Spike selection in SARS-CoV-2 variants across different geographical regions reveals unique signature patterns and differential stability with drug interaction

Devang Haresh Liya[1#], Nithishwer Mouroug Anand[1#], Ashwin K. Jainarayanan[2,3*], Mirudula Elanchezhian[4], Madhumati Seetharaman[1], Dhanuush Balakannan[4], Arpit Kumar Pradhan[5,6]*

1 Department of Physical Sciences, Indian Institute of Science Education and Research, Mohali, India

2 Kennedy Institute of Rheumatology, University of Oxford, Oxford OX3 7FY, UK

3 Interdisciplinary Bioscience Doctoral Training Program and Exeter College, University of Oxford, Oxford OX3 7DQ, UK

4 Department of Biological Sciences, Indian Institute of Science Education and Research, Mohali, India

5 Klinik für Anaesthesiologie und Intensivmedizin der Technischen Universität München, Klinikum rechts der Isar, Germany

6 Graduate School of Systemic Neuroscience, Ludwig Maximilian University of Munich, Germany

#These authors contributed equally to the work

***Correspondence: arpit.pradhan@tum.de, ashwin.jainarayanan@dtc.ox.ac.uk**

Keywords

Spike Protein, SARS-COV-2, COVID 19 variants, Mutational frequency, Drug docking, MD simulation, Spike evolution, Geographical spike variants

Abstract

The evolution of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus since its emergence in 2019 has yielded several new viral variants with varied infectivity, disease severity, and antigenicity. Although most mutations are expected to be relatively neutral, mutations at the Spike region of the genome has shown to have a major impact on the viral transmission and infection in humans. Therefore, it is crucial to survey the structures of spike protein across the global virus population to contextualize the rate of therapeutic success against these variants. In this study, high-frequency mutational variants from different geographic regions were pooled in order to study the structural evolution of the spike protein through drug docking and MD simulations. We investigated the mutational burden in the spike sub regions and have observed that the different variants harbour unique signature patterns in the spike sub regions, with certain domains being highly prone to mutations. Further, the MD simulations and docking study revealed that different variants show differential stability when docked for the same set of drug targets. This work sheds light on the mutational burden and the stability landscape of the spike protein across the variants from different geographical regions.

Introduction

Since its emergence in December 2019, the novel coronavirus disease 2019 (COVID-19) has created an unprecedented public health, social and economic challenge for humankind. The etiological agent of COVID-19 is the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) which belongs to the β coronavirus family (1). The unique pathophysiology of the SARS-CoV-2 virus has enabled it to spread globally and has led to COVID-19 being declared a global pandemic (2).

As of February 25, 2022, nearly 432 million cases were reported worldwide, and nearly 5.9 million individuals lost their lives due to three globally identified waves of infection, each led by a different variant of the virus (3). Alpha, Beta, Gamma, Delta and Omicron are current 'variants of concern', whereas Mu and Lambda are current 'variants of interest' recognized by the World Health Organization (https://www.who.int/). In the context of 'variants of concern', SARS-CoV-2 evolution has been characterized by the appearance of sets of mutations that affect viral features like transmissibility and antigenicity. In contrast, a variant of interest is a SARS CoV-2 variant with a genetic capability that affects the virus' severity, immune escape, transmissibility and diagnostic escape.

The SARS-CoV-2 genome encodes 27 proteins from 14 ORFs, including 15 non-structural, 4 major structural, and 8 accessory proteins, according to the first investigation of the full length genomic sequence of the virus (4). The four major structural proteins of SARS-CoV-2 are spike glycoprotein (S), membrane (M), envelope (E), and nucleocapsid (N) which are essential to viral pathogenicity(5). Out of these structural proteins, the Spike protein is the major antigen as it mediates attachment of the virus to host cell-surface receptors.

The SARS-CoV-2 spike protein is heavily glycosylated, with 66 potential N-glycosylation sites per trimer and is post-translationally cleaved into two subunits: S1 and S2 by mammalian furin (6-8). The S1 subunit predominantly comprises the amino-terminal domain and the receptor-binding domain (RBD), while the S2 subunit, which includes the protein's trimeric core, is responsible for membrane fusion (9). These properties of the spike protein make it a lucrative target for therapeutic drug design and hence most vaccines and therapeutic medications target the spike glycoprotein (S), which forms homotrimers on the virion's surface and is responsible for allowing the virion to enter the host cell by binding to the receptor ACE2 (RBD) (10). However, spike glycoprotein (S) evolution makes it challenging to develop a potential vaccine to eliminate SARS-CoV-2. Pereson et al. estimated the evolutionary rate for the spike protein as $1.08 \times 10-3$ nucleotide substitutions/site/year (11). High mutational diversity within the spike subunit enables the virus to escape potential therapeutics and helps increase its transmissibility (9, 12). An extensive study that looks into the region-specific variants and the potential targets to the spike becomes an absolute necessity.

Genomic analyses of the recently identified omicron variant has highlighted specific mutations in the spike protein of the Omicron variant which makes it easily transmissible and less deadly (13, 14). One of such studies suggests that due to the mutations acquired, the Omicron variant infection does not require the presence of TMPRSS2 receptor and uses endosomal route, while the older variants required TMPRSS2 receptor in addition to ACE2 receptor to inject their genome into the cell (14). This maybe the reason for high spread of Omicron variant as this enables the virus to infect other cells along their entry in the upper airway which expresses ACE2 alone as opposed to the conventional method of infecting the cells with both TMPRSS2 and ACE2 (TMPRSS2- rich lung cells) expression (15). In addition to the aforementioned mutations, several new mutations have been reported in the RBD and N terminal domain of the Omicron variant's spike protein which make it difficult for the antibodies to bind and neutralize them (16). Though some mutations, for example, substitution of asparagine for lysine at position 417 (K417N) led to reduction in ACE2 binding affinity, it appears that mutated residues arginine-493, serine-496, and arginine-498 in the RBD forms new salt bridges and hydrogen bonds with ACE2. These interactions appear to compensate for other Omicron mutations that are known to reduce ACE2 binding affinity (17). Thus, it becomes crucial to understand the mutations in every variant in order to target drugs effectively.

This article comprehensively studies the characteristics of COVID-19 variants designated as "variants of concerns" and "variants of interest" in a region-specific manner to reveal potential mutation hotspots in the Spike subunit. This multi-array analysis also addresses other aspects of the COVID 19 variants such as transmissibility and infection rate as a consequence of the mutations. First, we screen through the mutation in the spike region, and each variant based on their geographical predominance. The point mutation D614G that was found to have the highest mutation probability was chosen for further analysis. Along with this, we screened the variant along different geographic locations and identified the variant with the highest mutation in the spike protein. The Spike reconstructed with the above mutations was compared with the Spike from the initial strain of COVID19 as well as the spike from the omicron to identify the basal level mutations. Using this as the base, we screened through the FDA-approved drugs and studied how the mutations in the spike region affect the drug binding. Given the high mutational diversity of the Spike protein, our study forms the first-of-its-kind high throughput analysis of the region-specific mutation analysis detailing the hotspot zones of mutations in the spike. This study also gives new insights on the evolution of the different variants and their correlation with the mutation sites in the Spike subunit, which has affected the individual variants infection as well as transmissibility rate. Finally, it also explores the possible therapeutic applications by comparing the drugs that target the initial strain of COVID and also the variants with the highest mutation in the spike. Our study would be beneficial and a piece of first-ground information that compares the different strains of the COVID along with the recently emerged variant of interest Omicron.

Methods

Data Retrieval

The viral genomes used in this study were downloaded in FASTA format from the GISAID (Global Initiative on Sharing All Influenza Data; https://www.gisaid.org/), a global open-source repository of data pertinent to strains of influenza and coronaviruses. A total of 8367 genome sequences were sorted based on geographical location, i.e., the continent on which the sample was obtained, and variant type. Of these, 1017 were from Africa, 1502 from Asia, 2601 from Europe, 2321 from North America and 926 from South America. The countries of origin of these samples include Nigeria, Ghana, The Democratic Republic of Congo, Botswana, South Africa, Malaysia, Iran, Oman, India, China, Japan, France, United Kingdom, Belgium, Italy, Brazil, Canada, USA and Mexico amongst several others. With respect to variant types, 1424 were of the Alpha variant (B.1.1.7), 774 of the Beta variant (B.1.351), 977 of the Gamma variant (B.1.1.28.1), 1550 of the Delta variant (B.1.617.2), 1177 of the Epsilon variant (B.1.429) and 2465 of the Omicron variant (B.1.1.529). As of January 2022, the Alpha, Beta, Gamma, Delta and Omicron variants are labeled current VOCs (variants of concern) by the WHO, with the Epsilon variant having been a VOC until July 2021. All samples were collected from human sources. A comprehensive table with the number of sequences of each variant type, from each geographical location, is given in Supplementary Table 1.

Data preprocessing and alignment

All 8245 sequences were divided into subgroups according to their geographical location and the variant type (eg. Alpha variant in Africa, Beta variant in Europe). These individual subgroups were used separately in the further mutation analysis. Since the focus of this work is the spike protein region, we first need to isolate its sequence from the entire genome. The spike sequences were thus obtained by matching the regular expression ATGTTTGTT.[A-Z]+CATTACACA. 172 sequences out of initial 8245 sequences did not yield a match using the aforementioned method and they were removed from the subsequent analysis. Each subgroup of sequences was then aligned using online MAFFT(18) with NC_045512.2 as the reference sequence. The alignment length was kept the same as the reference sequence, scoring matrix was 1PAM/k=2 and all the other settings were fixed to their default values. The entire command for this alignment is mafft --inputorder --

adjustdirection --keeplength --compactmapout --anysymbol --maxambiguous 0.05 --kimura 1 --addfragments fragments --auto input.

