Medicinal Chemistry & Drug Discovery

Identification of a Potential mRNA-based Vaccine Candidate against the SARS-CoV-2 Spike Glycoprotein: A Reverse Vaccinology Approach

Olanrewaju Ayodeji Durojaye+,*[a, b, d] Divine Mensah Sedzro+,[a, b] Mukhtar Oluwaseun Idris,*[b] Abeeb Abiodun Yekeen,*[b] Adeola Abraham Fadahunsi,[c] and Oluwaseun Suleiman Alakanse[b, e]

The emergence of the novel coronavirus (SARS-CoV-2) in December 2019 has generated a devastating global consequence which makes the development of a rapidly deployable, effective and safe vaccine candidate an imminent global health priority. The design of most vaccine candidates has been directed at the induction of antibody responses against the trimeric spike glycoprotein of SARS-CoV-2, a class I fusion protein that aids ACE2 (angiotensin-converting enzyme 2) receptor binding. A variety of formulations and vaccinology approaches are being pursued for targeting the spike glycoprotein, including simian and human replication-defective adenoviral vaccines, subunit protein vaccines, nucleic acid vaccines and whole-inactivated SARS-CoV-2. Here, we directed a reverse vaccinology approach towards the design of a nucleic acid (mRNA-based) vaccine candidate. The "YLQPRTFLL" peptide sequence (position 269–277) which was predicted to be a B cell epitope and likewise a strong binder of the HLA*A-0201 was selected for the design of the vaccine candidate, having satisfied series of antigenicity assessments. Through the codon optimization protocol, the nucleotide sequence for the vaccine candidate design was generated and targeted at the human toll-like receptor 7 (TLR7). Bioinformatics analyses showed that the sequence "UACCUGCAGCCGCGUACCUUCCUGCUG" exhibited a strong affinity and likewise was bound to a stable cavity in the TLR7 pocket. This study is therefore expected to contribute to the research efforts directed at securing definitive preventive measures against the SARS-CoV-2 infection.

Introduction

The 21st century since inception has experienced three coronavirus types (severe acute respiratory syndrome coronavirus, Middle-East respiratory syndrome coronavirus, and severe acute respiratory syndrome coronavirus 2) that have crossed the species barrier to cause fatal human pneumonia.[1,2] The Guangdong province of China was the first to experience the emergence of SARS-CoV in 2002 which later spread through air travel to five other continents, infected 8,098 people and led to the death of 774.[3] The Arabian Peninsula in 2012 experienced the emergence of MERS-CoV where it remains a major concern for public health, claiming 858 lives with a total of 2,494 infections. SARS-CoV-2, a previously unknown coronavirus, in December 2019 was discovered in the Hubei province of China. Isolation and sequencing of the virus were completed by January 2020.[4,5]

About 77% amino acid sequence identity has been observed between the spike glycoprotein of the SARS-CoV and that of the SARS-CoV-2, while the MERS-CoV and SARS-CoV-2 share only an identity of 31%, suggesting a more distant relationship between both coronaviruses. An even much lower sequence identity exists between the spike glycoprotein of the common cold virus and that of the SARS-CoV-2 with a 25% to 30% variation. Phylogenetic analyses confirm the existence of an evolutionarily closer relationship between the SARS-CoV

[a] O. A. Durojaye,+ D. M. Sedzro+
MOE Key Laboratory of Membraneless Organelle and Cellular Dynamics,
Hefei National Laboratory for Physical Sciences at the Microscale,
University of Science and Technology of China,
Hefei, Anhui 230027, China
E-mail: lanredurojaye@mail.ustc.edu.cn

[b] O. A. Durojaye,+ D. M. Sedzro,+ M. O. Idris, A. A. Yekeen, O. S. Alakanse
School of Life Sciences,
University of Science and Technology of China,
Hefei, Anhui 230027, China
E-mail: idrisolaitan2009@gmail.com
yekeenaa@mail.ustc.edu.cn

[c] A. A. Fadahunsi
Department of Biomedical Engineering,
University of Science and Technology of China,
Hefei, Anhui 230027, China

[d] O. A. Durojaye+
Department of Chemical Sciences,
Coal City University, Emene,
Enugu State, Nigeria

[e] O. S. Alakanse
Department of Biochemistry,
Faculty of Life Sciences, University of Ilorin,
Ilorin, Kwara State, Nigeria

[+] OAD and DMS contributed equally to this work.

Supporting information for this article is available on the WWW under https://doi.org/10.1002/slct.202103903

and SARS-CoV-2 than other human coronaviruses.[6] The receptor-binding domains show different degrees of sequence identity, which ranges from 13 % between the HCoV-NL63 and SARS-CoV-2 to 74 % between the SARS-CoV and SARS-CoV-2. Nevertheless, there is a similarity between the 3D structure of their spike glycoprotein trimer ectodomains and the similarity extends to the 3D structure of other coronavirus spike glycoprotein, including the discovery of flexible receptor binding domains that can be in different "up" conformations or in the "down" of the closed pre-fusion trimer.[7] The differences in the amino acid sequence of the SARS-CoV-2 receptor binding domain and that of the SARS-CoV compared with the MERS-CoV result in the differential binding to host receptors (dipeptidyl peptidase 4 for MERS-CoV and angiotensin-converting enzyme 2 (ACE2) for SARS-CoV and SARS-CoV-2).[8,9] HCoV-NL63 also uses its receptor-binding domain to bind the angiotensin-converting enzyme 2, although there is a structural difference when compared to the SARS-CoV and SARS-CoV-2 interaction,[9] whereas HCoV-HKU1 and HCoV-OC43 use their receptor binding domain to bind the 9-*O*-acetylated sialic acids of the host.[10]

The emergence of nucleic acid therapeutics has provided a promising alternative to the conventional approaches in vaccine design. In 1990, the first successful animal usage of the IVT (*in vitro* transcribed) mRNA was published by Wolff *et al.*[11] In the study, protein production was observed upon injection of a reporter gene mRNA into mice.[11] In 1992, another study demonstrated that vasopressin-encoding mRNA administration in the hypothalamus of rats could trigger a physiological response.[12] However, these early promising results, because of concerns associated with *in vivo* delivery inefficiency, high innate immunogenicity and instability of the mRNA, could not attract substantial investment towards mRNA vaccine development, and the field for these reasons rather pursued protein-based and DNA-based therapeutics.[13] In the past decade, major research investments and technological innovations in the fields of protein replacement therapy and vaccine development have enabled the mRNA to become a potential and promising therapeutic tool. The mRNA-based vaccine usage has several benefits over live attenuated and subunit virus, likewise the DNA-based vaccines. With an emphasis on its safety, the mRNA is a non-integrating and non-infectious platform, which poses no potential insertional mutagenesis or infection risk. In addition, the degradation of mRNA is through regular cellular processes, and its half-life can be regulated *in vivo* via the usage of various delivery and modification methods.[13] To further improve the mRNA safety profile, the inherent immunogenicity can be down-modulated.[13,14] As regarding efficacy, several modifications make mRNA highly translatable and more stable.[14] *In vivo* delivery efficiency of the mRNA can be achieved through its formulation into carrier molecules, which allows quick uptake and cytoplasmic expression. Since mRNA is the minimal genetic vector, repeated administration of the mRNA-based vaccines can be possible, as anti-vector immunity is completely avoided.[14] In respect to production, because of high yields in *in vitro* transcription reactions, mRNA-based vaccines have the potential for scalable, inexpensive and rapid manufacturing.[15]

As the spike glycoprotein of the coronavirus is exposed on its surface and facilitates host cell entry, it is the major focus of vaccine and therapeutic design, likewise the main target of the Abs (neutralizing antibodies) upon infection. Toll-like receptors are membrane receptors that play an important role in the innate immune system.[16] They are structurally characterized type I membrane glycoprotein to house extracellular LLR (leucine-rich repeat) domain, a cytoplasmic TIR (Toll/interleukin-1 receptor) and a single transmembrane domain.[17] The toll-like receptors function for molecular patterns associated with danger and as a sensor for limited number of molecular patterns that are associated with pathogens. Once the toll-like receptors recognize specific molecules through their leucine-rich repeat domains, such as ssRNA (single-stranded RNA) for TLR7, 8 and 13,[18] and CpG-containing DNA for TLR9,[19] the cytoplasmic toll-interleukin receptor (TIR) domains recruit downstream toll-interleukin receptor domain-containing adaptor molecules such as MyD88 (myeloid differentiation factor 88) and TRIF (TIR domain-containing adaptor inducing interferon-b).[20] The initiated signal-transduction pathways lead to type I interferons and proinflammatory cytokine production, which mobilizes the host immune responses.[21]

There has been a continuous report of mutations in the gene encoding the SARS-CoV-2 spike glycoprotein, even though the virus was just recently discovered in humans.[22] Different naturally existing variants of the spike glycoprotein have recently been reported. These emerging variants and the rapid global spread have raised concerns regarding the possibility of reduced COVID-19 vaccine protection.[23] The emergence of notable variants that harbor series of spike glycoprotein mutations have been reported in Brazil (P.1), South Africa (B.1.351/501Y.V2), United Kingdom (B.1.1.7) and most recently in India (B.1.617). The United Kingdom variant, which is currently the most globally spread variant of concern with an associated increased transmissibility, possess a N-terminal domain H69/V70 deletion, an adjacent P681H mutation to spike glycoprotein furin cleavage site, and a receptor-binding domain N501Y substitution. The South African variant also harbors different mutations which include the K417 N, E484 K, and N501Y substitution. The variation between the South African and the Brazilian variant can be observed in the K417 mutation, which unlike the observed "N" substitution in the South African variant, is a "T" substitution in the Brazilian variant. Both variants possess the same E484 K, and N501Y substitution. Finally, the recent Indian variant has been reported to possess the L452R, E484Q and T478 K substitution. However, these variants of concern all share the D614G substitution which has been reported to confer on the virus an increased ability to spread rapidly.[23] In this study, we designed from an antigenic region of the SARS-CoV-2 spike glycoprotein a potential mRNA-based vaccine candidate which is capable of triggering immune response through its interaction with the human toll-like receptor 7. In addition, we analyzed the effect of the N501Y mutation on the stability dynamics of the various SARS-CoV-2 variants possessing it, with the aim of predicting the occurrence of possible stability-linked mutations at the selected antigenic epitope of the viral spike glycoprotein.

Materials and methods

Sequence and structural data retrieval

The amino acid sequence and 3-dimensional structures of the SARS-CoV-2 spike glycoprotein, the human major histocompatibility complex (class I) and the toll-like receptor (TLR7) were retrieved from the National Center for Biotechnology Information (NCBI) database[24] and the Protein Data Bank (PDB)[25] respectively. The SARS-CoV-2 primary sequence was retrieved in FASTA format with an accession number of P0DTC2 while the 3D structures were downloaded with codes 4U6Y, 6VXX and 7CYN, representing the crystal structure of the HLA*A-0201 and the Cryo-EM structures of the SARS-CoV-2 spike glycoprotein and the human TLR7 respectively.

T cell epitope prediction and processing

Prediction of the T cell epitopes was based on the identification of major histocompatibility complex class I (MHC–I)- binding molecules. The significance of binding of each peptide to a given MHC–I molecule is on the premise of the estimated binding strength exhibited by the predicted nested core peptide at a set threshold level. NetMHC 4.0 predicted the binding of epitope peptides to the HLA*A-0201 allele using artificial neural networks.[26] We repeated the same analysis using the Immune Epitope Database (IEDB) MHC–I epitope prediction module, selecting the IEDB recommended default prediction method. In this case, consideration is given to all alleles and their corresponding length of peptide, for a specific species. For each combination of allele length, the consensus method is used, which includes SMM (stabilized matrix method), Comblib (derived scoring matrices from combinatorial peptide libraries) and ANN (artificial neural networks).[27] The selection of epitopes presented by the MHC (major histocompatibility complex) class I molecules follows a multi-step process. The IEDB MHC–I processing platform presents a computer-based prediction of this process based on *in vitro* experiments. The MHC–I processing platform characterizes cleavage by proteasomes, transport by TAP (transporter associated with antigen processing) and MHC class I binding.[28] Finally, a multi-step algorithm "EpiJen" was used for T cell epitope prediction and processing. The application of the method is directed at a set of overlapping peptides that are derived from the sequence of a whole protein and acts as a filter that reduces successfully the number of potential epitopes. The set of final peptides needed for epitope testing rarely includes more than 5% of the whole sequence. Quantitative matrices for each step were developed using the additive method.[29] The same method was also used in the generation of quantitative matrices for proteasome cleavage and TAP binding.

Epitope promiscuity prediction

The predicted top binder of the T cell was analyzed for its potential to bind a wide range of alleles in the class I of the major histocompatibility complex. The ProPred1 tool was utilized to achieve this objective. Propred is a tool for predicting the binding of peptides to alleles of MHC–I. It is a matrix-based method which allows the prediction of binding sites for the MHC in an antigenic sequence that contains 47 alleles of the MHC class I. The matrices employed in the ProPred1 have been derived from literature and the tool also allows standard immunoproteasome and proteasome cleavage site prediction in an antigenic sequence. Matrices described by Toes *et al.*, 2001 have been implemented by the server to identify cleavage sites for proteasomes in an antigenic sequence. Filtering of MHC class I binders with cleavage site at the C terminus is also allowed.[30]

B cell epitope prediction

cell epitopes play a crucial role in disease diagnosis, allergy research and in the development of vaccines. The B cell epitope potential of the predicted top binder of the T cell was determined using the BepiPred-2.0, ABCpred and BcePred tools. BepiPred-2.0 is trained based on epitope data that has been derived from crystal structures, with a presumed higher quality and indeed a significantly improved predictive power.[31] ABCpred implements the FNN (standard feed-forward neural network) and RNN (recurrent neural network) for the prediction of B cell epitopes in an antigenic sequence. The networks have been tested and trained on a clean data set, which is composed of 700 non-redundant B cell epitopes that were derived from the Bcipep database and likewise 700 non-epitopes derived randomly from the Swiss-Prot database.[32] The approach used by Bcepred for the prediction of B cell epitopes is based on physiochemical properties (mobility/flexibility, hydrophilicity, turns, exposed surface, polarity and accessibility). Results are presented in a tabular and graphical frame. In the graphical frame, residue properties are plotted along protein backbone, which facilitates rapid visualization of B cell epitopes on protein sequence.[33]

Antigenicity assessment of predicted epitope

The antigenicity evaluation took into consideration properties such as allergenicity, transmembrane topology and the N-glycosylation propensity of the predicted epitope. Prediction of the selected epitope as a protective antigen was conducted using VaxiJen.[34] The transmembrane topology of the spike glycoprotein was predicted using the TMHMM v2.0 and TOPCONS, while N-glycosylation site prediction was achieved using the NetNGlyc 1.0, which predicts N-Glycosylation sites in proteins using artificial neural networks that examine the sequence context of Asn-Xaa-Ser/Thr sequons.[35] TMHMM is a transmembrane helices prediction method which is anchored on a hidden Markov model and the development was by Erik Sonnhammer and Anders Krogh.[36] TOPCONS is a membrane protein topology consensus prediction server. The algorithm is a combination of arbitrary predictions of topology which are merged into a consensus prediction, while its prediction reliability is quantified on the basis of agreement level between the underlying methods both on protein and transmembrane region level.[37]

Peptide preparation and simulation

The epitope peptide was designed using the "build structure" function of the Chimera software.[38] The "minimize structure" function was used in the geometry optimization protocol at steepest descent and conjugate gradient steps of 100 and 10 respectively, while size remained fixed at 0.02 angstroms for both steps. Modeling of the structural flexibility of the peptide was achieved with the use of the CABS-flex tool which generates protein dynamics at highly reduced (3 orders of magnitude) computational cost.[39] This follows the Jamroz *et al.*[40] work where the authors were able to demonstrate that the consensus view of the near-native protein which was obtained from a molecular dynamics simulation production time of 10 nanoseconds, is consistent with the CABS dynamics. Protein residue fluctuations obtained from CABS-flex have also been demonstrated to be well correlated to those generated from NMR ensembles.[40]

Peptide docking against HLA*A-0201

Molecular docking of the candidate peptide antigenic epitope models against the HLA*A-0201 allele was conducted with the aid of the HPEPDOCK[41] and validated with the ClusPro[42] server. In the HPEPDOCK docking algorithm, flexibility of the peptide is considered through the generation of an ensemble of peptide conformations, using the MODPEP program.[43] The sampled peptide conformations are then docked globally against the whole protein using the MDock docking protocol.[44]

