# Structural Topological Analysis of Spike Proteins of SARS-CoV-2 Variants of Concern Highlight Distinctive Amino Acid Substitution Patterns

Filips Peisahovics, Mohammed A. Rohaim, Muhammad Munir*

Division of Biomedical and Life Sciences, Lancaster University, Lancaster, Lancashire, LA1 4YG, United Kingdom

*Corresponding Author

**Prof Muhammad Munir**

Lancaster University, Lancaster, United Kingdom.

Email: muhammad.munir@lancaster.ac.uk

## Abstract

Since the onset of pandemic in 2019, SARS-CoV-2 has diverged into numerous variants driven by antigenic and infectivity-oriented selection. Some variants have accumulated fitness-enhancing mutations, evaded immunity and spread despite global vaccination campaigns. The spike (S) glycoprotein of SARS-CoV-2 demonstrated the greatest immunogenicity and amino acid substitution diversity owing to its importance in the interaction with human angiotensin receptor 2 (hACE2). The S protein consistently emerges as an amino acid substitution (AAS) hotspot in all six lineages, however, in Omicron this enrichment is significantly higher. This study attempts to design and validate a method of mapping S-protein substitution profile across variants to identify the conserved and AAS regions. A substitution matrix was created based on publicly available databases, and the substitution localization was illustrated on a cryo-electron microscopy generated S-protein model. Our analyses indicated that the diversity of N-terminal (NTD) and receptor-binding (RBD) domains exceeded that of any other regions but still contained extended low substitution density regions particularly considering significantly broader substitution profiles of Omicron BA.2 and BA.4/5. Finally, the substitution matrix was compared to a random sample alignment of variant sequences, revealing discrepancies. Therefore, it was suggested to improve matrix accuracy by processing a large number of S-protein sequences using an automated algorithm. Several critical immunogenic and receptor-interacting residues were identified in the conserved regions within NTD and RBD. In conclusion, the structural and topological analysis of S proteins of SARS-CoV-2 variants highlight distinctive amino acid substitution patterns which may be foundational in predicting future variants.

# 1. Introduction

COVID-19, a disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), first appeared in the Hubei province of China in 2019. On 11th March 2020, the World Health Organization (WHO) declared the disease a pandemic (WHO, 2020). By 16th June 2022, the confirmed cases count surpassed 535 million, including 6.3 million deaths (WHO, 2022a). Furthermore, the COVID-19 pandemic inflicted long-term damage to several aspects of international societies, including hardly quantifiable psychological impact. Yeyati and Filippini, (2021) have estimated that global GDP would be 54.68% lower in 2020-2030 compared to pre-pandemic trends as a consequence of educational loss, deaths and economic shrinking. Governments relied on vaccination as the major countermeasure, and by June 16th, 2022, more than 65% of the world population (>5 billion people) had received at least one dose of COVID-19 vaccine. The most commonly used vaccines around the world, include Oxford-AstraZeneca, Pfizer, Moderna, Sinopharm, J&J, Sputkin-V, and Sinovac, which were designed to target the spike (S) glycoprotein, which is the most immunogenic protein of the virus (Das and Roy, 2021; Holder, 2022).

Throughout the pandemic, SARS-CoV-2 accumulated subsets of mutations driven by antigenic drift and or selection favoring the virus infectivity and spread (Altmann et al., 2021). WHO established Greek-letter nomenclature for variants and classified them into three groups: variants under monitoring, variants of interest (VOI) and, most importantly, variants of concern (VoC), which demonstrated increased transmissibility and pathogenicity or decreased countermeasures efficiency (WHO, 2022b). The emergence of novel variants raised concerns about vaccines efficacy, which were later confirmed by a number of variants demonstrating different degrees of immune evasion (Altmann et al., 2021; Jangra et al., 2021; Munir, et al., 2021; Singh et al., 2021; Wang et al., 2021).

European Centre for Disease Prevention and Control (2022) was monitoring SARS-CoV-2 variants in Europe, assessing the impact of variants' substitution portfolio on severity and transmission in comparison to the previously circulating variants. Each dominant variant demonstrated high transmissibility combined with immune evasion and increased severity, excluding Omicron SARS-CoV-2. The variant domination patterns could be judged from the representation of sequence submission dynamics to the GISAID database. Alpha's transmission was surpassed by Delta variant, which was itself eventually surpassed by Omicron variants (Figure 1).

SARS-CoV-2 is a Betacoronavirus of 65-125 nm in diameter, has positive single-stranded RNA of 30-kilo base pairs genome size, encoding four structural and 15 accessory proteins (Jungreis et al., 2021; Astuti, 2020). S protein is a structural, transmembrane glycoprotein, accommodating a homotrimer structure, each monomer-1273 amino acids (141.2 kDa), its binding to the human angiotensin-converting enzyme-2 receptor (hACE2) leads to viral internalisation (UniProt, 2022). The receptor-binding domain (RBD) of the S protein adapted two conformations: UP – receptor accessible, and DOWN – receptor inaccessible. The DOWN conformation decreased hACE2 recognition potential, compensated by the high affinity of RBD, and also complicated antibody access (Cai et al., 2020; Shang et al., 2020). Multiple studies have reported the most immunogenic and key receptor-binding residues (often overlapping) in the S protein RBD, including 417, 452, 477, 484, 490, 493, 496, 498, 501 and 505, which were present in several VoCs of SARS-CoV-2 (Mercurio et al., 2021; Pavlova et al., 2021; Sharma et al., 2021; Watanabe et al., 2021; Yi et al., 2021; Yang et al., 2020). The S protein demonstrated high diversity and mutation rates (Miao et al., 2021; Forni and Mantovani, 2021; Agarwal et al., 2022) such as Omicron (VOC) carried 32 mutations in S protein, leading to immune evasion in vaccinated and convalescent patients (Planas et al., 2021).

