Identification of vaccine targets & design of vaccine against SARS-CoV-2 coronavirus using computational and deep learning-based approaches.

Bilal Ahmed Abbasi [1,#], Devansh Saraf [1], Trapti Sharma [1], Robin Sinha [1], Shachee Singh [1], Shriya Sood[1], Pranjay Gupta[1], Akshat Gupta[1], Kartik Mishra[1], Kamal Rawal [1,#,*]

1. Amity Institute of Biotechnology, Amity University Uttar Pradesh, India.

***Corresponding Author**

#Equal Contribution

Email Id: kamal.rawal@gmail.com

Center for Computational Biology and Bioinformatics, AIB

Amity University, Noida.

Keywords: SARS-CoV-2, Reverse Vaccinology, Epitopes, Vaccine-designing, Deep Learning, Network biology.

Supplementary Data: https://sites.google.com/view/corona7000/home

Abstract:

An unusual pneumonia infection, named COVID-19, was reported on December 2019 in China. It was reported to be caused by a novel coronavirus which has infected approximately 8.7 million people worldwide with a death toll of 463000 till date. This study is focused on finding potential vaccine candidates and designing an *in-silico* subunit multi-epitope vaccine candidates using a unique computational pipeline, integrating reverse vaccinology and molecular docking methods. A protein named SARS-CoV-spike [S] protein of SARS-CoV-2 having GenBank ID- **QHD43416.1** was shortlisted, as a potential vaccine candidate and was examined for presence of B-cell and T-cell epitopes. We also investigated antigenicity and interaction with distinct polymorphic alleles of the epitopes. High ranking epitopes/peptides such as **DLCFTNVY** (B cell class), **KIADYNKL** (MHC Class-I) and **VKNKCVNFN** (MHC class-II) were shortlisted for subsequent analysis. Digestion analysis verified the safety and stability of the shortlisted peptides. Docking study reported a strong binding of proposed peptides with HLA-A*02 and HLA-B7 alleles. We used standard methods to construct vaccine model and this construct was evaluated further for its antigenicity, physicochemical properties, 2D and 3D structure prediction and validation. Finally, the vaccine construct was reverse transcribed and adapted for *E. coli* strain K 12 prior to the insertion within the pET-28-a (+) vector for determining translational and microbial expression. Also, six multi-epitope subunit vaccines were constructed using different strategies containing immunogenic epitopes, appropriate adjuvants and linker sequences. We propose that our vaccine constructs can be used for downstream investigations using *in-vitro* and *in-vivo* studies to design effective and safe vaccine against COVID19.

1. Introduction

Coronavirus belongs to a large family of viruses called “Coronaviridae” (order Nidovirales) which are characterised by crown-like spikes (Figure 1) on their surface and usually infect the respiratory system of humans and other vertebrates. The epidemiological studies indicate the viral transmission from animal to human and thereafter from seeding clusters of human-human transmissions with the reproduction number (R_0) ranges between 2.2- 2.9 for humans (1).

Coronaviruses can come under any of the 4 genera: Alphacoronavirus, Betacoronavirus Gammacoronavirus, and Deltacoronavirus. The first incidence of human coronaviruses can be traced back to the mid-1960s. In the recent past, scientists have identified 7 sub-types of the coronavirus that are known to cause infection in human beings. These include 229E (alpha coronavirus); NL63 (alphacoronavirus); OC43 (betacoronavirus); HKU1 (betacoronavirus); MERS-CoV (the betacoronavirus that causes MERS); SARS-CoV (betacoronavirus causing SARS) and SARS-CoV-2 (n-2019-CoV, betacoronavirus). The first four viruses cause infection in the upper section of the respiratory tract that results in a mild infection while the last three viruses affect the lower section of the respiratory tract and result in severe respiratory syndrome in human beings (2).

SARS-CoV-2 is the most recently evolved coronavirus that was first reported in Wuhan, China, which led to a mysterious pneumonia-like disease in humans and has been named COVID-19 by WHO. It has an incubation period of 4-7 days (3). The pandemic, as of June 20, 2020 has resulted in more than 8.7 million cases worldwide and a death toll of approximately 463000. The worst hit nations are the USA, UK, Brazil, Italy, France, and Spain; all having crossed more than 20,000 deaths with the USA having more than 110,000 deaths (4).

The epidemiological studies have shown the Huanan seafood market to be the source of this outbreak, indicating an animal-to-human route, also known as zoonosis, as the prime transmission mode (5). Similar outbreaks in 2002-03 and in 2012 of Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS), have shown a fatality rate of ~10% and ~35% respectively. SARS and MERS viruses were known to transmit from animal-to-human (6). Due to this reason, extensive studies were conducted to understand the transmission of viral infections in humans and animals.

At the molecular level, coronaviruses are non-segmented, enveloped, positive-sense single stranded RNA viruses (~30kb), having a 5’ cap and 3’ poly-A tail. This virus propagates by

forming a replication-transcription complex (RTC) using its gRNA as a template. The RTC further encodes all the structural and non-structural proteins required for viral propagation. The viral genome is found to contain 6 ORFs. The first ORF (ORF1a/b) encodes 16 non-structural proteins and the rest encodes the 4 main structural proteins: spike (S), membrane (M), envelope (E) and nucleocapsid (N) (7,8).

Presently, scientists have submitted 49721 genomes of SARS-CoV-2 in GISAID (Global Initiative on Sharing All Influenza Data) and one of these has been released on GenBank with Accession ID: MN908947. In a recent phylogenetic study by Jiang et al, SARS-CoV-2 was found to be very similar to the bat SARS-like coronavirus, with 89 % similarity at genomic level (9).

2. Materials and Methods:

2.1 Data Acquisition:

Severe Acute Respiratory Syndrome Coronavirus-2 or SARS-CoV-2 isolate Wuhan-Hu-1, complete genome (Accession ID: NC_045512), and its coding sequences were retrieved from NCBI database (10) in FASTA format.

Crystal structures of human alleles, HLA-A*02 (PDB ID: 6O4Y) (11) and HLA-B7 (PDB ID: 3VCL) (12) were retrieved from Protein Data Bank (PDB) (13) to conduct the binding affinity studies with the predicted epitopes. HLA-A*02 was selected due to its presence in the majority of population in Wuhan region whereas HLA-B7 was selected as it is one of the predominant alleles in the world (14)**.**

Peptide sequences of three different adjuvants were extracted from NCBI Database. These sequences includes L7/L12 50s ribosomal protein (Accession ID: WP_088359560.1, Flavobacteriaceae bacterium JJC), β-defensin and HABA proteins (Accession ID: AGV15514.1; M. tuberculosis).

2.2 Workflow: Flow chart representation showing the workflow adopted has been made (Figure 2) and the whole approach is summarised in subsequent sections.

2.2.1 Identification of Exposed Surface Proteins:

Among the proteins encoded by pathogens (including viruses and bacteria), the surface and secretory proteins play important roles in the pathogenesis process, which include alterations in the host cell to the advantage of the pathogen, adhesion & invasion of host cell, host cell toxicity and defence against the host-immune response. Furthermore, the outer membrane proteins of the pathogen are involved in interactions with B-cells and Antigen Presenting Cells (APCs) (15)**.** These attributes of the surface and secretory proteins make them attractive drug & vaccine targets. Retrieval and selection of the outer cellular membrane proteins for the purpose of vaccine design and construction was performed using the state-of-the-art pipeline called VaxELAN. The pipeline uses three different tools namely: - CELLO (16), Virus-mPloc tool (17) and PSORTb (18) to determine the location of a given protein.

2.2.2 Trans-Membrane (TM) Analysis:

Several studies have indicated that it is difficult to purify proteins with more than one transmembrane helix, so it seems reasonable to exclude these proteins from the selection process (19). Therefore, tools such as HMMTOP (20), TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) and TMPred (21) were used to screen those proteins that have less than or equal to 1 transmembrane alpha-helices in their structure.

2.2.3 Non-Homology Analysis:

Those viral proteins which are dissimilar to human proteins are considered to be good vaccine candidates since the vaccines based upon these proteins would minimize any kind of side-effect and cross-reactivity.

To find such proteins, SARS-CoV-2 proteome was screened against the proteome of *Homo sapiens* (NCBI Database) using the BLASTp tool (22). The proteins having ≥35% identity, query coverage ≥35 % and E value <10e-5 were filtered.

2.2.4 Physicochemical Property Analysis:

With the help of ProtParam tool (23)**,** various physicochemical properties of viral proteins were computed. Based on these properties, those proteins were selected which were predicted to be stable in nature (i.e. instability index less than 40).

2.2.5 Antigenicity Prediction:

Doytchinova and Flower had proposed an alignment free approach in their VaxiJen v2.0 server (24); which is based on auto cross covariance (ACC) transformation of protein sequences into uniform vectors of principal amino acid properties. Using this approach, proteins whose antigenicity scores were greater than the threshold value of 0.4 were selected for further evaluation.

2.2.6 Adhesion Prediction:

Adhesin proteins play a significant role in the establishment of pathogen-based infections. Therefore, targeting the adhesin and adhesin-like proteins in vaccine development can help in combating such infections by blocking their function and preventing their adherence to the host cells (25). To achieve this objective, a tool named FungalRV (26) with the threshold value of greater than or equal to -1.2 was employed. Though this tool was developed using proteins drawn from the fungal system, still, it provides a detailed analysis as well as clues for effective vaccine design.

2.2.7 Non-Allergenicity Analysis:

Vaccines, just like drugs, also have the potential to cause allergic reactions. Therefore, it is important to check if the protein candidate acts as an allergen or not. In this study, a resource named AllergenOnline (27) was used for the identification of proteins having potential allergenic action. Here, only those proteins were selected which were labelled as non-allergen using BLASTp tool against the AllergenOnline database.

2.2.8 Evaluation of Filtered Protein:

Physicochemical characterization of the shortlisted protein was performed using ProtParam and DiANNA (28) tools. Protparam computes half-life, amino acid atomic composition, GRAVY (Grand average of hydropathicity), molecular weight and instability index. DiANNA is a neural network-based prediction system which was used to find the existence of disulphide-bonds in the viral proteins before subjecting them to B-cell and T-cell epitope predictions.

2.2.8.1 B-Cell Epitope Prediction:

B-cell epitope prediction is performed to identify any surface-exposed regions in an antigen that can interact with an antibody. The primary sequence of the selected protein [QHD43416.1; Surface glycoprotein or Spike S protein (SARS CoV-2)] was examined using BcePred server (29) for prediction of continuous B-cell epitopes. Parameters including antigenicity, accessibility of surface, flexibility and hydrophilicity were also determined.

Antigenic propensity and conservancy rate using IEDB Conservancy analysis tool (30) were also measured. Next, shortlisted epitopes were subjected for antigenicity evaluation using VaxiJen Server.

2.2.8.2 T-Cell Epitope Prediction:

The T-cell epitope prediction was performed to identify those immunogenic peptides of an antigen that can stimulate CD4+ (HTL, Helper T- Lymphocyte) and CD8+ (CTL, Cytotoxic T-Lymphocyte) cells.

T-cells operate by recognizing the antigen as peptides which are associated with major histocompatibility complex (MHC) molecules (31). Cytotoxic T lymphocytes (CTL or CD8+ cells) curbs proliferation of antigens in the body by directly killing the viral infected cells or secreting antiviral cytokines.

Tools such as ProPred-I (for MHC class-I alleles binding epitopes) (32) and ProPred (for MHC class-II alleles binding epitopes) (33) were used for T-cell epitope prediction. Using ProPred-I, filters were applied for proteasomal and immuno-proteasomal cleavages on the predicted MHC binding peptides (34). Finally, only the high-scoring unique epitopes with 100% conservancy rates were considered in subsequent analysis. Furthermore, these epitopes were also subjected to toxicity analysis using the ToxinPred server (35).

