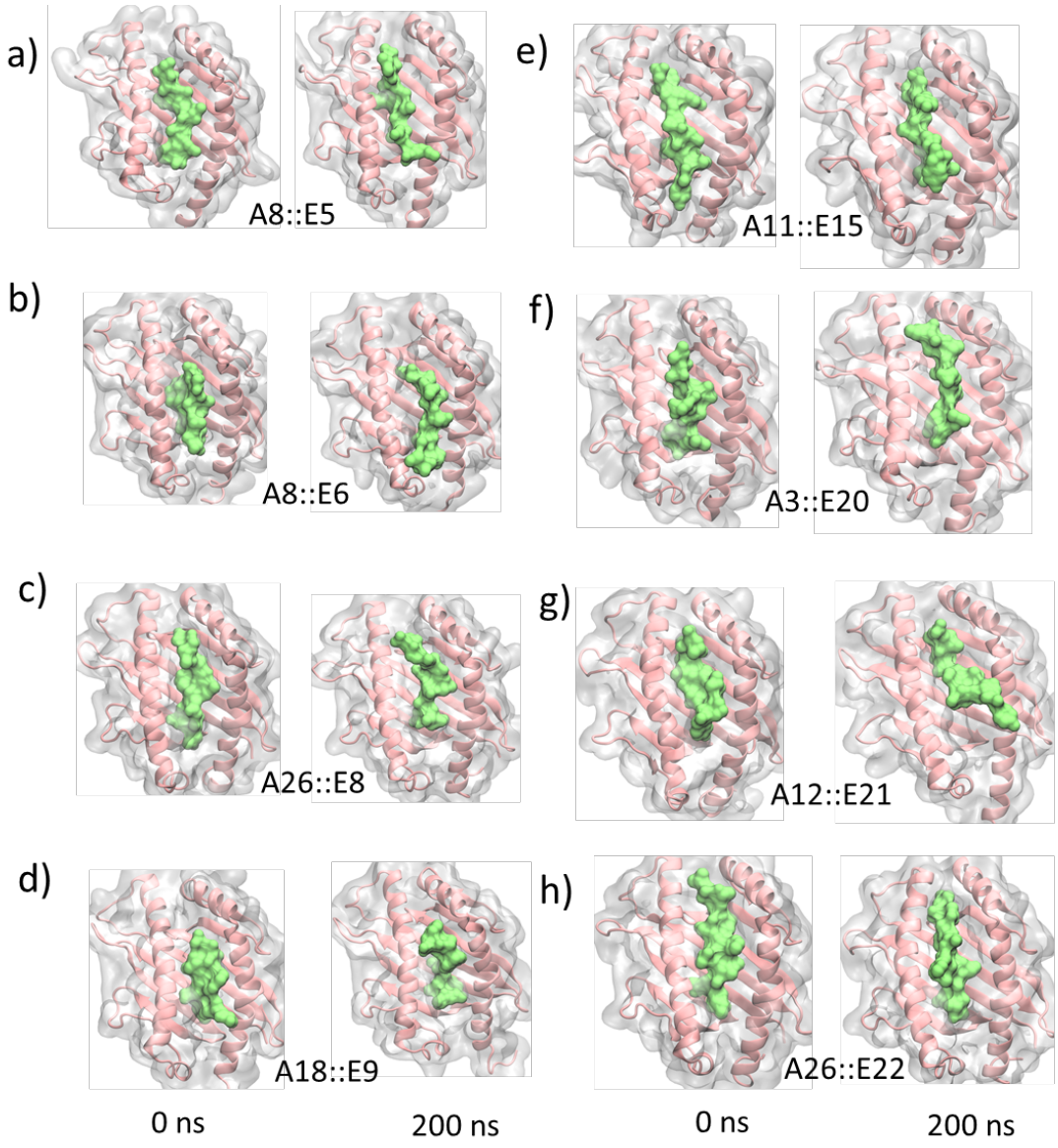


Figure 5.2: Snapshots of allele(pink)-epitope(green) complexes at the beginning 0 ns (initial minimized structure) and end (200 ns) of the MD simulations. GP: (a) A8::E5, (b) A8::E6, (c) A26::E8, (d) and A18::E9, and NP: e) A11::E15, f) A3::E20, g) A12::E21 and h) A26::E22.

5.5.1 RMSD of allele, epitope and complex

In Figure 5.3, I display the RMSD of the epitope-allele complexes during the 200 ns MD simulations. The RMSD is calculated with respect to the initial structures obtained from the docking servers. The noticeable size of the RMSD in Figure 5.3 shows that in order to investigate molecular level details of the epitope-allele binding, MD simulations are crucial for obtaining the precise structural arrangement.

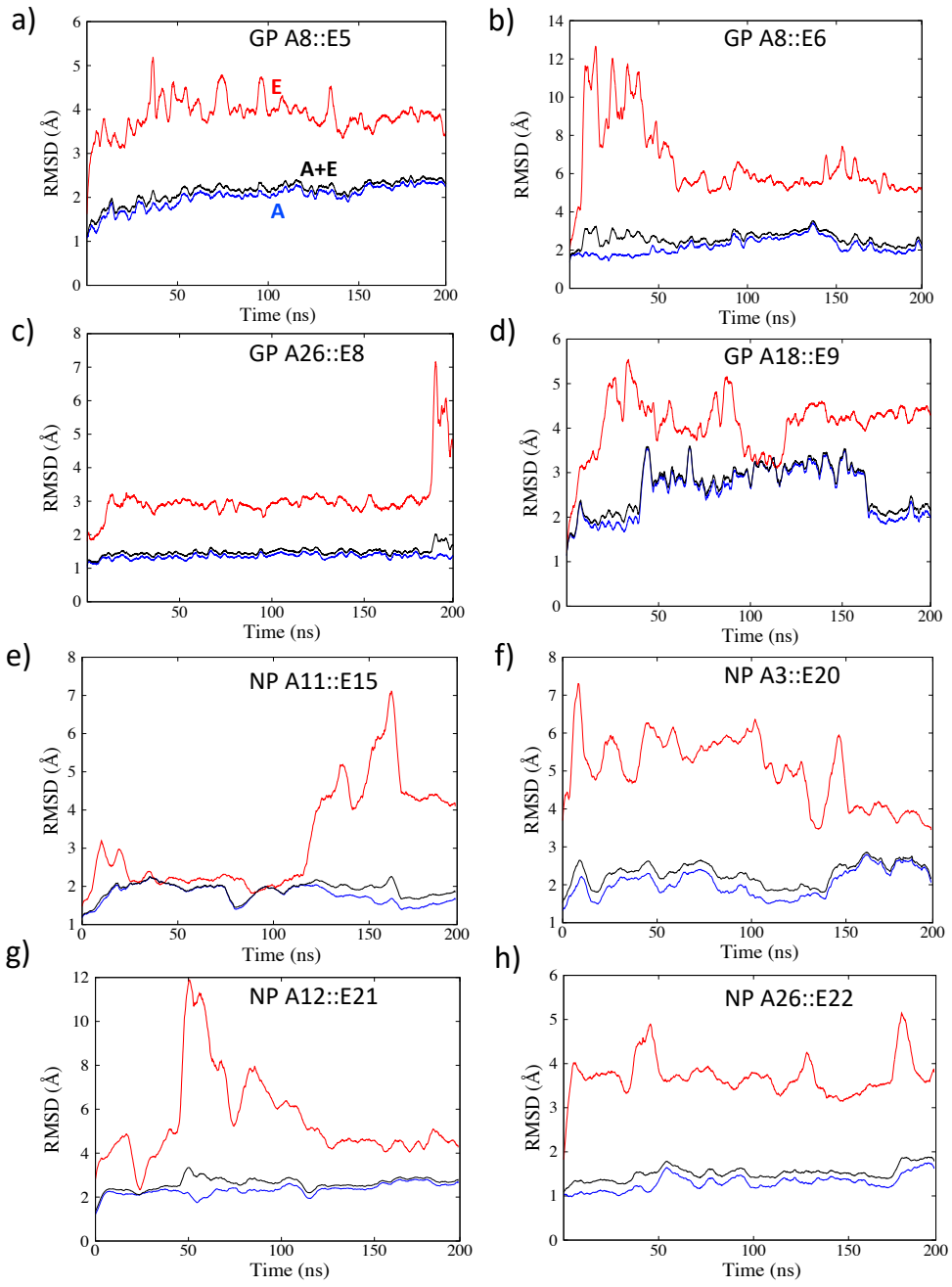


Figure 5.3. RMSD of backbone atoms from 200 ns MD simulations of selected epitope-allele complexes showing movement from the initial structure obtained from docking. As expected, the majority of the epitope-allele relative motion is performed by the epitope, which is smaller than the allele.

5.5.2 Hydrogen bonds and interaction energy

The RMSD analysis shows that the MHC-I T-cell epitope-allele complexes have stabilized by the end of our 200 ns MD simulations. To understand which molecular interactions are responsible for tight binding, I calculated the number of hydrogen bonds (H-bond) between the epitope and allele. The hydrogen bonds were calculated between the protein interface atoms within a cut-off distance of 3.5 Aº and a cut-off angle of 30º between the donor and acceptor heavy atoms. The number of H-bonds as a function of time are shown in Figure 5.4a for the GP complexes and Figure 5.4c for the NP complexes. In addition to the H-bonds, I calculated the interaction energy (electrostatic and van der Waals) between the allele and epitope in allele-epitope complexes. Figure 5.4 shows the interaction energy as a function of the MD simulation time. The interaction energy fluctuates similarly to the number of H-bonds. The strongest interaction energy for GP is for complex A8::E5 and for NP it is for complex A11::E15.

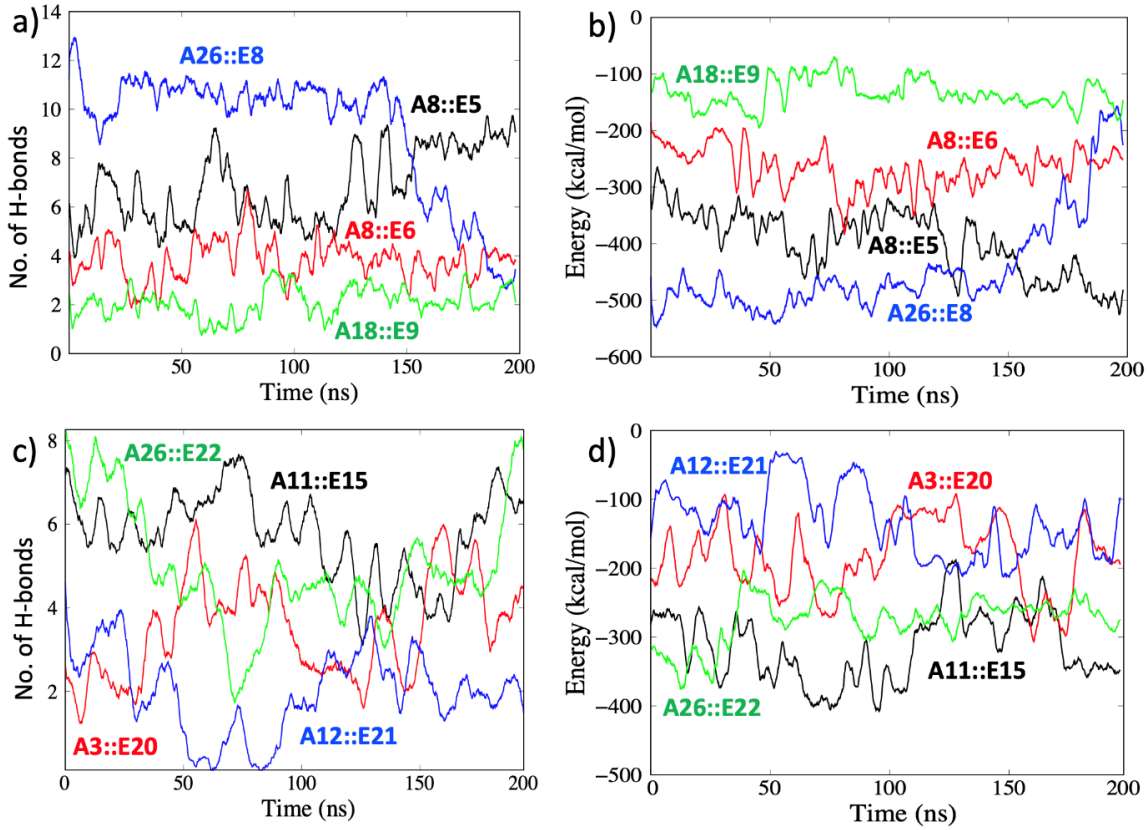


Figure 5.4. Epitope-allele H-bond number and interaction energy as a function of MD simulation time. GP complexes: (a) and (b); NP complexes: (c) and (d). As explained earlier, the MHC-I T-cell epitope prediction servers focus on epitopes that are nanomers for both GP and NP.

5.5.3 Surface Accessible Surface Area

I also calculated solvent-accessible surface area (SASA) at the epitope-allele interface to characterize the water excluded region and hydrophobic interactions. I used a 1.4 Å water radius. I probed the epitope region that is within 10 Å of any allele amino acid, and likewise probed the allele region that is within 10 Å of any epitope amino acid. The results are shown in Figure 5.5.

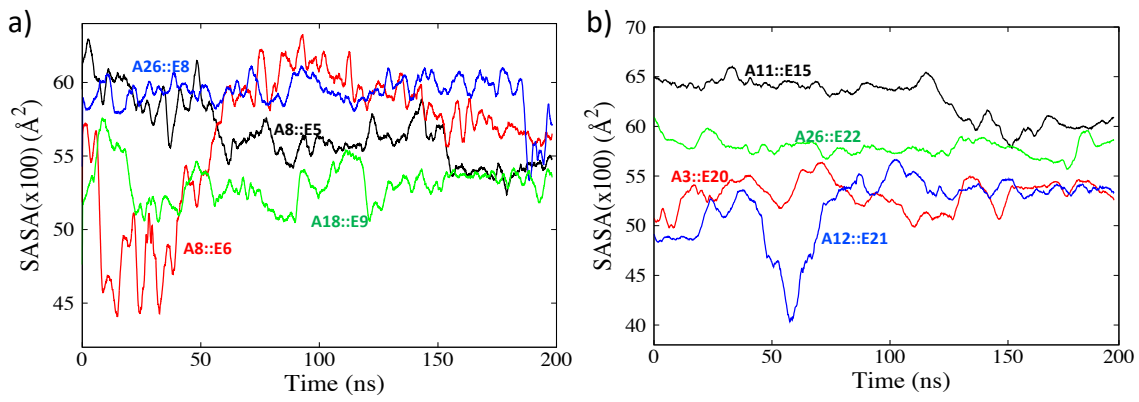


Figure 5.5. SASA calculated for epitope-allele interfaces as a function of simulations time for a) GP and b) NP.

Interestingly, the GP A18::E9 complex has the lowest SASA, implying a tight interface, even though Figure 5.4 shows it has the lowest number of H-bonds and weakest interaction energy among the GP complexes. Similarly, for NP, the two complexes with the fewest number of H-bonds and weakest interaction energy (A3::E20 and A12::E21) have the smallest SASA value. Thus, the weakest bound complexes have the tightest epitope-allele interface with respect to excluding water. To explain this seeming paradox, I counted the number of hydrophobic amino acids in all eight nanomer epitopes examined in Figure 5.4 and Figure 5.5.

Table 5.5.2 shows that the epitope-allele complexes with the lowest interface SASA have the highest number of hydrophobic residues. Thus, the interactions between epitopes and alleles depend not only on enthalpic interactions such as H-bonds, but also on entropically driven effects such as hydrophobic interactions.

