Analysis of variation in SARS-CoV-2 regarding Spike mutations D614G & N501Y and polymorphism of *in vitro* isolates.

Differentially Conserved Positions (DCPs) are found between two related groups of protein sequences at selective loci that diverge between the two groups that often signify evolutionarily valuable mutations that affect structure and/or function. This study aims to capture the evolution of DCPs of SARS-CoV-2 regarding two historic S protein mutations; 1. (S)D614G – the most widespread mutation found in all variants and 2. (S)N501Y - a marker of both the Alpha/Beta variants; the intention is to discover new concurrent mutations, exclude "assumed" concurrent mutations and define the subsequent structure-function changes for these two flagship mutations using publicly available EM and crystal structures.

Using GISAID data collected up till December 2020, (S)D614G was found to have two mildly conserved DCPs - (S)L18F and (S)A222V. (S)N501Y was separately examined for DCPs which recovered (S)A570D, (S)P681H, (S)T716I, (S)S982A and (S)D1118H. All DCPs were found in S protein and structurally examined from PDB structures in both the "open" (ACE-2 receptive state) and "closed" (non-receptive state). Three S mutations were calculated to specifically destabilise the "closed" form of S (A570D, D614G, and S982A). Of note, inter-protomer salt bridges appear severed for (S)D614 to (S)K853/K854 specifically in the "closed" form, indicating protomer destabilisation.

Additionally, sequencing data from a pair of infected CaCo-2 cultures were analysed for RNA mutations to detect *in vitro* mutations and identify areas of the SARS-CoV-2 genome that are prone to mutagenesis limited at the cellular level, i.e., without the pressures of the humoral immune system or host transmission. As well as silent mutations, cultivation resulted in two B.1.1.7-associated features – (ORF8)Q27STOP and (S)A570D, deleterious zones in NSP1 and NSP12 (pos. 508-522nt and 14,408-14,414) and several nonsynonymous mutations within NSP3, NSP6, NSP13, NSP15, S, N, ORF3a and ORF9c.

Author: Charlotte Louisa Brazil

Signed: 

M.Sc. by Research in Biochemistry **Year of Submission: 2021The University of Kent – School of Biosciences**

Total word count: 25,237

Declaration

No part of this thesis has been submitted in support of any other degree or qualification of the University of Kent, or any other University or Institution of learning.

Acknowledgments

I would like to thank both of my supervisors Dr. Mark Wass and Professor Martin Michaelis of the School of Biosciences for their guidance and encouragement throughout this project, alongside them I would like to thank their Ph.D. students Jake McGreig and Magda Antczak for their patience and kind assistance in helping the project run smoothly in the strangest of academic years. I also extend thanks to Dr. Gary Thompson and his team of helpers of the Software Carpentry Group for the online lessons that introduced me to coding in Linux shell scripting, Python, and GitHub.

Thank you to my fellow MSc students - Andrew, Charlotte, and Naoki for additional support in coming to grips with the project; it was a pleasure to work and grow alongside you all this year.

I am eternally grateful for my family and Isaiah, with special mention to Nikki and Kevin who were always on hand to answer my coding-related questions.

Contents

Declaration **2**

Acknowledgments **3**

List of Figures **6**

List of Tables **8**

Abbreviations **9**

Abstract **11**

1.0 Introduction **12**

1.1 SARS-CoV-2 – a short history *12*

1.2 Symptoms of COVID-19 and co-morbidities *12*

1.3 Lethality of SARS-CoV-2 ..... *13*

1.4 Related coronaviruses *13*

1.5 The RNA genome of SARS-CoV-2 ..... *14*

1.6 Structural proteins: Spike, Envelope, Matrix and Nucleoprotein *15*

1.7 (S)D614G: a pivotal spike mutation *16*

1.8 (S)D614G - "one-up" conformation preference? *17*

1.9 SARS-CoV-2: Ever growing clades *18*

1.10 Alpha variant (B.1.1.7 lineage) – The "UK variant" *19*

1.11 (S)N501Y Spike RBD mutation *20*

2.0 Aims **21**

3.0 Methods **23**

3.1 Preparation of GISAID data with shell scripting *23*

3.2 Finding and Analysing DCPs *25*

3.3 Generation of structural models and stability predictions *25*

3.4 Discovery of in vitro SARS-CoV-2 mutations from sequencing data *26*

4.0 Results **27**

4.1 (S)D614G *27*

4.1.1 D614G - finding DCPs *27*

4.1.2 L18F – structural analysis *29*

4.1.3 A222V – structural analysis *36*

4.1.4 D614G – structural analysis *42*

4.2 (S)N501Y *48*

4.2.1 (S)N501Y– finding DCPs *48*

4.2.2 N501Y – structural analysis *50*

4.2.3 A570D – structural analysis *52*

4.2.4 Summary of P681H *57*

4.2.5 T716I – structural analysis *58*

4.2.6 S982A – structural analysis *61*

4.2.7 D1118H – structural analysis *66*

4.2.8 N501Y-associated deletions HV69-70del & Y144del *69*

4.3 Cell culture *70*

4.3.1 t=0 ..... *71*

4.3.2 t=1 ..... *72*

4.3.3 t=2 ..... *79*

5.0 Discussion **87**

5.1 (S)D614G *87*

5.2 (S)N501Y *88*

5.3 Cell culture *91*

5.4 Validation of structural models *93*

5.5 Conclusion *94*

5.6 Significance *96*

6.0 References **98**

7.0 Supplementary Figures **117**

List of Figures

Fig. 1 – SARS-CoV-2 RNA genome (29,903nt).....15
Fig. 2: (A) The S protein "closed" pre-fusion trimer (PDB: 6XR8) (B) The S protein "one-up" pre-fusion trimer (PDB: 6VSB) (C) Chain A of the closed trimer and (D) Chain A of the one-up trimer
.....28
Fig. 3 - Close-up of the Spike L18F DCP from Chain A of the "closed" prefusion trimer PDB:6XR8
.....30
Fig. 4 - Close-up of the Spike L18F DCP from Chain A of the" one-up" post-fusion trimer PDB:6ZXN.....32
Fig. 5 - Close-up of the Spike L18F DCP from Chain B of the "one-up" post-fusion trimer PDB:6ZXN.....33
Fig. 6 - Close-up of the Spike L18F DCP from Chain C of the "one-up" post-fusion trimer PDB:6ZXN.....34
Fig. 7 - Spike "one-up" conformation: interaction of the RBD of the "up" monomer (Chain A) with the NTD of Chain B with nearby disulfide bridges (C131-C166 and C15-C136) PDB: 6ZXN....35
Fig. 8 – Close-up of the Spike A222V DCP from Chain A of the "closed" prefusion trimer (PDB:6XR8).....37
Fig. 9 – Close-up of the Spike A222V DCP from Chain A of the "one-up" prefusion trimer (PDB:6VSB).....38
Fig. 10 - Close-up of the Spike A222V DCP from Chain B of the "one-up" prefusion trimer (PDB:6VSB).....39
Fig. 11 - Close-up of the Spike A222V DCP from Chain C of the "one-up" prefusion trimer (PDB:6VSB).....40
Fig. 12 – Close-up of the Spike D614G DCP from Chain A of the "closed" prefusion trimer (PDB:6XR8)..... **Error! Bookmark not defined.**
Fig. 13 Close-up of the Spike D614G DCP from Chain A of the "one-up" prefusion trimer (PDB:6VSB).....44
Fig. 14 - Close-up of the Spike D614G DCP from Chain B of the "one-up" prefusion trimer (PDB:6VSB).....45
Fig. 15 - Close-up of the Spike D614G DCP from Chain C of the "one-up" prefusion trimer (PDB:6VSB).....46
Fig. 16 - (A) The S protein "closed" pre-fusion trimer (PDB: 6XR8) and **(B) The S protein "one-up" pre-fusion trimer (PDB: 6VSB) (C) Chain A of the "closed" trimer and (D) Chain A of the "one-up" trimer with DCPs**.....49
Fig. 17 – Close-up of the Spike N501Y DCP from Chain A of the "closed" prefusion trimer (PDB:6XR8).....50
Fig. 18 - Close-up of the Spike N501Y DCP from Chain B of the "one-up" post-fusion trimer (PDB:6M17).....51
Fig. 19 - Close-up of the Spike A570D DCP from Chain A of the "closed" prefusion trimer (PDB:6XR8).....53

Fig.20 Close-up of the Spike A570D DCP from Chain A of the "one-up" prefusion trimer (PDB:6VSB). 54
Fig.21 - Close-up of the Spike A570D DCP from Chain B of the "one-up" prefusion trimer (PDB:6VSB). 55
Fig. 22 - Close-up of the Spike A570D DCP from Chain C of the "one-up" prefusion trimer (PDB:6VSB). 56
Fig. 23 - Close-up of the Spike T716I DCP from Chain A of the "closed" prefusion trimer (PDB:6XR8) 59
Fig. 24 Close-up of the Spike T716I DCP from Chain A of the "one-up" prefusion trimer (PDB:6VSB) 60
Fig. 25 - Close-up of the Spike S982A DCP from Chain A of the "closed" prefusion trimer (PDB:6XR8). 62
Fig. 26 - Close-up of the Spike S982A DCP from Chain A of the "one-up" prefusion trimer (PDB:6VSB). 63
Fig. 27 - Close-up of the Spike S982A DCP from Chain B of the "one-up" prefusion trimer (PDB:6VSB). 64
Fig. 28 - Close-up of the Spike S982A DCP from Chain C of the "one-up" prefusion trimer (PDB:6VSB) 65
Fig. 29 - Close-up of the Spike D1118H DCP from Chains A-C of the "closed" prefusion trimer (PDB:6XR8). 67
Fig. 30 - Close-up of the Spike D1118H DCP from Chains A-C of the "one-up" prefusion trimer (PDB:6VSB) 68
Fig. 31 Close-up of the Spike T76I mutation derived from CaCo-2 cultivation at t=1 from Chain C of the "one-up" Ty1-bound prefusion trimer (PDB:6ZXN) 74
Fig. 32 – Electrostatic map of N protein NTD monomer with (N)N126Y mutation derived from CaCo-2 cultivation at t=1 (PDB:6VYO - Chain A only) 75
Fig. 33 - Close-up of the Spike F157S mutation derived from CaCo-2 cultivation at t=1 from Chain A of the "closed" prefusion trimer (PDB:6XR8) 77
Fig. 34 - Close-up of the Spike F157S mutation derived from CaCo-2 cultivation at t=1 from Chain A of the "one-up" prefusion trimer (PDB:6XR8) 78
Fig. 35 – Crystal Structure of NSP1 protein showing GHVMV82-86del site derived from CaCo-2 cultivation at t=2 (PDB: 7K7P) 81
Fig. 36 – EM Structure of RdRP:RNA showing (NSP12)P323-G327 site affected by pos.14,408-14,414 deletion derived from CaCo-2 cultivation at t=2 (PDB: 6XQB) 82
Fig. 37 – EM structure of mini RTC (RdRP:RNA:NSP13) showing (NSP13)T481M derived from CaCo-2 cultivation at t=2 (PDB: 7CXM) 84
(S2) Fig. 38 - Spike "one-up" conformation: interaction of the RBD of the "up" monomer (Chain A) with the NTD of Chain B with nearby disulfide bridges (C131-C166 and C136) PDB: 6VSB). 118
(S3) Fig. 39 – S protein alignment between 6VSB (blue) and 6ZXN (red) from the side and top view. 118
(S4) Fig. 40 – Close-up of the Spike A222V DCP from Chain C of the "one-up" prefusion trimer (PDB:6VSB) with the lowest strain rotamer of Y38 selected. 119
(S5) Figure 41 – Compound bar chart of post-mutation frequencies across time for FFM3 120

(S6) Figure 42– Compound bar chart of post-mutation frequencies across time for FFM7.Error! Bookmark not defined......121

List of Tables

Table 1 – Detailed list of reported non-synonymous B.1.1.7 mutations.....19
Table 2 – Overview of criteria of sequence selection and method of comparison.....22
Table 3 – Stages of Linux one-line commands used to separate D614-associated sequences in the case for Spike and as an example - Envelope protein.....24
Table 4 - Identification of nt. base changes in FFM3 & FFM7 before cultivation (t=0) relative to the Wuhan reference sequence. 72
Table 5 - ***Identification of nt. base changes in FFM3 & FFM7 midway cultivation (t=1) relative to before cultivation (t=0)***.....79
Table 6 - Identification of nt. base changes in FFM3 & FFM7 at end of cultivation (t=2) relative to midway cultivation (t=1).....86
(S1) Table 7 – Dynamut2 ΔΔG values (kcal/mol) for discussed SARS-CoV protein mutations.116

Abbreviations

2019-nCoV – 2019 novel coronavirus

ACE2 – Angiotensin-Converting Enzyme 2

CoV - Coronaviruses

COVID-19 – Coronavirus disease (2019)

CTD – C-terminal domain

DCP – Differentially Conserved Position

E – Envelope protein

EM – Electron Microscopy

GISAID – Global Initiative on Sharing All Influenza Data

HCoV – Human Coronaviruses

M – Matrix/Membrane protein

mAb – monoclonal antibodies

MR – Morbidity rate

MERS – Middle Eastern Respiratory Syndrome

N – Nucleoprotein

NSP – Non-Structural Protein

NTD – N-terminal domain

ORF – Open reading frame

PCR – Polymerase Chain Reaction

Pp1a – Polyprotein 1a

Pp1ab – Polyprotein 1ab

Pp1b – Polyprotein 1b

RBD – Receptor Binding Domain

RDM – Receptor Binding Motif

RdRP - RNA-dependent RNA polymerase

RMSD – Root Mean Square Deviation

RTC – Replication Transcription Complex

RT-PCR – Reverse Transcription PCR

S – Spike protein

S1/S2 – S cleavage protein products 1 and 2

SARS-CoV - Severe Acute Respiratory Syndrome Coronavirus

SARS-CoV-2 - Severe Acute Respiratory Syndrome Coronavirus 2

VAT – Virus Analysis Tool

VOC – Variant Of Concern

UTR – Untranslated Region

Abstract

Differentially Conserved Positions (DCPs) are found between two related groups of protein sequences at selective loci that diverge between the two groups that often signify evolutionarily valuable mutations that affect structure and/or function. This study aims to capture the evolution of DCPs of SARS-CoV-2 regarding two historic S protein mutations; 1. (S)D614G – the most widespread mutation found in all variants and 2. (S)N501Y - a marker of both the Alpha/Beta variants; the intention is to discover new concurrent mutations, exclude "assumed" concurrent mutations and define the subsequent structure-function changes for these two flagship mutations using publicly available EM and crystal structures.

Using GISAID data collected up till December 2020, (S)D614G was found to have two mildly conserved DCPs - (S)L18F and (S)A222V. (S)N501Y was separately examined for DCPs which recovered (S)A570D, (S)P681H, (S)T716I, (S)S982A and (S)D1118H. All DCPs were found in S protein and structurally examined from PDB structures in both the "open" (ACE-2 receptive state) and "closed" (non-receptive state). Three S mutations were calculated to specifically destabilise the "closed" form of S (A570D, D614G, and S982A). Of note, inter-protomer salt bridges appear severed for (S)D614 to (S)K853/K854 specifically in the "closed" form, indicating protomer destabilisation.

Additionally, sequencing data from a pair of infected CaCo-2 cultures were analysed for RNA mutations to detect *in vitro* mutations and identify areas of the SARS-CoV-2 genome that are prone to mutagenesis limited at the cellular level, i.e., without the pressures of the humoral immune system or host transmission. As well as silent mutations, cultivation resulted in two B.1.1.7-associated features – (ORF8)Q27STOP and (S)A570D, deleterious zones in NSP1 and NSP12 (pos. 508-522nt and 14,408-14,414) and several nonsynonymous mutations within NSP3, NSP6, NSP13, NSP15, S, N, ORF3a and ORF9c.

1.0 Introduction

1.1 SARS-CoV-2 – a short history

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is a previously unknown enveloped positive-sense ssRNA viral species (Order: *Nidovirales*, Family: *Coronaviridae*, Subfamily: *Orthocoronavirinae*, Genus: *Betacoronavirus*, Subgenus: *Sarbecovirus)*.

Until 11th February 2020, the disease was termed "2019 novel coronavirus" (2019-nCoV) by Chinese scientists studying the initial outbreak that was thought to originate from the Wuhan market in December 2019 [1], [2]; initial sequencing of the virus, identified a novel species of coronavirus (now referred to as SARS-CoV-2), which was most closely related to bat coronavirus strain RaTG13 (96% identical) and shared 79.6% sequence identity with SARS-CoV which caused the 2002-2003 SARS epidemic [1]. The SARS-CoV-2 outbreak was classified as a pandemic on 11th March 2020 and has since had an unbelievable impact on global research, health and economic sectors. SARS-CoV-2 has affected all continents, with 93 countries experiencing initial lockdowns: and several regions, mainly European - underwent their third lockdown in early 2021 (e.g. UK/Ireland, France, Israel, Austria...). SARS-CoV-2 continues to have millions of monthly cases in September 2021 and has amassed over 4.7 million deaths [3]

1.2 Symptoms of COVID-19 and co-morbidities

SARS-CoV-2 causes the coronavirus disease (COVID-19) which is characterised by three pathogenic phases: firstly a short-lived asymptomatic infectious stage (whereby respiratory droplets are the primary route of infection), followed by propagation in the upper respiratory tract and subsequent activation of the host innate immune response and finally pneumonia and alveolar tissue damage with possible respiratory

failure [4]; only 20% of patients experience more than moderate symptoms with the elderly categorised as high-risk [5]. Chronic conditions are correlated to poor prognosis; including but not limited to - COPD (4x risk of severe disease [6]), diabetes (5.1% increased morbidity [7]), and hypertension (10.3% higher disease severity, 22.1% higher admission to ICU/mechanical ventilation/death in a singular study [8]).

1.3 Lethality of SARS-CoV-2

The true MR of SARS-CoV-2 is still being debated, initially reported to be higher (January 2020 rates - 11% [9], 15% [10]) before dropping to a lower rate (February - 3.4% [11]; March – 4.3% [12], 0.3% [13]; April – 4.5 - 5.8% [14]). In 2021, global MR was calculated to be 2.14% (90.4 million infections/1.94 million deaths - accessed 11th Jan 2021) approximately thrice the severity of the yearly influenza MR (~600,000 deaths per annum) [15], [16]; in June 2021 the MR of SARS-CoV-2 rose marginally to 2.17% before dropping to 1.54% in September after some 6 billion recorded vaccinations [3].

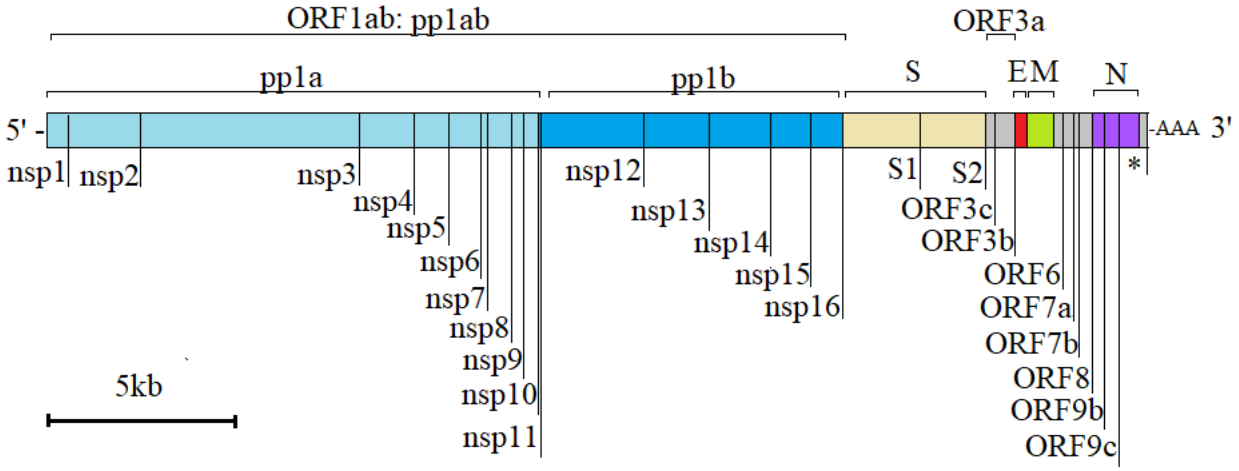
1.4 Related coronaviruses

SARS-CoV-2 is the seventh known coronavirus (CoV) to spread between humans, previous CoV species include: HCoV-229E, HCoV-OC43, HCoV-NL63, CoV-HKU1, SARS-CoV and MERS-CoV (Middle Eastern Respiratory Syndrome) [17]–[19]. The human coronavirus (HCoV) species generally cause mild common cold symptoms [20], whereas the latter two β-CoVs were responsible for epidemic-level outbreaks and compared to SARS-CoV-2 were deadlier with respective morbidity rates (MR) of ~10% and ~35% respectively [21], [22] (MERS was initially believed to be 65% in comorbid patients [23], [24]).

1.5 The RNA genome of SARS-CoV-2

(Ref. Fig. 1) Two-thirds of the ~30kb genomic structure of SARS-Cov-2 comprises of open reading frames ORF1a and ORF1b, which are translated into the corresponding ORF polypeptide1a (pp1a) and ORF polypeptide1b (pp1b). Continuous translation of ORF1ab, resulting in the replicase pp1ab is achieved via the ribosome bypassing a STOP codon by a -1 frameshift occurring just after the non-structural protein 10 (NSP10) exon [25]. ORF1ab can be cleaved into 16 NSPs - a portion of which assemble into the replicating-complex, RNA-dependent RNA polymerase (RdRP) - NSP12 plus cofactors NSP7/NSP8 [26] which combine with additional NSPs (e.g. NSP13) to form the Replication Transcription Complex (RTC). NSP3 and NSP4 can function as proteases – Papain-like protease (PLpro) and 3-chymotrypsin like protease (3CLpro) [27], [28] which cleave the NSPs contained within pp1ab. NSP13 acts as a helicase [29], NSP14 is a 3'-5' exoribonuclease (ExoN) used in proofreading and RNA recombination [30].

The remaining 10kb RNA genome encodes the structural and accessory proteins which are created by the discontinuous synthesis of antisense strands in the 5' direction by an RdRP 'skipping' mechanism permitted by interspersed transcription regulatory sequences; the antisense strands are then utilised by the RdRP to create subgenomic RNA sense strands for the host ribosomes to generate the structural proteins - Spike protein (S), Envelope protein (E), Matrix/Membrane protein (M), Nucleoprotein (N), and accessory proteins ORF3a, ORF3b, ORF6, ORF7a, ORF7b, ORF8, ORF9b, ORF9c and ORF10 [25], [31]–[33]. Current evidence of ORF9c (AKA uncharacterised protein 14/ORF14) and 10 expression is poor.



* ORF10 - unconfirmed to be expressed

Fig. 1 – SARS-CoV-2 RNA genome (29,903nt) *translation divisions are shown to scale. Showing non-structural proteins (NSP1-16), structural proteins (SEMN) with S protein S1/S2 cleavage site, accessory proteins are shaded in grey (ORF3a-ORF9c), untranslated regions border each end of the genome (5' end: 1-798nt and 3' end: 29,675-29,903nt including the poly-A tail).*

1.6 Structural proteins: Spike, Envelope, Matrix and Nucleoprotein

The S protein has a predicted mass of 141.2kDa and appears as an integral glycosylated homotrimer [34], S is of key research interest as it binds to the host angiotensin-converting enzyme 2 (ACE2) which allows host cell entry [35],[36] - a shared trait with SARS-CoV [25]. ACE2-bound SARS-CoV-2 is further activated by proteases furin at the S1/S2 site and by TMPRSS2 or endosomal cathepsins B/L at the S2' site [34]; furin cleavage is not typically an adaptation of SARS-CoV [37], [38].

S protein trimers can be divided into several areas of interest. S1 contains the N-terminal domain (NTD) and Receptor-Binding Domain (RBD) which holds the Receptor-Binding Motif (RBM) for the human ACE2 receptor; in 3D structures the NTD interlocks with the RBD of the neighbouring protomer (unless the RBD is in the "up" position to receive ACE2). S2 is described as the 'stalk' and contains the S2' cleavage site, Fusion Peptides 1 & 2 (FP1, FP2) which mediate Ca^{2+}-dependent membrane fusion of S protein [39]) and Heptad Repeats 1 & 2 (HR1. HR2) that are coiled-coil regions that re-arrange into a 6-

helical bundle during cell-fusion to stabilise merging of host and viral membranes [40]–[42]). Each S protomer resembles a contorted “Y” shape that stacks laterally - whereby S1’s NTD and RBD form two prongs and the S2’ stalk is linear with the distal end anchored in the plasma membrane. The S protein adopts several conformations according to the stage of cell-fusion: (1) pre-fusion - whereby RBDs can be “down” or have one or more in the ACE2 receptive “up” position, (2) fusion intermediate (“pre-hairpin”) and (3) fusion (“hairpin”) state – whereby after binding ACE2 S1 is shed and S2 is re-folded to a linear three-stranded coiled-coil stabilised by HR1/HR2, then FP1/FP2 is inserted into the host membrane [43].

The E protein is a small (75 amino acids long, 8.4kDa) single-pass type III membrane protein [44] that functions as a homopentameric cation channel that interacts with the larger M protein (25.2kDa) to aid with S membrane localisation and maturation in the ER-Golgi secretory pathway [45], [46]. The N protein (45.6kDa) has RNA-binding capabilities to package genomic RNA into ribonucleoprotein complexes for viral particles and is also recruited to RNA transcription complexes to aid with the synthesis of genomic and subgenomic [47]; in addition to this N has been demonstrated to interfere with innate immune response and interfere with cellular kinases to re-organise the actin cytoskeleton and trigger apoptosis [48], [49].

1.7 (S)D614G: a pivotal spike mutation

(S)D614G mutation is caused by a nucleotide base change of A=>G at pos. 23,403nt and G614 give the “G clade” its namesake and marks the start of the B.1 lineage. (S)D614G was first seen in singular cases in Wuhan and Thailand and has become globally widespread despite travel restrictions, (S)D614G is associated with higher viral loads in pseudotyped virions and human patients (correlated from lower PCR cycle numbers) [14], [50]–[52]. (S)D614G was noticed to have three co-occurring mutations where three additional C=>T mutations made up most cases: i.e., pos. 241nt in the 5’-UTR, pos. 3,037nt in NSP3 and

pos. 14,408nt in NSP12 (resulting in P323L). In January, Germany and China recorded these G clade-associated mutations excluding the RdRP mutation, the first recorded sequence with all four C=>T mutations at pos. 241, 3,037 and 14,408nt mutations alongside (S)D614G was identified in Italy on February 20th [50].

Studies using early genomes (WA1 strain) with the D614G mutation without the previously mentioned co-occurring mutations, resulted in greater viral replication *in vitro* in airway epithelial cells, and *in vivo* enhancement of viral replication in the nasal and upper respiratory tract in hamsters [53], [54]; in addition, transmission studies showed that airborne infection occurred more quickly for G614-infected hamsters [54]. D614/G614 co-infection studies undertaken by the same two research groups indicated that the G614 variant was shown to dominate over D614 – *in vivo* and *"ex-vivo"*; e.g. G614:D614 co-infection (1:1) resulted in a 4 fold greater viral titre for G614-infected hamsters a week post-infection [53]. Also, in large airway epithelial cells, even a x10 greater introduction of D614 still resulted in out-competition by the G614 variant [54]. This indicates that D614G alone, without the other 3 co-occurring mutations, has implications of a more contagious phenotype.

There is not yet significant evidence that disease severity has changed due to the mutation [14], [50] - one report found a positive correlation between case fatality rate and G614 in a restricted data window between March 30th – 4th June [55]. The risk of reduced vaccine efficacy due to D614G appears small as neutralisation assays show little difference or slightly higher sensitivity for G614 than D614 [51]–[53], [56].

1.8 (S)D614G - "one-up" conformation preference?

(S)D614G was not correlated with an increased binding affinity to ACE2 [52], but the increase in infectivity may be due to preference for a "one-up" conformation (whereby one of the trimeric S proteins

is outstretched/open due to increased protomer symmetry and possible loss of a hydrogen bond between inter-protomer residues D614-T859 [50], causing a more energy-favourable form for “one-up” S state which leads to greater exposure of the receptor-binding domain (RBD) [57].

The “one-up” state is the main conformation for G614 promoters; e.g. in two studies: 82% and 58% of protomers were in open conformation at any given time point compared to the D614 form at 42% and 18%, respectively [52], [56]. This could offer reasoning for G614’s increased responsiveness to neutralisation by RBD-directed antibodies (Abs) - a further complication is that cryo-EM also revealed elevated proportions (59%) of “two-up” and “all-up” S trimer conformations [52].

1.9 SARS-CoV-2: Ever growing clades

According to GISAID (Global Initiative on Sharing All Influenza Data [58]), the G clade is just one branch of uncovered groups of SARS-CoV-2, the rest are named the “L”, “S” and “V” clades (with an additional unclassed category “O” for “other”). The L clade is the original Wuhan reference sequence, the mutations that give S and V their respective naming are: NSP8-L84***S*** and NSP3-G251***V***. Since August 2020, the L, S, V and O categories have almost statistically vanished and the global pattern has shifted entirely to the G clade and its derivatives: GH, GR and GV (NSP3-Q57***H***, N-G204***R*** and S-A222***V***, respectively - which is not an exhaustive list of all their logged co-mutations).

More recent WHO nomenclature designates major variants of concern (VOCs) using the greek alphabet: (1) Alpha variant – also known as “B.1.1.7 lineage” with Pango nomenclature and as “GRY” clade with GISAID, (2) Beta variant (B.1.3351, GH/501Y.V2), (3) Gamma (P.1, GR/501Y.V3) and (4) Delta variant (B.1.617.2, G/478K.V1) [59].

1.10 Alpha variant (B.1.1.7 lineage) – The "UK variant"

The G clade and its tributaries are ever-growing in complexity, since the start of 2021, new G clade categories have been added (i.e. GRY and +RBDx). The highly divergent Alpha variant (B.1.1.7 lineage or GRY clade) from South-East UK was first detected with GISAID entries EPI_ISL_601443 and EPI_ISL_581117, first emerging in September 2020 and classified as a new VOC (i.e. VOC-202012/01) in December [60], [61]. Alpha variant/B.1.1.7 lineage carried 17 non-silent mutations/deletions - 8 of which in S protein which raised concerns of increased contagiousness, lethality and antibody resistance (ref. Table 1) [62]–[66]. In May 2021, the Alpha variant lineage forms over 2/3 of GISAID sequence submissions. Concerning GISAID clade naming, the Alpha variant was previously known as GR/501Y.V1, before being renamed the new clade category "GRY" with "+RBDx" being used as a suffix for all G clades in cases of clade-specific RBD/Antibody binding region mutations.

Genome Location	Δ nucleotide	Δ amino acid	Protein description
ORF1ab	C3267T	T1001I	NSP3
	C5388A	A1708D	NSP3
	T6954C	I2230T	NSP3
	11288-11296del	SGF3675-3677del	NSP6
S protein	21765-21770del	HV69-70del	NTD of S1
	21991-21993del	Y144del	NTD of S1
	A23063T	N501Y	RBM in CTD of S1
	C23271A	A570D	CTD of S1
	C23604A	P681H	CTD of S1, adjacent to S2 furin cleavage site
	C23709T	T716I	S2
	T24506G	S982A	S2, S2'
	G24914C	D1118H	S2, S2'
ORF8	C27972T	Q27stop	
	G28048T	R52I	
	A28111G	Y73C	
Nucleoprotein	28280 GAT=>CTA	D3L	
	C28977T	S235F	

Table 1 – Detailed list of reported non-synonymous B.1.1.7 mutations *recreated from Table 1 from Rambaut et el.* [60]

1.11 (S)N501Y Spike RBD mutation

The key mutation of the Alpha variant (B.1.1.7 lineage), as well as the South African B.1.351 lineage, is N501Y, occurring in the receptor-binding motif (RBM) of S which has been reported to increase binding affinity to ACE2 [67]–[69]. The first isolated occurrence of N501Y occurred in April in GISAID's database in Brazil (EPI_ISL_500467). With the B.1.1.7 lineage, N501Y is associated with the deletion of two residues in the N-terminal domain (NTD) HV69-70del - this deletion was first seen in March separately to the B.1.1.7 lineage in Italy (EPI_ISL_542148). HV69-70del was significant in evasion of detection by RT-PCR testing, is predicted to alter the immune response and is linked to RBD mutations such as N501Y and N439K [70], [71].