Three computational steps are performed by the ClusPro, which include sampling of billions of conformations through rigid docking, RMSD (root-mean-square deviation)-based clustering of the 1000 structures with the lowest energy, which are generated to detect the largest clusters that will represent the closest models of the complex, and the energy minimization refinement of selected structures. Docking of the rigid body by ClusPro involves the use of PIPER, a docking program based on the fast Fourier transform (FFT) correlation approach. The fast Fourier transform approach which was implemented by Katchalski-Katzir *et al.*[45] has introduced a major progress in the protein-protein docking of rigid body. In the method, a protein (the receptor) is stationed on a fixed grid at the origin of the coordinate system and another protein (the ligand) is stationed on a moveable grid while the energy of interaction is written as a correlation function. The numerical efficiency is supported by the fact that such energy functions can be calculated efficiently and this results in the ability to exhaustively sample many conformations of protein-protein interaction, likewise evaluating grid point energies. Thus, fast Fourier transform-based algorithm allows the docking of proteins without prior information of their structures.[45]

Conformational stability of the protein-peptide complex

The conformational behaviors of the HLA*A-0201 in its free form and when in complex with the top-ranked peptide antigenic epitope were determined using molecular dynamics simulations. The HLA*A-0201 binding potency of the top-ranked peptide was assessed by using the HPEPDOCK-derived HLA*A-0201-peptide complex, as a starting structure for the conformational behavior analysis. The peptide binds the HLA*A-0201 binding groove in a position close to the A, D and F-pocket of the HPEPDOCK complex. The molecular dynamics simulations were performed on the free HLA*A-0201 and HLA*A-0201 peptide complex using GROMACS 2019 with the Charmm36 force field.[46] The systems were solvated in a cubic box using the simple point charge, SPC water molecules and were neutralized with appropriate sodium and chloride ions. The systems were energy minimized using 50000 steps steepest descent method with a maximum force less than 1000 KJ/mol and was equilibrated at 300 K NVT and 1 bar NPT. The particle mesh ewald method was used for the electrostatic interaction calculations using 1.2 nm cutoff.[47] Then production run of 100 ns was done on the prepared free and peptide-bound HLA*A-0201. Equilibrium properties such as the Root-Mean-Square-Deviation (RMSD), Root-Mean-Square-Fluctuation (RMSF), Radius of gyration (Rg), intramolecular hydrogen bond, Solvent Accessibility Surface Area (SASA) were used to assess the conformational stability of the HLA*A-0201 in its peptide-bounded and free states. Principal component analysis was employed to understand the collective motions of atoms in the peptide-bound and free forms of HLA*A-0201, using the *gmx covar* and *gmx anaeig* tools. Vmd, Ligplot, PyMOL and Xmgrace were used for visualization and image preparation.[48]

Codon optimization

Adaptation of the peptide epitope gene for optimum expression in *E. coli* was achieved using the JCat (Java Codon Adaptation Tool)[49] and ExpOptimizer tools[50]. The calculations of JCat are made in advance using a proposed algorithm by Carbone *et al.*[51]. Calculation results are stored in the PRODORIC database, which hosts the data of most freely available sequenced prokaryotic genomes.[51] ExpOptimizer on the other hand is developed to highly express any protein of interest in any mainstream expression host. The codon optimization algorithm gives considerations to crucial gene transcription and translation factors.[50]

Mutagenicity study

The effect of mutation on the stability of the SARS-CoV-2 spike glycoprotein was predicted using the DynaMut[52] and I–Mutant2.0.[53] DynaMut implements two well established and distinct normal mode approaches, which can be used for the analysis and visualization of protein dynamics through conformational sampling and the assessment of mutational impact on protein stability and dynamics resulting from changes in vibrational entropy. DynaMut along with the normal mode dynamics also integrates graph-based signatures to generate a consensus prediction of mutational impact on the stability of a protein.[52] I–Mutant2.0 is a SVM (support vector machine)-based tool for the prediction of changes in protein stability upon single point mutations. The predictions of I–Mutant2.0 are performed either from the protein sequence or, less importantly, the protein structure. The method was tested and trained on ProTherm-derived data set which at the moment is the most comprehensive database of thermodynamic experimental data of changes in free energy of protein stability under different conditions upon mutation. I–Mutant2.0 can be used as an estimator of regression for the prediction of related values of DeltaDeltaG and as a classifier for the prediction of sign of protein stability change upon mutation.[53]

mRNA preparation and simulation

The codon-optimized sequence was used for the design of the mRNA vaccine candidate. Energy minimization of the mRNA was conducted using the previously described steepest descent and conjugate gradient steps parameters and the native form was simulated using SimRNA.[54] SimRNA is a recent method for the computational prediction of RNA 3D structures. It uses a coarse-grained representation, samples the conformational space using the Monte Carlo method, and employs a statistical potential for energy approximation and the identification of conformations which correspond to biologically relevant structures. For complex 3D structure modeling, it uses derived additional restraints from computational or experimental analyses, including long range contacts and/or secondary structure information. The SimRNA also analyzes conformational landscapes for the identification of potential alternative structures.[54]

mRNA model docking and binding pocket dynamics analysis

The top five mRNA models generated through the SimRNA simulation protocol were docked against the human toll-like receptor 7, using the HDOCK docking tool.[55] The HDOCK is an integrated package with multiple components that include several third party programs. The first of the HDOCK workflow steps is the input of data that accepts protein structures. The second step involves the search for sequence similarity which is conducted against the PDB sequence in order to locate the homologous sequences for both ligand and receptor molecules. It moves on to the third step by conducting a comparison between two template sets to know if they have similar records with the same PDB codes. If such PDB codes exist, a common template for both ligand and receptor will be selected. In a case where there exists no overlap between two homologous template sets, the templates will be selected for the ligand protein and/or the receptor protein from two template sets, respectively.[55] Analysis of the TLR7 binding pocket dynamics was conducted using the D3Pockets server.[56] The development of this tool was targeted at the analysis and detection of ligand binding pocket dynamic properties for a target protein. In addition to its ability to detect all potential ligand-binding pockets on the surface of a protein based on a PDB file, it can also analyze the pocket dynamic properties through correlation, stability and continuity, based on a conformation ensemble or a molecular dynamics trajectory. Results obtained from this tool can be used for the design of ligands on a novel binding pocket and for studying the functional mechanism of a target protein.[56]

Results

T cell epitope prediction and processing

The names of each column are displayed in the first row of the table, as adapted from the original NetMHC output. "pos" in the first column is the position of the first amino acid of the predicted peptide within the sequence of the epitope. "HLA" and "peptide" indicate the columns for the target MHC class I allele, and the peptide primary sequence respectively. The "log score" column is the raw prediction output, which for artificial neural networks is 1-log50k to the affinity in nanomolar units. An additional column is included for the artificial neural network predictions (Affinity), which depicts the predicted affinity in nanomolar units. "BindLevel" is the column that specifies if the predicted peptide binding affinity is stronger than a specific threshold. For the artificial neural network prediction, the affinity score of a strong binder (SB) is less than 50 nM and the corresponding %Rank is less than 0.5 while the affinity score of a weak binder (WB) is less than 500 nM and the corresponding %Rank is less than 2

The threshold for strong binders was set at 0.5 while the weak binders threshold was set at 2. The peptide length was set at 9 and sorted by predicted affinity. The displayed output shows the prediction for the top five strong binders and the strongest predicted binding peptide being YLQPRTFLL with a predicted affinity of 5.36 nM; which is well below the threshold (50 nM) for a peptide expected to have a strong affinity, and a %Rank of 0.04; which is also below the threshold (Table 1). In addition, selection of candidate peptide based on %Rank is a better criterion as this parameter is not influenced by possible inherent molecular biases towards binding affinities.

The IEDB output for MHC class I binders shows prediction results from various predictors that are transformed first into percentile scores, in order to allow a uniform scale comparison across the predictors. For a given predictor and a peptide, a percentile score is termed as the percentage of randomly sampled peptides from naturally occurring proteins with a better score than the peptide. The final predicted peptide binding affinity score, using the consensus approach is a median percentile scores from various predictors. The YLQPRTFLL peptide in the IEDB prediction also appears as the strongest binder (Table 2).

For a peptide to be recognized by the immune system, there are additional steps the peptide has to pass in the MHC class I pathway. These steps include the cleavage by protea-

Table 1. NetMHC 4.0 prediction output for the potential MHC–I binders from the full length amino acid sequence of the SARS-CoV-2 spike glycoprotein.

pos	HLA	peptide	1-log50k	Affinity (nM)	%Rank	BindLevel
268	HLA–A*0201	YLQPRTFLL	0.845	5.36	0.04	SB
132	HLA–A*0201	FQFCNDPFL	0.795	9.18	0.10	SB
1219	HLA–A*0201	FIAGLIAIV	0.785	10.29	0.12	SB
690	HLA–A*0201	SIIAYTMSL	0.759	13.54	0.17	SB
999	HLA–A*0201	RLQSLQTYV	0.740	16.66	0.25	SB

Table 2. IEDB prediction output for the top five potential MHC–I (HLA–A*-0201) strong binders.

Allele	Length	Peptide	Score	Percentile Rank
HLA–A*0201	9	YLQPRTFLL	0.971198	0.02
HLA–A*0201	9	VLNDILSRL	0.938498	0.03
HLA–A*0201	9	TLDSKTQSL	0.914998	0.03
HLA–A*0201	9	RLQSLQTYV	0.87376	0.05
HLA–A*0201	9	RLDKVEAEV	0.825045	0.06

somes and transportation by the transporter associated with antigen processing (TAP), which have been used for the identification of T cell epitopes in combination with MHC class I binding predictions. The ratings of the top five 9mer peptides predicted to be strong binders of the MHC class-I with a high probability for proteasomal cleavage and transportation by TAP are depicted in the IEDB processing (Table 3) and EpiJen (Table 4) outputs.

To delineate the EpiJen dataflow, the protein initially is cut into overlapping decamers and processed by the quantitative matrices (QM) for proteasome cleavage. The model only takes into account contributions of residues next to the C-terminus (cleavage site) and the amino acid residues that follow. This first step has a high filtering potential, which leads to the elimination of half or two-third of the generated true negatives. The cleaved peptides which are presented as nanomers are passed on to the TAP binding quantitative matrices, which is the next filter in the dataflow. For both partially and fully TAP-dependent alleles, a 5.00 threshold is recommended. Although the TAP step filtering ability is slow, about ten percent of the true negatives are still eliminated in this step. The transported peptides are then moved to the MHC binding process, which is the next filter. EpiJen included eighteen quantitative matrices with predictive potentials for different HLA allele binding. Continuous values like IC_{50}s, which are quantitative data were available for the allele of interest, which in this case is the HLA–A*0201, but for some other alleles, only discontinuous values (sequence of binders) are known. The ability of this step towards filtering is highly significant, leading to the elimination of approximately twenty-five to thirty percent of the true negatives. A default threshold of 0.5 and 5.3 are set for the discriminant analysis model and the multiple linear regression model respectively. These thresholds seek to limit the number of false positive outputs in proteins with long sequences. The final EpiJen ranking as shown in Table 4 is based on the IC_{50} and corresponding IC_{50} negative logarithm of the generated peptides, with the YLQPRTFLL peptide displaying the highest T cell epitope processing potential.

Table 3. IEDB output for the top five T cell epitope processing peptides from the SARS-CoV-2 spike glycoprotein.

Allele	Start	End	Peptide Length	Peptide	Proteasome Score	TAP Score	MHC Score	Processing Score	Total Score	MHC IC_{50} [nM]
HLA–A*0201	269	277	9	YLQPRTFLL	1.45	0.39	−0.66	1.84	1.17	4.6
HLA–A*0201	417	425	9	KIADYNYKL	1.68	0.51	−1.20	2.19	0.99	15.9
HLA–A*0201	133	141	9	FQFCNDPFL	1.55	0.38	−0.95	1.94	0.99	8.9
HLA–A*0201	691	699	9	SIIAYTMSL	1.50	0.50	−1.18	2.00	0.82	15.3
HLA–A*0201	976	984	9	VLNDILSRL	0.43	0.43	−1.29	1.72	0.43	19.7

Table 4. EpiJen output for the top five T cell epitope processing peptides from the SARS-CoV-2 spike glycoprotein.

Starting position	Peptide	Predicted -logIC_{50} (M)	Predicted IC_{50} Value (nM)
269	YLQPRTFLL	10.034	0.09
28	YTNSFTRGV	9.977	0.11
983	RLDKVEAEV	9.779	0.17
1220	FIAGLIAIV	9.765	0.17
857	GLTVLPPLL	9.709	0.20

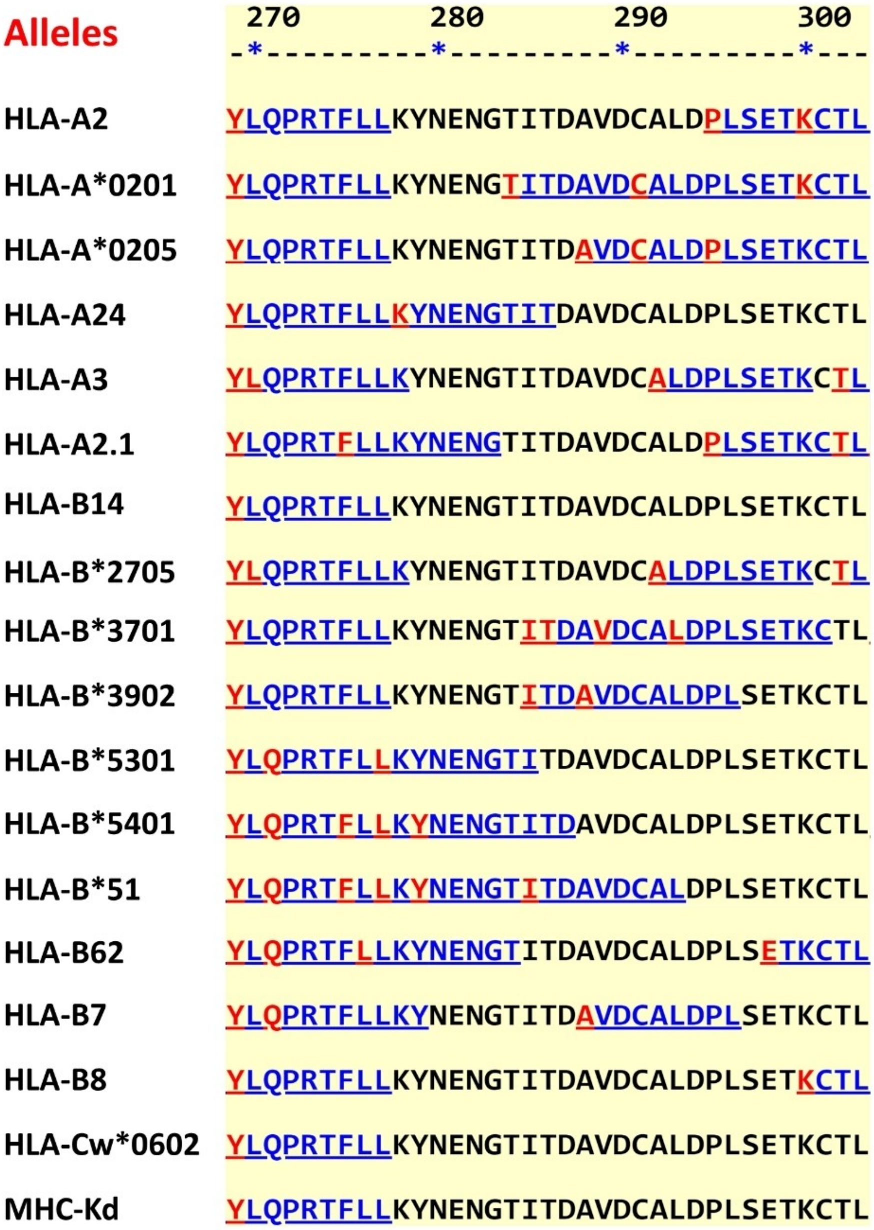
Epitope promiscuity prediction

For the identification of MHC class I binders in an antigen sequence, Propred1 first generates the overlapping nanomer peptides. This step is followed by the quantitative matrix-based calculation of the scores of the nanomer peptides, where the quantitative matrix is of selected MHC class I alleles. Finally, all nanomer peptides with higher scores than the selected threshold score are considered predicted binders of the selected MHC class I allele. Predicted binders on the antigen sequence are presented along the primary sequence and in a different sequence.

The server has a default threshold of 4 % on the observation that most of the alleles have nearly the same specificity and sensitivity at 4 %. The peptide "YLQPRTFLL" is predicted to bind 18 alleles (Figure 1).