2.2.9 Structural Modelling and Molecular Docking:

Molecular docking is used to investigate the interaction of the predicted peptides with the MHC molecules using binding energies and contact residues (36). With the help of PEP-FOLD (37) server at RPBS MOBYLE (38) portal, the 3-D structure of the predicted peptides was determined. Next, 3D structure of human allele HLA A*02 (crystallized at the resolution of 1.58 Angstrom) was retrieved from PDB (ID: 6O4Y). Since allele HLA A*02, is found mostly

in the population of Wuhan, therefore we used the 6O4Y structure for docking studies using the HPEPDOCK server (39). HLA-B7 protein structures were also used for comparative studies.

2.2.10 Construction of Final Vaccine:

Six potential multi-subunit vaccines against COVID-19 were constructed by using high-scoring CTLs, HTLs, and B-cell epitopes. The immunogenic peptides of length 9-12 amino acids were obtained from the shortlisted Spike protein and merged together to formulate the vaccine candidates using distinct strategies. To differentiate between various constructs, the constructed vaccine sequences were labelled as V1, V2, V3, V4, V5 and V6. The strategy for constructing V1, V2 and V3, has been discussed in this section while the strategies of V4-V6 constructs is described in the (Supplementary File A).

Each sequence starts from a distinct adjuvant sequence namely β- defensin, L7/L12 50s ribosomal protein and HABA protein, respectively. Each of these adjuvants have been reported to accentuate protective immune response (40).

The adjuvant was linked to the first CTL epitope via EAAAK linker and all the CTL epitope repeats were linked with each other by the GGGS linker. Conjugation of the CTL epitope with the HTL epitope and the HTL epitope repeats among themselves was carried out using AAY linker, whereas, conjugation of the HTL epitope with the B-cell epitope and B-cell epitope repeats among themselves was performed using KK linker. (Figure 3).

To determine the order of different components in the vaccine construct, information previously reported in studies namely Ebola virus (41), Avian influenza A (H7N9) (42), Monkeypox virus (43) and *Marburg marburgvirus* (44) was utilised.

2.2.11 Antigenicity, Allergenicity, Solubility and Physicochemical Analysis of Vaccine Constructs:

Antigenicity of vaccine constructs or chimeric protein was evaluated using the Vaxijen v2.0 server with a threshold of 0.4. Further, non-allergenic nature of all the constructs was evaluated using Algpred (45). This tool incorporates methods based on SVM, motif searching, and BLAST searches on allergen representative peptides (ARPs). The solubility of these constructs was also determined using Protein-Sol. (46).

Moreover, various physicochemical properties of the vaccine constructs were also determined using the ProtParam tool including their isoelectric pH, GRAVY values, molecular weight, instability index, estimated half-life, and aliphatic index.

2.2.12 Secondary and Tertiary Structure Prediction:

Secondary structures of the constructed vaccines were obtained using PSIPRED (47) server and the structural composition was determined by employing CFSSP (48) server. The tertiary structure of the vaccine constructs was predicted by using the I-TASSER server (49).

2.2.13 3D Structure Refinement and Validation:

The I-TASSER server predicted the 3-D model of the vaccine construct. It is based on a hierarchical approach for predicting high resolution protein structure and function. Among all the predicted 3D models of a vaccine construct, the model having the highest C-score was selected. Refinement of the predicted model was performed to improve its accuracy by using online refinement tools namely, ModRefiner (50) and 3Drefine (51). The refined protein structure was further validated using the Ramachandran plots generated by online tool RAMPAGE (http://mordred.bioc.cam.ac.uk/~rapper/rampage.php).

2.2.14 Molecular Docking of Subunit Vaccine with Immune Receptor:

Molecular docking analysis is an essential tool for determining the interaction between a receptor and ligand molecule. The binding affinity of the vaccine construct with human toll-like receptors (TLR-8) was determined via several online docking servers. These servers include HDOCK server (52), ClusPro 2.0 server (53) and PatchDock (54). The models obtained by PatchDock were further refined by FireDock (55).

The 3D structures of TLR-8, obtained from RCSB protein data bank, were used to analyse a desirable protein- protein complex in terms of better electrostatic interaction and binding energy.

2.2.15 Characterisation of Immune Profile of the Construct:

To simulate the real response of an immune system to our final vaccine construct, the C-immSim immune server was employed (56). This is a freely accessible web-server (http://150.146.2.1/C-IMMSIM/index.php) that works on the basis of Position-Specific Scoring Framework (PSSM) to simulate and predict immune interactions along with

immunogenic epitopes. The tool was run on default parameters with time step of injection being 1, 42, 84 i.e. three times and vaccine injection without LPS (Lipopolysaccharide).

2.2.16 Codon adaptation and *in silico* cloning of the chimeric protein.

Java Codon Adaptation Tool or JCat server was employed (http://www.jcat.de/) (57) for the purpose of *in silico* codon adaptation in model organism *E. coli* strain K12 for the expression of protein vaccine. Vaccine constructs were reverse transcribed to possible DNA sequence and filters were applied to avoid rho-independent transcription termination, prokaryotic ribosome binding sites and cleavage sites of various restriction enzymes (BamHI and XhoI). The reverse-transcribed DNA sequence (RT-DNA) thus obtained is conjugated with XhoI and BamHI restriction sites at N-terminal and C-terminal sites, respectively. Next this adapted DNA sequence is incorporated into the multiple cloning site (MCS) of pET-28a(+) vector using the SnapGene tool (58).

2.3 AI in Potential Vaccine Detection:

100 proteins were extracted and labelled as positive dataset- which were reported to be antigenic candidates using text mining and deep curation strategies (59). Similarly, various control datasets, labelled as negative datasets were constructed which consist of proteins not known to produce any immune response in the host system. Subsequently, several bioinformatics, reverse vaccinology and immunoinformatics tools such as PSORTb, FungalRV, SignalP, TargetP, IEDB, BLASTp, ProtParam, Vaxijen, etc. were utilised to characterise proteins into positive and negative datasets. Thereafter, distributions of scores as well as ROC curves were generated to determine the cut-off.

Further, each protein was converted into a feature vector. Next, the data was normalised using min-max normalisation function. This step was followed with training of the algorithm on two datasets: Model-1 was trained on viral proteins and Model-2 was trained on bacterial proteins. Thereafter, an LSTM network was constructed which consisted of two LSTM nodes, along with 2 fully connected nodes with leaky Relu activation function and a single fully connected node with sigmoid function as an output layer. Each hidden node in the network has weight and bias maximum normalize constraint of value 3 and was regularized using L2 regularization function to prevent overfitting during training. Cross validation was performed, and the dataset was divided into two parts. The first part had 170 equally weighted examples as used during

training and 30 examples were used in testing or cross validation purposes (https://sites.google.com/view/corona7000/home).

2.4 Viral-Host Protein Interactions:

The interactions of spike glycoprotein with other host proteins were also investigated using String (v11.0 protein-protein interaction database) (60). Since, SARS-CoV-2 interaction data was not available, hence, the data derived from Coronavirus 229E (NCBI taxonomy Id: 11137), Human SARS coronavirus (NCBI taxonomy Id: 694009) and Homo sapiens (host) (NCBI taxonomy Id: 9606) was used.

3. Results:

In this study, several computational strategies such as reverse vaccinology, deep learning and immunoinformatics tools were used to find the most suitable protein vaccine candidate against SARS-CoV-2. Using the above-mentioned approaches, a protein named as spike S (Surface glycoprotein) was shortlisted, and B-cell and T-cell epitopes were predicted for the construction of an epitope-based vaccine.

3.1 Reverse Vaccinology Pipeline:

Out of the 10 proteins of SARS-CoV-2, the integrated pipeline shortlisted one protein (Spike S- Surface Glycoprotein with accession no. **QHD43416.1**) as a potential vaccine candidate.

The physicochemical properties of this protein were predicted by ProtParam and DiANNA (Table S1). The description of secondary structure was predicted by PSIPRED (Table S2).

3.2 Recognition of B-cell Epitopes:

B-cell epitopes play a crucial role in the activation of B-cell mediated immune response against viral infections. The BcePred server was utilised to predict the continuous B cell epitopes. A total of 41 B-cell epitope sequences were predicted using BcePred. Physicochemical parameters like hydrophilicity (61), exposed surface (62), turns (63), accessibility (64), flexibility (65) and antigenic propensity (66) were also evaluated for prediction of linear epitopes. Furthermore, the IEDB conservancy tool was used to evaluate the predicted epitopes.

Out of these, only 15 peptides were predicted to be highly antigenic in nature determined by the Vaxijen server. For instance, a peptide "**DLCFTNVY**" is predicted to be the highest-ranking peptide (with a score: 1.85) amongst the other shortlisted peptides (Table 1).

3.3 Recognition of T-cell Epitopes:

3.3.1 MHC- I Allele Binding T-cell Epitopes:

The ProPred-I tool was used to predict MHC-I binding T-cell epitopes (Table 2). Using Proteasome and ImmunoProteasome filters set at the threshold of 5%, all alleles were selected and only the top 10 peptides were chosen to be displayed by the ProPred-I server result.

Only peptides with 100% conservancy rate were considered. Out of the 46 predicted MHC class-I binding epitopes, 45 epitopes were found to be conserved. For instance, we found that **KIADYNYKL** has the highest antigenicity score of 1.66 and binds to a number of alleles including HLA-A2, HLA-A*0201, HLA-A*0205, HLA-A3, HLA-B*0702 (Table S3). Physicochemical properties of top 8 selected epitopes were obtained by ToxinPred (Table S4).

3.3.2 MHC- II Allele Binding T-cell Epitopes:

The ProPred tool was used to predict the MHC-II binding T-cell epitopes (Table 2). Among the 94 predicted epitopes, 90 were found to have 100% conservancy rate. Out of which, **VKNKCVNFN** was found to have the highest antigenicity score of 2.05 and binds to several alleles (Table S5). The physicochemical properties of top 8 selected epitopes were obtained by ToxinPred (Table S6).

3.4 Structural Modelling and Molecular Docking:

The 3D structure of MHC class-I epitopes was predicted using PEP-FOLD. Molecular docking is a vital tool to understand protein-peptide interaction. Top 4 antigenic CTL epitopes were docked against various Human Leukocyte Antigen (HLA) using the web HPEPDOCK server with default settings (Table 3, Figure 4).

3.5 Construction of vaccine:

High scoring Linear B cell, CTL and HTL epitopes were used to construct multi epitope vaccines. Adjuvant sequences were used for enhancing immune interaction by utilizing its advantageous feature to act as an agonist and perform a significant part in improving the efficacy of vaccines (67). (Supplementary File A)

3.6 Antigenicity, Allergenicity, Solubility and Physicochemical Analysis of Vaccine Constructs:

Vaccine construct V1 was predicted to be highly antigenic (Score 1.16 using Vaxijen web server) (Table 4). In addition, V1 was also predicted to be non- allergenic by the Algpred tool. The solubility value of V1 was estimated to be 0.71 by Protein-sol tool (threshold value of 0.45) indicating that the constructed vaccine is more soluble than average soluble E. coli protein from the hypothetical dataset utilised by that tool. The molecular weight of the construct V1 with beta-defensin as an adjuvant (326 amino acids) was estimated to be 36.83 kDa with a theoretical isoelectric point value (pI) of 9.58. The half-life was estimated at 30 hrs in mammalian reticulocytes in vitro, and more than 20 hours in yeast and more than 10 hours in *E. coli in vivo*. The instability index (II) was estimated at 9.76, indicating that the vaccine is stable (Threshold II less than 40 indicates stability). The predicted aliphatic index was calculated to be equal to 67.61, indicating the thermostability of the proposed vaccine (68). The predicted hydropathicity came to be -0.467 which denotes the vaccine construct V1 is hydrophilic in nature and can bind with molecules of water (69). The information regarding these parameters for remaining constructs can be retrieved from (Supplementary File B).

3.7 Secondary Structure prediction:

The secondary structure of the vaccine candidate V1 was predicted by PSIPRED (Figure 5). It was predicted to have 62.9% helix, 29.8 % beta-sheets and 12.6% turns by using the CFSSP tool (For secondary structure of V2-V6 see Supplementary File C).