2.0 Aims

This study has two major projects, the first using GISAID data to define conserved positions in two pairs of sequence populations i.e. (S)D614 vs. (S)G614, and (S)N501 vs. (S)Y501 – see Table 2 for a visual explanation. The second element of this study is a time-based analysis of RNA base changes in isolated SARS-CoV-2 cultures.

Analysis of differentially conserved positions (DCPs) is the process of examining the difference in conserved positions between the sequences of genetic counterparts, this gives the ability to define areas that can be related to structure and/or function and explain differences in the conduct of viruses – for example, DCP analysis has previously identified sequence-derived drivers of pathogenicity in ebolaviruses and behaviour of infection in SARS-CoV vs SARS-CoV-2 [72], [73]. This investigation aims to study the DCPs between non-G clade to G clade (D614 vs. G614) which should highlight what amino acid positions are specifically associated with G614 and possibly allude to what base changes have led (S)D614G has become more transmissible. Similarly, investigation of DCPs from sequences groups split by the presence/absence of the prominent RBD mutation (S)N501Y which present in Alpha, Beta and Gamma variants, could reveal why variants containing (S)Y501 transmission is enhanced and what are the true co-occurring mutations of (S)N501Y.

The study of mutations in the case of SARS-CoV-2 has far-reaching applications in areas such as vaccines and drug development. Bioinformatics-based analysis of DCPs of D614 vs. G614 as well as N501 vs. Y501 in SARS-CoV-2 has several benefits: 1) it avoids bias to certain areas or proteins, 2) can re-assess pre-existing mutation-groupings, and 3) systematically define regions in which to scrutinise the effect of these evolutionary-central changes in 3D models. All DCPs recovered related to D614G and N501Y evolution were isolated only within the S protein and 3D structural investigation was performed in both the inactive (“closed”) and pre-fusion (“one-up”) trimers to assess any impacts to ACE-2 binding.

Pairs of sequence-separated groups			**Expectation from comparison of sequence group pairs**
1st group	**vs.**	*2nd group*	
(S)D614 containing sequences		(S)G614 containing sequences	Identify DCPs that define G-clade, the major parent clade of almost all current sequences
(S)N501 containing sequences		(S)Y501 containing sequences	Identify DCPs that define a major Alpha/Beta variant mutation (/B.1.1.7/B.1.351 lineages) commonly referred to as the UK/South African variants.

Table 2 – Overview of criteria of sequence selection and method of comparison *with reasons for pair selection and expected outcome of the analysis.*

Examination of SARS-CoV-2 mutations that arise in the sequencing of lab-cultured generational CaCo-2 cell colonies was performed to discriminate mutations that can arise *in vitro* as opposed to the various selective pressures of acting *in vivo* (e.g. B and T-cell epitopes, heterogeneous cell types) which comprises the GISAID data. This can highlight areas of SARS-CoV-2 that are prone to mutagenesis in the limitation of cellular infection and transmission. This could also reveal reactionary evolution regarding innate immunity (e.g. interferon/cytokine response), cell entry, RNA amplification rates and RNA fidelity,

3.0 Methods

3.1 Preparation of GISAID data with shell scripting

Protein sequences were acquired from the GISAID database (*accessed 3rd Dec 2020*). The S protein IDs were split into two groups: those containing D614 vs. those containing G614. Due to variance in the length of spike sequences, the search criteria for splitting the groups were not dependent on exact numbered positioning and relied upon residue pattern matching of Met1, Val608, Tyr612 Asp/Gly614 and Asn616 – this pattern in UNIX is written `"^M.*V...Y.D.N"` (^M) include sequences with the first occurrence of M (Met1), (.) is a singular character and (.*) is any number of random characters,

Met1 was selected to include the signal sequence and the latter residue pattern does not appear in this order anywhere other than flanking D614/G614. The latter residues had a very low frequency of mutation (between 1-2 mutations each globally) according to the GISAID 'Spike Glycoprotein Mutation Surveillance' resource, so this approach preserves any possible DCPs neighbouring position 614 – this method also had the lowest loss of S ids (228,319 in total compared to a possible 234,086).

The IDs of the spike groups (e.g. EPI_ISL_XXXXXX) were then used to extract and split the remaining protein sequences into the spike D614 and G614-associated categories. This gave in excess of 20,000 and 200,000 sequences in the D614 and G614 categories, respectively, for each SARS-CoV-2 protein. The categorisation of sequences was achieved using the "grep" search command in the Linux emulator 'Git BASH' for Windows 10 - see Table 3.

Extraction of N501Y followed the same method as seen in Table 3, however S protein ID pattern matching followed: Met1, Phe497, Gln498, Thr500 and Asn/Tyr501 (captured total of 244,516 out of 258,936 S ids) and a later GISAID mega file from 12th Dec 2020 was used. Approximately 240,000 and 1,400 sequences were used for N501 vs. Y501 DCP analysis, respectively.

Stage	Linux command inputs	Description
1)	*$ grep -A1 ">Spike|" allprot1203.fasta > spikes.fasta*	Select all Spike sequences from GISAID mega file and write to spikes.fasta
2)	*$* grep -B1 "^M.*V...Y.**D**.N" spikes.fasta > spikes_**D**.fasta	Select D614 matching sequences (with their headers) from spikes.fasta file and write to new spikes_D.fasta
3)	*$ grep ">" spikes_**D**.fasta | cut -d "|" -f4-4 > **D**_ids.txt*	For every header in spikes_D.fasta, select only the "EPI_ISL_XXXXXX" ID portion with the cut command and write to D_ids.txt
4)	*$ grep -A1 -f **D**_ids.txt envelope.fasta* > **D**_envelope.fasta	Using the D_ids.txt, pattern match the IDs to an extracted protein fasta file (made with the command in Stage 1 using ">E|"). Returns all D614-related sequences for envelope protein including headers.

Table 3 – Stages of Linux one-line commands used to separate D614-associated sequences in the case for Spike and as an example - Envelope protein, *instances whereby "D" is written in bold indicates that it may be replaced with "G" to gather G614-associated sequences with the same method. D614G IDs can be pattern-matched on any required protein.fasta extracted from the GISAID mega file as in Stage 1.*

3.2 Finding and Analysing DCPs

DCPs of each SARS-CoV-2 protein was then extracted using the Virus Analysis Tool (VAT) which uses Python2.7 and 3.9 with a minimum conservation threshold set to 70% to find positions that differ significantly in the divided groups. VAT relies on sequence conservation scoring based on Jensen-Shannon divergence with a BLOSUM62 substitution matrix using alignments created by the desktop alignment tool Clustal Omega with HHsearch [74]–[77].

3.3 Generation of structural models and stability predictions

The naming of primary structures follows Uniprot annotations, the viewing of 3D structures and DCPs were achieved with PDB structures (predominantly EM derived) on PyMOL. Possible structural impacts of mutation on Van der Waal's clashing (set to a ratio of 0.89), nearby saccharide chains, as well as changes to polar, covalent, aromatic, and hydrophobic associations, were considered. The PyMOL mutagenesis wizard was used to generate the non-native residues, of which the lowest strain orientations were selected and the APBS plugin with pdb2pqr was used for the generation of electrostatic maps [78].

ΔΔG stability prediction of point mutations used online tool DynaMut2, which reports a Pearson correlation of ≥0.72 [79]. Refer to (S1) in supplementary information for a complete table of stability results. DynaMut2 also provides a structural viewer which aided in identifying less obvious aromatic/hydrophobic zones that could be impacted by mutations.

3.4 Discovery of in vitro SARS-CoV-2 mutations from sequencing data

Nucleic acid extracts from SARS-CoV-2 infected CaCo-2 cell cultures were DNase treated, reverse transcribed and the amplified using a Sequence Independent Single-Primer Amplification (SISPA) method and later underwent Illumina sequencing using Nextera protocol (2 x 150bp paired-end sequencing on a MiSeq) [80]; these tasks were was performed by Public Health England and provided the raw sequencing data was used in this investigation.

Base changes were detected if the called IUPAC code was altered between datasets which are set to change when a nucleotide population reading experiences a 20% increase in favour of a new code at the next timeframe (codes called either being absolute nucleotide readings A, C, G, T or ambiguous callings of M, R, W, S, Y, K, V, H, D, B or N). The base changes through datasets were found using simple Linux commands (cut and grep with the "-f" flag) and was performed three times per isolate (*FFM3* & *FFM7*) at timepoints t=0 (before cultivation), t=1 (mid-way through cultivation at 30 passages), and t=2 (end of cultivation at 60 passages). Mutations occurring in *FFM3 and FFM7* were compared to global frequencies in GISAID data was accessed in July 2021.

4.0 Results

4.1 (S)D614G

4.1.1 D614G - finding DCPs

For the D614 vs. G614 groups (the total number of sequences included in the alignment were 22,083 and 206,236, respectively), two DCPs were revealed in the S protein by the VAT program, these were L18F and A222V which are both found on the extracellular surface of the N-terminal domain (NTD - pos.13-285) of S1 (the receptor-binding subunit - pos. 13-685) whereas D614G is found on the C-terminal domain (CTD) of S1 (see Fig. 2). The F18 mutation occurred in 0.12% of D614-matching samples and 9.2% of G614-matching samples, V222 occurred at 0.1% then 18.9% in the D614 and G614 datasets, respectively; both DCPs have BLOSUM scores of zero which do not rule out that these could have mutated by chance.

If the GISAID mega-data is separated by the same method into L18 vs. F18-separated sequences; A222V appears as a DCP (V222 found in 10.2% vs. 95.9% in the L18 and F18 datasets, respectively) and vice-versa if the set is split by A222 vs V222 - L18F appears (F18 found in 0.38% vs. 46.5% in the A222 and V222 datasets respectively); increasing the likelihood that these two DCPs are evolutionarily linked.

The lack of DCPs detected in the rest of Spike, as well as in the other SARS-CoV-2 transcriptome would suggest that other nonsynonymous mutations such as (NSP12)P323L are of lower confidence, however, L323 makes up 1.14% and 99.1% of D614 and G614-matching samples, respectively which makes P323L not a “differentially” conserved but almost wholly conserved position between the two groups. A limitation of the GISAID dataset, VAT and the alignment tools is that it cannot analyse the untranslated regions of the RNA genome (e.g. 5’UTR) and that it does not highlight silent mutations.

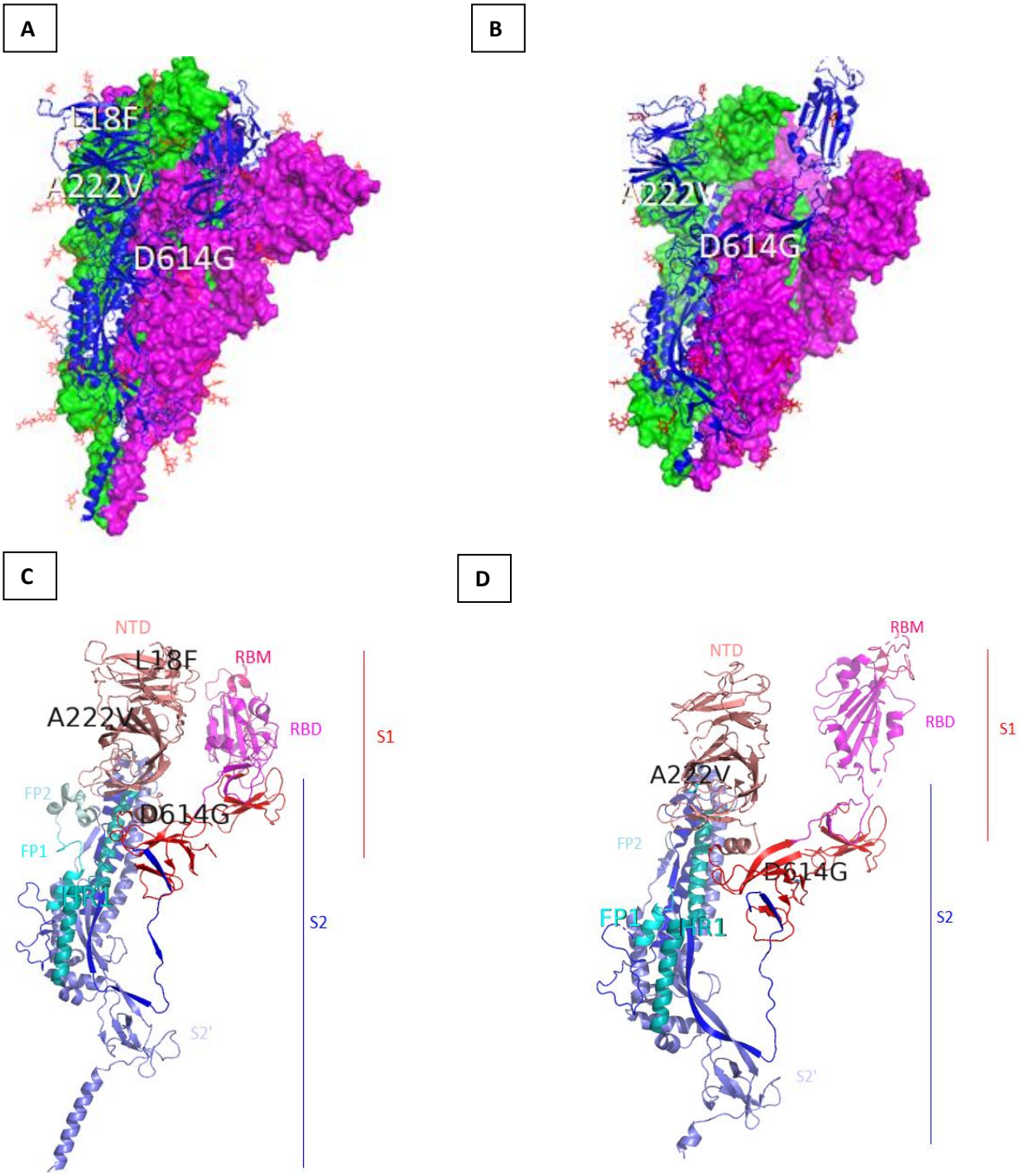


Fig. 2: (A) The S protein "closed" pre-fusion trimer (PDB: 6XR8 [43]***)*** *with chains A-C coloured in blue, magenta and green, sugars are in red sticks. Chain A displays the positioning of the D614G mutation with associated DCPs L18F and A222V.* ***(B) The S protein "one-up" pre-fusion trimer (PDB: 6VSB*** [81]***)*** *with chains A-C coloured as in (1A) with D614G and DCP A222V (L18F was not resolved), Chain A is the"up" monomer* ***(C) Chain A of the closed trimer and (D) Chain A of the one-up trimer*** *with DCPs labelled in black, S1 domain features are labelled with warm colours (i.e. red-pink) whereas S2 domain and its features are in cool colours (i.e. blue-green) with non-dark blue regions being included as part of S2'. NTD = N-terminal domain of S1 (salmon), RBD = Receptor binding domain (magenta), RBM = Receptor binding motif (hot pink), FP1 = Fusion peptide 1 (cyan), FP2 = Fusion peptide 2 (light cyan), HR1 = Heptad repeat 1 (teal)*

4.1.2 L18F – structural analysis

Closed form

L18F is not often captured in structures generated by X-ray diffraction/EM but is more commonly resolved in the “closed” Spike formation (an exception is used in the L18F “one-up” section where S trimer is bound to nano-Abs), which would suggest it is a highly flexible region. (Refer to Fig. 3) L18F is a nearby disulphide bond (C15-C135) which suggests intrachain stabilisation of the NTD; L18F also neighbours an N-linked glycosylation site (N17-GlcNAc-GlcNAc).

(Ref. Fig. 3) Both L18 and F18 sidechains have similar uncharged/hydrophobic properties, F18’s R group is larger and causes steric clashing of the larger phenol attachment against F140 and S255 (in other Spike structures such as 6ZB4, 6ZB5, 6ZGE and 6ZGI; F140 clashes are a common theme but L244 clashes instead of S255). To speculate, there may be a reorganisation of this area - perhaps F18 could have the R-group rotated by the combined clashing of F140 and S255, which puts pressure on the backbone to twist - resulting in 1) different angle of projection for the N17-glycan, 2) possibly altering the distance and ability of the disulphide bond between C15-C136 to form, and 3) could push the loop turn (pos.254-256) on which S255 sits towards the solvent; there also may be a possibility of aromatic stacking of F18 with F140 and/or W258 (shown in Figs.3-6) which could cause also cause structural changes to the N-glycan and disulphide bond.

A positional change in the N17-glycan may impact immune detection, folding or binding, whereas the disturbance of C15-C136 and nearby C131- C166 could alter NTD stability and/or folding, which may impact how the NTD (positions 13-303) interacts with the RBD (positions 319-541) in the “one-up” conformation (refer to the end of L18F “one-up” section for detail).

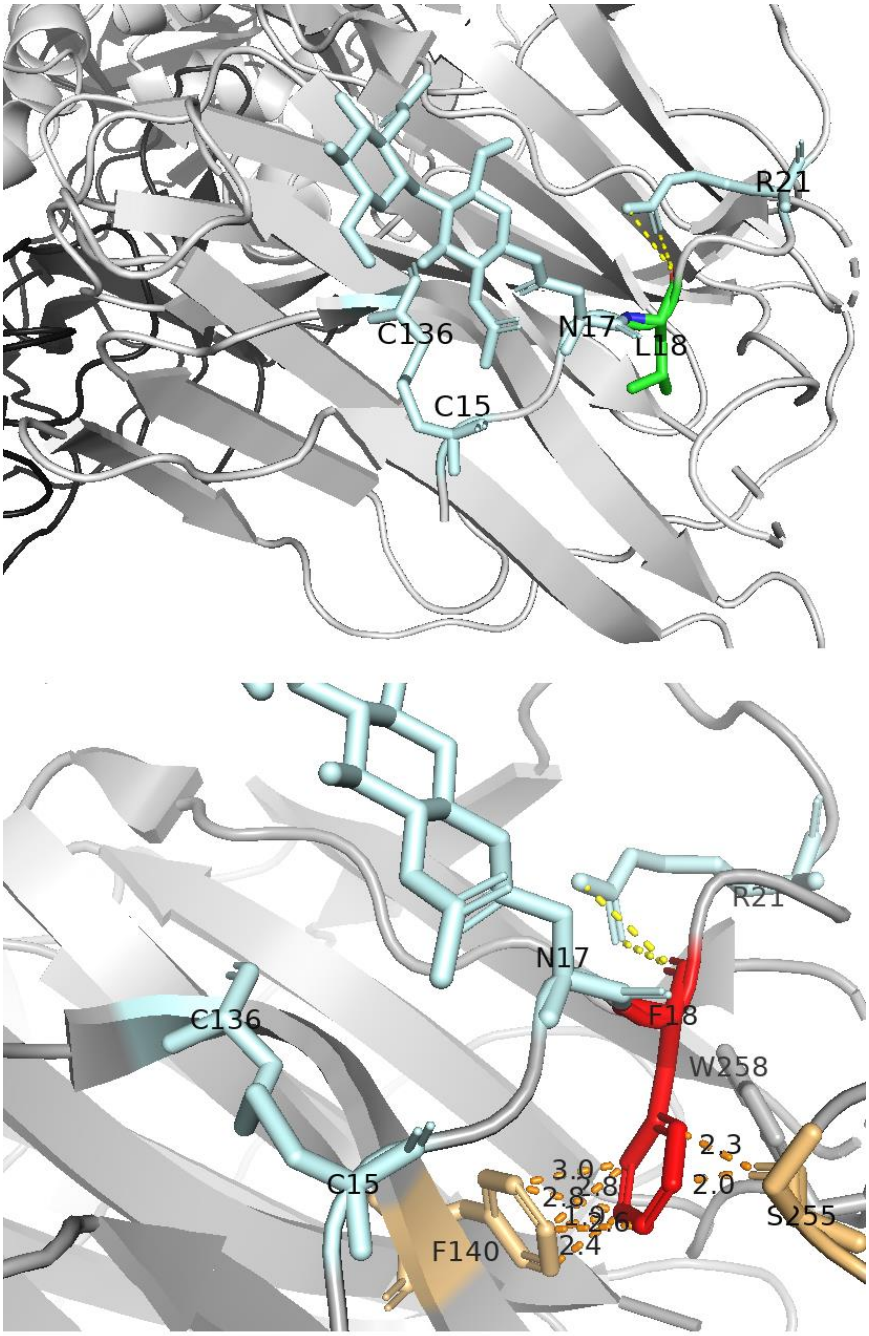


Fig. 3 - Close-up of the Spike L18F DCP from Chain A of the "closed" prefusion trimer PDB:6XR8: *The top image shows pre-mutation (L18) and the bottom image shows a zoom-in of the post-mutation (F18). Chain A backbone (light grey) is used to represent all 3 chains in the closed configuration. Colour scheme: pre-mutation DCP = green, post-mutation DCP = red. Residues of interest (e.g. disulphide bridges, N-glycans and hydrogen-bonded residues) are in light cyan, clashing residues are in light orange. Polar interactions are in yellow dashes and steric clashes are orange dashes labelled with distance in Å. W258 is left uncoloured as interaction is unknown. Chains A-C are shaded in sequential order from light grey -> mid-grey -> black. i.e. The Chain of interest (A) = light grey, (B) = mid-grey and (C) = black*

“One-up” form (fused to nano-antibodies)

Due to low availability of “one-up” pre-fusion trimers that contain L18F, a nano-Antibodyfragment (Ty1) bound structure (PDB: 6ZXN [82]) was used to study L18F in the “one-up” state. 6ZXN alignment to 6VSB is available in (S3), from the alignment the calculated RMSD (Root Mean Square Deviation) values = 5.587Å (Ty1 atoms not removed) and RMSD = 0.590 Å (Ty1 atoms removed).

Chain A: (Refer to Fig. 4) In the “one-up” Ab-fused state, F18 on Chain A clashes with F140, L244 and R246; the latter two belonging to a β-strand (β19) that occurs just before the turn that was previously affected in the “closed state” (loop pos.254-256). F140 sits on β12 (pos. 133-146, which notably overlaps with C131) which forms part of a six-stranded mixed β-sheet which also includes: β7, β9, β11, β13, β18-19 (β18-19 are continuous in 6ZXN but are distinct sheets in Uniprot annotation). Although F18 clashes with one more additional residue in comparison to the “closed” chains, the calculated steric strain on F18 is weaker in the “one-up” conformation (PyMOL Mutagenesis tool calculation: 21.78 vs. 27.65).

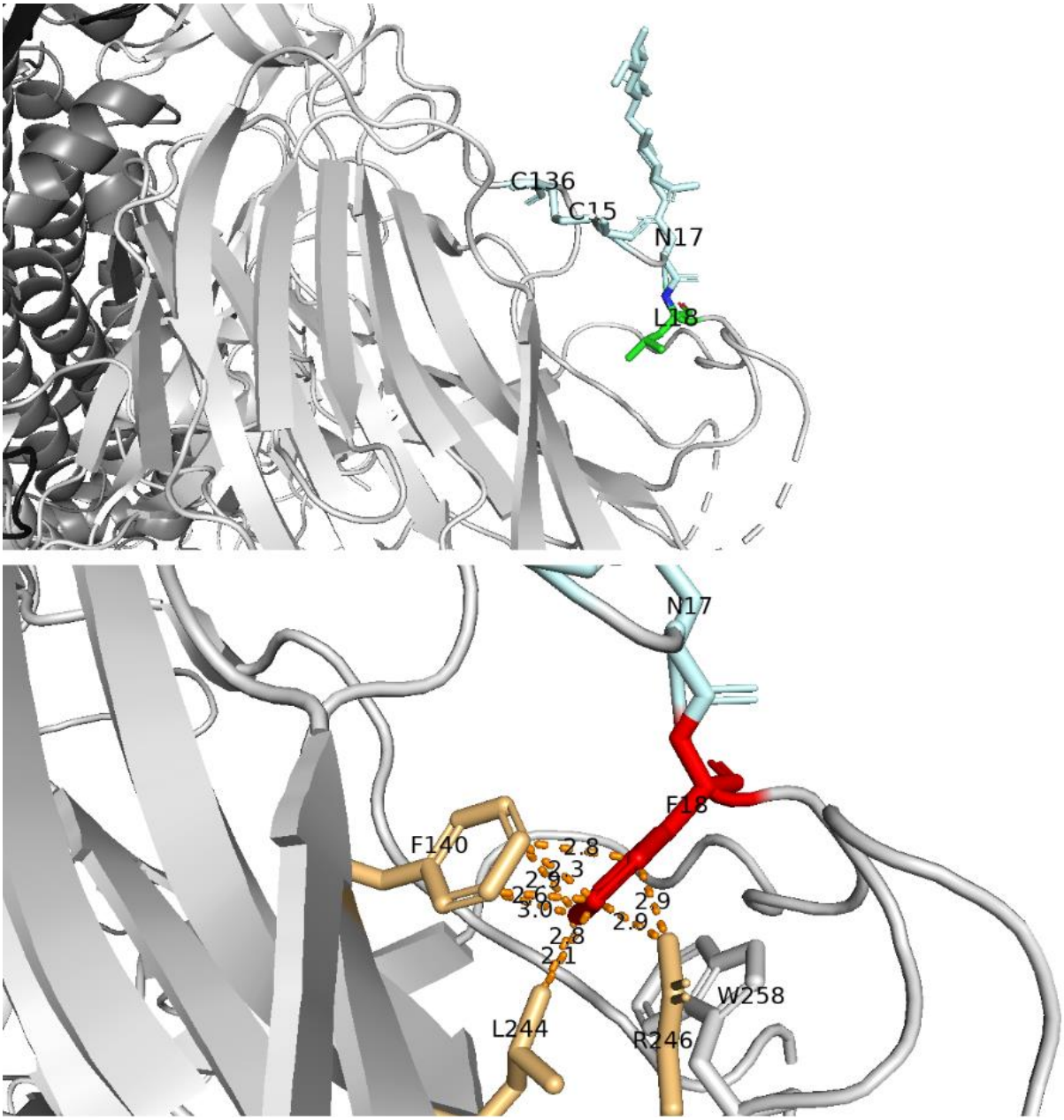


Fig. 4 - Close-up of the Spike L18F DCP from Chain A of the" one-up" post-fusion trimer PDB:6ZXN: *The left image shows pre-mutation (L18) and the right image shows a zoom-in of the post-mutation (F18). Colour scheme – same as in Fig.3.*

Chain B: (Ref. to Fig. 5) Steric clashing occurs with three nearby residues, again F140 being the most obvious but the remaining clashing residues have shifted to R246 and S254; the former at the border between β19 and the latter within the previously mentioned proceeding turn. This Chain during the "one-up" conformation experiences the least steric tension of all possible situations (Strain: 19.88).

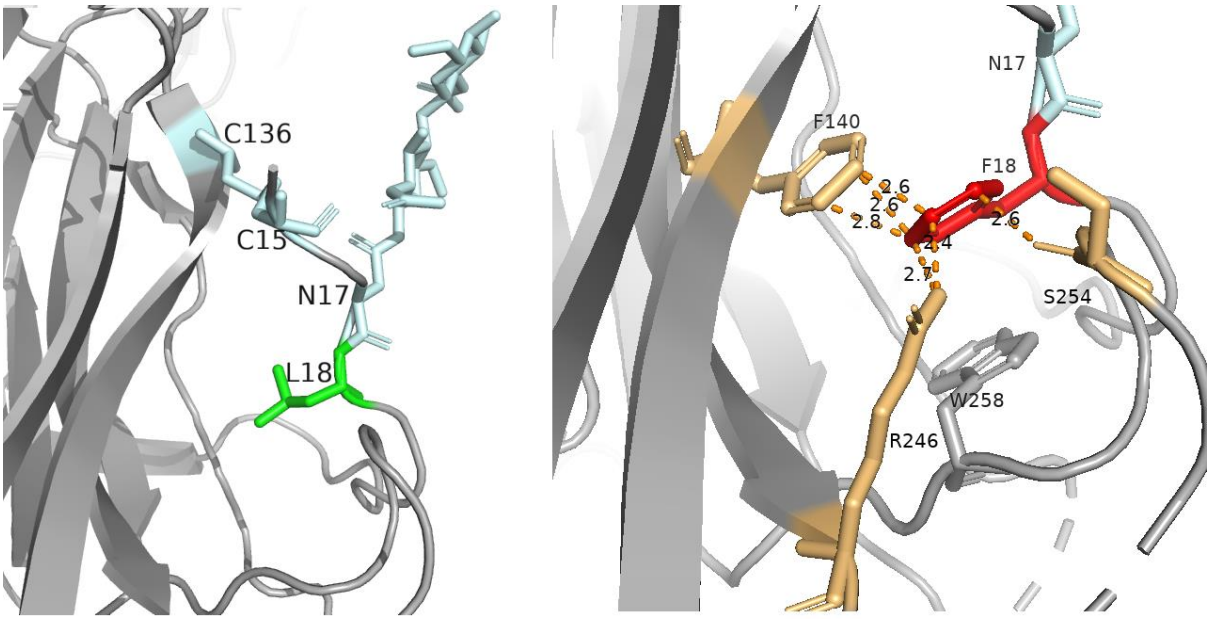


Fig. 5 - Close-up of the Spike L18F DCP from Chain B of the "one-up" post-fusion trimer PDB:6ZXN: *The left image shows pre-mutation (L18) and the right image shows a zoom-in of the post-mutation (F18). Colour scheme – same as in Figs.3-4 except Chain of interest (light grey) = B, mid-grey = C, black = A. Note: Chains A/C are not visible.*

Chain C: (Ref. to Fig. 6) Contains the only occurrence of polar interaction in the "one-up" trimer conformation via a backbone amine to the carboxyl group of D138. F18 in Chain C has the fewest clashing residues but undergoes the strongest level of strain (48.89) against F140 which would cause heavy repulsion of F18.

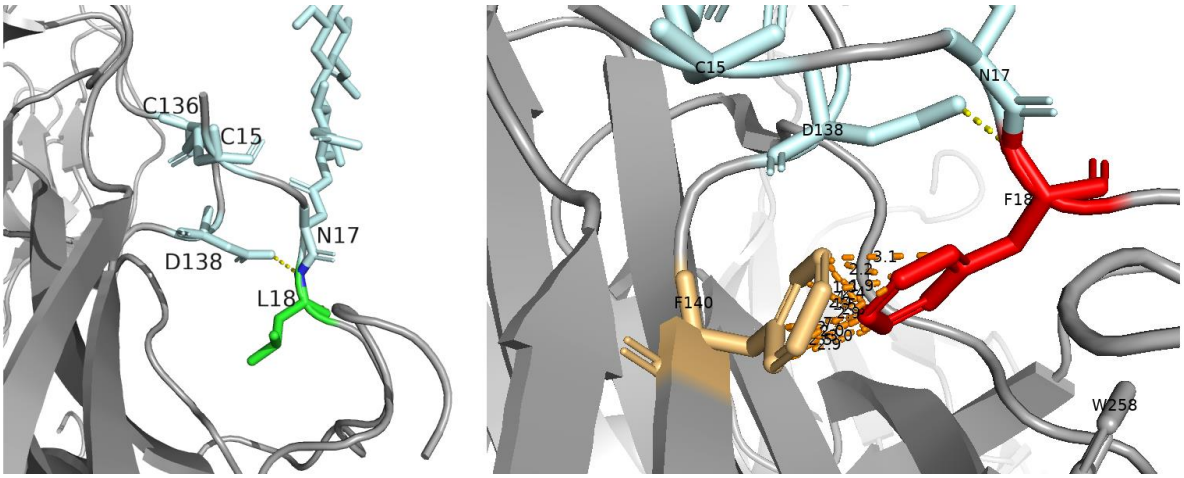


Fig. 6 - Close-up of the Spike L18F DCP from Chain C of the "one-up" post-fusion trimer PDB:6ZXN: *The left image shows pre-mutation (L18) and the right image shows a zoom-in of the post-mutation (F18). Colour scheme – same as in Figs.3-4 except Chain of interest (light grey) = C, mid-grey = A, black. Note: Chains B-C are not visible.*

Possible Impact of F18 on nearby disulphide bonds and NTD-RBD tether

In the Spike "closed" form there are 5 interchain polar contacts between NTD of Chain A to the RBD of Chain C: K113-S469, Y200-R355, I231-R466, G232-R466, and N234-K462 which locks the RBD in the closed position (not shown) - all of which are too distant to be affected by C15-C136.

