

Immunoinformatics-Guided Designing of an Epitope-Based Vaccine Against Severe Acute Respiratory Syndrome-Coronavirus-2 Spike Glycoprotein

Ahmed Rakib¹, Saad Ahmed Sami¹, Arkajyoti Paul^{2,3}, Asif Shahriar⁴, Abu Montakim Tareq⁵, Nazim Uddin Emon⁵, Nusrat Jahan Mimi¹, Md. Mustafiz Chowdhury¹, Taslima Akter Eva¹, Sajal Chakraborty³, Sagar Shil³, Sabrina Jahan Mily⁶, Talha Bin Emran^{2,3*}

¹Department of Pharmacy, Faculty of Biological Sciences, University of Chittagong, Chittagong-4331, Bangladesh

²Drug Discovery, GUSTO A Research Group, Chittagong-4000, Bangladesh

³Department of Pharmacy, BGC Trust University Bangladesh, Chittagong-4381, Bangladesh

⁴Department of Microbiology, Stamford University Bangladesh, 51 Siddeswari Road, Dhaka-1217, Bangladesh

⁵Department of Pharmacy, International Islamic University Chittagong, Chittagong-4318, Bangladesh

⁶Department of Gynaecology and Obstetrics, Banshkhali Upazila Health Complex, Jaldi Union, Chittagong-4390, Bangladesh

Running title: *de novo* vaccine design against SARS-CoV-2

*Correspondence:

Dr. T.B. Emran

talhabmb@bgctub.ac.bd or talhabmb@gmail.com

Keywords: COVID-19, SARS-CoV-2, *de novo* vaccine, epitope, immunity

Abstract

Currently, with a large number of fatality rates, coronavirus disease-2019 (COVID-19) has emerged as a potential threat to human health worldwide. It has been well-known that severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2) is responsible for COVID-19 and World Health Organization (WHO) proclaimed the contagious disease as a global pandemic. Researchers from different parts of the world amalgamate together in quest of remedies for this deadly virus. Recently, it has been demonstrated that the spike glycoprotein (SGP) of SARS-CoV-2 is the mediator behind the entrance into the host cells. Our group has comprehensively analyzed the SGP of SARS-CoV-2 through multiple sequence analysis along with the phylogenetic analysis. Further, this research work predicted the most immunogenic epitopes for both B-cell and T-cell. Notably, we focused mainly on major histocompatibility complex (MHC) class I potential peptides and predicted two epitopes; WTAGAAYY and GAAYYVGY, that bind with the MHC class I alleles which are further validated by molecular docking analysis. Furthermore, this study also proposed that the selected epitopes were shown availability in a greater range of the population. Hence, our study comes up with a strong base for the implementation of designing novel vaccine candidates against SARS-CoV-2, however adequate laboratory works will need to be conducted for the appropriate application.

1 Introduction

Pandemics caused by many life-threatening human pathogens have played a significant role in shaping human history throughout the ages. The world is currently battling a global pandemic that burst onto the scene in late December 2019. A cluster of pneumonia cases of unknown etiology was reported in Wuhan, the capital city of the Hubei province of the People's Republic of China. Later it was revealed that the causative agent of this outbreak was a novel coronavirus named SARS-CoV-2 (previously named as 2019-nCoV). The clinical condition associated with this novel coronavirus is referred to as COVID-19 (1–3). On March 11, 2020 WHO (World Health Organization) categorized the current outbreak of COVID-19 as a pandemic. This viral infection appears to constitute a major global threat to humans and has shown its devastating effects in all the WHO regions. As of April 29, 2020, more than 3 million cases and over 207 thousand deaths have been reported across the globe with developed countries like the USA, Italy, Spain, France, Germany, the United Kingdom bearing the large burden of morbidity and mortality (4). The number of COVID-19 cases has continued to escalate exponentially and can be considered as the largest outbreak of atypical pneumonia in recent times.

Tyrell and Bynoe first described the coronaviruses back in 1966 (5). Coronaviruses are pleomorphic or spherical particles with a positive single-strand RNA (+ssRNA) and are very common among mammals, birds and can be transmitted to the human body by pathogen spillover. They were named coronavirus as these virions consist of a core-shell and crown like 9-12 nm-long surface spikes on the outer surface of the virus resembling a solar corona. Their genome size is the largest among all the RNA viruses ranging from 27 to 32 kb in length that encodes structural and non-structural proteins of the coronavirus. Phenotypic and genotypic diversity gives them the ability to adapt to the new environment through mutation and

recombination. These mutations lied on the surface of the protein induce its sustainability and leave the immune system in the blind spot which makes the current COVID-19 infection more superior than other previous strains (6). Among the four genera (alpha , beta, gamma, delta) of coronaviruses, the beta-coronaviruses may cause severe disease and fatalities to the human (7). Including SARS-CoV-2, seven subtypes of coronaviruses were identified in the last few decades that can infect humans.

SARS-CoV-2 differs from other beta coronaviruses in terms of its significantly higher infectivity and low mortality rate. It belongs to the B lineage of the beta-coronavirus in the Nidovirales order, *Coronaviridae* family, and *Orthocoronavirinae* subfamily. After the two previously reported coronavirus- SARS and MERS, this is the third zoonotic coronavirus breakout of the 21st century and closely related to its predecessors (8,9). As it is a zoonotic virus, there will be an intermediate host by which it can be transmitted to a human. Preliminary investigations predict that it was caused due to zoonotic transfer from bats (9). Primarily, ome environmental specimens of the Huanan wet market in Wuhan were deemed positive for this infection but no specific association with an animal is confirmed yet based on the WHO report (3).

As this contagious disease is mainly a respiratory disease, it might affect only the lungs in most of the cases. The infection is transmitted between people during close contact, which occurs via spraying of droplets from the infected individual by coughing or sneezing (10,11). These droplets usually fall into the surfaces after breathing out and touching the contaminated surface followed by touching other parts of the face like eyes, nose, or mouth may result in spreading of the infection (12,13). The symptoms of this coronavirus may range from mild or moderate fever to severe pneumonia. The time from exposure to onset of symptoms may range from 2-14 days with an average of five days. The spreading of the infection may be possible before symptoms

appear. The clinical pathology greatly resembled its two predecessors with less upper respiratory and gastrointestinal symptoms. The most common symptoms or combinations of symptoms including fever, fatigue, dry cough, dyspnea, alveolar edema, sore throat, new loss of taste or smell and shortness of breath (14). Older people with medical comorbidities or multi-organ failure such as hypertension, cardiovascular disease, and diabetes are more likely to get infected with worse outcomes (14). Severe cases can lead to cardiac injury, respiratory failure, acute respiratory syndrome, hepatic injury, neurological complications, and death (15).

The structure of SARS-CoV-2 generally includes a polyprotein (the open reading frame 1a and 1b, Orf1ab), four major structural proteins such as S protein (Spike glycoprotein-SGP), E protein (Envelope), M protein (Membrane), and N protein (Nucleocapsid) and five accessory proteins namely Orf3a, Orf6, Orf7a, Orf8 and Orf10 (16). This novel coronavirus particularly encodes an additional glycoprotein having acetyl esterase and hemagglutination (HE) attributes which were not observed in its predecessors (17). Among the structural proteins, the S protein is a multifunctional molecular machine that can attach to the specific human host receptor, ACE2 on the surface of human cells and mediates entry of viral particles into the host cells with the aid of its protease which cleaves the Spike protein into S1 and S2 subunits (18,19). The actual binding to the ACE2 receptor on the host cell surface occurs through the RBD found in the S1 subunit of the virus. After that, viral and host membranes are fused by the S2 subunit. The viral genome RNA is released into the cytoplasm after this membrane fusion. So for developing neutralizing antibodies the RBD of SARS-COV-2 SGP might be an ideal target.

The identification of B-cell and T-cell epitopes for spike glycoproteins are critical for developing an effective vaccine. Although humans may mount an antibody response against viruses normally, only neutralizing antibodies can block the entry of viruses into human cells completely

(20). Antibody binding sites location on a viral protein strongly affects the body's ability to produce neutralizing antibodies (21). It is crucial to find out whether SARS-CoV-2 has any potential antibody binding sites (B-cell epitopes) with its known human entry receptor, ACE2 near their interacting surface.