3.8 Tertiary Structure prediction, Refinement and Validation:

A total of five 3D models (tertiary structures) of the vaccine construct V1 were predicted by I-TASSER server based on 10 best threading templates namely, 6buaA, 3du1X, 1kj6, 1kj6, 4plaA, 1kj6, 2xtwA, 1kj6A, 1kj6A and 4n9nA as identified by LOMETS (70) from the PDB library. These best templates were selected from the LOMETS threading programs using the Z- score values during the I-TASSER modelling. The 5 models thus predicted had C-score values ranging between -3.36 and -4.19. Since the C score normally ranges from -5 to 2, with a higher value indicating higher confidence, the model with the highest C-score (Model 4 in case of vaccine construct V1 has highest C-score of -3.36) was chosen for further refinement by online refinement tool ModRefiner followed by 3Drefine. The refined 3D models of all vaccine constructs were validated by referring to the Ramachandran plot generated using

RAMPAGE **(**http://mordred.bioc.cam.ac.uk/~rapper/rampage.php). The results for remaining vaccine constructs are available in supplementary data. The Ramachandran plot assessment of V1 predicted 73.8%, 18.8% and 7.4% residues to be in favoured, allowed and outlier regions, respectively. (Figure 6).

3.9 Molecular Docking of Subunit Vaccine with Immune Receptor:

With docking analysis, the binding affinity between the chimeric vaccine construct and Toll-like receptor (TLR-8) were studied. Various online tools for protein-protein docking were employed such as HDOCK, ClusPro 2.0, and PatchDock server.

The ClusPro server produced 30 protein-ligand complexes with their corresponding free binding energy as output. The lowest energy of -1277.5 kcal/mol was obtained for the complex 2 that indicates spontaneous binding between the Toll-Like Receptor and the vaccine component. The HDOCK server predicted the binding energy for the protein-peptide complex as -330.04. The PatchDock generated a range of solutions, and among them, the docking assembly with the highest negative atomic contact energy (ACE) value was selected for analysis. The ACE value of the docking complex was -353.27 for solution 36 which were further evaluated for refining the complexes using FireDock, which gives the ACE value and lowest Global energy of the refined model to be 1.28 and -38.62 respectively, as obtained for solution 9. (Figure 7).

3.10 Characterisation of immune profile of the construct:

C-ImmSim simulator was used to analyse the immune response produced by the final vaccine construct. The tool generated the immune response simulations that match the response of a real immune system. Results of simulated immune responses indicate an increased surge in the induction of secondary immune responses. B-cell population surge was observed during secondary and tertiary responses which was accompanied with rise in the levels of IgM, IgG1 + IgG2, and IgG + IgM along with the reduction in the antigen concentration (Figure 8).

3.11 Codon Adaptation and *in silico* Cloning of the Chimeric Protein:

JCat (Java Codon Adaptation Tool) was used for the optimization of codon of chimeric protein construct in *E. coli* (K12). It turned out that the optimized codon sequence has a length of 978 nucleotides and its CAI (Codon Adaptation Index) was predicted to be 1.00, with an average

of 41.21 % GC content (Optimal range lies between 30% to 70%) for the adapted sequence. These resultant values act as determining properties indicating potentially stable expression of the constructed vaccine in the selected microbial host.

For optimal gene expression, SnapGene software was employed to incorporate the adapted DNA sequence of the designed chimeric protein vaccine V1 into the E.coli pET-28a(+) vector by adding restriction sites which were followed by cloning of genetic sequence into the vector (Figure 9).

3.12 Multi-Layered Network of ACE2 and Spike S Protein:

Construction of network of interacting partners of Spike S proteins (Virus) and Human Interacting proteins was also performed which was used to make a multilayer network between Angiotensin Converting Enzyme 2 (ACE2) protein (Human) to study the viral host interaction (71) (Supplementary File D).

Since ACE2 is a critical molecule and a potent regulator of blood pressure, body fluids and electrolyte homeostasis (72). Further, it was also reported that loss of ACE2 accelerates the diabetic kidney injury (73). Studies have indicated that ACE2 displays strong interaction with dipeptidyl peptidase-4 molecules. Dipeptidyl peptidase-4 (DPP4) have been shown to play a significant role in T-cell receptor (TCR)-mediated T-cell activation. Importantly, Raj et al have shown that DPP4 is an emerging functional receptor for hCoV-EMC (74).

In the recent outbreak, it was reported that 60 % of hospitalised patients had one or more co-existing conditions such as hypertension, cardiovascular and diabetes (75). In several studies, it has been reported that people with obesity are at a significant risk factor to suffer from complications due to COVID-19. Further, the relationship between obesity and mortality rate due to COVID-19 was investigated. We also found diseases associated with obesity such as type-2 diabetes, cardiovascular diseases and hypertension are also linked with poor prognosis in COVID-19 (76,77). We also found that several molecules implicated in obesity, diabetes, and hypertension, appear to show interactions with SARS-CoV-2 proteins as well as human proteins (ACE2 and DPP4). Thus, it might be possible to correlate the high rate of mortality of COVID-19 patients with comorbidities such as obesity, diabetes, and hypertension due to involvement of common sets of molecules. Further studies (genomic, molecular etc) are warranted to test the hypothesis about selective disadvantage of patients suffering from metabolic syndrome X in context of coronavirus.

3.13 Randomisation Experiments:

To find out the biological significance of the predicted epitopes, random shuffling experiments were performed for constructing different vaccine models to compare their solubility, antigenicity allergenicity and physicochemical properties.

This was achieved using two different methods: Firstly, it includes shuffling of the shortlisted protein [**QHD43416.1**] and another is the shuffling of the final epitopes using Sequence Manipulation Suite (78).

This was followed by comparing their physicochemical properties with the original proposed vaccine candidate. In results, it was found that the vaccine obtained from the shuffling experiment has different properties from the original proposed one; antigenic scores of epitopes from the shuffled protein sequence were significantly different from the antigenic scores of epitopes predicted from the original protein sequence (Supplementary File E). Further, randomization of the top-ranking epitopes is conducted. Also, computation of the interaction score of peptides with the molecular docking process is done to study the impact of shuffling on binding scores.

4. Discussion:

As the world is embracing a crisis, the computational community is playing an important role in fighting against the pandemic (79–81). With the recent advancements of *in-silico* based approaches and sequence-based technology; a collection of proteomic and genomic data of viral pathogens have been possible. This has made feasible the designing of peptide vaccines based on neutralizing epitopes. The vaccinomics strategy has been extensively studied and applied in Avian influenza A (H7N9), Monkeypox virus, Ebola virus and *Marburg marburgvirus*.

This study incorporates reverse vaccinology, bioinformatics, immunoinformatics and AI-based strategies to build a computational framework for identifying probable vaccine candidates and constructing an epitope-based vaccine against COVID-19. The framework consists of identifying surface-exposed proteins, transmembrane helices analysis, Non homology to humans, Instability analysis, antigenicity analysis, adhesion prediction and allergenicity analysis.

The screening of viral proteome sequences resulted in shortlisting of Spike protein or Surface Glycoprotein of SARS-CoV-2 (Accession ID. **QHD43416.1)** as a potential protein target that can be used to design the vaccine.

The Spike protein plays an integral role in the SARS CoV-2 life cycle by cleaving into S1 glycoprotein (N-terminal) and S2 glycoprotein (C-Terminal) and exhibiting high amounts of glycosylation. S1 glycoprotein attaches the virion to the cell membrane by interacting with the host receptor, which neutralizes the antibodies in the host environment, thus causing infection. Also, S1 glycoprotein mediates the conformational changes in protein structure. The S2 glycoprotein is used in mediating the fusion of virion and cell membranes by enacting the role of class 1 viral fusion protein.

The shortlisted protein was subjected to computation of various physicochemical properties like number of amino acids, GRAVY value, extinction coefficient, molecular weight, instability index, theoretical pI, aliphatic index and cysteine disulfide bond score.

Tools namely ProPred, ProPred-I and BcePred were employed for the determination of all the possible epitopes for T cells and B cells. B and T lymphocytic cells play an important role in developing acquired immunity. The antigens, after being recognised by APC (Antigen Presenting Cells) are presented via MHC- II molecule to helper-T cells which further activates B-cells. The B-cells produce antibodies whereas T-helper cells also activate macrophages and cytotoxic T-lymphocytes. All these epitopes were found in Receptor Binding Domain (RBD) within SARS-CoV-2 S protein (Figure S1).

In a study by Ong et al (82), authors investigated entire proteome, including the S protein and five non-structural proteins (nsp3, 3CL-pro, and nsp8-10) and labelled them as adhesins, which are crucial to the viral adhering and host invasion. They also found nsp3 to be more conserved among SARS-CoV-2, SARS-CoV, and MERS-CoV than among 15 coronaviruses infecting human and other animals. The protein was also predicted to contain promiscuous MHC-I and MHC-II T-cell epitopes, and linear B-cell epitopes localized in specific locations and functional domains of the protein. They also used a pipeline called Vaxign-ML for target predictions.

Using immunoinformatics and docking studies, Bhattacharya et al (83) have identified potential epitopes and docking complexes of constructed vaccines and TLR5. Another group of scientists have also identified a set of B-cell and T-cell epitopes derived from the spike (S) and nucleocapsid (N) proteins that map identically to SARS-CoV-2 proteins under the

assumption that no mutation was seen in limited dataset of 120 available SARS-CoV-2 sequences (as of 21 February 2020). This assumption of zero mutation rate has changed in the light of new data submitted since February 2020.

Robson et al reported a specific sequence motif "KRSFIEDLLFNKV" as a conserved and interesting target. He also reported that this region is associated closely with known cleavage sites of the SARS virus that are believed to be required for virus activation for cell entry (84). In another study, Grifoni et al (85) used bioinformatics approaches to identify *a priori* potential B and T cell epitopes for SARS-CoV-2 using IEDB resources. They also described immune-dominant regions located in the S1 subunit in the CTD2 and CTD3 (C-terminal domain), and in the HR1 domain of the S2 subunit. Kiyotani et al (86) comprehensively screened potential SARS-CoV-2-derived, HLA-class I- and II-presented epitopes for 43 *HLA* alleles that are common in the Japanese population, and identified 2013 and 1399 epitopes, respectively. They found that 781 HLA-class I and 418 HLA-class II epitopes were common between SARS-CoV-2 and SARS-CoV. Researchers have tested 15 epitope-HLA-binding prediction tools, and using an *in vitro* peptide MHC stability assay, and assessed 777 peptides that were predicted to be good binders across 11 MHC allotypes (87). A research group recently found a cross-protective epitope between the spike proteins of SARS-CoV-2 and SARS-CoV, and successfully found the cross-protective epitopes in the RBDs of the spike proteins (88). Further, another study found that the spike RBD of SARS-CoV-2 bound potently to angiotensin-converting enzyme 2 (ACE2), the host cell receptor of SARS-CoV (89).

Studies indicate that HLA variations are associated with susceptibility or resistance to malaria, tuberculosis, leprosy, HIV, and hepatitis virus persistence (90). A report also suggests that human coronavirus OC43 interacts with HLA class I molecules at the cell surface to establish infection (91). Further, a study (92) indicates the association of HLA-B* 4601 with the severity of SARS infection in Asian population. In our work, we employed computational strategies (i.e. molecular docking) to check interaction of viral peptides with the commonly found human allele (HLA*B7) and Wuhan region (HLA*A2).

Furthermore, *in vitro* and *in vivo* study should be conducted to confirm the safety and potency of the predicted vaccine candidates. We suggest further wet lab-based studies and procedures, using animal models for experimental validation of our predicted vaccine candidates.

5. Conclusion:

In this study, the whole proteome of SARS-CoV-2 was screened using reverse vaccinology, bioinformatics and immunoinformatics approaches to identify potential vaccine candidates. Through our investigation, we arrive at a conclusion that Spike glycoprotein is one of the major protein responsible for pathophysiology of SARS-CoV-2.

The potential epitopes were identified through a robust process and employed for vaccine construction, using which, several potential vaccine constructs were obtained. Therefore, our study will ease the development of appropriate therapeutic and prompt the future vaccine development against COVID-19 and this could serve an important milestone in developing an antiviral vaccine against SARS-CoV-2.