The nucleotide sequences were then translated to the amino acid sequences using the translate() method in Biopython (19) and aligned using online MAFFT with default settings given by the command mafft --inputorder --adjustdirection --keeplength --compactmapout --anysymbol --maxambiguous 0.05 --addfragments fragments --auto input. Sequences with more than 5% ambiguous letters were removed at this stage – for both nucleotides and amino acids – and the remaining number of sequences in each subgroup are shown in Table 1 (For nucleotides Supplementary Table 2).

Mutation analysis

The mutations were identified by comparing each amino acid/nucleotide in the alignment to the one in the reference sequence. As a consequence of the fixed alignment length, this method only identifies the substitution and deletion mutations. However, we performed different checks to verify that the insertion mutations were indeed rare, with only the Omicron variant showing one significant insertion mutation. We further note that translating the nucleotide sequence to the amino acid sequence suppresses all the non-synonymous mutations. The mutation frequency for a given site is defined as the number of sequences having a mutation – not necessarily the same mutation – on that site. Similarly, the normalized mutation frequency is defined as the mutation frequency divided by the total number of sequences in the alignment. These quantities were then visualized as line plots and heat maps. We also combined all the subsets and aligned them to find the most mutated sequences irrespective of the geographical region or variant. This was done by calculating the number of mutations for each sequence.

Spike Protein Structure modeling

The structure of the spike protein was obtained from RCSB PDB (7CWL). The point mutation was introduced at position 614 involving D→G substitution was obtained using Chimera. The cryoEM structure of the Omicron spike 7T9J from the pdb was used for analysis. In order to reconstruct the spike from Spain/MD-11674437/2021|EPI_ISL_882632 sequence, we used homology modeling by Swiss Model using the spike from initial strain as the template. The 3D structure model for the protein were modeled by comparative protein modeling methods using the SWISS-

MODEL server (http://swissmodel.expasy.org)(20). We used the structure-based alignment and optimized it to minimize energy using SWISS-MODEL. Models are made according to the target template alignment and the per-residue and the global model quality was assessed using the QMEAN and Global Model Quality Estimate (GMQE) scoring functions. The GMQE score gives an idea about the accuracy of tertiary structure of the protein models. The QMEAN however, details on the quality of the submitted model based on its physicochemical properties and then generates a value referring to the overall quality of the structure. The obtained models were validated for their accuracy using ProSA, PROCHECK and Verify3D (21-23). PROCHECK analyzes the stereo chemical quality of the models based on the phi/psi angle arrangement and constructs a Ramachandran plot with residues being highlighted in favorable or unfavorable zones. ProSA calculates the potential energy by comparing them with the experimental structures submitted in pdb. Verify3D evaluates the local quality of the protein model on the basis of structure-sequence compatibility to generate a compatibility value for each residue of the protein. If a given model has 80% of their residues with a 3D-1D score equal to or higher than 0.2, it is considered to be a high quality structure.

Virtual screening and molecular docking

In order to perform a structure-based virtual screening e-LEA3D, (http://chemoinfo.ipmc.cnrs.fr/) which uses PLANTS algorithm was used (24). We screened across the list of FDA approved drugs. The docking score provided by e-LEA3D was used to screen the drugs for further analysis. From the docking scores the top four drugs, which had a higher docking score, were chosen further for the MD analysis.

Molecular Dynamics Simulations:

MD simulation was performed for the free protein and the drug docked protein complexes. To mimic the physiological state of the protein molecules, the simulation was performed for 5 ns in water, using the CHARMM27 all-atom force field (CHARM22 plus CMAP for proteins) (25). Modeling of the topology files of the drugs was done using the SwissParam web server (26). The system was made electrostatically neutral by adding the counter ions and the solvation of complexes was done within the SPC water cube (27).

The steepest descent method, an easy and robust algorithm, was used to carry out energy minimization of the whole system involving multiple steps. When increasing the temperature, which had a time period of 100ps, 300 K was kept as the upper bound to the temperature of the whole system. NPT (Constant pressure and temperature) and NVT (Constant volume and temperature) equilibration were performed in order (28). Rg (The radius of Gyration), H-bonding (Intermolecular Hydrogen Bonding), RMSF (Root Mean Square Fluctuations), RMSD (Root Mean Square Deviation), and SASA (Solvent- Accessible Surface Area) was calculated using the trajectory file of the system that was simulated in order to study the protein-drug complexes' behavior (29).

Results

The spike subunit, which binds to the ACE2 receptor, is directly related to the transmission and infection rate of the virus. Over the time, the mutation in the spike region has affected the infection rate and has enabled the emergence of several variants of COVID 19. The mutational diversity and burden is directly correlated to the evolution of viruses as a whole (Figure 1A). We have studied all these variants and their mutational diversity in different geographical locations to gain more information on the evolution of the virus.

Overall Spike protein is composed of several subunits: FP, fusion peptide; HR1, heptad repeat 1; HR2, heptad repeat 2; IC, intracellular domain; NTD, N-terminal domain; SD1, subdomain 1; SD2, subdomain 2; TM, trans membrane region as shown in Figure 1B. In this work, we study the mutational frequencies across different domains of the Spike protein (Figure 1C). Out of these, mutations with normalized mutation frequency >0.6 are provided in Supplementary File 2. We identified various stretches of amino acid residues, which are highly prone to mutations. Notably, the residues K417, L452, E484, and N501 were the mutation hotspots in the receptor-binding domain (RBD) across different geographical locations and in different spike variant. Similarly, the residues L18, W152, L241, L242, A243, I68, H69, V70 and D215 were the residues in the N terminal domain, which had the higher frequency of mutation. Interestingly enough, there was a zone of three consecutive amino acid residues across the stretch 241-243, which was the hotspot for mutation in the spike protein. The residue D614 was mutating across all variants in all the geographical locations and was thereby considered for the further analysis. A pictorial representation of these hotspots along with their occurrence in different domains in the spike has been shown in Figure 1. The location of these residues across the spatial domain in the spike subunit sheds light on the functionality as well as correlates with the transmissibility and the ability of the virus to evolve.

Identification of mutational Hotspots and geographical diversity in Spike Subunit

Different spike variants were grouped according to the region and their mutational burden was studied at the nucleotide and amino acid level. Figures 2A and 2B shows the heat map of the mutational hotspots in spike arranged according to their occurrence in different geographical locations grouped on the variant level both on amino acid as well as on nucleotide respectively. Similarly, Figures 2C and 2D are grouped in accordance with their occurrence in different geographical locations. The heat maps are colored based on the normalized mutation frequency in different regions in spike. The individual mutation frequency plots for different variants grouped by geographical locations are shown in Supplementary Figure 1. Figures 2E and 2F on the other hand show the mutational frequency for the combined set of all the sequences used in this study.

Structural Variants in Spike

We tried to look closely at the mutation hotspots in the spike subunit, which affect the stability, and thereby the transmission and the infection rate of the virus as a whole. The mutation was studied both in nucleotide as well as in the amino acid level in the Spike protein. We further tried to look at the evolution of spike in a spatio-temporal manner and how the mutational burden on the spike affects its functionality. In order to achieve this we took the omicron variant which has the highest frequency of mutation (Figure 3D) and compared to the reference spike sequence obtained from the initial strain (Figure 3A). In order to increase the diversity, we also studied the other variants for their mutational burden and picked the highly occurring non-omicron variant (Belongs to the Alpha strain- Spain/MD-11674437/2021|EPI_ISL_882632) which was found to have the highest number of mutations in Spike (39 mutations) among all the geographical sub locations (Figure 3C). These variants were used to understand the evolution of spike and its correlation with the stability and functionality. We also synthetically reconstructed the spike with a single-mutation D→G at position 614, which was occurring consistently in all the variants across all the geographical locations (Figure 3B). This synthetic variant was constructed to study how the single mutation, which is consistently present across all the variants, affects the functionality.

Drug Docking and MD Simulation

A list of FDA approved drugs were virtually screened over the spike protein. The drugs Alvimopan, Elvitegravir, Bictegravir and Nebivolol, which had the highest docking score, were chosen for further analysis to see how their docking properties changed due to the mutations in different Spike subunit. The superimposed images of pre MD and post MD proteins are shown in Figure 4.

D614G- Single Variant Mutant

Root mean square deviation (RMSD)

The stability of the ligand docked protein complex can be inferred from calculating the Root mean square deviation (RMSD). RMSD stability is an indicator of complex stability and is used to quantify the difference between the initial and final positions of the protein backbones.

The RMSD plots of D614G mutated protein reveal that the D614G-Nebivolol complex remains stable after attaining a peak at 1ns while the D614G-Elvitegravir complex continues to remain stable from 1.2ns. The Free protein and D614G-Alvimopan remain stable except for a couple of peaks and troughs (Figure 7A).

The radius of gyration (Rg)

The radius of gyration plots reveals that the compact nature of the protein is unaffected by the docked drug molecules. The mean radius of gyration values for Free protein, D614G-Alvimopan complex, D614G-Elvitegravir, and D614G-Nebivolol is 5.378 nm, 5.336 nm, 5.305 nm, and 5.376 nm respectively. These minor changes are an effect of the conformational changes due to the drug docking. The differences in mean values are hence negligible as they are well within the standard deviation of the respective complexes. This indicates that the docking of the drug molecules does not seem to affect the secondary structure of the protein molecule. The Free-protein and D614G-Nebivolol complex has an almost similar radius of gyration plots. While the D614G-Alvimopan and D614G-Elvitegravir complexes differ slightly from the Free-protein radius of gyration plots

(Figure 5A). Hence, the docking of Nebivolol to the D614G protein has affected the protein structure to a lesser extent when compared to the docking of Alvimopan and Elvitegravir.