Apart from neutralizing antibodies, the human body also relies on cytotoxic $CD8^+$ T-cells and helper $CD4^+$ T-cells to eliminate viruses from the body. The presentation of a peptide will allow a B-cell to receive stimulation from a helper T-cell and become a plasma cell so that it can generate antibodies. T-cell epitopes in coalition with the MHC proteins are presented to T-cells for anti-viral T-cell responses. Cytotoxic T-cells recognize peptides that are received from the intracellular space presented by MHC class I molecules ($CD8^+$ T-cell epitopes) while helper T-cells recognize extracellular peptides presented by MHC class II molecules ($CD4^+$ T-cell epitopes). The pMHC (peptide: MHC complex) interacts with the T-cell receptor and activates the cellular immune response. The inclusion of $CD4^+$ T-cell epitopes plays a key role in vaccine design as it provides cognate help and elicit vigorous humoral and cytotoxic $CD8^+$ T-cell responses and neutralizing antibodies (22). T-cell epitope-based vaccines have been explored in recent years as they can target conserved regions of the virus T cell responses and provide long-term protection against different strains of viruses (23).

Effective promiscuous epitopes binding to a variety of HLA alleles for wider dissemination with no human cross-reactivity is crucial since COVID-19 infection has affected populations all over the world. Our present study has been embarked upon with the clear objective of designing an epitope-based peptide vaccine against COVID-19 infection using in silico methods and considering SGP of the SARS-CoV-2. Here, we targeted the epitopes in the SGP because of their lingering immune response reported against SARS coronavirus previously (26). For the

identified epitopes, we incorporated the information on the associated MHC alleles so that we can provide a list of epitopes that would maximize population coverage across the world. Therefore, we designed an epitope-based peptide vaccine in the quest for finding potent targets against SARS-CoV-2 SGP using different computational tools with an expectation that the wet laboratory research will validate the outcome of our investigation.

2 Materials and Methods

The methodologies used for peptide vaccine development for SARS-CoV-2 SGP are shown in **Figure 1**.

2.1 Protein sequence retrieval and sequence analysis

The SARS-CoV-2 SGP sequence was extracted from UniProt database (UniProt entry: P0DTC2) in FASTA format (24). The understanding of the features, function, structure, and evaluation is mainly based on the process of sequence analysis which depicts the process of subjecting DNA, RNA, or peptide sequences to wide ranges of analytical methods. We implied BLASTp that screens homologous sequences from its database and selected those sequences that are more similar to our SARS-CoV-2 SGP; we also performed multiple sequence alignment (MSA) using the ClustalW web server with default settings and a phylogenetic tree was established using Clustal tree format using EMBL-EBI web server (25).

2.2 Protein antigenicity prediction

To determine the potent antigenic protein of the SARS-CoV-2 SGP, we used the online server VaxiJen v2.0, with a default threshold value (26). All the antigenic proteins of SARS-CoV-2 SGP with their respective scores were obtained then sorted in Notepad++.

2.3 Protein secondary and tertiary structure prediction

The secondary and tertiary structure of the SARS-CoV-2 SGP was predicted by using Phyre2 web server, which create a model protein from the given sequence and predict both 2D and 3D structure of the protein (27). The three-dimensional model was validated using PROCHECK and ERRAT web servers (28,29).

2.4 T-cell epitope prediction

2.4.1 CD8⁺ T-cell epitope identification

NetCTL 1.2 server was used in this experiment for the identification of the T-cell epitope, using a 0.95 threshold to maintain sensitivity and specificity of 0.90 and 0.95, respectively (30). The tool expands the prediction for 12 MHC-I supertypes and unified the prediction of peptide MHC-I binding, proteasomal C-terminal cleavage with TAP transport efficiency. These predictions were performed by an artificial neural network, weighted TAP transport efficiency matrix and a combined algorithm for MHC-I binding and proteasomal cleavage efficiency was then used to determine the overall scores and translated into sensitivity/specificity. Based on this overall score, ten best peptides (epitopes) were selected for further evaluation.

For the prediction of peptides binding to MHC-I, we used a tool from the Immune Epitope Database (IEDB) and calculate IC₅₀ values for peptides binding to specific MHC-I molecules.²⁸ For the binding analysis, all the frequently used alleles were selected with a word length of nine residues and binding affinity < 200 nm for further analysis. Another tool (named as MHC-NP) provided by the IEDB server was used to assess the probability that a given peptide was naturally processed and bound to a given MHC molecule (31).

2.4.2 Epitope conservancy and immunogenicity prediction

The degree of similarity between the epitope and the target (i.e. given) sequence is elucidated by epitope conservancy. This property of epitope gives us the promise of its availability in a range of different strains. Hence for the analysis of the epitope conservancy, the web-based tool from IEDB analysis resources was used (32). Immunogenicity prediction can uncover the degree of influence (or efficiency) of the respective epitope to produce an immunogenic response. The T-cell class I pMHC immunogenicity predictor at IEDB, which uses amino acid properties as well

as their position within the peptide to predict the immunogenicity of a class I peptide MHC (pMHC) complex (33).

2.4.3 Prediction of population coverage and allergenicity assessment

The population coverage tool from IEDB was applied to determine the population coverage for every single epitope by selecting HLA alleles of the corresponding epitope.

Allergenicity of the predicted epitope was calculated using AllerTop v2.0, which is an alignment-free server, used for *in silico* based allergenicity prediction of a protein-based on its physiochemical properties (34).

2.4.4 HLA and epitope interaction analysis using molecular docking studies

2.4.4.1 Epitope model generation

A web-based server, PEP-FOLD were used to predict the 3D structures of the selected epitopes (35). For each sequence, the server predicted five probable structures. The energy of each structure was determined by SWISS-PDB VIEWER and the structure with the lowest energy was chosen for further analysis (36).

2.4.4.2 Retrieval of HLA allele molecule

The three-dimensional structure of the HLA-B*15:25 was not available on Protein Data Bank (RCSB-PDB). We selected homology modeling using SWISS-MODEL web server to generate the three-dimensional structure of the HLA-B*15:25 (Accession id: HLA00188) (37). The validation of the predicted structure was done using PROCHECK, VERIFY 3D, ERRAT (28,29,38).

2.4.4.3 Molecular docking analysis

Molecular docking analysis was performed using Autodock vina in PyRx 0.8, by considering the HLA-B*15:25 modeled protein as receptor protein and identified epitopes as ligand molecule (39). Firstly, we used the protein preparation wizard of UCSF Chimera (Version 1.11.2) to prepare the protein for docking analysis by deleting the attached ligand, adding hydrogens and Gasteiger–Marsili charges. The file was then converted into pdbqt format using OpenBabel (40,41). The energy form of the ligand was minimized and converted to pdbqt format by OpenBabel in PyRx 0.8. The parameters used for the docking simulation were set to default. The size of the grid box in AutoDock Vina was kept at $50.9455 \times 43.5609 \times 61.2260$ Å respectively, for X, Y, and Z-axis. AutoDock Vina was implemented via the shell script offered by AutoDock Vina developers. Docking results were observed by the negative score in kcal/mol, as the binding affinity of ligands (42).

2.5 B-cell epitope identification

The prediction of B-cell epitopes was performed to find the potential antigen that assures humoral immunity. To detect B-cell epitope, various tools from IEDB were used to identify the B-cell antigenicity, together with the Emini surface accessibility prediction, Kolaskar and Tongaonkar antigenicity scale, Karplus and Schulz flexibility prediction, and Bepipred linear epitope prediction analysis (43–46). Since antigenic parts of a protein belonging to the beta-turn regions, the Chou and Fasman beta-turn prediction tool was also used (47).

3 Results

3.1 Sequence retrieval and analysis

In this study, the protein sequence of the SARS-CoV-2 SGP was retrieved from the UniProt database and then performed BLASTp. From plenty of homologues, we have selected only 17 homologues having more than 60% identical sequences. MSA were performed (**Supplementary Data-1**) and a phylogenetic tree was constructed (**Figure S1**). The findings from the MSA and phylogenetic data documented that the protein sequences have a closer relationship.

3.2 Antigenic protein prediction

The most potent antigenic protein of SARS-CoV-2 SGP was predicted by VaxiJen v2.0, which is based on the auto-cross covariance transformation of protein sequences into uniform vectors of principal amino acid properties. The overall antigen prediction score was 0.4683 (probable antigen) at 0.4 threshold value.