6. Author Contributions:

Conceived and designed the experiments: KR. BA. Performed the experiments: KR, BA, DS, TS, SS, RS, SS, PG. Analyzed the data: KR, BA, DS, SS, SS, PG. Contributed reagents/materials/analysis tools: KR, BA. Deep Learning: RS, AG, KM. Wrote the paper: KR BA DS SS SS PG.

7. Funding:

We are supported by the Department of Biotechnology, Ministry of Science and Technology, Government of India (Grant Id: BT/PR17252/BID/7/708/2016), the Robert J. Kleberg, Jr. and Helen C. Kleberg Foundation, USA and Baylor College of Medicine, Houston USA. Funders have no role in the design of this study.

8. References:

1. Lai A, Bergna A, Acciarri C, Galli M, Zehender G. Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2. *J Med Virol* (2020) **92**:675–679. doi:10.1002/jmv.25723

2. Cdc.gov. (2020). Coronavirus | Human Coronavirus Types | CDC. [online]. Available at: https://www.cdc.gov/coronavirus/types.html [Accessed February 22, 2020]

3. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N Engl J Med* (2020) **382**:1199–1207. doi:10.1056/NEJMoa2001316

4. No Title. Available at: https://www.worldometers.info/coronavirus/ [Accessed June 20, 2020]

5. Nishiura H, Jung S, Linton NM, Kinoshita R, Yang Y, Hayashi K, Kobayashi T, Yuan B, Akhmetzhanov AR. The Extent of Transmission of Novel Coronavirus in Wuhan, China, 2020. *J Clin Med* (2020) **9**:330. doi:10.3390/jcm9020330

6. Guo Q, Li M, Wang C, Fang Z, Wang P, Tan J, Wu S, Xiao Y, Zhu H. Host and infectivity prediction of Wuhan 2019 novel coronavirus using deep learning algorithm. *bioRxiv* (2020)2020.01.21.914044. doi:10.1101/2020.01.21.914044

7. Spaan W, Cavanagh D, Horzinek MC. Coronaviruses: Structure and Genome Expression. *J Gen Virol* (1988) **69**:2939–2952. doi:10.1099/0022-1317-69-12-2939

8. Chen Y, Liu Q, Guo D. Emerging coronaviruses: Genome structure, replication, and pathogenesis. *J Med Virol* (2020) **92**:418–423. doi:10.1002/jmv.25681

9. Jiang S, Du L, Shi Z. An emerging coronavirus causing pneumonia outbreak in Wuhan, China: calling for developing therapeutic and prophylactic strategies. *Emerg Microbes Infect* (2020) **9**:275–277. doi:10.1080/22221751.2020.1723441

10. Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, Bourexis D, Brister JR, Bryant SH, Canese K, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* (2018) **46**:D8–D13. doi:10.1093/nar/gkx1095

11. Mishto M, Mansurkhodzhaev A, Ying G, Bitra A, Cordfunke RA, Henze S, Paul D, Sidney J, Urlaub H, Neefjes J, et al. An in silico—in vitro Pipeline Identifying an HLA-A*02:01+ KRAS G12V+ Spliced Epitope Candidate for a Broad Tumor-Immune Response in Cancer Patients. *Front Immunol* (2019) **10**: doi:10.3389/fimmu.2019.02572

12. Brennan RM, Petersen J, Neller MA, Miles JJ, Burrows JM, Smith C, McCluskey J, Khanna R, Rossjohn J, Burrows SR. The Impact of a Large and Frequent Deletion in the Human TCR β Locus on Antiviral Immunity. *J Immunol* (2012) **188**:2742–2748. doi:10.4049/jimmunol.1102675

13. Rose PW, Prlić A, Bi C, Bluhm WF, Christie CH, Dutta S, Green RK, Goodsell DS, Westbrook JD, Woo J, et al. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res* (2015) **43**:D345–D356. doi:10.1093/nar/gku1214

14. Gonzalez-Galarza FF, McCabe A, Santos EJM dos, Jones J, Takeshita L, Ortega-Rivera ND, Cid-Pavon GM Del, Ramsbottom K, Ghattaoraya G, Alfirevic A, et al. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res* (2019) doi:10.1093/nar/gkz1029

15. Hizbullah, Nazir Z, Afridi SG, Shah M, Shams S, Khan A. Reverse vaccinology and subtractive genomics-based putative vaccine targets identification for Burkholderia pseudomallei Bp1651. *Microb Pathog* (2018) **125**:219–229. doi:10.1016/j.micpath.2018.09.033

16. Yu C-S, Lin C-J, Hwang J-K. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n -peptide compositions. *Protein Sci* (2004) **13**:1402–1406. doi:10.1110/ps.03479604

17. Shen H-B, Chou K-C. Virus-mPLoc: A Fusion Classifier for Viral Protein Subcellular Location Prediction by Incorporating Multiple Sites. *J Biomol Struct Dyn* (2010) **28**:175–186. doi:10.1080/07391102.2010.10507351

18. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, et al. PSORTb 3.0: improved protein subcellular localization prediction with

refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* (2010) **26**:1608–1615. doi:10.1093/bioinformatics/btq249

19. He Y, Xiang Z, Mobley HLT. Vaxign: The First Web-Based Vaccine Design Program for Reverse Vaccinology and Applications for Vaccine Development. *J Biomed Biotechnol* (2010) **2010**:1–15. doi:10.1155/2010/297505

20. Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server. *Bioinformatics* (2001) **17**:849–850. doi:10.1093/bioinformatics/17.9.849

21. Hofmann K, Stoffel WT. TMpred, Prediction of Transmembrane Regions and Orientation. (1993) Available at: https://embnet.vital-it.ch/software/TMPRED_form.html

22. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* (1990) **215**:403–410. doi:10.1016/S0022-2836(05)80360-2

23. Gasteiger E. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* (2003) **31**:3784–3788. doi:10.1093/nar/gkg563

24. Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics* (2007) **8**:4. doi:10.1186/1471-2105-8-4

25. Wizemann TM, Adamou JE, Langermann S. Adhesins as Targets for Vaccine Development. *Emerg Infect Dis* (1999) **5**:395–403. doi:10.3201/eid0503.990310

26. Chaudhuri R, Ansari FA, Raghunandanan MV, Ramachandran S. FungalRV: adhesin prediction and immunoinformatics portal for human fungal pathogens. *BMC Genomics* (2011) **12**:192. doi:10.1186/1471-2164-12-192

27. Goodman RE, Ebisawa M, Ferreira F, Sampson HA, van Ree R, Vieths S, Baumert JL, Bohle B, Lalithambika S, Wise J, et al. AllergenOnline: A peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reactivity. *Mol Nutr Food Res* (2016) **60**:1183–1198. doi:10.1002/mnfr.201500769

28. Ferre F, Clote P. DiANNA: a web server for disulfide connectivity prediction. *Nucleic Acids Res* (2005) **33**:W230–W232. doi:10.1093/nar/gki412

29. Saha S, Raghava GPS. “BcePred: Prediction of Continuous B-Cell Epitopes in Antigenic Sequences Using Physico-chemical Properties,” in, 197–204. doi:10.1007/978-3-540-30220-9_16

30. Bui H-H, Sidney J, Li W, Fusseder N, Sette A. Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines. *BMC Bioinformatics* (2007) **8**:361. doi:10.1186/1471-2105-8-361

31. Rötzschke O, Falk K, Stevanovic S, Jung G, Walden P, Rammensee H-G. Exact prediction of a natural T cell epitope. *Eur J Immunol* (1991) **21**:2891–2894. doi:10.1002/eji.1830211136

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

33. Singh H, Raghava GPS. ProPred: prediction of HLA-DR binding sites. *Bioinformatics* (2001) **17**:1236–1237. doi:10.1093/bioinformatics/17.12.1236

34. Sutmuller RPM, van Duivenvoorde LM, van Elsas A, Schumacher TNM, Wildenberg ME, Allison JP, Toes REM, Offringa R, Melief CJM. Synergism of Cytotoxic T Lymphocyte–Associated Antigen 4 Blockade and Depletion of Cd25+ Regulatory T Cells in Antitumor Therapy Reveals Alternative Pathways for Suppression of Autoreactive Cytotoxic T Lymphocyte Responses. *J Exp Med* (2001) **194**:823–832. doi:10.1084/jem.194.6.823

35. Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R, Raghava GPS. In Silico Approach for Predicting Toxicity of Peptides and Proteins. *PLoS One* (2013) **8**:e73957. doi:10.1371/journal.pone.0073957

36. Rawal K, Khurana T, Sharma H, Verma S, Gupta S, Kubba C, Strych U, Hotez PJ, Bottazzi ME. An extensive survey of molecular docking tools and their applications using text mining and deep curation strategies. (2019) doi:10.7287/peerj.preprints.27538

37. Maupetit J, Derreumaux P, Tuffery P. PEP-FOLD: an online resource for de novo peptide structure prediction. *Nucleic Acids Res* (2009) **37**:W498–W503. doi:10.1093/nar/gkp323

38. Neron B, Menager H, Maufrais C, Joly N, Maupetit J, Letort S, Carrere S, Tuffery P, Letondal C. Mobyle: a new full web bioinformatics framework. *Bioinformatics* (2009) **25**:3005–3011. doi:10.1093/bioinformatics/btp493

39. Zhou P, Jin B, Li H, Huang S-Y. HPEPDOCK: a web server for blind peptide–protein docking based on a hierarchical algorithm. *Nucleic Acids Res* (2018) **46**:W443–W450. doi:10.1093/nar/gky357

40. Meza B, Ascencio F, Sierra-Beltrán AP, Torres J, Angulo C. A novel design of a multi-antigenic, multistage and multi-epitope vaccine against Helicobacter pylori: An in silico approach. *Infect Genet Evol* (2017) **49**:309–317. doi:10.1016/j.meegid.2017.02.007

41. Ullah MA, Sarkar B, Islam SS. Exploiting the reverse vaccinology approach to design novel subunit vaccines against Ebola virus. *Immunobiology* (2020) **225**:151949. doi:10.1016/j.imbio.2020.151949

42. Hasan M, Ghosh PP, Azim KF, Mukta S, Abir RA, Nahar J, Hasan Khan MM. Reverse vaccinology approach to design a novel multi-epitope subunit vaccine against avian influenza A (H7N9) virus. *Microb Pathog* (2019) **130**:19–37. doi:10.1016/j.micpath.2019.02.023

43. Farjana S, Islam N, Taiebah A. Scrutinizing surface glycoproteins and poxin-schlafen protein to design a heterologous recombinant vaccine against monkeypox virus Scrutinizing surface glycoproteins and poxin-schlafen protein to design a heterologous recombinant vaccine against monkeypox. (2020)

44. Hasan M, Azim KF, Begum A, Khan NA, Shammi TS, Imran AS, Chowdhury IM, Urme SRA. Vaccinomics strategy for developing a unique multi-epitope monovalent vaccine against Marburg marburgvirus. *Infect Genet Evol* (2019) **70**:140–157. doi:10.1016/j.meegid.2019.03.003

45. Saha S, Raghava GPS. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res* (2006) **34**:W202–W209. doi:10.1093/nar/gkl343

46. Hebditch M, Carballo-Amador MA, Charonis S, Curtis R, Warwicker J. Protein–Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics* (2017) **33**:3098–3100. doi:10.1093/bioinformatics/btx345

47. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics* (2000) **16**:404–405. doi:10.1093/bioinformatics/16.4.404

48. Ashok Kumar T. CFSSP: Chou and Fasman Secondary Structure Prediction server. *Wide Spectr* (2013) **1**:15–19. doi:10.5281/zenodo.50733

49. Zhang Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* (2008) **9**:40. doi:10.1186/1471-2105-9-40

50. Xu D, Zhang Y. Improving the Physical Realism and Structural Accuracy of Protein Models by a Two-Step Atomic-Level Energy Minimization. *Biophys J* (2011) **101**:2525–2534. doi:10.1016/j.bpj.2011.10.024

51. Bhattacharya D, Nowotny J, Cao R, Cheng J. 3Drefine: an interactive web server for efficient protein structure refinement. *Nucleic Acids Res* (2016) **44**:W406–W409. doi:10.1093/nar/gkw336