These mutations have severely undermined vaccine efficacies and antiviral therapies. The Imdevimab which targets the linear epitope (440–449 amino acid of S protein), Cilgavimab and Bebtelovimab have capabilities to neutralize newest variants of SARS-COV-2 including BA.2 and BA.4/BA.5 (Cao et al., 22, Ahmed et al., 2022). However, antibodies such as Adintrevimab and Sotrovimab showed markedly reduced neutralization against BA.4/BA.5 subvariants. Similarly, neutralization by the antibodies induced by Wuhan antigen-based vaccines or through natural infections showed weaker protection against Omicron subvariants particularly after four months of recovery or vaccination (Cao et al., 22).

The aim of this study is to identify the regions of the highest immunogenicity and conservation in the S protein of all major VoCs including BA.4 and BA.5 using a range of *in silico* tools and models. It has been hypothesized that the functional constraint on the virus divergence could result in the conservation of regions of the S protein surface to preserve hACE2-interaction ability. Moreover, this study aims to apply genetic analyses methods for mapping the conserved regions across all VoCs and project genetic conservation in a parallel comparison fashion. The provided information is fundamental to predict future evolutionary trajectories and training better vaccine and

therapeutic candidates.

## 2. Methods

### *2.1. Sequences Acquisition*

Sequences pertaining to SARS-CoV-2 variants were downloaded from the GISAID database in FASTA format. A query was made for each variant by inputting the Pango Lineage index according to WHO and Pango Lineage nomenclature, selecting a high-coverage, complete sequence without any unidentified nucleotide regions (Table 1).

### *2.2. Gene Sequence Extraction and Alignment*

The S gene sequences was extracted from full-length sequences using SnapGene Viewer software (www.snapgene.com). Each nucleotide sequence was opened as a single-stranded linear DNA sequence, and then were translated to amino acid sequence. Using default parameters of the SnapGene Viewer, protein-encoding regions (ORF) were identified. According to Yoshimoto (2020), the S gene is located in the region spanning from nucleotide 21,563 to 25,384. Both nucleotide and amino acid sequences were extracted and saved as FASTA format in individual datasets. Next, the translated amino acid sequences and a Wuhan reference S protein amino acid sequence were aligned using MEGAX software by the MUSCLE method (with parameters including Gap open: -2.9; Gap extend: 0; Hydrophobicity multiplier: 1.2; Max iterations: 16; Cluster method: UPGMA; Lambda: 24) (Edgar, 2004) and the alignment was exported in FASTA format.

### *2.3. Analysis of Amino Acid Substitutions in the S protein*

A table of amino acid substitutions (data not presented) was made for each variant according to the CoVariants Online Database (Hodcroft, 2022), excluding Zeta and Theta, that were acquired directly from GISAID from the earliest accession in the Pango lineage-filtered tree (Khare et al., 2021). The substitution recurrence across variants (cumulative substitution count) was calculated per each identified site and summed by the S protein regions according to the GenBank designation (Accession ID: NC_045512.2). The relative frequency was calculated as the ratio of cumulative substitution count in sites of a region against the total number of substitutions in the protein. These results were organized in a substitution matrix format.

The substitutions present in the alignment were compared against the substitution matrix to identify discrepancies. To construct substitution sequence WebLogo representation, a sequence set of varying residues (according to the

substitution matrix) was constructed in FASTA format and imported into the WebLogo online tool (Crooks et al., 2004). Finally, the alignment was uploaded to the Phylogeny.fr online tool ("One Click" Mode) to construct a maximum-likelihood phylogenetic tree (Dereeper et al., 2008).

### *2.4. Spike protein Substitutions 3D*

A cryo-electron microscopy 3D structure model of the Wuhan's S protein with RBD in the UP conformation was downloaded from RCSB Protein Data Bank (accession ID: 6SVB) (Wrapp et al., 2020). Each variant was designated with a random, visually-distinguishable color to help in the comprehension of the models. All the substitution sites were highlighted at one subunit of the Wuhan S protein using UCSF Chimera software. Despite every substitution was being highlighted with a contrasting color, some of the substitutions were not visible from the chosen perspective, therefore, not indicated.

Next, 3D models were generated for each variant, highlighting the substitutions localisation on every subunit. Substitution sites were highlighted according to the substitution matrix; if the substitution residue was not resolved in the model, the two flanking residues were highlighted, unless one of those was above ten residues apart, then only the closer residue was highlighted.