Intermolecular hydrogen bonding

The number of hydrogen bonds is a crucial parameter to study the affinity of the protein and the drug to bind. The number of hydrogen bonds will indicate the strength of the bond between the protein and the docked drug molecule. Hence, more hydrogen bonds will stipulate that there is a strong binding between the drug and protein molecules. From the plot, we can infer the maximum number of hydrogen bonds formed by the drug molecules Alvimopan, Elvitegravir, and Nebivolol is 4,3, and 2 respectively. The plot indicates that Alvimipan binding affinity is more than Elvitegravir and Nebivolol (Figure 6A). The formation of Hydrogen bonds indicates the presence of an affinity between the drug and the protein.

Root mean square fluctuations (RMSF)

The fluctuations and rigidity of the docked protein complexes can be quantified using the root mean square values (RMSF). RMSF helps to study the conformational flexibility of the docked protein complexes, as it is the measurement of deviations of the residue from the initial position.

The highest fluctuations are observed at 20000-25000 and 38000-42000 atoms stretch. The mean RMSF values of the Free protein, D614G-Alvimopan, D614G-Elvitegravir, and D614G-Nebivolol is 0.233 nm, 0.238 nm, 0.214 nm, and 0.251 nm respectively (Figure 8A). These values indicate that the binding of the three drugs Alvimopan, Elvitegravir, and Nebivolol preserves the flexibility of the protein.

Solvent Accessible Surface Area analysis (SASA)

Performing solvent accessible surface area analysis (SASA) helps to study the solvent hydrophilic and hydrophobic nature of the docked protein complexes. The mean SASA values for Free protein, D614G-Alvimopan, D614G-Elvitegravir, and D614G-Nebivolol are 1500.31 nm^2, 1504.92

nm^2, 1494.40 nm^2, and 1510.66 nm^2 respectively (Figure 9A). These values suggest that the docked protein complexes are solvated after the binding, except for the binding of Elvitegravir which has a lower value than the Free-protein.

Spike-Initial Strain

Root mean square deviation (RMSD)

The RMSD plots of Spike protein reveal that the Spike-Bictegravir complex remained almost stable from 2 ns except for two peaks at 3 ns and 4.5 ns. While the Free protein and Spike-Alvimopan complex begin to stabilize around 1.3 ns and fluctuate around 2.5 ns. Thereafter they continue to reach the stable state around 3 ns (Figure 7B).

The radius of gyration (Rg)

The mean values of Free-protein, Spike-Alvimopan, and Spike-Bictegravir are 5.30 nm, 5.36 nm, and 5.33 nm respectively. From the radius of gyration plots and the mean value references, we infer that the docking of Alvimopan affects the stability of the protein stability when compared to the docking of Bictegravir (Figure 5B).

Intermolecular hydrogen bonding

From the plot, we can infer the maximum number of hydrogen bonds formed by the drug molecules Alvimopan and Bictegravir is 4 and 3 respectively. These results indicate that Alvimopan has a higher binding affinity as compared to Bictegravir (Figure 6B).

Root mean square fluctuations (RMSF)

The highest fluctuations for the spike protein were observed at 38000-42000 atoms stretch. The mean RMSF values of the Free protein, Spike-Alvimopan, and Spike-Bictegravir is 0.229 nm, 0.227 nm, and 0.221 nm respectively. These values indicate that the binding of the drugs Alvimopan, and Bictegravir preserves the flexibility of the protein (Figure 8B).

Solvent Accessible Surface Area analysis (SASA)

The mean SASA values for Free protein, Spike-Alvimopan, and Spike-Bictegravir are 1504.99 nm^2, 1507.10 nm^2, and 1506.54 nm^2 respectively. Both the docked protein complexes are observed to have solvated better than the Free-protein from the results of SASA (Figure 9B). These differences in the SASA values can be attributed to the relatively large values of the radius of gyration of the complexes.

Omicron

Root mean square deviation (RMSD)

The RMSD plot for the omicron variant indicates that the drug-protein complexes have a stabilized RMSD profile similar to that of the free protein. However, one has to note that the magnitude of RMSD with the drug protein complexes is higher than that of the free protein (Figure 7C). While the mean RMSD of the free protein was 0.31 A, the mean RMSD of Omicron-Alvimopan, Omicron, Bictegravir and Omicron-Elvitegravir were 0.61 A, 0.64 A and 0.34 A respectively. Therefore, it is to be noted that the docking of Bictegravir and Alvimopan has slightly affected the stability of the protein.

The radius of gyration (Rg)

The Radius of Gyration plots of the Omicron protein complexes are slightly higher compared to the mean Radius of Gyration of the free protein (4.90). This indicates that the docking of ligands has not affected the compactness of the protein. The mean radius of gyration of the protein-ligand complexes are 4.95 nm, 4.98 nm and 4.92 nm for Omicron-Alvimopan, Omicron, Bictegravir and Omicron-Elvitegravir respectively (Figure 5C).

Intermolecular hydrogen bonding

From the plot, we can infer the maximum number of hydrogen bonds formed by the drug molecules Alvimopan, Bictegravir, and Elvitegravir are 4, 1, and 3 respectively (Figure 6C). The mean value of the number of intermolecular hydrogen bonds is more than two for Alvimopan while for the other cases it is lesser suggesting that Alvimopan has a higher affinity for the protein molecule.

Root mean square fluctuations (RMSF)

The RMSF plots suggest that the Omicron-Elvitegravir and Omicron Free protein simulations have a similar profile. However, the Omicron-Bictegravir complex is found to have a very high RMSF illustrating very high flexibility (Figure 8C). This is in line with the RMSD plots which indicate a very high RMSD for the Omicron-Bictegravir complex and Omicron-Alvimopan complex.

Solvent Accessible Surface Area analysis (SASA)

The mean SASA values for Free protein, Omicron-Alvimopan, D614G-Elvitegravir, and Omicron-Bictegravir are 1206.98 nm^2, 1468.06 nm^2, 1450.61 nm^2, and 1215.67 nm^2 respectively (Figure 9C). The high SAS values indicate that all the complexes are adequately solvated. The Omicron-Alvimopan and Omicron-Bictegravir complexes have unusually high levels of solvation indicating that the docking of the drug has interfered with the conformation of the protein.

Spain Variant (Spain/MD-11674437/2021|EPI_ISL_882632)

Root mean square deviation (RMSD)

The RMSD of Alvimopan, Elvitegravir and Nebivolol docked Spain variant protein are drastically higher than that of the free protein. The difference in the profiles illustrate that the docking of the ligands has caused significant destabilization to the structure of the protein. The mean RMSD of the Free protein, the Spain-Alvimopan complex, Spain-Elvitegravir complex and Spain-Nebivolol complex are 0.88, 11.42, 11.39 and 11.36 A respectively (Figure 7D).

The radius of gyration (Rg)

The respective mean values of radius of gyration of Free protein, Spain variant-Alvimopan complex, Spain variant-Elvitegravir, and Spain variant-Nebivolol are 5.246 nm, 5.250 nm, 5.279 nm, and 5.234 nm. These reveal that the docking of the drug molecules has not affected the structure of the protein much (Figure 5D). This is due to the fact that the differences in the mean values are well within the standard deviation of the respective complexes.

Intermolecular hydrogen bonding

From the plot, we can infer the maximum number of hydrogen bonds formed by the drug molecules Alvimopan, Elvitegravir, and Nebivolol with the Spain variant are 7, 3, and 2 respectively. The mean value of the number of intermolecular bonds is three for Alvimopan while it is lesser for other drug molecules (Figure 6D). This indicates that Alvimopan has a higher affinity to the Spain variant protein when compared to Elvitegravir and Nebivolol. From the plot, it is also possible to infer that more hydrogen bonds begin to form from the beginning for Alvimopan.

Root mean square fluctuations (RMSF)

The highest fluctuations for the spike protein were observed at 17000-22000 atoms stretch. The mean RMSF values of the Free protein, Spain variant-Alvimopan, Spain variant-Elvitegravir, and Spain variant-Nebivolol is 0.359 nm, 0.361 nm, 0.412 nm, and 0.388 nm respectively (Figure 8D). These values suggest that the binding of the drugs Alvimopan, and Bictegravir may lead to loss of flexibility.

Solvent Accessible Surface Area analysis (SASA)

The mean SASA values for Free protein, Spain variant-Alvimopan, Spain variant-Elvitegravir, and Spain variant-Nebivolol are 1877.71 nm^2, 1850.94 nm^2, 1834.03 nm^2, and 1877.73 nm^2 respectively (Figure 9D). These values illustrate that the docked protein complexes are solvated after the binding, except for the binding of Elvitegravir, which has a lower value than the Free-protein.

Discussion

The mutational frequency across the spike subunit is directly related to its infection and transmission rates. From our previous study, the spike subunit was characterized as unstable with higher number of mutations among all regions (30). Over the time period there have been multitudes of mutations in spike, some beneficial and some that had no or less effect on its binding with the ACE2 receptor. This accumulation of mutation in the spike has resulted in the evolution of new strains of the virus as well as have led to multiple waves across different geographical locations. The characterization of the mutational hotspots is important in determining the course of virus transmission, reinfection and mapping the efficiency of therapeutics across different variants. Previous studies have indicated the mutations Q493R, N501Y, S371L, S373P, S375F, Q498R, and T478K as the ones that significantly affect the binding with the ACE2 receptor (31). The evolution of the recent Omicron strain has definitely increased the speculation of increasing the diversity of mutations in the different sub regions of spike protein.