52. Yan Y, Zhang D, Zhou P, Li B, Huang S-Y. HDOCK: a web server for protein–protein and protein–DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res* (2017) **45**:W365–W373. doi:10.1093/nar/gkx407

53. Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C, Beglov D, Vajda S. The ClusPro web server for protein-protein docking. *Nat Protoc* (2017) **12**:255–278. doi:10.1038/nprot.2016.169

54. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res* (2005) **33**:W363–W367. doi:10.1093/nar/gki481

55. Andrusier N, Nussinov R, Wolfson HJ. FireDock: Fast interaction refinement in molecular docking. *Proteins Struct Funct Bioinforma* (2007) **69**:139–159. doi:10.1002/prot.21495

56. Castiglione F, Bernaschi M. C-immsim: playing with the immune response. *Proc Sixt ...* (2004)1–7. Available at: http://www.math.ucsd.edu/~helton/MTNSHISTORY/CONTENTS/2004LEUVEN/CDROM/papers/316.pdf

57. Grote A, Hiller K, Scheer M, Munch R, Nortemann B, Hempel DC, Jahn D. JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res* (2005) **33**:W526–W531. doi:10.1093/nar/gki376

58. Biotech G. SnapGene Viewer. *Glick B*

59. Jagannadham J, Jaiswal HK, Agrawal S, Rawal K. Comprehensive Map of Molecules Implicated in Obesity. *PLoS One* (2016) **11**:e0146759. doi:10.1371/journal.pone.0146759

60. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* (2015) **43**:D447–D452. doi:10.1093/nar/gku1003

61. Parker JMR, Guo D, Hodges RS. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and x-ray-derived accessible sites. *Biochemistry* (1986) **25**:5425–5432. doi:10.1021/bi00367a013

62. Janin J, Wodak S, Levitt M, Maigret B. Conformation of amino acid side-chains in proteins. *J Mol Biol* (1978) **125**:357–386. doi:10.1016/0022-2836(78)90408-4

63. Pellequer J-L, Westhof E, Van Regenmortel MHV. Correlation between the location of antigenic sites and the prediction of turns in proteins. *Immunol Lett* (1993) **36**:83–99. doi:10.1016/0165-2478(93)90072-A

64. Emini EA, Hughes J V, Perlow DS, Boger J. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *J Virol* (1985) **55**:836–839. doi:10.1128/jvi.55.3.836-839.1985

65. Karplus PA, Schulz GE. Prediction of chain flexibility in proteins. *Naturwissenschaften* (1985) **72**:212–213. doi:10.1007/BF01195768

66. Kolaskar AS, Tongaonkar PC. A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett* (1990) **276**:172–174. doi:10.1016/0014-5793(90)80535-Q

67. Marciani DJ. Vaccine adjuvants: role and mechanisms of action in vaccine immunogenicity. *Drug Discov Today* (2003) **8**:934–943. doi:10.1016/S1359-6446(03)02864-2

68. Ikai A. Thermostability and Aliphatic Index of Globular Proteins. *J Biochem* (1980) doi:10.1093/oxfordjournals.jbchem.a133168

69. Ali M, Pandey RK, Khatoon N, Narula A, Mishra A, Prajapati VK. Exploring dengue genome to construct a multi-epitope based subunit vaccine by utilizing immunoinformatics approach to battle against dengue infection. *Sci Rep* (2017) **7**:9232. doi:10.1038/s41598-017-09199-w

70. Wu S, Zhang Y. LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Res* (2007) **35**:3375–3382. doi:10.1093/nar/gkm251

71. Liu Z, Xiao X, Wei X, Li J, Yang J, Tan H, Zhu J, Zhang Q, Wu J, Liu L. Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2. *J Med Virol* (2020) **92**:595–601. doi:10.1002/jmv.25726

72. Donoghue M, Hsieh F, Baronas E, Godbout K, Gosselin M, Stagliano N, Donovan M, Woolf B, Robison K, Jeyaseelan R, et al. A Novel Angiotensin-Converting Enzyme–Related Carboxypeptidase (ACE2) Converts Angiotensin I to Angiotensin 1-9. *Circ Res* (2000) **87**: doi:10.1161/01.RES.87.5.e1

73. Wong DW, Oudit GY, Reich H, Kassiri Z, Zhou J, Liu QC, Backx PH, Penninger JM, Herzenberg AM, Scholey JW. Loss of Angiotensin-Converting Enzyme-2 (Ace2) Accelerates Diabetic Kidney Injury. *Am J Pathol* (2007) **171**:438–451. doi:10.2353/ajpath.2007.060977

74. Raj VS, Mou H, Smits SL, Dekkers DHW, Müller MA, Dijkman R, Muth D, Demmers JAA, Zaki A, Fouchier RAM, et al. Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-EMC. *Nature* (2013) **495**:251–254. doi:10.1038/nature12005

75. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, Wang B, Xiang H, Cheng Z, Xiong Y, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–

Infected Pneumonia in Wuhan, China. *JAMA* (2020) **323**:1061. doi:10.1001/jama.2020.1585

76. Kassir R. Risk of COVID-19 for patients with obesity. *Obes Rev* (2020) **21**: doi:10.1111/obr.13034

77. Guo W, Li M, Dong Y, Zhou H, Zhang Z, Tian C, Qin R, Wang H, Shen Y, Du K, et al. Diabetes is a risk factor for the progression and prognosis of COVID-19. *Diabetes Metab Res Rev* (2020)e3319. doi:10.1002/dmrr.3319

78. Stothard P. The Sequence Manipulation Suite: JavaScript Programs for Analyzing and Formatting Protein and DNA Sequences. *Biotechniques* (2000) **28**:1102–1104. doi:10.2144/00286ir01

79. Cleemput S, Dumon W, Fonseca V, Abdool Karim W, Giovanetti M, Alcantara LC, Deforche K, de Oliveira T. Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. *Bioinformatics* (2020) **36**:3552–3555. doi:10.1093/bioinformatics/btaa145

80. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* (2020) **395**:565–574. doi:10.1016/S0140-6736(20)30251-8

81. Xu Z, Shi L, Wang Y, Zhang J, Huang L, Zhang C, Liu S, Zhao P, Liu H, Zhu L, et al. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *Lancet Respir Med* (2020) **8**:420–422. doi:10.1016/S2213-2600(20)30076-X

82. Ong E, Wong MU, Huffman A, He Y. COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. *bioRxiv* (2020)2020.03.20.000141. doi:10.1101/2020.03.20.000141

83. Bhattacharya M, Sharma AR, Patra P, Ghosh P, Sharma G, Patra BC, Lee S, Chakraborty C. Development of epitope-based peptide vaccine against novel coronavirus 2019 (SARS-COV-2): Immunoinformatics approach. *J Med Virol* (2020) **92**:618–631. doi:10.1002/jmv.25736

84. Robson B. Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus. *Comput Biol Med* (2020) **119**:103670. doi:10.1016/j.compbiomed.2020.103670

85. Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. *Cell Host Microbe* (2020) **27**:671-680.e2. doi:10.1016/j.chom.2020.03.002

86. Kiyotani K, Toyoshima Y, Nemoto K, Nakamura Y. Bioinformatic prediction of potential T cell epitopes for SARS-Cov-2. *J Hum Genet* (2020) **65**:569–575. doi:10.1038/s10038-020-0771-5

87. Prachar M, Justesen S, Steen-Jensen DB, Thorgrimsen SP, Jurgons E, Winther O, Bagger FO. COVID-19 Vaccine Candidates: Prediction and Validation of 174 SARS-CoV-2 Epitopes. *bioRxiv* (2020)2020.03.20.000794. doi:10.1101/2020.03.20.000794

88. Qiu T, Mao T, Wang Y, Zhou M, Qiu J, Wang J, Xu J, Cao Z. Identification of potential cross-protective epitope between a new type of coronavirus (2019-nCoV) and severe acute respiratory syndrome virus. *J Genet Genomics* (2020) **47**:115–117. doi:10.1016/j.jgg.2020.01.003

89. Tian X, Li C, Huang A, Xia S, Lu S, Shi Z, Lu L, Jiang S, Yang Z, Wu Y, et al. Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody. *Emerg Microbes Infect* (2020) **9**:382–385. doi:10.1080/22221751.2020.1729069

90. Blackwell JM, Fakiola M, Castellucci LC. Human genetics of leishmania infections. *Hum Genet* (2020) **139**:813–819. doi:10.1007/s00439-020-02130-w

91. Collins AR. “Virus-Ligand Interactions of OC43 Coronavirus with Cell Membranes,” in, 285–291. doi:10.1007/978-1-4615-2996-5_44

92. Lin M, Tseng H-K, Trejaut JA, Lee H-L, Loo J-H, Chu C-C, Chen P-J, Su Y-W, Lim KH, Tsai Z-U, et al. Association of HLA class I with severe acute respiratory syndrome coronavirus infection. *BMC Med Genet* (2003) **4**:9. doi:10.1186/1471-2350-4-9

Tables:

Table 1: B-cell epitopes present on surfaces predicted via BCPRED.

S. No.	Antigenic propensity	Antigenic Score
1.	DLCFTNVY	1.85
2.	YYVGYLQPR	1.46
3.	EPVLKGVKLHYT	1.41
4.	LIDLQEL	1.39
5.	TEILPVS	1.26
6.	EILDITPCSFGGVSVITPG	1.13
7.	SVVNIQK	1.08
8.	YQPYRVVVLSFELLH	0.97
9.	PHGVVFLHVTYVP	0.93
10.	YNYLYRLFR	0.86
11.	ECSNLLLQYGSFC	0.86
12.	MFVFLVLLPLVSSQCVNLTT	0.83
13.	LEPLVDLPIGI	0.82
14.	FNCYFPLQSY	0.82
15.	FSTFKCYGVSPT	0.8

Table 2: List of top scoring MHC class I and MHC class II binding T-cell epitopes.

S. No.	MHC class I binding (CTL)epitopes	MHC class II binding (HTL) epitopes
1.	KIADYNYKL	VKNKCVNFN
2.	VVVLSFELL	YRFNGIGVT
3.	TLDSKTQSL	VVFLHVTYV
4.	GKQGNFKNL	FKCYGVSPT
5.	VRDLPQGFS	VNLTTRTQL
6.	PWYIWLGFI	IGINITRFQ
7.	NFGAISSVL	LVKNKCVNF
8.	QGFSALEPL	VVIGIVNNT

Table 3: Protein-Peptide docking using web server HPEPDOCK of MHC-I with crystal structure of HLA*A2 and HLA*B7:

For MHC-Class I with HLA-A2		
Peptide	**Human allele (PDB ID)**	**Docking Score (kcal mol-1)**
KIADYNYKL	6O4Y	-205.89
VVVLSFELL	6O4Y	-157.77
TLDSKTQSL	6O4Y	-150.05
GKQGNFKNL	6O4Y	-178.48
For MHC-Class I with HLA-B7		
Peptide	**Human allele (PDB ID)**	**Docking Score (kcal mol-1)**
KIADYNYKL	3VCL	-171.84
VVVLSFELL	3VCL	-166.31
TLDSKTQSL	3VCL	-136.22
GKQGNFKNL	3VCL	-152.84

Table 4: Protein sequence of vaccine constructs V1, V2 and V3 along with their antigenicity analysis.