3.3 Protein structure prediction and validation

In this study, we determined the secondary structure and tertiary of the SARS-CoV-2 SGP using Phyre2 web server, which depicts that 26% of the residues were from α -helix, whereas 37% of the residues were from β -strands. Previous research has already unveiled that the antigenic part of the protein mostly remains as β -sheet. Moreover, we carried out the homology modeling of SARS-CoV-2 SGP using Phyre2 web server. Ramachandran Plot analysis using PROCHECK web server showed that the modeled SARS-CoV-2 SGP has 74.7% residues in most favored region and 25% residues in the allowed region (**Figure S2**) and ERRAT predicted the quality factor of 68.5484 (Data not shown).

3.4 T-Cell epitope identification

3.4.1 CD8⁺ T-cell epitope identification

On the basis of highest combinatorial score and MHC class I binding, top twenty three epitopes were predicted by NetCTL 1.2 server from the given protein sequence. Further, the antigenicity of each selected peptides was predicted by VaxiJen v2.0 server and we found that total ten epitopes has shown being antigenic (**Table S1**). We utilized the MHC-I binding prediction tool from the IEDB server, which is based on stabilized matrix method (SMM), we chose those MHC-I alleles for which the epitopes showed the highest affinity ($IC_{50} < 200$ nm). Proteasomes play an important role during the conversion of protein into peptide and these peptide molecules are homogeneous to class I MHC molecules and the peptide-MHC molecule after the proteasomal cleavage were presented as T-helper cells after the transportation into the cell membrane. The total score of each epitope-HLA interaction was taken into consideration and higher processing efficiency was meant by obtaining a higher score. In this experiment, the epitope WTAGAAAYY interacted with total 17 MHC class I alleles, including HLA-A*29:02 , HLA-A*26:01, HLA-A*68:01, HLA-C*12:03, HLA-B*15:25, HLA-B*35:01, HLA-C*03:02, HLA-A*30:02, HLA-A*01:01, HLA-B*15:01, HLA-B*15:02, HLA-C*16:01, HLA-A*25:01, HLA-C*02:09, HLA-C*02:02, HLA-C*12:02, HLA-C*14:02 (**Table 1**). In addition, two other epitopes, namely GAAAYYVGY and STQDLFLPF interacted with MHC I alleles, the former interacted with five HLAs and the number was exhibited nine in case of later, but, like WTAGAAAYY, both the epitopes interacted with HLA-B*15:25, where the tendency of binding were mostly shown by epitope GAAAYYVGY, whereas the epitope WTAGAAAYY delineated greatest immunogenicity predicted by the I-pMHC immunogenicity prediction analysis (**Table 1**). Furthermore, all the predicted epitopes had a maximum identity for conservancy hit and 100% maximum identity was found (**Table 1**). Therefore, we considered the aforementioned three

epitopes for further analysis. In addition, MHC-NP finding unleashed that the epitopes were bound with the MHC class I alleles naturally (**Table S2**).

3.4.2 Population Coverage

The cumulative amount of the population coverage was obtained for the three predicted epitope, WTAGAAAYY, GAAAYYVGY and STQDLFLPF respectively. The results of the population coverage analysis represented that with 77.98% coverage, West Africa found the highest coverage region and Europe has demonstrated the second highest coverage region with a percentage of 77.91. The results of the population coverage were shown in **Table 2** and **Figures S3-S6**.

3.4.3 Allergenicity assessment

The AllerTop server was used for the identification of the allergic reaction caused by a vaccine in an individual which might be harmful or life-threatening. The allergenicity of the selected epitope was calculated using the AllerTop tool and predicted as probable non-allergen.

3.4.4 Molecular docking analysis for HLA and epitope interaction

In this experiment, the validation of the interaction between the HLA molecules and our predicted potential epitope was done by molecular docking simulation using Autodock Vina in PyRx 0.8. We previously mentioned that the allele HLA-B*15:25 interacted with the three predicted epitopes. However, the three-dimensional structure of the HLA-B*15:25 was not available on the Protein Data Bank. As a result, the HLA-B*15:25 was modeled using SWISS-MODEL web server (**Figure 2**). The modeled HLA-B*15:25 was then verified using PROCHECK, ERRAT and VERIFY-3D web servers. The Ramachandran plot analysis using PROCHECK server depicted that 94.2% amino acid residues were from favored region and 5.4% residues from allowed region (**Figure 2**). Further, the ERRAT server predicts the overall quality

for the non-atomic bond interactions and the score found for the modeled HLA-B*15:25 allele was 95.4717, which was greater than the threshold value of 50 (**Figure 2**). Finally, the validation of sequence-to-structure agreement was depicted by VERIFY 3D. The results from the VERIFY 3D server for the modeled HLA-B*15:25 alleles asserted pass, showing that 98.91% of the residues have averaged 3D-1D score more than 0.2 (**Figure 2**).

For docking analysis, we considered the modeled HLA molecule as receptor and the selected three epitopes as ligand molecules. The results of the docking experiments showed that the epitope GAAAYYVGY binds with HLA-B*15:25 allele with the greatest binding affinity which was calculated as -8.8 kcal/mol, and the binding affinity of the epitope WTAGAAAYY were almost equal to the epitope GAAAYYVGY, but the binding affinity of the epitope STQDLFLPF were found only -7.8 kcal/mol (**Table 3**). From the visualization of the docking results, it has been lucid that the epitope WTAGAAAYY formed conventional hydrogen bonds (H-bonds) with five amino acid residues, including, Asn94, Arg121, Trp171, Glu176, Gln179 and pi-pi stacking with Tyr123 and Tyr183 residues with an unfavorable bond with Ser148 respectively (**Figure 3** and **Figure S7**). However, no unfavorable bond was visualized for the epitope GAAAYYVGY and more H-bonds were visualized for the epitope GAAAYYVGY as well as attractive charges with Lys170 and Arg121 residues (**Figure 4** and **Figure S8**).

3.5 B-cell epitope prediction

In the current study, by utilizing amino acid scale-based method, the B-cell epitope was identified. Different analysis methods were used for the prediction of continuous B-cell epitope, which were shown in **Table 4**, **Figure 5** and **Figures S9-S10**.

First of all, Bepipred linear epitope prediction was used, which is regarded as the best single method for predicting linear B-cell epitopes using a Hidden Markov model. Our analysis revealed that the peptide sequences from 805 to 816 amino acid residues were able to induce the desired immune response as B-cell epitopes.

The β -turns were predicted by Chaus and Fasman β -turn prediction method. The region 807-813 residues were predicted as a β -turn region with a score of 1.484, which was higher than the average score.

For antigenicity prediction, the Kolaskar and Tongaonkar antigenicity prediction methods were implied. The method evaluates the antigenicity based on the physicochemical properties of amino acids and their abundances in experimentally known epitopes. The average antigenic propensity of our SARS-CoV-2 SGP was 1.041 with a maximum of 1.261. Region 803-808 residues were found prone to antigenicity, scoring 1.064, which was greater than the average score. Besides, the average flexibility of 0.993 and a minimum of 0.866 were predicted by the Karplus and Schulz flexibility prediction method. The residues from 809-815 were found to be most flexible by scoring 1.101, that was almost equal in comparison with the highest score. The Parker hydrophilicity prediction tool predicts the hydrophilicity of SARS-CoV-2 SARS-CoV-2 SGP with an average score of 1.238, and in alignment with the previous B-cell epitope results, the region 808-814 amino acid residues possessed the highest score of 5.514.

For predicting the surface ability, this study includes the Emini surface accessibility prediction method. The average surface accessibility was 1.0 and a minimum 0.042. In alignment with the previous B-cell epitope results, we predicted the peptide sequence from 810-815 had the better surface ability, with the highest score of 5.662.

4 Discussion

Currently, no clinically proven vaccine grants immunity from the infection due to the elusive nature of SARS-CoV-2. Researchers have examined different repurposed compounds from other viral infections to treat COVID-19 infection but the treatment benefit derived was dubious in most cases (48). To date, with the advancement of plenty of experiments, it has been hypothesized that a vaccine development for SARS-CoV-2 is not far away and several platforms of the vaccine including DNA vaccine, RNA vaccine, protein subunit vaccine, virus-like particles (VLP) based vaccines, vector-based vaccine have been explored. However, further investigations are critical for developing a harmless vaccine that is applicable for different age groups, and for this purpose, extensive relevant data of the genomic and structural organization of the SARS-CoV-2 is crucial. Recent studies suggested that inactivated and live-attenuated vaccines prove unsuccessful against COVID-19 as the risk of infection is high and difficulty in isolating the virus. Many researchers have concentrated on making mRNA and DNA vaccines that eliminate the risk of any unwanted reactions. But these novel methods face many obstacles due to their experimental status and inherently carry very little antigenicity. Epitope-based vaccines offer a viable alternative since they can elicit potent immune responses without causing undesirable allergic reactions and have been already used against a variety of pathogens successfully. Designing of vaccines by conventional approaches are not only time consuming but also inefficient. To speed up this laborious process of vaccine design, different immunoinformatics tools can be significantly useful. Immunoinformatics tools such as epitope prediction can be used to construct vaccines in silico and then experimentally validate the results (49).