B cell epitope prediction

The sequence-based output of the B cell epitope prediction by BepiPred-2.0 is presented in Figure 2 alongside the surface representation of the SARS-CoV-2 spike glycoprotein to show the localization of the predicted antigenic region on the surface of the viral protein. BepiPred-2.0 is trained only on the data of an epitope that is derived from crystal structures, which



```
Alleles        270       280       290       300
               -*---------*---------*---------*---

HLA-A2         YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
HLA-A*0201     YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
HLA-A*0205     YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
HLA-A24        YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
HLA-A3         YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
HLA-A2.1       YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
HLA-B14        YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
HLA-B*2705     YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
HLA-B*3701     YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
HLA-B*3902     YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
HLA-B*5301     YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
HLA-B*5401     YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
HLA-B*51       YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
HLA-B62        YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
HLA-B7         YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
HLA-B8         YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
HLA-Cw*0602    YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
MHC-Kd         YLQPRTFLLKYNENGTITDAVDCALDPLSETKCTL
```

Figure 1. ProPred-I output for promiscuous MHC Class-I binding Peptides. The server represents predicted binders using color variations. The predicted binders are represented in blue color while the red-colored peptides represent the first position of each binder, for easy identification of overlapping peptides.

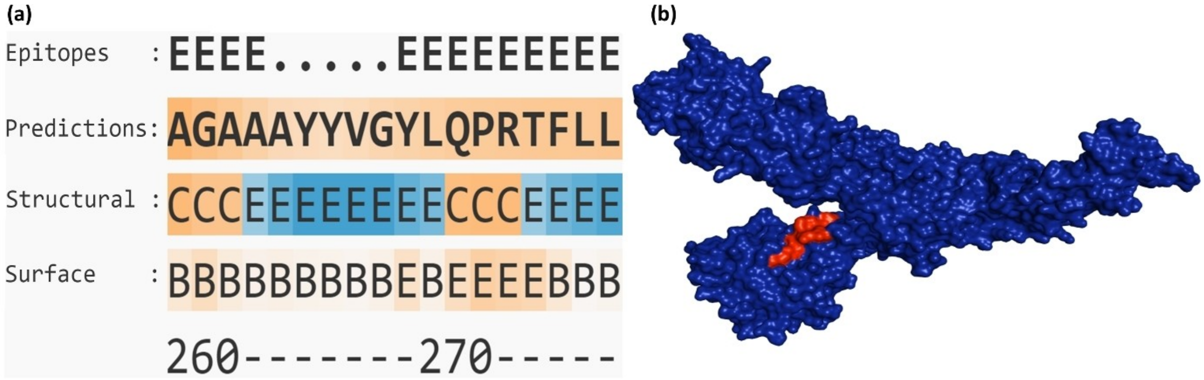

(a)
Epitopes : EEEE.....EEEEEEEEE
Predictions: AGAAAYYVGYLQPRTFLL
Structural : CCCEEEEEEEECCCEEEE
Surface : BBBBBBBBBEBEEEEBBB
260-------270-----
(b)

Figure 2. (a) BepiPred-2.0 B cell epitope prediction for the SARS-CoV-2 spike glycoprotein. The "Epitopes" row indicates the position of residues with scores above the set epitope threshold. "Predictions" illustrate the predictions of BepiPred-2.0 with the amino acid residues of the protein displayed in orange gradient. "Structural" illustrate the probability gradient of the alpha helix (H) (which is not covered in the figure), beta sheet (E) with blue gradient, and coils (C), with orange gradient. The exposed (E) and buried (B) regions of the sequence are illustrated by the "Surface" column which was obtained using the NetsurfP default threshold. The orange gradient in this column illustrates the predicted relative surface accessibility. (b) An illustration of the SARS-CoV-2 surface representation (PDB 6VXX), with the predicted antigenic region highlighted in red color.

presumably is to be of a better quality and indeed resulted in a significantly improved predictive power. At a BepiPred-2.0 threshold of 0.36, the "YLQPRTFLL" peptide, which is predominantly composed of beta sheets and coils, with a higher number of exposed residues than the buried, is predicted as a B cell epitope (Figure 2a).

The ABCpred B cell epitope prediction is a sequence-based predictor which presents results in a tabular frame and an overlap display (Table 5). For a tabular frame display, ABCpred ranks epitopes on the basis of obtained scores from the trained RNN (recurrent neural network). Peptides with higher score values are more likely to be predicted as B cell epitopes. The "YLQPRTFLL" peptide which falls among the top scorers, is ranked 8th with a score of 0.74 (Table 5).

The BcePred allows the prediction of B cell epitopes in protein sequences and presents it in graphical format as shown in Figure 3. The residue properties are plotted along the protein backbone, which facilitates the quick B cell epitope visualization on proteins. The peak of the segment of amino acid residue above the default threshold value (1.8) is considered as a predicted B cell epitope. The physiochemical property considered for this prediction is the antigenic propensity. The residues between 269 to 277, which make up the antigenic region of interest are predicted as a B cell epitope (Figure 3).

Antigenicity assessment

Using the VaxiJen server, the "YLQPRTFLL" peptide is predicted as a probable antigen. The model for the AllerTop server that was used for the allergenicity prediction of the epitope sequence of interest is based on the total set of allergens and non-allergens derived by the *k* nearest neighbors (*k*NN) algorithm, with the value of *k* being 3. The peptide epitope of

Table 5. The top ten representative ABCpred continuous B cell epitope prediction output ranked based on the prediction scores.

Rank	Sequence	Start position	Score
1	STFKCYGVSP	375	0.81
2	LQYGSFCTQL	754	0.80
2	QCVNLTTRTQ	14	0.80
3	IGAGICASYQ	666	0.79
4	FQQFGRDIAD	562	0.78
5	ICGDSTECSN	742	0.77
5	FKNIDGYFKI	194	0.77
6	VFRSSVLHST	42	0.76
7	KSNIIRGWIF	97	0.75
7	WMESEFRVYS	152	0.75
8	NNSYECDIPI	657	0.74
8	YLQPRTFLLK	269	0.74
9	NLDSKVGGNY	440	0.73
9	LALHRSYLTP	242	0.73
10	GVYFASTEKS	89	0.72
10	REFVFKNIDG	190	0.72
10	GNFKNLREFV	184	0.72

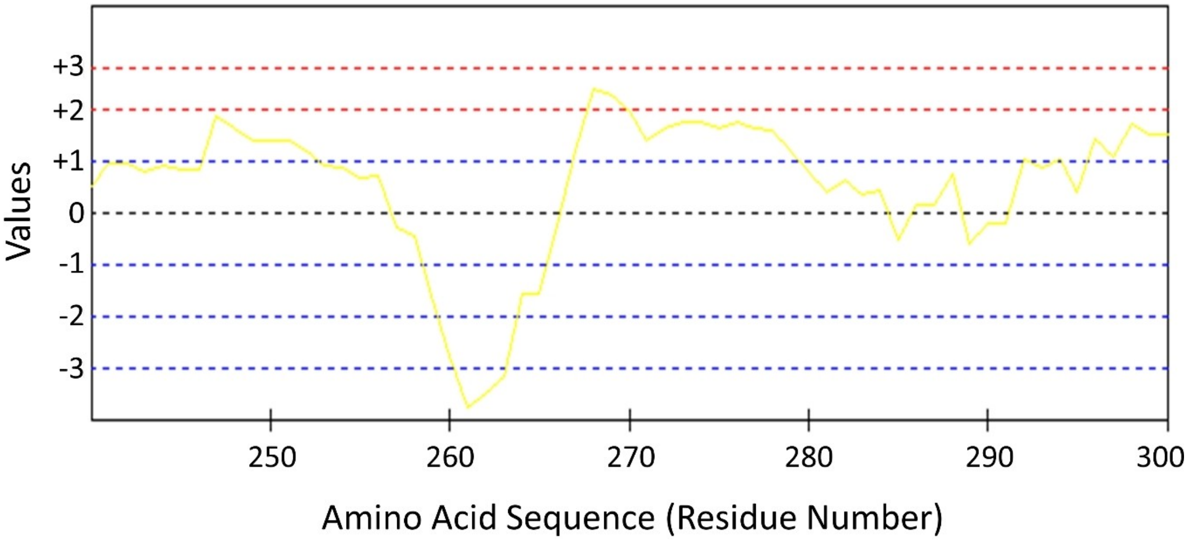


Figure 3. BcePred graphical output for the SARS-CoV-2 spike glycprotein B cell epitope prediction. The displayed graph covers the amino acid residues between 241 to 300 of the viral sequence.

interest was submitted in plain format also to this server which predicted a "non-allergen" status for the epitope. The probable orientation and location of transmembrane helices in the SARS-CoV-2 spike glycoprotein sequence were analyzed using the TMHMM *v*2.0 (Figure 4). In this program, the top (values between 1 and 1.2) of the probability plot is known as the best possible transmembrane helices prediction. As illustrated in Figure 4, the residues 269 to 277 were localized outside the residues with high transmembrane helical probability.

Considering the amino acid sequence of the SARS-CoV-2 spike glycoprotein, TOPCONS predicted the protein transmembrane topology, which specifies the membrane-spanning segments of the protein and their relative orientation (IN/OUT) to the membrane (Figure 5a and b). The prediction is a consensus from the algorithm of five different predictors (Philius, OCTOPUS, SPOCTOPUS and SCAMPI). The predictions from these five algorithms are used as input to the TOPCONS HMM (Hidden Markov Model), which gives a consensus-based prediction for the protein, alongside a reliability score on the

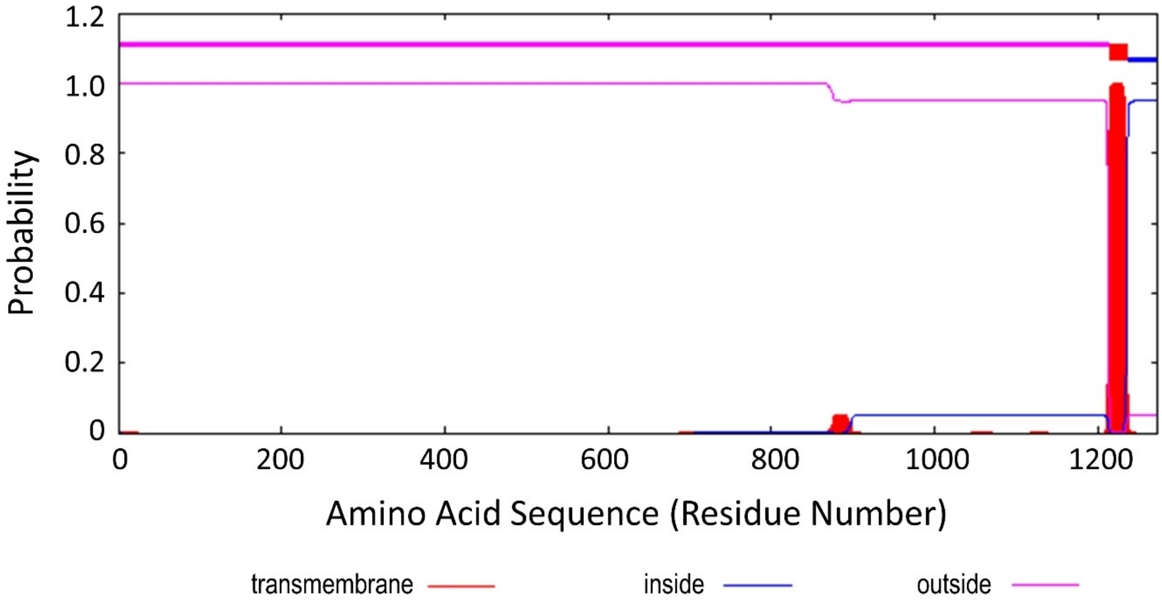


Figure 4. TMHMM graphical display of the SARS-CoV-2 spike glycoprotein transmembrane propensity prediction. The vertical red lines of the plot illustrate the predicted transmembrane helical segments of the protein. The blue line depicts the probability of a segment of the protein sequence to be intracellular while the outer membrane segment is denoted in pink line.

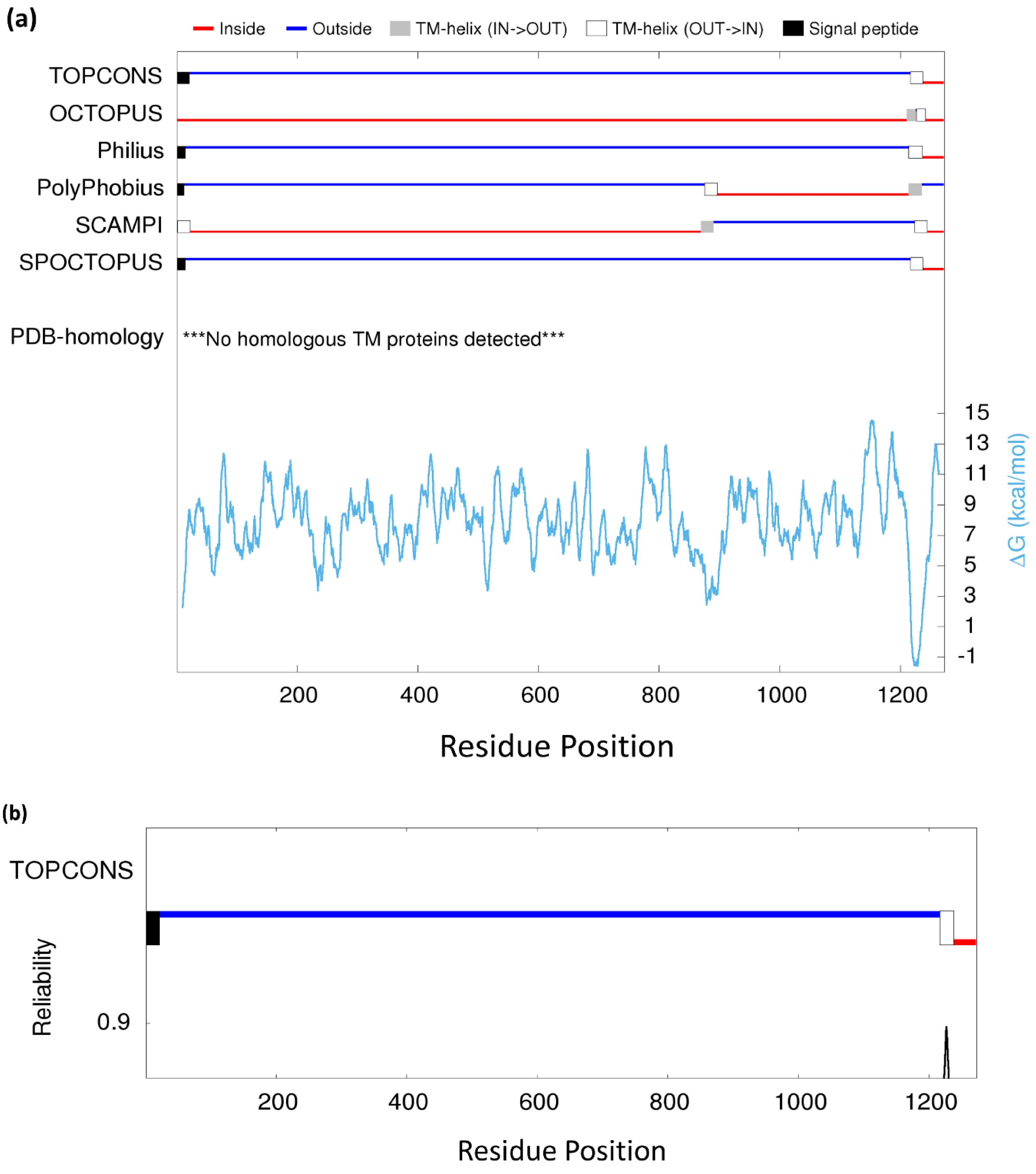


Figure 5. (a) TOPCONS predicted transmembrane topology with the predicted ΔG values. Transmembrane helices as denoted in the keys shown in the figure, are the grey and white rectangles. (b) Consensus prediction of the TOPCONS, detecting short sequence regions of the protein as predicted signal peptides (vertical black rectangle) and transmembrane helices (vertical white rectangle) respectively.

basis of the agreement of all included methods across the sequence. In addition, the ΔG-scale is used for the prediction of the membrane insertion free energy for a window of nineteen amino acids centered around each sequence position. The result shows no homologous transmembrane protein was detected among the five prediction algorithms (Figure 5a), although signal peptides were detected between the first 21 residues of the protein in the consensus prediction (Figure 5b).

To identify significant glycosylation patterns, N-linked glycosylation site of the SARS-CoV-2 spike glycoprotein was analyzed using the NetNGlyc 1.0. The default 0.5 value was selected as the prediction threshold. The result is displayed as a graph which illustrates the potential SARS-CoV-2 spike glycoprotein glycosylation sites. A total of 22 potential N-glycosylation sites, which does not include the predicted antigenic epitope, were observed on the spike glycoprotein of the SARS-CoV-2 as predicted by the NetNGlyc web server (Figure 6).