Vaccine Construct	Composition /Order	Sequence	Antigenicity Score (Threshold=0.4)
V1	Predicted CTL, HTL & BCL epitopes of Spike Glycoprotein with β defensin adjuvant	GIINTLQKYYCRVRGGRCAVLSCLPKEEQIGKCSTR GRKCCRRKKEAAAKKIADYNYKLGGGSKIADYNY KLGGGSKIADYNYKLGGGSKIADYNYKLGGGSKIA DYNYKLGGGSKIADYNYKLGGGSKIADYNYKLGG GSKIADYNYKLAAYVKNKCVNFNAAYVKNKCVNF NAAYVKNKCVNFNAAYVKNKCVNFNAAYVKNKC VNFNAAYVKNKCVNFNAAYVKNKCVNFNAAYVK NKCVNFNKKDLCFTNVYKKDLCFTNVYKKDLCFT NVYKKDLCFTNVYKKDLCFTNVYKKDLCFTNVYK KDLCFTNVYKKDLCFTNVY	**1.16**
V2	Predicted CTL, HTL & BCL epitopes of Spike Glycoprotein with L7/L12 ribosomal protein adjuvant	MSDINKLAETLVNLKIVEVNDLAKILKEKYGLDPSA NLAIPSLPKAEILDKSKEKTSFDLILKGAGSAKLTVV KRIKDLIGLGLKESKDLVDNVPKHLKKGLSKEEAES LKKQLEEVGAEVELKEAAAKKIADYNYKLGGGSKI ADYNYKLGGGSKIADYNYKLGGGSKIADYNYKLG GGSKIADYNYKLGGGSKIADYNYKLGGGSKIADYN YKLGGGSKIADYNYKLAAYVKNKCVNFNAAYVK NKCVNFNAAYVKNKCVNFNAAYVKNKCVNFNAA YVKNKCVNFNAAYVKNKCVNFNAAYVKNKCVNF NAAYVKNKCVNFNKKDLCFTNVYKKDLCFTNVYK KDLCFTNVYKKDLCFTNVYKKDLCFTNVYKKDLC FTNVYKKDLCFTNVYKKDLCFTNVY	**1.03**
V3	Predicted CTL, HTL & BCL epitopes of Spike protein with HABA adjuvant	MAENPNIDDLPAPLLAALGAADLALATVNDLIANL RERAEETRAETRTRVEERRARLTKFQEDLPEQFIEL RDKFTTEELRKAAEGYLEAATNRYNELVERGEAAL QRLRSQTAFEDASARAEGYVDQAVELTQEALGTVA SQTRAVGERAAKLVGIELEAAAKKIADYNYKLGGG SKIADYNYKLGGGSKIADYNYKLGGGSKIADYNYK LGGGSKIADYNYKLGGGSKIADYNYKLGGGSKIAD YNYKLGGGSKIADYNYKLAAYVKNKCVNFNAAY VKNKCVNFNAAYVKNKCVNFNAAYVKNKCVNFN AAYVKNKCVNFNAAYVKNKCVNFNAAYVKNKCV NFNAAYVKNKCVNFNKKDLCFTNVYKKDLCFTNV YKKDLCFTNVYKKDLCFTNVYKKDLCFTNVYKKD LCFTNVYKKDLCFTNVYKKDLCFTNVY	**0.98**

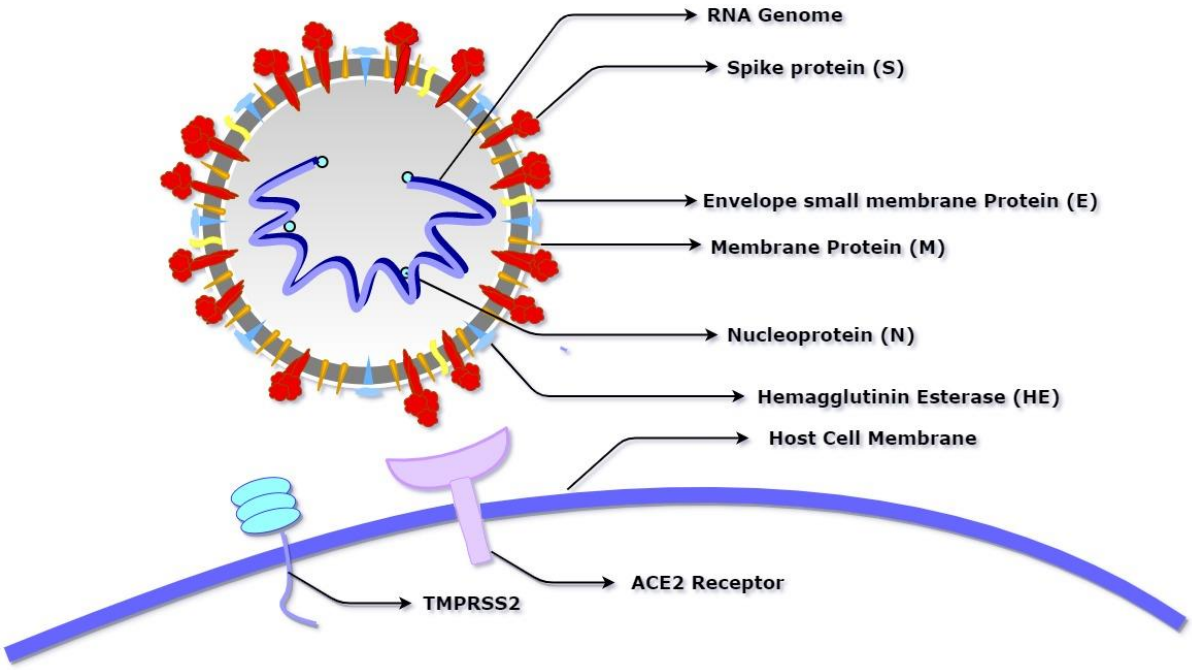


Figure 1: Schematic diagram of SARS-CoV-2 showing its basic component proteins along with its receptor binding site, Angiotensin-converting enzyme 2 (ACE2) and Transmembrane serine Protease TMPRSS2 . The virus consists of a spherical membrane (shown in white and grey) which constitutes membrane protein (shown in Orange), spike protein (shown in Red), hemagglutinin esterase (shown in Blue), and envelope small membrane protein (shown in Yellow) . The spike protein binds to the ACE2 receptor of the host cell after being activated by the proteolytic cleavage activity of TMPRSS2.

Figure 2: Flow chart depicting the multi-epitope subunit vaccine development against SARS-CoV-2.

Figure 2: Flow chart depicting the multi-epitope subunit vaccine development against SARS-CoV-2.

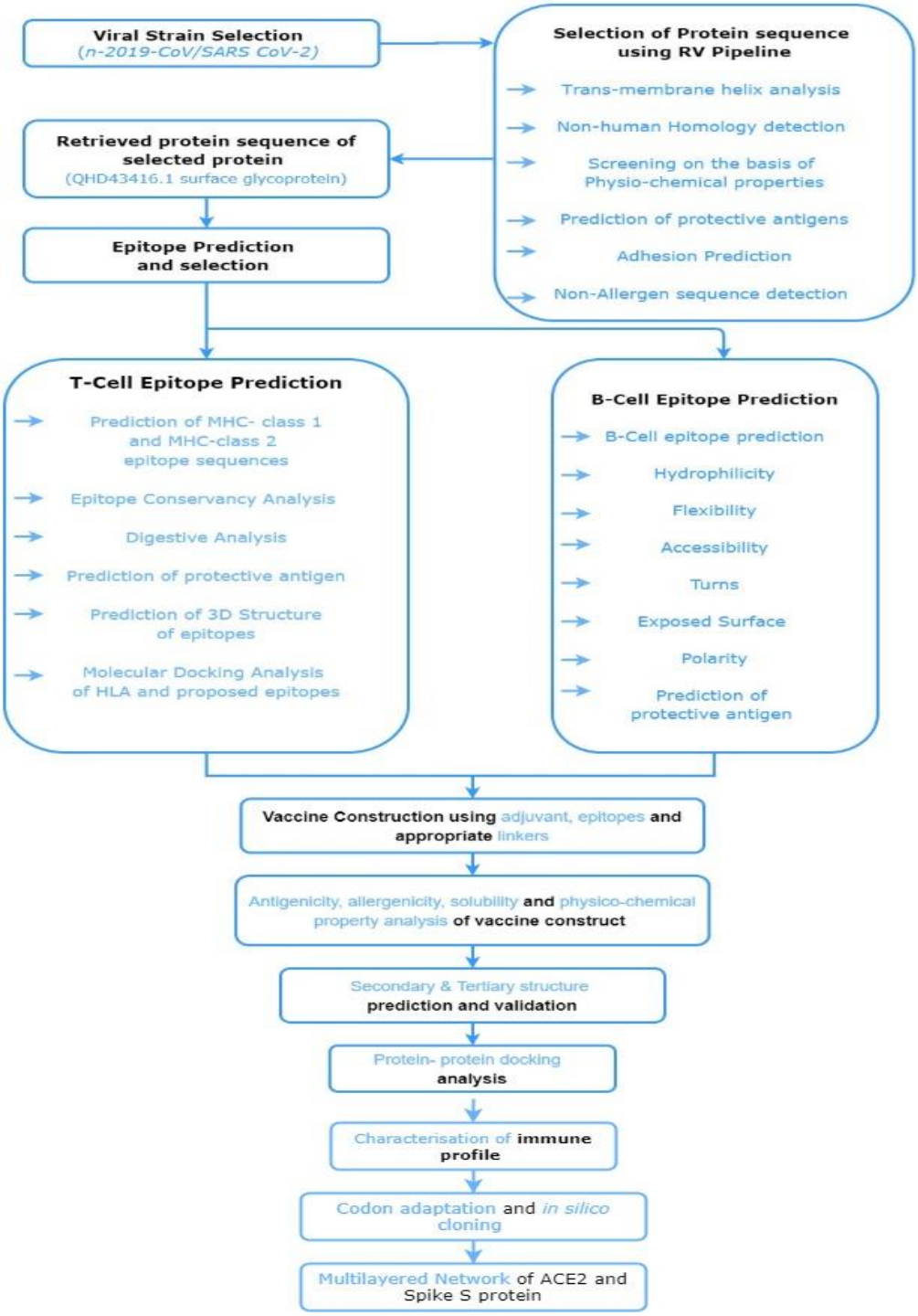


Figure 2: Flow chart depicting the multi-epitope subunit vaccine development against SARS-CoV-2.

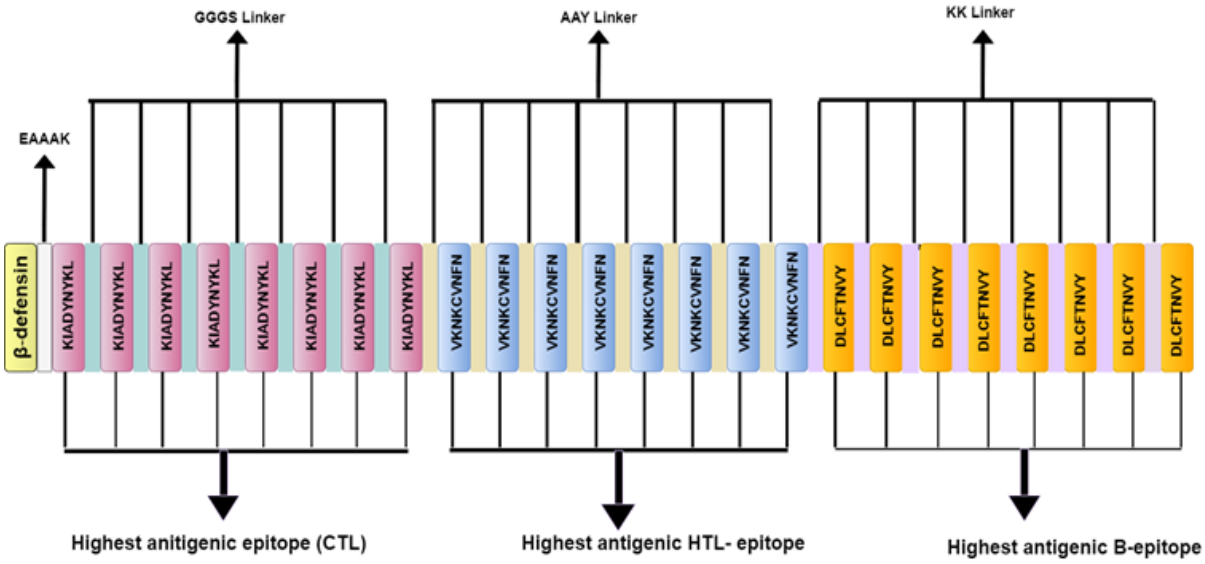


Figure 3: Schematic diagram of multi-epitope vaccine peptide. It is a 32 (insert amino acid number) amino acid long sequence having Beta-defensin as an adjuvant (Light canary yellow) which is connected to the highest antigenic CTL epitope sequence (Pink) through EAAAK linker (White). The CTL epitopes are linked to each other by GGGS linkers (Grayish Cyan), and to the highest antigenic HTL epitope (Light Blue) by AAY linkers (very soft yellow). Next, the HTL epitopes are linked to each other through AAY linkers, and to the highest antigenic B Epitope (Vivid Yellow) through KK linkers (Pale violet). The B epitopes are linked to each other using the KK linkers as well.