In the current study, we studied what makes the spike protein more susceptible to mutations in comparison to the rest of the viral genome and how different variants of the spike affect the stability and the functionality of the virus as a whole. We have also tried to narrow down the mutation hotspots to the different domains of the spike protein. Thereby, on studying the spike sequences from the different geographical locations we observed that there exists a unique signature pattern in spike that can be used as markers for the different variants. We observed that the recently emerged Omicron variant has the highest number of evenly distributed frequently occurring mutations across different subdomains of the spike protein. Epsilon and the Gamma variants have the least number of mutations. However, owing to the higher transmissibility of these variants these mutations tend to occur in critical positions that affect the binding of spikes with

ACE2 receptors. Alpha variant on the other hand has mutations, which occur at different frequencies across the similar residues in different geographical locations. This could possibly explain the difference in infection rate and the intensity of the first wave across different countries. Delta subunit has a greater accumulation of mutation in the N Terminal domain and these occur at different frequencies across the globe. The position 614 was observed to be mutated across all the subunits as well as across all the geographical locations. Interestingly, we observed the highest number of mutations in the receptor-binding domain of the Omicron variant whereas the lowest number of mutations were observed in the Alpha, which could possibly correlate with the evolution of the virus over the time scale. The omicron variant in South America has many low frequency mutations which could be attributed to bad quality sequences. Consistent with the previous studies (31), residue number 501 in the receptor binding domain and at position 614 had a greater mutational probability. This defines the unique signature pattern of every variant which is essential in understanding the past course of the infection and which would also give an estimation of further mutation accumulation and changes in the transmissibility of the virus.

Furthermore we wanted to study how the spike selection varied across the globe on spatio-temporal scale. Thereby, we investigated selected variants of the spike based on their abundance as well as their unique signature to evaluate their stability and their functionality through drug docking and MD simulations. These variants were docked with the same set of drugs to see how the free protein profile varied with the drug docked profile. The properties and strength of the drug docked protein complexes and their conformational analysis were studied by performing MD simulations. Calculation of RMSD, RMSF, Intermolecular hydrogen bonds, Radius of gyration, and SASA parameters gave insights into the structural stability of the proteins. The MD simulations were carried out for 5 ns to illustrate the dynamics and to study the conformational stability of the drug docked complexes. The Spike protein from the initial strain was stabilized when docked with Alvimopan and Bictegravir drugs. The D614G mutated protein was stabilized when docked with the drugs Nebivolol, Alvimopan, and Elvitegravir. Whereas, the Omicron protein tends to be stabilized when docked with Elvitegravir but destabilized when docked with Alvimopan and Bictegravir. All three drug candidates Alvimopan, Elvitegravir, and Nebivolol affected the stability of the Spain variant protein molecule. However, they did not affect the conformational properties of the protein in most cases.

Conclusion

Emergence of different variants of SARS-CoV-2 have always been a threat to the therapeutic measures. Spike Protein, which binds to the ACE2 receptor, has been a site for high mutational diversity. These mutations if occurring in critical regions, which affect the binding with ACE2 receptors, can affect the overall infection as well as the transmissibility of the virus. The current study highlights the regions in spike susceptible to mutations that affects its functionality. One of the important aspects of this study lies in the identification of unique "fingerprint" patterns of mutation in spike, which shape the identity of different variants across the globe. The drug docking approach to different spike enables us to have a clear correlation on how the mutations affect the binding of Spike with ACE2 and how this offers a differential selectivity to each variants and the ones to come.

Conflict of interest

The authors declare no conflict of interest. Funding: The authors did not receive funding from any source for this work.

Acknowledgement

We would like to thank GISAID (Acknowledgement table in the supplementary file 1) for the database of SARS-CoV-2 genome sequences. The authors would like to thank Ishita Agarwal and Swaroopa Nakkeeran from Indian Institute of Science Education and Research, Mohali for their help with data mining and literature review. DHL, NMA, MS and DB are supported by the INSPIRE scholarship (https://online-inspire.gov.in/). A.J. is supported by UKRI-BBSRC grant BB/M011224/1, Clarendon Fund (University of Oxford), and Oxford Interdisciplinary Bioscience DTP. AKP is supported by the DFG Grant from TUM and GSN LMU Munich.

Author's contribution

AKJ and AKP conceptualized and designed the project. ME and MS retrieved the data from the GISAID. DHL performed mutation frequency and variant analysis. AKP performed the docking analysis and the drug interaction study. NMA, AKJ and DB worked on MD simulation and its analysis. AKP, NMA, DHL, AKJ and ME wrote the manuscript. AKJ and AKP provided intellectual support in interpreting the results and editing the manuscript.

Funding Information

The authors have not received any funding for this work. The work was conducted on mere scientific interest.

Figure Legends

Figure 1. A. Diversity of mutation across the SARS-COV-2 viral genome. B. Spike subunits and the hotspot of mutation occurring across different domains marked in Gray FP, fusion peptide; HR1, heptad repeat 1; HR2, heptad repeat 2; IC, intracellular domain; NTD, N-terminal domain; SD1, subdomain 1; SD2, subdomain 2; TM, trans membrane region C. Structure of Spike subunit highlighting the highly mutable residues across different subunit. For illustration purposes, the NTD is marked in blue, the receptor-binding domain in green and the S2 subunit in orange.

Figure 2. Heat maps showing the normalized mutational frequencies across different variants A. on amino acid level B. on nucleotide level. Across different geographical locations C. on amino acid level and D. on nucleotide level. Mutation frequencies across different E. amino acid residues of spike studied from 5492 sequences and F. nucleotide position of spike sequence studied from 6100 sequences.

Figure 3: Illustration of the spike constructed from the 4 different type of variant with NTD marked in green, receptor binding domain marked in blue and the S2 subunit marked in orange A. Spike protein obtained from the initial strain B. D614G mutant synthetically reconstructed spike C. Spike from Spain/MD-11674437/2021|EPI_ISL_882632 D. Spike from the omicron strain

Figure 4. An image illustrating superimposition of pre MD (pink) and post MD (blue) structures of Free protein and docked protein complexes of A. D614G protein B. Spike Protein C. Omicron Protein D. Spain Variant Protein.

Figure 5. The radius of gyration plots over the entire simulation, where the ordinate is Rg (nm) and the abscissa is time (ps) with Free protein and docked protein complexes of A. D614G protein B. Spike Protein C. Omicron Protein D. Spain Variant Protein.

Figure 6. The intermolecular hydrogen bond plots with docked protein complexes of A. D614G protein B. Spike Protein C. Omicron Protein D. Spain Variant Protein.

Figure 7. The root-mean-square deviation plots of the atoms with Free protein and docked protein complexes of A. D614G protein B. Spike Protein C. Omicron Protein D. Spain Variant Protein.

Figure 8. The root-mean-square fluctuation plots over the entire simulation, where the ordinate is RMSF (nm) and the abscissa is residue position with Free protein and docked protein complexes of A. D614G protein B. Spike Protein C. Omicron Protein D. Spain Variant Protein.

Figure 9. The solvent-accessible surface area analysis (SASA), where the ordinate is SASA (nm2) and the abscissa is time (ps) with Free protein and docked protein complexes of A. D614G protein B. Spike Protein C. Omicron Protein D. Spain Variant Protein.

Supplementary Figure 1: Mutation frequency in different variants grouped by geographical locations.