Antigenic peptide flexibility simulation

Having built the antigenic peptide using the Chimera software and likewise completed the energy minimization protocol, we performed simulation protocols to generate near-native dynamic conformations of the antigenic peptide. The asymmetric Metropolis scheme and Monte Carlo dynamics which satisfies the Boltzmann distribution of generated ensembles and microscopic reversibility requirements was employed by the CABS-flex for the completion of this protocol. The top five generated models [Figure 7] were then selected for molecular docking against the HLA*A-0201 allele of the human major histocompatibility complex Class I.

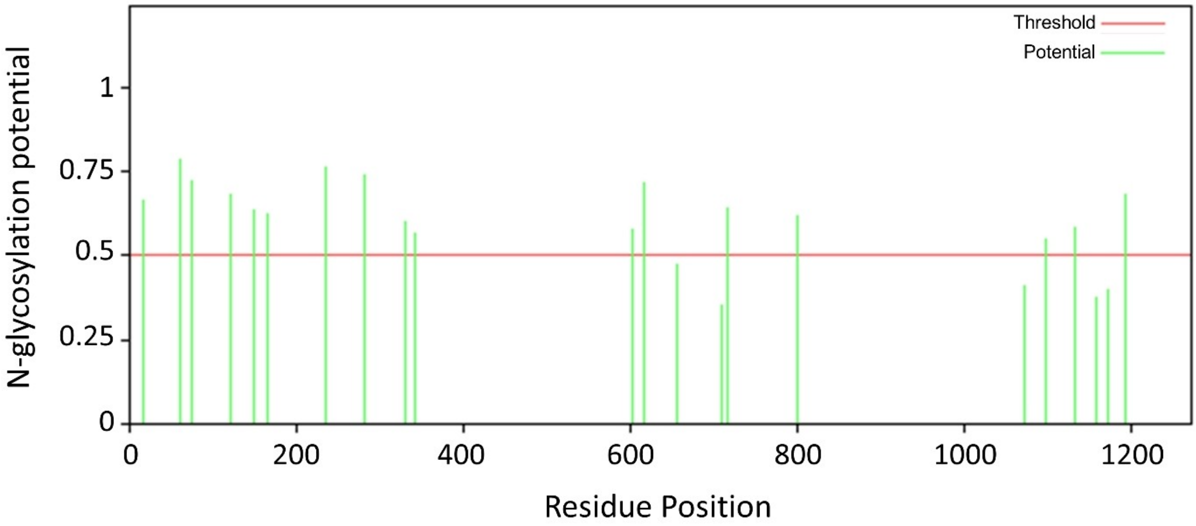


Figure 6. Predicted N-glycosylation sites in the SARS-CoV-2 spike glycoprotein.

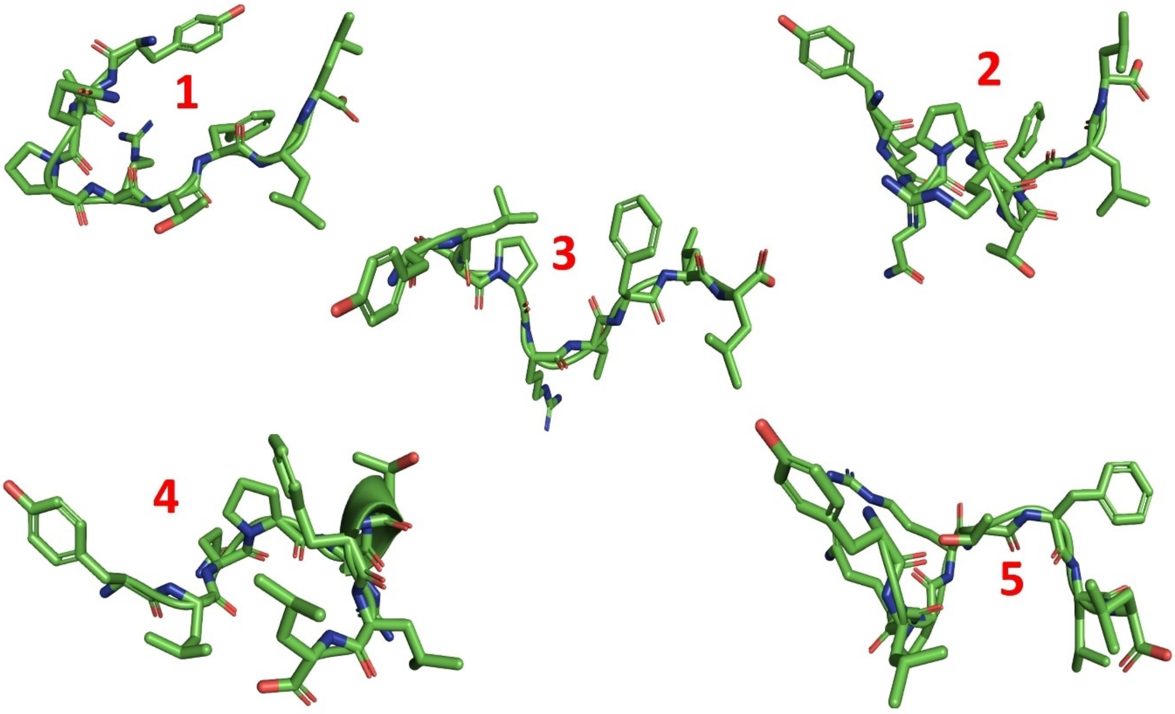


Figure 7. Stick representation of the top five generated models from the CABS-flex simulation protocol.

*Peptide docking against HLA*A-0201*

The generated peptide models were docked against the HLA*A-0201 protein in order to validate the strong binding predictions of the T cell epitope predictors. Table 6 shows the individual score for each peptide model as predicted by HPEPDOCK and ClusPro. In the HPEPDOCK docking algorithm, the flexibility of the peptide is considered through the generation of an ensemble of peptide conformations, after which the sampled conformations are docked globally against the whole protein using the rigid docking protocol. The binding of "model 2" to the HLA*A-0201-peptide binding site produced the best score for both docking tools. Figure 8 therefore, displays the pose of the "model 2" peptide epitope in the HLA*A-0201 binding pocket.

For the descriptive purpose of specificity and selectivity of epitope recognition by the HLA*A-0201, we conducted a similar molecular docking study on two additional peptides that were selected from the list of predicted processed T cell epitopes. These peptides were selected on the premise that intracellular peptides for MHC class I presentation can only be made by cytosolic proteasomes and proteases, followed by transportation into the endoplasmic reticulum through the Transporter associated with Antigen Processing (TAP) for further processing and an eventual transformation into antigenic peptides. For this reason, the second-ranked peptides by both predictive tools for the T cell epitope processing (IEDB and EpiJen) were selected. Both peptides (KIADYNYKL and YTNSFTRGV) were used as the positive control group for this study.

Following the CABS-flex flexibility simulation protocol on both peptides in order to generate their near-native conformation (Supplementary Figures 1 and 2), the molecular docking protocol was conducted on all generated models (Supplementary Tables 1 and 2). The binding poses for top-scoring models upon binding to the HLA*A-0201, as evaluated by both HPEPDOCK and ClusPro were also displayed (Supplementary Figures 3 and 4). The obtained binding energy output for each displayed peptide model suggests a strong affinity upon binding to the HLA*A-0201.

Conformational stability of the protein-peptide complex

Understanding the nature of crucial interactions that facilitate the binding and stabilization of the highest energy peptide within the HLA*A-0201 binding groove to trigger an immunological response is a necessity. From the interaction analysis shown in Figure 9, hydrogen bond formation exists between the HLA*A-0201-peptide complex, with Arg65, Lys66, Asp77, Lys146 and Ala150 amino acid residues of the HLA*A-0201 interacting with other residues across the length of the peptide. Hydrophobic interactions also mediate the binding of the peptide to the HLA*A-0201 via residues Lys66, Ala69, Asp77, Leu81, Tyr123, Ile124, Thr143, Trp147, Val152, Leu156,

Table 6. ClusPro and HPEPDOCK binding energies for the top 5 CABS-flex simulation models.

Peptide models	Receptor protein	Docking score (ClusPro)	Docking score (HPEPDOCK)
Model 1	HLA*A-0201	−852.9	−246.811
Model 2	HLA*A-0201	−934.7	−257.458
Model 3	HLA*A-0201	−916.7	−254.871
Model 4	HLA*A-0201	−799.2	−242.794
Model 5	HLA*A-0201	−910.2	−237.752

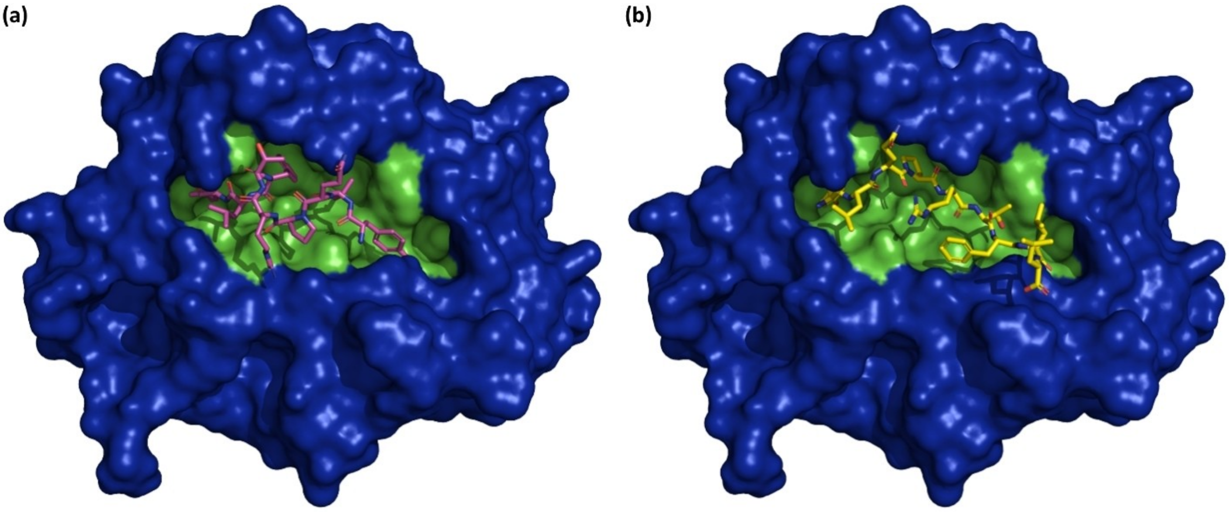


Figure 8. Binding poses for the peptide model with the strongest binding energy (model 2) in the HLA*A-0201 binding pocket as obtained from (a) the ClusPro model (purple sticks) and (b) the HPEPDOCK peptide binding model (yellow sticks).

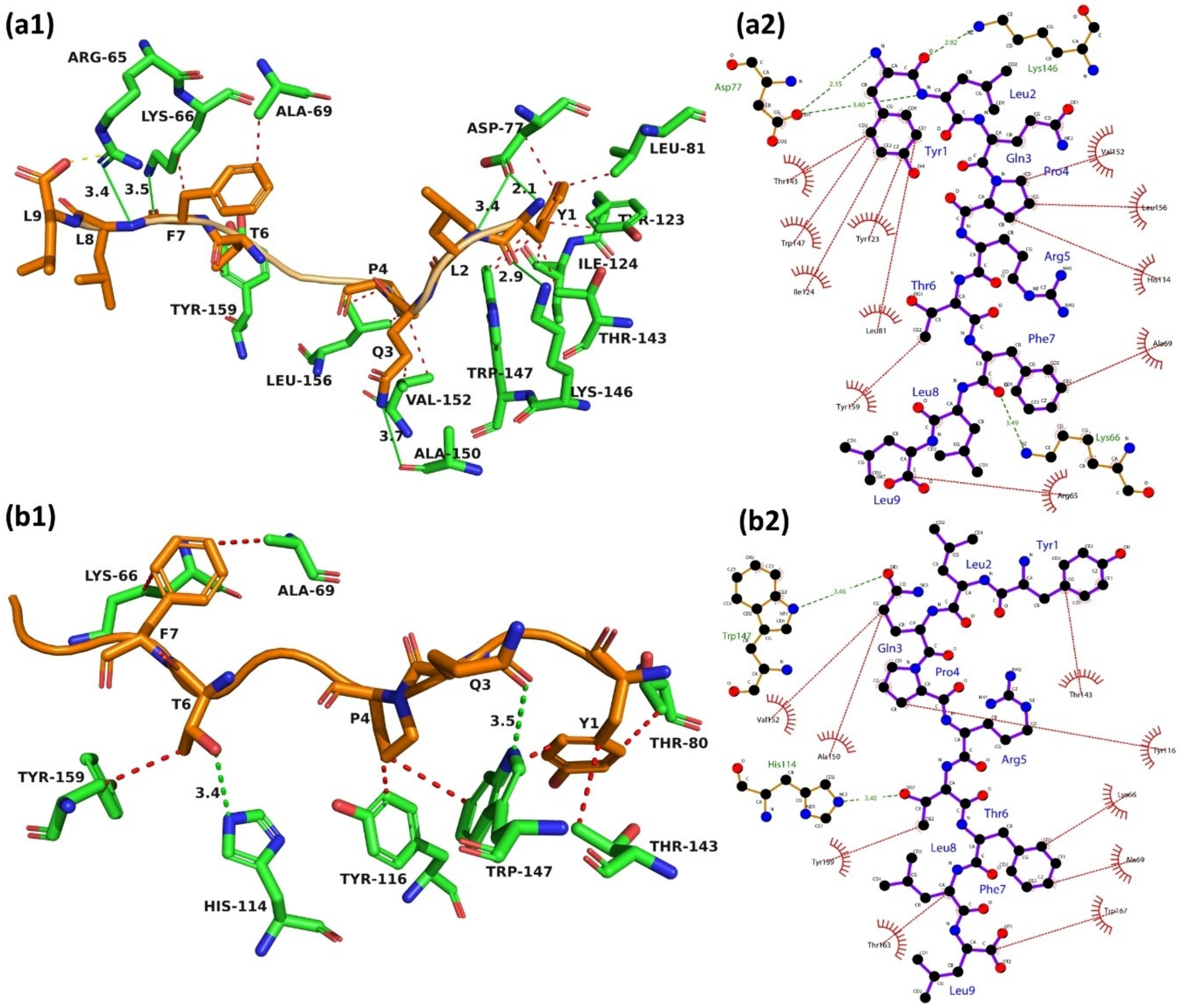


Figure 9. Interaction analysis of HLA*A-0201-peptide complex in 3D and 2D representations. (a) The HPEPDOCK-derived docked HLA*A-0201-peptide complex used as the starting complex for the molecular dynamics simulation. (b) A representative snapshot of the simulated HLA*A-0201-peptide complex.

and Tyr159 of the HLA*A-0201 receptor. There are six pockets within the binding site of HLA*A-0201, and these pockets might provide specificity for peptide side chain interactions. These pockets exists at the junction of the beta sheets and a helix, with the center of the binding cleft having no deep depression.

Met5, Tyr7, Tyr59, Glu63, Tyr159, Glu163, Trp167 and Tyr171 have been identified as amino acid residues of the pocket A of HLA*A-0201, and its pocket B is made up of His9, Thr24, Glu45, Leu66, Cys67 and Tyr99. Pocket C is composed of His9, Lys70, Thr73, Asp74 and Arg97, while Tyr99, His114, Leu156, Tyr159 and Leu160 are found at pocket D. Pocket E has residues His114, Trp133, Trp147, Val152 and Leu156.[57] Residues Lys66, Tyr159 of the HLA*A-0201-peptide complex are among the residues surrounding the opening rim of pocket A (Figure 9a). The side chain of Leu156, the ring of Tyr159 and the ring of Tyr99 also contribute to the formation of pocket D. Pocket F is formed from Asp77, Leu81, Tyr123, Thr143, and Trp147 of the HLA–A2-peptide complex.