## 3. Results

### *3.1. Substitution Localisation*

Our results revealed that the S protein fusion peptide (798-806), internal fusion peptide (816-833), transmembrane domain (1213-1236) and cytoplasmic domain (1237-1273) regions did not have any reported substitutions in the sample alignment. The majority of the reported substitutions were localized in the N-terminal half-part of the S protein, especially in the N-terminal domain (NTD) and RBD (Figure 2A). The C-terminal half-part contained the minor part of the substitutions that were localised in four function-unidentified intermediate regions (IR), heptad repeats (HR) I and II, and the cleavage site (CS) (Figure 2B). In addition, the level of conservation varied across regions (Figure 2).

Although, NTD has the highest number of substitutions, the consensus sequence remained considerably conserved compared to other regions. Meanwhile, RBD and CS showed the highest variability; particularly, RBD has highly variable sites 417, 452, 484 and 501, while CS has sites 681 and 701. Substitution sites 142, 253, 339, 371, 655

and 1176 also demonstrated outstanding variability, yet to a lower extent compared to the aforementioned (Figure 3).

*3.2. Phylogenetic Analysis*

Maximum likelihood tree for the S protein amino acid sequences as well as phylogenetic tree based on full length sequence of the SARS-COV-2 revealed that Omicron BA.1, BA.2, BA.4 and BA.5 subvariants were significantly deviated from all other variants (Figure 4, A, B). Notably, the variant clustering did not correspond to the chronological outbreak order (WHO, 2020b), while the Wuhan clustering closely with Lambda, Delta, Kappa and Epsilon. Additionally, all variants including Omicron sub-variants formed a successive, single branching pattern; every earlier branch diverged only once, into a single later branch (Figure 4, A and B).

*3.3. Cumulative Substitutions*

The proportion of cumulative substitutions prevailed in NTD and RBD – 33.88% and 35.95% respectively. Moreover, out of total 93 overall substitution sites, RBD and NTD contained 66 substitutions. The RBD showed a greater number of substitutions than NTD (87 versus 82 subs) and more substitutions per residue (0.39 VS 0.28) due to the shorter length of the region (223 versus 292 amino acids). Out of 93 unique substitution sites, most of these substitutions were present in one to few variants where 47 sites were unique to one variant, 34 were present in two to four, and the remaining 12 sites – in five or more variants. Several sites have an outstanding manner of substitutions: site 142 and 144 – in 5 variants, site 417 – in 6, sites 501 and 681 – in 9 variants, site 484 – in 12 variants, and all 16 variants have an Asp-to-Gly substitution at site 614 (Table 2).

*3.4. Substitution Patterns*

3.4.1. Substitutions in NTD

Deletions constitute around 50% of the total amino acid substitutions in the NTD where some of these deletions were observed in several SARS-CoV-2 variants. Variants including Alpha, Eta and Omicron BA.1, BA.4 and BA.5 have H69del and V70del deletions. One specific deletion (Y144del) also appeared in Alpha, Eta, Omicron BA.1 as well as within a deleted region 141-144 in Theta. Similarly, region from 142-145 amino acids was substituted in Omicron BA.1 while Alpha and Theta variants shared a pair of identical deletions; L241del and A243del. Lambda variant

showed the highest number deletions as a distinct, where seven amino acids long sub-region, spanning from site 246 to 252 inclusively, along with substitution D253N, followed by Theta and Omicron BA.1 variants with six deletions, and Omicron BA.4 and BA.5 with five deletions (Table 2). The most common substitution in NTD, apart from deletions, was T95I, appearing in Iota, Mu and Omicron BA.1, while Omicron subvariants BA.2, BA.4 and BA.5 also shared T19I, A27S and G142D mutations. Finally, Lambda and Omicron BA.1 showed the highest overall number of substitutions in NTD (ten substitutions), followed by Omicron BA.4 and BA.5 (nine substitutions). In contrast, Zeta carried no mutations in the NTD while Kappa has only one, and Epsilon and Iota have two substitutions (Table 2).

3.4.2. Substitutions in RBD

Omicron subvariants carried 15-17 substitutions in the RBD, significantly outperforming all other variants, which accumulated only one to three substitutions in the RBD of S-protein. Interestingly, just one substitution (F486V) of total 17 was unique to the most-recent Omicron subvariants BA.4 and BA.5, which shared identical substitution portfolio in S protein. Apart from Omicron-specific substitutions, most of the substitutions in RBD were present in multiple variants, and R364K and F490S substitutions appeared in Mu and Lambda variants.

Residue 484 is the highest diversified among all the reported substitution sites within the S-protein. Alpha, Delta, Epsilon and Lambda variants retained the original Wuhan's Glu at this site, while other variants, except for both Omicrons and Kappa were substituted by Lys – the most common form across variants. Omicrons have Ala at this site, but Kappa has a unique Gln. Six variants that have accumulated N501Y substitution included Alpha, Beta, Gamma, Theta, Mu and Omicron subvariants, making it the second most common substitution in the RBD, followed by L452R that is in six variants. Notably, substitutions L452R and E484(K/A) did not overlap (Table 2).