Table 5.5.2. Number of hydrophobic residues in each epitope investigated in Figure 5.4 and Figure 5.5. All epitopes are nanomers. Hydrophobic residues are underlined.

GP		NP	
Epitope	No. of hydrophobic residues	Epitope	No. of hydrophobic residues
E5 NIAAMIVNK	6	E15 YLRDAGYEF	4
E6 MTTSDITSK	2	E20 GLYPQLSAI	6
E8 QEDDLAAGL	5	E21 YPQLSAIAL	6
E9 AAGLSWIPF	8	E22 TEITHSQTL	2

5.6 T- and B-cell epitope prediction for MARV VP35, VP40, VP30, VP24 and L proteins

In addition to the prediction and detailed analysis of GP and NP epitopes above, I performed epitope prediction for the remaining MARV Proteins which includes proteins VP35, VP40, VP30, VP24 and L using the above-mentioned MHC-I and -II and B-cell epitope prediction servers on the Lake Victoria strain. The prediction results are listed in Table 5.6.1. After determining the epitopes for the Lake Victoria strain, I inspected the same sequence in the other strains. The Angola strain proteins always had the same sequence as the Lake Victoria strain. For some epitopes, the Musoke strain or Ravn strain had slightly different sequences, which are listed in Table 5.6.1. The antigenicity and allergenicity of consensus epitopes predicted by three or more methods for VP35, VP40, VP30, VP24 and L proteins are listed in Table 5.6.2.

The consensus, antigenicity and allergenicity analyses show that VP35 E78 epitope has the best antigenic value of 0.911 and is a non-allergic epitope, while VP40 E85 and E86 have antigenicity values of 1.093 and 1.112, respectively, and both are non-allergic epitopes, indicating that they could be promising epitope candidates among VP35 and VP40 epitopes. In addition, of the predicted VP30 and VP24 epitopes, E93 and E97 (Ravn) have high antigenicity of 1.1239 and 1.6259, respectively, and both are non-allergic. Among the predicted epitopes listed in Table T9 that are non-allergenic, the L protein's E202 has an especially high antigenicity of close to 2.0. These analyses clearly suggest that VP35, VP40, VP30, VP24 and L proteins contain promising epitopes.

Table 5.6.1: Epitope predictions for VP35, VP40, VP30, VP24 and L proteins using MHC-I, MHC-II and B cell prediction methods for the Lake Victoria strain. Predicted epitope sequences were identical in all four strains except where noted by R (Ravn) or M (Musoke).

		MHC-I T cell		MHC-II T cell		B-cell	
Proteins	**Sequence**	**Pro Pred-1**	**CTL-Pred**	**Epi-TOP**	**NetMHCII-2.3**	**Bepi-pred**	**Bce-Pred**
VP35	YMQQVSEGL	✓	×	✓	×	✓	×
	LEKLYKRRK	×	×	×	✓	✓	✓
	LYKRRKPKG	×	×	✓	✓	✓	✓
	YKRRKPKGT	×	×	✓	✓	✓	✓
	KRRKPKGTV	×	✓	×	×	✓	✓
	KMGRTLE KMGKTLE(R)	×	×	✓	✓	×	✓
	MSKVLELSE	×	×	✓	×	✓	✓
	LELSEETFS	×	×	✓	×	✓	✓
	LSEETFSKP	×	×	✓	×	✓	✓

Proteins	Sequence	MHC-I T cell		MHC-II T cell		B-cell	
		Pro Pred-1	CTL-Pred	Epi-TOP	NetMHCII-2.3	Bepi-pred	Bce-Pred
	FSKPNLSAK	×	×	✓	×	✓	✓
	FQTVPRPCQ	×	×	✓	✓	✓	✓
	VPRPCQKSL	✓	×	×	×	✓	✓
VP40	NYNTYMQYL	×	✓	×	×	✓	✓
	YNTYMQYLN	×	×	✓	×	✓	✓
	YMQYLNPPP	×	×	×	✓	✓	✓
	MQYLNPPPY	✓	×	×	✓	✓	✓
	QYLNPPPYA	×	×	×	✓	✓	✓
	YVGDLNLDD	×	×	✓	✓	✓	×
	SNFEYPL	✓	×	✓	×	✓	×
	FIQNMVIPR FVQNMVIPR(R)	×	×	✓	✓	✓	×
	FSTNQFTYN	×	×	✓	✓	×	✓
	HPNLPPIVL	✓	✓	×	×	×	✓
	YRQHKNPNN	×	×	×	✓	✓	✓
	LRVEKVPEK	×	✓	×	×	✓	✓
VP30	GKLDETS	✓	×	✓	×	×	✓
	IHLDKGGQF	×	×	✓	✓	✓	
	DKGGQFE	×	×	✓	×	✓	✓
VP24	LSTRYNLPV LSTRYNLPA(M) LSTRYNLPT(R)	×	×	✓	×	✓	✓
	NVTEKSINL NVTENSINL(M) NITEKSINL(R)	✓	✓	×	×	✓	✓
	VTEKSINLD VTENSINLD(M) ITEKSINLD(R)	×	×	✓	×	✓	✓

Proteins	Sequence	MHC-I T cell		MHC-II T cell		B-cell	
		Pro Pred-1	CTL-Pred	Epi-TOP	NetMHCII-2.3	Bepi-pred	Bce-Pred
	INLDLNSTA	×	×	✓	✓	×	✓
	LLHHLKSNF	×	×	✓	✓	×	✓
	LKSNFVVPE	×	×	✓	×	✓	✓
	DQELQQS	✓	×	✓	×	✓	✓
	HISPNLLG	×	×	✓	✓	✓	
	FLVEVRRID	×	×	✓	✓	×	✓
	EPCCGETVL	✓	×	×	×	✓	✓
	LSESVVFGL	✓	×	✓	×		✓
	MEKGQPLNL IERGQPLNL(R)	×	×	✓	×	✓	✓
	LNLTQYMNS	×	×	✓	×	✓	✓
L	HYSHNPKLR	×	✓	×	×	✓	✓
	YSHNPKLRN	×	×	✓	×	✓	✓
	SHNPKLRNC	×	✓	×	×	✓	✓
	NPKLRNCRI	×	✓	×	×	✓	✓
	RNCRIPHHI RNCRIPYHI(R)	×	✓	×	✓	✓	✓
	RIPHHIYRL RIPYHIYRL(R)	×	×	×	✓	✓	✓
	IPHHIYRLR IPYHIYRLR(R)	×	×	✓	×	✓	✓
	PHHIYRLRN PYHIYRLRN(R)	×	×	×	✓	✓	✓
	HIYRLRNST	×	×	×	✓	✓	✓
	IYRLRNSTA	×	×	×	✓	✓	✓
	YRLRNSTAL	×	×	✓	×	✓	✓
	RLRNSTALK	×	×	×	✓	✓	✓
	LRNSTALKT	×	×	✓	✓	✓	
	NHVDDFKY NHINDFKY(R)	×	×	✓	×	✓	✓

Proteins	Sequence	MHC-I T cell		MHC-II T cell		B-cell	
		Pro Pred-1	CTL-Pred	Epi-TOP	NetMHCII-2.3	Bepi-pred	Bce-Pred
	HVDDFKYLL	✓	✓	×	×	✓	×
	HINDFKYLL(R)						
	FKYLLPSEL	×	×	✓	✓	✓	×
	LARRIKGQR	×	×	✓	✓	×	✓
	IKGQRGSLR	×	×	✓	×	✓	✓
	IIRNAVSLQ	×	×	✓	✓	✓	×
	IIRNATSLQ(R)						
	IRNAVSLQA	×	×	✓	✓	✓	×
	IRNATSLQA(R)						
	FKLIKHLEP	×	×	✓	✓	×	✓
	LIKHLEPLC	×	×	✓	✓	×	✓
	IKHLEPLCV	×	×	✓	✓	×	✓
	FSLQKHWGH	×	×	✓	✓	×	✓
	YHSQGSWYK	×	×	✓	×	✓	✓
	WYKTTHDLH	×	×	✓	×	✓	✓
	IISDLSIFI	✓	×	✓	✓	×	×
	WDSVFDRSV	×	✓	✓	×	✓	×
	LGYNPPVRF	×	×	✓	✓	✓	×
	SLKEKELNI	×	✓	×	×	✓	✓
	YRVRNVQTL	✓	✓	✓	×	×	×
	VTEREQKEA	×	×	✓	×	✓	✓
	LLHQASWHH	×	×	✓	×	✓	✓
	YNLAFRYEF	×	✓	✓	×	×	✓
	LKTKLKLKS	×	×	✓	✓	×	✓
	YQENEAELN	✓	×	✓	×	×	✓
	WIASFHSML	✓	×	✓	✓	×	×
	WIAAFHSML(R)						
	FHSMLAINL	✓	×	✓	✓	×	×
	FHSMLAVNL(R)						
	LIALITPQV	✓	×	✓	✓	×	×
	LFQLKNALE	✓	×	✓	✓	×	×

Proteins	Sequence	MHC-I T cell		MHC-II T cell		B-cell	
		Pro Pred-1	CTL-Pred	Epi-TOP	NetMHCII-2.3	Bepi-pred	Bce-Pred
	LFQLRNALE(R)						
	LEFLEKEEL	✓	×	✓	×	×	✓
	LEFLRKEEL(R)						
	ILIAKKPGL	✓	×	✓	✓	×	×
	ILISKKPGL(M)						
	LRQTVRENI	×	×	✓	✓	×	✓
	LEDQRVCEW	×	✓	✓	×	✓	×
	FSRTPSGKR	×	×	✓	✓		✓
	LTTEGTMLM	×	×	✓	✓	✓	×
	MLMKLRELT	×	×	✓	✓	×	✓
	MLMRLRELT(R)						
	LMKLRELTR	×	×	✓	✓	×	✓
	LMRLRELTK(R)						
	MKLRELTRN	×	×	✓	✓	×	✓
	MRLRELTKS(R)						
	LDDDLSESL	×	×	✓	×	✓	✓
	LSESLEKFT	✓	×	✓	×	✓	✓
	LSESLEKFI(R)						
	WSDVLKG	×	×	✓	×	✓	✓
	LSEDLREQF	×	×	✓	×	✓	✓
	LSEDLKEQF(R)						
	LREQFNLSS	×	×	✓	✓	✓	×
	LKEQFKLSS(R)						
	FLPYDCKEL	✓	×	✓	×	✓	×
	LLPYDCKEL(M)						
	LLQYDCNGL(R)						
	YDCKELRLG	×	×	✓	×	✓	✓
	YDCKELRLE(M)						
	YDCNGLHSK(R)						
	VVQKHPSVN	×	×	✓	×	✓	✓
	IVQKHPSDN(R)						

Proteins	Sequence	MHC-I T cell		MHC-II T cell		B-cell	
		Pro Pred-1	CTL-Pred	Epi-TOP	NetMHCII-2.3	Bepi-pred	Bce-Pred
	VQKHPSVNR	×	✓	✓	×	✓	✓
	VQKHPSDNR(R)						
	KIGYPPLRV	×	✓	×	×	✓	✓
	IGYPPLRVN	×	×	✓	×	✓	✓
	FQNTINLGV	✓	×	✓	✓	×	×
	LFFDKPLDV	×	×	✓	✓	✓	×
	LYFDKPLDV(M)						
	LDVDLNKYM	×	×	✓	×	✓	✓
	LNKYMDNEL	×	×	✓	×	✓	✓
	LCSGIKGRL	✓	×	✓	×	✓	✓
	IKGRLGRVS	×	×	✓	×	✓	✓
	SRSTLSLSL	✓	×	×	✓	✓	×
	SRSTLTLSL(M)						
	IFYAFGANL	✓	×	✓	✓	×	×
	IRNLSHRSL	✓	✓	✓	✓	×	✓
	NLSHRSLRI	×	✓	×	✓	×	✓
	LSHRSLRIL	×	✓	✓	✓	×	✓
	LRILQSTFR	×	×	✓	✓	×	✓
	LQSTFRHEL	×	×	✓	✓	×	✓
	RHELVLTRL	✓	×		✓	×	✓
	LVLTRLAHH	×	×	✓	✓	×	✓
	VLTRLAHHI	×	×	✓	✓	×	✓
	GGSAGEKS	×	×	✓	×	✓	✓
	LIKKGQSS	×	×	✓	×	✓	✓
	LMKKGQSS(M)						
	LMRKNQSP(R)						
	DKVQKRKIL	✓	×	×	×	✓	✓
	DKIQKRKIL(M)						
	KKVQNHRPV(R)						
	VQKRKILAD	×	×	✓	×	✓	✓
	IQKRKILAD(M)						

Proteins	Sequence	MHC-I T cell		MHC-II T cell		B-cell	
		Pro Pred-1	CTL-Pred	Epi-TOP	NetMHCII-2.3	Bepi-pred	Bce-Pred
	VQNHRPVAD(R)						
	ILADTCYPV	×	×	✓	✓	✓	×
	ILADTCCPI(M)						
	PVADTCFLT(R)						
	YYASLNYWR	×	×	✓	✓	×	✓
	YASLNYWRD	×	✓	×	✓	×	✓
	LMNYGSTTL	✓	×	✓	×	✓	×
	LINYGSTAL(M)						
	STNHEPTAL(R)						
	LSEQNLVEN	×	×	✓	×	✓	✓
	YSEQDPAKS(R)						
	LVENCRPSK	×	×	✓	×	✓	✓
	LVENCRPSE(M)						
	PAKSYLLLE(R)						
	IRCKDNQKI	×	×	✓	×	✓	✓
	IRFKDNQKI(M)						
	TRFRDDQKI(R)						
	IIKHDQRYG	×	×	✓	×	✓	✓
	ITKHDQRCE(M)						
	ILRHDQKAE(R)						
	MLPKDNMQT	×	×	✓	×	✓	✓
	MFPEDNMQT(M)						
	VSSRGCLQA(R)						
	IIKSLDVHE	×	×	✓	✓	✓	×
	LIKSLDAHE(M)						
	TTEPLSMLR(R)						
	FLTTLTGTE	✓	×	✓	×	✓	×
	LLTTPTRTE(M)						
	SLKTPMRIE(R)						
	LQPSRYSST	×	×	✓	×	✓	✓
	LQSSRYSST(M)						

Proteins	Sequence	MHC-I T cell		MHC-II T cell		B-cell	
		Pro Pred-1	CTL-Pred	Epi-TOP	NetMHCII-2.3	Bepi-pred	Bce-Pred
	LQPSECLST(R)						
	LSREQASYL	×	×	✓	✓	×	✓
	ISLDPGFRN	×	×	✓	✓	✓	×
	ISLDPGFRS(M)						
	ISLDSGFRN(R)						
	FCRFTGVVS	×	×	✓	✓	×	✓
	FTGVVSSMH	×	×	✓	✓	×	✓
	YDLLPPGEL	×	×	✓	×	✓	✓
	YDLLPPGKL(M)						
	YDLLPAGKL(R)						
	LSGRVIPRM	×	×	✓	✓	✓	×
	MLYNIDKLS	×	×	✓	✓	✓	×
	MLYNIDRLS(M)						
	LYNIDKLSV	×	×	✓	✓	✓	×
	LYNIDRLSA(M)						
	IDKLSVLLE	×	×	✓	✓	✓	×
	IDRLSALLE(M)						
	LWLDSVIQY	×	×	✓	✓	✓	×
	RTSPNI	×	✓	✓	×	✓	✓
	IQYYGQVQL	×	✓	✓	✓	×	✓
	VQLKKPYSS			✓		✓	✓
	AMSRQRQAI	×	✓	×	✓	×	✓
	KNYPAS	×	×	✓	✓	✓	✓
	YVRQGKQHL	×	×	✓	✓	×	✓
	YVRQGRQHL(R)						
	LRGKITKYY	×	×	✓	✓	×	✓
	YNDILKLNL	✓	×	✓	✓	×	×
	FIRNTKIAE	×	×	✓	✓	×	✓
	FVRNTKIAE(R)						

Table 5.6.2. Antigenicity and allergenicity of VP35, VP40, VP30, VP24 and L protein consensus epitopes predicted by three or more methods (Rank) in Table 5.6.1.