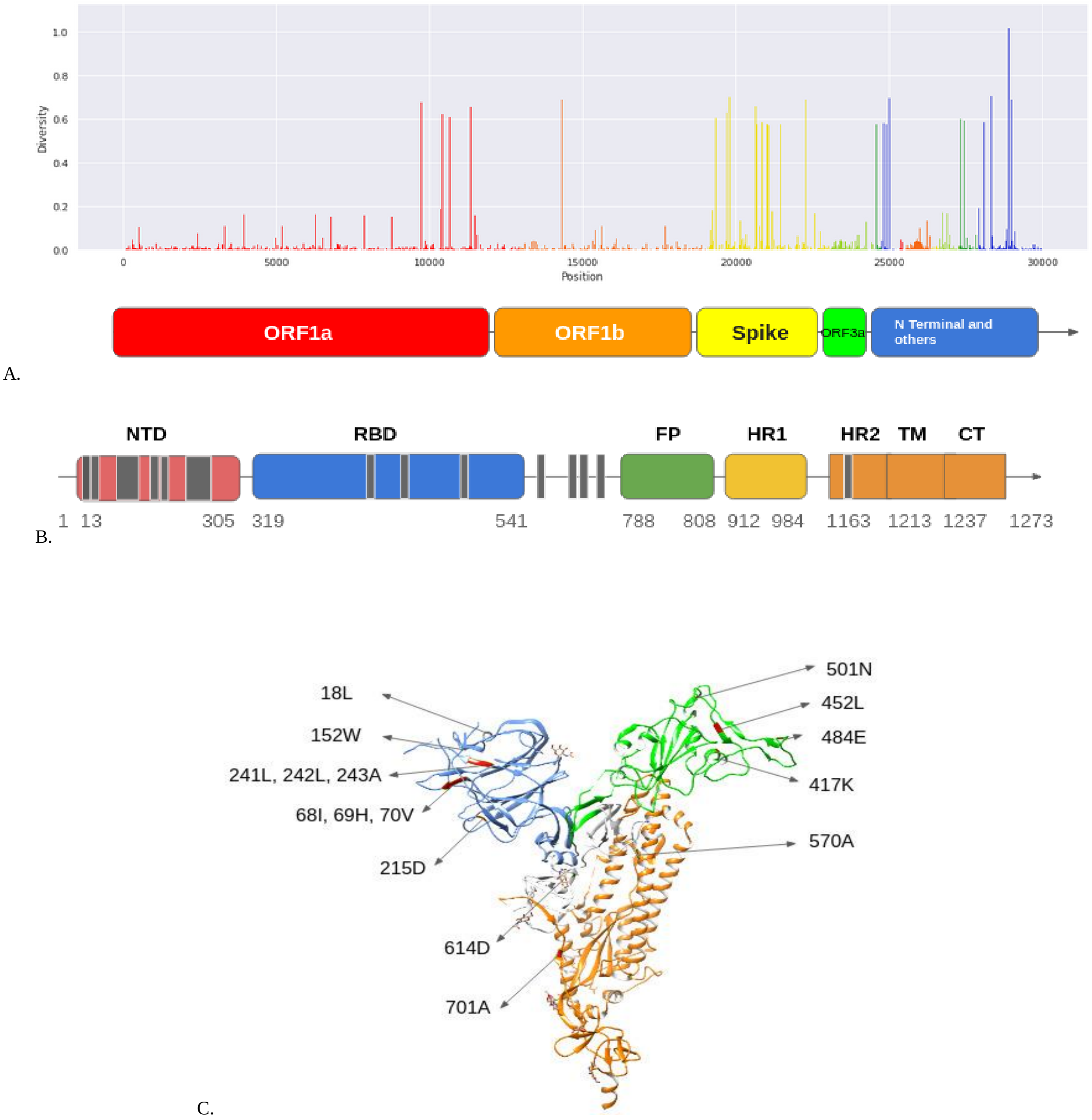


Figure 1. A. Diversity of mutation across the SARS-COV-2 viral genome. B. Spike subunits and the hotspot of mutation occurring across different domains marked in Gray FP, fusion peptide; HR1, heptad repeat 1; HR2, heptad repeat 2; IC, intracellular domain; NTD, N-terminal domain; SD1, subdomain 1; SD2, subdomain 2; TM, trans membrane region C. Structure of Spike subunit highlighting the highly mutable residues across different subunit. For illustration purposes, the NTD is marked in blue, the receptor-binding domain in green and the S2 subunit in orange.

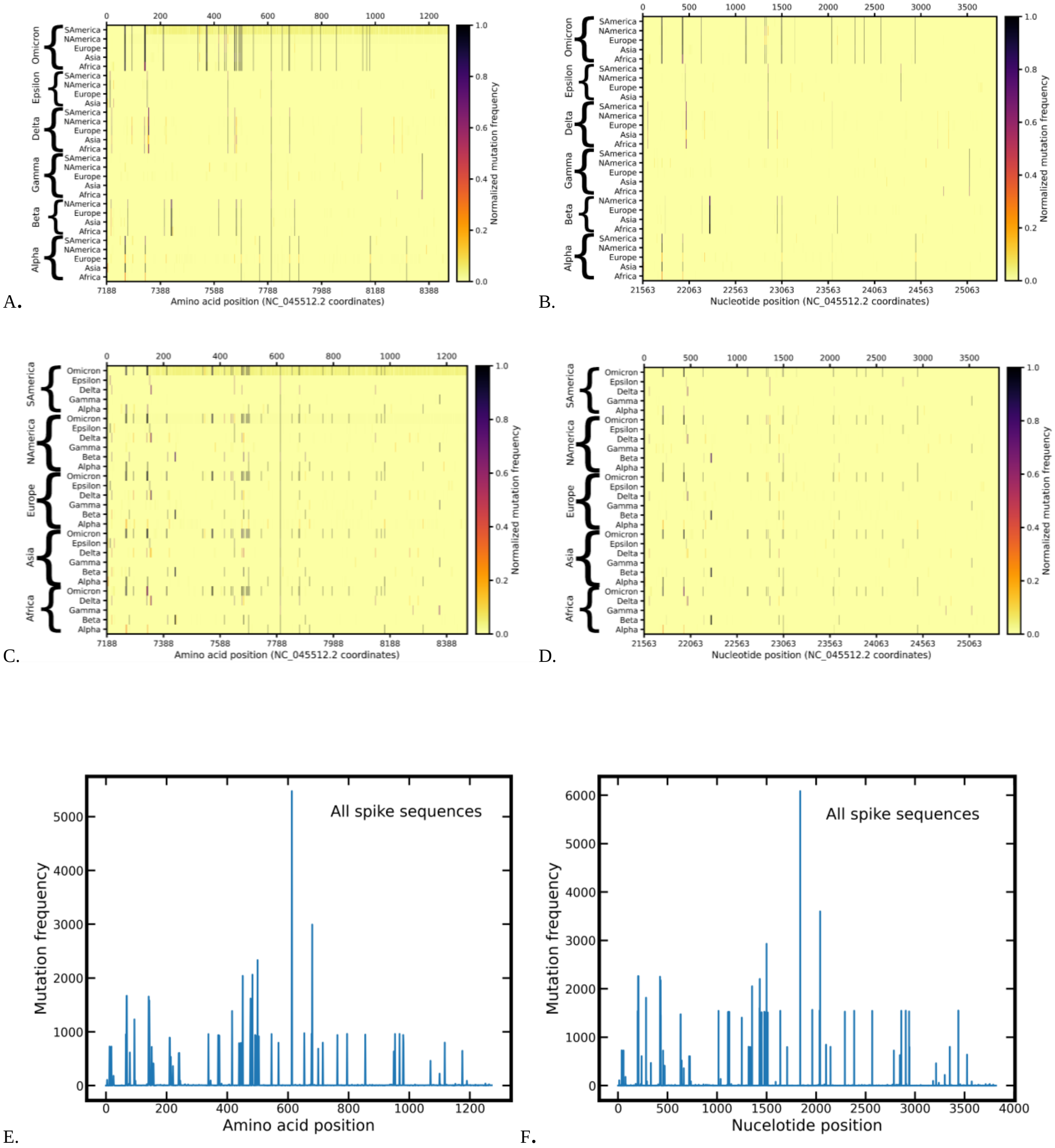


Figure 2. Heat maps showing the normalized mutational frequencies across different variants A. on amino acid level B. on nucleotide level. Across different geographical locations C. on amino acid level and D. on nucleotide level. Mutation frequencies across different E. amino acid residues of spike studied from 5492 sequences and F. nucleotide position of spike sequence studied from 6100 sequences.

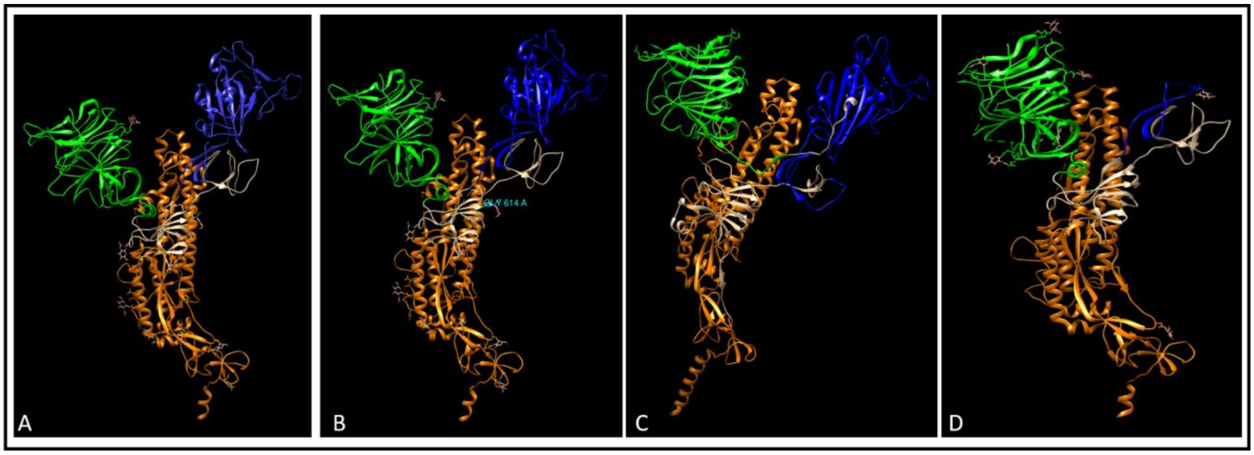


Figure 3: Illustration of the spike constructed from the 4 different type of variant with NTD marked in green, receptor binding domain marked in blue and the S2 subunit marked in orange A. Spike protein obtained from the initial strain B. D614G mutant synthetically reconstructed spike C. Spike from Spain/MD-11674437/2021|EPI_ISL_882632 D. Spike from the omicron strain