Molecular dynamics simulation is considered an important method for understanding the physical basis of the structure and functions of biomolecules. Additionally, this method is commonly used to study the structure, dynamics, and conformational changes in macromolecular systems, as well as ligand binding among others.[58] Most proteins function by interacting with other proteins, and understanding these complexes is crucial to almost all physiological processes. The molecular dynamics simulation trajectories of the free HLA*A-0201 and HLA*A-0201-peptide complex were analyzed to get insight into their conformational behavior over a period of time. The RMSD of the backbone atoms RMSD describes the conformational changes between the free HLA*A-0201 and HLA*A-0201-peptide complex. Figure 10a shows that throughout the 100 ns simulation, the systems RMSD values range between 0.10 to 1.20 nm. The HLA*A-0201-peptide complex deviated below the free HLA*A-0201 throughout the simulation. The complex reached equilibrium just after 100 ns and maintain the same trend till the end of the simulation. The free HLA*A-0201 and HLA*A-0201-peptide complex have an aver-

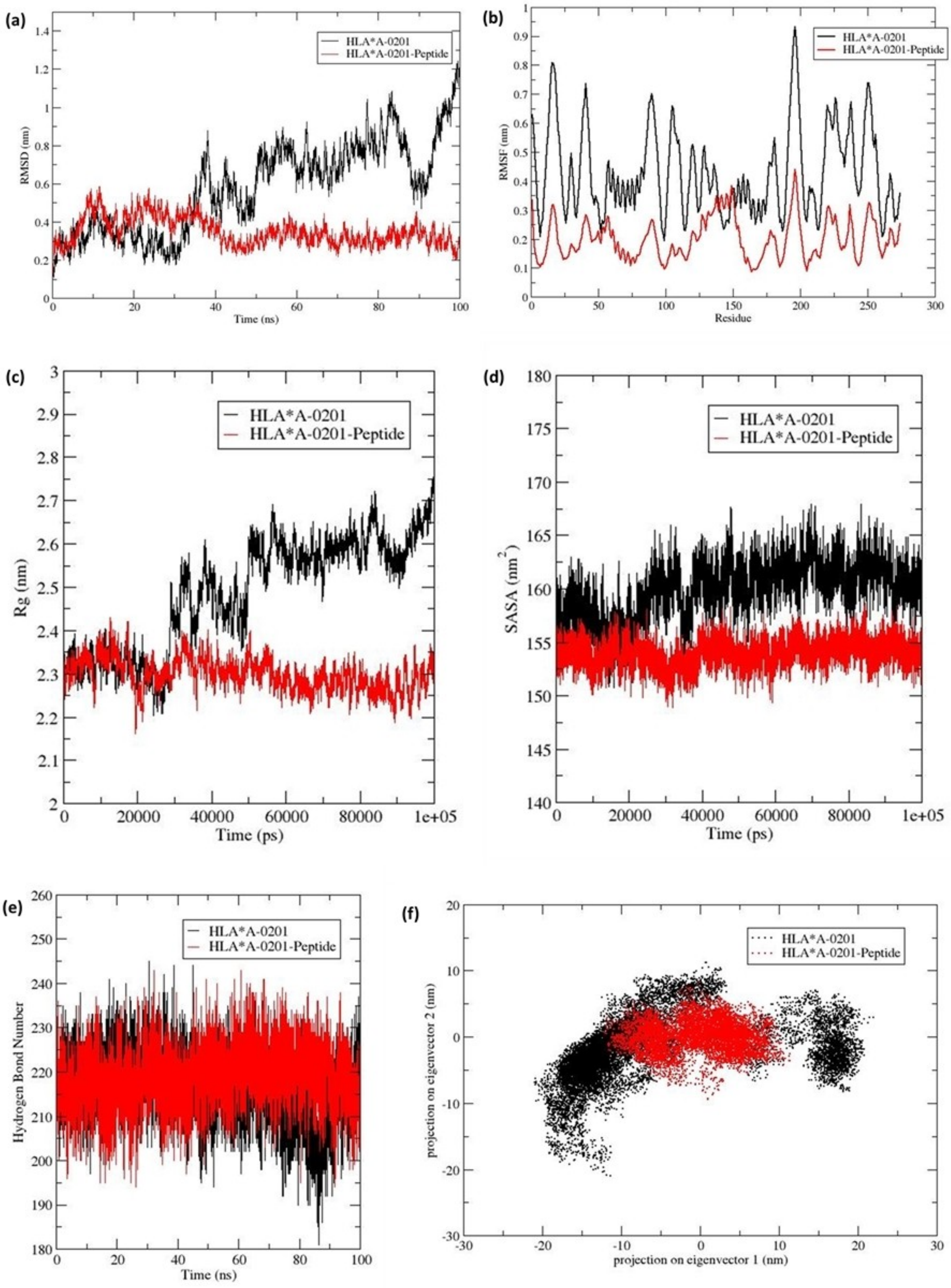


Figure 10. Molecular dynamics simulation and trajectory analysis of the free HLA*A*0102 and HLA*A*0102-peptide complex. Plots of the (a) Root-mean-square deviation, (b) Root-mean-square fluctuation, (c) Radius of gyration, (d) Solvent-accessible surface area, (e) Intramolecular hydrogen bonding, and (f) principal component analyses.

age RMSD of 0.757 nm and 0.350 nm respectively. The HLA*A-0102 peptide complex exhibited lower conformational mobility during the simulation, with the regions between 11–22 and 193–200 amino acid residues exhibiting almost similar patterns to those of the HLA*A-0102 free. The other regions of the free HLA*A-0201 exhibited higher fluctuations than the HLA*A-0102-peptide complex (Figure 10b).

The radius of gyration has been used as a measure of the compactness of proteins. Protein stability is known to be affected by the packing of amino acid residues. The lower degree of fluctuation of the HLA*A-0102-peptide complex throughout the simulations indicates a higher compactness of the system when compared with the free HLA*A-0102 protein.[59] The HLA*A-0201-peptide complex maintained low fluctuation level throughout the simulation time, unlike the free HLA*A-0201 which has a high radius of gyration. This indicate that the free HLA*A-0201is a less compacted system (Figure 10c). The compactness has been defined as a ratio of the accessible surface area of a protein to the surface area of the ideal sphere of the same volume.[60] The surface area of the protein exposed to the solvent molecules was examined using the solvent accessibility surface area (SASA). The SASA of the free HLA*A-0201 was higher when compared with that of the HLA*A-0201-peptide complex, describing the latter shrunken nature. The average SASA of the free HLA*A-0201 and HLA*A-0201-peptide complex are 160.311 nm^2 and 153.93 nm^2 respectively (Figure 10d). The interaction between the peptide and the HLA*A-0201 are mostly hydrophobic in addition to the identified hydrogen bonds, both of which were contributing to the stability of the complex (Figure 9b). Notwithstanding, an increased number of intramolecular hydrogen bonds was observed in the HLA*A-0201-peptide complex compared with the free HLA*A-0201 throughout the simulation time (Figure 10e).

The principal component analysis (PCA) gives information about the prominent modes of the trajectory of HLA*A-0102 free and HLA*A-0102 peptide complex. PCA can help to minimize the dimensionality of molecular dynamics simulation trajectory data. It gives a simple way to examine, analyze, and compare large-scale collective motion seen during the simulation. The eigenvectors with the largest eigenvalues have the highest contribution to the observed covariance.[61] The analysis was done by diagonalizing and solving the covariance matrix's eigenvalues and eigenvectors using the backbone atoms of the HLA*A-0102 free and HLA*A-0102 bound. The HLA*A-0102 bound shows a distinct cluster with less collective motion while the HLA*A-0102 free has a widespread cluster. The low flexibility observed in HLA*A-0102 peptide complex can account for the lesser space occupied in the phase space compared with the HLA*A-0102 free and as such the HLA*-0102 peptide complex appears to be more stable than the free HLA*A-0102 (Figure 10f).[62]

Codon optimization

For various biotechnological applications and in basic biochemical research, the production of heterologous protein is of major importance. However, there exist a limited number of eukaryotic and prokaryotic production hosts for many organisms under investigation. The codon usage of the gene in focus and that of its desired production host usually differ significantly. In most cases, this ends up in low protein recovery. In this context, JCat and ExpOptimizer offer the possibility for the adaptation of the codon usage of the gene that codes for the "YLQPRTFLL" peptide, in the *E. coli*. For this purpose, the codon adaptation index of the antigenic epitope sequence was displayed after calculation. The basis for the codon optimization is provided by the obtained value. Here, the JCat and ExpOptimizer calculated the codon adaptation index (CAI) and GC content of the optimized nucleotide sequence of the antigenic epitope of interest. The CAI values were noted as 1.0 and 0.87, for the calculation results that were obtained from JCat and ExpOptimizer respectively with respective 56.66 and 55.56 GC content values for both tools (Supplementary Figure 5a and b).

Based on these results, we compared the nucleotide sequence output from both tools. A sequence of 27 nucleotides in length was generated by both JCat and ExpOptimizer (TACCTGCAGCCGCGTACCTTCCTGCTG and TATCTGCAGCCGCGTACTTTCCTGCTG). A careful study of the output shows that the difference in both sequences occur on the 3rd and 18th nucleotide, where cytosine (C) in the JCat sequence is substituted for thymine (T) in the ExpOptimizer sequence. Based on the codon adaptation index result, we selected the JCat sequence for the design of the mRNA-based vaccine candidate, as it produced an optimum adaptation result. The codon adaptation protocol was likewise conducted on the control group peptides, (KIADYNYKL and YTNSFTRGV) in order to generate a set of optimized nucleotide sequences which were utilized for the design of their 3D structures. However, because the calculated CAI value for the two selected peptides that represent the control group is the same as that of the antigenic peptide of interest (1.0), a single graphical output was displayed to represent the relative adaptiveness of the three peptide sequences after adaptation by the JCat tool (Supplementary Figure 5b). The codon optimization protocol for both control group peptides also generated sequences with 27 nucleotides each (AAAATCGCTGACTACAACTACAAACTG was generated for the KIADYNYKL peptide while TACACCAACTCTTTCACCCGTGGTGTT was generated for the YTNSFTRGV peptide), with both producing GC content values of 37.04 and 48.15 respectively.

Mutagenicity study

Generally, the stability of proteins is a critical biophysical property that drives their evolution. Although the realized fitness of a specific strain of virus is a complex phenomenon which is the outcome of the interaction between host and viral molecules, a large fraction of the observed effect of mutation on fitness parameters, such as cell receptor fusion are also related to alterations in the viral protein stability.[62] In this study, we used DynaMut and I-mutant 2.0 for the evaluation of the effect of single mutation on the stability of the SARS-CoV-2 spike glycoprotein. For the N501Y single mutation, the server outputs the predicted change in stability (0.427 kcal/mol), along with the variation in entropy between the wild-type and mutant structures (−4.008 kcal/mol/K) (Figure 11). For the

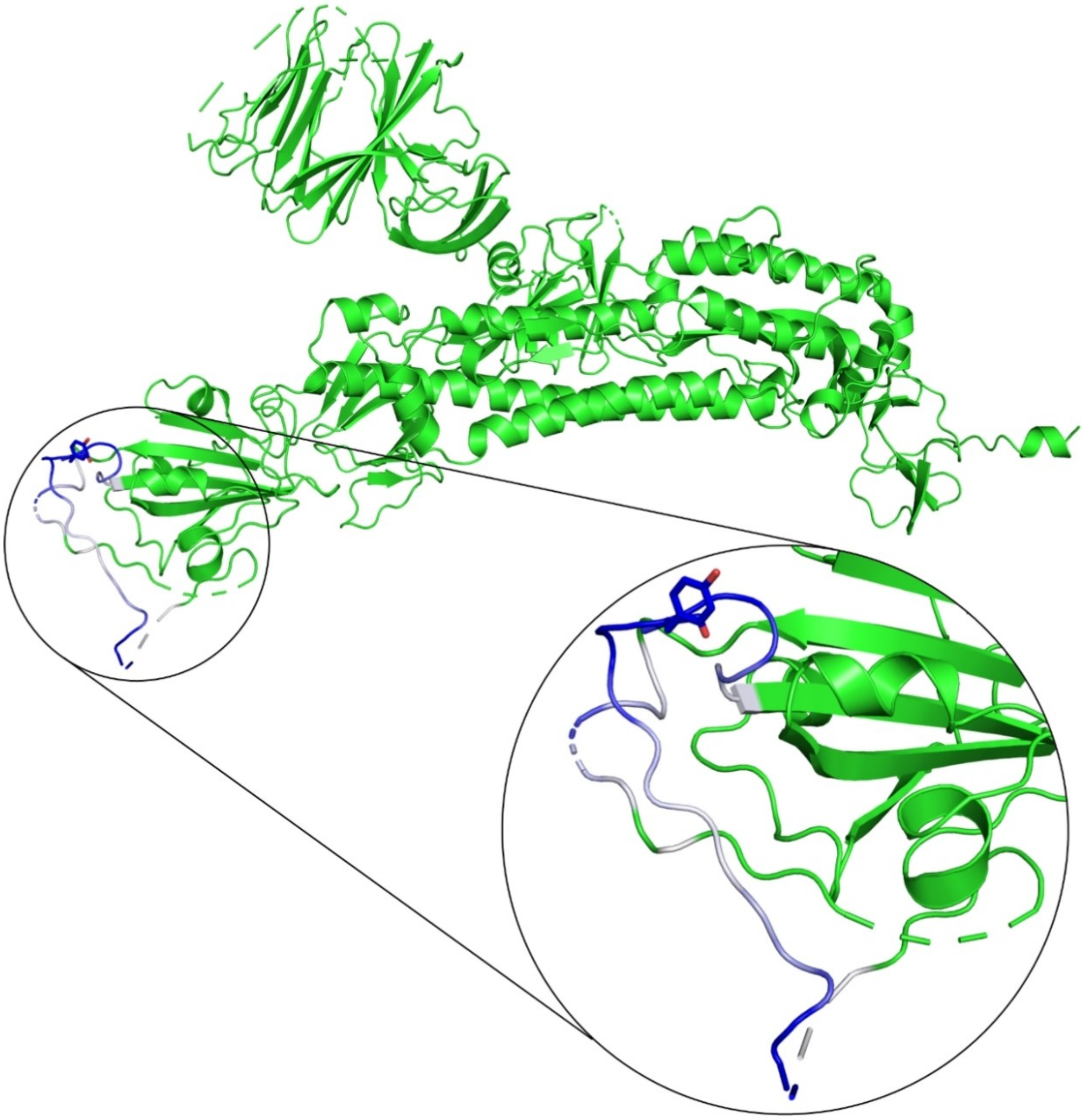


Figure 11. Amino acids colored based on the change in vibrational entropy upon mutation. Blue color signifies structural rigidification while red signifies flexibility gain. The output (blue colored) is an indication of an increase in stability upon the N501Y mutation on the SARS-CoV-2 spike glycoprotein.

purpose of comparison, the changes in stability calculated by normal mode analysis (NMA) and structure-based methods are displayed in Table 7. The output shows the N501Y mutation has a stabilizing effect on the viral protein. Visualization of non-covalent interactions (Figure 12), deformation energies and atomic fluctuations (Figure 13) of wild-type and mutant residues in their respective 3D structures, is also enabled by DynaMut. Deformation energy provides an estimate for the amount of local flexibility in the protein while the atomic fluctuation provides the amplitude of the absolute atomic motion. Estimations were performed over the first ten non-trivial molecule modes.

Table 7. Structure-based and NMA calculations of changes in stability upon mutation.

Methods	Prediction (kcal/mol)	Effect
DynaMut	0.427	Stabilizing
ENCoM	4.406	Stabilizing
SDM	0.660	Stabilizing

The first column shows the predictive methods. DynaMut and SDM uses a structure-based predictive model while the ENCoN bases its calculation on the normal mode analysis (NMA). Predictions from the three methods depict a stabilizing effect upon the N501Y mutation.

The I-mutant 2.0 prediction (Table 8) unlike the DynaMut prediction, is sequence-based, where a table of nineteen rows is returned as result to illustrate the mutational sign of free energy change. The nineteen rows in the I-mutant 2.0 output contains residues that differ from the one present in the corresponding position in the wild-type sequence. This corresponds to the differential number of columns returned as

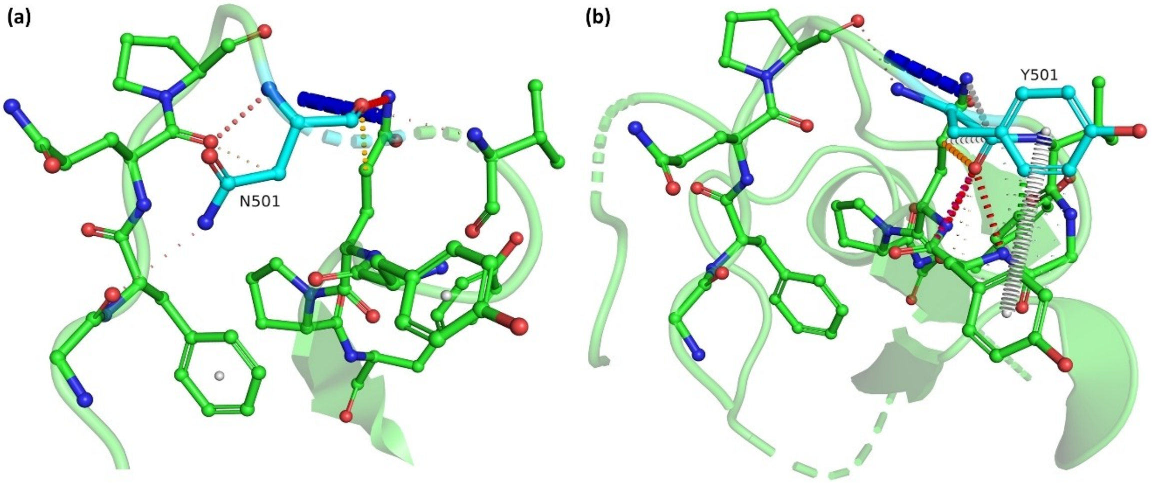


Figure 12. Interatomic interaction prediction upon the N501Y mutation. Residues of the wild-type (a) and mutant (b) are colored in light-green and are also presented along with the surrounding residues as sticks. The surrounding residues are involved in other types of interaction. The bonds are colored according to interaction types. The red, yellow and pink represent hydrogen bonds, ionic interactions and carbonyl contacts respectively.