However, in the "one-up" Spike trimer conformation (refer to Fig. 7) – on the NTD of Chain B (anticlockwise of Chain A in Fig. 2A-B) there is a single interchain hydrogen bond to the RBD of Chain A and NTD of Chain B (T167-R357) appearing to act as a tether for the outstretched RBD (in the PDB 6VSB the NTD tether is N165, see (S2) in supplementary figures). Perhaps a lack of stabilisation from intrachain C15-C136 as a result of L18F could have a knock-on effect on C131-C166. It is not confirmed if C15-C136 influences the stability of C131-C166, however in another "closed" structure 6ZB4, where there is no C15-C136 bridge - C131-C166 was also missing (not shown).

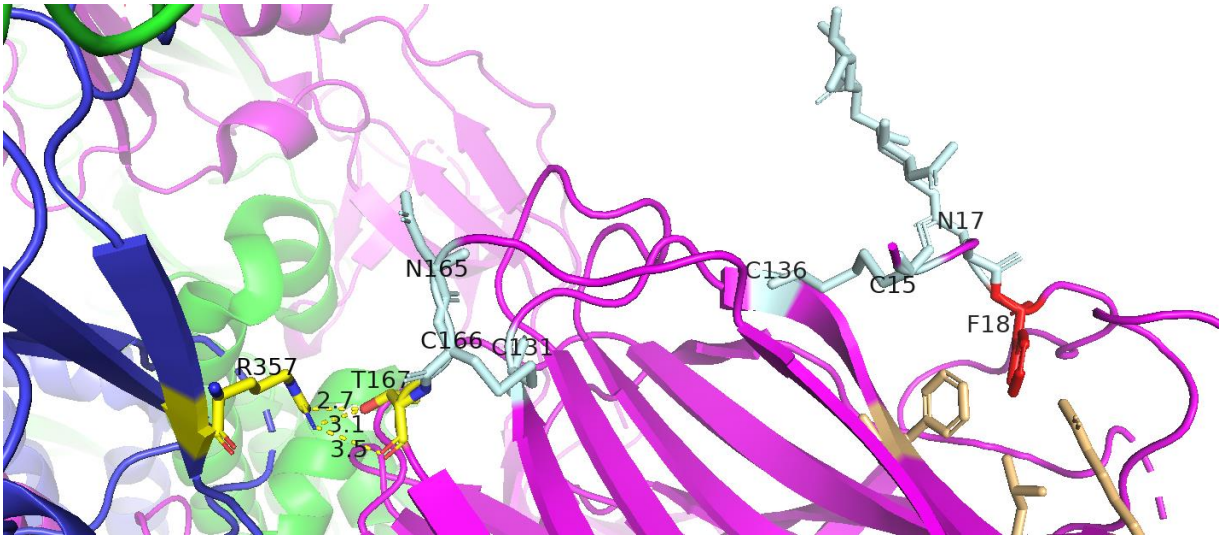


Fig. 7 - Spike "one-up" conformation: interaction of the RBD of the "up" monomer (Chain A) with the NTD of Chain B with nearby disulphide bridges (C131-C166 and C15-C136) PDB: 6ZXN. Chains are coloured as in Fig. 1A-B, interchain interaction (Chain A's RBD to Chain C's NTD) - R357-N165 coloured elemental yellow, features such as N-glycans/disulphide bridges are light cyan

Summary of L18F findings:

L18F is near the disulphide bond C15-C136, implying possible impact on NTD-stabilisation with possible secondary effects on neighbouring C131-C166 which is associated with a "tether-acting" residue (i.e. T167 or N165) that could stabilise the "one-up" conformation. Possible interactions with glycans (N17 directly and indirectly N165) may impact folding, immune recognition or even binding; current evidence for the replicative advantage of L18F suggests poorer neutralisation via reduced Ab-binding to the NTD [83]–[85].

Clashing with sidechains (i.e. L244, R246, S254, S255) and F140 - which may have stacking potential (in addition to W258), is more likely to push F18 and the unstructured loop (pos.254-256) towards the solvent rather than disorganising the β-sheets (β7, β9, β11, β12, β13, β18-19) which is less energy-favourable; the repulsive strain of F18's larger R-group was slightly more pronounced in the "one-up" conformation than the "closed" conformation (21.78 vs. Chains A-C: 27.65, 19.88, 48.89), which was

partially reflected in the Dynamut2 calculated ΔΔG values (6XR8 “closed” chains A-C: -0.55, -0.62, -0.6; 6ZXN “one-up” chains A-C: -0.52, -0.72, -0.64 respectively with units in kcal/mol)

4.1.3 A222V – structural analysis

Closed form

(Ref. to Fig. 8) A222V is the representative mutation of the GV clade division, when Spike is in the closed state, both A222 and V222 share a backbone polar interaction to the sidechain of K206 in all Spike monomers – this bond is not present in any Chains in the “one-up” conformation. The singular difference of the mutation is that V222 causes a steric clash with I285; nearby N282 which carries a complex glycan - N-GlcNAc-GlcNAc-Man (note that only N-GlcNAc is resolved in Fig. 9). If the loop on which N282-glycan sits (loop is from Y279 up to I285) is pushed away from V222, there could be modest impacts on immune shielding, folding or bonding of this area if the glycan chain is moved.

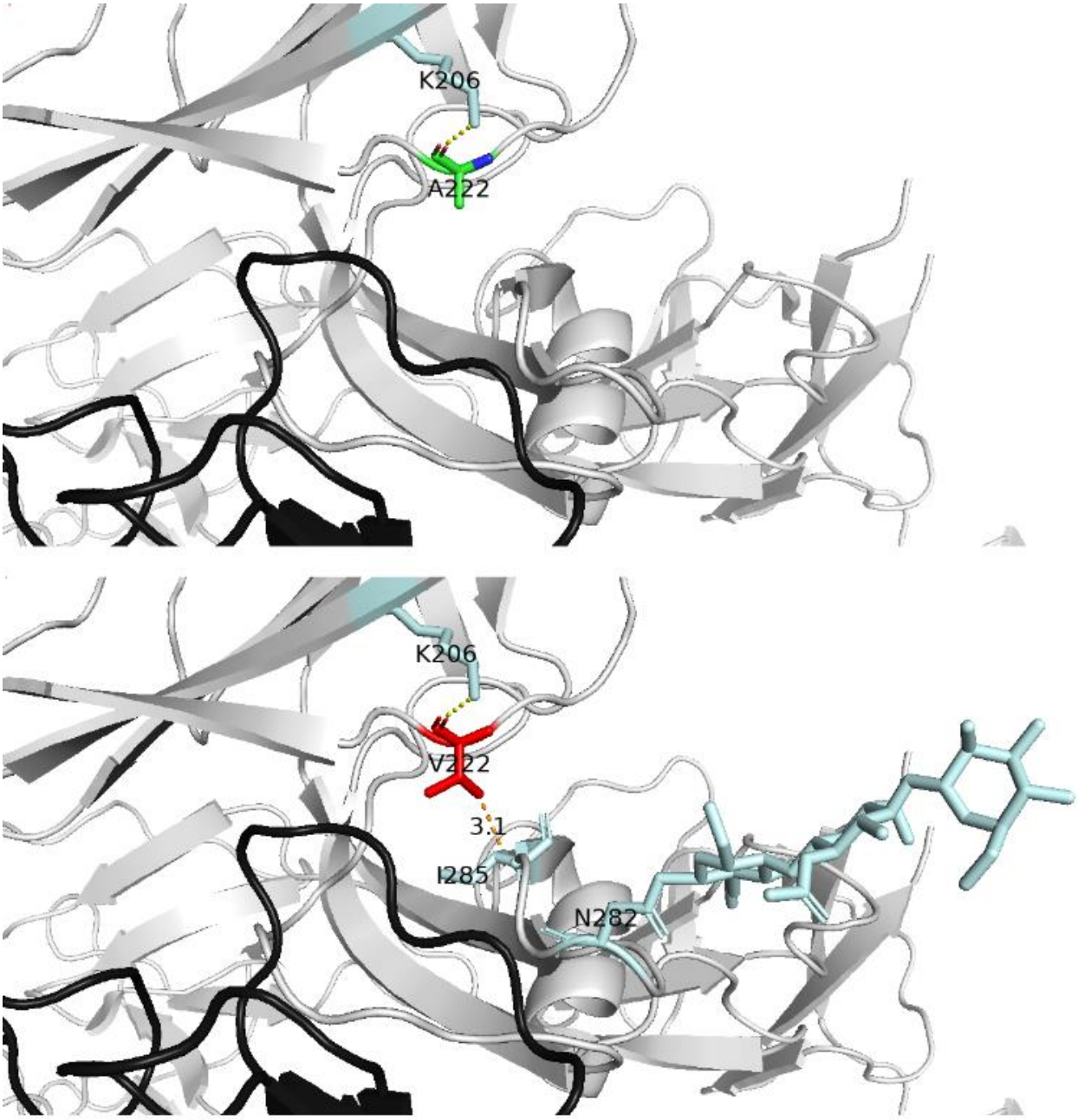


Fig. 8 – Close-up of the Spike A222V DCP from Chain A of the "closed" prefusion trimer (PDB:6XR8) *The upper image shows pre-mutation (A222) and the bottom image shows the post-mutation (V222). Colour scheme – same as in Figs.3-4 Note: N282-glycan is not fully glycosylated, and Chain B is hidden for clarity.*

"One-up" form

Chain A: (Ref. Fig. 9) Although A222 nor V222 on Chain A (the "up" extended monomer) have similar appearances, with V222 there is intrachain clashing against Y38 which could cause minor positional shifting with ring rotation. Chain C's F562 is close in distance to Y38, however, no interchain π-π interaction was found for Y38 in its original rotation nor when a lower-strain rotamer was chosen, however aromatic stacking does occur naturally and post-mutation later in Chain C). Y38 shares backbone interaction with L221 which is not altered with choosing the lowest-strain Y38 rotamer (not shown).

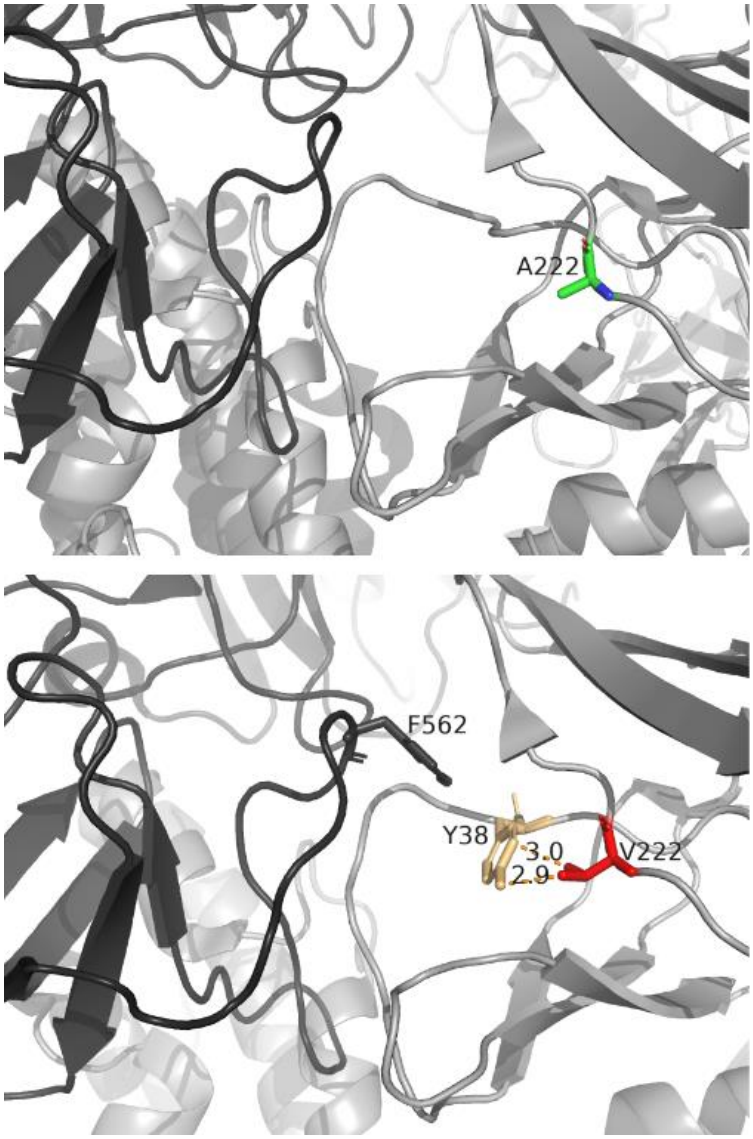


Fig. 9 – Close-up of the Spike A222V DCP from Chain A of the "one-up" prefusion trimer (PDB:6VSB) *The upper image shows pre-mutation (A222) and the bottom image shows the post-mutation (V222). Colour scheme – same as in Figs.3-4. Note: Chain B is hidden for clarity.*

Chain B: (Ref. to Fig. 10) As with the "closed" trimer chains, no polar contacts were found pre- or post-mutation in Chain B of the "one-up" trimer, however similarly to the "closed" trimer form, there is clashing with I285 – only more slightly more pronounced with only a 0.1Å distance difference and greater strain (21.6 strain Chain B compared to strain 17.39 in the "closed" Chains). There is low significance clashing with Y37 and Y38 which is not shown.

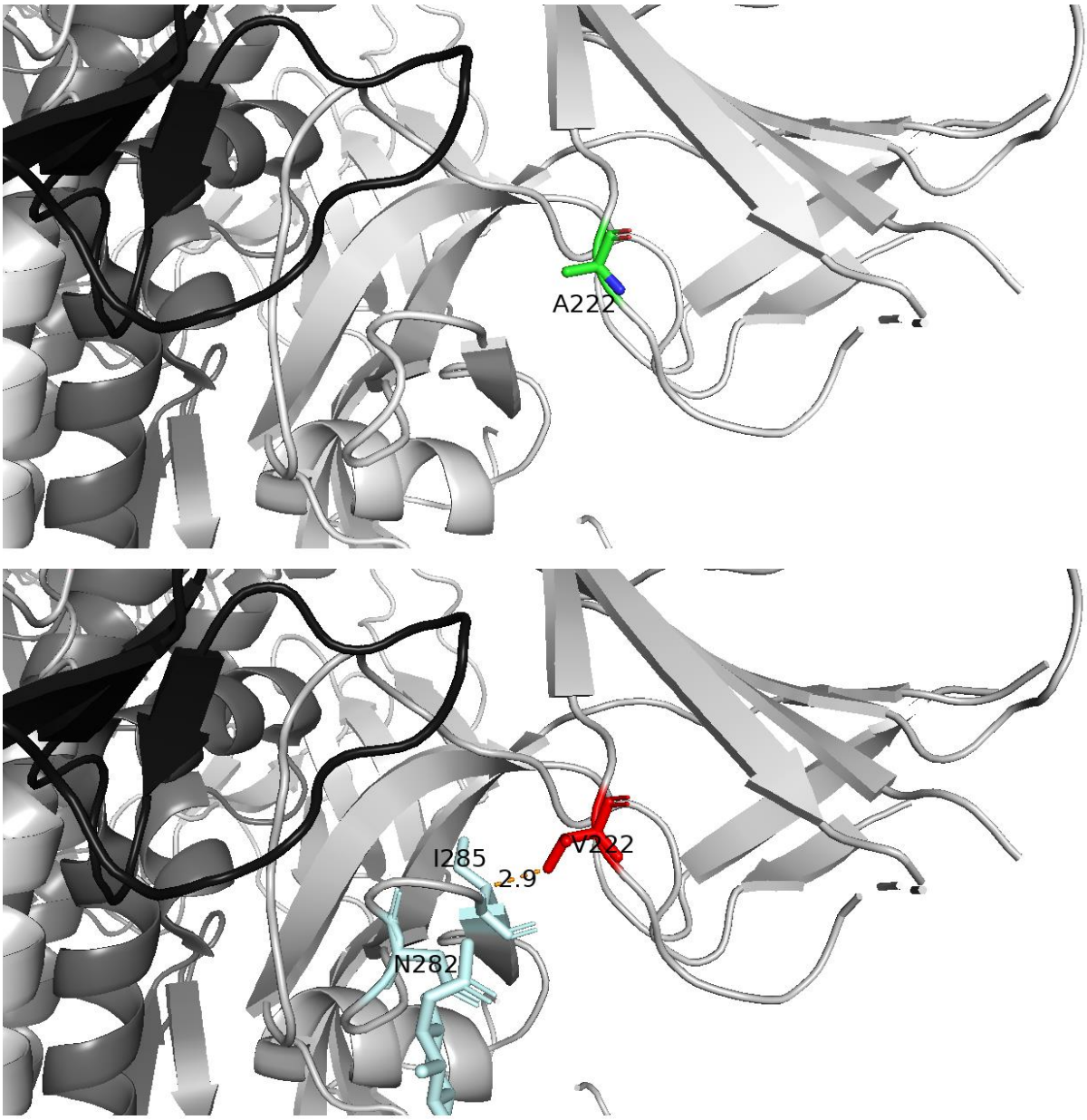


Fig. 10 - Close-up of the Spike A222V DCP from Chain B of the "one-up" prefusion trimer (PDB:6VSB) *The upper image shows pre-mutation (A222) and the bottom image shows the post-mutation (V222). Colour scheme – same as in Figs.3-4. however, Chain (A-C) shading follows Fig.5*

Chain C: (Refer to Fig. 11) This monomer holds the only occurrence of A222 and/or V222 sharing a hydrogen bond with neighbouring S221. In the "one-up" trimer form, Chain C (in "down" position, clockwise of Chain A in Fig. 2B) has Van Der Waals overlap with Y38. The steric clash with Y38 on Chain C is more pronounced than Chain A (strain values: 25.71 vs. 16.30) as the ortho and meta positions on the phenyl ring clash instead of against the meta position and para-hydroxyl group.

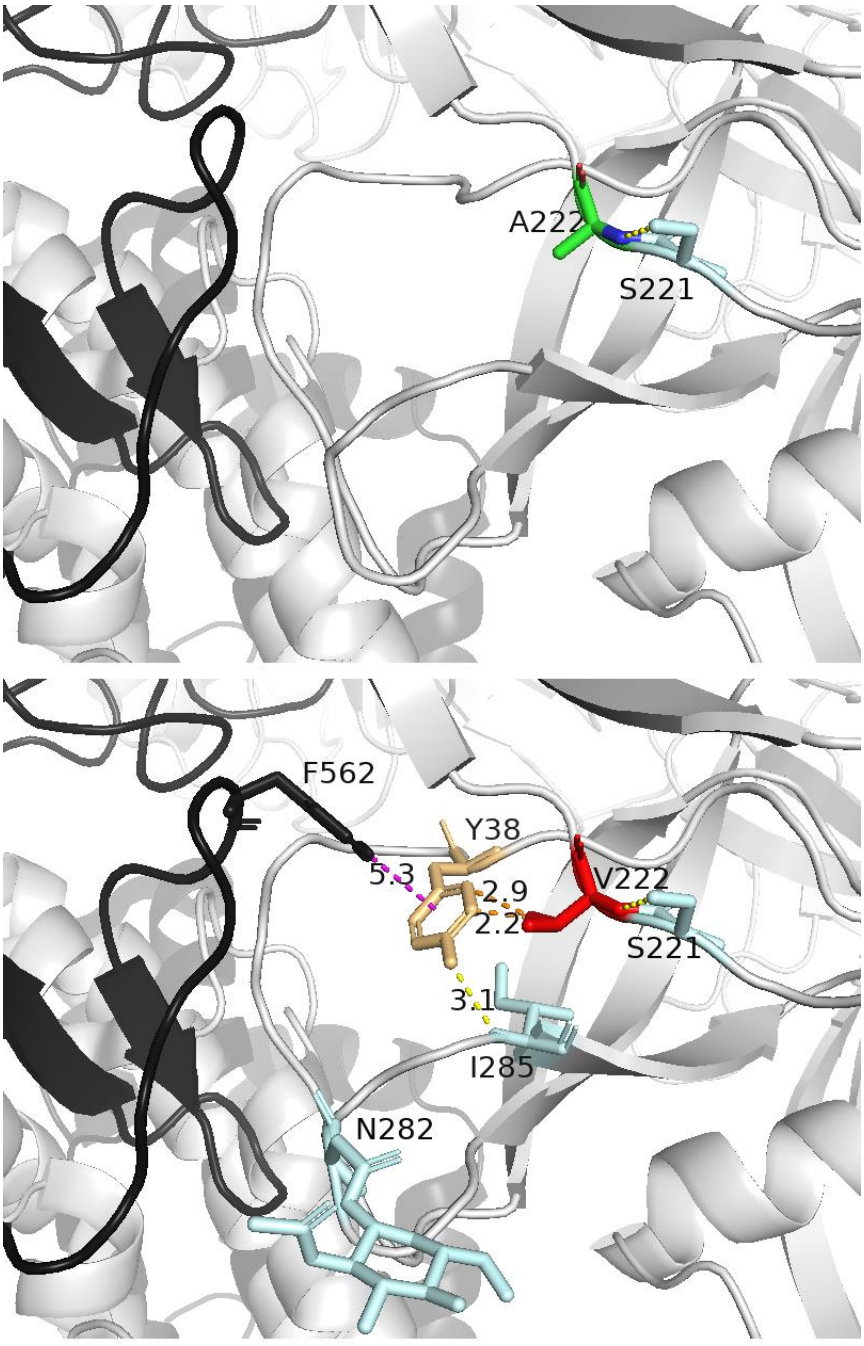


Fig. 11 - Close-up of the Spike A222V DCP from Chain C of the "one-up" prefusion trimer (PDB:6VSB) *The upper image shows pre-mutation (A222) and the bottom image shows the post-mutation (V222). Colour scheme – same as in Figs.3-4. however, Chain (A-C) shading follows Fig.6 pattern. Magenta dashes represent aromatic π- π interactions.*

For Chain C, there is the detection of π - π aromatic stacking for Y38 to Chain B's F562 when Y38 is left in its original rotamer or is adjusted by the steric clashes induced by V222. In addition, Y38 on Chain C shares a backbone hydrogen bond to L221 (bond distance not affected by rotamer choice) and another polar bond via the hydroxyl group to the intrachain backbone amine of I285, the length of which is reduced from 3.1 to 3.0Å when the lowest strain rotamer is chosen (displayed in (S4).

Summary of A222V findings

A222V introduces minor clashes and very subtle changes, the most significant being against 1) I285 (in both the "closed" or "one-up" trimer state in Chain B) which connects to a short loop containing a large glycan chain (N282-GlcNAc-GlcNAc-Man) which could offer slightly different immune protection, folding or binding of S protein and 2) Y38 (in the "one-up" trimer state of Chain C which engages in interchain T-shaped aromatic stacking with F562 and simultaneously shares a slightly shortened hydrogen bond to I285. Order of strain strength is as follows: Chain C (25.71), Chain B (21.6), all "closed" state Chains (17.39) then finally Chain A (16.30).A222V could have a hydrophobic-based stabilising effect on Y38/I285 (and possibly V36 and F220 which is not shown in Figs.7-10) which was predicted in the 6XR8 "closed" and 6VSB "one-up" PDB structures by the online stability prediction tool DynaMut2 which gave an average ΔΔG increase (+0.23 and +0.27 kcal/mol, respectively); A222V may be advantageous for modulating immunogenic response as A222V is suspected to be within a region of B-cell epitope recognition or could be involved in allosteric binding [86].

4.1.4 D614G – structural analysis

Closed form

(Ref. Fig 12) When in the "closed" form, D614 forms a salt bridge with K854 and with the backbone amine of K835; both Lysine residues belong to the anticlockwise monomer (i.e. Chain B in Fig. 2A), these interchain interactions are lost entirely with the G614 mutation which is likely to be destabilising. Nearby residue features of D614G which may be impacted by the mutation are the disulphide bridge C617-C649 and a complex N616-glycan (N-GlcNAc, GlcNAc, Mannose) and on Chain B there is another disulphide bridge, C840-C851 which resides close to the previously mentioned Lysine residues that connect Chain A to Chain B.

"One-up" form

Chain A: (Ref Fig. 13), D614 in Chain A was found to share a hydrogen bond with T859 which is lost on the replacement by G614 which is already known [50], [87], [88]. There is an unchanged stabilising intrachain bond between D614G and A647 and in this structure, the nearby N616-glycan (N-GlcNac) is less glycosylated than the "closed" form – although this could be due to failure to resolve the whole sugar chain.

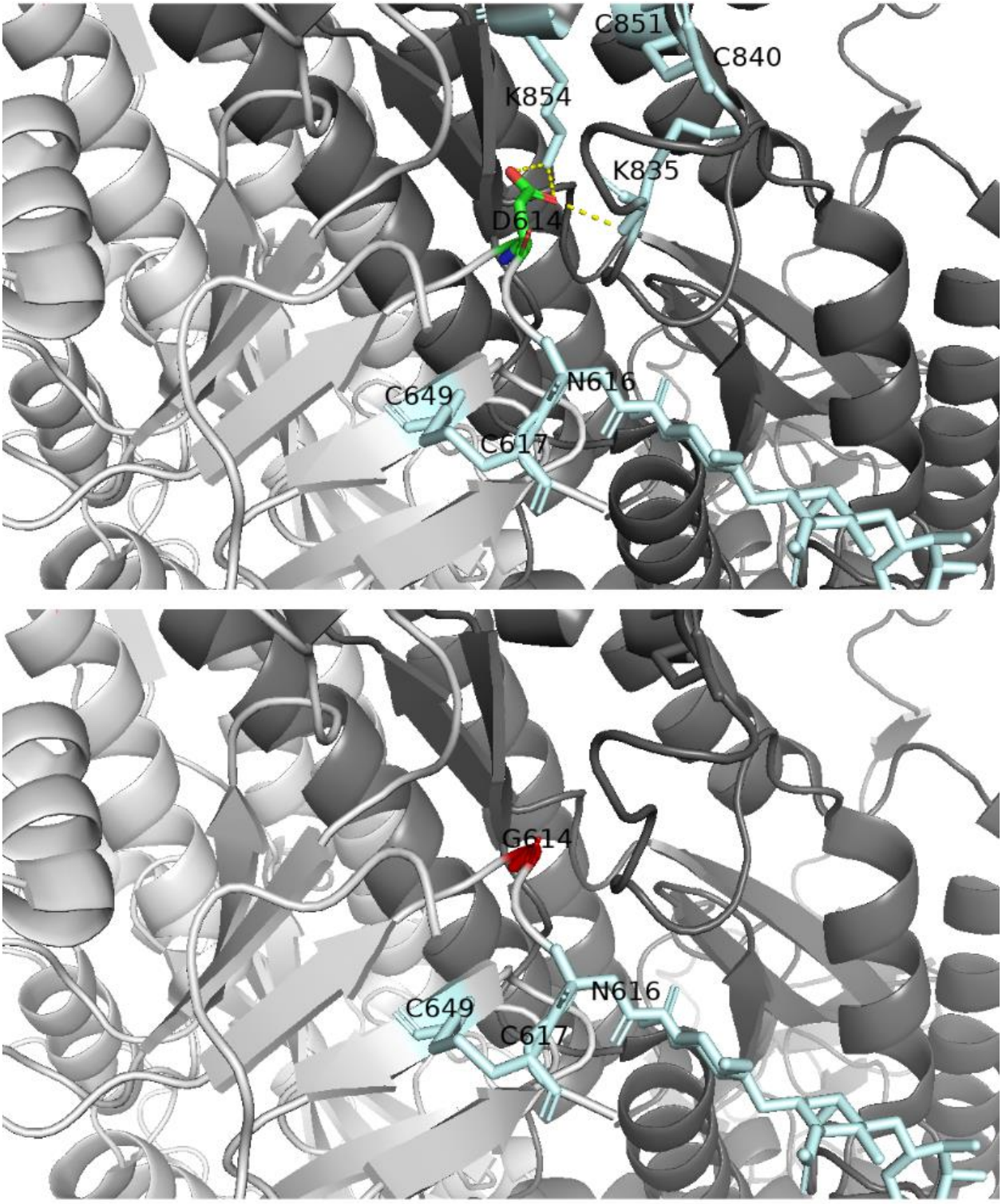


Fig. 12 – Close-up of the Spike D614G DCP from Chain A of the "closed" prefusion trimer (PDB:6XR8) *The upper image shows pre-mutation (D614) and the bottom image shows the post-mutation (G614). Colour scheme – same as in Figs.3-4.*

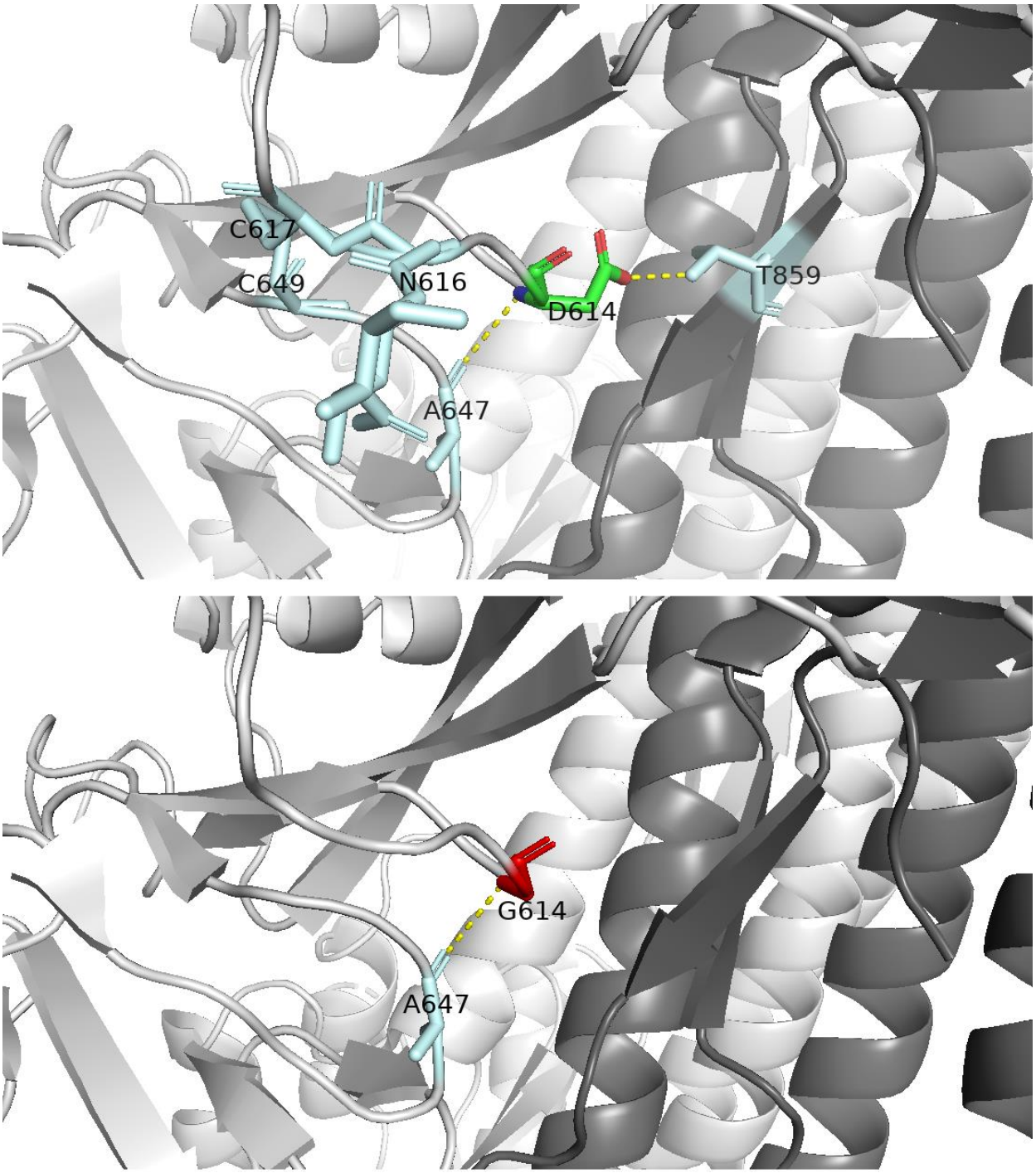


Fig. 13 Close-up of the Spike D614G DCP from Chain A of the "one-up" prefusion trimer (PDB:6VSB) *The upper image shows pre-mutation (D614) and the bottom image shows the post-mutation (G614). Colour scheme – same as in Figs.3-4.*

Chain B: (Ref. Fig. 14) in the "one-up" state there is backbone intrachain stabilisation to A647, similarly to Chain A and Chain C. However, only a single high-strain D614 rotamer out of seven was able to bond to Chain C with almost double the strain compared to the other rotations (e.g., the other six rotamer strains range from 13.61-16.51 whereas the only interchain-connecting rotamer strain was 28.26) which indicates lower probability that there are any interchain interactions mediated by D614 on Chain B during the "one-up state". There is likely to be a minimal impact on stability in this case although there would be an overall tendency to increase flexibility as glycine's smaller size will reduce strain induced by sidechains.