3.4.3. Substitutions in Other Regions

The remaining regions, including SP, IR 1-4, S1/S2 CS and HR I and II, carried additional unique 27 substitution sites out of a total of 93 substitutions. Disrespecting Omicron-specific substitutions, only six were present in two or more variants, including (1) D614G that as present in every variant, (2) H655Y, (3) P681 (H/R) and (4) A701V in S1/S2 CS, (5) D950N in HR I and (6) V1176F in HR II.

Residue 681 showed the second-highest diversified site after reside 484 in the

RBD. Variants Alpha, Theta, Mu and all Omicron subvariants carried a Pro-to-His substitution at this site, while Delta and Kappa have a Pro-to-Arg one. Omicron subvariants demonstrated the greatest substitution numbers in the remaining regions – 8 in BA.2, BA.4 and BA.5, and 10 in BA.1. Variants such as Alpha and Theta revealed six substitutions in these regions while Gamma contains four, and others - one to three (Table 2).

3.4.4. Conserved Regions

Regions within the S-protein were only partially diversified while NTD spanned from residue 13 to 304. There were numerous subregions of low substitution density, including subregions 27-67 that contain a single cumulative substitution (Q52R), 95-138, 158-211 (R190S) and 215-241. Next, inter-region 265-339 located between NTD and RBD (spanning 319-541) has no substitutions. Furthermore, from site 265 to 371, there were only two substituted residues; G339D and R346K. Going downstream of the RBD, several subregions were of a low substitution density: 376-405; 417-440; and 505-547. The longest substitution-less subregion of the NTD- RBD region was the inter-region 265-339 (74 amino acids long), followed by 95-138 (43), 505-547 (42) and 339-371 (32). Additionally, subregions 158-211 (58) and 27-67 (40) contain only one substitution (Table 2).

3.4.5. Sample Alignment and Matrix Discrepancies

Most of the reported substitutions were present in the sample alignment with four exceptions, and additional substitutions were discovered in six cases. Sequences from Beta, Iota, Kappa and Lambda contain two additional substitutions each, while Mu had three, and Delta – four; also, Delta and Theta sequences were missing two reported substitutions each, Mu missed one, and Omicron BA.1 was missing eight reported substitutions in the NTD. In contrast, sequences of Alpha, Gamma, Epsilon, Zeta, Eta and Omicron BA.2, BA.4 and BA.5 corresponded to the reported substitutions profile perfectly (Table 3).

3.4.6. Substitutions Pattern Visualization

The localization of substitutions on the surface of S-protein subunits is shown in Figure 4. Overall, 23 out of 93 substitution sites were not visible on the surface of the S-protein from three selected perspectives. The top-centre region (Figure 7) of the protein surface contains a dense substitution localization that accommodated most of

the substitution sites in the RBD, except for 405, 408 and 446. However, RBD sub-regions 471-487 and 371-376 were unresolved in the cryo-EM structure, hence, the surface protrusion of the contained six substitution residues could not be visualized precisely.

The pyramidal, outer-protruding structures at the topsides of the protein demonstrated substitution sites in the NTD, missing sites up to site 26 due to unresolved N-terminal gap in the cryo-EM structure as well as substitution site 265. The NTD subregions including 66-80, 140-158, and 245-261 flanked 22 substitution sites, whose positioning in the cryo-EM model was unresolved. The remaining 15 surface-visible sites were distributed across other S-protein region that was downstream in primary sequence, and ten other sites in these regions were not visible on the surface. Sites 677, 679 and 681 were unresolved and, therefore, indicated by a flanking subregion 672-687.

In comparison to substitution clustering in the RBD and NTD, the remaining substitutions were distributed individually. Most of these substitutions were localised on the top surface (Figure 7) were common across variants with a few exceptions: R346K in Mu, T376A in Omicron BA.2, K417T in Gamma, E484Q in Kappa, F490K in Lambda and G496S in Omicron BA.1, BA.4 and BA.5. Remarkably, the highest substitutions density across variants was localized in the centre of the top surface, where the RBD (Figure 7 and 8). Omicron subvariants have the highest substitution density on the top surface. Uniquely, variant-specific mutations on the S-protein sides mainly appeared in the NTD, particularly in unresolved subregions spanning from amino acid 140-157 and 245-261. Overall, the substitution localization density decreased from the RBD on the top surface down to the cytosolic domain (Figure 7 and 9).

## 4. Discussion

Early investigations of SARS-CoV-2 genomes predicted an evolutionary rate of roughly 0.001 subs site$^{-1}$ year$^{-1}$ (two to three mutations per month) (Duchene et al., 2020); however, there is significant divergence from this pace across the phylogeny, with certain outlier lineages, particularly VOC, acquiring multiple mutations at a considerably faster rate. The mutations analysis from virus genome data is important for basic virology (Houldcroft et al. 2017), as it identifies evolutionary signals associated

with mutations prior to experimental and real-world data on clinical outcomes or vaccine effectiveness, and it documents and tracks changes that may affect therapeutic effectiveness. Therefore, it is imperative to assess the tendencies and trends in the topological and structural differences of major variants of concerns to predict future evolutionary trajectories (Tariq et al., 2022).