A recent review documented that the particular genomic and proteomic analysis of respective proteins of SARS-CoV-2 including M, SGP, E, N proteins propounds the eligibility in representing protective immune responses against SARS-CoV-2 (50). Additionally, recent studies also proposed that the epitopes that are present on the SGP and N proteins of SARS-CoV-2 elicited strong T-cell immune responses for a longer period and therefore, can be executed as an ideal vaccine candidate against SARS-CoV-2. Significantly, the receptor-binding domain (RBD) of the SGP of the two predecessors; SARS-CoV and MERS-CoV delineated strongly potent neutralizing antibody responses and has efficient to be developed vaccine for the treatment of SARS and MERS infection, respectively (51,52). In alignment with the previous analysis on the ancestors of SARS-CoV-2 and recent literature shreds of evidence, the current study includes the application of computational biology techniques as well as immunoinformatics tool for the identification of epitope-based vaccine utilizing the SARS-CoV-2 SGP.

Previously, it has been well recognized that the development of vaccines was rudimentarily dependent on B-cell immunity, but recently, it has been well established that T-cell epitopes provide more long-lasting immune response which is mainly mediated by CD8⁺ T-cells and as a result of antigenic drift (53). In the present study, concentrating on MHC class I potential peptide epitopes, we were mainly predicted not only T-cell but also B-cell epitopes, that are capable of showing immune responses in several ways. Many criteria are needed to consider while identifying a protein sequence-based epitope into a vaccine candidate of which allergenicity is regarded as one of them. We served the SARS-CoV-2 SGP sequence into the VaxiJen server and predicted the antigenicity of the protein sequence. Further, a total of ten potent epitopes have been predicted from the NetCTL 1.2 server and the epitopes were further taken for the

progressive analysis which also showed MHC class I interaction respectively. Besides, all peptides except GAEHVNNSY were able to interact with the MHC class I alleles, and WTAGAAAYY interacted with the most MHC class I alleles. Amongst them the allele HLA-A*29:02 represented the highest binding score of 1.51. Interestingly, three of the epitopes including WTAGAAAYY, GAAAYYVGY, STQDLFLPF showed binding interaction with the same MHC class I allele i.e. HLA-B*15:25, of which, GAAAYYVGY exerted the greater affinity towards HLA-B*15:25. Also, the conservancy of the epitopes which was predicted by the IEDB conservancy analysis tool delineated that all of our predicted epitopes had the maximum identity of 100%. Hence, the three aforementioned epitopes were taken into consideration for further analysis.

During vaccine development, allergenicity is regarded as a notable obstacle. Most importantly, CD4⁺ T-cells are responsible for provoking an allergic reaction and along with immunoglobulin E, type 2 T helper cell stimulates the allergic reactions (54). In the current experiment, we evaluated the allergenicity using AllerTop 2.0, which is based on its high sensitivity while identifying new allergens in comparison with the known allergens (34). AllerTop predicted our selected epitope as non-allergen.

MHC class I and class II molecules play a pivotal role in presenting peptides on the cell surface with a view to identifying by the T cells, where MHC class molecules present the shorter molecules (generally 8-11 amino acid long) and MHC class II molecules present longer peptides with 13-17 amino acid residues (55). In this present study, we determined the binding affinity of the predicted peptide sequences with modeled HLA-B*15:25. The findings from the molecular docking analysis revealed that the epitope GAAAYYVGY and WTAGAAAYY interacted with the greatest affinity with the modeled HLA-B*15:25 allele, as the more negative result, implied

to the greatest interaction (56). Besides, the abovementioned two epitopes were able to interact with the allele through H-bond and hydrophobic (pi-pi stacking, pi-alkyl) interactions, whereas attractive charges were also responsible for the binding in case of GAAAYYVGY. Conversely, the epitope STQDLFLPF has possessed a less negative binding affinity; hence the binding affinity was less in comparison with the other two epitopes.

Furthermore, population coverage is considered most noteworthy during vaccine development due to the variability of HLA according to ethnicity and geographical region. For the analysis of population coverage, we utilized the population coverage tool from IEDB and predicted that our selected three epitopes covered almost all available regions throughout the world, where the highest coverage was observed in West Africa. Surprisingly, from which regions the most cases of the infection were reported, Europe and North America delineated coverage of more than 65%, where the coverage of Europe was almost the same as West Africa. In addition, the epitopes showed coverage of 62.39% in East Asia, where the first COVID-19 case reported.

Importantly, the B-cell epitope elicits a stronger immune response but does not cause any side effects. As a corollary we also predicted the B-cell epitope utilizing the IEDB database and found that the protein sequences from 805-816 residues as B-cell epitope. The predicted region mayhap stimulate the potential immune response and presumably important for the development of a vaccine candidate.

5 Conclusions

The improvement in the field of immunoinformatics has become a potential field for predicting peptide-based vaccines. Viruses show not only T-cell but also humoral immunity. As a result, our predicted epitope presumably improves immunity against SARS-CoV-2. The presumption is

based on the fundamental principles of immunity, that confers the attachment of foreign bodies with the host immune cell, provoking immune responses, and transfers the relevant information to T-cell and B-cell. In the present study, our investigated epitopes simulate the interaction to CD8⁺ cells antigen presentation utilizing in silico approaches. However, the present study is a preliminary approach for predicting epitope-based vaccine against SARS-CoV-2 and we hope that the predicted epitope will pave the way for further laboratory analysis for designing novel vaccine candidates against SARS-CoV-2.

6 Conflict of Interest

The authors report no conflicts of interests in this work.

7 Author contributions

Study concept and design: A.R., S.A.S., and T.B.E. Acquisition of data: A.R., S.A.S., A.S., and A.P. Analyses and interpretation of data: A.R., S.A.S., A.M.T., N.U.E., N.J.M., M.M.C., T.A.E., S.C., S.S., and T.B.E. Drafting the manuscript: A.R., S.A.S., S.J.M., and T.B.E. Critical revision of the manuscript for important intellectual content: S.J.M., and T.B.E. Technical or material support: A.R. Study supervision: T.B.E.

8 Funding

This work is conducted with the individual funding of all authors.

9 Acknowledgments

The authors would like to thank Cambridge Proofreading® & Editing LLC. (<https://proofreading.org/>) for editing a draft of this manuscript.

10 Data availability statement

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

11 References

1. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus--infected pneumonia in Wuhan, China. *Jama*. 2020; doi: 10.1001/jama.2020.1585
2. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus--infected pneumonia. *New England Journal of Medicine*. 2020; doi: 10.3410/f.737281536.793571806
3. Gralinski LE, Menachery VD. Return of the Coronavirus: 2019-nCoV. *Viruses*. 2020;12(2):135. doi: 10.3390/v12020135
4. Organization WH, others. Coronavirus disease 2019 (COVID-19): situation report, 72. 2020.
5. Tyrrell DAJ, Bynoe ML, others. Cultivation of viruses from a high proportion of patients with colds. *Lancet*. 1966;76–7. doi: 10.1016/s0140-6736(66)92364-6
6. Christie JM, Chapel H, Chapman RW, Rosenberg WM. Immune selection and genetic sequence variation in core and envelope regions of hepatitis C virus. *Hepatology*. 1999;30(4):1037–44. doi: 10.1002/hep.510300403
7. Yang D, Leibowitz JL. The structure and functions of coronavirus genomic 3' and 5' ends. *Virus research*. 2015;206:120–33. doi:10.1016/j.virusres.2015.02.025
8. Xiong C, Jiang L, Chen Y, Jiang Q. Evolution and variation of 2019-novel coronavirus. *Biorxiv*. 2020.
9. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak

- associated with a new coronavirus of probable bat origin. *nature*. 2020;579(7798):270–3.
doi: 10.1038/s41586-020-2012-7
10. Kanne JP. Chest CT findings in 2019 novel coronavirus (2019-nCoV) infections from Wuhan, China: key points for the radiologist. Radiological Society of North America; 2020.
 11. Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner N, Ke R. Early Release-High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2. doi:10.3201/eid2607.200282.
 12. Organization WH, others. Q&A on coronaviruses (COVID-19). Retrieved April 6th. 2020;
 13. Organization WH, others. Modes of transmission of virus causing COVID-19: implications for IPC precaution recommendations: scientific brief, 29 March 2020. 2020.
 14. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*. 2020;395(10223):507–13.
 15. Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, et al. First case of 2019 novel coronavirus in the United States. *New England Journal of Medicine*. 2020; doi: 10.1016/s0140-6736(20)30211-7
 16. Ahmed SF, Quadeer AA, McKay MR. Preliminary identification of potential vaccine targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. *Viruses*. 2020; doi: 10.3390/v12030254
 17. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Hu Y, et al. Complete genome characterisation of a novel coronavirus associated with severe human respiratory disease in Wuhan, China. *bioRxiv*. 2020; doi: 10.1101/2020.01.24.919183

18. Tortorici MA, Veerler D. Structural insights into coronavirus entry. *Advances in virus research*. 2019;105:93–116. doi: 10.1016/bs.aivir.2019.08.002
19. Zhu X, Liu Q, Du L, Lu L, Jiang S. Receptor-binding domain as a target for developing SARS vaccines. *Journal of thoracic disease*. 2013;5(Suppl 2):S142.
20. Suarez DL, Schultz-Cherry S. Immunology of avian influenza virus: a review. *Developmental & Comparative Immunology*. 2000;24(2–3):269–83. doi: 10.1016/s0145-305x(99)00078-6
21. Briney B, Sok D, Jardine JG, Kulp DW, Skog P, Menis S, et al. Tailored immunogens direct affinity maturation toward HIV neutralizing antibodies. *Cell*. 2016;166(6):1459–70. doi: 10.1016/j.cell.2016.08.005
22. Rosa DS, Ribeiro SP, Cunha-Neto E. CD4+ T cell epitope discovery and rational vaccine design. *Archivum immunologiae et therapiae experimentalis*. 2010;58(2):121–30. doi: 10.1007/s00005-010-0067-0
23. Liu J, Zhang S, Tan S, Zheng B, Gao GF. Revival of the identification of cytotoxic T-lymphocyte epitopes for immunological diagnosis, therapy and vaccine development. *Experimental Biology and Medicine*. 2011;236(3):253–67.
24. UniProt: the universal protein knowledgebase. *Nucleic acids research*. 2017;45(D1):D158–D169. doi: 10.1093/nar/gkh131
25. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, et al. The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic acids research*. 2015;43(W1):W580–W584. doi: 10.1093/nar/gkv279
26. Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC bioinformatics*. 2007;8(1):4. doi:

10.1186/1471-2105-8-4

27. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols*. 2015;10(6):845.
28. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*. 1993; doi: 10.1038/nprot.2015.053
29. Colovos C, Yeates TO. ERRAT: an empirical atom-based method for validating protein structures. *Protein Sci*. 1993;2(9):1511–9.
30. Bugembe DL, Ekii AO, Ndembu N, Serwanga J, Kaleebu P, Pala P. Computational MHC-I epitope predictor identifies 95% of experimentally mapped HIV-1 clade A and D epitopes in a Ugandan cohort. *BMC Infectious Diseases*. 2020;20(1):1–16. doi: 10.1186/s12879-020-4876-4
31. Giguère S, Drouin A, Lacoste A, Marchand M, Corbeil J, Laviolette F. MHC-NP: predicting peptides naturally processed by the MHC. *Journal of immunological methods*. 2013;400:30–6. doi: 10.1016/j.jim.2013.10.003
32. Bui H-H, Sidney J, Li W, Fusseder N, Sette A. Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines. *BMC bioinformatics*. 2007;8(1):361. doi: 10.1186/1471-2105-8-361
33. Moutaftsi M, Peters B, Pasquetto V, Tschärke DC, Sidney J, Bui H-H, et al. A consensus epitope prediction approach identifies the breadth of murine T CD8⁺-cell responses to vaccinia virus. *Nature biotechnology*. 2006;24(7):817–9. doi: 10.1038/nbt1215
34. Dimitrov I, Flower DR, Doytchinova I. AllerTOP-a server for in silico prediction of allergens. In: *BMC bioinformatics*. 2013. p. S4. doi: 10.1186/1471-2105-14-S6-S4

35. Maupetit J, Derreumaux P, Tuffery P. PEP-FOLD: an online resource for de novo peptide structure prediction. *Nucleic acids research*. 2009;37(suppl_2):W498--W503. doi: 10.1093/nar/gkp323
36. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modeling. *electrophoresis*. 1997;18(15):2714–23. doi: 10.1002/elps.1150181505
37. Robinson J, Soormally AR, Hayhurst JD, Marsh SGE. The IPD-IMGT/HLA Database--New developments in reporting HLA variation. *Human immunology*. 2016;77(3):233–7. doi: 10.1016/j.humimm.2016.01.020
38. Eisenberg D, Lüthy R, Bowie JU. [20] VERIFY3D: assessment of protein models with three-dimensional profiles. In: *Methods in enzymology*. Elsevier; 1997. p. 396–404. doi: 10.1016/S0076-6879(97)77022-8
39. Dallakyan S. PyRx-python prescription v. 0.8. The Scripps Research Institute. 2008;2010.
40. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. *Journal of cheminformatics*. 2011;3(1):33. doi: 10.1186/1758-2946-3-33
41. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*. 2004; doi: 10.1002/jcc.20084
42. Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*. 2009; doi: 10.1002/jcc.21334
43. Emini EA, Hughes J V, Perlow D, Boger J. Induction of hepatitis A virus-neutralizing

- antibody by a virus-specific synthetic peptide. *Journal of virology*. 1985;55(3):836–9.
44. Kolaskar AS, Tongaonkar PC. A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS letters*. 1990;276(1–2):172–4.
 45. Karplus PA, Schulz GE. Prediction of chain flexibility in proteins. *Naturwissenschaften*. 1985;72(4):212–3.
 46. Larsen JEP, Lund O, Nielsen M. Improved method for predicting linear B-cell epitopes. *Immunome research*. 2006;2(1):2. doi: 10.1186/1745-7580-2-2
 47. Chou PY, Fasman GD. Empirical predictions of protein conformation. *Annual review of biochemistry*. 1978;47(1):251–76.
 48. Cao B, Wang Y, Wen D, Liu W, Wang J, Fan G, et al. A trial of lopinavir--ritonavir in adults hospitalized with severe Covid-19. *New England Journal of Medicine*. 2020; doi: 10.3410/f.737578995.793572690
 49. Poland GA, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Jacobson RM. Vaccinomics and a new paradigm for the development of preventive vaccines against viral infections. *Omics: a journal of integrative biology*. 2011;15(9):625–36. doi: 10.1089/omi.2011.0032
 50. Patel SK, Pathak M, Tiwari R, Yattoo MI, Malik YS, Sah R, et al. A vaccine is not too far for COVID-19.
 51. He Y, Zhou Y, Liu S, Kou Z, Li W, Farzan M, et al. Receptor-binding domain of SARS-CoV spike protein induces highly potent neutralizing antibodies: implication for developing subunit vaccine. *Biochemical and biophysical research communications*. 2004;324(2):773–81. doi: 10.1016/j.bbrc.2004.09.106
 52. Du L, Kou Z, Ma C, Tao X, Wang L, Zhao G, et al. A truncated receptor-binding domain

- of MERS-CoV spike protein potently inhibits MERS-CoV infection and induces strong neutralizing antibody responses: implication for developing therapeutics and vaccines. *PloS one*. 2013;8(12). doi: 10.1371/journal.pone.0081587
53. Chiou S-S, Fan Y-C, Crill WD, Chang R-Y, Chang G-JJ. Mutation analysis of the cross-reactive epitopes of Japanese encephalitis virus envelope glycoprotein. *Journal of general virology*. 2012;93(6):1185–92. doi: 10.1099/vir.0.040238-0
 54. Kallinich T, Beier KC, Wahn U, Stock P, Hamelmann E. T-cell co-stimulatory molecules: their role in allergic immune reactions. *European Respiratory Journal*. 2007;29(6):1246–55. doi:10.1183/09031936.00094306
 55. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Peter Walter P. *Molecular Biology of the Cell*, New York: Garland Science.[Google Scholar]. 2002;
 56. Ahmed S, Rakib A, Islam MA, Khanam BH, Faiz FB, Paul A, et al. In vivo and in vitro pharmacological activities of *Tacca integrifolia* rhizome and investigation of possible lead compounds against breast cancer through in silico approaches. *Clinical Phytoscience*. 2019; doi:10.1186/s40816-019-0127-x

12 Table legends

TABLE 1 The potential CD8⁺ T-cell epitopes along with their interacting MHC class I alleles and total processing score, epitopes conservancy hits and pMHC-I immunogenicity score

TABLE 2 Analysis of the population coverage for the proposed epitopes against SARS-Cov-2

TABLE 3 Binding affinities of the selected epitopes with HLA-B*15:25

TABLE 4 Combined B-cell linear epitope prediction

13 Figure legends

FIGURE 1 Workflow of the methodologies used in epitope-based vaccine design from SARS-CoV-2 Spike Glycoprotein.