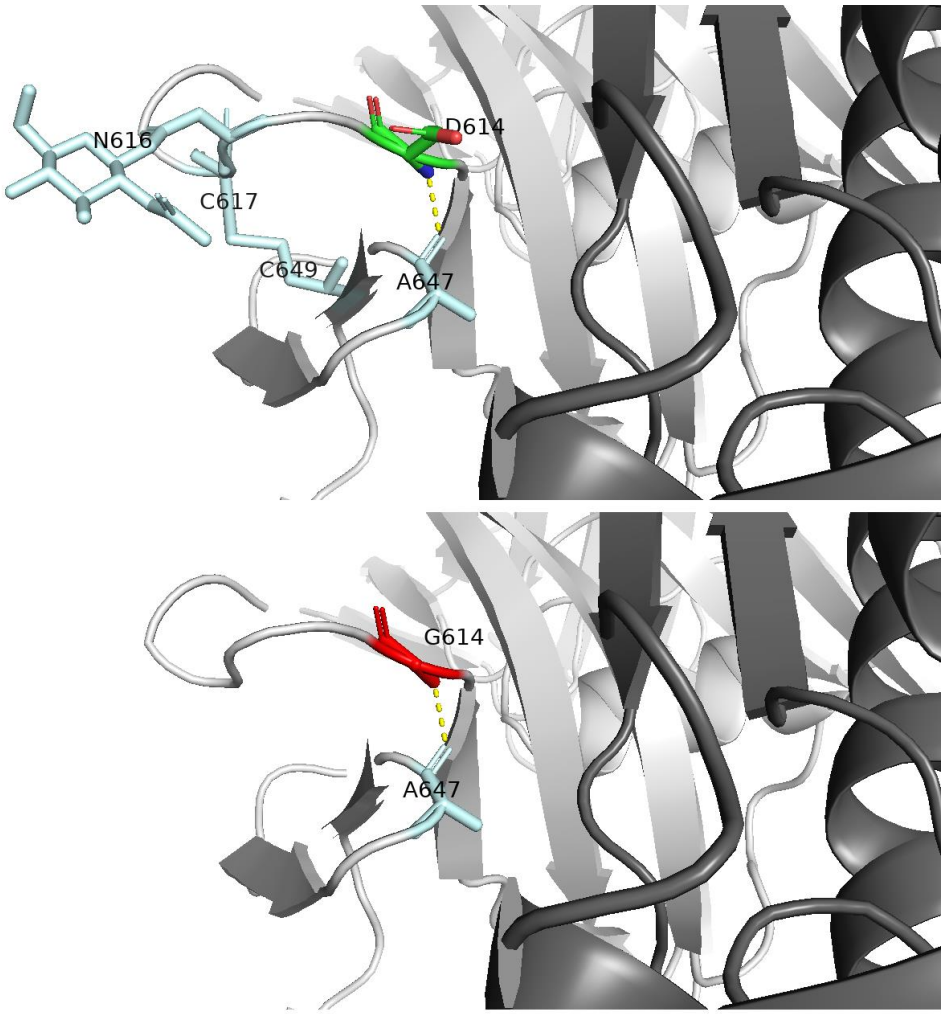


Fig. 14 - Close-up of the Spike D614G DCP from Chain B of the "one-up" prefusion trimer (PDB:6VSB) *The upper image shows pre-mutation (D614) and the bottom image shows the post-mutation (G614). Colour scheme – same as in Figs.3-4 however, Chain (A-C) shading follows Fig.5 pattern.*

Chain C: (Ref. Fig. 15) D614 behaves similarly as in Chain A bonding with A647 (intrachain) and Chain B's T859 (interchain to Chain A), the introduction of G614 results in loss of the interchain connection and likely increases chain flexibility.

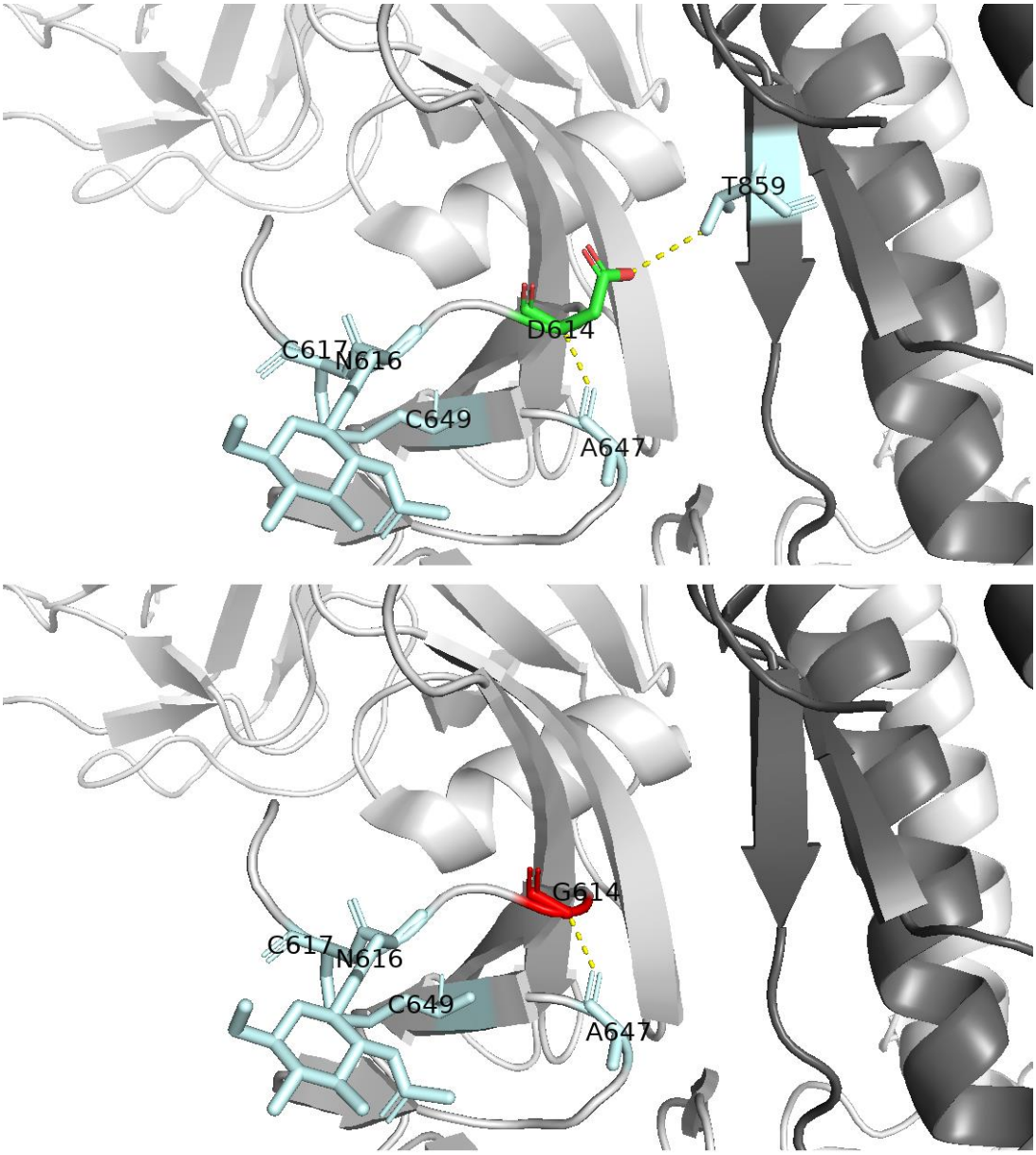


Fig. 15 - Close-up of the Spike D614G DCP from Chain C of the "one-up" prefusion trimer (PDB:6VSB) *The upper image shows pre-mutation (D614) and the bottom image shows the post-mutation (G614). Colour scheme – same as in Figs.3-4 however, Chain (A-C) shading follows Fig.6 pattern.*

Summary of (S)D614G findings

As a result of D614G, in the “one-up” state - Chain A D614-T859 and the Chain C D614-T859 is lost with the G614 mutation which lessens association with the monomers on either side which could lead to increased flexibility and capability of a monomer to exist in the “one-up” state as mentioned in other papers due to reduced anchorage.

However, there is an additional ramification of D614G in the “closed” state with the loss of salt bridges to interchain lysine residues (K835 and K854). The loss of ionic bonds would strongly suggest destabilisation of S protein specifically in the “closed” state, which offers a lower energy requirement to shift to the “one-up” state. This disparity in stability for the “closed” and “one-up” states was confirmed by Dynamut2 with 6XR8 “closed” Chains A-C: -0.51, -0.78, -0.63 and 6VSB: “one-up” Chains A-C: -0.3, -0.23, -0.29 (kcal/mol).

Aside from this, there are nearby disulphide bridges (C617-C649 and C840-C851) and glycan N616 which may have unknown structural or binding alterations.

4.2 (S)N501Y

4.2.1 (S)N501Y– finding DCPs

For the N501 vs. Y501 DCP analysis, the total number of sequences used in the alignment were 243,092 and 1,424, respectively taken on 14th Dec 2020. The VAT program calculated five total DCPs (A570D, P681H, T716I, S982A and D1118H), the incidence of native residues decreased by ~66.4-66.5% across the separated groups; these DCPs are identical to those found in a sub-lineage from a previous study that occurred alongside the HV69-70del and Y144del (also described as Y145del due to adjacent tyrosine residues) deletions [60]; all other mutations in NSP3/6, accessory protein 8 and N highlighted in [60] were not found to be DCPs (see Table 1).

The post-mutation percentage of each DCP in the N501 vs. Y501-respective groupings are as follows D570 = 0.0008% and 66.5%, H681 = 0.12% and 66.4%, I716 = 85 and 84%, A982= 0% and 66.4% and H1118 = 85 and 88%; there is only a singular positive integer BLOSUM score in this set (S982A = +1). Referring to the two deletions, HV69-70del occurrence rate rose from 1.7% to 66.5% and Y144del increased from 0.1% to 66.5% in comparisons against N501 vs. Y501 groupings.

(Ref. Fig. 16) N501Y is found on the RBM (pos.437-508) which the region responsible for binding ACE2, A570D and P618H lay in the CTD of the S1 domain; P681H sits directly in front of the “RRAR” motif which is the S1/S2 boundary furin cleavage site [35], [38], [60], [89], [90]. T716I is found early in the S2 domain upstream of the TMPRSS2 S2’ cleavage site (S2’ - pos.816-1273) [34], [91]. The coiled-coil domain (pos.949-993) contains S982A and similarly to D1118H, lies between the heptad repeats (HR1 and HR2; pos.920-970 and 1163-1202, respectively).

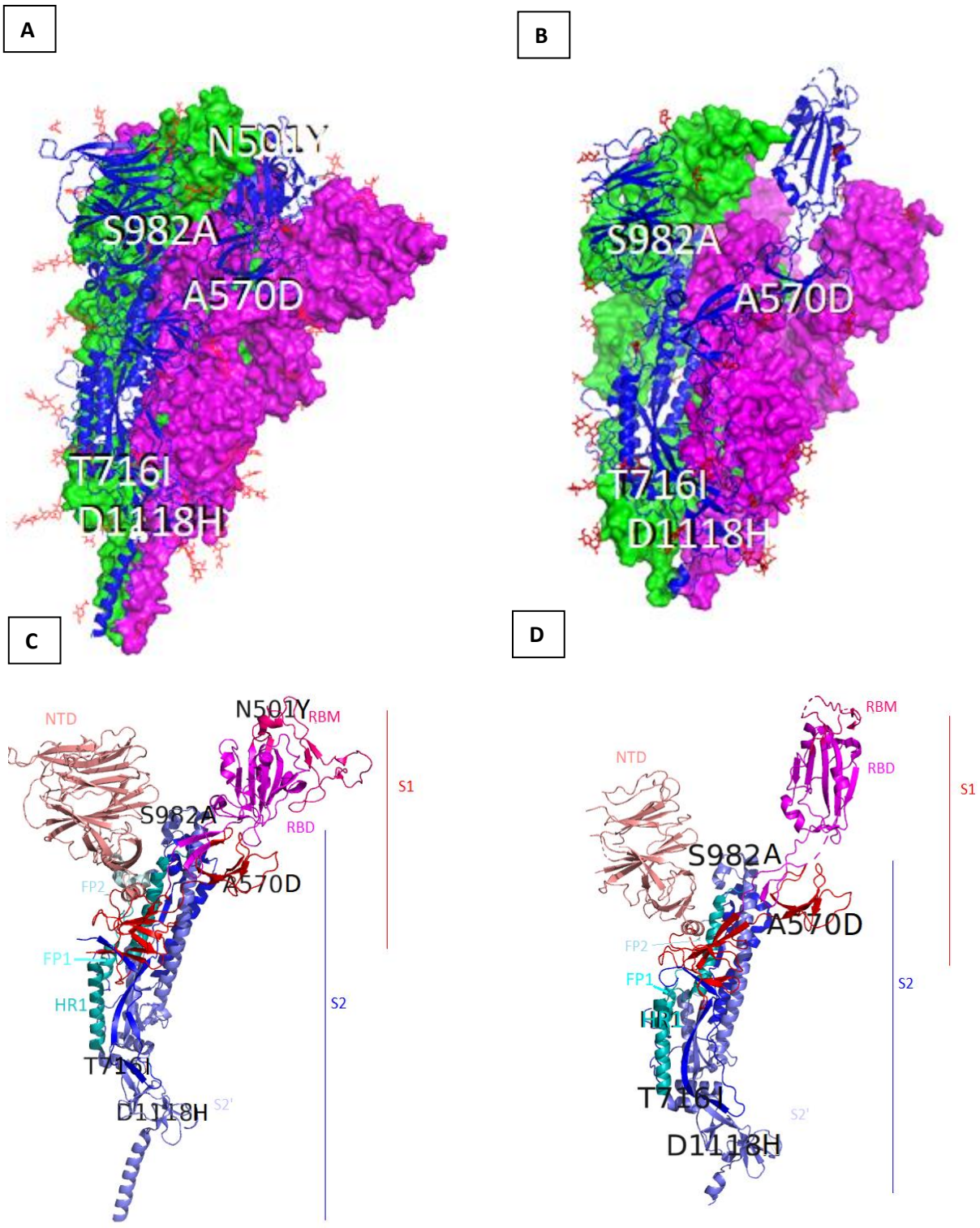


Fig. 16 - (A) The S protein "closed" pre-fusion trimer (PDB: 6XR8) *and* ***(B) The S protein "one-up" pre-fusion trimer (PDB: 6VSB)*** *with an identical scheme as Fig. 1A-B with N501Y and DCPs A570D, T716I, S982A and D1118H (N501Y on the "one-up" structure and P681H position in both structures were not resolved) Grey labels indicate that the DCP is obscured from view within the 3D structure.* ***(C) Chain A of the "closed" trimer and (D) Chain A of the "one-up" trimer with DCPs*** *has identical labelling to Fig.2C-D*

4.2.2 N501Y – structural analysis

Closed form

(Ref. Fig. 17) In the non-mutated form, N501 stabilises other residues of the RBM (G496, Q498, Y505 and Q506). Y501 with its larger R group clashes heavily with the three residues between G496-Q498 but can perform aromatic stacking with Y505 and is suspected to form a hydrogen bond with the sidechain amine group of Q498 via the hydroxyl group (not visible in Fig. 17 due to clashing).

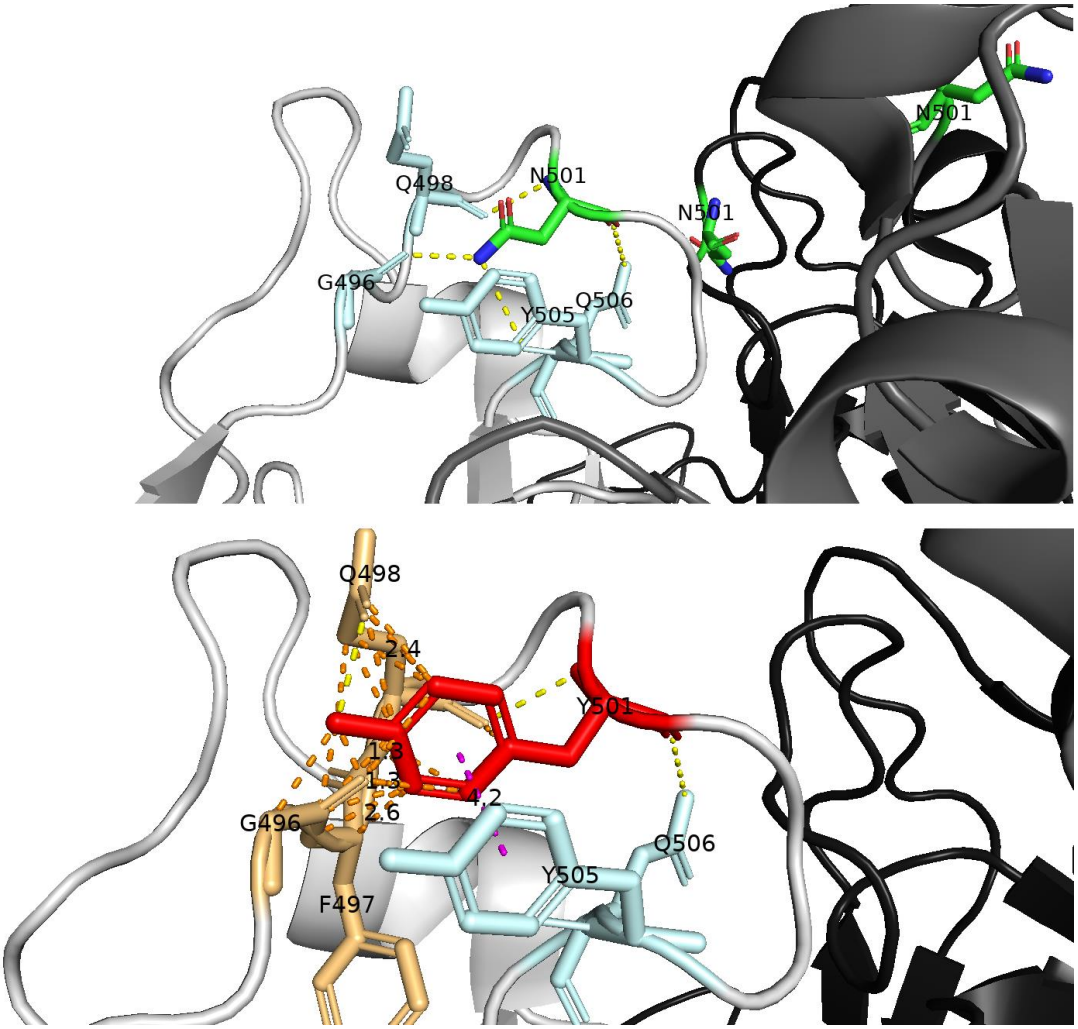


Fig. 17 – Close-up of the Spike N501Y DCP from Chain A of the "closed" prefusion trimer (PDB:6XR8) *The upper image shows pre-mutation (N501 of all chains shown with focus on Chain A) and the bottom image shows a zoom-in of the post-mutation (Y501), only the closest clashes per residue are numerically shown to increase overall clarity. Colour scheme – same as in Figs.3-4 with magenta dashes representing aromatic π- π interactions.*

“One-up” form

The N501Y position is unresolved in 6VSB, in place of this a different “one-up” structure is used whereby the RBD of two S proteins are bound to dimerised ACE2-B0AT (PDB: 6M17 [92]) which has the additional advantage of seeing impacts on receptor-binding of the “one-up” state.

(Ref. Fig. 18) S protein N501 holds a singular contact to intrachain Q506 and two interchain hydrogen bonds to ACE2 (Y41 and K353) via its sidechain. Y501 allows for two intrachain interactions (G496 and Q506) and loses the polar contact to Y41 instead of a π-π interaction. Clashing occurs within (S) between G496-Q498 and (ACE2) Y41. Residues that are directly involved with the RBD-ACE2 fusion that also either clash or share polar bonds with Y501 include (S)G496, (S)Q498, (ACE2)K353 and (ACE2)Y41.

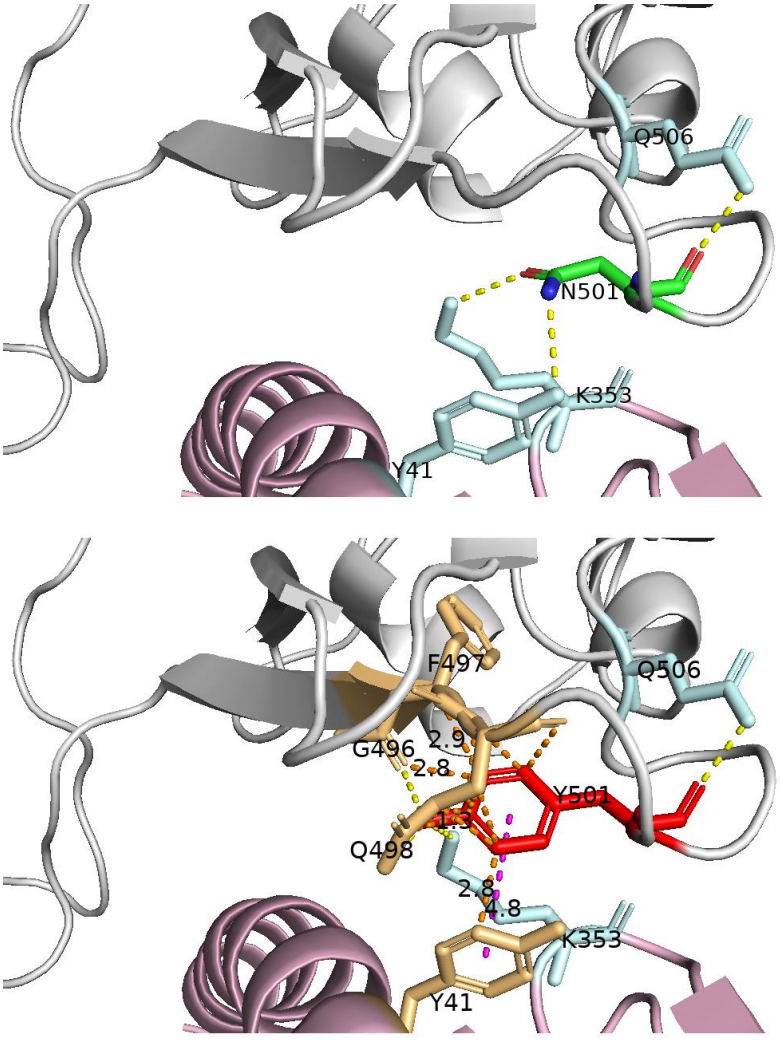


Fig. 18 - Close-up of the Spike N501Y DCP from Chain B of the “one-up” post-fusion trimer (PDB:6M17) *The upper image shows pre-mutation (N501) and the bottom image shows the post-mutation (Y501). Colour scheme – same as in Figs.3-4 except light pink represents ACE2 and magenta dashes represent aromatic π- π interactions.*

Summary of N501Y findings

The closed form could have some minor re-arrangement due to the clashing of the larger Y501 R group and possibly a new aromatic interaction to Y505, but as an unstructured region, the impact is likely discreet. In the Y501 "one-up" form there is one extra intrachain stabilisation (to the carbonyl group of G496) which could aid in rigidifying the ACE2-receptive state, and altered interchain interaction from a hydrogen bond to π-π stacking toward (ACE2)Y41. Y501 via clashing or bonding directly affects two pairs of residues each from the S protein (G496, Q498) and ACE2 (K353, Y41) that partake in the action of S-ACE2 binding; perhaps their re-arrangement could enhance the available binding surfaces. Dynamut2 stability prediction for N501Y was that the "closed" state was not impacted (avg. ΔΔG for Chains A-C = -0.03kcal/mol), however, the "one-up" state was destabilised (6M17: ΔΔG for S protein Chains E-F -0.42, -0.44 kcal/mol) which could indicate that in the "one-up" state N501Y introduces conformational flexibility in the RBD and/or the binding site of ACE2 which may relate to improved site recognition.

4.2.3 A570D – structural analysis

Closed form

(Ref. Fig. 19) The residue change to D570 results in new polar intrachain interactions towards D568/T572, and an additional interchain polar bond to N856 of Chain B (found anticlockwise of Chain A, see Fig. 15) which lies next to the FP2 region (pos.835-855). Intrachain clashing occurs with nearby residues D568 and T572 which sit on the short turn between β47-β48 alongside D570; re-arrangement is likely a minor consequence due to the unstructured nature.

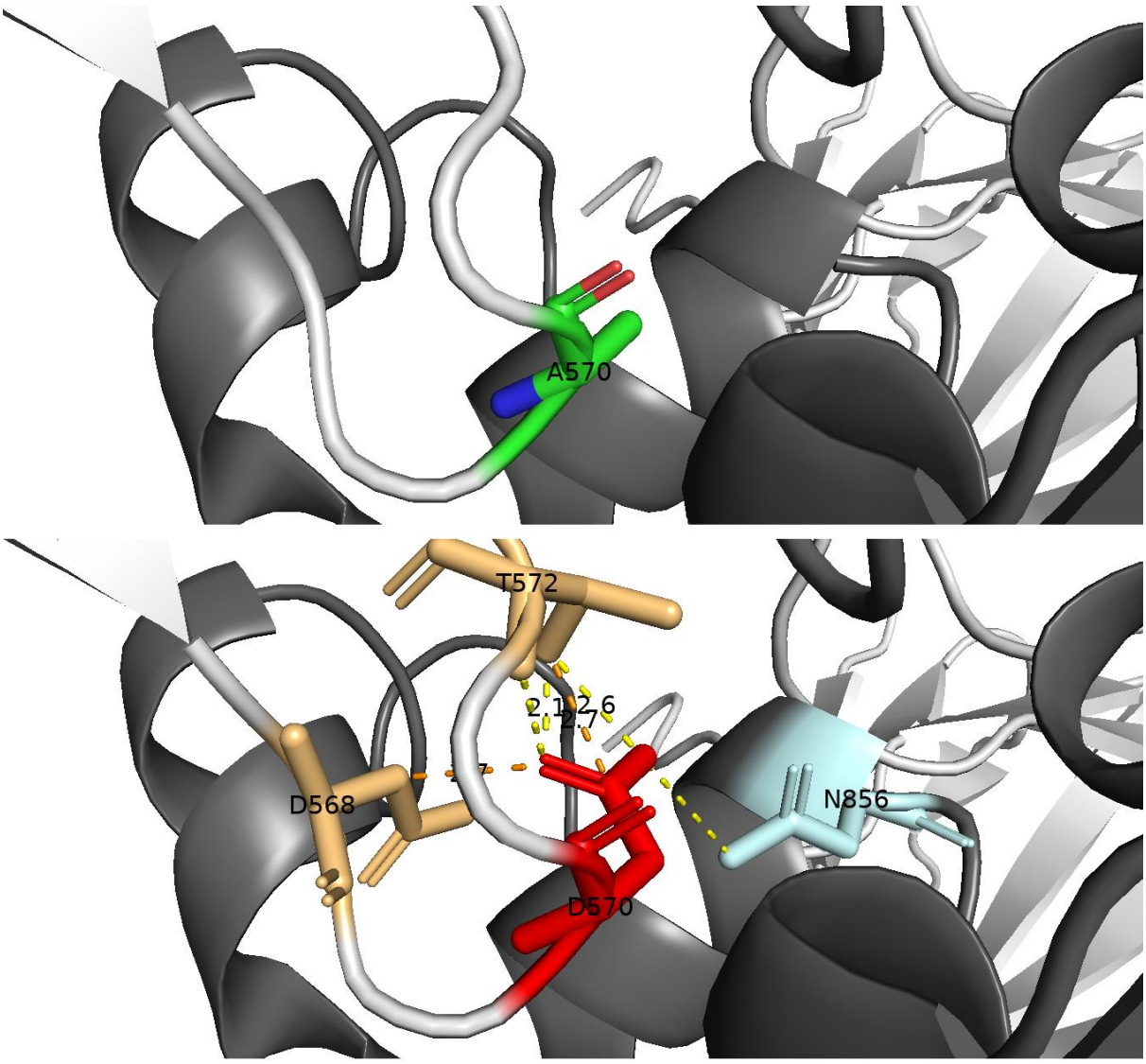


Fig. 19 - Close-up of the Spike A570D DCP from Chain A of the "closed" prefusion trimer (PDB:6XR8) *The upper image shows pre-mutation (A570) and the bottom image shows the post-mutation (D570). Colour scheme – same as in Figs.3-4*

"One-up" form

Chain A: (Ref. Fig. 20) D570 did not introduce new polar contacts, instead the sidechain clashed with neighbouring I569 and the hydrophobic isopropyl group of V963 found on the α23-helix of Chain B which lies in HR1 (pos.920-970) of the coiled-coil region; possibly reducing stabilising hydrophobicity of the region.