Proteins	Epitopes	Sequence	Interval	Rank	Antigenicity	Allergenicity
VP35	E69	YMQQVSEGL	6-14	3	0.5049	A
	E70	LEKLYKRRK	30-38	3	-0.1292	N
	E71	LYKRRKPKG	33-41	4	0.5383	N
	E72	YKRRKPKGT	34-42	4	0.5099	A
	E73	KRRKPKGTV	35-43	3	0.3384	A
	E74	KMGRTLE	112-118	3	0.4695	N
		KMGKTLE(R)			0.4892	N
	E75	MSKVLELSE	192-200	3	0.6089	N
	E76	LELSEETFS	196-204	3	0.6478	A
	E77	LSEETFSKP	198-206	3	-0.5029	N
	E78	FSKPNLSAK	203-211	3	0.911	N
	E79	FQTVPRPCQ	289-297	4	0.1446	A
	E80	VPRPCQKSL	292-300	3	-0.0486	A
VP40	E81	NYNTYMQYL	6-14	3	0.1981	A
	E82	YNTYMQYLN	7-15	3	0.382	A
	E83	YMQYLNPPP	10-18	3	0.9681	N
	E84	MQYLNPPPY	11-19	4	1.1548	A
	E85	QYLNPPPYA	12-20	3	1.0931	N
	E86	YVGDLNLDD	43-51	3	1.112	N
	E87	SNFEYPL	90-96	3	0.269	N
	E88	FIQNMVIPR	145-153	3	-0.5711	N
		FVQNMVIPR(R)			-0.5135	N
	E89	FSTNQFTYN	155-163	3	-0.2792	N
	E90	HPNLPPIVL	198-206	3	1.3067	A
	E91	YRQHKNPNN	214-222	3	0.775	N
	E92	LRVEKVPEK	235-243	3	0.7323	N

Proteins	Epitopes	Sequence	Interval	Rank	Antigenicity	Allergenicity
VP30	E93	GKLDETS	167-173	3	1.1239	N
	E94	IHLDKGGQF	221-229	3	0.5053	A
	E95	DKGGQFE	224-230	3	0.7489	N
VP24	E96	LSTRYNLPV	4-12	3	0.7188	A
		LSTRYNLPA(M)			0.8147	A
		LSTRYNLPT(R)			0.2484	A
	E97	NVTEKSINL	13-21	4	1.4052	N
		NVTENSINL(M)			0.7172	N
		NITEKSINL(R)			1.6259	N
	E98	VTEKSINLD	14-22	3	1.3665	A
		VTENSINLD(M)			0.6296	N
		ITEKSINLD(R)			1.5109	A
	E99	INLDLNSTA	19-27	3	1.0645	N
	E100	LLHHLKSNF	56-64	3	0.2882	A
	E101	LKSNFVVPE	60-68	3	0.3223	A
	E102	DQELQQS	104-110	4	0.5253	N
	E103	HISPNLLG	135-142	3	1.153	N
	E104	FLVEVRRID	197-205	3	1.4757	A
	E105	EPCCGETVL	207-215	3	0.0058	N
	E106	LSESVVFGL	215-223	3	0.7713	N
	E107	MEKGQPLNL	235-243	3	1.1714	N
		IERGQPLNL(R)			1.1941	A
	E108	LNLTQYMNS	241-249	3	0.5211	N
L	E109	HYSHNPKLR	33-41	3	0.6254	N
	E110	YSHNPKLRN	34-42	3	0.5835	A
	E111	SHNPKLRNC	35-42	3	-0.2883	A
	E112	NPKLRNCRI	37-45	3	-0.2517	N
	E113	RNCRIPHHI	41-49	4	0.611	N
		RNCRIPYHI(R)			0.6674	N
	E114	RIPHHIYRL	44-52	3	-0.5708	A
		RIPYHIYRL(R)			-0.2678	N

Proteins	Epitopes	Sequence	Interval	Rank	Antigenicity	Allergenicity
	E115	IPHHIYRLR	45-53	3	-0.1482	N
		IPYHIYRLR(R)			0.2546	N
	E116	PHHIYRLRN	46-54	3	0.3827	A
		PYHIYRLRN(R)			0.6255	N
	E117	HIYRLRNST	48-56	3	-0.1264	A
	E118	IYRLRNSTA	49-57	3	0.0614	A
	E119	YRLRNSTAL	50-58	3	0.6006	N
	E120	RLRNSTALK	51-59	3	0.3758	N
	E121	LRNSTALKT	52-60	3	0.1505	N
	E122	NHVDDFKY	89-96	3	0.2997	N
		NHINDFKY(R)			0.4078	A
	E123	HVDDFKYLL	90-98	3	-0.1657	N
		HINDFKYLL(R)			0.1446	N
	E124	FKYLLPSEL	94-102	3	0.1635	N
	E125	LARRIKGQR	158-166	3	-0.9381	N
	E126	IKGQRGSLR	162-170	3	0.5061	N
	E127	IIRNAVSLQ	198-206	3	-0.1146	A
		IIRNATSLQ(R)			-0.1076	N
	E128	IRNAVSLQA	199-207	3	0.3678	A
		IRNATSLQA(R)			0.3531	A
	E129	FKLIKHLEP	295-303	3	0.2525	N
	E130	LIKHLEPLC	297-305	3	1.1063	A
	E131	IKHLEPLCV	298-306	3	2.0584	A
	E132	FSLQKHWGH	372-380	3	0.2425	A
	E133	YHSQGSWYK	423-431	3	0.0589	A
	E134	WYKTTHDLH	429-437	3	0.4746	A
	E135	IISDLSIFI	477-485	3	-0.1933	N
	E136	WDSVFDRSV	497-505	3	-0.2104	N
	E137	LGYNPPVRF	506-514	3	0.389	A
	E138	SLKEKELNI	554-562	3	2.1217	A
	E139	YRVRNVQTL	571-579	3	0.8591	N
	E140	VTEREQKEA	599-607	3	1.3117	N
	E141	LLHQASWHH	608-616	3	0.3526	A

Proteins	Epitopes	Sequence	Interval	Rank	Antigenicity	Allergenicity
	E142	YNLAFRYEF	639-647	3	1.4719	N
	E143	LKTKLKLKS	732-740	3	1.7743	N
	E144	YQENEAELN	763-771	3	0.628	N
	E145	WIASFHSML	862-870	3	-0.5153	A
		WIAAFHSML(R)			-0.4685	A
	E146	FHSMLAINL	866-874	3	0.758	N
		FHSMLAVNL(R)			0.5972	N
	E147	LIALITPQV	903-911	3	0.3176	N
	E148	LFQLKNALE	936-944	3	0.5234	A
		LFQLRNALE(R)			0.5332	A
	E149	LEFLEKEEL	943-951	3	1.2825	A
		LEFLRKEEL(R)			1.1677	A
	E150	ILIAKKPGL	954-962	3	0.906	A
		ILISKKPGL(M)			1.1551	A
	E151	LRQTVRENI	987-995	3	-0.0895	N
	E152	LEDQRVCEW	1014-	3	1.002	A
	E153	FSRTPSGKR	1038-	3	-0.6935	N
	E154	LTTEGTMLM	1067-	3	0.4855	N
	E155	MLMKLRELT	1073-	3	0.9029	A
		MLMRLRELT(R)			0.8888	N
	E156	LMKLRELTR	1074-	3	0.3875	A
		LMRLRELTK(R)			0.3298	N
	E157	MKLRELTRN	1075-	3	0.5569	A
		MRLRELTKS(R)			0.8635	N
	E158	LDDDLSESL	1095-	3	-0.1793	A
	E159	LSESLEKFT	1099-	4	-0.581	N
		LSESLEKFI(R)			-1.065	A
	E160	WSDVLKG	1121-	3	0.0364	N
	E161	LSEDLREQF	1149-	3	0.2338	N
		LSEDLKEQF(R)			0.1637	A
	E162	LREQFNLSS	1153-	3	0.9138	N
		LKEQFKLSS(R)			0.9195	N
	E163	FLPYDCKEL	1169-	3	0.9262	A

Proteins	Epitopes	Sequence	Interval	Rank	Antigenicity	Allergenicity
		LLPYDCKEL(M)			0.833	A
		LLQYDCNGL(R)			0.1975	A
	E164	YDCKELRLG	1172-	3	2.1862	A
		YDCKELRLE(M)			1.3367	N
		YDCNGLHSK(R)			0.821	A
	E165	VVQKHPSVN	1198-	3	0.3595	N
		IVQKHPSDN(R)			0.061	A
	E166	VQKHPSVNR	1199-	4	-0.4402	A
		VQKHPSDNR(R)			-0.1778	A
	E167	KIGYPPLRV	1226-	3	1.8042	A
	E168	IGYPPLRVN	1227-	3	1.4882	A
	E169	FQNTINLGV	1341-	3	0.7179	A
	E170	LFFDKPLDV	1386-	3	-0.1492	A
		LYFDKPLDV(M)			-0.1914	A
	E171	LDVDLNKYM	1392-	3	-0.4207	N
	E172	LNKYMDNEL	1396-	3	-0.4571	A
	E173	LCSGIKGRL	1411-	4	0.715	N
	E174	IKGRLGRVS	1415-	3	0.6917	N
	E175	SRSTLSLSL	1423-	3	0.8438	N
		SRSTLTLSL(M)			0.7127	N
	E176	IFYAFGANL	1489-	3	-0.0163	N
	E177	IRNLSHRSL	1524-	5	0.1208	N
	E178	NLSHRSLRI	1526-	3	1.5845	A
	E179	LSHRSLRIL	1527-	4	1.194	N
	E180	LRILQSTFR	1532-	3	-0.0861	A
	E181	LQSTFRHEL	1535-	3	0.3704	N
	E182	RHELVLTRL	1540-	3	0.4249	A
	E183	LVLTRLAHH	1543-	3	0.5564	N
	E184	VLTRLAHHI	1544-	3	0.0366	N
	E185	GGSAGEKS	1560-	3	1.5382	A
	E186	LIKKGQSS	1590-	3	0.6216	A
		LMKKGQSS(M)			0.6309	N
		LMRKNQSP(R)			1.1302	A

Proteins	Epitopes	Sequence	Interval	Rank	Antigenicity	Allergenicity
	E187	DKVQKRKIL	1628-	3	-0.4933	N
		DKIQKRKIL(M)			-0.5909	N
		KKVQNHRPV(R)			-0.087	A
	E188	VQKRKILAD	1630-	3	0.1502	N
		IQKRKILAD(M)			0.1857	N
		VQNHRPVAD(R)			0.2615	N
	E189	ILADTCYPV	1635-	3	-0.1008	A
		ILADTCCPI(M)			-0.1206	A
		PVADTCFLT(R)			0.0751	A
	E190	YYASLNYWR	1659-	3	-0.0632	N
	E191	YASLNYWRD	1660-	3	-0.0853	A
	E192	LMNYGSTTL	1723-	3	0.6583	A
		LINYGSTAL(M)			0.9146	N
		STNHEPTAL(R)			0.97	N
	E193	LSEQNLVEN	1740-	3	0.1488	N
		YSEQDPAKS(R)			0.3537	N
	E194	LVENCRPSK	1745-	3	-0.0456	A
		LVENCRPSE(M)			0.3709	A
		PAKSYLLLE(R)			0.1367	A
	E195	IRCKDNQKI	1755-	3	1.6386	A
		IRFKDNQKI(M)			1.447	N
		TRFRDDQKI(R)			0.3472	A
	E196	IIKHDQRYG	1763-	3	0.8441	A
		ITKHDQRCE(M)			1.5254	N
		ILRHDQKAE(R)			0.5727	N
	E197	MLPKDNMQT	1780-	3	-0.1419	N
		MFPEDNMQT(M			-0.1272	A
		VSSRGCLQA(R)			0.5021	N
	E198	IIKSLDVHE	1801-	3	0.1808	A
		LIKSLDAHE(M)			-0.0162	A
		TTEPLSMLR(R)			0.5715	A
	E199	FLTTLTGTE	1836-	3	0.4921	A
		LLTTPTRTE(M)			0.6984	N

Proteins	Epitopes	Sequence	Interval	Rank	Antigenicity	Allergenicity
		SLKTPMRIE(R)			0.6735	A
	E200	LQPSRYSST	1851-	3	0.9356	A
		LQSSRYSST(M)			0.6506	A
		LQPSECLST(R)			0.0135	A
	E201	LSREQASYL	1866-	3	0.3382	A
	E202	ISLDPGFRN	1884-	3	1.8819	A
		ISLDPGFRS(M)			2.0084	N
		ISLDSGFRN(R)			0.7252	A
	E203	FCRFTGVVS	1919-	3	0.4115	A
	E204	FTGVVSSMH	1922-	3	0.5127	A
	E205	YDLLPPGEL	1934-	3	0.76	N
		YDLLPPGKL(M)			0.9062	N
		YDLLPAGKL(R)			1.0864	A
	E206	LSGRVIPRM	1984-	3	0.7647	A
	E207	MLYNIDKLS	1992-	3	0.0446	N
		MLYNIDRLS(M)			0.0892	N
	E208	LYNIDKLSV	1993-	3	0.2364	A
		LYNIDRLSA(M)			0.2175	N
	E209	IDKLSVLLE	1996-	3	0.1114	A
		IDRLSALLE(M)			0.0732	N
	E210	LWLDSVIQY	2024-	3	-0.2598	N
	E211	RTSPNI	2065-	4	1.4815	N
	E212	IQYYGQVQL	2094-	4	0.4979	A
	E213	VQLKKPYSS	2100-	3	-0.0975	N
	E214	AMSRQRQAI	2143-	3	0.8871	N
	E215	KNYPAS	2160-	4	0.7669	A
	E216	YVRQGKQHL	2203-	3	0.7262	A
		YVRQGRQHL(R)			0.786	N
	E217	LRGKITKYY	2235-	3	-0.336	A
	E218	YNDILKLNL	2243-	3	-0.1403	N
	E219	FIRNTKIAE	2297-	3	0.8862	N
		FVRNTKIAE(R)			0.9244	N

5.7 Novelty analysis

The novelty of the MARV epitopes identified in this study was analyzed by using IEDB. The IEDB database contains the epitopes that are annotated based on the scientific literature. Analyses of the eleven GP and thirteen NP MHC-I T-cell epitopes with IEDB showed that the GP epitopes E5, E6, E10 and E11 were listed as a partial sequence, while the E1 was listed with the same sequence in the database. I found that the epitopes E2, E4, E8 and E9 investigated in this study were never reported previously in IEDB. Similarly, the exact sequences of NP E19, E20, E21 and E23 epitopes are in the IEDB database, while the epitopes E12, E13, E14, E15, E17, E18, E22 and E24 are not in the IEDB.