FIGURE 2 (A) Three dimensional structure of model HLA-B*15:25 and evaluation of structure superiority by **(B)** Ramachandran plot analysis, **(C)** ERRAT and **(D)** VERIFY 3D assessment.

FIGURE 3 Molecular docking analysis of epitope WTAGAAYY with HLA-B*15:25 allele. The interacting residues were shown as ball and stick, conventional hydrogen bonds were shown as green line, pi-pi/pi-alkyl stacking were shown as pink lines, unfavorable bumps were shown as red lines.

FIGURE 4 Molecular docking analysis of epitope GAAYYVGY with HLA-B*15:25 allele. The interacting residues were shown as ball and stick, conventional hydrogen bonds were shown as green line, pi-pi/pi-alkyl stacking were shown as pink lines, carbon-hydrogen bonds were shown as white lines, attractive charges were shown as golden lines.

FIGURE 5 Combined B-cell linear epitope prediction showed the region from 803 to 816 amino acid residues had the highest antigenic propensity for B-cell linear epitopes. Surrounded by six differently coloured lines, which cover the region 800-820 amino acid residues in SARS-CoV-2 SGP, each line indicating different analysis methods with the maximum scores.

13.1 Supplementary data

Supplementary data 1 Multiple sequence alignment of SARS-CoV-2 spike glycoprotein.

Supplementary tables

TABLE S1 Antigenicity prediction of selected epitopes using VaxiJen 2.0 server

TABLE S2 MHC-NP probability score for the selected epitopes

Supplementary figures

FIGURE S1 Evolutionary divergence analysis of available spike glycoproteins of different strains of SARS-CoV-2; results are represented in a phylogenetic tree.

FIGURE S2 Ramachandran plot analysis for the tertiary structure of the SARS-CoV-2 SGP.

FIGURE S3 Population coverage based on MHC restriction data for **(A)** Central Africa, **(B)** Central America, **(C)** East Africa, **(D)** East Asia - using the Immune Epitope Database analysis resource.

FIGURE S4 Population coverage based on MHC restriction data for **(E)** Europe, **(F)** North Africa, **(G)** North America, **(H)** Northeast Asia - using the Immune Epitope Database analysis resource.

FIGURE S5 Population coverage based on MHC restriction data for **(I)** Oceania, **(J)** South Africa, **(K)** South America, **(L)** South Asia - using the Immune Epitope Database analysis resource.

FIGURE S6 Population coverage based on MHC restriction data for **(M)** Southeast Asia, **(N)** Southwest Asia, **(O)** West Africa, **(P)** West Indies - using the Immune Epitope Database analysis resource.

FIGURE S7 2D representation of the interaction between epitope WTAGAAYY and HLA-B*15:25 allele.

FIGURE S8 2D representation of the interaction between epitope GAAYYVGY and HLA-B*15:25 allele.

FIGURE S9 Combined B-cell linear epitope prediction using **(A)** Bepipred linear epitope prediction, **(B)** Chou & Fasman beta-turn prediction, **(C)** Emini surface accessibility prediction methods.

FIGURE S10 Combined B-cell linear epitope prediction using **(D)** Karplus & Schulz flexibility prediction, **(E)** Kolaskar & Tongaonkar antigenicity, **(F)** Parker hydrophilicity prediction methods.

TABLE 1 The potential CD8⁺ T-cell epitopes along with their interacting MHC class I alleles and total processing score, epitopes conservancy hits and pMHC-I immunogenicity score

Epitopes	NetCTL Combined score	Epitope_ Conservancy_Hit (MAX. Identity %)	MCH-I interaction with an affinity of IC50 < 200 and the total score (proteasome score, TAP score, MHC-I score, processing score)	pMHC-I immunogenicity score
WTAGAAAYY	3.1128	100	HLA-A*29:02 (1.51), HLA-A*26:01 (1.43), HLA-A*68:01 (1.12), HLA-C*12:03 (0.99), HLA-B*15:25 (0.97), HLA-B*35:01 (0.96), HLA-C*03:02 (0.91), HLA-A*30:02 (0.90), HLA-A*01:01 (0.89), HLA-B*15:01 (0.78), HLA-B*15:02 (0.68), HLA-C*16:01 (0.62), HLA-A*25:01 (0.56), HLA-C*02:09 (0.53), HLA-C*02:02 (0.53), HLA-C*12:02 (0.52), HLA-C*14:02 (0.24)	0.15259
CNDPFLGVY	1.3355	100	HLA-A*01:01 (0.36)	0.15232
GAAAYYVGY	1.2194	100	HLA-B*15:25 (1.03), HLA-A*29:02 (0.81), HLA-B*15:01 (0.69), HLA-A*30:02 (0.56), HLA-B*15:02 (0.39)	0.09963
ITDAVDCAL	1.1680	100	HLA-C*05:01 (0.62), HLA-C*08:02 (0.13), HLA-C*08:01 (0.08), HLA-C*03:04 (-0.41), HLA-C*03:03 (-0.41), HLA-C*16:01 (-0.43)	0.08501
STQDLFLPF	1.0468	100	HLA-B*15:25 (0.77), HLA-B*15:01 (0.60), HLA-B*15:02 (0.50), HLA-A*32:01 (0.48), HLA-C*16:01 (0.42), HLA-C*03:02 (0.35), HLA-B*35:01 (0.02), HLA-C*12:03 (-0.03),	0.06828

			HLA-A*29:02 (-0.11)	
TSNQVAVLY	3.0758	100	HLA-A*01:01 (0.76), HLA-A*29:02 (0.68), HLA-A*30:02 (0.54), HLA-B*58:01 (0.33)	-0.01327
KTSVDCTMY	2.3795	100	HLA-A*30:02 (1.04), HLA-B*58:01 (0.47)	-0.11115
MTSCCCLK	1.0963	100	HLA-A*68:01 (0.26), HLA-A*11:01 (-0.07), HLA-A*03:01 (-0.74), HLA-A*30:01 (-0.92), HLA-A*31:01 (-0.93), HLA-A*33:03 (-1.19)	-0.36816
STECNLLL	2.3492	100	HLA-C*05:01 (-0.27)	-0.20478
GAEHVNSY	1.9960	100	-	-0.00296

Notes: MHC-I alleles that have an interacting affinity lower than 200 nm are represented, and total processing scores are shown as enclosed numbers.

TABLE 2 Analysis of the population coverage for the proposed epitopes against SARS-Cov-2

Population/Area	Coverage(%) ^a	Average hit ^b	PC90 ^c
Central Africa	63.97	1.56	0.28
Central America	2.19	0.04	0.10
East Africa	60.78	1.40	0.25
East Asia	62.39	1.49	0.27
Europe	77.91	1.86	0.45
North Africa	72.00	1.77	0.36
North America	65.50	1.59	0.29
Northeast Asia	56.06	1.54	0.23
Oceania	35.97	0.67	0.16
South Africa	70.08	1.71	0.33
South America	46.65	0.91	0.19
South Asia	69.11	1.53	0.32
Southeast Asia	44.14	1.08	0.18
Southwest Asia	60.13	1.14	0.25
West Africa	77.98	2.17	0.45
West Indies	56.56	1.22	0.23

Notes: ^a Projected population coverage.

^b Average number of epitope hits/HLA combinations recognized by the population.

^c Minimum number of epitope hits/HLA combinations recognized by 90% of the population.