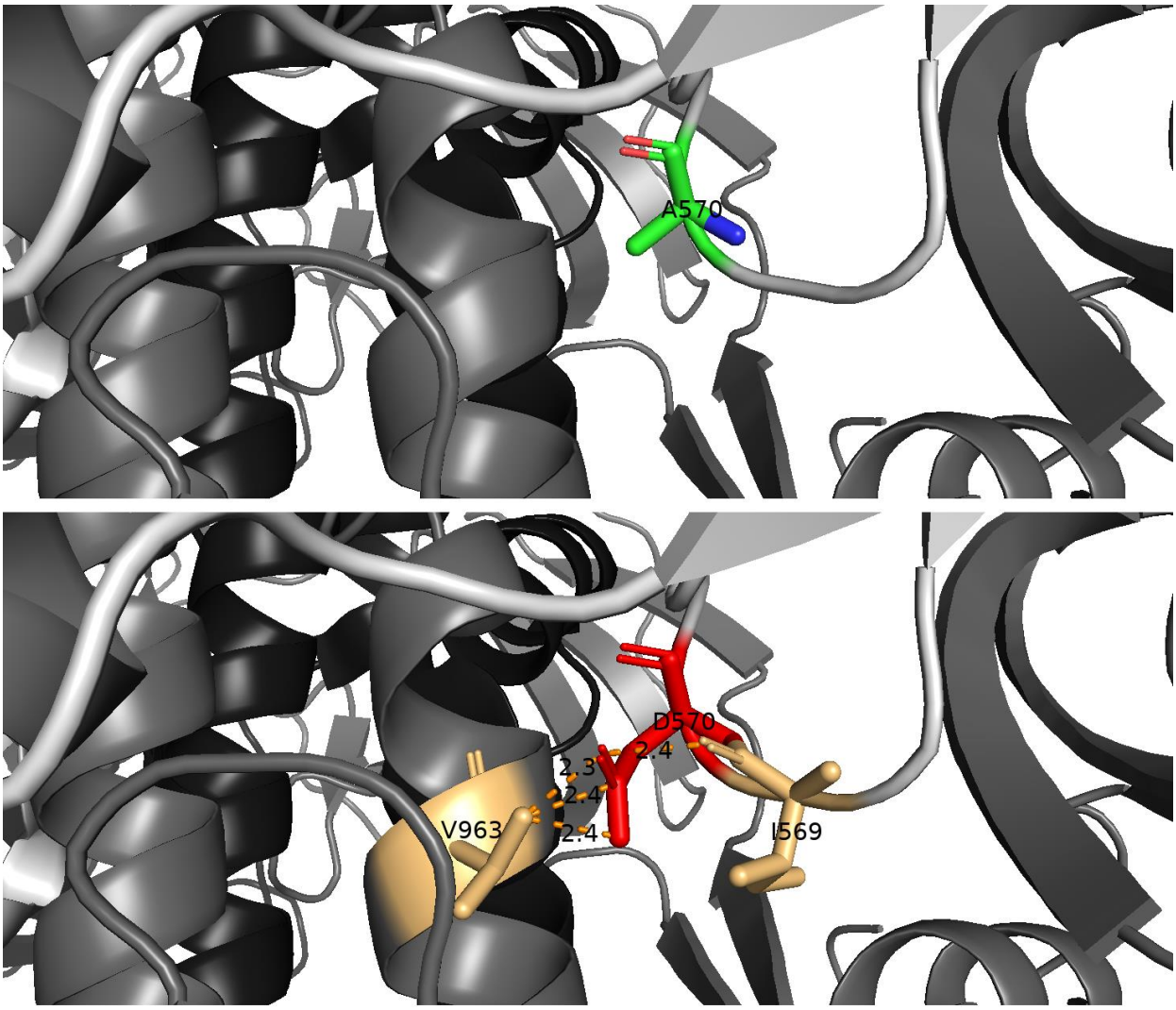


Fig.20 Close-up of the Spike A570D DCP from Chain A of the "one-up" prefusion trimer (PDB:6VSB) *The upper image shows pre-mutation (A570) and the bottom image shows the post-mutation (D570). Colour scheme – same as in Figs.3-4.*

Chain B: (Ref. Fig. 20) D570 interacts with HR1-region residues of α23, with possible polar contacts forming to the backbone amines L966 and S967 and clashing occurring with V963 and L966; if this were to occur the helix would likely be broken although it is possible that the turn D570 sits on would reposition further away from the helix. A570 may allow the helix to form in the wild-type structure by shielding the hydrophobic Val/Leu from the solvent

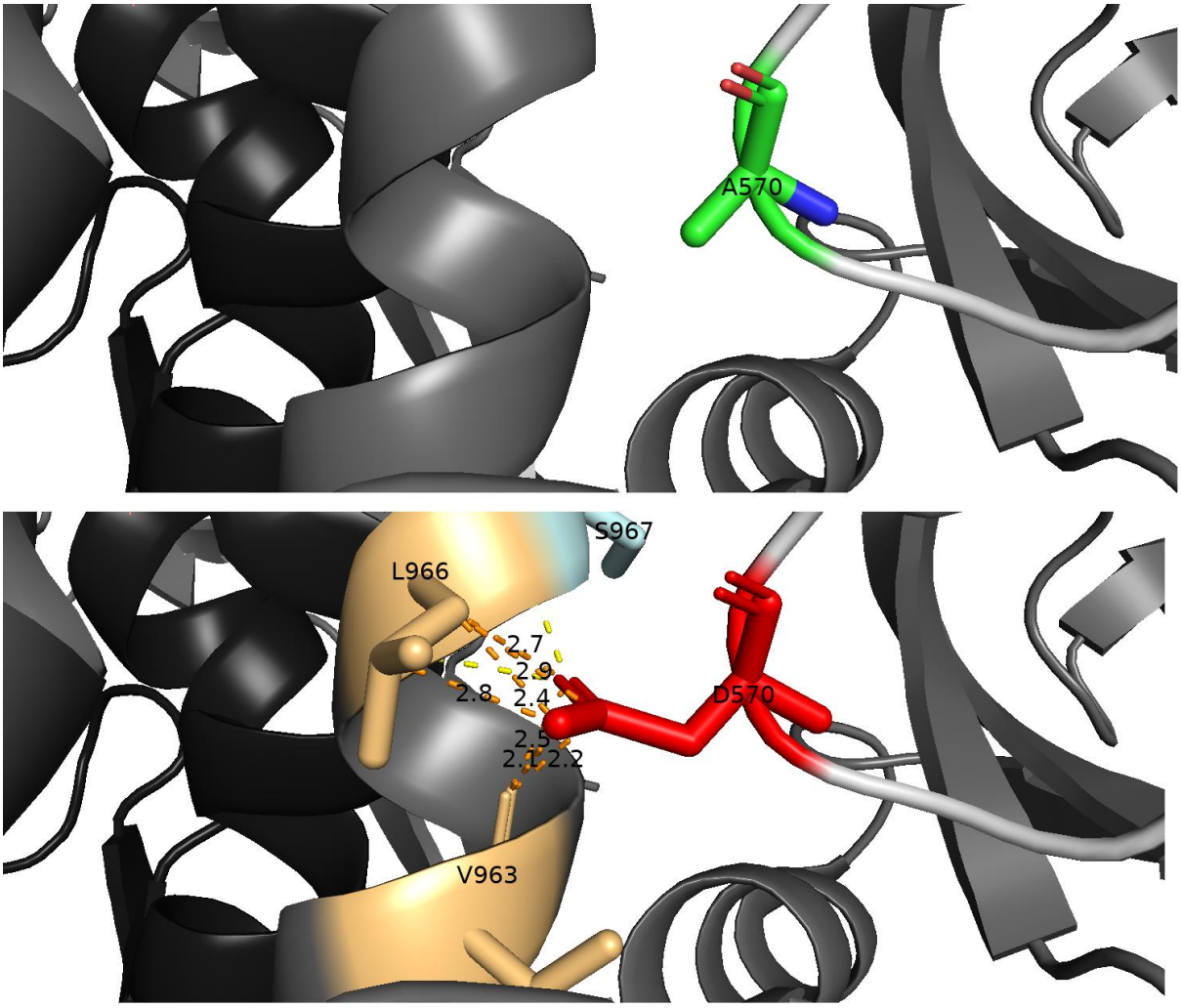


Fig.21 - Close-up of the Spike A570D DCP from Chain B of the "one-up" prefusion trimer (PDB:6VSB) *The upper image shows pre-mutation (A570) and the bottom image shows the post-mutation (D570). Colour scheme – same as in Figs.3-4 however, Chain (A-C) shading follows Fig. 5 pattern.*

Chain C:

(Ref. Fig. 22) similarly to Chains A-B, D570 causes clashing with α23 but instead to the Cα of K964. As D570 is sitting on a flexible loop perhaps a salt bridge could be possible between these two residues, although this could not be shown artificially with rotamers in 6VSB.

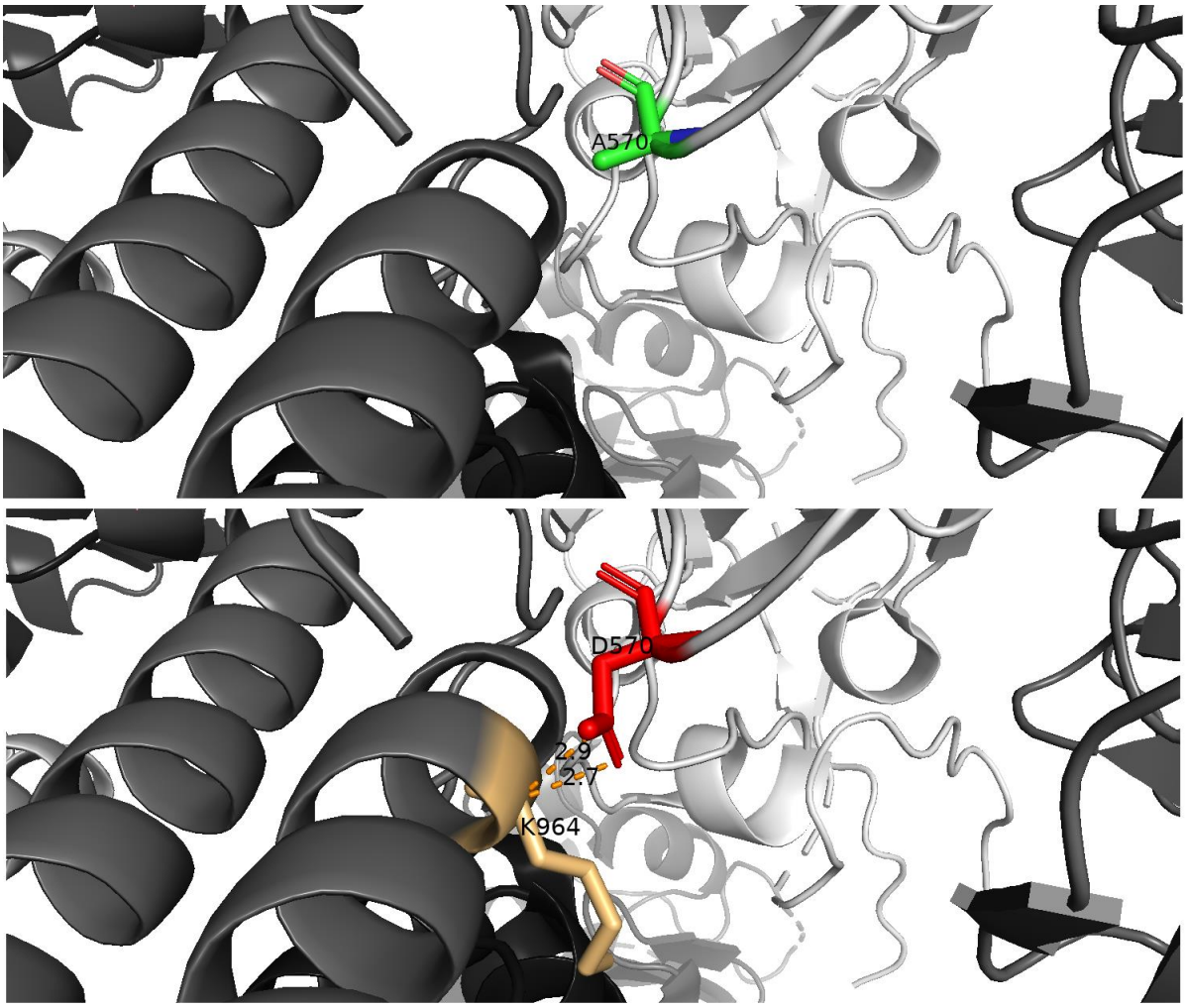


Fig. 22 - Close-up of the Spike A570D DCP from Chain C of the "one-up" prefusion trimer (PDB:6VSB) The upper image shows pre-mutation (A570) and the bottom image shows the post-mutation (D570). Colour scheme – same as in Figs.3-4. however, Chain (A-C) shading follows Fig. 6 pattern.

Summary of findings for A570D

There is an interchain polar interaction from D570-N856 in the closed state, stabilisation of the region (i.e. the CTD of S1 to the terminus of the FP2) by this bond is expected to be outweighed by neighbouring residue clashes (D568 and T572) as Dynamut2 predicts a large average ΔΔG destabilisation of -1.28 kcal/mol for all chains of 6XR8.

It is relatively unlikely that breakage of the α23 helix occurs in the "one-up" state as the D570 loop could simply move away with lower energy demand. D570 may be protomer-destabilising in the "one-up" state - disrupting a hydrophobic pocket between V963, L966 and the hydrophobic portion of K964 (i.e. Cβ/Cγ) but this could be somewhat reduced by an ionic bond to the amine group of K964. For 6VSB, Dynamut2 predicts uneven destabilisation of -0.28 kcal/mol for Chain A, -0.69kcal/mol for Chain B and -0.53kcal/mol for Chain C. Perhaps the less exaggerated destabilisation in the "one-up" state in contrast to the "closed" state could induce a greater likelihood of S1 adapting the "one-up" conformation via protomer destabilisation; this is also supported via Chain A – the "up" monomer, experiencing the least destabilisation of all Chains.

A570D could act as a pivoting region for the RBD due to the relative positioning of A570D relative to the S1 heads (ref. Fig. 16) and could explain the disparity between the ΔΔG values in the "one-up" and "closed" states – which was similarly seen in Dynamut2 stability predictions for another suspected hinge-region - D614G. A570D has been described as part of a "pedal-bin" mechanism allowing RBD displacement thought to be allowed by D570 salt bridges – in addition, D614G laterally flanks the opposite face of α23 and it was been suggested the D570 salt bridges are a form of compensation for the loss of D614G salt bridges [93].

Interestingly, A570D was also mutated spontaneously *in vitro*, see Table 6.

4.2.4 Summary of P681H

P681 is not resolved in any S PDB structure, therefore structural analysis is limited – P681 is assumed to lie on an unstructured region between β56 - β57 facing out towards the solvent to be presented for cleavage by furin; H681 may be advantageous for furin binding due to its basic nature and may have other

binding implications (e.g. Neuropilin-1 can bind ^{682}RRAR at the C-terminal of S1 [94]). In-silico structural prediction of P681H has only suggested minor changes to secondary structure [95].

The first occurrences of P681H alongside N501Y (and all other N501Y-related DCPs plus the HV69-70del and Y144del) was found in samples from Milton Keynes, UK (20/09/2020, EPI_ISL_601443; 21/09/2020, EPI_ISL_581117) which marked the origins of the B.1.1.7 lineage [60], [63]. P681H can be found first in Washington USA (12/03/2020, EPI_ISL_430887) and the second appearance occurred in the UK (23/03/2020, EPI_ISL_423723) with both examples not occurring with any other DCPs mentioned in this study, including the HV69-70del and D614G-related DCPs. Nigerian, Israeli and the US (Hawaii and New York) studies have also found isolated cases of P681H [96]–[99]. Evidence has indicated that although furin-induced cleavage may be increased - this did not translate to increased viral fusion in pseudotyped virions or cell fusion assays and Antibody (Ab) neutralisation of B.1.1.7 and the Israeli P681H strain was also not significantly impacted [97], [100]–[102]. P681H, therefore, has an unknown function but perhaps has a role in host immune response or S1/S2 cleavage - such as adaption to host TMPRSS2 variation - which could also apply to A570D and T716I [95].

4.2.5 T716I – structural analysis

Closed form

(Ref. Fig. 23) The T716 transition to isoleucine results in the loss of a hydrogen bond in exchange for a steric clash to the carbonyl of Q1071. This would suggest destabilisation, however, there could be lessened by hydrophobic association formed between residues I714, K1073 (i.e. Cβ-Cδ) and Y1110 as Dynamut2 suggests effects ranging from neutral to stabilising for 6XR8 (ΔΔG values for chains A-C = -0.04, 0.38, 0.42 kcal/mol, respectively). A notable feature includes the N-glycan chain on the neighbouring N717.

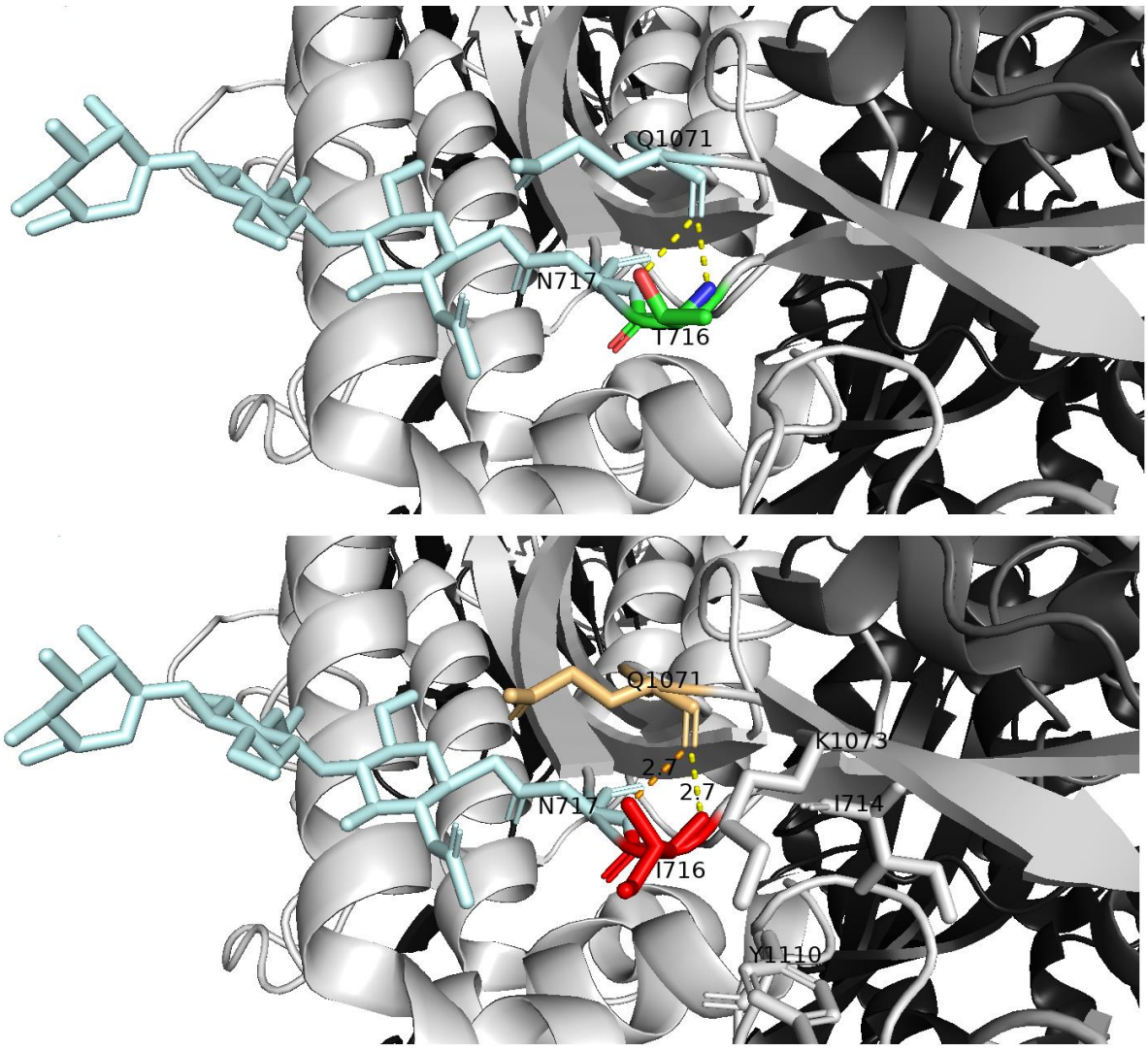


Fig. 23 - Close-up of the Spike T716I DCP from Chain A of the "closed" prefusion trimer (PDB:6XR8) *The upper image shows pre-mutation (T716) and the bottom image shows the post-mutation (I716). Colour scheme – same as in Figs.3-4.*

"One-up" form

Chains A-C:

(Ref. Fig. 24) there is little difference between the "one-up" vs. closed states with the loss of the hydrogen bond to Q1071 and introduced clashing. The β-strand that T716I sits on in 6VSB (pos.711-728) is continuous whereas it is broken in 6XR8 (pos. 711-715 and 718-728) which resembles Uniprot secondary structure labelling which suggests that T716I exists within β63 pos.709-716 and flanked by β64 pos.718-

728. In the “one-up” state, there is little difference between Chains A-C in the case of T716I, there was only detection of an extra hydrogen bond of T716 from the sidechain hydroxyl group to the backbone amine of the N717 in Chain B; in all other respects, the Chains in the “one-up” state are identical to Chain A in Fig. 23. Dynamut2 predicted similar stabilising effects for each Chain in the “one-up” state (ΔΔG values for Chains A-C = 0.54, 0.45, 0.55 kcal/mol, respectively).

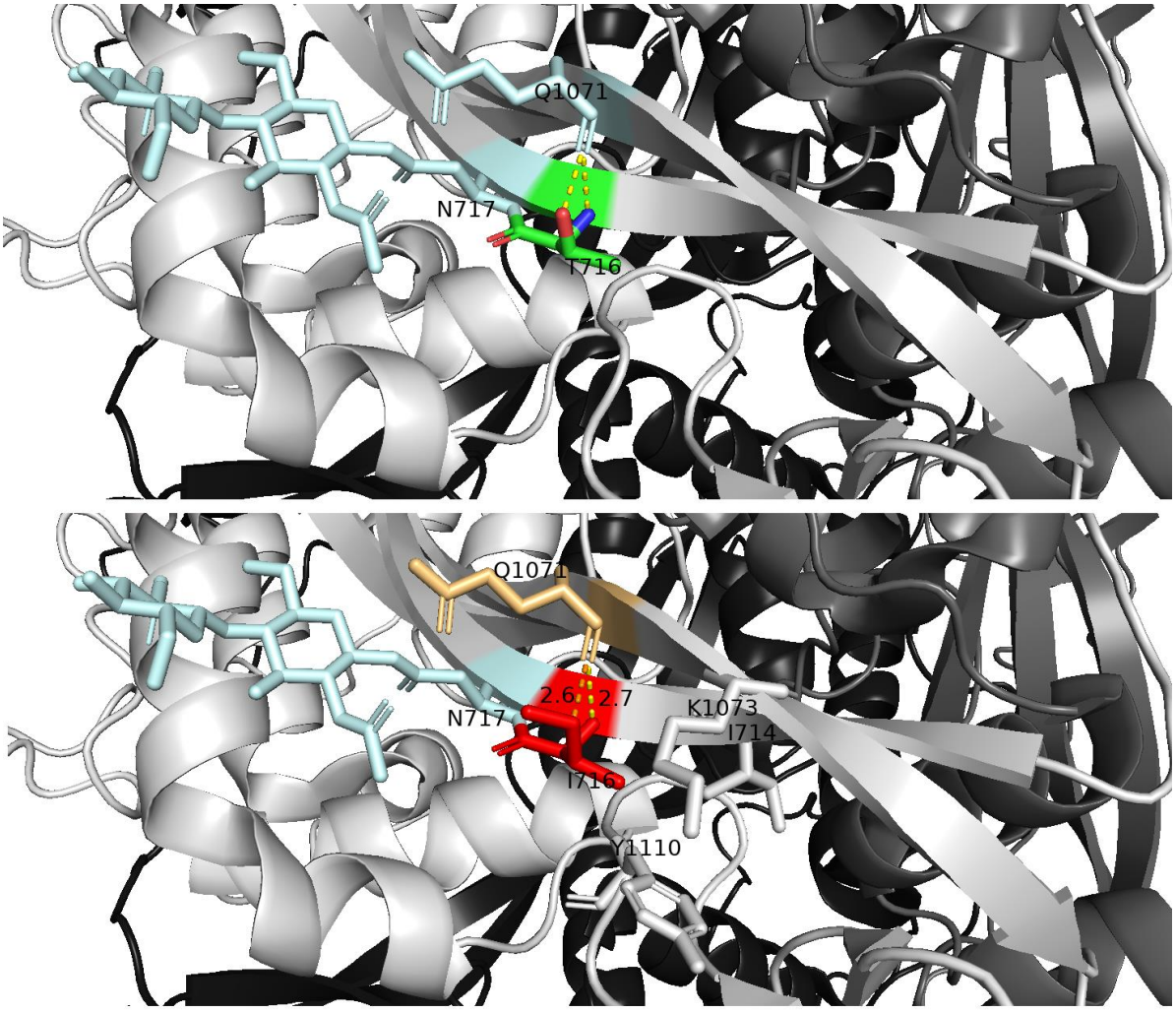


Fig. 24 Close-up of the Spike T716I DCP from Chain A of the “one-up” prefusion trimer (PDB:6VSB) *The upper image shows pre-mutation (T716) and the bottom image shows the post-mutation (I716). Colour scheme – same as in Figs.3-4 with the grey labelled sticks in the post-mutation image refers to residues that are suspected to interact hydrophobically with I716*

Summary of findings for T716I:

There were only a few differences between, the "one-up" vs. "closed" states – including that β59-60 may only remain continuous in the "one-up" form; other computational evidence has alluded to the loss of these β-sheets due to T716I [95]. Chains A-C in the "one-up" state were highly similar in appearance and each predicted ~0.5kcal/mol increase in stability - strangely, although the Chains A-C in the "closed" state appears identical, Chain A was given a neutral ΔΔG value by Dynamut2 (avg. ΔΔG=0.28kcal/mol); it appears that T716I increases stability in both states with a slight advantage to the "one-up" state. The similarity between "one-up" and "closed" states could be the result of T716I being in an area of S2 not involved with ACE2 binding and remains stationary during the up/down state shift of an S1 monomer. T716I could instead stabilise the S2 stalk hinge region whose usual function is to allow greater rotational freedom of the S1 head and is noted to be shielded by glycan chains [103]. T716I lies near a proposed "hip joint" of the stalk and is positioned next to an N-linked GlcNAc (N717) which blends seamlessly to the surrounding glycosylation of the "hip joint" area. T716I mutation was shown to decrease infectivity by 7-fold in pseudotyped virions so it likely acts in concordance with other B.1.1.7 mutations [95], [102].

4.2.6 S982A – structural analysis

"Closed" form:

(Ref. Fig. 25) S982A results in loss of contacts towards the backbone carbonyl groups of N978 and D979 via the serine hydroxyl sidechain within the α24-helix (pos. 976-982). Dynamut2 predicted an average ΔΔG = -0.58 kcal/mol, suggesting that the smaller hydrophobic A982 will equally increase destabilisation of this area, likely due to helix breakage. There may be some association with other hydrophobic residues of the adjacent anticlockwise chain (L390, L518 and Cγ of T547).

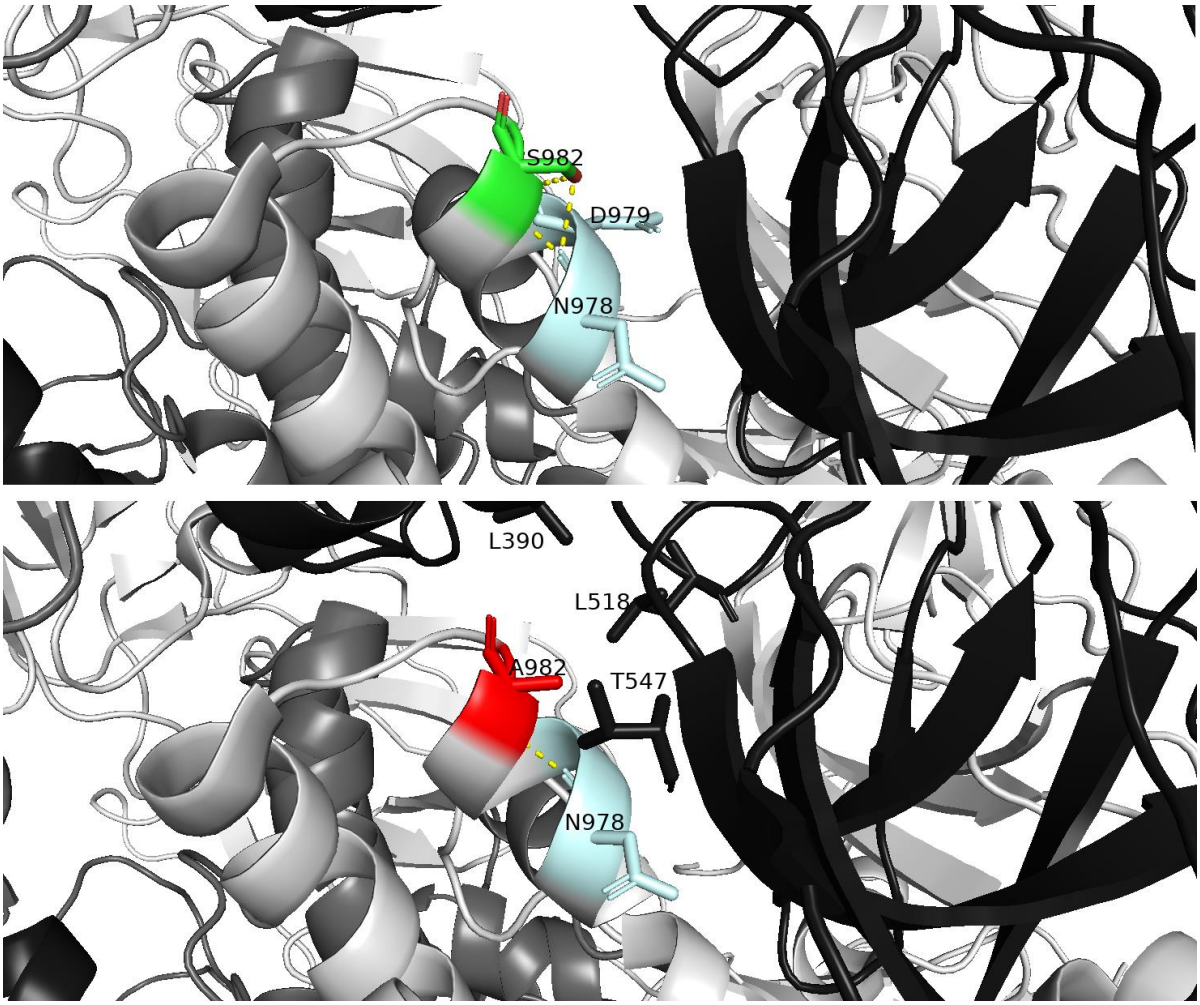


Fig. 25 - Close-up of the Spike S982A DCP from Chain A of the "closed" prefusion trimer (PDB:6XR8) *The upper image shows pre-mutation (S982) and the bottom image shows the post-mutation (A982). Colour scheme – same as in Figs.3-4.*

"One-up" conformation:

Chain A: (Ref. Fig. 26) S982A does not appear to have a different structural implication to the "closed" trimer form and appears similar to the other "one-up" conformations, however, Dynamut2 predicted variance in stability across the chains for 6VSB (ΔΔG values for chains A-C = -0.15, 0.00, -0.37 kcal/mol, respectively).