The GP MHC-II T-cell epitopes, E27 and E28 match exactly with epitopes in the IEDB, while the epitopes E29-E34 partially overlap. The epitopes E25 and E26 are not in the IEDB database. The NP MHC-II epitope E38 has partial overlap with an IEDB sequence, while E35, E36, E37, E39 and E40 have no overlaps with the sequences in the IEDB database. In addition to these MHC-I and MHC-II T-cell epitopes, the search of the predicted B-cell epitopes with the IEDB database showed that the GP E53 and the NP E68 epitopes have partial overlap, while GP epitopes E41-E52, E54 and NP epitopes E55-E67, do not have any overlaps and therefore are novel.

In addition to GP and NP epitopes, the VP35 epitopes E69 and E80, VP40 epitope E84, and the L protein epitopes E112, E133, E135, E139, E211 and E215 in Table 5.6.2 were previously determined experimentally and are available in the IEDB

database. The other epitopes that I found for VP35, VP40, L, and for VP24 and VP30 were not listed in the IEDB.

In addition to the cross-referencing of epitopes with IEDB, I also compared the predicted epitopes with previously predicted epitopes reported in the literature. The predicted GP epitopes E3, E7, E8, E9, E25 (Musoke), E26 (Musoke), E27, E48, E49, E50, E51, E53 were previously reported[144-148]. Similarly, the predicted NP epitope[145,147] E16, VP40 epitopes[146] E86, E87, E90, E91, and L protein epitope[149,150] E119, E120, E139 were reported in the literature.

Taken together, the GP epitopes E2, E4, E41, E42, E43, E44, E45, E46, E47, E52, and E54, and the NP epitopes E12, E13, E14, E15, E17, E18, E22, E24, E35, E36, E37, E39, E40, E55-E67 identified in this study have not been previously reported either in the IEDB database or in the literature. Therefore, these 11 GP epitopes and 26 NP epitopes, as well as 135 other epitopes in the VP35, VP40, VP30, VP24 and L proteins in Table T9 are novel and they remain to be verified experimentally.

6. MUTATION-INDUCED CHANGES IN THE RECEPTOR-BINDING INTERFACE OF THE SARS-COV-2 DELTA VARIANT B.1.617.2 AND IMPLICATIONS FOR IMMUNE EVASION

This chapter was published in 2021: P. Baral, N. Bhattarai, Md Hossen, V. Stebliankin, B.S. Gerstman, G. Narasimhan and P.P. Chapagain, Mutation-induced Changes in the Receptor-binding Interface of the SARS-CoV-2 Delta Variant B.1.617.2 and Implications for Immune Evasion, Biochemical and Biophysical Research Communications, Volume 574, 2021, Pages 14-19, ISSN 0006-291X

I investigate the effects of mutations in the Delta variant on the structure of the receptor-binding interface of RBD as well as the RBD-ACE2 interactions and RBD-neutralizing Abs interactions. I examine the SARS-CoV-2 Ab-RBD complexes available in the protein data bank (PDB) and compare the differences in the RBD-Ab interactions due to the mutations in the Delta variant. Our results suggest that the Delta variant features a stable but slightly reorganized receptor-binding interface that can lead to weakened interactions with some neutralizing Abs resulting in immune evasion.

6.1 Structural changes due to mutation in Delta variant

To investigate the RBD dynamics and the structural changes due to mutations, I performed MD simulations of the RBD of the Delta variant B.1.617.2 and compared the results with the WT, B.1.1.7, and B.1.351 variants. Both of the mutations, L452R and T478K, in the Delta RBD are in the receptor-binding interface comprised of a motif spanning residues 438 to 508. The same interface is a target

for many neutralizing antibodies. Therefore, any changes in the receptor-binding interface can affect both the receptor binding to the host ACE2 as well as Ab-binding. To assess the structural changes in this interface, I analyzed different regions of the interface as shown in Figure 6.1.

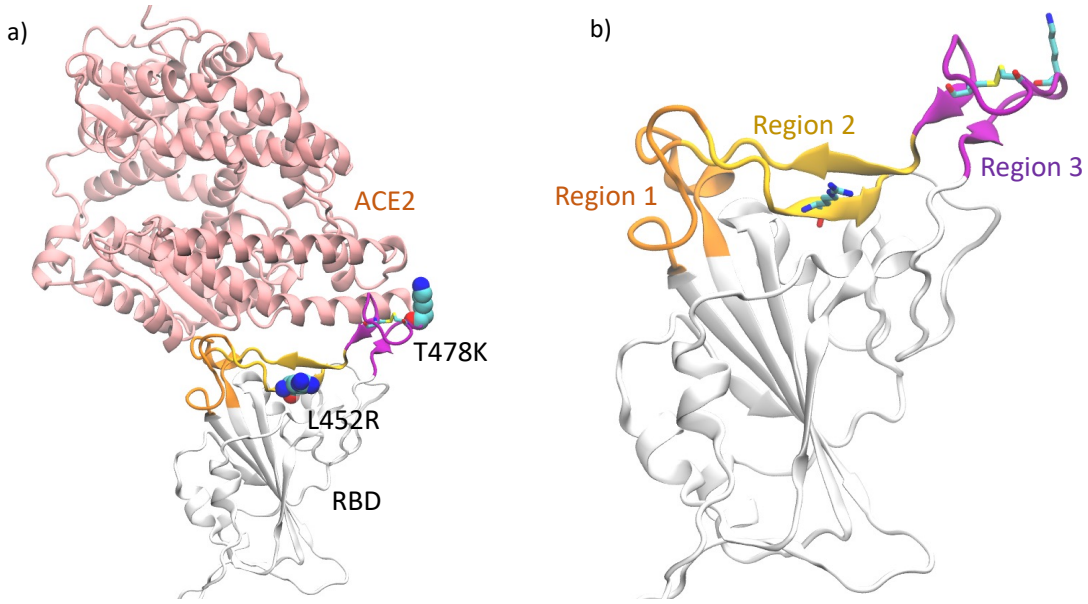


Figure 6.1 a) RBD complexed with ACE2. The locations of the mutations in the RBD of Delta variant are highlighted in VDW representation b) The loop segments consisting of residues 438-447 and 499-508 (Region 1) are highlighted in orange, the β-sheet region consisting of residues 448-455 and 491-498 (Region 2) are highlighted in yellow and the receptor-binding loop consisting of residues 472−490 (Region 3) in purple. The disulfide bond in the loop as well as the mutations in the Delta variant are shown as sticks.

I calculated the root mean square fluctuations (RMSF) of the RBD and plotted the results in Figure 6.2, which shows that the amino acid residues in the β-loop-β motif (Region 3, residues 472-490) have the largest flexibility for all variants. The

loop segments in Region 1 are not found to change significantly compared to WT, and therefore I focused on the β-sheet in Region 2 and the β-loop-β motif of Region 3.

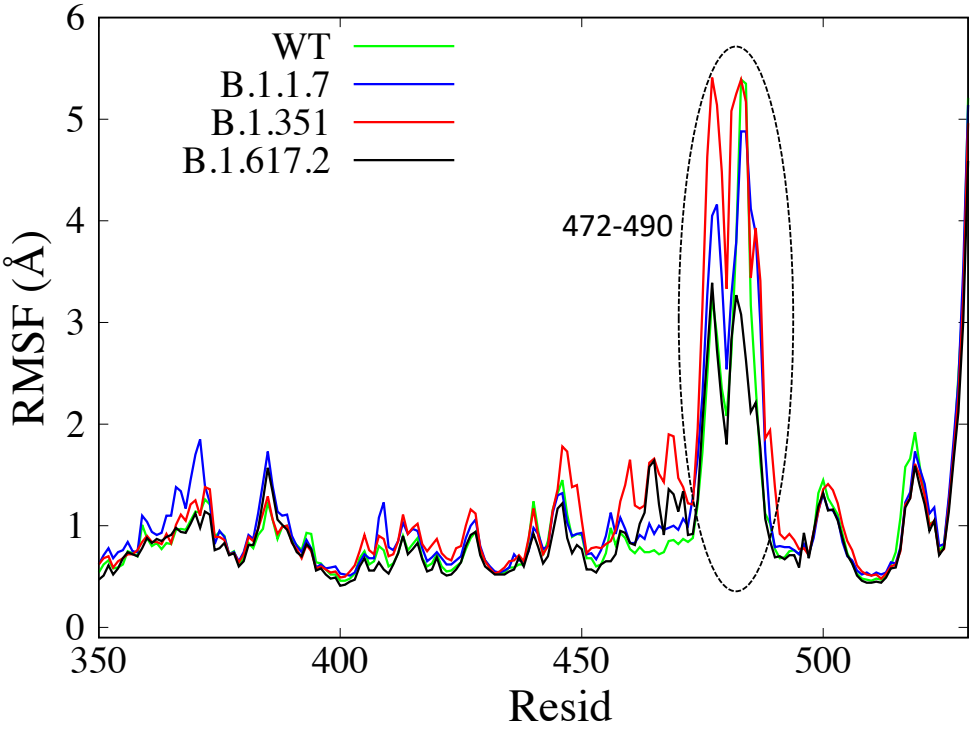
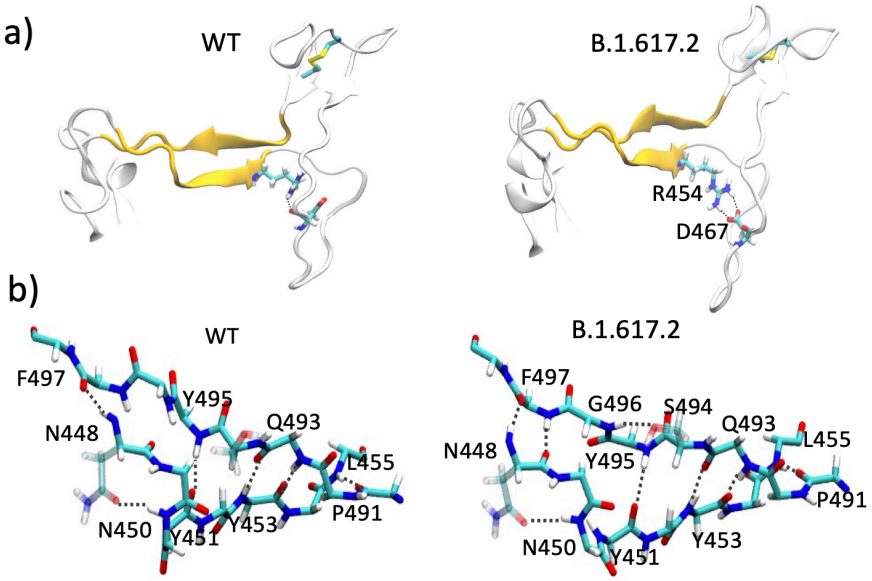


Figure 6.2 Root Mean Square Fluctuations (RMSF) of amino acids calculated from the last 300 ns of 600 ns MD simulations. The flexible β-loop-β region at the RBM interface is indicated with a dotted ellipse.

6.2 Structural rearrangements in the interfacial beta sheet region

Figure 6.3a shows the β-sheet region of the receptor-binding motif (RBM) interface (Region 2 comprised of residues 448-455, 491-498) containing a hydrogen-bond network (Fig. 2b) that creates a stable interface. In the WT, residues in each segment have β-secondary structure (β5: 452-455 and β6: 491-495)[151] with backbone hydrogen-bonds, whereas the additional residues in each segment

(448-451 and 496-498) are mostly unstructured. Figure 6.3b shows the hydrogen-bonding in the RBM in the WT and the Delta B.1.617.2 variant at the end of 600 ns MD simulations. The hydrogen-bond analysis of Region 2 of the interface (Figure 6.3c) shows that WT and B.1.1.7 have similar H-bond interaction patterns, whereas B.1.351 and B.1.617.2 have noticeably different H-bond patterns.



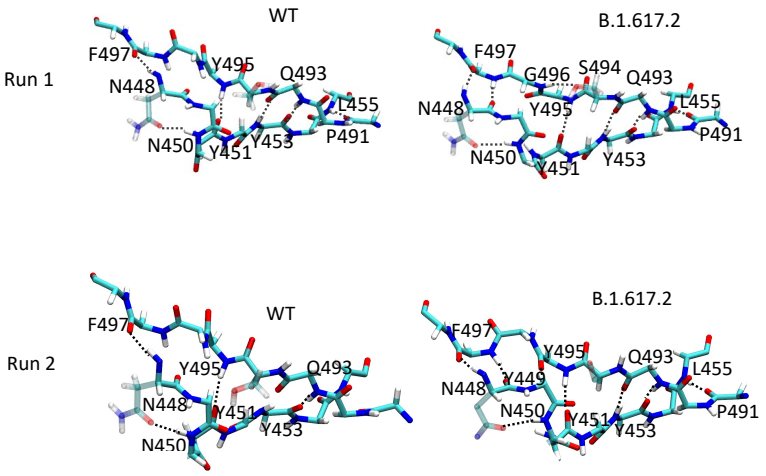
c)

Donor	Acceptor	WT	B.1.1.7	B.1.351	B.1.617.2
N448-N	F497-O	46%	29%	-	85%
F497-N	N448-O	-	-	-	61%
N450-N	N448-OD (SC)	75%	77%	50%	72%
S494-OG (SC)	N448-O	-	-	43%	-
Y495-N	Y451-O	81%	86%	90%	84%
Y453-N	Q493-O	80%	75%	68%	82%
Q493-N	Y453-O	88%	88%	79%	87%
L455-N	P491-O	63%	62%	81%	80%
G496-N	S494-OG (SC)	12%	9%	52%	39%
R454-NH (SC)	D467-OD (SC)	-	-	87%	164%
R454-NH (SC)	D467-O	61%	61%	-	-

Figure 6.3 a) RBM showing the antiparallel β-strands. Residues R454 and D467 participating in ionic interactions in the Delta variant are shown as sticks. b) Hydrogen-bond network in the β-sheet region of the RBM for the WT and the Delta

variant. c) % hydrogen bond occupancy obtained from the last 300 ns for the interactions in WT, B.1.1.7, B.1.351, and B.1.617.2. The sidechain interactions are denoted as SC.

Interestingly, the analysis of Figure 6.3 shows that a slight reorientation of residue G496 in the Delta variant results in much stronger hydrogen bonding between the β-strands. Most notably for the Delta variant, 1) a more stable network of backbone hydrogen bonds is established, with a new hydrogen bond N448(N)-F497(O) formed, and 2) a significantly enhanced salt-bridge interaction between the R454 side chain and D467 side chain is observed. These two changes appear to be due to the mutation L452R which gives a slightly enhanced β-structural propensity[152,153] in β5. It has recently been shown that the L452R mutation in another variant of concern, B.1.427/B.1.429, caused reduction in nearly half of the tested monoclonal Abs.[154], highlighting the dangerous consequences of this mutation. I analyzed 300 ns re-runs for each of these variants and compared in Figure 6.4. Although the R454-D467 backbone hydrogen bond has not switched to a side chain interaction by 300 ns in the re-run of the Delta variant, the presence of the N448(N)-F497(O) hydrogen-bonding between the β-strands is consistent (Figure 6.4).