Currently, about 12 million genome sequences are accessible through the GISAID Initiative, allowing for real-time monitoring of the epidemic (Shu and McCauley 2017; Meredith et al. 2020). Since the cumulative substitution count was based on the number of substitution recurrences in WHO-named variants. Theoretically, it indicates the importance of the S protein in terms of phenotypical diversification, gaining enhanced transmissibility, pathogenicity, evading immunity or adapting for a particular epidemiological niche (Wright et al., 2022). Hence, a high cumulative substitution count of a region would suggest that substitution accumulation in this region contributed to the emergence of SARS-CoV-2 variants with the aforesaid qualities more substantially than substitution accumulation in regions of a low cumulative substitution count.

Both NTD and RBD of the S protein demonstrated the highest cumulative substitution count – 82 (33.88%) and 87 (35.95%), respectively, and contained the majority of all substitution sites – 66 out of 93. Multiple sites augmented the diversity of RBD by varying substitutional outcomes, such as Glu-to-Lys, Glu-to-Gln and Glu-to-Ala substitutions emerged at site 484 in various variant combinations. Besides, deletions prevailed in the NTD, making 43 out of 82 substitutions in total. Finally, the visualization of substitution on cryo-EM models suggested that the substitution density prevailed in the RBD and NTD, the distal-most domains of the S-protein, which were reported to be targeted by antibodies (Dejnirattisai et al., 2021; McCallum et al., 2021; Liu et al., 2020; Ahmed et al., 2022). Thus, RBD and NTD, to a lesser extent, could be considered as the critical S protein region regarding viral adaptation, immune response, and treatment design.

Omicrons shared substitutions that were associated with enhanced hACE2 affinity and immune evasion as in Alpha variant - N501Y and P681H (Khan et al., 2021; Luan et al., 2021). While Omicrons had one shared substitution in RBD of unclear importance with Delta, thought to be linked to viral fitness - T478K (Di Giacomo et al.,

2021; Jhun et al., 2021). Despite substitution E484K being reported to have a role in evading immunity (Wu et al., 2022; Jangra et al., 2021), neither Alpha nor Delta contained it, while it reported in seven other variants. Omicrons have an E484A substitution, which similarly to E484K, removed the carboxyl's negative charge. Perhaps, the removal of negative charge at site 484 resulted in the immune evasion by decreasing antibody recognition ability, but the introduction of Lys positive charge created more epitope potential than the introduction of neutral Ala residue as in Omicrons (Wu et al., 2022; Jangra et al., 2021; Altaf et al., 2022). Interestingly, recently emerged Omicron variants consists of identical RBD sequences compared to BA.2 with the L452R/F486V mutations in BA.4 and BA.5. These mutations provided a transmission advantage and antibodies escape characteristics to BA.4 and BA.5 over BA.2. This aligned with the recent research highlighting those individuals who had recovered from SARS infection displayed a systematic reduction in neutralization activity against BA.4 and BA.5 variants (Cao et al., 2022; Agarwal et al., 2022). Overall, Omicron subvariants' substitution profile was explicitly distinct from others, clearly illustrated by the phylogenetic analyses. The unprecedented number of substitutions in the RBD could explain the enhanced transmissibility and immune evasion ability of Omicron variants (ECDC, 2022; Planas et al., 2021).

Strong dependence on the host's protein machinery set functional constraints on the adaptive capability of RNA viruses, leaving a genetic vulnerability as RNA viruses could keep substituting only a fraction of their genome to evade immunity before being forced out of the niche (Simmonds et al., 2019; Holmes, 2003). This functional constraint might press on the SARS-CoV-2 to conserve particular on critical regions in the S protein throughout variants to be able to bind with hACE2 – these regions were arguably the low substitution density subregions of NTD and RBD.

Several studies have reported immunogenic and receptor-binding key residues in the RBD. Six of them were not substituted in any of the observed variants and situated in the extended, low substitution density subregions of RBD: 403, 418, 421, 426, 439 and 506 (Pavlova et al., 2021; Watanabe et al., 2021; Yi et al., 2021; Yang et al., 2020). Additionally, Mercurio et al. (2021) and Sharma et al. (2021) have reported other critical residues located in shorter low substitution density regions of the RBD. Despite eliciting

strong immune response, these residues were not substituted, which might indicate that these residues and corresponding subregions were under functional constraint, therefore, treatment targeted them would potentially be less variant-biased. Even highly mutated Omicron subvariants contained low substitution density subregions in the RBD, where the six key residues remained unsubstituted.

Our analyses revealed that individual virions attributed to a variant could contain additional substitutions and lack those reported to be variant-specific, so the variant designation might only represent a particular fraction of viruses in the lineage. Besides, previous study estimated the variant doubling period at 71.67 (± 0.06 SE), one novel variant per 600,000 infections, which signified the divergence potential of SARS-CoV-2 and, hence, the necessity for universal treatment (Duarte et al., 2021).

Results of substitution matrix and cumulative substitution count support the importance of NTD and RBD in the antigenic drift of SARS-CoV-2. Additionally, the substitution matrix clearly illustrated the differences between substitution profiles across variants, especially of Omicron subvariants, which could be studied further regarding adaptation patterns, biochemical and epidemiological effects. Overall, this study has demonstrated a method of genetic analysis that would hypothetically aid in revealing sites and regions in the S protein of high immunogenicity and conservation if improved in terms of scope and accuracy.