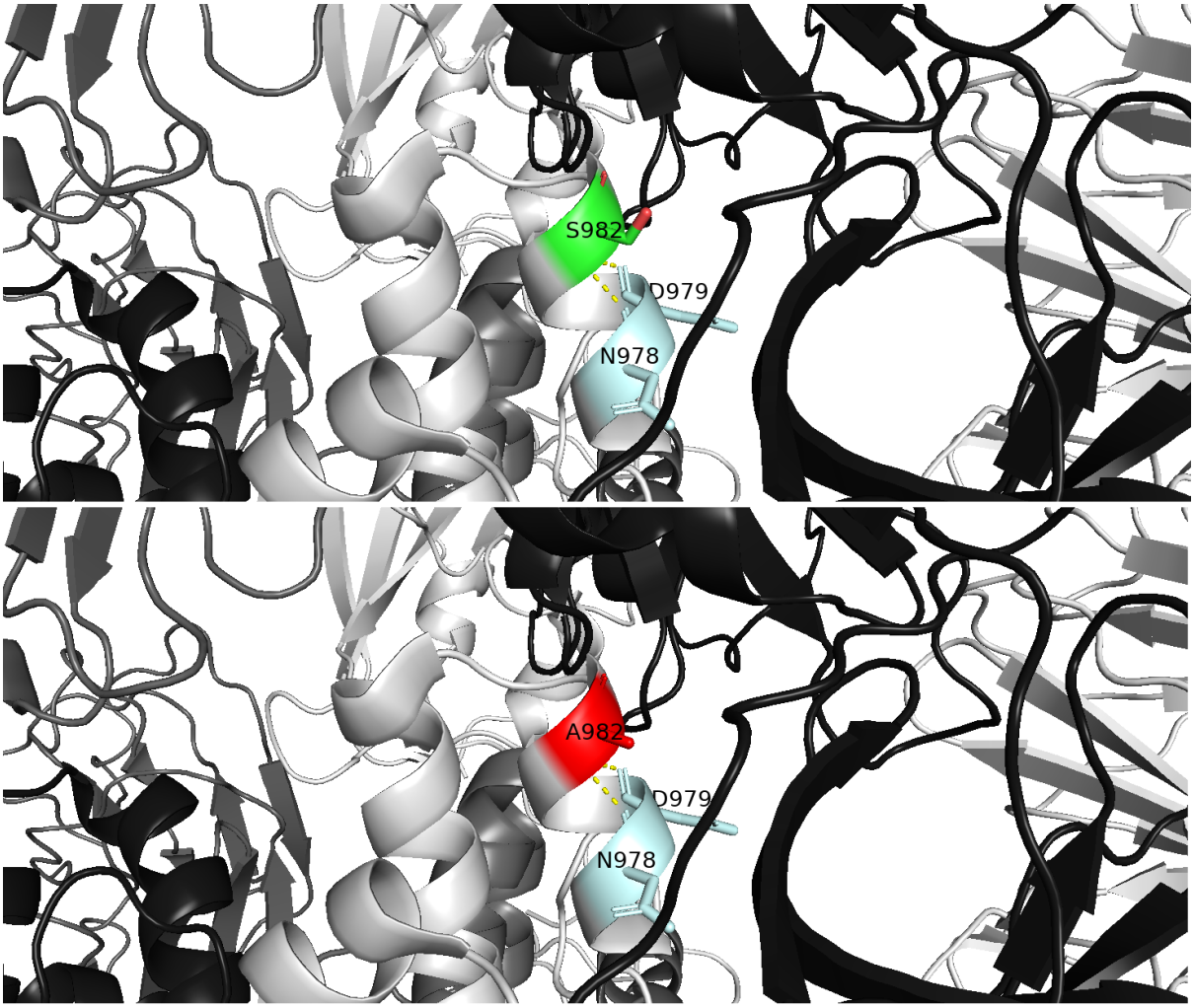


Fig. 26 - Close-up of the Spike S982A DCP from Chain A of the "one-up" prefusion trimer (PDB:6VSB) *The upper image shows pre-mutation (S982) and the bottom image shows the post-mutation (A982). Colour scheme – same as in Figs.3-4.*

Chain B: (Ref. Fig. 27) S982A has a similar outcome to Chain A and even to the "closed" format, however it is noticeable that the DCP position is relatively close to Chain A's RBD linker region (particularly pos. 383-390 – which corresponds to α6) which may allude to a role in stabilisation of the RBD.

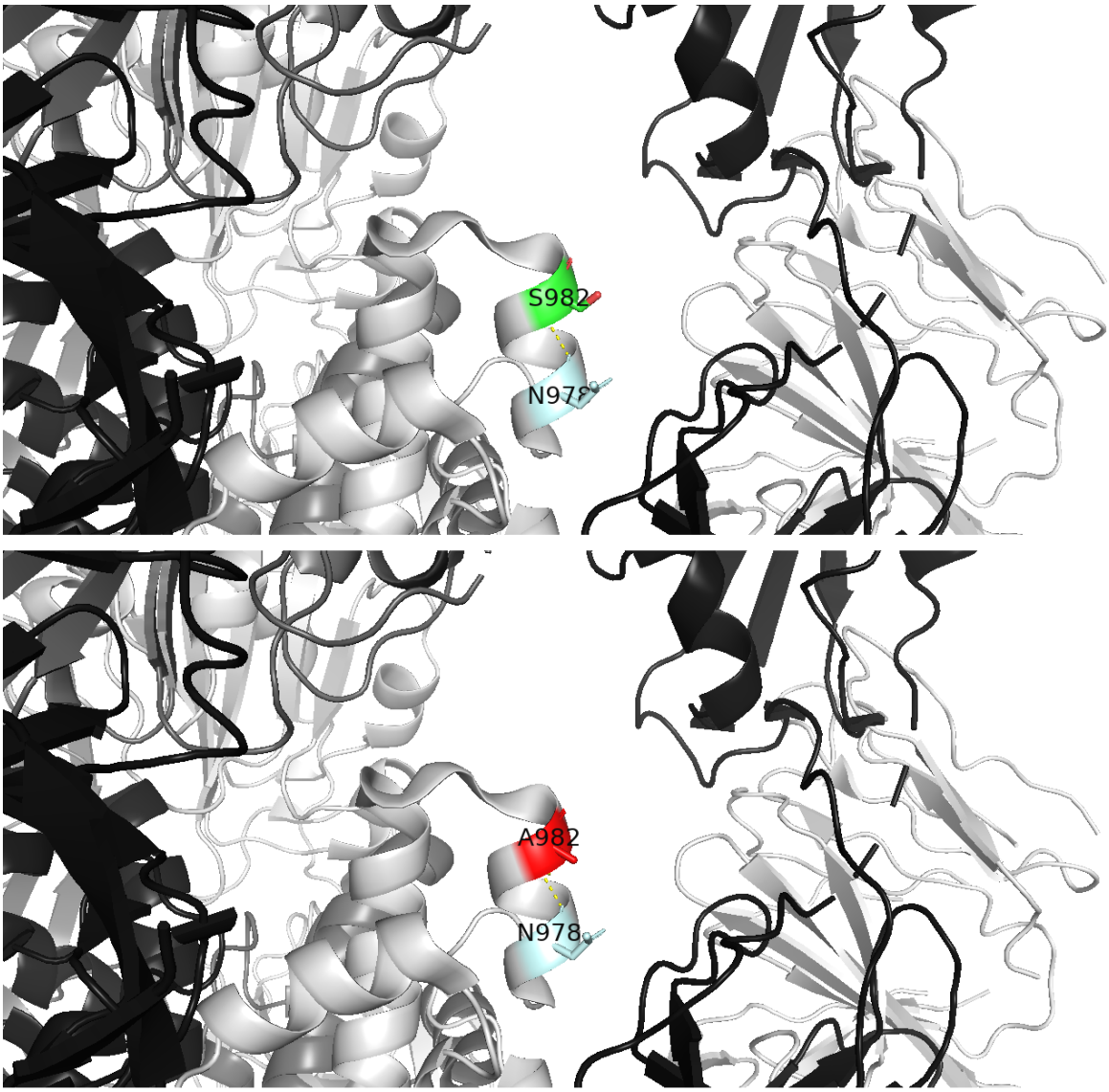


Fig. 27 - Close-up of the Spike S982A DCP from Chain B of the "one-up" prefusion trimer (PDB:6VSB) *The upper image shows pre-mutation (S982) and the bottom image shows the post-mutation (A982). Colour scheme – same as in Figs3-4. however, Chain (A-C) shading follows Fig. 5 pattern.*

Chain C: (Ref. Fig. 28) S982A in Chain C is the only chain that showed a confirmed bond to the RBD linker region via K386 (of Chain B). S982A is suggested to contact the sidechain amine group of K386 via its carbonyl backbone which is possible in both DCP forms. Perhaps the S982-OH is slightly more stabilising for K386 compared to the A982-CH_3 – additionally, A982 could be hydrophobically drawn toward the aliphatic portion of T547 on Chain B.

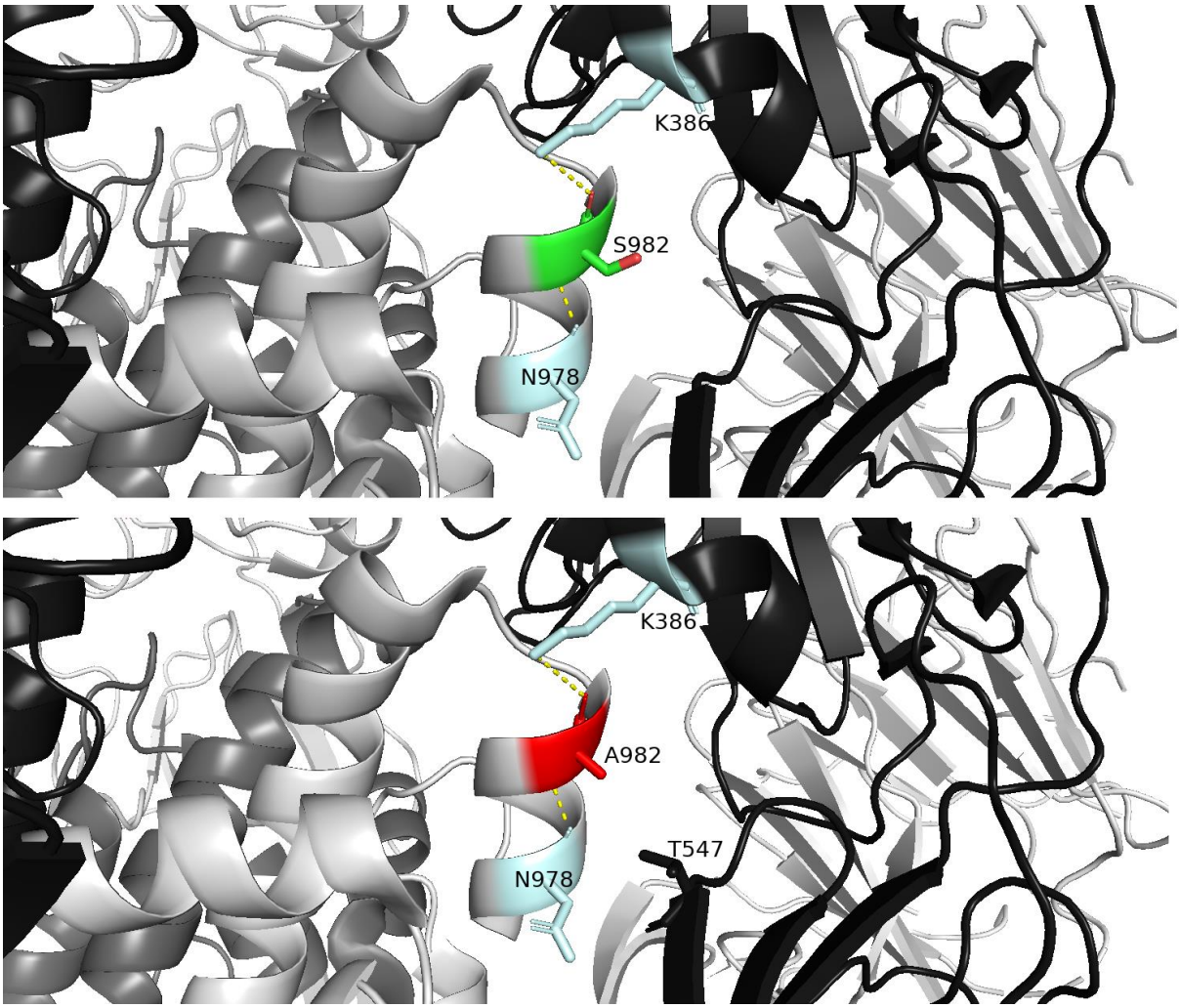


Fig. 28 - Close-up of the Spike S982A DCP from Chain C of the "one-up" prefusion trimer (PDB:6VSB) *The upper image shows pre-mutation (S982) and the bottom image shows the post-mutation (A982). Colour scheme – same as in Fig. 3-4. however, Chain (A-C) shading follows Fig. 5 pattern.*

Summary of findings for S982A:

Initially, S982A seemed to be limited to maintaining the small α-helix which forms part of the first coiled-coil region, with alanine probably reducing the rigidity of the region and breaking the helix prematurely. Other than an unknown effect on membrane fusion based upon the location of S982 in the coiled-coil, it is possible that during the "one-up" state that pos. 982 of normally supports the inactive "down" position of the RBDs, which would explain why the average stability of the "closed" trimers

(6XR8 ΔΔG Chains A-C= -0.56, -0.57, -0.61 kcal/mol) was more unstable compared to the "one-up" state (6VSB ΔΔG Chains A-C= -0.15, -0.27, 0.0 kcal/mol); perhaps the mutation aids in having an S1 unit extended at any one time or reduces the tethering of the RBD in general which could be a result of losing contacts to polar residues such as K382/T547 in exchange for hydrophobic contacts (e.g. L390, L518 and Cγ of T547). S982A shares a similar disparity in ΔΔG values that were present for D614G and A570D which suggests that S982A along with the latter two, possibly have roles in protomer instability - suggested elsewhere to impact a range of viral functions including S1/S2 cleavage, structural dynamics and cell fusion [104].

4.2.7 D1118H – structural analysis

"Closed" form:

(Ref. Fig. 29) The D1118H DCP is in a thin neck region which can be viewed simultaneously in the trimers, each mutation was calculated to have similar destabilisation values of (avg. ΔΔG =) -0.36 kcal/mol. The introduction of H1118 on Chains A-C resulted in a triangular inwards-projecting sidechain display with an average distance of 5.2Å between the closest atoms of the histidine rings which clash with each P1140; intrachain backbone polar interactions to R1091 and T1116 appear unchanged with the mutation.

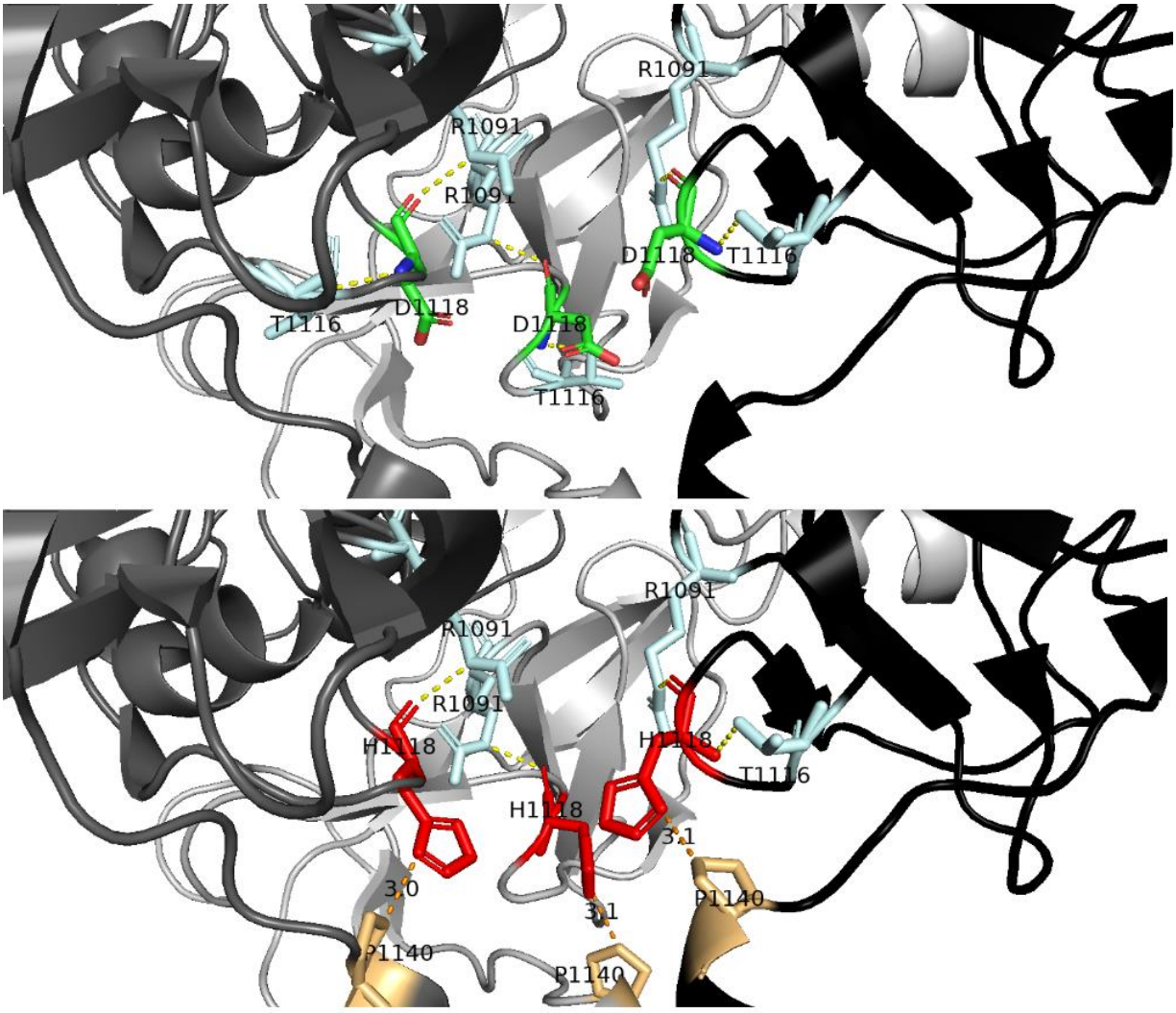


Fig. 29 - Close-up of the Spike D1118H DCP from Chains A-C of the "closed" prefusion trimer (PDB:6XR8) *The upper image shows pre-mutation (D1118) and the bottom image shows the post-mutation (H1118). Colour scheme – same as in Figs.3-4.*

"One-up" form:

(Ref. Fig. 30). The D1118 backbone is associated with R1091 for Chain B only and with T1116 in Chains A-B via the aspartate sidechain. The "one-up" triangular histidine display had a lower average distance of 4.5Å between the closest atoms of the rings compared to the "closed" and loses the trimeric symmetry on Chain B when placed in the lowest strain rotamer. Dynamut2 prediction was identical to the "closed" state (avg. $\Delta\Delta G$ =) -0.36 kcal/mol.

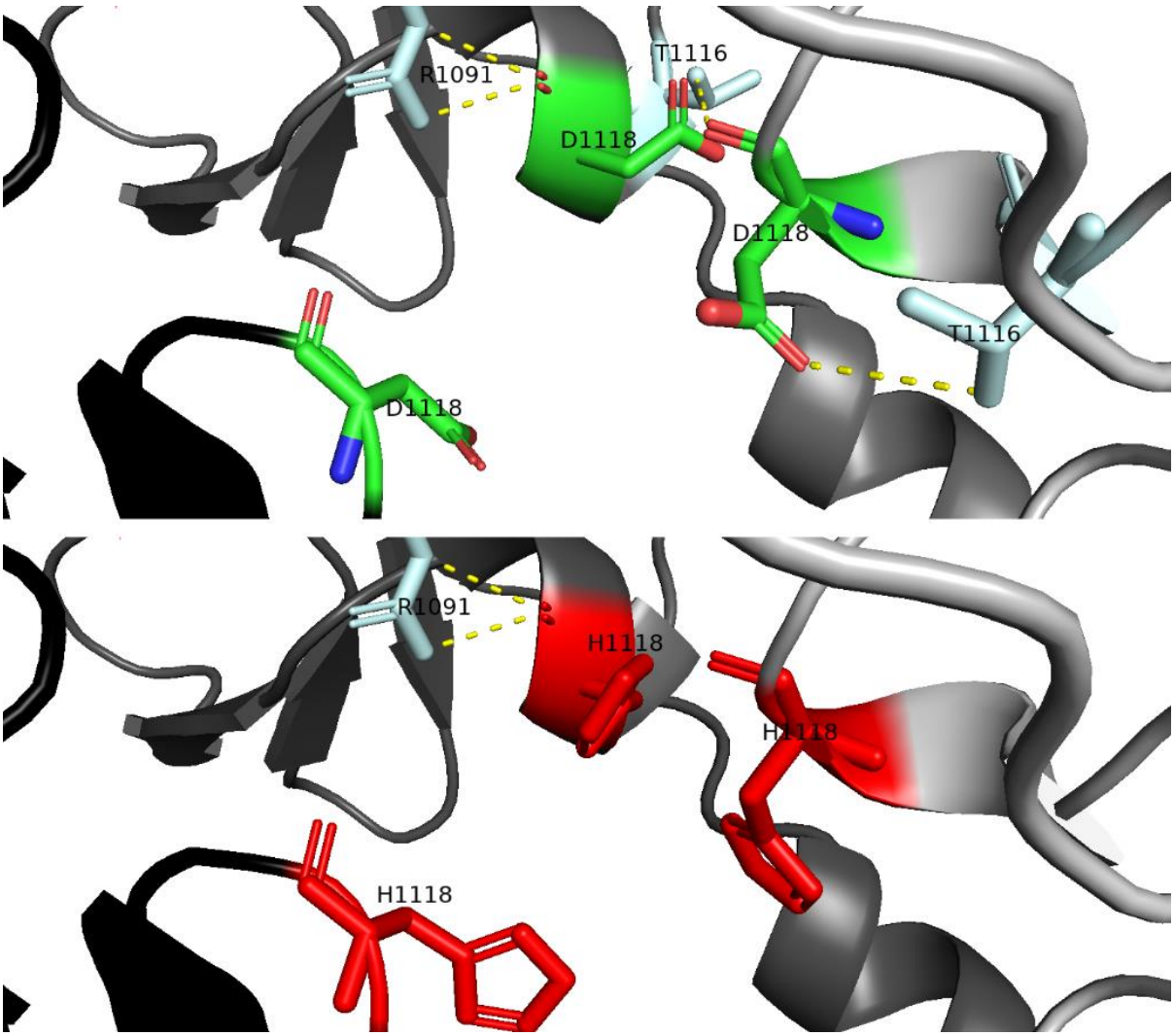


Fig. 30 - Close-up of the Spike D1118H DCP from Chains A-C of the "one-up" prefusion trimer (PDB:6VSB) *The upper image shows pre-mutation (D1118) and the bottom image shows the post-mutation (H1118). Colour scheme – same as in Figs.3-4.*

Summary for D1118H:

D1118H is found facing inwards, upstream of the HR2 region within a hollow area between the intersecting S2 regions (see Fig. 16A). The introduction of the D1118H mutation only affected the hydrogen bond to T1116 and caused clashing only in the "closed" state with P1140. Between the "one-up" and "closed" states there was little difference in contacts made and are supposedly equally destabilising (-0.36kcal/mol), however, there is a 0.7Å decrease in distance between each H1118 in the

transition to the “one-up” state that is aided with H1118 of Chain B breaking symmetry and facing upright - implying that the S1 extension causes compression in this area of the S2 stalk; in addition, the bulky histidine residues when viewed with the molecular surface shows a reduction in the size of the hollow area. D1118H could be in a load-bearing position under tension from elongation or twisting of the stalk, the area between HR1-HR2 is largely populated by β-sheets and disordered regions, D1118H has elsewhere been predicted to cause breakage of β-sheets, perhaps the destabilising histidine allows for alternation of spring-like or rotational motion – which could also relate to the close-by “hip” joint earlier mentioned for the T716I DCP [95], [103].

4.2.8 N501Y-associated deletions HV69-70del & Y144del

If the GISAID dataset is separated based upon the presence/absence of ^{69}HV in the same method as N501Y, a single DCP within the RBD is uncovered - N439K (K439 occurs in 0.63% in ^{69}HV-positive samples and 59.5% in ^{69}HV-negative samples). N439K is developed in lineages B.1.141 and B.1.258, N439K may enhance ACE2 binding affinity via a salt bridge to (ACE2)E329 and reduce Ab-recognition [105]. Co-occurrence between HV69-70del and N439K has previously been noted and is evident in B.1.258 where ~69% of the lineage GISAID entries carry N439K and HV69-70del simultaneously [71], [106].

DCP analysis for Y144del results in a return of the six N501Y-associated DCP positions with conservations >80% (i.e. N501Y, A570D, P681H, S982A and D1118H) as well as HV69-70del. This result suggests that Y144del was more strongly associated with the B.1.1.7 lineage whereas HV69-70del tends to occur spontaneously in several lineages (e.g. appears in the B.1.258 and mink lineages [107]) and the two deletion events may have a combined effect against NTD-directed Antibody recognition but Y144del is likely to be more powerful, e.g. indirect immunofluorescence using mAb 4A8 indicated that HV69-70del plus Y144del was sufficient to block mAb binding and in cryo-EM/ELISA NTD antigenic

mapping Y144del was able to block several mAbs binding (S2M28, S2X28, S2X333 and 4A8) whereas HV69-70del did not demonstrate mAb-blocking [84], [108]. In addition, computational prediction concluded that HV69-70del would not alter secondary structure whereas Y144del would alter a β-sheet to a coiled-coil [95].

4.3 Cell culture

To look at areas of SARS-CoV-2 that are prone to mutation during *in vitro* conditions, two cultivated colonies of CaCo-2 (colorectal cancer) cells were infected with SARS-CoV-2, the two SARS-CoV isolates (*FFM3 and FFM7*) were sequenced and then analysed for mutations at three timepoints: t=0 (mutations that were present initially that differ from the Wuhan reference sequence), t=1 (mutations that developed mid-way through the evolution after 30 passages) and t=2 (mutations present at 60 passages). Nucleotide base change in the two strains was detected if the called IUPAC code was altered between datasets, which is set at an 80% baseline. The mutations detected in *FFM3* and *FFM7* across t=0, t=1 and t=2 is discussed in the next sections and detailed per time stage in Tables 4-6. A graphical summary of per-isolate mutations in *FFM3* and *FFM7* across the whole experiment can be seen in the supplementary figures section (Figs. 41-42).

4.3.1 t=0

FFM3 and *FFM7* each had six single-base transition substitutions, ¾ of which were C=>T mutations and ⅓ were silent - all *FFM3* and *FFM7* mutations and their frequencies for t=0 can be found in Table 4.

FFM3 and *FFM7* shared several mutations that differ from the Wuhan reference sequence pre-cultivation which are found at pos. 241, 3,037, 14,408 and 23,403nt. A=>G at 23,403nt results in D614G and the former three refer to C=>T mutations which result in: a mutation within the 5'-UTR, a silent mutation in NSP3 and (NSP12)P323L, respectively – these mutations have previously been associated with D614G (G clade) [50]. pos. 241nt is positioned in the 5'-UTR has been calculated to be a binding site for TAR DNA binding protein 43 which regulates transcription, RNA stability and splicing [109]. Pos. 3,037nt is silent and pos. 14,408nt results in (NSP12)P323L which is almost completely conserved in G clade. Unique mutations for *FFM3* include pos. 28,882nt (G=>A) which translates to (ORF9c)G50E and pos. 28,883 (G=>C) which results in (N)G204R; these two mutations have previously been identified to co-occur with the pos. 241, 3,037, 14,408 and 23,403 mutations in the timeframe between Dec 2019 – May 2020 and is correlated to Ab-binding regions of SARS-CoV N protein [110], [111]. *FFM7* has two unique starting mutations: pos. 15,324 and 23,179 which are both synonymous (C=>T) occurring in NSP12 and S protein, the former mutation was universally present and retained at >99% of the base reads whereas the latter began with only 43.3% at t=0 before rising to 99.4% of reads by t=2.

All these starting mutations at the beginning of the experiment were retained up to t=2 (>99% population of base readings per position).

t	Strain	Pos (nt)	Mut type	Postmut nt.% (t=0, t=1, t=2)	Δ amino acid	Notes:
t=0	**FFM3**	241	C=>T	**99.4**, 99.7, 99.4	(5'-UTR)	
		3,037	C=>T	**99.7**, 99.8, 99.3	(pp1a)F924 (nsp3)F106	Silent
		14,408	C=>T	**98.5**, 91.8, 31.3	(pp1ab)P4714=>L4714) (nsp12)P323=>L323	
		23,403	A=>G	**99.7**, 99.7, 99.5	(S)D614=>G614	D614G mutation
		28,882	G=>A	**99.3**, 99.6, 99.1	(N)R203 (ORF9c)G50=>E50	Silent in N only
		28,883	G=>C	**99.6**, 99.8, 99.8	(N)G204=>R204 (ORF9c)G50	Silent in ORF9c only
	FFM7	241	C=>T	**99.7**, 99.6, 99.5	(5'-UTR)	
		3,037	C=>T	**99.6**, 99.6, 99.4	(pp1a)F924 (nsp3)F106	Silent
		14,408	C=>T	**99.5**, 98.7, 99.5	(pp1ab)P4714=>L4714 (nsp12)P323=>L323	
		15,324	C=>T	**99.7**, 99.7, 99.8	(pp1ab)N5019 (nsp12)N628	Silent
		23,179	C=>T	**43.3**, 87.2, 99.4	(S)V539	Silent
		23,403	A=>G	**99.7**, 99.7, 99.7	(S)D614=>G614	D614G mutation

Table 4 - Identification of nt. base changes in FFM3 & FFM7 before cultivation (t=0) relative to the Wuhan reference sequence. *Postmut nt.% column describes the percentage of the base reads that match the postmutation state in the three timeframes separated by commas - before cultivation t=0 (in bold), midway cultivation t=1 and end of cultivation t=2. E.g. for FFM3 at pos.241 the percentage of T base reads 99.4% for t=0, 99.7% for t=1 and 99.4% for t=2*

4.3.2 t=1

Similarly to t=0, the most common mutation type detected at the midway stage are C=>T point mutations (Ref. Table 5 for list of t=1 mutations). Three were silent mutations (one of which is in the 5-UTR), seven were non-silent point mutations and includes an A=>T transversion and *FFM3* has an indication of a 3nt deletion in NSP1 (pos. 516-518.; M85del), at t=0, M85del was only present in 6.9% of base reads before rising to 20.5%. M85del has been found at low levels in global sequencing with the greatest concentration in the UK and is currently averaging 0.45% of GISAID entries. [112], [113].

FFM3 and *FFM7* share a single C=>T mutation at pos. 21,789 which results in T76I in the NTD of S protein which developed more rapidly between t=0 to t=1 and was retained well in both strains to t=2 (99.9% and 69.3% respectively*)*. T76I is comparatively uncommon but maintains a presence in all continents (0.19% of GISAID sequences) [114]. Interestingly this mutation is first present in the Wuhan bat CoV RaTG13 sample EPI_ISL_402131, unfortunately (S)T76 is not visible in either 6XR8 or 6VSB but is available in 6ZXN which was used to view the L18F DCP in the "one-up" S conformation. (S)T76I may cause minor destabilisation in 6ZXN (Dynamut2 chains A-C ΔΔG values: -0.2, -0.04, -0.44 in kcal/mol), with chain C possibly being destabilised by T76I by the loss of polar interaction to the backbone amine of K77 in addition to instability caused by displaying a hydrophobic residue to the solvent face – both factors could indicate increased flexibility (see Fig. 31).

FFM3 has one silent C=>T mutation in M protein at pos. 27,131 which unlike most of the other mutations was slightly selected against by the time of t=2 sampling. (N)N126Y, caused by an A=>T transversion in the RNA-binding region of N (pos. 28,649) was selected more rapidly than all other mutations detected at this stage and retained almost up to 100% of base reads – surprisingly, no literature has detected this mutation in the global population and GISAID only contains 16 sequences with this mutation as of September 2021 with the earliest from the Philippines (EPI_ISL_2155244, EPI_ISL_2155347 both collected in November 2020); this may suggest that this mutation that it is poorly selected for *in vivo.*

The relative importance of (N)N126Y is unknown as it is conserved in SARS-CoV but not in other CoV species such as MERS or HCoV-OC43 [115]. DynaMut2 predicts an average destabilisation of -0.34 kcal/mol for N126Y in an NTD structure of N (6VYO) but also shows no loss/gain of interactions as the sidechain faces out into the solvent nor does Y126 appear to cause clashing or significant strain (calculated by Pymol as 2.09). Asn=>Tyr results in negligible surface charge alteration at native conditions within N and the actual location of the RNA-binding region lies outside of the main RNA-binding sites that correlate with grouped basic residues (see Fig. 32) [115]. (N)N126Y is within an

antigenic region mapped in SARS-CoV N so may instead be involved in adaptive immunity response or protein binding [111].