Donor	Acceptor	WT	Alpha	Beta	Delta
N448-N	F497-O	46%	29%	-	85%
F497-N	N448-O	-	-	-	61%
N450-N	N448 (SC)	75%	77%	50%	72%
S494 (SC)	N448-O	-	-	43%	-
Y495-N	Y451-O	81%	86%	90%	84%
Y453-N	Q493-O	80%	75%	68%	82%
Q493-N	Y453-O	88%	88%	79%	87%
L455-N	P491-O	63%	62%	81%	80%
G496-N	S494 (SC)	12%	9%	52%	39%
R454 (SC)	D467 (SC)	-	-	87%	164%
R454 (SC)	D467-O	61%	61%	-	-

Donor	Acceptor	WT	Alpha	Beta	Delta
N448-N	F497-O	46%	-	50%	76%
F497-N	N448-O	-	-	-	43%
Y495-N	Y451-O	31%	66%	78%	89%
Y453-N	Q493-O	79%	87%	78%	81%
Q493-N	Y453-O	81%	84%	87%	87%
L455-N	P491-O	84%	52%	71%	66%
S494 (SC)	Y451-O	43%	-	-	-
S494 (SC)	N450-O	-	14%	-	-

Figure 6.4 Hydrogen bonding network in beta sheet at the interface for 600 ns MD simulations (run 1) and rerun for 300 ns (run2). The backbone hydrogen bonding between the β-strands for the Delta variant is consistent in both run 1 and run 2.

As seen in Figure 6.3b-c, the interactions in the antiparallel β-strands are enhanced in Delta B.1.617.2, thereby stabilizing the receptor-binding interface. This may directly affect how the RBD binds with ACE2 and with neutralizing antibodies. I note that the recently solved crystal structure of the L452R variant B.1.617.1 (PDB ID 7orb) does not show these hydrogen bonds, suggesting that the changes observed here are perhaps a result of a dynamic reorganization. Similarly, almost all RBD structures show that R454 side chain makes hydrogen bonds with backbones of D467 and/or S469 but some structures do show possible side chain interactions with D467 (e.g. PDB ID 7n1q[155], 7kdj[156]), suggesting an agile network of hydrogen-bonding in this region.

6.3 Structural rearrangements in the β-loop-β motif

The flexible β-loop-β motif (Figure 6.1, Region 3, residues 472-490) contains a disulfide bond between resides C480-C488, and the mutation T478K in the Delta variant also lies in this loop. I explored the mutation-induced changes in the flexibility of this region, and rearrangements in the hydrogen bonding for the different variants (WT, B.1.1.7, B.1.351, B.1.617.2). I find that the Delta variant features a significantly different loop structure. While all variants have a flexible loop in this region[157], the Delta variant shows a reduced flexibility (Figure 6.2) as it adopts a more stable yet different conformation compared to other variants. The difference in the conformational change in the loop can be seen from the changes in the disulfide bond dihedral angle C-C_α-C_β-SG for C480 as displayed in Figure 6.5a. Compared to the WT, the dihedral angle as a function of time for the Delta variant shows a quick flip early and then remains stable in a new orientation.

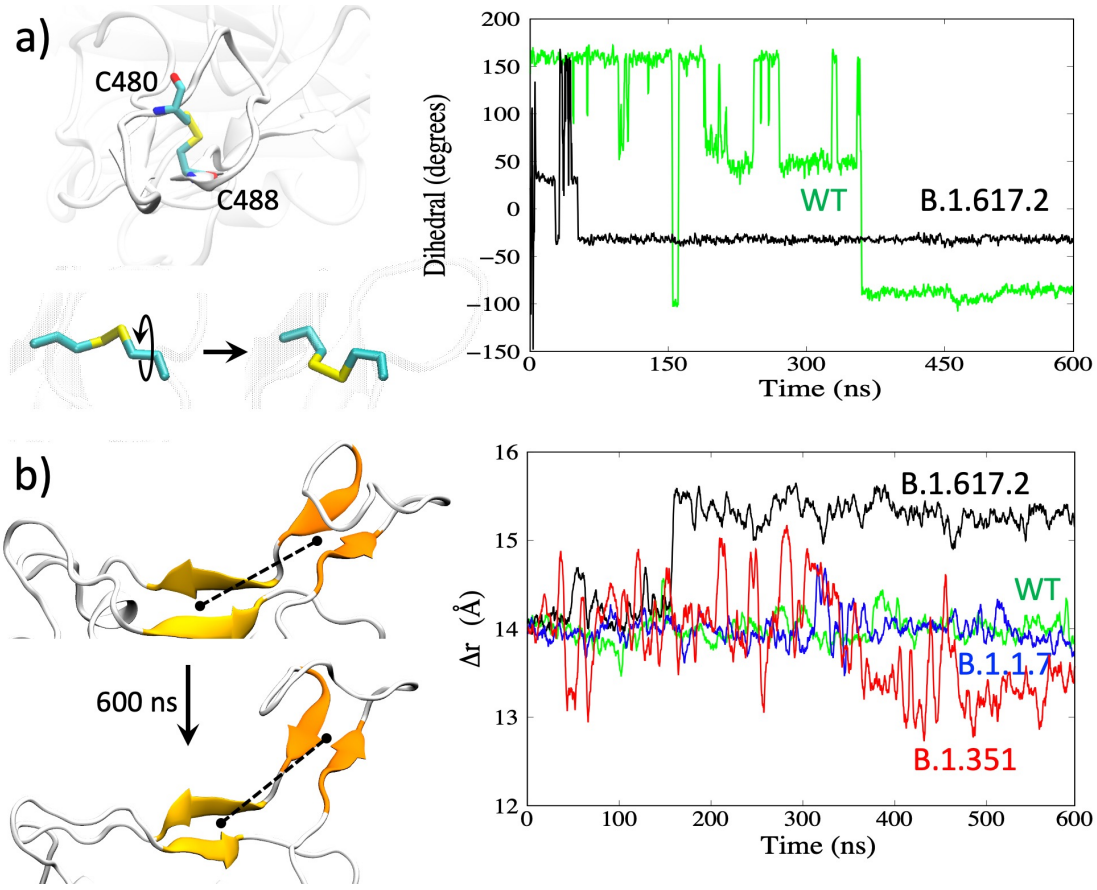


Figure 6.5 a) Reorientation of the disulfide bond. Right: changes in the dihedral angles for WT and the Delta variant. b) Distance between the center-of-mass between the two β-sheets (shown as the dotted line on the left).

The combination of the changes in the β-sheet region (Figure 6.1b, Region 2) and the β-loop-β motif (Figure 6.1b, Region 3) appears to result in an overall change in the receptor-binding interface. In the Delta variant, Regions 2 and 3 are farther apart. This is shown by the separation distance (dotted lines in Figure 6.5b) between the center-of-mass (COM) of the two β-strands in Region 2 (Figure 6.1b: β-strands 452-455 and 492-495) versus the COM of the two β-strands in Region 3 (Figure 6.1b: 472-475 and 487-490). The distance plot in Figure 6.5b (right) shows a stable but slightly extended Region 2 – Region 3 receptor-binding interface for

the Delta variant (black curve) with a 1.5 Å increase in the COM separation distance. Overall, these structural changes and differences in loop flexibility can impact the ACE2 and Ab-binding. The reduced fluctuations in the Delta RBD with an altered receptor-binding interface could result in weaker interactions with neutralizing antibodies leading to immune evasion.

6.4 Antibody binding to the Delta RBM and possible mechanism of immune evasion

A recent study illustrated the mechanism of immune evasion by a variant of concern B.1.427/B.1.429[154]. Specifically, the same mutation found in the Delta variant, L452R, was responsible for reduced neutralizing activities in many of the monoclonal Abs tested, whereas re-grouping of a disulfide bond in a different RBD site caused the loss of activities for all Abs tested. To assess the impact on the Ab binding due to the changes in the receptor-binding interface caused by the mutations in the Delta RBD, I first examined the interfacial interactions in the Ab-RBD complexes available in the protein data bank (PDB) in the WT. Of the 118 RBD-Ab complexes with Ab bound in the receptor-binding interface retrieved from the Protein Data Bank, 47 non-repeating complexes were considered for further analysis. The Ab-RBD complexes (pdb IDs) are listed in Table 6.4.1. I identified the RBD residues involved in ionic or hydrogen bond interactions in each complex and plotted in Figure 6.6 the frequency of occurrences of the important residues in all complexes. While this distribution may be inherently biased due to the available pdb structures of the complexes, it provides a general idea of the preferred

interfacial RBD binding epitope sites for a sample of Abs. From Figure 6.6, I see that most of the Abs have interactions with the β-loop-β residues Y473, A475, N487, E484, among others.

Table 6.4.1 List of the Ab-RBD complexes obtained from the Protein Data Bank. The representative structures from each of the non-repeating groups considered for structural analysis are highlighted in boldface. The structures from the same family of complexes are underlined.

6xc2, 6xc3, 6xc4, 6xc7, **6xdg**, **6xe1**, **6xey**, **6xkp**, **7b3o**, **7beh**, 7bei, 7bej, 7bek, 7bel, 7bem, 7ben, 7beo, 7bep, **7bwj**, **7bz5**, **7c01**, **7cdi**, 7cdj, **7ch4**, 7ch5, **7chb**, **7chc**, 7che, 7chf, 7chh, **7cho**, 7chp, 7chs, **7cjf**, **7cm4**, 7cwl, 7cwm, 7cwn, **7cwo**, 7cwu, **7czp**, 7czq, 7czr, 7czs, 7czt, 7czu, 7czv, 7czw, 7czx, 7czy, 7czz, 7d00, **7d03**, 7deo, **7deu**, **7dk4**, 7dk5, 7dk6, 7dk7, **7e23**, **7jmo**, 7jmp, **7k43**, 7k45, 7k4n, **7k8m**, 7k8s, 7k8t, 7k8x, **7k90**, **7k9z**, **7kfv**, 7kfw, 7kfx, 7kfy, **7klg**, 7klh, 7kmg, 7kmh, **7kmi**, 7kmk, **7kml**, **7kn6**, 7kn7, **7ks9**, **7kxj**, 7kxk, **7kzb**, **7l3n**, **7l56**, 7l57, 7l58, 7l5b, **7laa**, **7ljr**, **7lop**, **7m6d**, 7m6f, 7m6h, 7m6i, **7mf1**, **7mjj**, 7mjk, 7mjl, **7nd4**, 7nd5, 7nd6, 7nd7, 7nd8, 7nd9, 7nda, 7ndb, **7neg**, 7neh, **7nx6**, 7nx7, 7nx8, 7nx9.

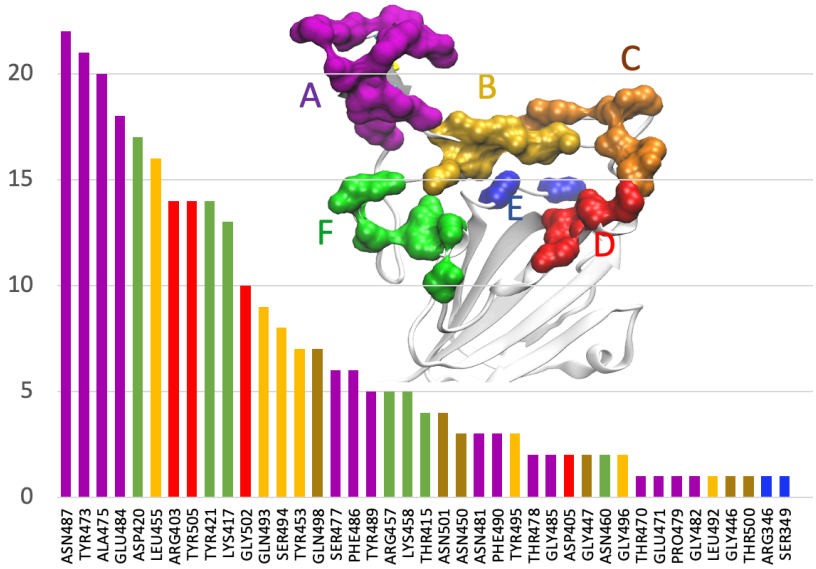


Figure 6.6 Frequency of occurrences of the RBD residues involved in hydrogen-bonding with Ab in 47 complexes from the Protein Data Bank.

Inspection of the Ab-RBD complexes for the WT shows that many of the Abs anchor at multiple sites. For example, in many Ab-RBD complexes, including 6xe1, 7b3o, 7cdi, and 7cjf Abs bind at A475/G485 at one site (site A in Figure 6.6) and R457/K458 (site F in Figure 6.6) at another site, as shown in the figure. I grouped different sites and color-coded as shown in Figure 6.6. To examine how the changes in these sites may affect the Ab binding, I plotted the C_α distance between the residues K458 and A475 belonging to two Ab-binding sites in Figure 6.7 for both the WT and the Delta variant. As shown in Figure 6.7b, the distance between these residues (458-475) mostly remains at ~9 Å for the WT. However, the same (458-475) distance for the Delta variant in Figure 6.7c increases to ~14 Å by 150 ns and remains stable at that distance. This 4-5 Å increase in the Ab-binding sites suggests that the Ab-binding will be severely affected, and the Ab becomes insensitive to Delta RBD binding. While ACE2 binds at the site of 475/487 (site A

in Figure 6.6), it does not bind at the site of 457/458 (site F) and therefore the increase in the K458-A475 distance does not affect the ACE2 binding. Instead, ACE2 binds at 475/487 and Q493 (in the middle of β6). Therefore, I also plotted the distance between the residues N487 and Q493. Interestingly, despite the structural changes, this distance in both the WT and the Delta variant remains nearly the same (16-17 Å) as seen in Figure 6.7b-c. This suggests that the structural changes have not affected the ACE2 binding sites but significantly affected the Ab-binding sites, suggesting a possible immune evasion by the Ab while maintaining the ability of receptor binding.

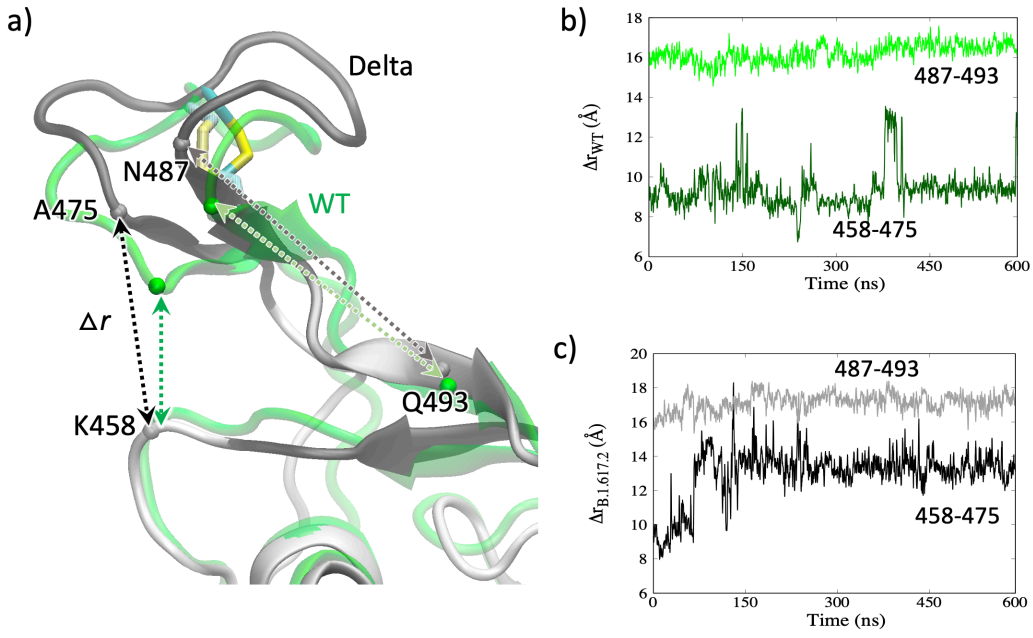


Figure 6.7 a) Amino acid residues involved in Ab-binding or ACE2-binding. b) The Cα-Cα distances between the residue pairs K458-A475 and N487-Q493 in WT. c) The Cα-Cα distances in the Delta variant.