We offered in this study an algorithm that would compute a substitution matrix with cumulative substitution count per residue based on the large alignment of all SARS-CoV-2 sequences. In addition, the hypervariable residues could be mapped using the cryo-EM model to check the antibody binding sites and corrected for glycan shielding that can cover 40% of the S-protein (Grant et al., 2020). However, an enormous computational capacity would be required to analyze such query, as only for the alignment stage, the best US supercomputer operated a full week to align only 17,000 virus genomes (Garvin et al., 2020), and the sequence number is the primary accuracy-limiting factor. Therefore, to facilitate such calculation, the amino acid sequences can be trimmed down to include NTD and RBD due to their immunogenicity and receptor-interaction importance.

## 5. Conclusions and limitations

Tracking the virus evolution play an important role in providing clear and accessible information to those who are tackling the pandemic, including through public health actions and the development of vaccines and therapeutics. Although amino acid sequence analyses alone are insufficient to determine the functional effect of a single mutation on SARS-CoV-2 fitness, computational analysis of existing SARS-CoV-2 mutations provided substantial information on possible phenotypic changes and projected mutations may confer on variants.

**CRediT authorship contribution statement**

Filips Peisahovics, Mohammed A. Rohaim and Muhammad Munir: Conceptualization, Methodology, Investigation, Writing – original draft, Visualization. Muhammad Munir: Resources, Data curation. Filips Peisahovics and Mohammed A. Rohaim, Writing – review & editing. Muhammad Munir Supervision and Funding Acquisition. All authors read and confirmed the submission of the paper.

**Data Availability**

Data will be made available on request.

**Acknowledgement or Funding**

*Declarations of interest*

None

## References

Agarwal, D, Zafar, I, Ahmad, S.U, et al., 2022. Structural, genomic information and computational analysis of emerging coronavirus (SARS-CoV-2). Bull Natl Res Cent. 46(1): 170.

Ahmad, S.U, Kiani, B.H, Abrar, M. et al., 2022. A comprehensive genomic study, mutation screening, phylogenetic and statistical analysis of SARS-CoV-2 and its variant omicron among different countries. J Infect Public Health. 15(8):878-891.

Altmann, D.M., et al., 2021. Immunity to SARS-CoV-2 variants of concern. *Science*,

371(6534): p. 1103-1104.

Astuti, I., 2020. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS- CoV-2): An overview of viral structure and host response. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(4): 407–412

Cai, Y., et al., 2020. Distinct conformational states of SARS- CoV-2 spike protein. *Science*, 369(6511): 1586-1592

Cao, Y., Yisimayi, A., Jian, F. et al. BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. Nature (2022). https://doi.org/10.1038/s41586-022-04980-y

Crooks, G.E., et al., 2004. WebLogo: a sequence logo generator. *Genome Research*, 14(6): 1188-1190

Altaf M., et al., 2022. Wildlife as a Source of SARS-CoV-2 Evolution- A Review. Pakistan journal of Zoology. 54 (4): 1899-1904

Das, J.K. and Roy, S., 2021. A study on non-synonymous mutational patterns in structural proteins of SARS-CoV-2. *Genome*, 99(999): 1-14

Dejnirattisai, W., et al., 2021. The antigenic anatomy of SARS-CoV-2 receptor binding domain. *Cell*, 184(8): 2183-2200

Dereeper, A., et al., 2008. Phylogeny. fr: robust phylogenetic analysis for the non-specialist. *Nucleic acids research*, 36(2): 465-469

Di Giacomo, S., et al., 2021. Preliminary report on severe acute respiratory syndrome coronavirus 2 (SARS- CoV-2) Spike mutation T478K.*Journal of medical virology*, 93(9): 5638-5643

Duarte, C.M., et al. 2022. Rapid evolution of SARS-CoV-2 challenges human defenses. *Scientific Reports*, 12(1); 6457.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5): 1792-1797.

European Centre for Disease Prevention and Control (ECDC), 2022. SARS-CoV-2 variants of concern as of 24 February 2022. Available at: https://www.ecdc.europa.eu/en/covid-19/variants-concern [Accessed on 1st March 2022].

Forni, G. and Mantovani, A., 2021. COVID-19 vaccines: where we stand and

488 challenges ahead. *Cell Death & Differentiation*, 28(2): 626-639.

489 Garvin, M.R., et al., 2020. A mechanistic model and therapeutic interventions for
490 COVID-19 involving a RAS-mediated bradykinin storm. *elife*, 9: e59177.

491 Grant, O.C., et al., 2020. Analysis of the SARS-CoV-2 spike protein glycan shield
492 reveals implications for immune recognition. *Scientific reports*, 10(1): 1-11.

493 Jangra, S., et al., 2021. The E484K mutation in the SARS-CoV-2 spike protein reduces
494 but does not abolish neutralizing activity of human convalescent and post-vaccination
495 sera. *MedRxiv*. 01.26.21250543. doi: 10.1101/2021.01.26.21250543.