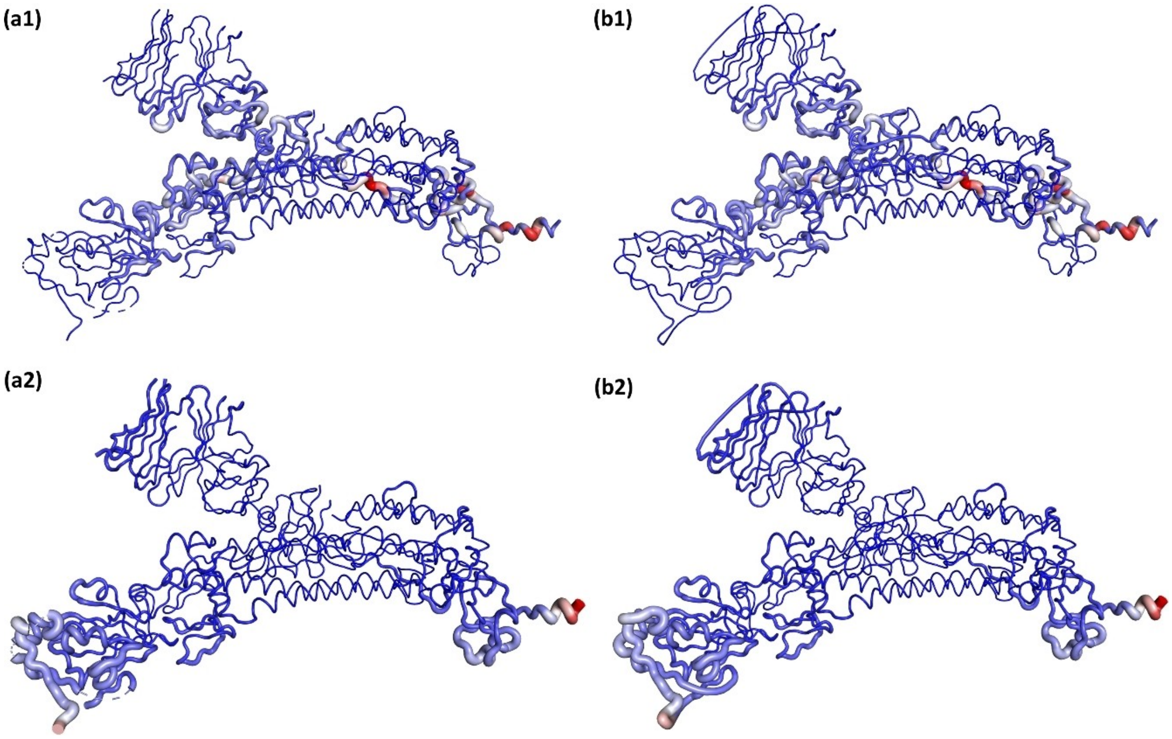


Figure 13. Visual analysis of deformation energies and atomic fluctuation. The deformation magnitude is illustrated by thin to thick colored tubes. Blue white and red represent the low, moderate and high magnitudes respectively (a1 and b1). The fluctuation magnitude is also illustrated by thin to thick colored tubes, where the blue white and red represent the low, moderate and high magnitudes respectively (a2 and b2).

output. The most common number of columns is 6, listing respectively the position in the sequence under consideration, the original name of the residue in one-letter code, one-letter code representation of the mutated residue, the predicted DDG (value of change in free energy) or the prediction sign (increase or decrease in stability), the pH, and temperature in which the prediction was conducted. One more column is included in the output table that in turn lists the RI (reliability index) value of the prediction. This occurs only when the prediction of the sign of the stability change starts from the protein sequence. The result presented in Table 8 shows only mutations with increasing stability effect on the protein as predicted by I-mutant 2.0. Mutations with decreasing effects were excluded.

Table 8. Sequence-based stability prediction upon residue mutation as obtained from the I-mutant 2.0.

Position	Wild-type	Mutant	Stability	RI	pH (−log [H$^+$])	T (°C)
501	N	V	Increase	6	7.0	25
		L	Increase	0	7.0	25
		I	Increase	4	7.0	25
		M	Increase	2	7.0	25
		F	Increase	2	7.0	25
		W	Increase	0	7.0	25
		Y	Increase	2	7.0	25
269	Y	V	Increase	3	7.0	25
271	Q	V	Increase	0	7.0	25
		L	Increase	1	7.0	25
		I	Increase	1	7.0	25
		E	Increase	2	7.0	25

The presented results are specifically for mutations with increasing stability effect

mRNA preparation and simulation

Following the codon optimization protocol for the antigenic peptide of interest, the generated nucleotide sequence (the JCat output) was selected for the design of the potential mRNA-based vaccine candidate. The nucleotide sequence was converted into the standard (A, U, C, G) RNA code before generating the PDB structure that was used as the SimRNA input file (UACCUGCAGCCGCGUACCUUCCUGCUG). The SimRNA simulation output is recorded as a trajectory file made up of the conformations of energies selected from a series of consecutive simulation steps. This protocol is conducted to fold the mRNA into its near-native 3-dimensional structure. The trajectory is converted into a series of PDB files which contains models in the reduced SimRNA output representation and the output by default also includes the secondary information (expressed in dots-and-brackets format) of the top five structural conformations (Figure 14). Detection of the secondary structure is through an in-built SimRNA classifier, which operates on the reduced 3D structural representation.

The same protocol was followed for the preparation and simulation of the control group mRNAs after the nucleotide sequence conversion into standard RNA codes (AAAAUCGCUGACUACAACUACAAACUG and UACACCAACUCUUUCACCCGUGGUGUU). The simulation protocol for the control group mRNAs also produced five top clusters each (Supplementary Figures 6 and 7), out of which their respective secondary structures were generated (Supplementary Figures 8 and 9).

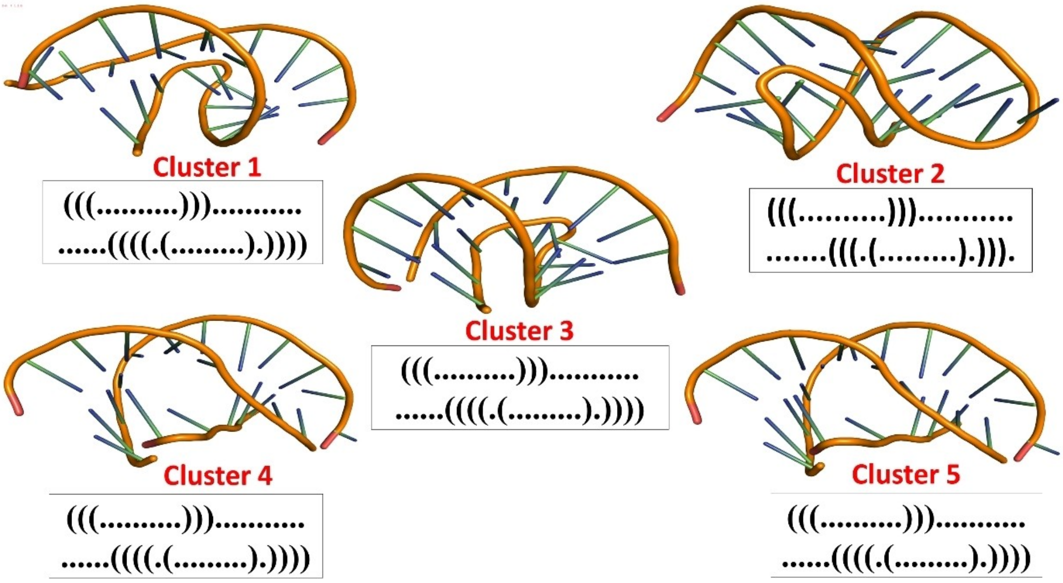


Figure 14. Secondary and tertiary structure depiction of the top five SimRNA simulation output. For each cluster, the tertiary structures are displayed in their 3D format above the corresponding secondary structures.

mRNA model docking and binding pocket dynamics analysis

The HDOCK utilizes a fast Fourier transform (FFT) based global docking program for the global sampling of putative modes of binding, where a pairwise scoring function which is shape-based, has been used. A 15° angle interval is used for rotational sampling, while a 1.2 Å spacing is adopted for the fast Fourier transform-based translational search. Optimization for the top ten translations for each rotation, with best shape complementarities from the fast Fourier transform search is conducted, using knowledge-based scoring functions. The top five clusters from the SimRNA simulation for the mRNA vaccine candidate and control group mRNAs were docked against the human TLR7 protein and the scores presented in Table 9 and Supplementary Tables 3 and 4. The ranked binding modes are clustered with an RMSD cutoff of 5 Å. The binding orientations of the clusters with the strongest affinity are also displayed in Figure 15a and Supplementary Figures 10a and 11a. Observation from the binding pocket dynamics output also shows that cluster 1 mRNA vaccine candidate is bound to the stable cavity of the TLR7 pocket, likewise the cluster 4 and cluster 3 of the control group mRNAs respectively. D3Pockets colors the grid points with pocket cavities. The higher the red gradient at each point, the more frequent such points in the cavity are observed throughout the molecular dynamics trajectory. The higher the blue gradient at each point, the less frequent such points in the cavity are observed throughout the molecular dynamics trajectory. The subpocket region predominantly composed of red points is more stable than other regions (Figure 15b, Supplementary Figures 10b and 11b).

The resulting complex interaction formed from the binding of the "Cluster 1" mRNA vaccine candidate and the human TLR7 was analyzed using the Protein-Ligand Interaction Profiler (PLIP)[63] to give insight into the nature of amino acid residues that facilitates binding. There are nine hydrogen bonds formed between the TLR7 and the "Cluster 1" mRNA vaccine candidate as shown in Figure 16 and Table 10. Also, salt bridges exist between the Cluster 1 mRNA vaccine candidate and the TLR7 complex. This suggests that the identified amino-acid residues might be important for the Cluster 1 mRNA vaccine candidate binding to the TLR7 to exert its innate immunity.

Discussion

As a pandemic-response strategy, the mRNA vaccine platform has an advantage, considering its efficiency and flexibility in

Table 9. Docking scores for the top 5 generated SimRNA clusters against the human toll-like receptor 7 protein.

Clusters	Receptor	Docking score (Kcal/mol)	Ligand RMSD (Å)
1	TLR7	−415.23	170.98
2	TLR7	−390.82	177.96
3	TLR7	−414.31	170.18
4	TLR7	−376.84	187.11
5	TLR7	−375.04	188.76

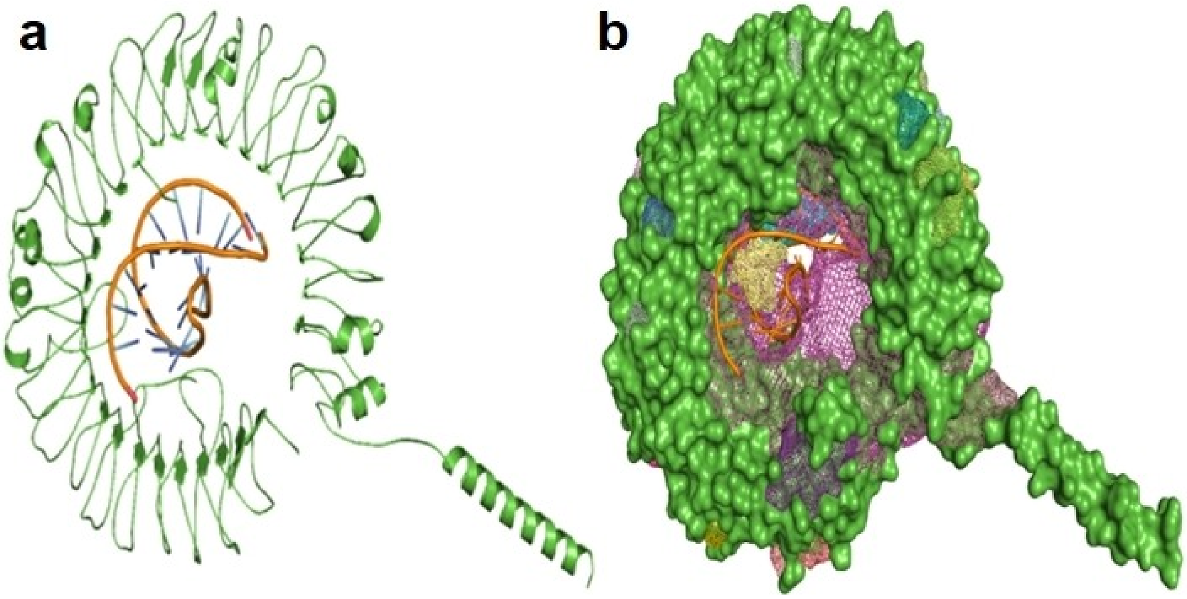


Figure 15. Cartoon representation of the binding pose of "Cluster 1" mRNA vaccine candidate in the human TLR7 binding pocket (a). Surface representation of the human TLR7 is displayed, with meshes in different color gradients denoting the different degrees of stability. Green mesh denotes a metastable pocket while blue and red denote the unstable and stable pockets respectively (b).

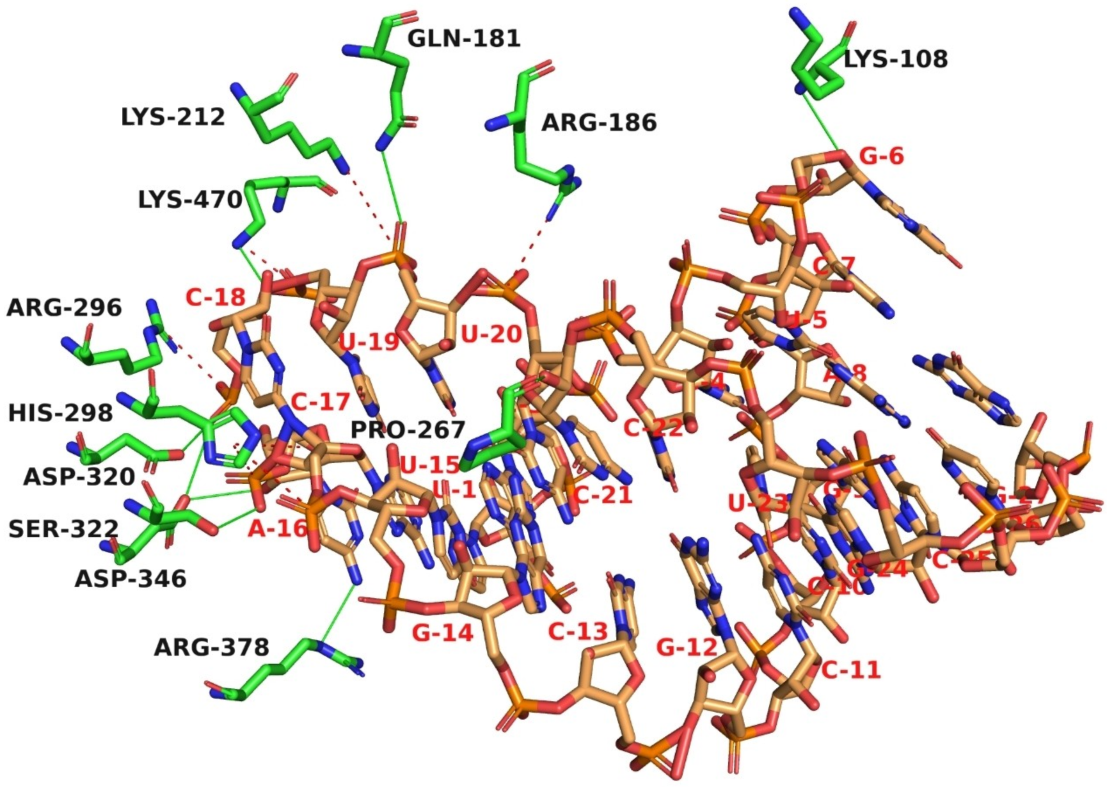


Figure 16. 3-D interaction analysis of the Cluster 1 mRNA with the TLR7 pocket residues.