6.5 ACE2 binding vs. antibody binding in the Delta variant

With the observation of the increase in the distance between the two sites in the Delta RBM, I investigated how the changes affect ACE2 and Ab binding. If the ACE2 binding is maintained or enhanced but the Ab binding is weakened, at least for a set of neutralizing Abs, that would mean that the virus is less sensitive to the Abs thereby making it more effective at infecting and spreading. To explore this, I performed simulations of the RBD-ACE2 complex and Ab-RBD complexes for the Delta variant and compared with the complexes of the WT. Since the complexes for the Delta variant were modeled from the RBD obtained from the 600 ns simulation, the interactions are expected to evolve, whereas those in the WT remain steady. The hydrogen bond interactions in the RBD-ACE2 as well as the Ab-RBD complexes are shown in Figure 6.8. In the WT RBD-ACE2 complex, the RBD residues that primarily participate in hydrogen-bond interactions include K417, Y489, G502, E484, T500 and N487. The RBD residue K417 forms a strong salt-bridge with D30 of ACE2 in WT and B.1.1.7 but not in B.1.351 due to the K417 mutation[157]. These WT interactions are still present in the complex with the Delta RBD, though the % occupancy are reduced (Table 6.5.1). With some of the major interactions, including the K417-D30 salt-bridge, still present in the complex, ACE2 binding seems tolerate the structural changes in the Delta RBD.

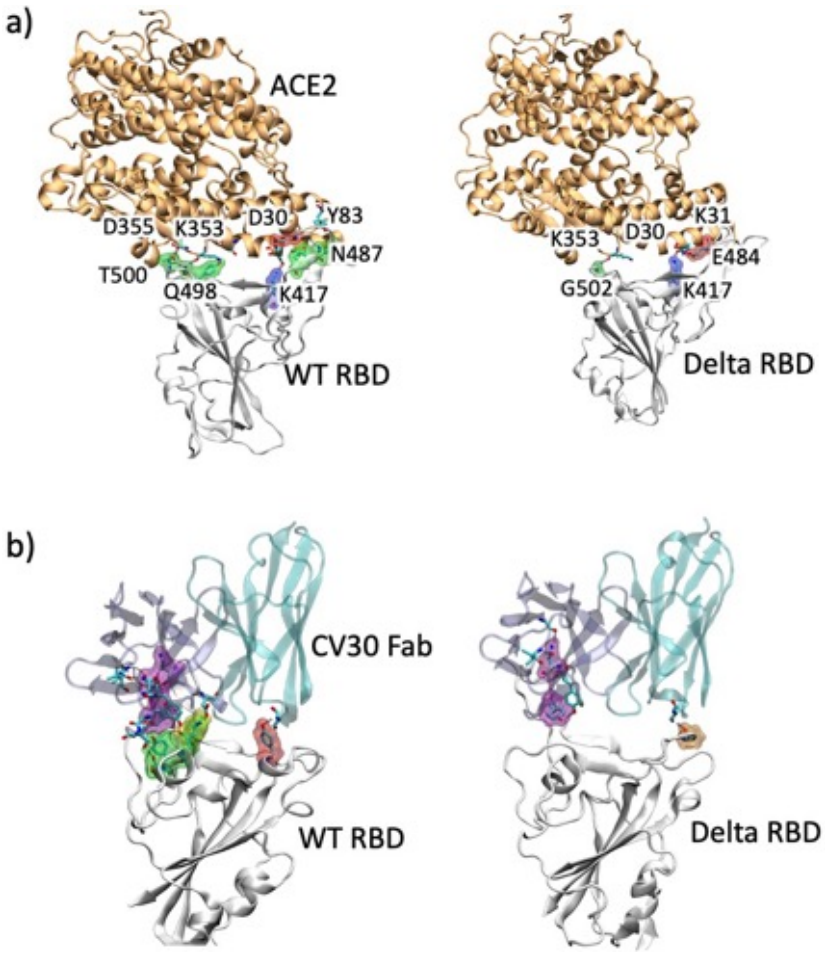


Figure 6.8 a) Structures of the ACE2-RBD complexes for WT and the Delta variant at the end of the 100 ns simulations. b) Ab-RBD complex for WT and the Delta variant, with CV30-Fab neutralizing Ab (PDB 6xe1) complexed with RBD. The interacting sites are highlighted in surface representation for the RBD and sticks for the ACE2 or Ab.

Table 6.5.1 Hydrogen bond details analysis of RBD-ACE2 for WT and B.1.617.2.

	WT		B.1.617.2	
RBD-ACE2	0-10 ns	90-100 ns	0-10 ns	90-100 ns
K417-D30	80%	84%	80%	67%
K417-H34			28.70%	
Y449-D38	31%	26%		
Y453-H34				36%
E484-K31		44%	12%	48%
N487-Y83	24%	44%		
N487-Q24				14%
Y489-T27		77%	14%	
L492-K31			12%	
Q493-K31			51%	40%
Q493-E35	32%		63%	
Q498-K353	31%			
Q498-Q42		25%		10%
T500-D355	62%	49%		
T500-Y41	19%	26%	52%	
T500-R357				11%
T500-N330				10%
G502-K353	86%	52%	13%	61%
Y505-E37	45%	35%		19%

I next compared the Ab binding in the WT and the Delta RBD. I considered two examples of the Ab-Delta RBD complexes modeled from the WT RBD complexed with the neutralizing Ab CV30 Fab (pdb ID 6xe1)[158] and complexed with the Ab BD-236 Fab (pdb ID 7chb)[159]. The 100 ns simulation of the CV30-Delta RBD model shows a less stable complex with significantly reduced interactions. As shown in Figure 6.8b, Ab in the WT has interactions with residues in three clusters that are intact during the simulation. However, in the Delta RBD, the Ab is only able to bind at site A or F but not both. This is consistent with the argument made based on Figure 6.7. The major hydrogen bonds, including those with Y473, Y421, L455,

A475, R457, R403, K417, and Y505 in WT are broken or weakened in the Delta variant. The Ab-RBD complex modeled from 7chb (complex with BD-236 Fab) also shows reduction in the number and strength of the hydrogen bonding. Some of the major hydrogen bonds withY453, G502, D420, and K417 in the WT BD-236 complex are still present in the Delta variant during the 100 ns simulation (Table 6.5.2). Figure 6.8b displays the final conformations of the Ab-RBD complexes for the CV30 Ab, with the RBD residues colored according to the grouping in Figure 6.6. The % hydrogen bonding occupancies are given in Table 6.5.2.

Table 6.5.2 Hydrogen bond details for the RBD complexed with a) CV30 Fab antibody (PDB 6xe1) and b) BD-236 Fab antibody (PDB 7chb) for WT and B.1.617.2.

a)

	WT		B.1.617.2	
RBD-6XE1 AB	0-10 ns	90-100 ns	0-10 ns	90-100 ns
R403-G92	75%	104%		
K417-D97	73%	67%	73%	
D420-S56	90%	92%	88%	
Y421-G54	53%	71%		
Y421-S56			76%	
L455-Y33	51%	80%		
R457-S53	94%	94%		
Y473-S31	94%	82%		
A475-I28	69%	51%		
A475-N32	49%	57%		
N487-R94	82%	96%		
N487-I28				34%
N487-G26				28%
Y489-Y33				31%
P499-S29				32%
Y505-S29		43%	64%	

b)

	WT		B.1.617.2	
RBD-7CHB AB	0-10 ns	90-100 ns	0-10 ns	90-100 ns
R403-N92			50%	
E406-N92				
E406-Y94		68%		
T415-S56		72%		
K417-E101	67%	72%	72%	66%
D420-S56	82%			
D420-Y33				49%
Y421-G54	75%			
Y453- E101	90%	90%	73%	82%
L455-Y33	84%	84%		
R457-S53		106%		
N460-S56	49%			
Y473- S31	88%	84%		
A475-T28	80%	66%		
A475-N32	69%	52%		
N487-R97	86%	82%		
Y489-L99			55%	
G502-G28		66%		
G502-Q27				80%
Y505-S30		52%		

7. CONCLUSIONS

I investigated the molecular details of proteins encoded by three different viruses, Lassa, Marburg, and the B.1.617.2 delta variant of SARS-CoV-2. These three systems were chosen because they have caused current or recent epidemics, and the crystallographic structure of the proteins are known.

LASV hemorrhagic fever is endemic in West Africa, and no approved effective therapeutics are currently available. Therefore, there is an urgent need for the discovery and development of potential antiviral therapeutics. The LASV GP (glycoprotein) spike is a promising selective target for the development of novel vaccines as it plays an essential role in the virus-host interaction. Several *in silico* studies[134,135,160-163] were performed to predict LASV GP epitopes with the use of a single prediction tool for each type of epitope. I have identified new T and B-cell epitopes using a variety of computational approaches, including twelve epitope prediction methods, protein-peptide docking, and MD simulations. The MHC I and II T-cell epitopes were separately predicted with the LASV GP sequence using well-known databases and prediction methods.

The predicted MHC I T-cell epitopes then were prioritized based on the consensus score, binding affinity, and antigenicity, while MHC II T and B-cell epitopes were prioritized based on the consensus score. Novelty analysis of the consensus-selected 33 epitopes showed that thirty of these predicted epitopes have either no overlap or only a partial overlap to previously reported sequences. Within this list of new epitopes, six sequences have no overlap with any known experimentally

tested epitopes in the IEDB. In addition, docking and MD simulations were performed to further validate the MHC I T-cell epitopes. The simulation results show that the allele-MHC-I epitope complexes are stable, with favorable hydrogen-bond and interaction energy. Of these, the epitope ^{233}FSRPSPIGY^{241} segment was found to be especially stable. This study demonstrates that the consensus epitope prediction strategy is valuable for *in silico* investigations of known epitopes and the identification of new epitopes. Experimental validation of these epitopes may lead to the design and development of effective LASV vaccines.

In addition to LASV, endemic MARV has resulted in many deaths since its initial outbreak in 1967. The lack of approved and effective therapeutics and treatment measures has underscored the need for the discovery and development of vaccine and therapeutic measures to mitigate Marburgvirus infections. Identification of potential vaccine candidate epitopes is the first step towards designing effective vaccines against the infection. In this work, I employed a total of nine different epitope prediction methods that predicted 35 GP, 33 NP, 12 VP35, 12 VP40, 3 VP30, 13 VP24 and 111 L protein epitopes. Among these, 11 GP, 26 NP, 10 VP35, 7 VP40, 3 VP30, 13 VP24, and 102 L protein predicted epitopes are novel. Most of these epitopes are conserved among all four strains: Lake Victoria, Angola, Musoke and Ravn-87. The non-allergenic epitopes with high value of antigenicity and binding affinity are considered as potential candidates for further studies. From the virtual screening of MHC-I epitopes, the non-allergenic GP epitope E9 and the NP epitope E15 are obtained to have the strongest binding scores (-10.9 kcal/mol and -8.9 kcal/mol respectively).

Sufficient information was available to allow us to perform MD simulations on highly ranked MHC-I T-cell epitopes to determine molecular details of the epitope-allele interface. The MD simulations of the epitope-allele complexes show that the GP epitope 164NIAAMIVNK172 and the NP epitope 80YLRDAGYEF88 have especially stable interactions with the alleles and a high number of hydrogen bonds. Both of these epitopes are also conserved among MARV strains, suggesting that these are promising candidate epitopes for vaccine development. I also found that hydrophobic amino acids in epitopes can create a tight, water excluding epitope-allele interface. This study demonstrates the value of the use of consensus methods and *in silico* MD screening for identification of novel MARV epitopes. Further exploration with experimental validation of these epitopes may lead to the design and development of effective MARV vaccines.

In addition to the identification of novel epitopes as possibly promising vaccine candidates, studies of antibody binding to viral proteins are important. A spike in the number of cases and deaths due to the coronavirus pandemic occurred due to the B.1.617.2 delta variant of SARS-CoV-2, implying the possibility of viral evasion of the antibodies created in response to earlier viral strains. In this work, I performed molecular dynamics simulations of the Delta variant's spike protein RBD with the mutations L452R/T478K and investigated the resulting structural changes in the receptor- and Ab-binding interfaces. I find that the Delta variant presents a noticeably different receptor-binding interface compared to the WT, B.1.1.7, and B.1.351. Specifically, the receptor-binding β-loop-β motif adopts an altered conformation which appears to cause shifts in the Ab-binding epitope regions that

can reduce the binding affinities for some neutralizing Abs. I investigated this by performing all-atom MD simulations of two Ab-RBD complexes and found that one of the complexes shows significantly reduced interactions between the Ab and the RBD, suggesting a possible mechanism of the immune escape by the Delta variant. Even though the Ab-resistant conformations obtained in these simulations may represent only a subset of the conformational ensemble, they can still contribute considerably to the reduced sensitivity of the Abs. Future work with a full mapping of the conformational space of the receptor-binding interface may shed further light on the nature of the interactions with the common anti-RBD Abs, providing useful information on vaccine efficacies. Understanding how the structural changes alter the RBD's ability to present itself in the up conformation in the spike trimer or its ACE2 binding affinity can also inform us about the variant's transmissibility.

REFERENCES

1 Wolkowicz, R., Schaechter, M., Raoult, D. & Forterre, P. What makes a virus a virus? *Nature Reviews Microbiology* **6** (2008).

2 Rutschman, A. S. The Vaccine Race in the 21st Century. *Ariz. L. Rev.* **61**, 729 (2019).

3 Cheng, V. C.-C., Chan, J. F.-W., Hung, I. F.-N. & Yuen, K.-Y. Viral Infections, an Overview with a Focus on Prevention of Transmission. *International Encyclopedia of Public Health*, 368 (2017).

4 Ahmad, S. I. *Human Viruses: Diseases, Treatments and Vaccines: The New Insights*. (Springer Nature, 2021).

5 Meganck, R. M. & Baric, R. S. Developing therapeutic approaches for twenty-first-century emerging infectious viral diseases. *Nature Medicine* **27**, 401-410 (2021).

6 Plotkin, S. & Reef, S. (Philadelphia: Saunders, 2013).

7 Plotkin, S. & Mortimer Jr, E. New Technologies for Making Vaccines. *Vaccines, WB Saunders Co* **571** (1988).

8 Hwang, J.-R., Byeon, Y., Kim, D. & Park, S.-G. Recent insights of T cell receptor-mediated signaling pathways for T cell activation and development. *Experimental & molecular medicine* **52**, 750-761 (2020).

9 Palatnik-de-Sousa, C. B., Soares, I. d. S. & Rosa, D. S. Epitope discovery and Synthetic Vaccine design. *Frontiers in immunology* **9**, 826 (2018).

10 Arora, S. & Arya, A. Epitope Based Vaccine Designing-A mini review. *J Vaccines Immunol* **6**, 038-041 (2020).

11 Janeway Jr, C. A., Travers, P., Walport, M. & Shlomchik, M. J. in *Immunobiology: The Immune System in Health and Disease. 5th edition* (Garland Science, 2001).

12 Huang, J. & Honda, W. CED: a conformational epitope database. *BMC Immunol* **7**, 7, doi:10.1186/1471-2172-7-7 (2006).

13 Barlow, D. J., Edwards, M. S. & Thornton, J. M. Continuous and discontinuous protein antigenic determinants. *Nature* **322**, 747-748, doi:10.1038/322747a0 (1986).

14 Li, Y., Yin, Y. & Mariuzza, R. A. Structural and biophysical insights into the role of CD4 and CD8 in T cell activation. *Front Immunol* **4**, 206, doi:10.3389/fimmu.2013.00206 (2013).