496 Jhun, H., et al., 2021. SARS-CoV-2 Delta (B.1.617. 2) variant: a unique T478K mutation
497 in receptor binding motif (RBM) of spike gene. *Immune Network*, 21(5): 32.

498 Jungreis, I., et al., 2021. Conflicting and ambiguous names of overlapping ORFs in the
499 SARS-CoV-2 genome: A homology-based resolution. *Virology*, 558: 145-151

500 Hodcroft, E.B., 2021. CoVariants: SARS-CoV-2 Mutations and Variants of Interest.
501 Available at: https://covariants.org/ [Accessed on 14th June 2022].

502 Holder, J., 2022. Tracking coronavirus vaccinations around the world. Available at:
503 https://www.nytimes.com/interactive/2021/world/covid-vaccinations-tracker.html
504 [Accessed on 16th June 2022].

505 Holmes, E.C., 2003. Error thresholds and the constraints to RNA virus evolution. *Trends*
506 *in microbiology*, 11(12): 543-546.

507 Khan, A., et al., 2021. Higher infectivity of the SARS-CoV-2 new variants is associated
508 with K417N/T, E484K, and N501Y mutants: an insight from structural data. *Journal of*
509 *cellular physiology*, 236(10): 7045-7057.

510 Khare, S., et al., 2021. GISAID's Role in Pandemic Response. *China CDC Weekly*,
511 3(49): 1049.

512 Liu, L., et al., 2020. Potent neutralizing antibodies against multiple epitopes on SARS-
513 CoV-2 spike. *Nature*, 584(7821): 450-456.

514 Luan, B., et al., 2021. Enhanced binding of the N501Y- mutated SARS-CoV-2 spike
515 protein to the human ACE2 receptor: insights from molecular dynamics simulations.

516 *FEBS letters*, 595(10): 1454-1461.

517 McCallum, M., et al., 2021. N-terminal domain antigenic mapping reveals a site of
518 vulnerability for SARS-CoV-2. *Cell*, 184(9): 2332-2347.

519 Mercurio, I., et al., 2021. Protein structure analysis of the interactions between SARS-
520 CoV-2 spike protein and the human ACE2 receptor: from conformational changes to
521 novel neutralizing antibodies. *Cellular and Molecular Life Sciences*, 78(4): 1501-1522.

522 Miao, M., et al., 2021. Genetic Diversity of SARS-CoV-2 over a One-Year Period of the
523 COVID-19 Pandemic: A Global Perspective. *Biomedicines*, 9(4): 412.

524 Nikolaidis, M.; et al., 2022. Comparative Analysis of SARS-CoV-2 Variants of Concern,
525 Including Omicron, Highlights Their Common and Distinctive Amino Acid Substitution
526 Patterns, Especially at the Spike ORF. Viruses, 14, 707.

527 Pavlova, A., et al., 2021. Machine Learning Reveals the Critical Interactions for SARS-
528 CoV-2 Spike Protein Binding to ACE2. *The journal of physical chemistry letters*, 12(23):
529 5494-5502.

530 Munir M., A. et al., 2021. Facts and Figures On Covid-19 Pandemic Outbreak.
531 Pakistan journal of Zoology, 53 (3): 801-1200

532 Tariq, M. H et al., 2022. In silico Screening of Bioactive Phytochemicals against Spike
533 Protein of COVID-19. Pakistan journal of zoology, 54 (1): 433-438

534 Planas, D., et al., 2021. Considerable escape of SARS-CoV-2 Omicron to antibody
535 neutralization. Available at: https:// www.nature.com/articles/s41586-021-04389-z
536 [Accessed on 10th February 2022].

537 Shang, J., et al., 2020. Cell entry mechanisms of SARS-CoV-2. *Proceedings of the*
538 *National Academy of Sciences*, 117(21): 11727-11734.

539 Sharma, D., et al., 2021. Elucidating important structural features for the binding affinity of
540 spike-SARS-CoV-2 neutralizing antibody complexes. *Proteins: Structure, Function,*
541 *and Bioinformatics*, 90(3): 824- 834.

542 Simmonds, P., et al., 2019. Prisoners of war - host adaptation and its constraints on
543 virus evolution. *Nature Reviews Microbiology*, 17(5): 321-328.

Singh, J., et al., 2021. Evolutionary trajectory of SARS-CoV-2 and emerging variants. *Virology journal*, 18(1): 1-21.

Universal Protein Database, 2022. P0DTC2 · SPIKE_SARS2. Available at: https://covid-19.uniprot.org/uniprotkb/P0DTC2 [Accessed on 10th February 2022].

Wang, L., et al., 2021. Ultrapotent antibodies against diverse and highly transmissible SARS-CoV-2 variants. *Science*, 373(6556): 818-823.

Watanabe, K., et al., 2021. Intermolecular Interaction Analyses on SARS-CoV-2 Spike Protein Receptor Binding Domain and Human Angiotensin-Converting Enzyme 2 Receptor-Blocking Antibody/Peptide Using Fragment Molecular Orbital Calculation. *The journal of physical chemistry letters*, 12(16): 4059-4066.

Wise, M.J., 2020. SARS, SARS-Cov-2 and MERS are the Same Viral Species, Clades within Bat Beta-Coronaviruses, *OSF Preprints*.