Table 10. Interaction analysis of Cluster 1 mRNA with the TLR7 pocket.

Nucleotides of the Cluster 1 mRNA	Interacting TLR7 amino acid residues	
	Hydrogen bonds (distance, Å)	Salt bridges (distance, Å)
G6	Lys108 (3.35)	–
A16	–	Arg298 (5.15), His298 (4.55)
C17	Ser322 (2.93), Asp346 (3.66), Arg378 (3.40)	His298 (3.92)
C18	Asp320 (3.09), Asp346 (3.81), Lys470 (2.70)	Arg296 (4.79)
U19	–	Lys470 (3.64)
U20	Gln181 (3.47)	Lys212 (5.13)
C21	Pro267 (2.44)	Arg186 (4.96)

the design and manufacturing of immunogen.[64] The spike glycoprotein of the coronavirus linked to the 2002 outbreak of SARS had been suggested by earlier works to be an ideal target for protective immunity while many vaccine candidates in various developmental stages are currently undergoing evaluation.[65] Shortly after the determination of the genetic sequence of the SARS-CoV-2 in January 2020, a LNP (lipid-nanoparticle)-encapsulated mRNA vaccine (mRNA-1273) which expresses the prefusion-stabilized spike glycoprotein, was developed by the NIAID (National Institute of Allergy and Infectious Diseases) vaccine research center, within the NIH (National Institutes of Health) and Moderna.[66] Animal challenge experiments have been conducted,[67] where the mRNA-1273 vaccine demonstrated protection.[68] Recently demonstrated also was the safety and efficacy of BNT162b2, which is another mRNA vaccine.[69]

Our reverse vaccinology approach in the design of a potential mRNA vaccine candidate followed series of steps, first of which is the determination of the most suitable antigenic epitope from the viral spike glycoprotein to serve this purpose. To achieve this, the SARS-CoV-2 spike glycoprotein sequence was analyzed for the detection of a single epitope with the potential of triggering immune response through its ability to interact with a wide range of MHC class-I alleles and likewise stimulate the adaptive immune system for the secretion of antibodies. T-cells scan ligands of the major histocompatibility complex presented to them on professional APCs (antigen-presenting cells), cells of the lymphoid lineage and nucleated cells surface, which expresses molecules of the MHC class I. This allows the detection of the presence of abnormal self-antigens such as those expressed by cancer cells, as well as the detection of pathogen-derived antigens. Complexes of the major histocompatibility complex and their ligands are generated by antigen processing and antigen presentation pathways, which consist of various enzymatic events with the inclusion of specialized processes and organelles, that are specific for different classes of the major histocompatibility complex.[70]

The main MHC class I processing pathway involves proteasome-mediated degradation of proteins, followed by the transport of the degradation products by the TAP (transporter associated with antigen processing) to the ER (endoplasmic reticulum), where the binding of peptides to molecules of the MHC class I takes place, and then presentation on the surface of the cell by the MHCs. The responsibility of the proteasome is directed at the generation of the C-terminus of the final presented peptide (not the N-terminus),[71] while the transporter associated with antigen processing is an ATP-dependent protein that transport peptides belonging to the ABC (ATP-binding cassette) family of transporters.[72] This family transport a wide range of molecules across the membrane, including sugars and large polypeptides. Identification of promiscuous or cross-reactive antigenic regions in a sequence is another area of interest for immunologists, and this makes the process an important step in the design of subunit vaccines. The promiscuous regions have the ability to bind to many alleles of the MHC class I and this binding is their recognition prerequisite by the cytotoxic T lymphocytes (CTLs).[73] Results from the T cell epitope prediction and processing shows the predicted antigenic epitope for the design of the potential mRNA candidate satisfies the described antigenic properties, with a high probability to exhibit promiscuous binding to many MHC class I alleles and as such was selected for further antigenicity validation studies.

The immune system of humans has an incredible pathogen-fighting ability (viral, fungal and bacterial infections). One of the most critical events of the immune system which is involved in the clearance of infectious organisms is the interaction between antigens (such as pathogenic organism proteins) and antibodies.[74] The binding of antibodies to antigens occurs at sites known as B cell epitopes hence, identification of B cell epitopes (areas on surface antigens which can bind to antibodies) may facilitate the development of different immune-related therapies, such as vaccine development.[75] The importance of the B cell epitopes also cuts across allergy research and the determination of cross-reactivity of the IgE-type allergen epitopes. These epitopes may be continuous (linear) or discontinuous (conformational). When linear synthetic peptides are capable of inducing antibodies that can cross-react with parent proteins or are observed to cross-react with anti-protein antibodies, then such peptides are termed continuous (linear) epitopes.[76] The protective linear epitopes of the B cell may facilitate the synthesis of efficient antiviral peptide vaccines. A dominant linear B cell epitope in an autoimmune disease state is also used as the target for the response of neutralizing antibodies.[77] The B cell epitope prediction results with the aid of several tools have also shown the potentials of the antigenic epitope for antibody recognition.

To further validate the antigenic potentials of the predicted epitope, the allergenic potentials, transmembrane topology and N-glycosylation profile were assessed. Allergy is a form of hypersensitivity to regular innocuous substances, such as pollen, dust, vaccines, foods or drugs. Allergens are known to provoke the response of IgE antibodies as small antigens. Such antigens do enter the body at minute doses through mucosal surface diffusion, then trigger a type 2 T helper (Th2) response.[78] The allergen-specific B cells are driven by the type 2 T helper cells specific for allergen recognition, for the production of antibodies (IgE) which binds to the FcεRI, a high-affinity surface receptor located on activated eosinophils, mast cells and basophils. Upon activation, these cells release mediators that have been stored, giving rise to tissue damage and inflammation.[79] TMDs (transmembrane domains) are predominantly composed of non-polar amino acid residues and may once or severally traverse the lipid bilayer. Transmembrane domains are also usually made up of alpha helices. The polar peptide bond is capable of forming internal hydrogen bonds between amide nitrogens and carbonyl oxygens, or either being hydrated.[80] At the position where water is essentially excluded within the lipid bilayer, peptides do adopt the alpha helical configuration which in turn maximizes their internal hydrogen bonding. It has been established that proteins with two or more transmembrane helices are not ideal for the development of vaccines because of the difficulty in expressing them in soluble form, and as such should be excluded during the process of screening.[81] NLG (N-linked glycosylation) is a complex biosynthetic process which regulates protein maturation through a secretory pathway. The regulation of this cotranslational modification is by a series of enzyme-catalyzed reactions, which leads to the transfer of a core lipid carrier glycan to a protein substrate. The N-linked glycoprotein biosynthesis is well characterized,[82] while amino acid residues in such highly glycosylated regions may be shielded by masking carbohydrates from presentation to antibodies.[83] With consideration given to the obtained results from the predictive tools used for the purpose of this study, we validate the antigenic relevance of the predicted epitope, having shown a non-allergenic potential, does not form a transmembrane helical structure and also not glycosylated.

Proteins are dynamic macromolecules with their function linked intricately to their biological motions. It has been established that genetic disease mutations and drug resistance can both act through alterations in protein dynamics and conformational equilibria.[84] For a complete understanding of the molecular consequences of a mutation, it is necessary to consider changes in the dynamics of the protein involved. On the 18th of December 2020, the South African national authorities announced the discovery of a new SARS-CoV-2 variant which is spreading rapidly in several South African provinces. This variant because of the N501Y mutation has been named 501Y.V2, and the mutation has been observed to be conserved across several variants of the virus. While highlights from genomic data have shown that the variants displaced rapidly other circulating lineages of the virus, suggestions from preliminary studies have also linked the variants with higher viral load, suggesting increased transmissibility potential.[85] In this study we analyzed the stability dynamics of these variants upon mutation using both structure and sequence-based predictive tools. Results from this study suggests that the N501Y mutation increases the stability of the SARS-CoV-2 spike glycoprotein, giving credence to earlier studies suggesting increased transmission potential and viral load. We aimed further to predict the possibility of having such stability-linked mutation within amino acid residues of the antigenic epitope of interest. Results from this study revealed two amino acid residues within the antigenic epitope with stability-linked potentials (Y269 and Q271) as displayed in Table 8. The results further suggest other possible mutations at position 501 with an increasing effect on the stability dynamics of the spike glycoprotein.

Interactions between peptides and proteins play a critical role in many biological processes such as immune responses, cellular regulation, and signal transduction. It has previously been reported that short peptides mediate about 40 % of the protein-protein interactions. Therefore, the determination of the protein-peptide complex structures involved in such interactions is important for understanding the molecular mechanism and thus modulations of the protein-protein interactions for the purpose of therapeutics.[86] Proteins and nucleic acids also are two important biological macromolecule types in the cell. The interactions between both macromolecules are essential for various biological processes such as RNA transcription, replication of the DNA and its repair, protein synthesis, cellular regulation and signal transduction. Therefore, determining the protein-nucleic acid complex structure is also essential for the understanding of the atomic level biological processes and thus the development of drugs or therapeutic interventions to target these interactions.[87] However, only a few of the protein-nucleotide and protein-peptide interactions have been experimentally determined because of the technical difficulties and high cost of experimental methods. Computational modeling such as molecular docking has a critical role in the process of determining protein-nucleotide and protein-peptide complex structures.[88] Here, two docking protocols were performed. Docking of the predicted antigenic peptides against the HLA*A-0201 protein was aimed at validating the predicted strong binding by the T cell epitope prediction tools (Table 1) and likewise to predict the near-native binding model of the protein-peptide interaction (Figure 8). Protein-RNA docking protocols were also conducted to predict the affinity of the designed mRNA vaccine candidate upon binding to the human TLR7 protein (Table 9), thereby suggesting its potential as an ideal vaccine candidate (Figure 15a).

Identifying the ligand-binding and druggable pockets of target proteins is of critical importance, especially in SBDD (structure-based drug design). Molecular docking also is an important virtual screening technology, which can speed up the rate of hit compound identification.[89] The first essential step in molecular docking is to identify a target and its binding pocket. Protein surfaces are usually composed of multiple cavities. Small molecules capable of binding to these cavities have the potential of adjusting the protein function.[90] Analysis conducted on the pocket dynamics of the TLR7 protein upon RNA binding indicate that the designed mRNA vaccine candidate is bound to a stable cavity, thereby suggesting an ideal interaction for immunological stimulation (Figure 15b).

However, mRNAs are vulnerable to degradation. A prevention mechanism against such vulnerability must therefore be engineered into the mRNA-based vaccines to increase its stability. Eukaryotic and viral mRNAs possess at the 5′ end, a methylguanosine cap, which contains two forms of methylation. The m7G (7-methylganosine) cap is added during the process of transcription, through a triphosphate bridge, for the prevention of premature degradation, and also important for the maturation, export and initiation of the mRNA translation.[91] Another stabilizing element of the mRNA is the poly(A) tail. Its deletion destabilizes the mRNA.[82] Doel and coworker[92] have reported that the polyncleotide phosphorylase-aided removal of poly(A) reduced the polysome size, number of translational rounds and peptide elongation rate. The poly(A) tail is therefore important for the maintenance of the mRNA vaccine stability and its successful translation.[93] The addition of the 5′ cap to the final vaccine construct can be achieved through the use of a nucleotide cap analog after *in vitro* transcription reaction or through the use of a capping enzyme.[91]

Bell *et al.*[94] reported an interesting UTR (untranslated region) modification for the development of mRNA vaccines, where engineered riboswitches were added to the 3′ UTR of a vaccine construct for the regulation of gene expression and RNA amplification. The riboswitches are made up of a hammerhead ribozyme from the satellite RNA of tobacco ringspot virus actuated by an aptamer sensor specific for theophylline. Upon the addition of the vaccine to a cell host, the gene expression was modulated by the riboswitches.[94] The approach consequently has the potential to produce a high expression of vaccine antigens, hence necessary to be incorporated into the final mRNA vaccine construct.

mRNA vaccines have been reported to be highly efficient in expressing antigens, but secondary structures and sequences from mRNAs can also be recognized by specific innate immune receptors, resulting in the inhibition of protein translation.[95] However, with the advancement in the understanding of RNA biology, various techniques can now be used to increase mRNA vaccine potency. These methods include the use of modified nucleosides and optimization of the mRNA sequence. To avoid innate immune sensor recognition, modified nucleosides like 5-methylcytidine, optimized codons, pseudouridine, and cap-1 structure, can be incorporated into mRNA vaccines to increase the efficiency of translation.[96] In addition, the mRNA vaccine formulation and the route of administration are also essential for the determination of the kinetics, immune response potency, as well as the magnitude of antigen expression. For instance, the intravenous administration of unmodified naked mRNA led to the stimulation of the innate immune response and a quick digestion by ribonucleases.[97] These limitations however, can be addressed through mRNA modification and an appropriate system of delivery. The administration of mRNA vaccines is through a local or systemic method based on the requirements of antigen expression localization. Subcutaneous, direct intramuscular or intradermal injection of *in vitro* transcribed mRNA are the major routes of delivery for mRNA vaccines against infectious diseases, while intravenous and intraperitoneal administration of mRNA vaccines are used when there is need for a systemic expression of antigens of interest, mostly for therapeutic purposes.[97]

While mRNA nucleoside modifications and FPLC or HPLC purification have been shown to reduce innate immune response attack against the mRNA vaccine[98] and consequently leading to greatly enhanced mRNA stability and antigen expression,[99] possible immune stimulation by both the mRNA (in an antigen-independent mechanism) and the expressed antigen is also desirable.[100] Because our computational modelling suggested that the proposed mRNA has a strong affinity for TLR7 (which could direct innate immune response stimulation) in addition to the potential activation of full-fledge adaptive immune response by the expressed antigen, it would be interesting to explore the potentials of the mRNA in achieving this dual activity. Undoubtedly, achieving both effects is highly challenging as the interaction of mRNA with the PPRs would directly result in the inhibition of its translation. Hence, most vaccine development efforts have always focused on one side of the mRNA vaccine potentials.

However, some research groups have been exploring the possibility of combining the two features to maximize the immunostimulatory ability of the mRNA.[101] Of particular interest is the study by Fotin-Mleczek and co-workers[100] where they developed two-component mRNA vaccine approach and demonstrated its effectiveness by showing that treatment of mice with a two-component antitumor mRNA vaccine elicited a strong antitumor response against OVA-expressing tumor cells both in a prophylactic as well as in a therapeutic context.[100] The two-component mRNA vaccine contained free and modified (protamine-complexed) mRNA. Their results based on studies from *in vitro* experiments as well as animal models showed that such two-component mRNA design could result in the stimulation of innate immunity via TLR7 activation whilst allowing antigen expression. Thus, it is of interest to us to explore these possibilities via follow-on experimental efforts in order to actualize the potentials of the identified mRNA vaccine candidate.

Conclusion

The SARS-CoV-2 in recent times has been the leading cause of the deadliest global pandemic, resulting in excess mortality especially among the vulnerable and older populations. Prevention of the infection is a mandatory task that has been quite challenging owing to the fast mutation rate of the virus, hence leading to the emergence of new variants suspected to be more infective than the wild-type. The reverse vaccinology approach can be harnessed for the discovery of the desired solution as it saves both cost and time. In this study, a potential mRNA-based vaccine candidate was designed with the aid of various computational tools which were directed at first predicting the most effective antigenic region of the viral spike glycoprotein to trigger the desired immune response. The nucleotide sequence obtained as a result of the codon adaptation of this antigenic peptide was then used in the mRNA-based vaccine candidate construction, which was targeted at the human toll-like receptor 7 protein. Results obtained from the study suggest the designed vaccine candidate might be an effective therapy to curb the SAR-CoV-2-linked infection and its spread. We also speculate that the N501Y mutation is linked to the stability of the viral protein and as such have predicted regions of the antigenic epitope with such potential mutations.

Supporting Information Summary

In this study, we have designed a potential mRNA-based vaccine candidate using several computational approaches which were directed first towards the prediction of an efficient antigenic region in the SARS-CoV-2 spike glycoprotein to trigger the desired immune response against infection. Additional computational methods such as, the molecular docking and molecular dynamics simulation were used to validate the stability of the antigenic peptide upon binding to the human HLA*A-0201, after which the mRNA vaccine candidate was designed to target the human TLR-7.

Competing Interest

Authors declare no competing interest.

Funding

Authors received no funding for this project from any organization.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords: Antigens · Reverse vaccinology · SAR-CoV-2 · Spike glycoprotein · Vaccine · Viruses

[1] C. Drosten, S. Günther, W. Preiser, S. Van Der Werf, H.-R. Brodt, S. Becker, H. Rabenau, M. Panning, L. Kolesnikova, R. A. Fouchier, *N. Engl. J. Med.* **2003**, *348* (20), 1967.