15 Zamoyska, R. CD4 and CD8: modulators of T-cell receptor recognition of antigen and of immune responses? *Curr Opin Immunol* **10**, 82-87, doi:10.1016/s0952-7915(98)80036-8 (1998).

16 Nielsen, M., Lundegaard, C. & Lund, O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics* **8**, 238, doi:10.1186/1471-2105-8-238 (2007).

17 Yang, X. & Yu, X. An introduction to epitope prediction methods and software. *Rev Med Virol* **19**, 77-96, doi:10.1002/rmv.602 (2009).

18 Zhang, L. Multi-epitope vaccines: a promising strategy against tumors and viral infections. *Cell Mol Immunol* **15**, 182-184, doi:10.1038/cmi.2017.92 (2018).

19 Shey, R. A. *et al.* In-silico design of a multi-epitope vaccine candidate against onchocerciasis and related filarial diseases. *Sci Rep* **9**, 4409, doi:10.1038/s41598-019-40833-x (2019).

20 Baral, P. *et al.* Mutation-induced changes in the receptor-binding interface of the SARS-CoV-2 Delta variant B. 1.617. 2 and implications for immune evasion. *Biochemical and biophysical research communications* **574**, 14-19 (2021).

21 Radoshitzky, S. R. *et al.* ICTV Virus Taxonomy Profile: Arenaviridae. *J Gen Virol* **100**, 1200-1201, doi:10.1099/jgv.0.001280 (2019).

22 Frame, J. D., Baldwin, J. M., Jr., Gocke, D. J. & Troup, J. M. Lassa fever, a new virus disease of man from West Africa. I. Clinical description and pathological findings. *Am J Trop Med Hyg* **19**, 670-676 (1970).

23 Gibb, R., Moses, L. M., Redding, D. W. & Jones, K. E. Understanding the cryptic nature of Lassa fever in West Africa. *Pathog Glob Health* **111**, 276-288, doi:10.1080/20477724.2017.1369643 (2017).

24 Hastie, K. M. *et al.* Structural basis for antibody-mediated neutralization of Lassa virus. *Science* **356**, 923-928, doi:10.1126/science.aam7260 (2017).

25 Buckley, S. M. & Casals, J. Lassa fever, a new virus disease of man from West Africa. 3. Isolation and characterization of the virus. *Am J Trop Med Hyg* **19**, 680-691 (1970).

26 McCormick, J. B. *et al.* A case-control study of the clinical diagnosis and course of Lassa fever. *J Infect Dis* **155**, 445-455 (1987).

27 Buchmeier, M. J. & Oldstone, M. B. Protein structure of lymphocytic choriomeningitis virus: evidence for a cell-associated precursor of the virion glycopeptides. *Virology* **99**, 111-120, doi:10.1016/0042-6822(79)90042-4 (1979).

28 Sogoba, N., Feldmann, H. & Safronetz, D. Lassa fever in West Africa: evidence for an expanded region of endemicity. *Zoonoses Public Health* **59 Suppl 2**, 43-47, doi:10.1111/j.1863-2378.2012.01469.x (2012).

29 Wolff, S. *et al.* Genome Sequence of Lassa Virus Isolated from the First Domestically Acquired Case in Germany. *Genome Announc* **4**, doi:10.1128/genomeA.00938-16 (2016).

30 Jahrling, P. B. & Peters, C. J. Passive antibody therapy of Lassa fever in cynomolgus monkeys: importance of neutralizing antibody and Lassa virus strain. *Infect Immun* **44**, 528-533 (1984).

31 Li, S. *et al.* Acidic pH-Induced Conformations and LAMP1 Binding of the Lassa Virus Glycoprotein Spike. *PLoS Pathog* **12**, e1005418, doi:10.1371/journal.ppat.1005418 (2016).

32 Bederka, L. H., Bonhomme, C. J., Ling, E. L. & Buchmeier, M. J. Arenavirus stable signal peptide is the keystone subunit for glycoprotein complex organization. *MBio* **5**, e02063, doi:10.1128/mBio.02063-14 (2014).

33 Klaus, J. P. *et al.* The intracellular cargo receptor ERGIC-53 is required for the production of infectious arenavirus, coronavirus, and filovirus particles. *Cell Host Microbe* **14**, 522-534, doi:10.1016/j.chom.2013.10.010 (2013).

34 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).

35 Robinson, J. E. *et al.* Most neutralizing human monoclonal antibodies target novel epitopes requiring both Lassa virus glycoprotein subunits. *Nat Commun* **7**, 11544, doi:10.1038/ncomms11544 (2016).

36 King, L. B. *et al.* The Marburgvirus-Neutralizing Human Monoclonal Antibody MR191 Targets a Conserved Site to Block Virus Receptor Binding. *Cell Host Microbe* **23**, 101-109 e104, doi:10.1016/j.chom.2017.12.003 (2018).

37 Dye, J. M. *et al.* Postexposure antibody prophylaxis protects nonhuman primates from filovirus disease. *Proc Natl Acad Sci U S A* **109**, 5034-5039, doi:10.1073/pnas.1200409109 (2012).

38 Kuhn, J. H. *et al.* Proposal for a revised taxonomy of the family Filoviridae: classification, names of taxa and viruses, and virus abbreviations. *Arch Virol* **155**, 2083-2103, doi:10.1007/s00705-010-0814-x (2010).

39 Clarke, E. C. *et al.* Production and Purification of Filovirus Glycoproteins in Insect and Mammalian Cell Lines. *Sci Rep* **7**, 15091, doi:10.1038/s41598-017-15416-3 (2017).

40 Hashiguchi, T. *et al.* Structural basis for Marburg virus neutralization by a cross-reactive human antibody. *Cell* **160**, 904-912, doi:10.1016/j.cell.2015.01.041 (2015).

41 Mittler, E., Kolesnikova, L., Herwig, A., Dolnik, O. & Becker, S. Assembly of the Marburg virus envelope. *Cell Microbiol* **15**, 270-284, doi:10.1111/cmi.12076 (2013).

42 Mittler, E., Kolesnikova, L., Strecker, T., Garten, W. & Becker, S. Role of the transmembrane domain of marburg virus surface protein GP in assembly of the viral envelope. *J Virol* **81**, 3942-3948, doi:10.1128/JVI.02263-06 (2007).

43 Volchkov, V. E. *et al.* Proteolytic processing of Marburg virus glycoprotein. *Virology* **268**, 1-6, doi:10.1006/viro.1999.0110 (2000).

44 Rahim, M. N. *et al.* Generation and Characterization of Anti-Filovirus Nucleoprotein Monoclonal Antibodies. *Viruses* **11**, doi:10.3390/v11030259 (2019).

45 Watanabe, S., Noda, T. & Kawaoka, Y. Functional mapping of the nucleoprotein of Ebola virus. *J Virol* **80**, 3743-3751, doi:10.1128/JVI.80.8.3743-3751.2006 (2006).

46 Niikura, M. *et al.* Analysis of linear B-cell epitopes of the nucleoprotein of ebola virus that distinguish ebola virus subtypes. *Clin Diagn Lab Immunol* **10**, 83-87, doi:10.1128/cdli.10.1.83-87.2003 (2003).

47 Niikura, M. *et al.* Detection of Ebola viral antigen by enzyme-linked immunosorbent assay using a novel monoclonal antibody to nucleoprotein. *J Clin Microbiol* **39**, 3267-3271, doi:10.1128/jcm.39.9.3267-3271.2001 (2001).

48 Liu, B. *et al.* Structural Insight into Nucleoprotein Conformation Change Chaperoned by VP35 Peptide in Marburg Virus. *J Virol* **91**, doi:10.1128/JVI.00825-17 (2017).

49 Ascenzi, P. *et al.* Ebolavirus and Marburgvirus: insight the Filoviridae family. *Mol Aspects Med* **29**, 151-185, doi:10.1016/j.mam.2007.09.005 (2008).

50 Bornholdt, Z. A. *et al.* Structural rearrangement of ebola virus VP40 begets multiple functions in the virus life cycle. *Cell* **154**, 763-774, doi:10.1016/j.cell.2013.07.015 (2013).

51 Gc, J. B., Gerstman, B. S. & Chapagain, P. P. Membrane association and localization dynamics of the Ebola virus matrix protein VP40. *Biochim Biophys Acta Biomembr* **1859**, 2012-2020, doi:10.1016/j.bbamem.2017.07.007 (2017).

52 Fraser, M. E. *Novel Anti-Viral Strategies for Lipid-Enveloped Viruses*. (University of Notre Dame, 2019).

53 Takamatsu, Y., Kolesnikova, L. & Becker, S. Ebola virus proteins NP, VP35, and VP24 are essential and sufficient to mediate nucleocapsid transport. *Proc Natl Acad Sci U S A* **115**, 1075-1080, doi:10.1073/pnas.1712263115 (2018).

54 Banadyga, L. *et al.* Ebola virus VP24 interacts with NP to facilitate nucleocapsid assembly and genome packaging. *Sci Rep* **7**, 7698, doi:10.1038/s41598-017-08167-8 (2017).

55 Bhattarai, N., Gerstman, B. S. & Chapagain, P. P. Role of k-loop cysteine residues in the marburg virus protein VP24–human Keap1 complex. *ACS Omega* **3**, 18639-18645 (2018).

56 Trunschke, M. *et al.* The L–VP35 and L–L interaction domains reside in the amino terminus of the Ebola virus L protein and are potential targets for antivirals. *Virology* **441**, 135-145 (2013).

57 Yasmin, T. & Nabi, A. N. B and T cell epitope-based peptides predicted from evolutionarily conserved and whole protein sequences of Ebola virus as vaccine targets. *Scandinavian journal of immunology* **83**, 321-337 (2016).

58 Sharma, S. & Malla, S. In silico Characterization and Selection of Epitope-based Peptide Vaccines Against Ebola Viruses.

59 Rota, P. A. *et al.* Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* **300**, 1394-1399, doi:10.1126/science.1085952 (2003).

60 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269, doi:10.1038/s41586-020-2008-3 (2020).

61 Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273, doi:10.1038/s41586-020-2012-7 (2020).

62 Cui, J., Li, F. & Shi, Z. L. Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol* **17**, 181-192, doi:10.1038/s41579-018-0118-9 (2019).

63 Krammer, F. SARS-CoV-2 vaccines in development. *Nature* **586**, 516-527 (2020).

64 WHO. COVID-19 vaccine tracker and landscape. (2021).

65 Tenforde, M. W. Effectiveness of Pfizer-BioNTech and Moderna Vaccines Against COVID-19 Among Hospitalized Adults Aged≥ 65 Years—United States, January–March 2021. *MMWR. Morbidity and mortality weekly report* **70** (2021).

66 Bracken, C. J. *et al.* Bi-paratopic and multivalent VH domains block ACE2 binding and neutralize SARS-CoV-2. *Nature Chemical Biology*, 1-9 (2020).

67 Cao, L. *et al.* De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science* **370**, 426-431 (2020).

68 GREIN, J., OHMAGARI, N. & SHIN, D. original: Compassionate Use of Remdesivir for Patients with Severe Covid-19. *View Article* (2020).

69 Joyner, M. J. *et al.* Effect of convalescent plasma on mortality among hospitalized patients with COVID-19: initial three-month experience. *MedRxiv* (2020).

70 Lv, Z. *et al.* Structural basis for neutralization of SARS-CoV-2 and SARS-CoV by a potent therapeutic antibody. *Science* **369**, 1505-1509 (2020).

71 Medina-Enríquez, M. M. *et al.* ACE2: the molecular doorway to SARS-CoV-2. *Cell & bioscience* **10**, 1-17 (2020).

72 Monteil, V. *et al.* Inhibition of SARS-CoV-2 infections in engineered human tissues using clinical-grade soluble human ACE2. *Cell* **181**, 905-913. e907 (2020).

73 Robbiani, D. F. *et al.* Convergent antibody responses to SARS-CoV-2 in convalescent individuals. *Nature* **584**, 437-442 (2020).

74 Walser, M. *et al.* Highly potent anti-SARS-CoV-2 multi-DARPin therapeutic candidates. *bioRxiv* (2020).

75 Wang, B. *et al.* Bivalent binding of a fully human IgG to the SARS-CoV-2 spike proteins reveals mechanisms of potent neutralization. *bioRxiv* (2020).

76 Abu-Saleh, A. A.-A. A., Awad, I. E., Yadav, A. & Poirier, R. A. Discovery of potent inhibitors for SARS-CoV-2's main protease by ligand-based/structure-based virtual screening, MD simulations, and binding energy calculations. *Physical Chemistry Chemical Physics* **22**, 23099-23106 (2020).

77 Wang, Q. *et al.* Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell* **181**, 894-904. e899 (2020).

78 Pinto, D. *et al.* Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. *Nature* **583**, 290-295 (2020).

79 Wu, Y. *et al.* A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. *Science* **368**, 1274-1278 (2020).

80 Hansen, J. *et al.* Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. *Science* **369**, 1010-1014 (2020).

81 Cao, Y. *et al.* Potent neutralizing antibodies against SARS-CoV-2 identified by high-throughput single-cell sequencing of convalescent patients' B cells. *Cell* **182**, 73-84. e16 (2020).

82 Ju, B. *et al.* Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature* **584**, 115-119 (2020).

83 Lim, S. A. *et al.* in *MAbs.* 1893426 (Taylor & Francis).

84 Liu, L. *et al.* Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature* **584**, 450-456 (2020).

85 Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260-1263 (2020).

86 Kemp, S. *et al.* Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion ΔH69. *V70* (2020).

87 Andreano, E. *et al.* SARS-CoV-2 escape in vitro from a highly neutralizing COVID-19 convalescent plasma. *bioRxiv*, doi:10.1101/2020.12.28.424451 (2020).

88 Iacobucci, G. (British Medical Journal Publishing Group, 2021).

89 Andrews, N. *et al.* Effectiveness of COVID-19 vaccines against the B. 1.617. 2 variant. (2021).

90 Naveca, F. *et al.* Phylogenetic relationship of SARS-CoV-2 sequences from Amazonas with emerging Brazilian variants harboring mutations E484K and N501Y in the Spike protein.

91 Control, C. f. D. & Prevention. SARS-CoV-2 variant classifications and definitions. *Retrieved March* **16**, 2020 (2021).

92 Pickett, B. E. *et al.* Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource

for the coronavirus research community. *Viruses* **4**, 3209-3226, doi:10.3390/v4113209 (2012).

93 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539, doi:10.1038/msb.2011.75 (2011).

94 UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **46**, 2699, doi:10.1093/nar/gky092 (2018).

95 Singh, H. & Raghava, G. P. ProPred1: prediction of promiscuous MHC Class-I binding sites. *Bioinformatics* **19**, 1009-1014, doi:10.1093/bioinformatics/btg108 (2003).

96 Bhasin, M. & Raghava, G. P. Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine* **22**, 3195-3204, doi:10.1016/j.vaccine.2004.02.005 (2004).

97 Larsen, M. V. *et al.* Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics* **8**, 424, doi:10.1186/1471-2105-8-424 (2007).

98 Singh, H. & Raghava, G. P. ProPred: prediction of HLA-DR binding sites. *Bioinformatics* **17**, 1236-1237 (2001).

99 Jensen, K. K. *et al.* Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* **154**, 394-406, doi:10.1111/imm.12889 (2018).

100 Dimitrov, I., Garnev, P., Flower, D. R. & Doytchinova, I. EpiTOP--a proteochemometric tool for MHC class II binding prediction. *Bioinformatics* **26**, 2066-2068, doi:10.1093/bioinformatics/btq324 (2010).

101 Dimitrov, I., Garnev, P., Flower, D. R. & Doytchinova, I. J. B. EpiTOP—a proteochemometric tool for MHC class II binding prediction. **26**, 2066-2068 (2010).