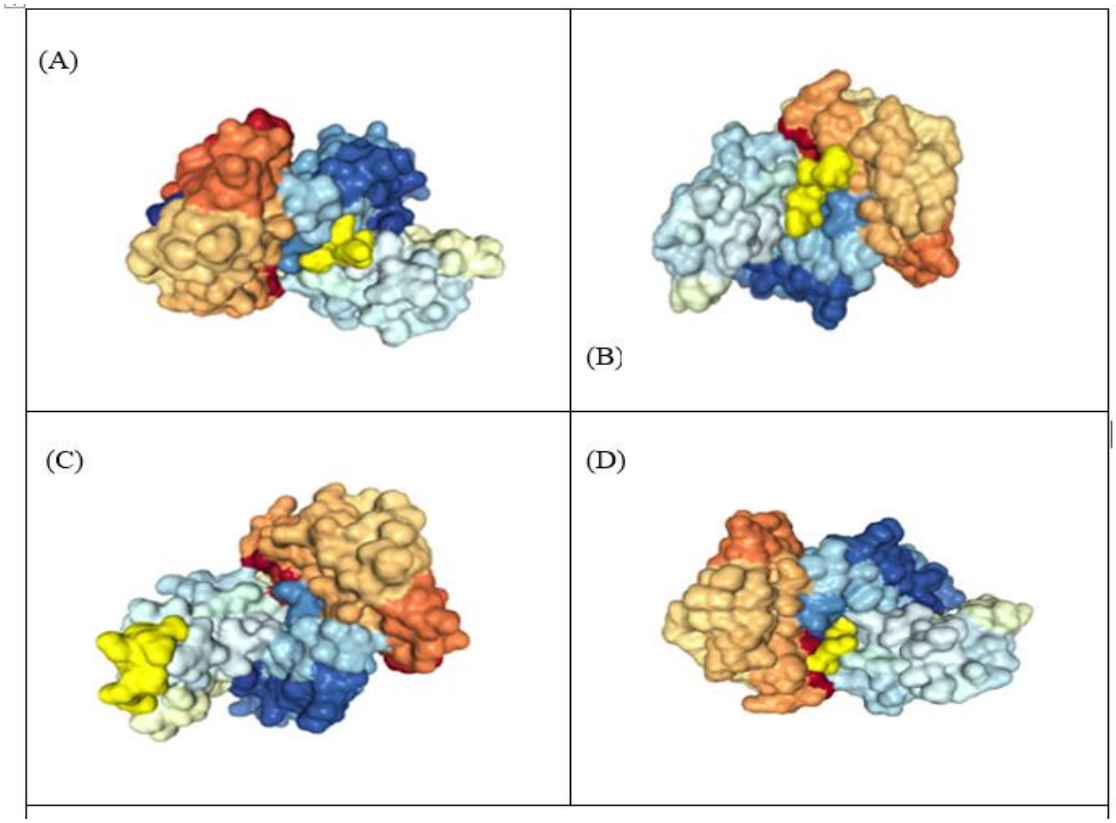


Figure 4: Representation of protein-peptide docked complex of top 4 MHC class-1 epitopes sequences **(A)** KIADYNYKL, **(B)** VVVLSFELL, **(C)** TLDSKTQSL and **(D)** GKQGNFKNL, shown in golden yellow) in association with the HLA-A*02 allele using HPEPDOC [Zhou et al.,2018]. The epitopes have a binding affinity of -205.89, -157.77, -150.05 and -178.48 respectively with HLA_A*02.

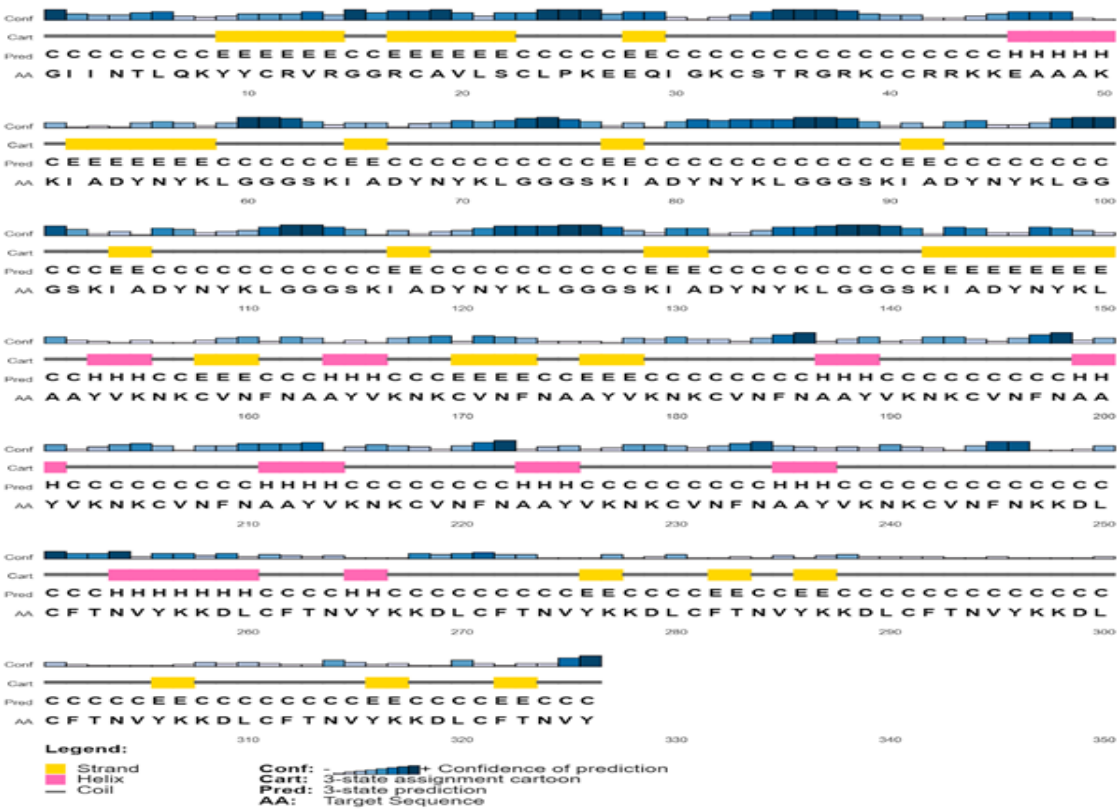


Figure 5: Graphical representation of secondary structure features of proposed subunit vaccine sequence using PSIPRED tool.

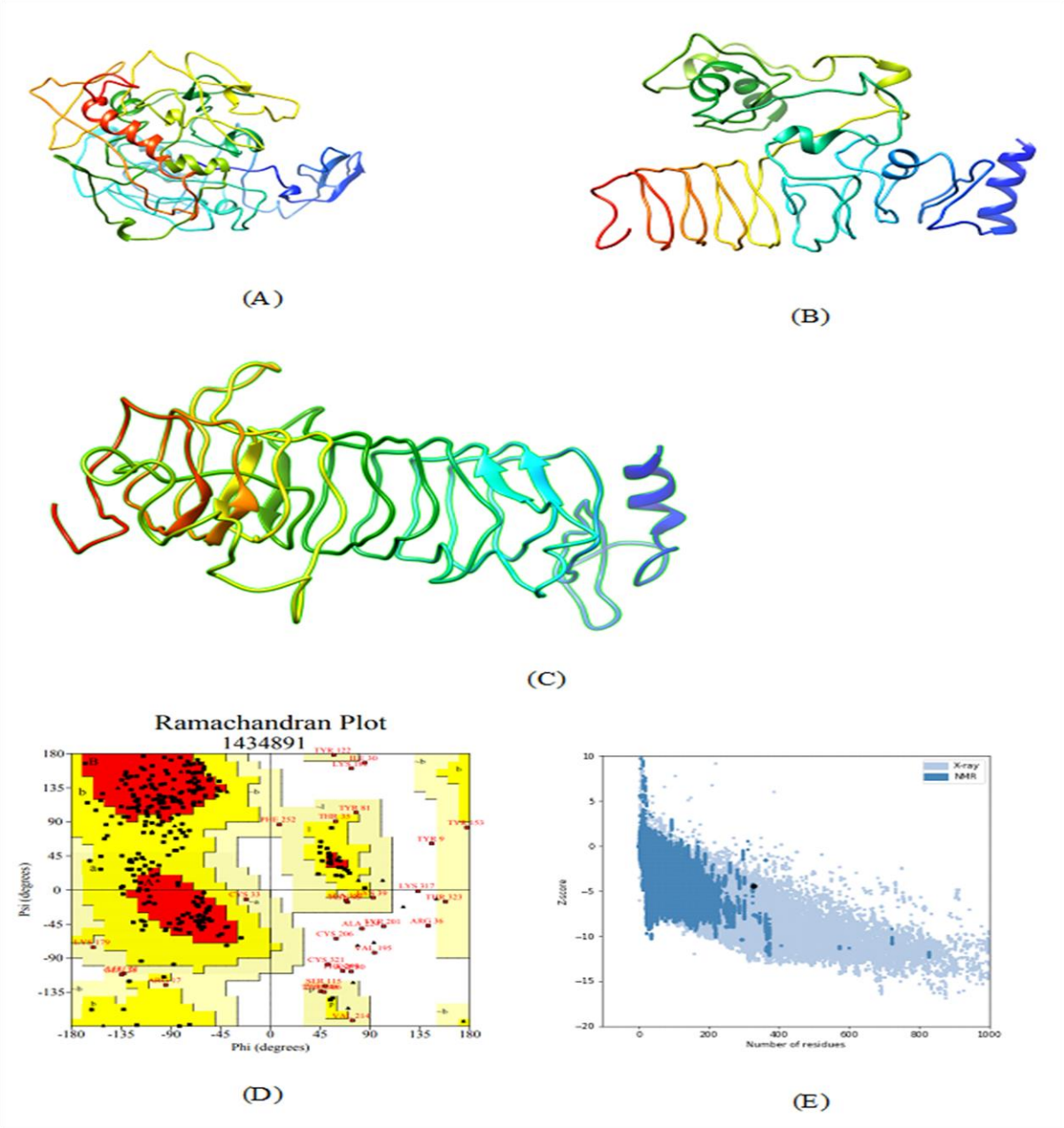


Figure 6: Tertiary structure modeling, refinement and validation**. (A)** The final 3D model of multi epitope vaccine chimeric protein generated via homology modelling on I-TASSER, **(B)** Refined model obtained via ModRefiner, **(C**)The refined 3D structure generated by 3DRefine **(D)** Ramachandran Plot Analysis signifying 57.0%, 38.9% and 4.0% of protein residues in favoured, allowed and disallowed (outlier) regions respectively, **(E)** ProSA-web, giving a Z-Score of -4.4.