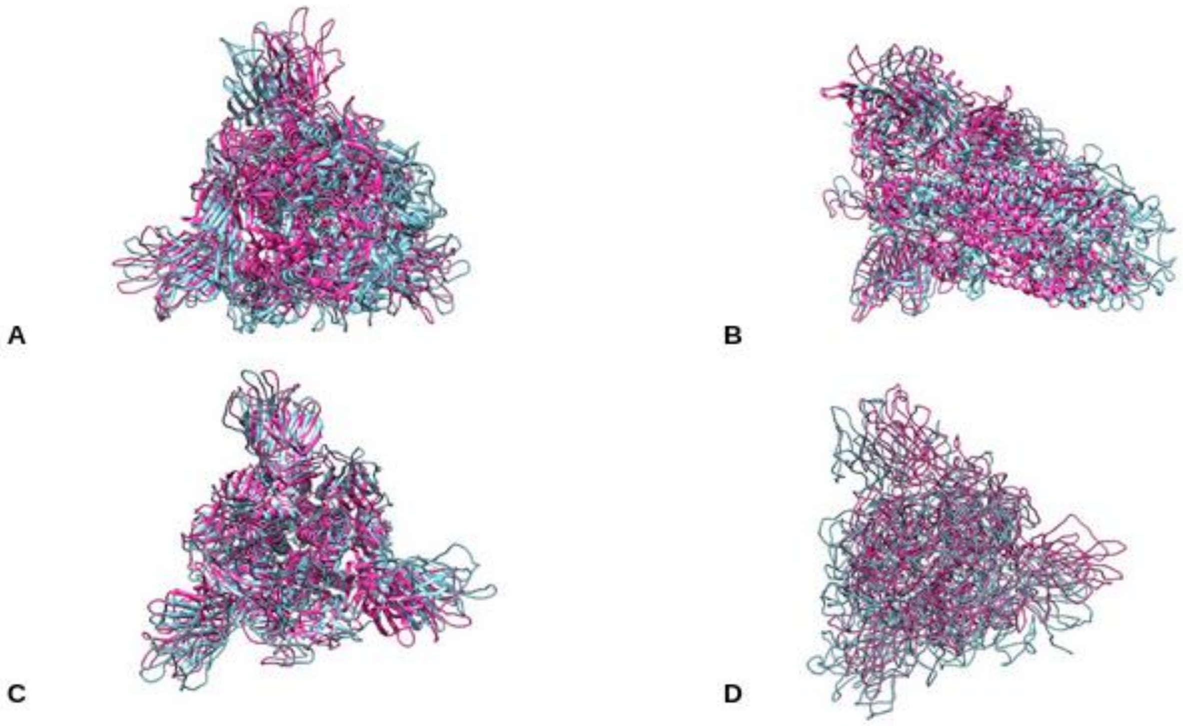


Figure 4. An image illustrating superimposition of pre MD (pink) and post MD (blue) structures of Free protein and docked protein complexes of A. D614G protein B. Spike Protein C. Omicron Protein D. Spain Variant Protein.

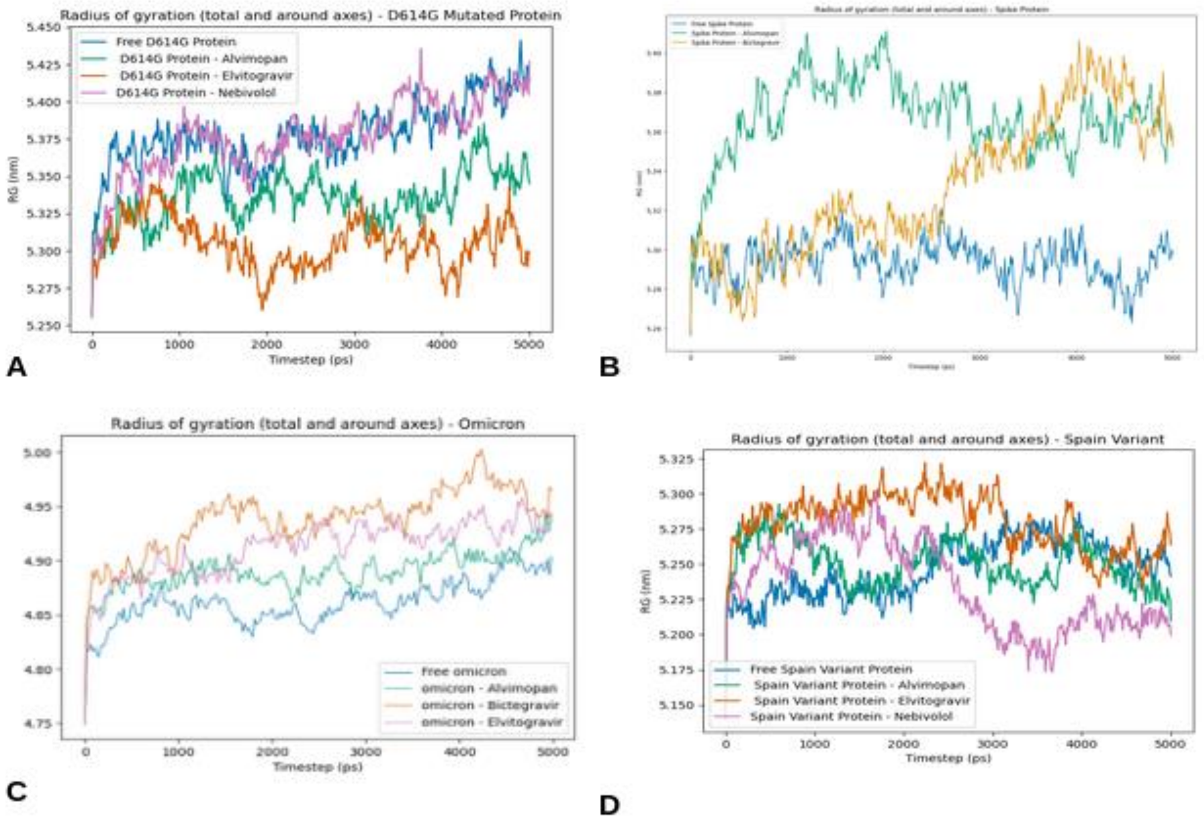


Figure 5. The radius of gyration plots over the entire simulation, where the ordinate is Rg (nm) and the abscissa is time (ps) with Free protein and docked protein complexes of A. D614G protein B. Spike Protein C. Omicron Protein D. Spain Variant Protein.

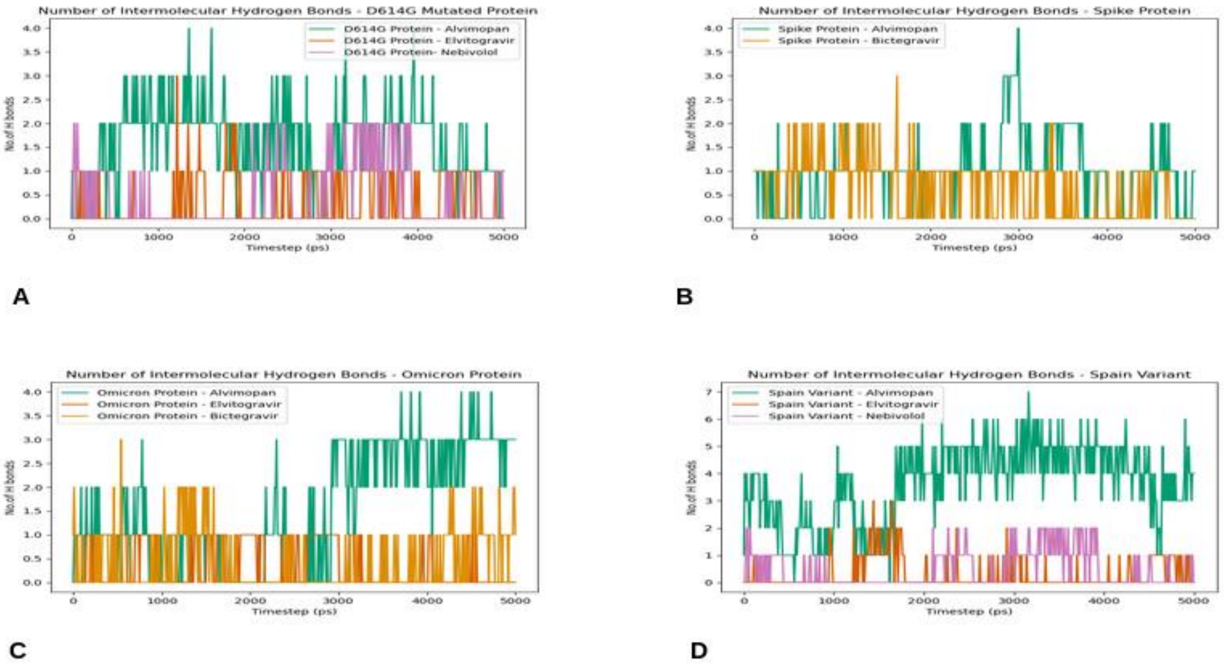


Figure 6. The intermolecular hydrogen bond plots with docked protein complexes of A. D614G protein B. Spike Protein C. Omicron Protein D. Spain Variant Protein.