102 Jespersen, M. C., Peters, B., Nielsen, M. & Marcatili, P. J. N. a. r. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. **45**, W24-W29 (2017).

103 EL-Manzalawy, Y., Dobbs, D. & Honavar, V. J. J. o. M. R. A. I. J. Predicting linear B-cell epitopes using string kernels. **21**, 243-255 (2008).

104 Saha, S. & Raghava, G. in *International Conference on Artificial Immune Systems.* 197-204 (Springer).

105 Ponomarenko, J. *et al.* ElliPro: a new structure-based tool for the prediction of antibody epitopes. **9**, 514 (2008).

106 Rubinstein, N. D., Mayrose, I., Martz, E. & Pupko, T. J. B. b. Epitopia: a web-server for predicting B-cell epitopes. **10**, 287 (2009).

107 Kringelum, J. V., Lundegaard, C., Lund, O. & Nielsen, M. J. P. c. b. Reliable B cell epitope predictions: impacts of method development and improved benchmarking. **8**, e1002829 (2012).

108 Doytchinova, I. A. & Flower, D. R. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics* **8**, 4, doi:10.1186/1471-2105-8-4 (2007).

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

110 Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* **43**, D405-412, doi:10.1093/nar/gku938 (2015).

111 Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* **30**, 2785-2791, doi:10.1002/jcc.21256 (2009).

112 Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and

multithreading. *J Comput Chem* **31**, 455-461, doi:10.1002/jcc.21334 (2010).

113 McCammon, J. A., Gelin, B. R. & Karplus, M. Dynamics of folded proteins. *Nature* **267**, 585-590 (1977).

114 De Vivo, M., Masetti, M., Bottegoni, G. & Cavalli, A. Role of molecular dynamics and related methods in drug discovery. *Journal of medicinal chemistry* **59**, 4035-4061 (2016).

115 Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *Journal of computational chemistry* **26**, 1781-1802 (2005).

116 Huang, J. & MacKerell Jr, A. D. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of computational chemistry* **34**, 2135-2145 (2013).

117 Huang, J. & MacKerell Jr, A. D. Force field development and simulations of intrinsically disordered proteins. *Current opinion in structural biology* **48**, 40-48 (2018).

118 Woo, H. *et al.* Developing a Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein Model in a Viral Membrane. *J Phys Chem B* **124**, 7128-7137, doi:10.1021/acs.jpcb.0c04553 (2020).

119 Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J Comput Chem* **29**, 1859-1865, doi:10.1002/jcc.20945 (2008).

120 Nosé, S. & Klein, M. Constant pressure molecular dynamics for molecular systems. *Molecular Physics* **50**, 1055-1076 (1983).

121 Essmann, U. *et al.* A smooth particle mesh Ewald method. *The Journal of chemical physics* **103**, 8577-8593 (1995).

122 Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular

dynamics of n-alkanes. *Journal of computational physics* **23**, 327-341 (1977).

123 Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *Journal of molecular graphics* **14**, 33-38 (1996).

124 Nielsen, M., Lund, O., Buus, S. & Lundegaard, C. MHC class II epitope predictive algorithms. *Immunology* **130**, 319-328, doi:10.1111/j.1365-2567.2010.03268.x (2010).

125 Sanchez-Trincado, J. L., Gomez-Perosanz, M. & Reche, P. A. Fundamentals and Methods for T- and B-Cell Epitope Prediction. *J Immunol Res* **2017**, 2680160, doi:10.1155/2017/2680160 (2017).

126 Ahmad, B., Ashfaq, U. A., Rahman, M. U., Masoud, M. S. & Yousaf, M. Z. Conserved B and T cell epitopes prediction of ebola virus glycoprotein for vaccine development: An immuno-informatics approach. *Microb Pathog* **132**, 243-253, doi:10.1016/j.micpath.2019.05.010 (2019).

127 Freire, M. C. *et al.* Mapping Putative B-Cell Zika Virus NS1 Epitopes Provides Molecular Basis for Anti-NS1 Antibody Discrimination between Zika and Dengue Viruses. **2**, 3913-3920 (2017).

128 Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* **46**, W296-W303, doi:10.1093/nar/gky427 (2018).

129 Benkert, P., Biasini, M. & Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* **27**, 343-350, doi:10.1093/bioinformatics/btq662 (2011).

130 Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7**, 95-99 (1963).

131 Colovos, C. & Yeates, T. O. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci* **2**, 1511-1519, doi:10.1002/pro.5560020916 (1993).

132 Antunes, D. A. *et al.* Structural allele-specific patterns adopted by epitopes in the MHC-I cleft and reconstruction of MHC: peptide complexes to cross-reactivity assessment. *PLoS One* **5**, e10353 (2010).

133 Kotturi, M. F. *et al.* A multivalent and cross-protective vaccine strategy against arenaviruses associated with human disease. *PLoS Pathog* **5**, e1000695, doi:10.1371/journal.ppat.1000695 (2009).

134 Botten, J. *et al.* Identification of protective Lassa virus epitopes that are restricted by HLA-A2. *J Virol* **80**, 8351-8361, doi:10.1128/JVI.00896-06 (2006).

135 Wauquier, N. *et al.* HLA-C-restricted viral epitopes are associated with an escape mechanism from KIR2DL2(+) NK cells in Lassa virus infection. *EBioMedicine* **40**, 605-613, doi:10.1016/j.ebiom.2019.01.048 (2019).

136 Kotturi, M. F. *et al.* Polyfunctional CD4+ T cell responses to a set of pathogenic arenaviruses provide broad population coverage. *Immunome Res* **6**, 4, doi:10.1186/1745-7580-6-4 (2010).

137 Meulen, J. *et al.* Old and New World arenaviruses share a highly conserved epitope in the fusion domain of the glycoprotein 2, which is recognized by Lassa virus-specific human CD4+ T-cell clones. *Virology* **321**, 134-143, doi:10.1016/j.virol.2003.12.013 (2004).

138 Baral, P., Pavadai, E., Gerstman, B. S. & Chapagain, P. P. In-silico identification of the vaccine candidate epitopes against the Lassa virus hemorrhagic fever. *Sci Rep* **10**, 7667, doi:10.1038/s41598-020-63640-1 (2020).

139 Kramer, A. *et al.* Molecular basis for the binding promiscuity of an anti-p24 (HIV-1) monoclonal antibody. *Cell* **91**, 799-809 (1997).

140 Frank, S. A. *Immunology and evolution of infectious disease*. (Princeton University Press, 2002).

141 Ponomarenko, J. *et al.* ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics* **9**, 514, doi:10.1186/1471-2105-9-514 (2008).

142 Kringelum, J. V., Lundegaard, C., Lund, O. & Nielsen, M. Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLoS Comput Biol* **8**, e1002829, doi:10.1371/journal.pcbi.1002829 (2012).

143 Freire, M. *et al.* Mapping Putative B-Cell Zika Virus NS1 Epitopes Provides Molecular Basis for Anti-NS1 Antibody Discrimination between Zika and Dengue Viruses. *ACS Omega* **2**, 3913-3920, doi:10.1021/acsomega.7b00608 (2017).

144 Hossain, M. S., Reza, H. A. & Hossain, M. S. Immunoinformatics Approach Identified Two Highly Conserved B and T Cell Epitopes, LEASKRWAF and DSPLEASKRWAFRTG, for Effective Vaccine Design against Ebola and Marburg Viruses. *Journal of Advances in Microbiology*, 1-16 (2019).

145 Kumar, A., Jain, A. & Verma, S. K. Screening and structure-based modeling of T-cell epitopes of Marburg virus NP, GP and VP40: an immunoinformatic approach for designing peptide-based vaccine. *Trends in Bioinformatics* **6**, 10 (2013).

146 Hasan, M. *et al.* Vaccinomics strategy for developing a unique multi-epitope monovalent vaccine against Marburg marburgvirus. *Infection, Genetics and Evolution* **70**, 140-157 (2019).

147 Kalina, W. V., Warfield, K. L., Olinger, G. G. & Bavari, S. Discovery of common marburgvirus protective epitopes in a BALB/c mouse model. *Virol J* **6**, 132, doi:10.1186/1743-422X-6-132 (2009).

148 Fusco, M. L. *et al.* Protective mAbs and Cross-Reactive mAbs Raised by Immunization with Engineered Marburg Virus GPs. *PLoS Pathog* **11**, e1005016, doi:10.1371/journal.ppat.1005016 (2015).

149 Mahmud, S. N., Rahman, M., Kar, A., Jahan, N. & Khan, A. Designing of an Epitope-Based Universal Peptide Vaccine against Highly Conserved Regions in RNA Dependent RNA Polymerase Protein of Human Marburg Virus: A Computational Assay. *Anti-Infective Agents* **18**, 294-305 (2020).

150 Pervin, T. & Oany, A. R. Vaccinomics approach for scheming potential epitope-based peptide vaccine by targeting l-protein of Marburg virus. *In silico pharmacology* **9**, 1-18 (2021).

151 Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215-220, doi:10.1038/s41586-020-2180-5 (2020).

152 Chemmama, I. E., Pelea, A. C., Bhandari, Y. R., Chapagain, P. P. & Gerstman, B. S. Structural propensities and entropy effects in peptide helix-coil transitions. *Physical Review E* **86**, 031915 (2012).

153 Beck, D. A., Alonso, D. O., Inoyama, D. & Daggett, V. The intrinsic conformational propensities of the 20 naturally occurring amino acids and reflection of these propensities in proteins. *Proceedings of the National Academy of Sciences* **105**, 12259-12264 (2008).

154 McCallum, M. *et al.* SARS-CoV-2 immune evasion by the B. 1.427/B. 1.429 variant of concern. *Science* (2021).

155 Cai, Y. *et al.* Structural basis for enhanced infectivity and immune evasion of SARS-CoV-2 variants. *Science*, doi:10.1126/science.abi9745 (2021).

156 Gobeil, S. M. *et al.* D614G Mutation Alters SARS-CoV-2 Spike Conformation and Enhances Protease Cleavage at the S1/S2 Junction. *Cell Rep* **34**, 108630, doi:10.1016/j.celrep.2020.108630 (2021).

157 Bhattarai, N., Baral, P., Gerstman, B. S. & Chapagain, P. P. Structural and Dynamical Differences in the Spike Protein RBD in the SARS-CoV-2 Variants B. 1.1. 7 and B. 1.351. *The Journal of Physical Chemistry B* (2021).

158 Hurlburt, N. K. *et al.* Structural basis for potent neutralization of SARS-CoV-2 and role of antibody affinity maturation. *Nature communications* **11**, 1-7 (2020).

159 Du, S. *et al.* Structurally resolved SARS-CoV-2 antibody shows high efficacy in severely infected hamsters and provides a potent cocktail pairing strategy. *Cell* **183**, 1013-1023. e1013 (2020).

160 Verma, S. K., Yadav, S. & Kumar, A. In silico prediction of B- and T- cell epitope on Lassa virus proteins for peptide based subunit vaccine design. *Adv Biomed Res* **4**, 201, doi:10.4103/2277-9175.166137 (2015).

161 Faisal, A. M., Imtiaz, S. H., Zerin, T., Rahman, T. & Shekhar, H. U. Computer aided epitope design as a peptide vaccine component against Lassa virus. *Bioinformation* **13**, 417-429, doi:10.6026/97320630013417 (2017).

162 Hossain, M. U. *et al.* Design of peptide-based epitope vaccine and further binding site scrutiny led to groundswell in drug discovery against Lassa virus. *3 Biotech* **8**, 81, doi:10.1007/s13205-018-1106-5 (2018).

163 Boesen, A., Sundar, K. & Coico, R. Lassa fever virus peptides predicted by computational analysis induce epitope-specific cytotoxic-T-lymphocyte responses in HLA-A2.1 transgenic mice. *Clin Diagn Lab Immunol* **12**, 1223-1230, doi:10.1128/CDLI.12.10.1223-1230.2005 (2005).

VITA

PRABIN BARAL

Born, Ilam, Nepal

2008-2011	B.Sc., Physics Tribhuvan University, Kathmandu, Nepal
2012-2015	M.Sc., Physics Tribhuvan University, Kathmandu, Nepal
2016-2022	Ph.D., Physics Florida International University, Miami, FL, USA

PUBLICATIONS AND PRESENTATIONS

1. Baral, P., Bhattarai, N., Hossen, M. L., Stebliankin, V., Gerstman, B. S., Narasimhan, G., & Chapagain, P. P. (2021). Mutation-induced changes in the receptor-binding interface of the SARS-CoV-2 Delta variant B. 1.617. 2 and implications for immune evasion. *Biochemical and biophysical research communications*, *574*, 14-19.

2. R. Pokhrel, N. Bhattarai, P. Baral, B. Gerstman, J. Park, M. Handfield, P.P. Chapagain, Lipid II binding and transmembrane properties of various antimicrobial lanthipeptides, *Journal of Chemical Theory and Computation* (2021)

3. Bhattarai, N., Baral, P., Gerstman, B. S., & Chapagain, P. P. (2021). Structural and Dynamical Differences in the Spike Protein RBD in the SARS-CoV-2 Variants B. 1.1. 7 and B. 1.351. *The Journal of Physical Chemistry B*.

4. Bhattarai, N., Pavadai, E., Pokhrel, R., Baral, P., Hossen, M.L., Stahelin, R.V., Chapagain, P.P. and Gerstman, B.S., 2021. Ebola Virus Protein VP40 Binding to Sec24c for Transport to the Plasma Membrane. *Proteins: Structure, Function, and Bioinformatics*.

5. Koirala, R. P., Pokhrel, R., Baral, P., Tiwari, P. B., Chapagain, P. P., & Adhikari, N. P. (2021). Structural insights into the repair mechanism of AGT for methyl-induced DNA damage. *Biological Chemistry*.

6. Baral, P., Pavadai, E., Gerstman, B. S., & Chapagain, P. P. (2020). In-silico identification of the vaccine candidate epitopes against the Lassa virus hemorrhagic fever. *Scientific reports*, *10*(1), 1-11.

7. Pavadai, E., Bhattarai, N., Baral, P., Stahelin, R. V., Chapagain, P. P., & Gerstman, B. S. (2019). Conformational Flexibility of the Protein–Protein Interfaces of the Ebola Virus VP40 Structural Matrix Filament. *The Journal of Physical Chemistry B*, *123*(43), 9045-9053.

8. Pokhrel, R., Bhattarai, N., Baral, P., Gerstman, B. S., Park, J. H., Handfield, M., & Chapagain, P. P. (2019). Molecular mechanisms of pore formation and membrane disruption by the antimicrobial lantibiotic peptide Mutacin 1140. *Physical Chemistry Chemical Physics*, *21*(23), 12530-12539.

9. P Baral, E Pavadai, B Gerstman, P Chapagain, In-silico epitope prediction against the Lassa virus glycoprotein, American Physical Society (APS) Meeting, Virtual, Mar 15-19, 2021. (Bulletin of the American Physical Society, 2021)

10. P Baral, E Pavadai, B Gerstman, PP Chapagain, Lassa Virus Epitope-Allele Complexes Identified Through Computational Modeling, Biophysical Society Meeting, San Diego, CA, Feb 15-19, 2020 (Biophysical Journal, Volume 118, Issue 3).

11. P Baral, N Bhattarai, R Pokhrel, B Gerstman, PP Chapagain, Molecular Dynamics Investigations of Enzyme Conformational Changes, Biophysical Society Meeting, Baltimore, MD, March 2-6, 2019, (Biophysical Journal, Volume 116, Issue 3).