[2] A. M. Zaki, S. Van Boheemen, T. M. Bestebroer, A. D. Osterhaus, R. A. Fouchier, *N. Engl. J. Med.* **2012**, *367* (19), 1814.

[3] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, *The Lancet* **2020**, *395* (10223), 497.

[4] P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang, *Nature* **2020**, *579* (7798), 270.

[5] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, *A novel coronavirus from patients with pneumonia in China, 2019. N. Engl. J. Med.* **2020**.

[6] Z. Li, A. C. Tomlinson, A. H. Wong, D. Zhou, M. Desforges, P. J. Talbot, S. Benlekbir, J. L. Rubinstein, J. M. Rini, *eLife* **2019**, *8*, e51230.

[7] D. Wrapp, N. Wang, K. S. Corbett, J. A. Goldsmith, C.-L. Hsieh, O. Abiona, B. S. Graham, J. S. McLellan, *Science* **2020**, *367* (6483), 1260.

[8] M. Hoffmann, H. Kleine-Weber, N. Krüger, M. Müller, C. Drosten, S. Pöhlmann, *BioRxiv* **2020**.

[9] M. A. Tortorici, D. Veesler, *Adv. Virus Res.* **2019**, *105*, 93.

[10] M. A. Tortorici, A. C. Walls, Y. Lang, C. Wang, Z. Li, D. Koerhuis, G.-J. Boons, B.-J. Bosch, F. A. Rey, R. J. de Groot, *Nat. Struct. Mol. Biol.* **2019**, *26* (6), 481.

[11] J. A. Wolff, R. W. Malone, P. Williams, W. Chong, G. Acsadi, A. Jani, P. L. Felgner, *Science* **1990**, *247* (4949), 1465.

[12] G. F. Jirikowski, P. P. Sanna, D. Maciejewski-Lenoir, F. E. Bloom, *Science* **1992**, *255* (5047), 996.

[13] J. J. Suschak, J. A. Williams, C. S. Schmaljohn, *Hum. Vaccin. Immunother.* **2017**, *13* (12), 2837.

[14] S. Guan, J. Rosenecker, *Gene Ther.* **2017**, *24* (3), 133.

[15] N. Pardi, M. J. Hogan, F. W. Porter, D. Weissman, *Nat. Rev. Drug Discovery* **2018**, *17* (4), 261.

[16] O. Takeuchi, S. Akira, *Cell* **2010**, *140* (6), 805.

[17] H. Tanji, U. Ohto, T. Shibata, K. Miyake, T. Shimizu, *Science* **2013**, *339* (6126), 1426.

[18] W. Song, J. Wang, Z. Han, Y. Zhang, H. Zhang, W. Wang, J. Chang, B. Xia, S. Fan, D. Zhang, *Nat. Struct. Mol. Biol.* **2015**, *22* (10), 782.

[19] H. Tanji, U. Ohto, T. Shibata, M. Taoka, Y. Yamauchi, T. Isobe, K. Miyake, T. Shimizu, *Nat. Struct. Mol. Biol.* **2015**, *22* (2), 109.

[20] U. Ohto, T. Shibata, H. Tanji, H. Ishida, E. Krayukhina, S. Uchiyama, K. Miyake, T. Shimizu, *Nature* **2015**, *520* (7549), 702.

[21] D. H. Song, J. O. Lee, *Immunol. Rev.* **2012**, *250* (1), 216.

[22] A. A. Dawood, *New Microbes New Infect.* **2020**, *35*, 100673.

[23] J. Y. Noh, H. W. Jeong, E.-C. Shin, *Signal Transduct. Target. Ther.* **2021**, *6* (1), 1.

[24] E. W. Sayers, J. Beck, E. E. Bolton, D. Bourexis, J. R. Brister, K. Canese, D. C. Comeau, K. Funk, S. Kim, W. Klimke, *Nucleic Acids Res.* **2021**, *49* (D1), D10.

[25] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28* (1), 235.

[26] M. Andreatta, M. Nielsen, *Bioinformatics* **2016**, *32* (4), 511.

[27] J. Sidney, E. Assarsson, C. Moore, S. Ngo, C. Pinilla, A. Sette, B. Peters, *Immunome Res.* **2008**, *4* (1), 1.

[28] S. Tenzer, B. Peters, S. Bulik, O. Schoor, C. Lemmel, M. Schatz, P.-M. Kloetzel, H.-G. Rammensee, H. Schild, H.-G. Holzhütter, *Cellular and Molecular Life Sciences CMLS* **2005**, *62* (9), 1025.

[29] I. A. Doytchinova, P. Guan, D. R. Flower, *BMC Bioinf.* **2006**, *7* (1), 1.

[30] R. Toes, A. Nussbaum, S. Degermann, M. Schirle, N. Emmerich, M. Kraft, C. Laplace, A. Zwinderman, T. Dick, J. Müller, *Exp. Med.* **2001**, *194* (1), 1.

[31] M. C. Jespersen, B. Peters, M. Nielsen, P. Marcatili, *Nucleic Acids Res.* **2017**, *45* (W1), W24.

[32] S. Saha, G. P. S. Raghava, *Proteins Struct. Funct. Bioinf.* **2006**, *65* (1), 40.

[33] S. Saha, G. P. S. Raghava, International Conference on Artificial Immune Systems, 2004, p 197.

[34] I. A. Doytchinova, D. R. Flower, *BMC Bioinf.* **2007**, *8* (1), 1.

[35] P. Ryt-Hansen, E. Hagberg, M. Chriél, T. Struve, A. Pedersen, L. Larsen, C. K. Hjulsager, *Virol. J.* **2017**, *14* (1), 1.

[36] A. Krogh, B. Larsson, G. Von Heijne, E. L. Sonnhammer, *J. Mol. Biol.* **2001**, *305* (3), 567.

[37] A. Bernsel, H. Viklund, A. Hennerdal, A. Elofsson, *Nucleic Acids Res.* **2009**, *37* (suppl_2), W465.

[38] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, *J. Comput. Chem.* **2004**, *25* (13), 1605.

[39] A. Kuriata, A. M. Gierut, T. Oleniecki, M. P. Ciemny, A. Kolinski, M. Kurcinski, S. Kmiecik, *Nucleic Acids Res.* **2018**, *46* (W1), W338.

[40] M. Jamroz, M. Orozco, A. Kolinski, S. Kmiecik, *J. Chem. Theory Comput.* **2013**, *9* (1), 119.

[41] P. Zhou, B. Jin, H. Li, S.-Y. Huang, *Nucleic Acids Res.* **2018**, *46* (W1), W443.

[42] I. T. Desta, K. A. Porter, B. Xia, D. Kozakov, S. Vajda, *Structure* **2020**, *28* (9), 1071.

[43] Y. Yan, D. Zhang, S.-Y. Huang, *J. Cheminf.* **2017**, *9* (1), 1.

[44] S. Y. Huang, X. Zou, *Proteins Struct. Funct. Bioinf.* **2007**, *66* (2), 399.

[45] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, I. A. Vakser, *PNAS* **1992**, *89* (6), 2195.

[46] H. J. Berendsen, D. van der Spoel, R. van Drunen, *Comput. Phys. Commun.* **1995**, *91* (1-3), 43.

[47] T. Darden, D. York, L. Pedersen, *J. Chem. Phys.* **1993**, *98* (12), 10089.

[48] A. C. Wallace, R. A. Laskowski, J. M. Thornton, *Protein Eng. Des. Sel.* **1995**, *8* (2), 127.

[49] A. Grote, K. Hiller, M. Scheer, R. Münch, B. Nörtemann, D. C. Hempel, D. Jahn, *Nucleic Acids Res.* **2005**, *33* (suppl_2), W526.

[50] A. Villalobos, J. E. Ness, C. Gustafsson, J. Minshull, S. Govindarajan, *BMC Bioinf.* **2006**, *7* (1), 1.

[51] A. Carbone, A. Zinovyev, F. Képes, *Bioinformatics* **2003**, *19* (16), 2005.

[52] C. H. Rodrigues, D. E. Pires, D. B. Ascher, *Nucleic Acids Res.* **2018**, *46* (W1), W350.

[53] E. Capriotti, P. Fariselli, R. Casadio, *Nucleic Acids Res.* **2005**, *33* (suppl_2), W306.

[54] M. J. Boniecki, G. Lach, W. K. Dawson, K. Tomala, P. Lukasz, T. Soltysinski, K. M. Rother, J. M. Bujnicki, *Nucleic Acids Res.* **2016**, *44* (7), e63.

[55] Y. Yan, D. Zhang, P. Zhou, B. Li, S.-Y. Huang, *Nucleic Acids Res.* **2017**, *45* (W1), W365.

[56] Z. Chen, X. Zhang, C. Peng, J. Wang, Z. Xu, K. Chen, J. Shi, W. Zhu, *J. Chem. Inf. Model.* **2019**, *59* (8), 3353.

[57] D. Rognan, L. Scapozza, G. Folkers, A. Daser, *Biochemistry* **1994**, *33* (38), 11476.

[58] M. Karplus, J. A. McCammon, *Nat. Struct. Biol.* **2002**, *9* (9), 646.

[59] L. Mlu, N. Bogatyreva, O. Galzitskaia, *Mol. Biol.* **2008**, *42* (4), 701.

[60] C. J. Tsai, R. Nussinov, *Protein Sci.* **1997**, *6* (7), 1426.

[61] X. Cheng, I. Ivanov, H. Wang, S. M. Sine, J. A. McCammon, *Biophys. J.* **2007**, *93* (8), 2622.

[62] E. Y. Klein, D. Blumenkrantz, A. Serohijos, E. Shakhnovich, J.-M. Choi, J. V. Rodrigues, B. D. Smith, A. P. Lane, A. Feldman, A. Pekosz, *Msphere* **2018**, *3* (1), e00554.

[63] M. F. Adasme, K. L. Linnemann, S. N. Bolz, F. Kaiser, S. Salentin, V. J. Haupt, M. Schroeder, *Nucleic Acids Res.* **2021**.

[64] L. R. Baden, H. M. El Sahly, B. Essink, K. Kotloff, S. Frey, R. Novak, D. Diemert, S. A. Spector, N. Rouphael, C. B. Creech, *N. Engl. J. Med.* **2021**, *384* (5), 403.

[65] Y. He, J. Li, S. Heck, S. Lustigman, S. Jiang, *J. Virol.* **2006**, *80* (12), 5757.

[66] K. S. Corbett, D. K. Edwards, S. R. Leist, O. M. Abiona, S. Boyoglu-Barnum, R. A. Gillespie, S. Himansu, A. Schäfer, C. T. Ziwawo, A. T. Di Piazza, *Nature* **2020**, *586* (7830), 567.

[67] K. S. Corbett, B. Flynn, K. E. Foulds, J. R. Francica, S. Boyoglu-Barnum, A. P. Werner, B. Flach, S. O'Connell, K. W. Bock, M. Minai, *N. Engl. J. Med.* **2020**, *383* (16), 1544.

[68] E. J. Anderson, N. G. Rouphael, A. T. Widge, L. A. Jackson, P. C. Roberts, M. Makhene, J. D. Chappell, M. R. Denison, L. J. Stevens, A. J. Pruijssers, *N. Engl. J. Med.* **2020**, *383* (25), 2427.

[69] F. Polack, S. Thomas, N. Kitchin, Safety and efficacy of the BNT162b2 Covid-19 vaccine. *N. Engl. J. Med.* https://doiorg/10.1056/NEJ-Moa2034577. .

[70] B. Peters, M. Nielsen, A. Sette, *Annu. Rev. Immunol.* **2020**, *38*, 123.

[71] P. Cascio, C. Hilton, A. F. Kisselev, K. L. Rock, A. L. Goldberg, *EMBO J.* **2001**, *20* (10), 2357.

[72] B. T. J. Van den Eynde, S. Morel, *Curr. Opin. Immunol.* **2001**, *13* (2), 147.

[73] M. Bhasin, G. Raghava, *Journal of biosciences* **2007**, *32* (1), 31.

[74] O. A. Durojaye, T. Mushiana, S. Cosmas, G. O. Ibiang, M. O. Ibiang, *Egypt. J. Med. Hum. Genet.* **2020**, *21* (1), 1.

[75] D. Castelletti, G. Fracasso, S. Righetti, G. Tridente, R. Schnell, A. Engert, M. Colombatti, *Clin. Exp. Immunol.* **2004**, *136* (2), 365.

[76] G. Clement, D. Boquet, Y. Frobert, H. Bernard, L. Negroni, J.-M. Chatel, K. Adel-Patient, C. Creminon, J.-M. Wal, J. Grassi, *J. Immunol. Methods* **2002**, *266* (1-2), 67.

[77] J. P. Langeveld, J. Martinez-Torrecuadrada, R. S. Boshuizen, R. H. Meloen, J. I. Casal, *Vaccine* **2001**, *19* (17-19), 2352.

[78] P. Cooper, *Parasite Immunol.* **2004**, *26* (11-12), 455.

[79] C. Emanuelsson, M. D. Spangfort, *Mol. Immunol.* **2007**, *44* (12), 3256.

[80] T. Harayama, *Curr. Biol.* **2020**, *30* (3), R122.

[81] A. Mehmood, S. Naseer, A. Ali, H. Fatimah, S. Rehman, A. K. Kiani, *Comput. Biol. Chem.* **2020**, *89*, 107380.

[82] C. A. West, H. Liang, R. R. Drake, A. S. Mehta, *J. Proteome Res.* **2020**, *19* (8), 2989.

[83] V. Bayrami, M. Keyhanfar, H. Mohabatkar, M. Mahdavi, V. Moreau, *Mol. Biol. Res. Commun.* **2016**, *5* (4), 201.

[84] A. T. Albanaz, C. H. Rodrigues, D. E. Pires, D. B. Ascher, *Expert Opin. Drug Discovery* **2017**, *12* (6), 553.

[85] R. Challen, E. Brooks-Pollock, J. M. Read, L. Dyson, K. Tsaneva-Atanasova, L. Danon, *BMJ* **2021**, *372*.

[86] N. London, B. Raveh, O. Schueler-Furman, *Curr. Opin. Struct. Biol.* **2013**, *23* (6), 894.

[87] S.-Y. Huang, *Drug Discovery Today* **2014**, *19* (8), 1081.

[88] K. Fosgerau, T. Hoffmann, *Drug Discovery Today* **2015**, *20* (1), 122.

[89] J. De Ruyck, G. Brysbaert, R. Blossey, M. F. Lensink, *Adv. Appl. Bioinform. Chem.* **2016**, *9*, 1.

[90] A. Borrel, L. Regad, H. Xhaard, M. Petitjean, A.-C. Camproux, *J. Chem. Inf. Model.* **2015**, *55* (4), 882.

[91] C. Iavarone, D. T. O'hagan, D. Yu, N. F. Delahaye, J. B. Ulmer, *Expert Rev. Vaccines* **2017**, *16* (9), 871.

[92] M. Doel, N. Carey, *Cell* **1976**, *8* (1), 51.

[93] Y. Wang, Z. Zhang, J. Luo, X. Han, Y. Wei, X. Wei, *Mol. Cancer* **2021**, *20* (1), 1.

[94] C. L. Bell, D. Yu, C. D. Smolke, A. J. Geall, C. W. Beard, P. W. Mason, *Virol.* **2015**, *483*, 302.

[95] C. Zhang, G. Maruggi, H. Shan, J. Li, *Front. Immunol.* **2019**, *10*, 594.

[96] M. Meyer, E. Huang, O. Yuzhakov, P. Ramanathan, G. Ciaramella, A. Bukreyev, *J. Infect. Dis.* **2018**, *217* (3), 451.

[97] K. A. Whitehead, J. E. Dahlman, R. S. Langer, D. G. Anderson, *Annu. Rev. Chem. Biomol. Eng.* **2011**, *2*, 77.

[98] N.-N. Zhang, X.-F. Li, Y.-Q. Deng, H. Zhao, Y.-J. Huang, G. Yang, W.-J. Huang, P. Gao, C. Zhou, R.-R. Zhang, *Cell* **2020**, *182* (5), 1271.

[99] M. Avci-Adali, A. Behring, H. Steinle, T. Keller, S. Krajeweski, C. Schlensak, H. P. Wendel, *J. Vis. Exp.* **2014**, (93), e51943.

[100] B. Scheel, S. Braedel, J. Probst, J. P. Carralot, H. Wagner, H. Schild, G. Jung, H. G. Rammensee, S. Pascolo, *Eur. J. Immunol.* **2004**, *34* (2), 537.

[101] M. Fotin-Mleczek, K. M. Duchardt, C. Lorenz, R. Pfeiffer, S. Ojkic-Zrna, J. Probst, K.-J. Kallen, *J. Immunother.* **2011**, *34* (1), 1.

Submitted: November 2, 2021
Accepted: February 4, 2022