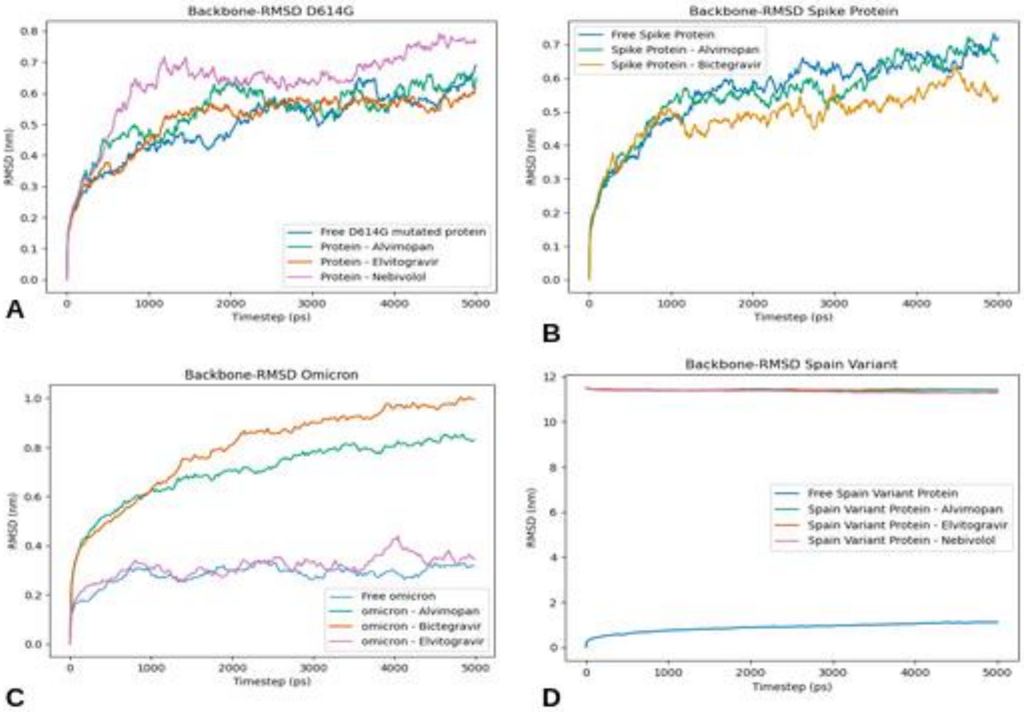


Figure 7. The root-mean-square deviation plots of the atoms with Free protein and docked protein complexes of A. D614G protein B. Spike Protein C. Omicron Protein D. Spain Variant Protein.

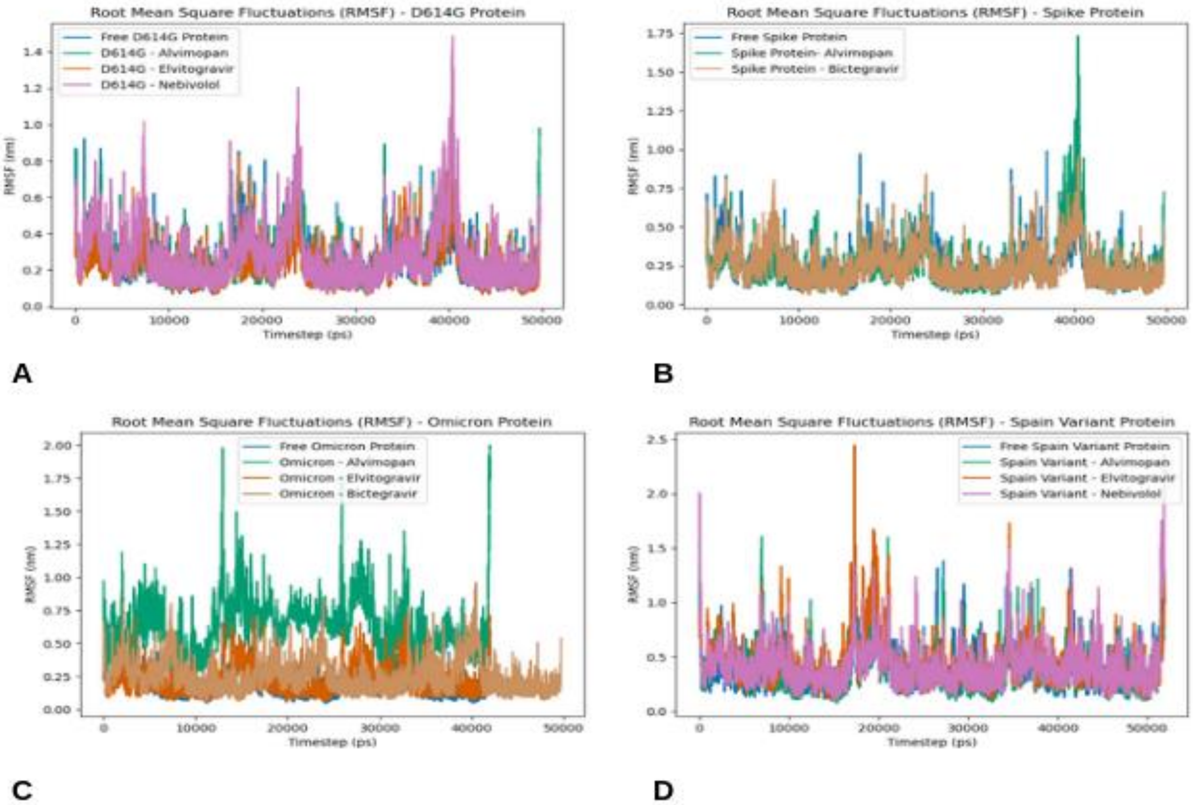


Figure 8. The root-mean-square fluctuation plots over the entire simulation, where the ordinate is RMSF (nm) and the abscissa is residue position with Free protein and docked protein complexes of A. D614G protein B. Spike Protein C. Omicron Protein D. Spain Variant Protein.

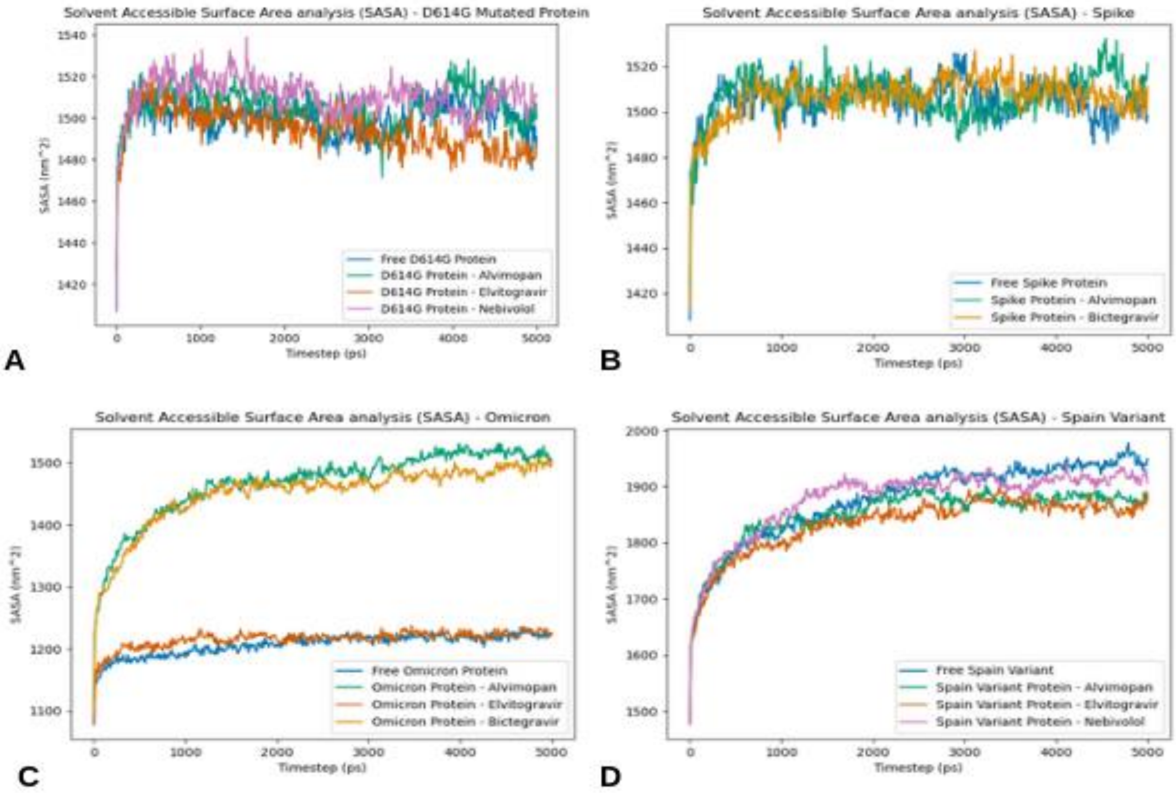
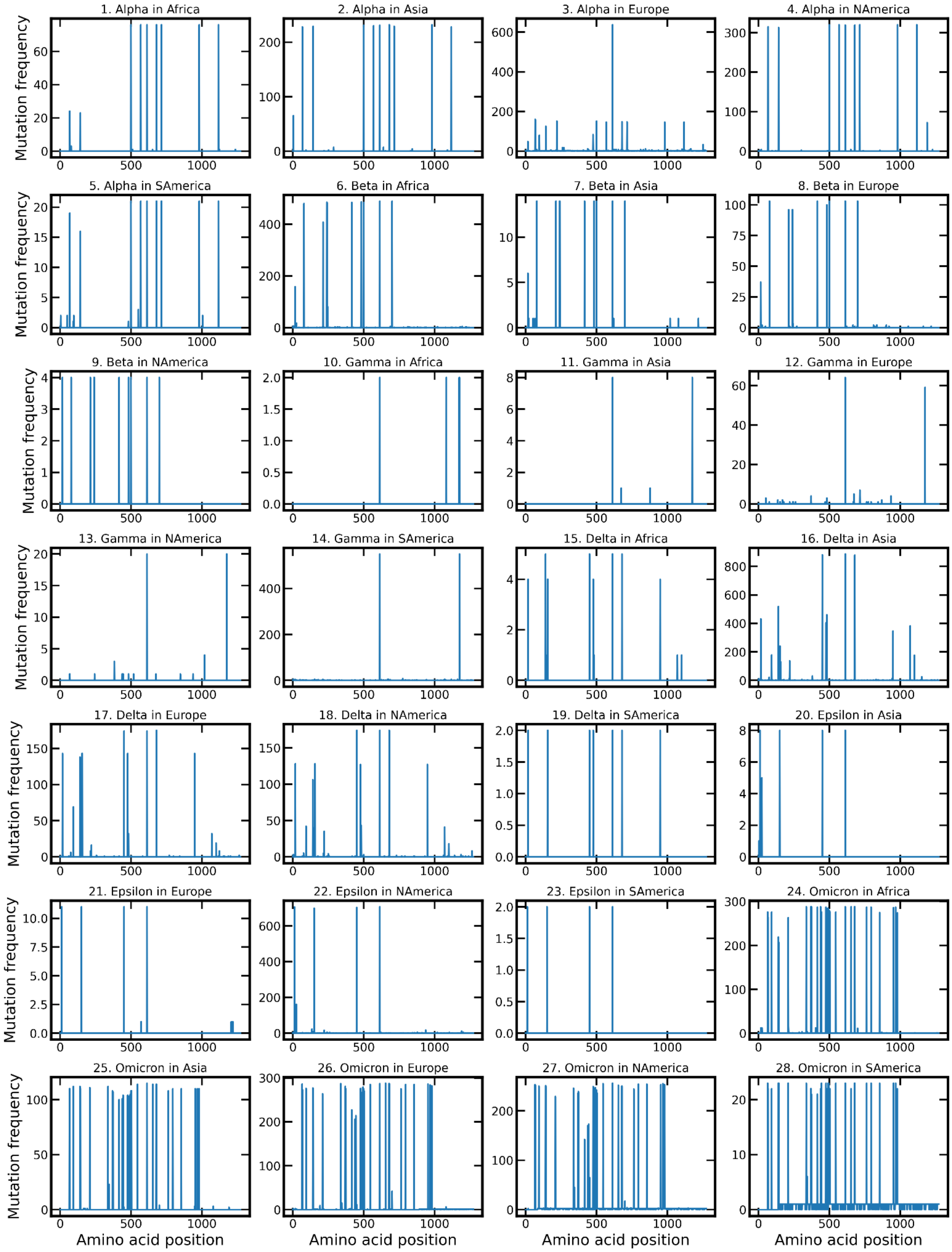