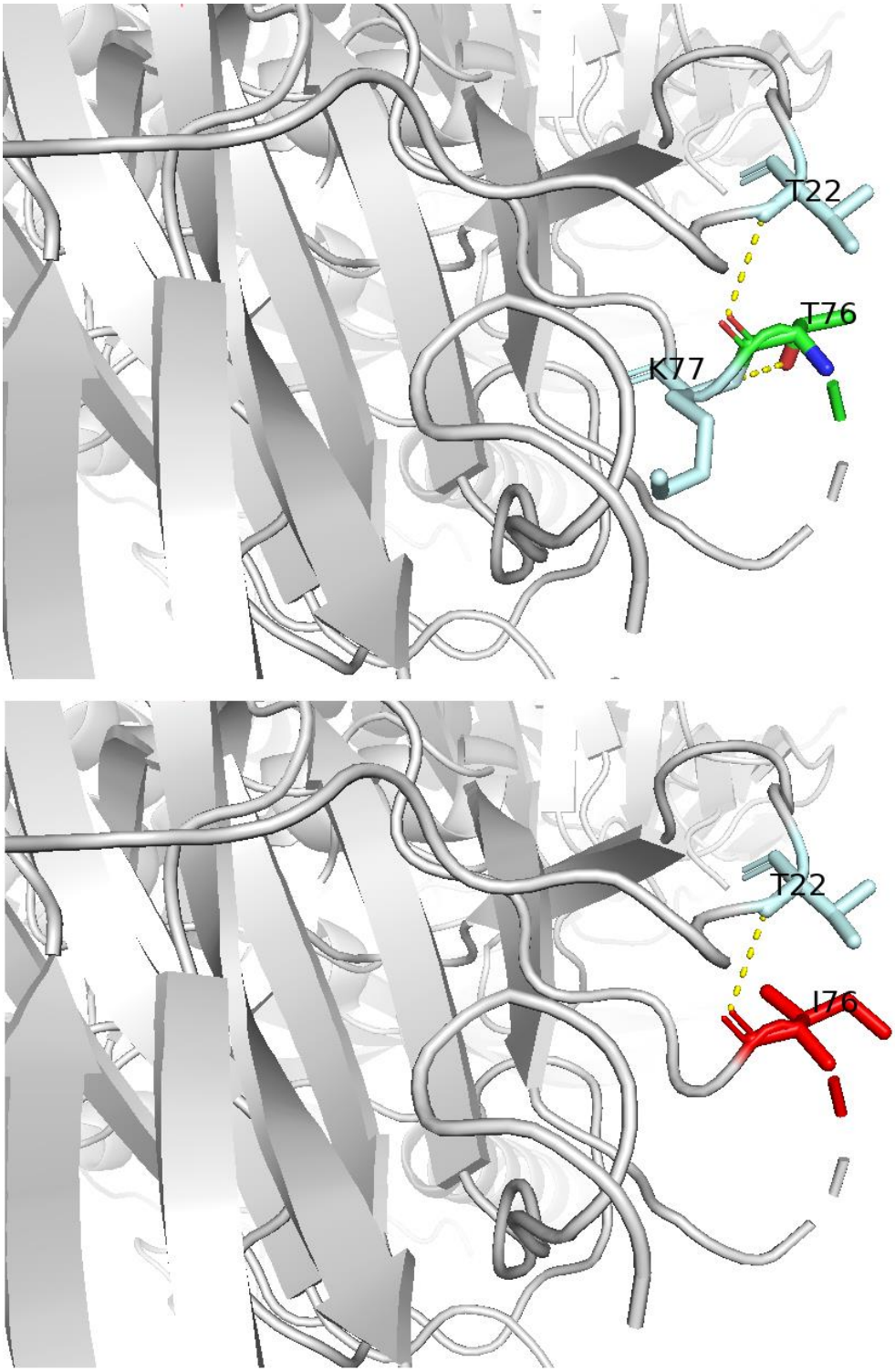


Fig. 31 Close-up of the Spike T76I mutation derived from CaCo-2 cultivation at t=1 from Chain C of the "one-up" Ty1-bound prefusion trimer (PDB:6ZXN) *The upper image shows pre-mutation (F157) and the bottom image shows the post-mutation (S157). Colour scheme – same as in Figs.3-4 with the grey labelled sticks in the pre-mutation image refers to residues that are suspected to interact hydrophobically with F157*

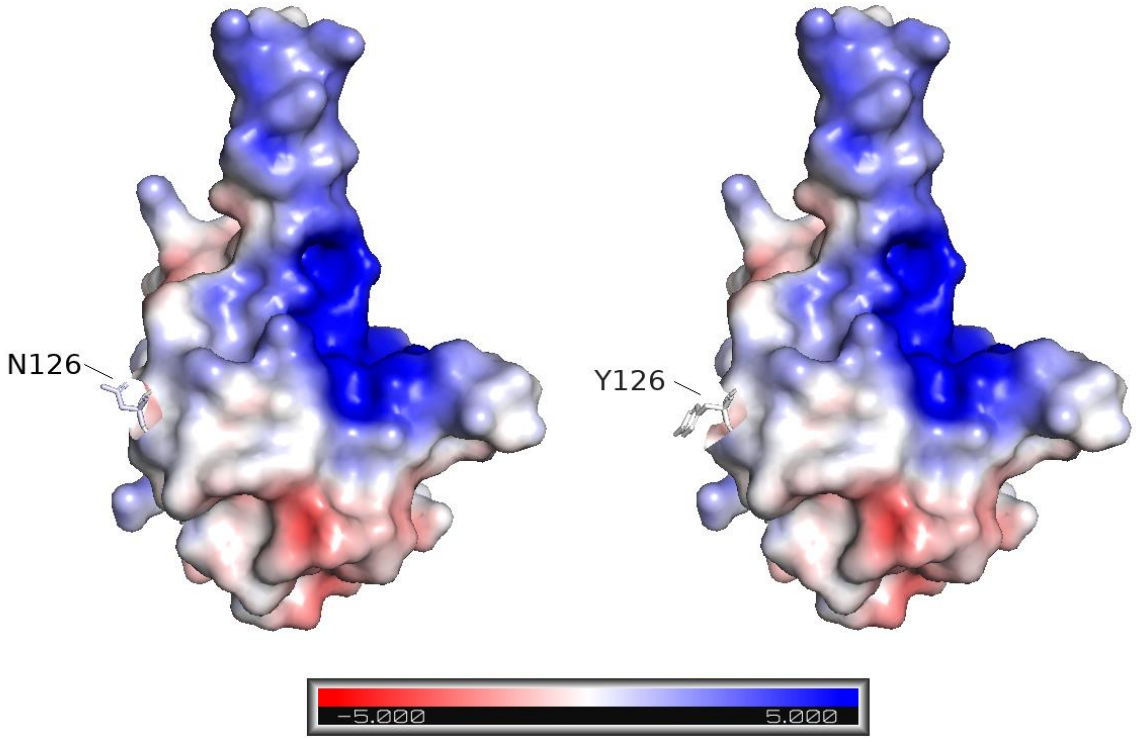


Fig. 32 – Electrostatic map of N protein NTD monomer with (N)N126Y mutation derived from CaCo-2 cultivation at t=1 (PDB:6VYO - Chain A only [116]***)*** *The left image shows pre-mutation (N126) and the right image shows the post-mutation (Y126), red signifies acidic/negatively charged amino acids and blue shows basic/positively charged with electrostatic potential in units (kT/e). APBS plugin with pdb2pqr for Pymol was used to create a map* [78].

FFM7 developed a C=>T 5'-UTR mutation at pos.44 which was maintained to 92.9% of reads in t=2, documentation of this mutation is not available and is likely at undetectable levels in surveillance of global mutations [117], [118].

Three *FFM7* mutations detected at t=1 that were rising in occurrence but were slightly less competitive and failed to reach >90% of base reads include; (S)T76I, (NSP6)M143V and (NSP13)T127I (pos. 21,789, 11,399 and 16616 respectively). (NSP6)M143V is not found in literature and accounts for 0.02% of GISAID samples with almost 8/10 cases occurring in North America, (NSP13)T127I is more frequent in the global population with 0.15% occurrence and has been noticed as a minor mutation [112], [119]. Another C=>T at pos. 28,311 was actively selected against as the cultivation progressed from t=1 to t=2

stage dropping from over 2/3 of base reads to just under 1/3, this mutation results in amino acid changes in two proteins: (N)P13L and (ORF9b)P10S which has a global occurrence of 0.98% with 2/3rd of hits concentrated in North America and has been mentioned briefly in surveillance studies [112], [119]. (N)P13L is correlated to reduced disease severity and may have an impact on the binding of RNA or to M and is found in an antigenic area of SARS-CoV N Ab-epitope mapping [111], [120]. The impact on ORF9b could pertain to antiviral IFN-I suppression via inhibition of adapter proteins such as TOM70 [121].

FFM7 developed a notable mutation in the NTD of S protein (T=>C pos. 22,032) which successfully populated 81.1% of base readings by t=2. The resulting missense mutation (S)F157S, has a global frequency of 0.52% with brief mentions in literature [114], [122]. (S)F157S has been identified to be potentially destabilising and may cause reduction/re-arrangement of hydrogen bonds between the S RBD and ACE2 during receptor binding [123]. If the (S)F157S is run through Dynamut2 the "closed" S conformation (6XR8) the calculated ΔΔG is a whole unit more negative than any other inputted mutations so far (avg. ΔΔG = -2.08kcal/mol), which is explainable by the loss of hydrophobic association to other NTD residues V120, L141, V159 and Y160; the latter of which could lose a possible π-π interaction with F157 (see Fig. 32). (S)F157S in the "one-up" conformation (6VSB) chains A-C have ΔΔG values of -0.36, -0.11 and -1.1 kcal/mol respectively, in Chain A only a minor hydrophobic association to L141 is lost and Chain B S157 is stabilised by the amine sidechain of Arg102 (not shown); Chain C is the most destabilised by S157 which is due to disruption of aromatic binding to Y160 and loss of hydrophobic interaction between V120 (see Fig. 33). Additionally, (S)F157S is located nearby N122-glycan and disulphide bond C15-C136 (Figs.33-34).

FFM7 also contained a singular silent C=>T mutation at t=1 stage detection concerning pos. 835 in t=2, which was well-tolerated reaching 92% frequency of base reads.

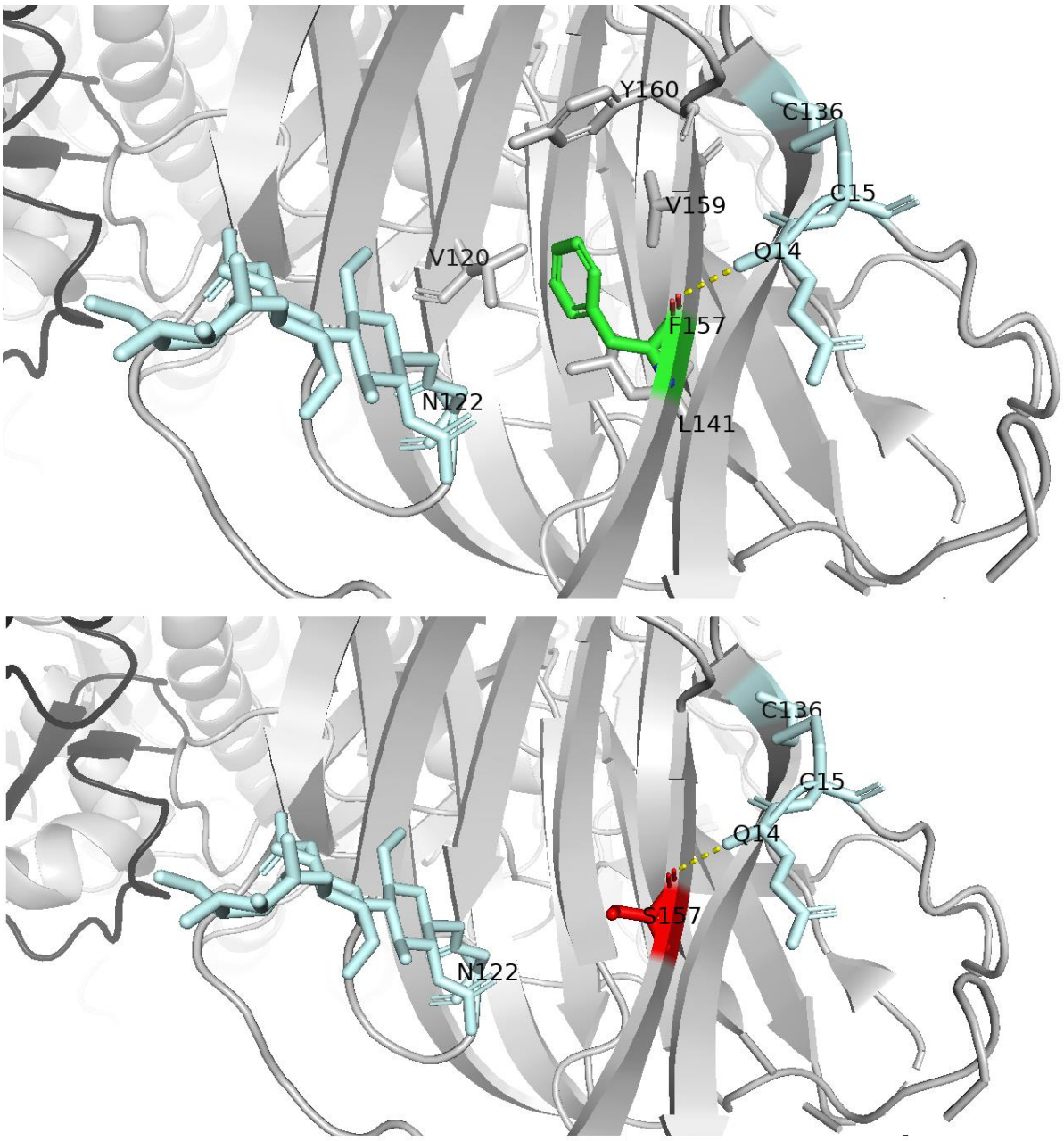


Fig. 33 - Close-up of the Spike F157S mutation derived from CaCo-2 cultivation at t=1 from Chain A of the "closed" prefusion trimer (PDB:6XR8) *The upper image shows pre-mutation (F157) and the bottom image shows the post-mutation (S157). Colour scheme – same as in Figs.3-4 with the grey labelled sticks in the pre-mutation image refers to residues that are suspected to interact hydrophobically with F157*

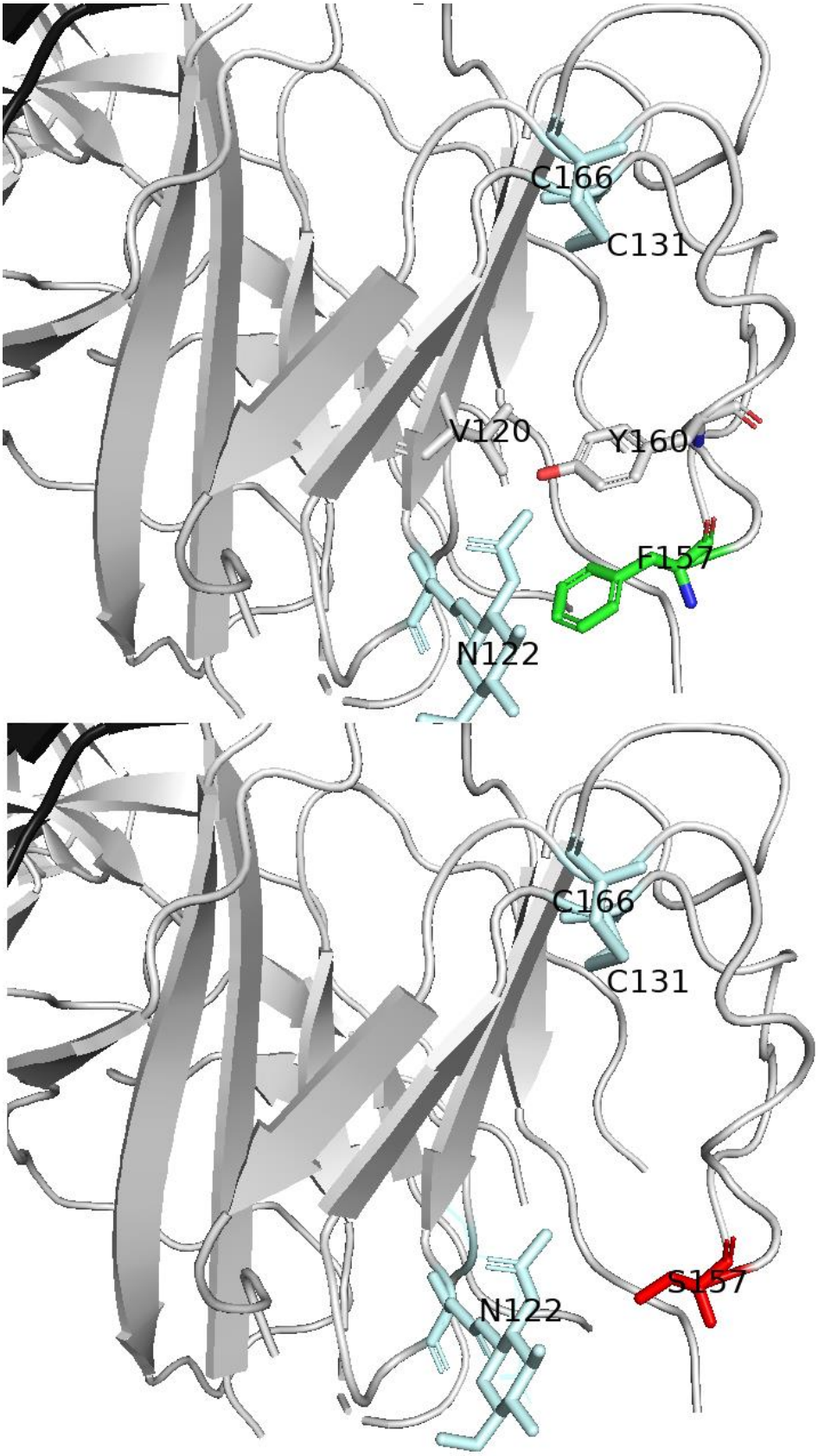


Fig. 34 - Close-up of the Spike F157S mutation derived from CaCo-2 cultivation at t=1 from Chain A of the "one-up" prefusion trimer (PDB:6XR8) *The upper image shows pre-mutation (F157) and the bottom image shows the post-mutation (S157). Colour scheme – same as in Figs.3-4 with the grey labelled sticks in the pre-mutation image refers to residues that are suspected to interact hydrophobically with F157*

t	Strain	Pos (nt)	Mut type	Postmut nt.% (t=0, t=1, t=2)	Δ amino acid	Notes:
t=1	**FFM3**	516-518	TTA_del	6.9, **20.5**, 67.3	(pp1a, nsp1)M85del	3nt deletion
		21,789	C=>T	0.4, **99.4**, 99.9	(S)T76=>I76	
		27,131	C=>T	0.1, **30.1**, 24.5	(M)N203	Silent
		28,649	A=>T	0.0, **99.5**, 99.7	(N)N126=>Y126	RNA-binding region
	FFM7	44	C=>T	0.0, **52.9**, 92.9	(5'-UTR)	
		835	C=>T	0.0, **54.1**, 92.0	(pp1a)F140 (nsp2)F10	Silent
		11,399	A=>G	0.0, **51.3**, 59.3	(pp1a)M3712=>V3712 (nsp6)M143=>V143	
		16,616	C=>T	0.0, **21.5**, 22.9	(pp1ab)T5450=>I5450 (nsp13)T127=>I127	
		21,789	C=>T	0.0, **51.1**, 69.3	(S)T76=>I76	RaTG13
		22,032	T=>C	0.0, **39.1**, 81.1	(S)F157=>S157	
		28,311	C=>T	0.0, **67.4**, 28.0	(N)P13=>L13 (ORF9b)P10=>S10	

Table 5 - ***Identification of nt. base changes in FFM3 & FFM7 midway cultivation (t=1) relative to before cultivation (t=0).*** *Postmut nt.% column describes the percentage of base reads that match the postmutation state in the three timeframes separated by commas - before cultivation t=0, midway cultivation t=1 (in bold) and end of cultivation t=2.*

4.3.3 t=2

(Ref. Table 6 for a full list of t=2 mutations). At t=2 *FFM3* and *FFM7* had twelve non-silent point mutations with a singular one being a G=>T transversion. Two mutations were silent, there were two truncation events in ORF8 and ORF9c caused by point mutations and two deletion events were in NSP1 and NSP12, the latter of which could truncate NSP12 if all predicted 7nt are deleted at once to cause frameshifting.

Mutations detected for *FFM3* at the end of the cultivation included deletions between pos. 508-522 in NSP1 of five sequential amino acids (GHVMV82-86del) which were retained in ⅔ of base reads. GHVMV82-86del occurs in 0.1% of all GISAID samples - of these samples approximately ⅔ were collected in Europe and just under ⅓ in North America with a minor presence in all other continents. If each deletion is considered individually, the occurrence is slightly higher and ranges between 0.2-0.45%

with the highest occurrence belonging to the previously detected M85del which appeared firstly in t=1 and after t=2 increased from 20.5% to 67.3% of base reads.Values for each deletion are as follows: G82del 0.2%, H83del 0.2%, V84del 0.3%, M85del 0.45%, V86del 0.2%.

The location of GHVMV82-86del on NSP1, is at the N-terminal side of β4 and partially on the preceding loop which resides in a highly basic face of the 3D surface of NSP1 (see Fig. 35). GHVMV82-86 share backbone interactions with residues S74, P80, K120, L122 and R124 (the former two are found on the loop between β3-β4 and the latter three are found on β7). There was a suggestion in a structural comparison of NSP1 SARS-CoV and SARS-CoV-2 that alterations of the loop between β3-β4 could affect host protein binding interactions [124]. Deletions in the NSP1 pos.500-532nt have been found to downregulate IFN-β responses, reduce viral load and was associated with less severe symptoms in infected patients, a deletion of 5 amino acids (pos.509-523) reduced IFN-I responses in HEK293T and A549 cell lines [125]; reduced severity can decrease awareness of infection which may lead to higher infection rates.

FFM3 also demonstrated a deletion in pos. 14,408-14,414 which was retained in 2/3rds of reads. The 7nt deletion is derived from C=>T point mutation and leads to a -1 frameshift and early STOP codon in the RdRP resulting in: (NSP12)^{323}PTSFG=>^{323}LLDH- ("-"is signifying a STOP codon at pos. 327). P323L has been noticed to co-occur with the *FFM3* t=0 starting mutations C=>T at pos. 241nt in the 5'-UTR and pos. 3,037nt in NSP3 [110], [126]. Although (NSP12)P323L is extremely prevalent, occurring in 95.6% of the global data, cases of P323L with T324L/S325D/F326H/G327STOP are comparatively rare - but have appeared (EPI_ISL_465707 – note this sample has a low coverage) and there are similar entries with NSP12 truncation at the same position ending with ^{323}VLDH- which can be caused by 1nt deletion at pos.14,408 but is not the case in any of these entries (EPI_ISL_460634, EPI_ISL_478284, EPI_ISL566076, EPI_ISL_500814, EPI_ISL_500824, EPI_ISL_500806, EPI_ISL_500805, EPI_ISL_500815, EPI_ISL_448969, EPI_ISL_478141).

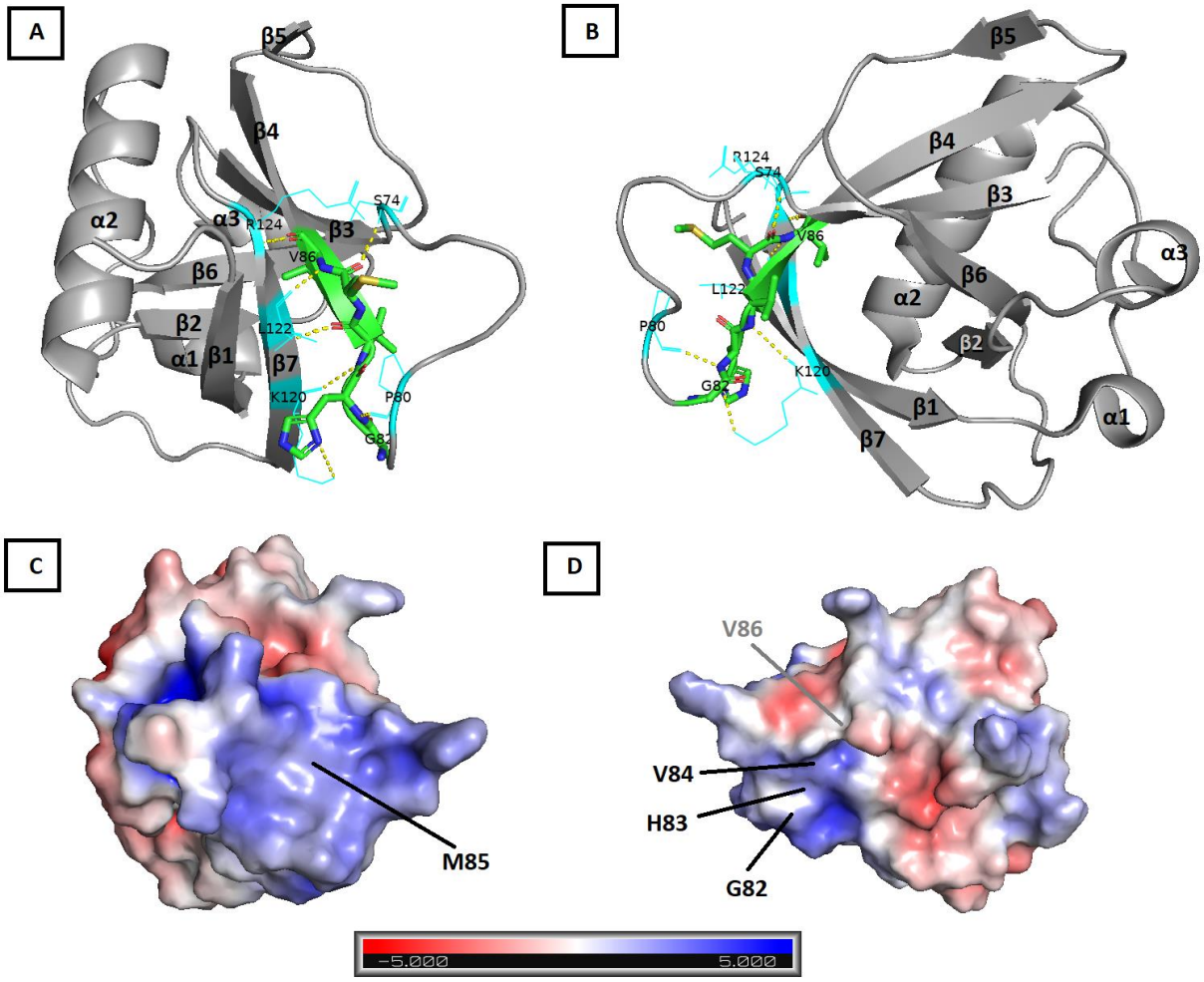


Fig. 35 – Crystal Structure of NSP1 protein showing GHVMV82-86del site derived from CaCo-2 cultivation at t=2 (PDB: 7K7P [124]***).*** *Fig. 35A-B shows NSP1 (residues 13-127) labelled secondary structure with 1B rotated left 180° relative to 1A, green sticks show the GHVMV82-86 stretch with yellow dashes to cyan wire residues representing polar interactions. Fig. 35C-D are corresponding electrostatic surface maps of A-B in (APBS plugin with pdb2pqr for Pymol* [78] *units in kT/e), black labels show residues found at the surface whereas grey indicates buried residues.*

Nearly all these samples are unusual, either being incomplete with large unknown stretches or a high frequency of mutations, for example, EPI_ISL_460634 submitted from Seattle, Washington in April 2020 has 112 mutations within NSP12 along with the similar nonsense mutation ^{323}VLDH- indicating the -1 frameshift. RdRP is rendered non-functional from truncation at G327STOP which is intolerable for

SARS-CoV-2 replication. It is far more likely that pos.14,408-14,414 is prone to more than one possible variation of in-frame deletion which belongs to separate populations in the *in vitro* sample, the key aspect of this area of NSP12 is that within the RdRP is that it is located next to the binding site of NSP8 whereby the backbone carbonyls of (NSP12)P323, T324 and F326 all stabilise the amide group of (NSP8)N118 (see Fig. 36); alteration to this site via deletions could alter the recruitment of NSP8 to the RdRP complex which could entail changes to RNA synthesis.

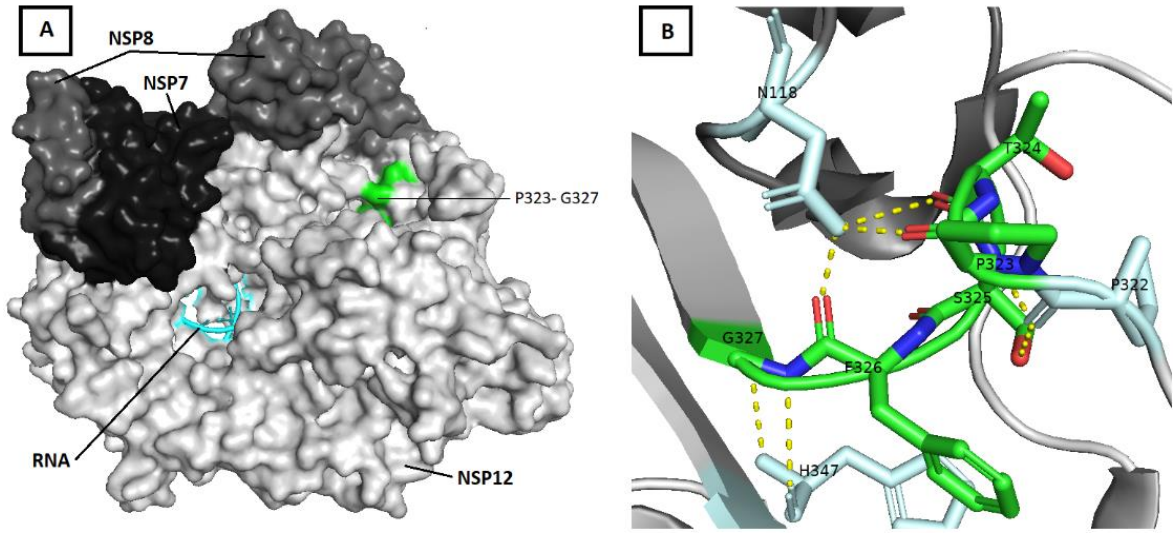


Fig. 36 – EM Structure of RdRP:RNA showing (NSP12)P323-G327 site affected by pos.14,408-14,414 deletion derived from CaCo-2 cultivation at t=2 (PDB: 6XQB [127]***).*** *Fig. 36A shows the surface structure of RdRP with NSP7/8/12 and bound RNA with the P323-G327 site in green, Fig. 36B shows a close-up of P323-G327*

(S)A570D was found in ½ of t=2 base reads for *FFM3*, (S)A570D is caused by C=>T at pos. 23,271 and has a high global occurrence of 42.8% as (S)A570D is part of the B.1.1.7 lineage (Alpha variant) repertoire of mutations (Ref. Table 1). (S)A570D was discussed earlier as a DCP for (S)N501Y with function in S RBD movement with a "pedal bin" mechanism and as a compensatory salt bridge for (S)D614G [93].

FFM3 at pos. 25,688 also carries (ORF3a)A99V which made up 45% of t=2 base reads. A99V occurs in the transmembrane region and has a global frequency of 0.15% - it is in the top 5 most common ORF3a mutations and tends to co-occur with (ORF3a)Q57H which is true for a third of A99V-positive GISAID samples but not for *FFM3,* A99V's impact is likely minor as it is predicted to only be marginally destabilising by Dynamut and neutral by PROVEAN [128], [129].

FFM3 carried a singular silent C=>T mutation within NSP15 at genomic pos. 20,178nt and five mutations with very low global frequencies which if ranked in global frequency are found at pos. 20,480, 7,521, 11,760, 17,146 and 20573 (with respective global percentages of: 0.09%, 0.01%, 0.003%, 0.002% and 0.002%); which result in the following amino acid changes - (NSP15)S287L, (NSP3)T1601I, (NSP6)K263R, (NSP13)I304V and (NSP15)V318A. For *FFM3,* (NSP15)V318A was the most frequent mutation of any detected at the t=2 stage (reaching 71.5% whereas the former mutations failed to reach over 50% of base reads) despite having little significance in global data.