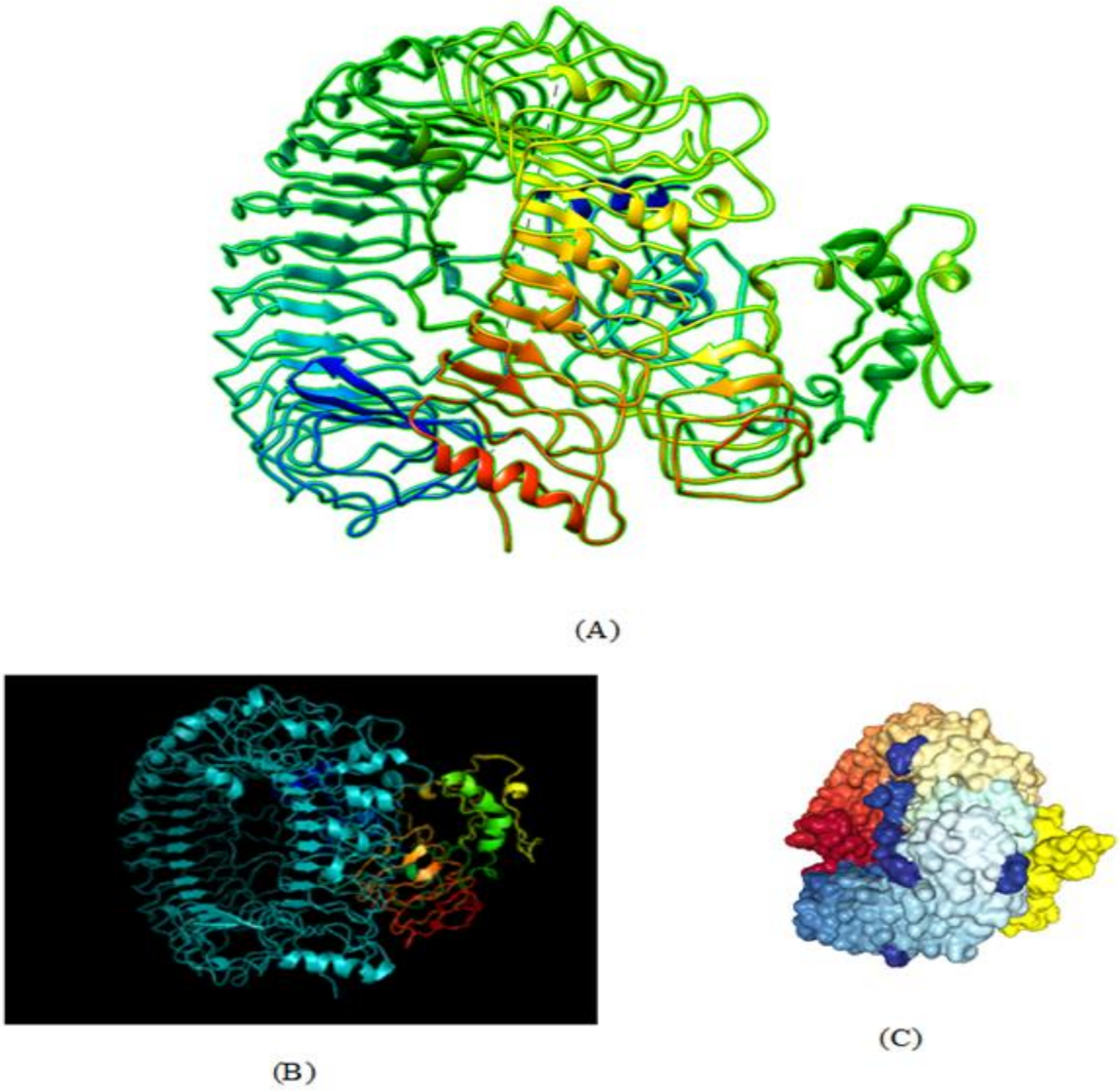


Figure 7: (A) Docked complex of TLR-8 with the chimeric vaccine construct. **(B)** Docking complex generated via Cluspro server illustrating binding affinity between TLR-8 (cyan) and vaccine component (rainbow). The lowest energy of -1277.5 kcal/mol was achieved for this model (complex 2). **(C)** Docking complex generated via HDOCK server which predicted the binding energy as -330.04 for protein (rainbow) and ligand (yellow).

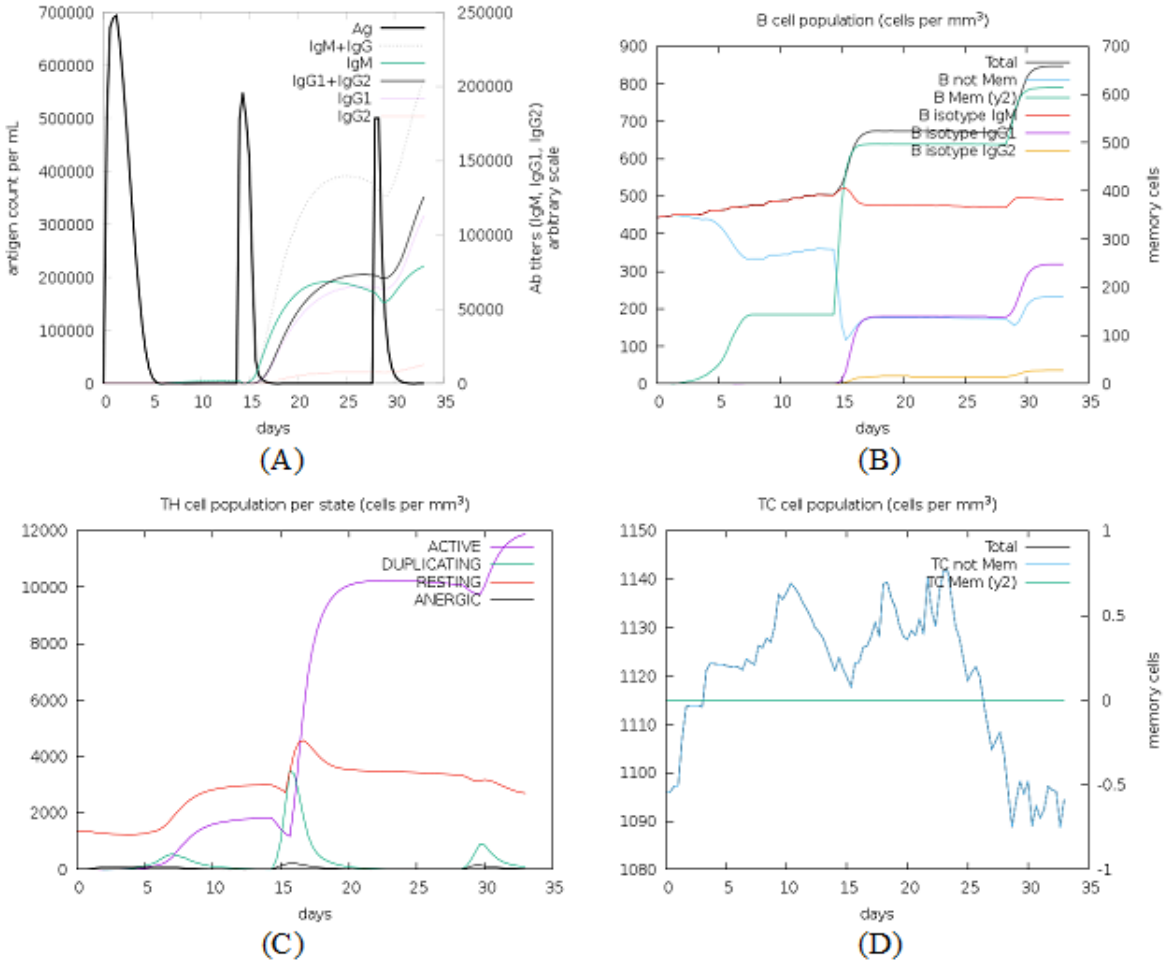
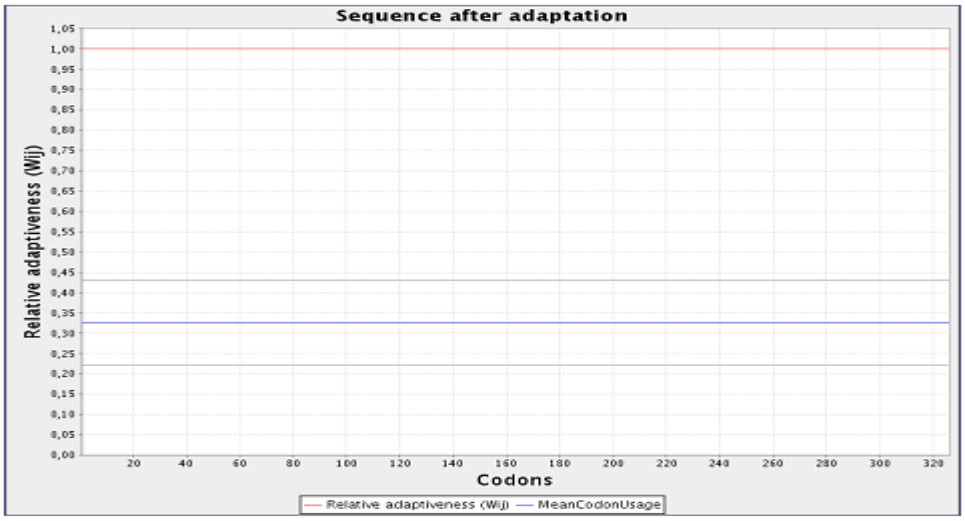
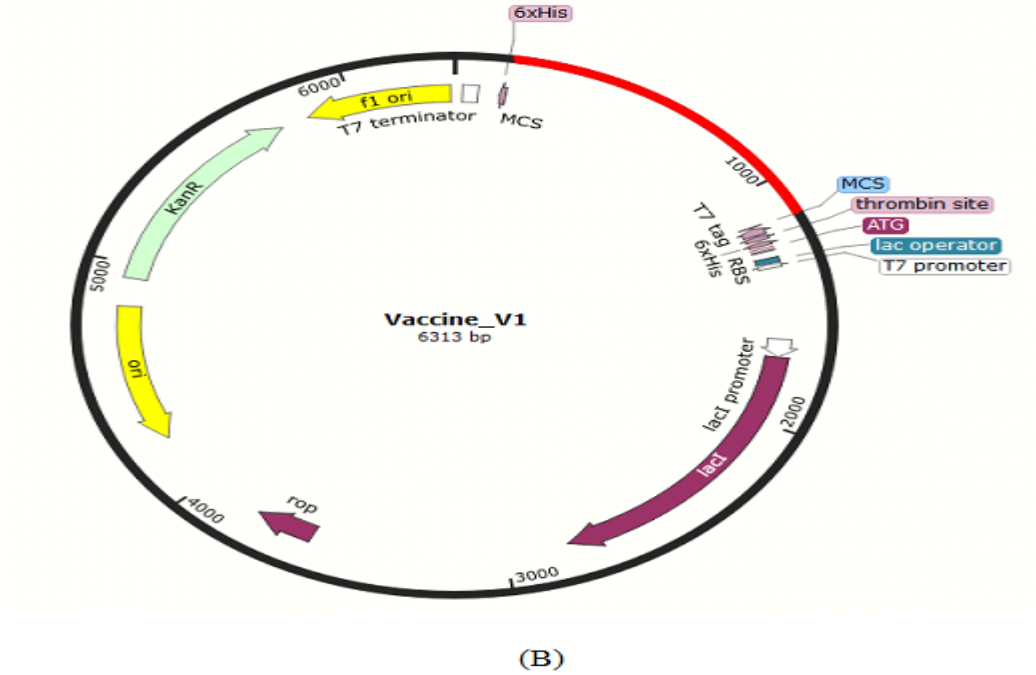


Figure 8: Immune simulations of the chimeric protein vaccine. **(A)** Production of Immunoglobulins in response to successive antigen injections (different coloured peaks corresponding to different sub-classes of immunoglobulins and antigen represented by black vertical lines). **(B)** Changes observed in B-cell population **(C)**T-helper cells per state (Resting state denotes the cells not presented with antigen while anergic state denotes cells showing tolerance to antigens due to repeated exposure.) **(D)** Changes in T-cytotoxic cell population after administration of vaccine construct V1.



(A)



(B)

Figure 9: (A) Codon adaptation result of vaccine construct V1 predicted by JCat tool predicting that the optimized codon sequence has a length of 978 nucleotides and its CAI (Codon Adaptation Index) was predicted to be 1.0, with an average of 41.21% GC for the adapted sequence. **(B)** Final protein in-silico restriction cloning into pET28a (+) vector. Here, the red portion represents the gene sequence of the designed vaccine and the black portion denotes the backbone of the vector. The DNA sequence is inserted into the MCS region of the cloning vector.