Figure 9. The solvent-accessible surface area analysis (SASA), where the ordinate is SASA (nm2) and the abscissa is time (ps) with Free protein and docked protein complexes of A. D614G protein B. Spike Protein C. Omicron Protein D. Spain Variant Protein.



Supplementary Figure 1: Mutation frequency in different variants grouped by geographical locations

Table and table legends

Variant \ Location	Alpha	Beta	Gamma	Delta	Epsilon	Omicron	Per Location Total
Africa	76	489	2	5	-	288	860
Asia	232	14	8	887	8	115	1264
Europe	648	103	64	175	11	288	1289
N. America	320	4	20	174	707	256	1481
S. America	21	-	550	2	2	23	598
Per Variant Total	1297	610	644	1243	728	970	**5492**

Table 1: Break down of the amino acid sequences used in this study after filtering according to the procedure described in the Data preprocessing and alignment section.

Variant \ Location	Alpha	Beta	Gamma	Delta	Epsilon	Omicron	Per Location Total
Africa	96	575	2	5	-	339	1017
Asia	248	15	7	1036	7	189	1502
Europe	699	175	70	253	13	1391	2601
N. America	357	9	25	254	1155	521	2321
S. America	24	-	873	2	2	25	926
Per Variant Total	1424	774	977	1550	1177	2465	**8367**

Supplementary Table 1: List of sequences retrieved from GISAID broken down according to their geographical location and the variant type.

Variant \ Location	Alpha	Beta	Gamma	Delta	Epsilon	Omicron	Per Location Total
Africa	76	492	2	5	-	294	869
Asia	232	14	8	890	8	116	1268
Europe	650	106	64	179	11	849	1859
N. America	324	4	20	174	709	275	1506
S. America	21	-	550	2	2	23	598
Per Variant Total	1303	616	644	1250	730	1557	**6100**

Supplementary Table 2: Break down of the nucleotide sequences used in this study after filtering according to the procedure described in the Data preprocessing and alignment section.

References

1. V. Zoumpourlis, M. Goulielmaki, E. Rizos, S. Baliou, D. A. Spandidos, [Comment] The COVID-19 pandemic as a scientific and social challenge in the 21st century. *Molecular medicine reports* **22**, 3035-3048 (2020).
2. C. S. G. of the International, The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature microbiology* **5**, 536 (2020).
3. Worldometers.info (25 February, 2022) (Dover, Delaware, U.S.A.).
4. F. Wu *et al.*, A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269 (2020).
5. A. A. T. Naqvi *et al.*, Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *Biochimica et biophysica acta. Molecular basis of disease* **1866**, 165878-165878 (2020).
6. S. Kumar, V. K. Maurya, A. K. Prasad, M. L. B. Bhatt, S. K. Saxena, Structural, glycosylation and antigenic variation between 2019 novel coronavirus (2019-nCoV) and SARS coronavirus (SARS-CoV). *VirusDisease* **31**, 13-21 (2020).
7. A. Shajahan, L. E. Pepi, D. S. Rouhani, C. Heiss, P. Azadi, Glycosylation of SARS-CoV-2: structural and functional insights. *Analytical and Bioanalytical Chemistry* **413**, 7179-7193 (2021).
8. Y. Watanabe, J. D. Allen, D. Wrapp, J. S. McLellan, M. Crispin, Site-specific glycan analysis of the SARS-CoV-2 spike. *Science* **369**, 330-333 (2020).
9. W. T. Harvey *et al.*, SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology* **19**, 409-424 (2021).
10. L. Dai, G. F. Gao, Viral targets for vaccines against COVID-19. *Nature Reviews Immunology* **21**, 73-82 (2021).
11. M. J. Pereson *et al.*, Evolutionary analysis of SARS-CoV-2 spike protein for its different clades. *J Med Virol* **93**, 3000-3006 (2021).
12. D. Van Egeren *et al.*, Risk of rapid evolutionary escape from biomedical interventions targeting SARS-CoV-2 spike protein. *PLOS ONE* **16**, e0250780 (2021).
13. F. Grabowski, M. Kochańczyk, T. Lipniacki, Omicron strain spreads with the doubling time of 3.2—3.6 days in South Africa province of Gauteng that achieved herd immunity to Delta variant. *medRxiv* 10.1101/2021.12.08.21267494, 2021.2012.2008.21267494 (2021).
14. T. P. Peacock *et al.*, The SARS-CoV-2 variant, Omicron, shows rapid replication in human primary nasal epithelial cultures and efficiently uses the endosomal route of entry. *bioRxiv* 10.1101/2021.12.31.474653, 2021.2012.2031.474653 (2022).
15. K. P. Y. Hui *et al.*, SARS-CoV-2 Omicron variant replication in human bronchus and lung ex vivo. *Nature* 10.1038/s41586-022-04479-6 (2022).
16. M. Hoffmann *et al.*, The Omicron variant is highly resistant against antibody-mediated neutralization: Implications for control of the COVID-19 pandemic. *Cell* **185**, 447-456.e411 (2022).
17. D. Mannar *et al.*, SARS-CoV-2 Omicron variant: Antibody evasion and cryo-EM structure of spike protein-ACE2 complex. *Science* **375**, 760-764 (2022).
18. K. Katoh, J. Rozewicki, K. D. Yamada, MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in bioinformatics* **20**, 1160-1166 (2019).
19. P. J. Cock *et al.*, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423 (2009).

20. A. Waterhouse *et al.*, SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research* **46**, W296-W303 (2018).

21. J. U. Bowie, R. Lüthy, D. Eisenberg, A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164-170 (1991).

22. R. A. Laskowski, M. W. MacArthur, D. S. Moss, J. M. Thornton, PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of applied crystallography* **26**, 283-291 (1993).

23. M. Wiederstein, M. J. Sippl, ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic acids research* **35**, W407-W410 (2007).

24. D. Douguet, H. Munier-Lehmann, G. Labesse, S. Pochet, LEA3D: a computer-aided ligand design for structure-based drug design. *Journal of medicinal chemistry* **48**, 2457-2468 (2005).

25. K. Vanommeslaeghe *et al.*, CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry* **31**, 671-690 (2010).

26. V. Zoete, M. A. Cuendet, A. Grosdidier, O. Michielin, SwissParam: a fast force field generation tool for small organic molecules. *Journal of computational chemistry* **32**, 2359-2368 (2011).

27. H. Berendsen, J. Grigera, T. Straatsma, The missing term in effective pair potentials. *Journal of Physical Chemistry* **91**, 6269-6271 (1987).

28. A. D. Elmezayen, A. Al-Obaidi, A. T. Şahin, K. Yelekçi, Drug repurposing for coronavirus (COVID-19): in silico screening of known drugs against coronavirus 3CL hydrolase and protease enzymes. *Journal of Biomolecular Structure and Dynamics* **39**, 2980-2992 (2021).

29. R. J. Khan *et al.*, Targeting SARS-CoV-2: a systematic drug repurposing approach to identify promising inhibitors against 3C-like proteinase and 2′-O-ribose methyltransferase. *Journal of Biomolecular Structure and Dynamics* **39**, 2679-2692 (2021).

30. N. M. Anand *et al.*, A comprehensive SARS-CoV-2 genomic analysis identifies potential targets for drug repurposing. *Plos one* **16**, e0248553 (2021).

31. S. Kumar, T. S. Thambiraja, K. Karuppanan, G. Subramaniam, Omicron and Delta variant of SARS-CoV-2: a comparative computational study of spike protein. *Journal of medical virology* (2021).