World Health Organisation, 2022a. WHO Coronavirus (COVID-19) Dashboard. Available at: https://covid19.who.int/ [Accessed on 16th June 2022].

World Health Organisation, 2022b. Tracking SARS-CoV-2 variants. Available at: https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/ [Accessed on 11th February 2022].

World Health Organisation, 2020. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. Available at: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-1911-march-2020 [Accessed on 9th February 2022]

Wrapp, D., et al., 2020. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, 367(6483): 1260-1263.

Wright, D.W., et al., 2022. Tracking SARS-CoV-2 mutations and variants through the COG-UK-Mutation Explorer. *Virus evolution*, *8*(1), veac023. https://doi.org/10.1093/ve/veac02

Wu, L., et al., 2022. Exploring the immune evasion of SARS-CoV-2 variant harboring E484K by molecular dynamics simulations. *Briefings in bioinformatics*, 23(1): 383.

572 Yang, Y., et al., 2021. Key residues of the receptor binding domain in the spike protein
573 of SARS- CoV-2 mediating the interactions with ACE2: a molecular dynamics study.
574 *Nanoscale*, 13(20): 9364-9370.

575 Yeyati, E.L. and Filippini, F., 2021. Social and economic impact of COVID-19. Available
576 at: https://www.brookings.edu/research/social-and-economic-impact-of-covid-19/
577 [Accessed on 9th February 2022].

578 Yi, C., et al., 2020. Key residues of the receptor binding motif in the spike protein of
579 SARS-CoV-2 that interact with ACE2 and neutralizing antibodies. *Cellular & molecular*
580 *immunology*, 17(6): 621-630.

581 Yoshimoto, F.K., et al., 2020. The proteins of severe acute respiratory syndrome
582 coronavirus-2 (SARS-CoV-2 or n-COV19), the cause of COVID-19. *The protein journal*,
583 39(3): 198-216.

**Figure legends:**

**Figure 1:** SARS-CoV-2 variants distribution in sequences submitted to GISAID database.

**Figure 2:** Localization of cumulative substitutions in regions of SARS-CoV-2 spike protein. (A) Protein regions correspond to GenBank designation (Accession ID: NC_045512.2). Cumulative substitutions were calculated as the sum of recurrence events in all the variants. The relative frequency (presented in brackets) represents the ratio of the cumulative substitution frequency in the sites of a region against the cumulative number of substitutions in the protein. (B) Schematic domain structure of S protein. Different domains including signal peptide (SP), N-terminal domain (NTD) receptor binding domain (RBD), spike protein subunit 1 (S1), spike protein subunit 2 (S2), fusion peptide (FP), heptad repeat 1 domain (HR1), heptad repeat 2 domain (HR2), and transmembrane domain (TM) are shown.

**Figure 3:** SARS-CoV-2 S protein substitution sequence WebLogo representation. Protein regions are indicated above the residues according to the GenBank designation (Accession ID: NC_045512.2). Abbreviations: CS - Cleavage Site; HR - Heptad Repeat; IR - Intermediate Region; SP - Signaling Peptide. (Adapted from: Crooks et al., 2004).

**Figure 4:** Phylogenetic analysis of SARS-CoV-2 variants based on Spike protein amino-acid sequence (A) and full-length sequences of the SARS-COV-2 (B). Phylogeny.fr online application was used to construct a maximum-likelihood tree. Sample sequences were obtained from the GISAID database, accession IDs presented in the branch names.

**Figure 5:** SARS-CoV-2 variants colour designation. The colour choice was completely random. This figure was designed to aid comprehension of the following figures.

**Figure 6:** Localisation of amino acid substitution sites on the surface of SARS-CoV-2 spike protein. Presented cryo-electron microscopy structure was obtained from Protein Data Bank (Accession ID: 6VSB) and processed in UCSF Chimera software. The receptor-binding domain in this model adapted an UP conformation. Three subunits are colored in shades of red, and the most saturated one is the subunit of interest, on which the substitution localisation was highlighted according to those reported at least in one Greek letter-designated variant.

**Figure 7:** Comparison of amino acid substitution localisation in spike proteins of SARS-CoV-2 variants. Presented cryo-electron microscopy structure was obtained from Protein Data Bank (accession ID: 6VSB) and processed in UCSF Chimera software. The receptor-binding domain in this model adapted an UP conformation. The localisation of reported substitution residues was highlighted on all three subunits (see Section 2.3). Unique, variant-specific substitutions were indicated. Panel A demonstrates top perspective, Panel B - side-front, and Panel C - another side.

**Figure 8:** Comparison of amino acid substitution localisation in spike proteins of SARS-CoV-2 variants. Presented cryo-electron microscopy structure was obtained from Protein Data Bank (accession ID: 6VSB) and processed in UCSF Chimera software. Unique, variant-specific substitutions were indicated in a side view.

**Figure 9:** Comparison of amino acid substitution localisation in spike proteins of SARS-CoV-2 variants. Presented cryo-electron microscopy structure was obtained from Protein Data Bank (accession ID: 6VSB) and processed in UCSF Chimera software. Unique, variant-specific substitutions were indicated in a side view.

630