TABLE 3 Binding affinities of the selected epitopes with HLA-B*15:25

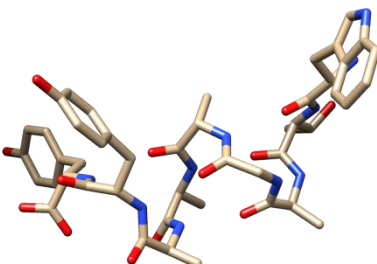
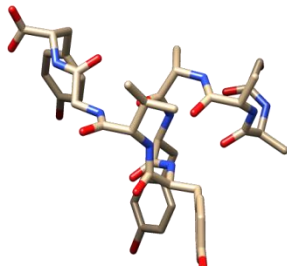
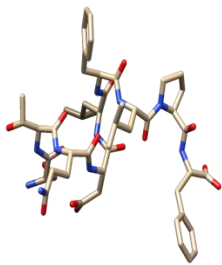
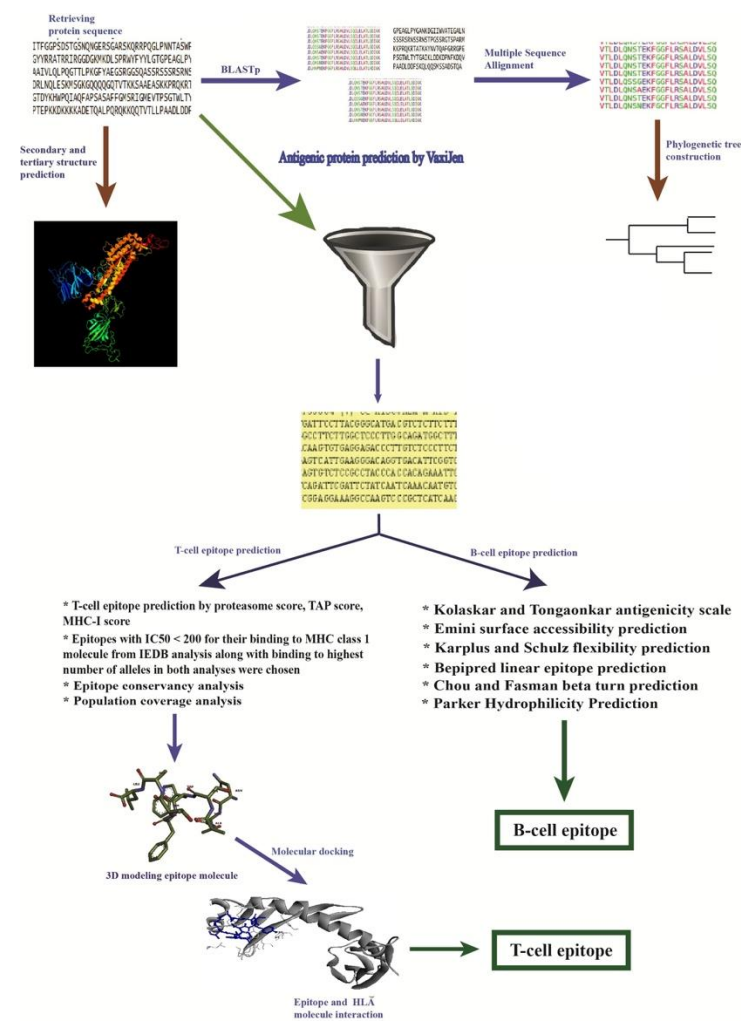
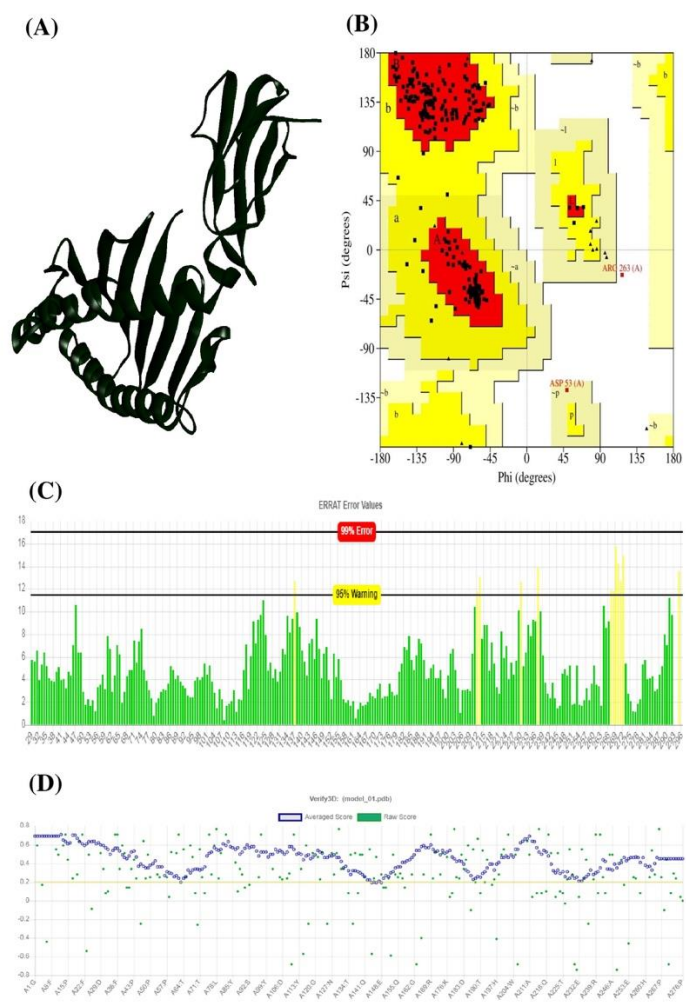
Epitope	Three dimensional structure of the epitope	Binding affinity (kcal/mol)
WTAGAAAYY		-8.5
GAAAYYVGY		-8.8
STQDLFLPF		-7.8

TABLE 4 Combined B-cell linear epitope prediction

Method	Region	Residues	Length	Score		
				Max.	Avg.	Min.
Bepipred Linear Epitope Prediction	805-816	ILPDPSKPSKRS	12	2.291	-0.066	-0.001
Chou & Fasman Beta-Turn Prediction	807-813	PDPSKPS	7	1.484	0.997	0.541
Emini Surface Accessibility Prediction	810-815	SKPSKR	6	6.051	1.00	0.042
Karplus & Schulz Flexibility Prediction	809-815	PSKPSKR	7	1.125	0.993	0.876
Kolaskar & Tongaonkar Antigenicity	803-808	SQILPD	6	1.261	1.041	0.866
Parker Hydrophilicity Prediction	808-814	DPSKPSK	7	7.743	1.238	-7.629



1 **FIGURE 1** Workflow of the methodologies used in epitope-based vaccine design from SARS-CoV-2 Spike Glycoprotein.



2

3 **FIGURE 2** (A) Three dimensional structure of model HLA-B*15:25 and evaluation of structure superiority by (B) Ramachandran
4 plot analysis, (C) ERRAT and (D) VERIFY 3D assessment.

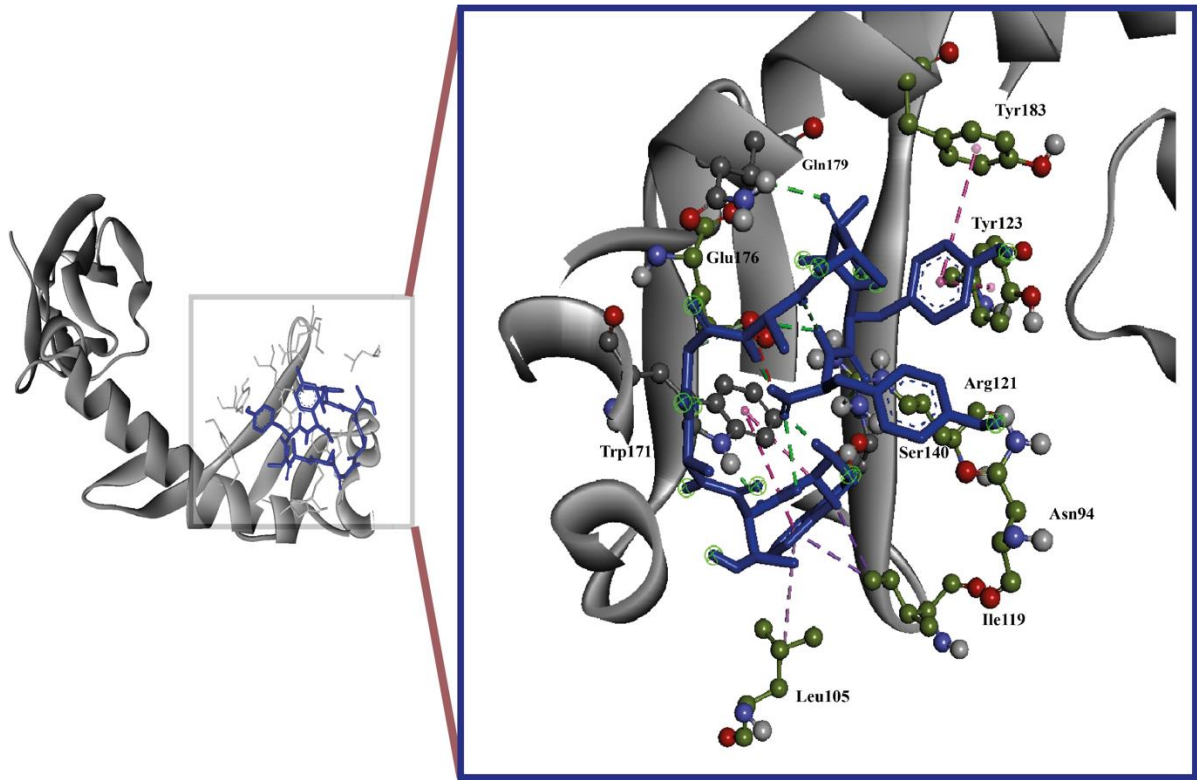
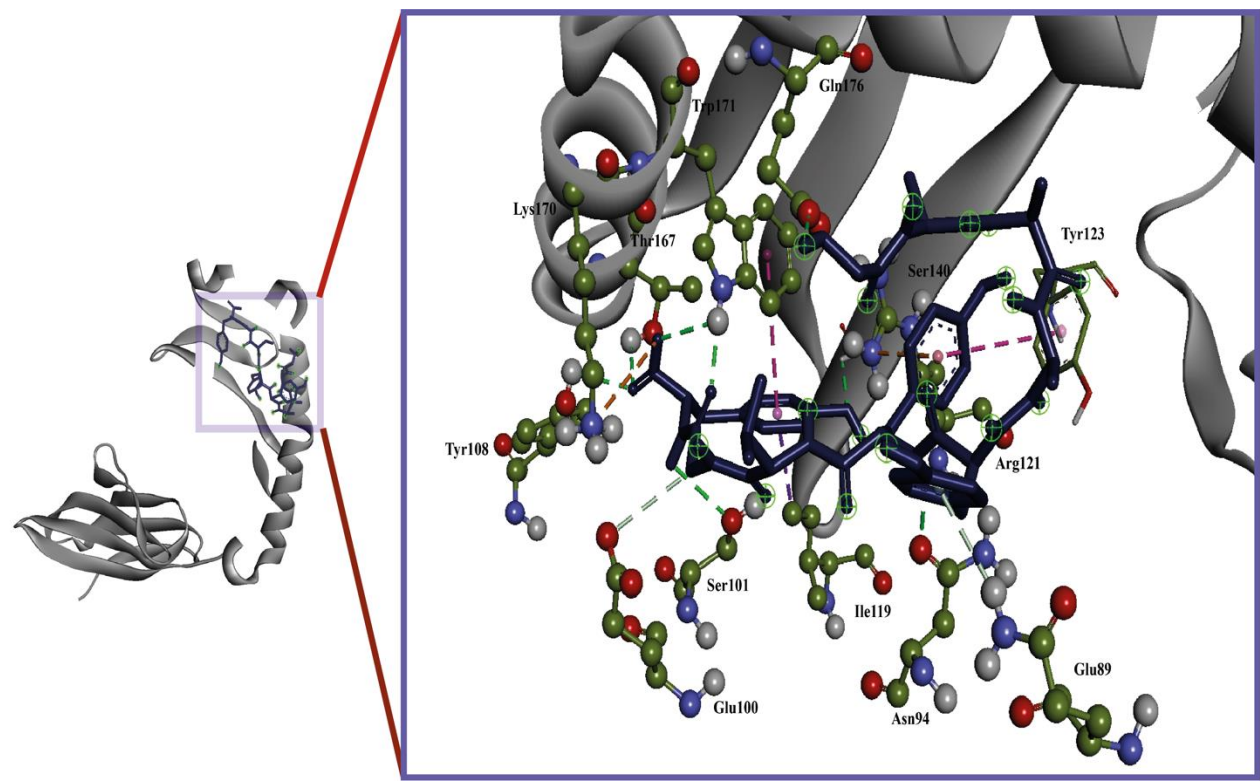
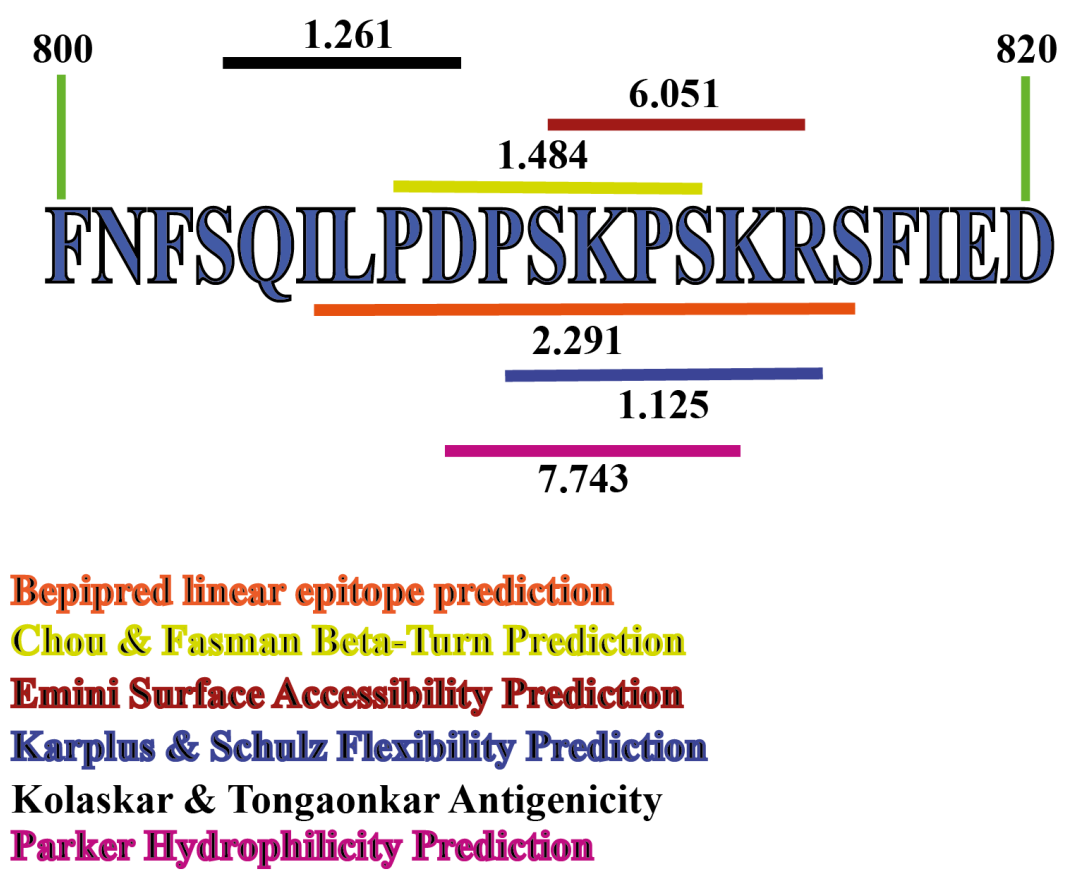


FIGURE 3 Molecular docking analysis of epitope WTAGAAAYY with HLA-B*15:25 allele. The interacting residues were shown as ball and stick, conventional hydrogen bonds were shown as green line, pi-pi/pi-alkyl stacking were shown as pink lines, unfavorable bumps were shown as red lines.



10 **FIGURE 4** Molecular docking analysis of epitope GAAAYVVG with HLA-B*15:25 allele. The interacting residues were shown as
11 ball and stick, conventional hydrogen bonds were shown as green line, pi-pi/pi-alkyl stacking were shown as pink lines, carbon-
12 hydrogen bonds were shown as white lines, attractive charges were shown as golden lines.



13

14 **FIGURE 5** Combined B-cell linear epitope prediction showed the region from 803 to 816 amino acid residues had the
15 highest antigenic propensity for B-cell linear epitopes. Surrounded by six differently coloured lines, which cover the region
16 800-820 amino acid residues in SARS-CoV-2 SGP, each line indicating different analysis methods with the maximum
17 scores.