FFM7 at the t=2 stage has a few C=>T mutations of global significance. Genomic pos. 17,678 (base read frequency of 60.7%) with resultant (NSP13)T481M occurs in 0.25% of global samples with almost 70% of these originating from North America [112]. Dynamut2 predicts (NSP13)T481M to have a stabilising effect in the RTC which concerns RdRP plus NSP7 and dimers of NSP8 and NSP13 (PDB: 7CXM, Chains E-F $\Delta\Delta G$ = 0.31, 0.35 kcal/mol), T481M is near the RNA-binding domain of NSP13, polar interactions seem unaffected but the transition to a hydrophobic residue could impact helicase activity.

C=>T at pos. 6,255nt leads to (NSP3)A1179V (base read frequency of 44.5%) has a global occurrence of 0.19% with a quarter of V1179-positive sequences in GISAID matching the B.1.36 lineages (e.g. the fourth infection surge Hong Kong occurring in November 2020 as part of B.1.36.27 [130]) and pos.19,955 (NSP15)T112I occurs globally at 0.14% but has not been reported on elsewhere as of yet.

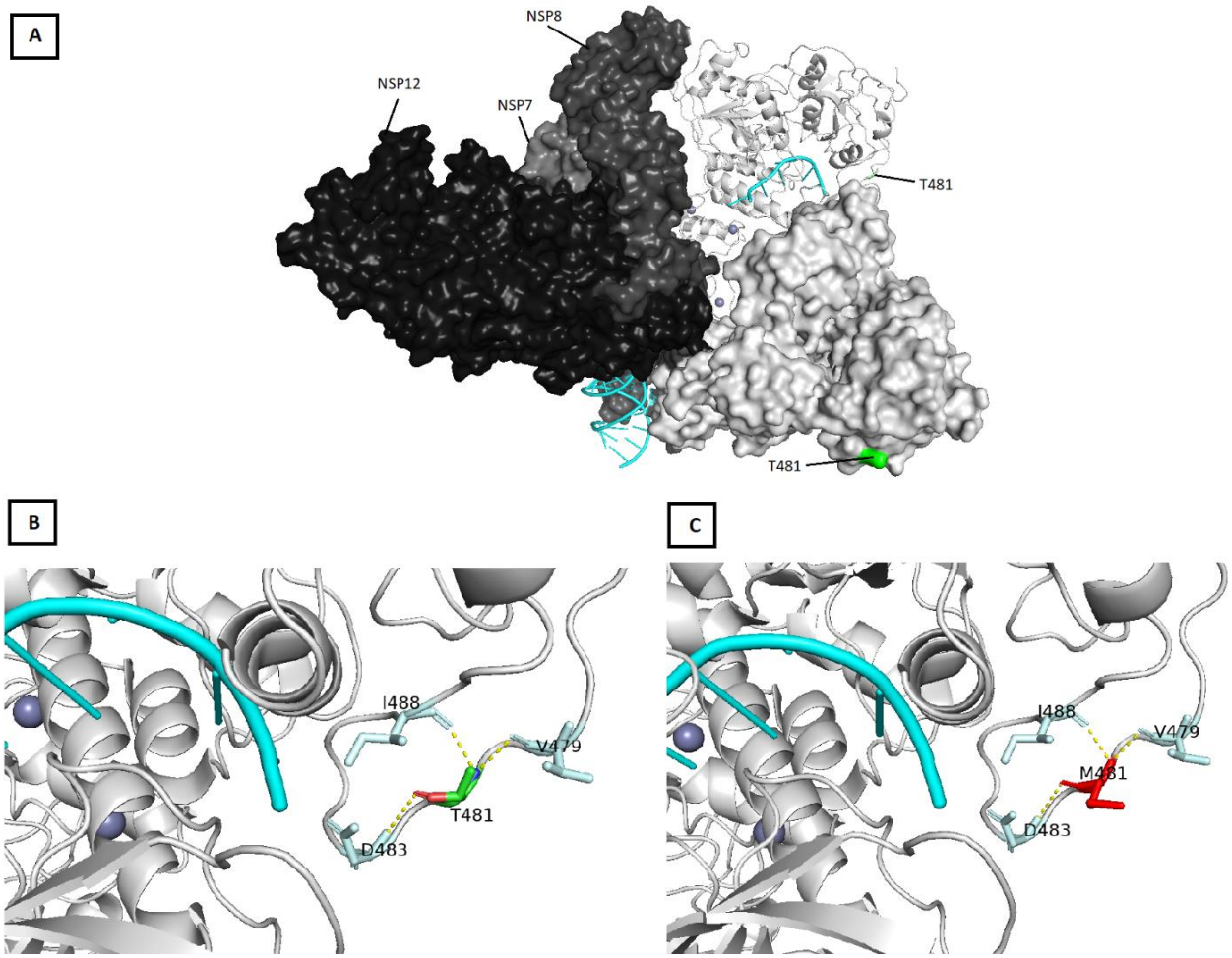


***Fig. 37 – EM structure of mini RTC (RdRP:RNA:NSP13) showing (NSP13)T481M derived from CaCo-2 cultivation at t=2 (PDB: 7CXM** [131]**).** Fig. 37A shows the RTC with NSPs7/8/12/13 of which NSP8 and NSP13 are dimers (Chains B-D and E-F, respectively), Chain E is in cartoon structure to reveal the RNA (cyan) binding site within NSP13. Fig. 37B-C show a close-up of T481M pre- vs. post-mutation.*

FFM7 developed a truncation of ORF8 which was described earlier [60] (see Table 1), was firmly established *in vitro* matching 99.6% of readings. (ORF8)Q27STOP is associated with the B.1.1.7 lineage which explains its high global significance (41.3%) but was not identified as a DCP relative to N501Y in this investigation; truncation of ORF8 has appeared independently of B.1.1.7 and non-functional ORF8 is thought to reduce innate immune evasion by MHC-I downregulation and may reduce disease severity [132]–[134].

FFM7 developed two mutations in N that are mentioned in the literature [112], [135]. Firstly, a C=>T mutation presented at pos. 28,887 in ⅓ of base reads that simultaneously results in (N)T205I and (ORF9c)L52F (global percentage of 1.05%) which is concerned with the South African VOC (501Y.V2) in the B.1.351 lineage [135], [136]. Secondly, a G=>C transversion at pos. 28,899 resulted in (N)R209I and (ORF9c)E56STOP (0.19% global percentage) which was present in 2/3rds of base reads by the end of the cultivation. These two neighbouring N mutations occur in an area mapped to antibody epitope regions in the SARS-CoV N protein and R09I specifically has been found in Italian samples with high viral loads that result in negative antibody tests [111], [137].

Finally, *FFM7* carries a silent C=>T mutation at pos. 2,509nt that occurred in 44.3% of t=2 reads.

t	Strain	Pos (nt)	Mut type	Postmut nt.% (t=0, t=1, t=2)	Δ amino acid	Notes:
t=2	**FFM3**	508-522	TGGTCATGTTATGGT_del	5.9, 13.8, **64.8**	(pp1a, nsp1) GHVMV82-86_del	5aa deletion
		7,521	C=>T	0.0, 3.4, **46.0**	(pp1a)T2419=>I2419 (nsp3)T1601=>I1601	
		11,760	A=>G	0.0, 1.5, **31.0**	(pp1a)K3832=>R3832 (nsp6)K263=>R263	
		14,408-14,414	CTACAAG_del	0.6, 7.6, **68.4**,	(pp1ab) ^{4714}PTSFG => ^{4714}LLDH- (nsp12) ^{323}PTSFG => ^{323}LLDH-	7nt deletion leads to -1 frameshift and early STOP
		17,146	A=>G	0.0, 0.0, **25.1**	(pp1ab)I5627=>VI5627) (nsp13)I304=>V304	
		20,178	C=>T	0.0, 5.9, **38.0**	(pp1ab)V6637 (nsp15)V186	Silent
		20,480	C=>T	0.0, 11.1, **28.4**	(pp1ab)S6738=>L6738 (nsp15)S287=>L287	
		20,573	T=>C	0.0, 10.7, **71.5**	(pp1ab)V6769=>A6769 (nsp15)V318=>A318	
		23,271	C=>T	0.0, 6.0, **47.4**	(S)A570=>D570	A570D
		25,688	C=>T	0.0, 3.0, **45.3**	(ORF3a)A99=>V99	
	FFM7	2,509	C=>T	0.0, 10.3, **44.3**	(pp1a)P748 (nsp2)P568	Silent
		6,255	C=>T	0.0, 0.0, **44.5**	(pp1a)A1997=>V1997 (nsp3)A1179=>V1179	
		17,678	C=>T	0.0, 6.8, **60.7**	(pp1ab)T5804=>M5804 (nsp13)T481=>M481	
		19,955	C=>T	0.0, 0.0, **55.1**	(pp1ab)T6563=>I6563 (nsp15)T112=>I112	
		27,972	C=>T	0.0, 2.0, **99.6**	(ORF8)Q27=>STOP27	B.1.1.7 lineage
		28,887	C=>T	0.0, 7.6, **30.1**	(N)T205=>I205 (ORF9c)L52=>F52	B.1.351 lineage
		28,899	G=>T	0.0, 9.9, **64.3**	(N)R209=>I209 (ORF9c)E56=>STOP56	

Table 6 - Identification of nt. base changes in FFM3 & FFM7 at end of cultivation (t=2) relative to midway cultivation (t=1). *Postmut nt.% column describes the percentage of base reads that match the postmutation state in the three timeframes separated by commas - before cultivation t=0, midway cultivation t=1 and end of cultivation t=2 (in bold).*

5.0 Discussion

5.1 (S)D614G

The (S)D614G mutation was attributed to greater infectivity of SARS-CoV-2 beginning in March 2020 and now represents >98% of GISAID entries. Due to the functional outcome of (S)D614G being attributed to a greater occurrence of “one-up” conformation in S [52], [56] all structures were analysed in both “closed” and “one-up” states – including the later separate co-mutations logged with N501Y to see if any may perturb mechanisms related to ACE2 binding.

D614G was previously grouped with C=>T mutations at pos. 241, 3,037 and pos. 14,408 (located in 5’-UTR, NSP3 and NSP12) by Korber et al. [50]. In this study which uses GISAID data up until early December 2020, two more (S)D614G-related DCPs were detected; (S)L18F and (S)A222V are found in the NTD of S and can be said to have been mildly concurrent with D614G. F18 occurrence rose from 0.12% to 9.2% and V222 occurrence rose from 0.1% to 18.9% when incidence is compared in the D614 vs. G614 datasets.

L18F was calculated to be destabilising, slightly more so in the “one-up” state (avg. “closed” and “one-up” $\Delta\Delta G$=-0.59kcal/mol, -0.63kcal/mol) and had possible ramifications for nearby glycans (N17, N165) as well as disulphide bridges (C15-C136, C131-C166) of which the latter disulphide bridge neighbours an RBD-NTD “tethering residue” (T167 or N165 – dependent on PDB structure chosen) which may be impacted by F18. Primary concerns in literature with L18F are Ab-dependent resistance and L18F has since been identified as a substrain to the B.1.1.7 lineage in 2021 but earlier was part of the A222V-carrying 20A.EU1 strain that was abundant pre-December 2020 [84], [85]. The L18F-containing B.1.1.7 substrain was calculated to have a 1.7-fold replicative advantage during December 2020 and continues to be relevant in established (e.g. South-African/Brazilian) and newer strains [138].

A222V lies nearby to glycan N282 and lends itself to the name of “GV” clade (20A.EU1 variant [139]) that arose in Europe of 2020 was conversely calculated to be a stabilising mutation (avg. “closed” and “one-up” ΔΔG=0.23kcal/mol, 0.27kcal/mol) which could be due to a network of hydrophobic interactions with I285, Y38 and a new interchain aromatic association with Y38-F562 caused by clashing of the slightly larger valine residue. A222V is suspected of residing in an antigenic region [86].

D614G had only previously been looked at in the “one-up” form whereby intrachain association from D614 to T859 was lost [50]. Inspection of D614G in the “closed” form identified that interchain loss of salt bridges to (S)K835 and (S)K854 is abolished with G614; this would undoubtedly cause greater chain flexibility and/or destabilisation in the “closed” state which could cause preference to the “one-up” conformation. Dynamut2 also predicted (S)D614G to be more destabilising in the closed state (avg. “closed” and “one-up” ΔΔG=-0.64kcal/mol, -0.27kcal/mol).

5.2 (S)N501Y

DCPs uncovered regarding the RBD mutation (S)N501Y attributed to the B.1.1.7 lineage are all localised in S and have previously been found to be concurrent with N501Y, namely A507D, P681H, T716I, S982A, D1118H and the two deletions HV69-70del and Y144del (also known as Y145del). In this case, all mutations were found with S by Rambaut et al.[60] (listed in Table 1) were linked with N501Y in this investigation but those within NSP3, NSP6 and accessory protein 8 were not identified to be DCPs. Of note, two B.1.1.7 lineage mutations A570D and the truncation of accessory protein 8 were found to occur spontaneously in the CaCo-2 cell cultures.

N501Y was viewed in the “closed” and ACE2 bound forms due to being in the RBM, it is postulated that Y501 can form a new π-π bond with (ACE2)Y41 and may impact other residues in the S/ACE2 binding site such as (S)G496, (S)Q498 and (ACE2)K353. Stability calculations for N501Y indicated an effect

solely on the ACE2 bound form (avg. “closed” and ACE2 bound ΔΔG=-0.03kcal/mol, -0.43kcal/mol) which is expected as the clashing of Y501 with nearby residues can be easily relieved by movement of the unstructured loop Y501 sits upon.

A570D is calculated to be highly destabilising in the “closed” form of S protein and clashes heavily with residues D568 and T572 (avg. ΔΔG=-1.28kcal/mol), whereas in the “one-up” form the destabilisation is weaker but has an uneven but distribution between S protomers (Chains A-C ΔΔG=-0.28, -0.69, -0.53 kcal/mol). D570 is postulated to have a role in protomer stabilisation to cause preference to the “one-up” state as 1) the “up” monomer was calculated to be the most stable in the “one-up” state and 2) “closed” state was equally highly destabilised, 3) the position of A570D underneath the globular RBD has potential as a pivoting region – which is further supported by the fact that the G614 mutation, that is already confirmed to have a hinge-like function - occurs on the opposite side of the same α23-helix. In addition, A570D was predicted a large range difference in their stability calculations which also occurred with D614G. A570D has elsewhere been related to a “pedal-bin” mechanism of the RBD extension and provides compensation for the loss of D614G salt bridges [93].

P681H remains elusive due to being unresolved in PDB structures, possible reasons behind P618H being retained in the population could include allosteric binding or possible adaptation to cleavage via furin (e.g. in [100] cleavage was improved but not thought to impact host cell entry or cell-fusion) or TMPRSS2

T716I is predicted to be one of few stabilising mutations in both S conformations (avg. “closed” and “one-up” ΔΔG=0.25kcal/mol, 0.51kcal/mol), T716I does not increase psuedovirion infectivity and is likely acting in concert with other B.1.1.7 lineage mutations [102]. T716I may interrupt β-sheet organisation in the S2 stalk, although this is speculation – T716I could be influencing rotational movement as part of a heavily glycosylated “hip” joint [103].

S982A likely results in the breaking of the α28 helix that resides in the coiled-coil region between the HR1 and HR2 segments. Similarly to D614G and A570D, which are thought to be involved in protomer instability, Dynamut2 stability calculations between the "closed and "one-up" state showed greater destabilisation in the "closed" state (avg. "closed" and "one-up" ΔΔG=-0.58, -0.14 kcal/mol). S982 may stabilise the "down" position of the RBD within adjacent chains via intermolecular hydrogen bonding to T547 and possible interaction with K386. S982A has poor coverage in research papers and tends to be mentioned briefly within the context of the B.1.1.7 mutation repertoire, computational structure predictions for S982A suggest no resultant structural changes and potential effects listed elsewhere include changes to protomer stability as well as to cleavage, fusion and antigenic recognition [95], [104].

D1118H was calculated to be equally destabilising in both "one-up" or "closed" states (avg. ΔΔG=-0.36kcal/mol), suggesting that D1118H does not cause a preference to either S protein conformation. D1118H occurs facing inside a bulbous hollow cavity mostly composed of β-sheets between the two coiled-coil regions, during the shift to "one-up" conformation the lowest strain H1118 rotamer of Chain B flips the direction of its sidechain relative to the other H1118 residues and the distance between the closest atoms between the histidine residues reduce by 0.7Å which indicates compression of the region. However, as destabilisation values are near identical this is unlikely due to the protomer "one-up" conformation but could be due to movement in a rotational or elongational fashion of the area – which could be akin to the "hip joint" [103]. In addition, Asp/His charged residues at pos. 1118 would aid solvent entry and prevent the Chains from annealing together which would allow for individual protomer movements. Computational structure analysis suggests the loss of two nearby β-sheets and the possible effect on T-cell epitopes although the effect on B-cell epitopes is unlikely due to the buried nature of D1118H [95].

Deletions found that are related to N501Y are both found in the NTD and are HV69-70del and Y144del (AKA Y145del), computationally generated structures suggest that HV69-70del does not lead to significant changes whereas Y144_del could cause β-sheet loss in favour of a coiled-coil structure [95].

HV69-70del itself can be said to have a singular DCP, N439K whereas Y144del shares DCPs associated with B.1.1.7 lineage (i.e. N501Y, A570D, P681H, S982A and D1118H); this would suggest that Y144del is more heavily-associated to B.1.1.7 lineage than HV69-70del, however, both deletions have appeared elsewhere on separate occasions [106], [107]. Like many NTD mutations, 69-70del and Y144del have been related to modulating antigenicity, however, it would appear that Y144del is the most potent counterpart in mAb studies and that 69-70del is likely acting in concordance with other mutations [108].

5.3 Cell culture

Two separate colonies of CaCo-2 cells infected with SARS-CoV-2 were cultivated *in vitro* and sequenced at three timepoints through passaging - which correspond to 0, 30 and 60 passages respectively. The resultant mutations from each time period were picked up at a sensitivity of 20% change in base readings. The most common mutation type was a C=>T transition which occurred in 70% of recorded point mutations; other changes include two truncations in ORF8 and ORF9c and three total instances of deletion occurring at lengths of 3 and 15nt in NSP1 and a 7nt deletion in NSP12.

Before cultivation (t=0), sequencing of SARS-CoV-2 in both colonies identified common C=>T mutations at genome positions 241, 3,037, 14,408nt that are associated with the pos. 23,403nt A=>G transition which results in (S)D614G. Both strains had pairs of unique mutations: *FFM3* carried G=>A at pos. 28,882nt and G=>C at 28,883nt, whereas *FFM7* carried two C=>T silent mutations at pos. 15,324 and 23,179nt within NSP12 and S protein, respectively at t=0.

At t=1 after 30 passages, *FFM3* developed a deletion at (NSP1)M85del and three point mutations which result in (S)T76I, a silent mutation at pos. 27,131 in M protein and (N)N126Y. (S)T76I is found in SARS-CoV as well as bat strain RaTG13, (S)T76I was ubiquitously established in *FFM3* and occupied 2/3rds of reads independently in *FFM7* by 60 passages, it is calculated to be destabilising for both the “one-up” S

protomer and to the anti-clockwise chain relative to the “one-up” protomer. (N)N126Y is very rarely found in the GISAID database, however in *FFM3* successfully dominated at almost 100% of reads, (N)N126Y occurs in the RNA-binding region but not at the main RNA-binding surface and may be antigenic (see Fig. 35) [111], [115]. *FFM7* at t=1 developed two synonymous mutations: C=>T at pos. 44nt which lies within the 5’-UTR and C=>T at pos. 835nt in NSP2. Non-synonymous mutations for FFM7 at t=1 include (NSP6)M143V, (NSP13)T127I, (S)F157S and finally C=>T at pos. 28,311nt which simultaneously results in (N)P13=>L13 and (ORF9b)P10=>S10.

By 60 passages, *FMM3* developed two stretches of deletions in NSP1 and NSP12, (NSP1)GHVMV82_86del can be found in ~4,000 GISAID entries as of September 2021, however, the individual amino acid deletions can be found in higher rates globally e.g. (NSP1)M85del currently detected at ~21,000 entries; therefore this 15nt region within NSP1 could be considered partial to deletions which may be beneficial in protein binding or minimising disease severity by altering the IFN-β host viral immune response [124], [125]. *FFM3* also developed an unbalanced 7nt deletion that results in (NSP12)323PTSFG => 323LLDH- which very rarely appears in global samples as the -1 frameshift results in early truncation of NSP12 would not support replication; it is more likely that an in-frame combination of base deletions occurring between pos. 14,408-14,414nt could be supported; deletions may impact the nearby NSP8 binding region as (NSP12)P323, T324 and F326 possibly interact with (NSP8)N118 (see Fig. 36).

Noteworthy mutations include (S)A570D and (ORF8)Q27STOP, which are associated with B.1.1.7 lineage [60] - this could indicate that these two mutations are not just minor correlations with the other more widely researched mutations of the lineage such as (S)N501Y. (ORF8)Q27STOP of *FFM7* occupied >99% of reads whereas (S)A570D of *FFM3* occupied half of all base reads, their *in vitro* development suggests that they have a shared benefit to a generalised function that can be applied *in vivo* and *in vitro*. (S)A570D was previously suggested to be involved in the mechanism for RBD extension and compensates for salt bridges lost by (S)D614G mutation [93] whereas (ORF8)Q27STOP could

downregulate MHC-1 expression and is correlated to milder infections [133], [134]; if (S)A570D is purely an aid in the mechanism of cell entry then both *in vitro* and *in vivo* infection would benefit, however, the benefit of (ORF8)Q27STOP - which was highly selected for, is not explained *in vitro* by MHC-I downregulation as this mainly relates to humoral immunity.

FFM7 developed a calculated stabilising mutation (avg. $\Delta\Delta G$= 0.33 kcal/mol) - (NSP13)T481M has relevance as a binding component in the RTC (Replication Transcription Complex) which concerns the RdRP plus NSP7, NSP8 and NSP13 – see Fig.37). T481M is near the RNA-binding domain of NSP13 which may allude to an alteration of replication performed by the RTC. *FFM7* has two mutations in N protein associated with epitopic regions (N)T205I and (N)R209I, for which their usage *in vitro* seems irrelevant, however, the same point mutations are also causative for the respective mutations (ORF9c)L52F and (ORF9c)E56STOP. ORF9c has been evidenced to evade innate immunity systems by regulation of interferon/cytokine signalling and interact with membranous host proteins which could be beneficial *in vitro*. [140].

Other nonsynonymous point mutations from the CaCo-2 cultures that occurred by 60 passages include (NSP3)T1601I, (NSP3)A1179V, (NSP6)K263R, (NSP13)I304V, (NSP15)S287L, (NSP15)T112I, (NSP15)V318A and (ORF3a)A99V; there were also two silent C=>T mutations at pos. 2,509nt and 20,178nt.

5.4 Validation of structural models

Regarding the trustworthiness of the PDB files used in this study - wwPDB validation reports provide percentile scoring of Clashscore, Ramachandran outliers and sidechain outliers to give a quality indication of how a structure compares to other matched PDB entries. 6XR8 (the “closed” S trimer), scores well in all categories (4 per 1000 atoms, 0.1% and 0.5% respectively). Only 1% of residues in Chains A-C are considered a “poor fit” to other EM structures. 6VSB (“one-up” S trimer) had a worse-than-average Clashscore – meaning that atoms are unusually close compared to other structures, however,

Ramachandran outliers were averagely scored, and sidechain outliers were better than average – the percentage of mapped outliers was uneven between Chains A-C with Chain A having 7% of residues as outliers and only 1% each in Chains B and C. As a comparative cross-check, PDB 6VYB (another "one-up" S trimer EM structure [35]) has a near-perfect calculated clashcore and the validation report similarly shows outliers only occurring in the "up" chain – this could indicate that the "up" chain tends to be variable and difficult to map precisely. If 6VYB and 6VSB structures are aligned the RMSD between the "up" chains is 0.727Å, and for the "closed" chains - 0.544 and 0.531Å; these are a very close match and go towards validating the 6VSB structure despite its' poor calculated clashscore.

Regarding other PDB structures used in this study, 6ZXN and 6VYO were well-validated whereas less-ideally validated structures include 6M17 and 7CXM (which may be due to their complexity). 7K7P, an x-ray crystal structure, had a relatively poor real-space R-value z-score and 6XQB had a poor clashscore with the bound RNA elements having slightly irregular torsion angles.

5.5 Conclusion

Surveillance of SARS-CoV-2 and its ever-evolving RNA genome will most likely need to continue into 2022, even as we surpass 3.5 billion global vaccinations we are still experiencing over 600,000 new infections daily in September 2021 [3] and tens of thousands of submissions per day continue to be uploaded to GISAID. This study has found that S in SARS-CoV-2 is currently the main source of DCPs - which would suggest changes in S protein contains many of the evolutionarily important mutations - the variability of S protein was recognised even at the very beginning of the pandemic [1]. mRNA vaccines use S protein mRNA (e.g. Pfizer, Moderna, AstraZeneca), so antibody-evasive mutations in S are a major research interest [102], [105], [107], [108]. Vaccination options containing more than just S, or fragments of, will likely become more available, such as the entire attenuated virion or other proteins, although more time is needed to assess biological safety [141].

Regarding the DCPs occurring alongside (S)N501Y, this investigation highlighted DCPs that all resided in S protein (A570D, P681H, T716I, S982A, D1118H plus deletions HV69-70del, Y144del) which suggests the other B.1.1.7 lineage mutations from the preliminary report [60] are less specific to B.1.1.7 which is supported by the truncation of ORF8 occurring in cell culture; however, (S)A570D also arose independently in *FFM3* which could mean that by chance that (S)A570D was pooled with the other more influential mutations that had higher selection pressures in the evolution of B.1.1.7 has a general neutral or positive selection pressure which would align with the (S)D614G compensation/"pedal-bin mechanism" theories [93].

This study uncovered deleterious regions, such as (NSP1)GHVMV82-86 and (NSP12)PTS323-325 which both arose simultaneously in untreated CaCo-2 cell culture, although one could have arisen as a consequence of the other (i.e. polymerase deletions leading to impaired RTC-mediated RNA synthesis and base-pair matching); NSP1 deletions seem to have roles in IFN-β or could generally alter protein association [124], [125]. The NTD deletions of S protein that are likely to affect antigenicity and did not arise spontaneously in cell culture are the HV69-70del and Y144del/Y145del which are associated with B.1.1.7 strains (with Y144del more often occurring alongside the defining lineage mutation (S)N501Y. It seems to be a common theme that S NTD mutations alter defences against the immune system as the same can be said for (S)D614G DCPs - L18F and A222V [84], [85].

To conclude, (S)D614G is likely to be a leading mutation that drove the increase in infection rates from March 2020 rather than a bystander mutation. No deletions were detected and only two DCPs (L18F and A222V) were found, which are relatively minor global mutations and appear to have a lower impact on the overall S structure. These DCPs could complement (S)D614G as (S)L18F is a monitored mutation associated with the Beta variant, that formed a substrain with a replicative advantage associated with the Alpha variant [59], [83]. A222V was substantial as it formed the "GV" clade that was dominant in UK/Europe in the latter months of 2020 [88], [142]–[144]. Both L18F and A222V appeared in the European substrain 20A.EU1 [83]. Analysis of DCPs from (S)N501Y revealed that the core S mutations

are truly concurrent and ruled out other mutations that appeared in the early investigations [60] which happened to appear in the first Beta variant samples from Milton Keynes (EPI_ISL_601443, EPI_ISL_581117). If this investigation were to be expanded, analysis of DCPs should be segmented over smaller time periods and updated with newer sequences to describe the gain/loss of DCPs over time. Finally, the study of mutations from cell culture provided an index of mutations that can occur *in vitro* which could be useful in future studies for deducing function, as dominating mutations found in *FFM3* and *FFM7* are unlikely to be related to humoral immunity or adaption to the diversity of host cell type - but could relate to cell entry, replicative speed and/or stability, innate immunity or substrate binding/allostery.

5.6 Significance

At time of corrections in February 2022, the globe is dominated by the Omicron variant (BA.1 or "GRA clade) which has previously phased out Delta (B.1.617.2 or "GK" clade) which took hold in the latter half of 2021. In the past month, out of 511,421 total GISAID submissions - 477 sequences contained (S)L18F, and (S)A222V appeared 605 times and (S)D614G still applies to 99.4% of sequences. One Delta-type sequence, EPI_ISL_9844246, from a patient in Indonesia contains (S)L18F, (S)A222V *and* (S)D614G.

Alpha variant mutations that are strongly shared with Omicron (>80% of all entries) include HV69-70del, (S)Y144del (S)N501Y and (S)P681H. Omicron entries for February 2022 do contain some of the other mutations shared with the (S)N501Y-DCPs from this study, albeit in low frequencies: i.e. (S)T716I, (S)S982A and (S)D1118H with 63, 20, 21 entries respectively.

Regardless of submission dates, (S)A507D was not found in any Omicron entries and appeared sparsely in all Delta entries (0.004%), which would indicate (S)A570D was mostly isolated to the Alpha variant. The (S)D614G-DCPs are more consistent, with (S)L18F appearing in 0.4% of all Delta entries and 0.08% of all Omicron entries and (S)A222V at 10% and 0.06% hits respectively; the appearance of both (S)L18F and (S)A222V together essentially stopped during the Delta wave with a total of 2,649 instances.

Considering that DCPs in this study were collected in the Winter of 2020, their longevity makes them highly noteworthy, particularly (S)D614G, and (S)N501Y with its relations HV69-70del, (S)Y144del (S)N501Y and (S)P681H; these mutations are highly relevant, as that they were carried over across successions of VOCs where many others were lost. A mutation in Omicron that may be of future interest to analyse with the same method for DCPs is (S)E484A - as a similar mutation to (S)E484K which was thought to enhance immune escape and is problematic to vaccine efficacy [65], [102]).

Regarding the cell culture research, there is only a single mutation that occurred in the *FFM7* isolate that can be found in 98% of all Omicron submissions. (N)P13L/(ORF9b)P10S was previously suggested in this study to have a role in reducing the IFN response and pathogenicity, which is a key characteristic of Omicron [145].

Currently in the UK, there has been a relaxation of COVID-19 policies as Omicron is considered a milder strain and total infections are beginning to wane again [146]. Although tracking of SARS-CoV-2 must continue as Omicron could potentially develop new mutations that increase lethality over time, or there could be the risk of recombination with a more deadly strain such as Delta which is still in circulation – a worst-case scenario being the combination of mutations that cause vaccine target escape with more pathogenic tendencies.

6.0 References

[1] P. Zhou *et al.*, "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature*, vol. 579, no. 7798, pp. 270–273, Mar. 2020, doi: 10.1038/s41586-020-2012-7.

[2] N. Zhu *et al.*, "A Novel Coronavirus from Patients with Pneumonia in China, 2019," *N. Engl. J. Med.*, vol. 382, no. 8, pp. 727–733, Feb. 2020, doi: 10.1056/nejmoa2001017.

[3] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *The Lancet Infectious Diseases*, vol. 20, no. 5. Lancet Publishing Group, pp. 533–534, May 01, 2020, doi: 10.1016/S1473-3099(20)30120-1.

[4] R. J. Mason, "Pathogenesis of COVID-19 from a cell biology perspective," *European Respiratory Journal*, vol. 55, no. 4. European Respiratory Society, Apr. 01, 2020, doi: 10.1183/13993003.00607-2020.

[5] Z. Wu and J. M. McGoogan, "Characteristics of and Important Lessons from the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases from the Chinese Center for Disease Control and Prevention," *JAMA - Journal of the American Medical Association*, vol. 323, no. 13. American Medical Association, pp. 1239–1242, Apr. 07, 2020, doi: 10.1001/jama.2020.2648.

[6] Q. Zhao *et al.*, "The impact of COPD and smoking history on the severity of COVID-19: A systemic review and meta-analysis," *J. Med. Virol.*, vol. 92, no. 10, pp. 1915–1921, Oct. 2020, doi: 10.1002/jmv.25889.

[7] L. Zhu *et al.*, "Association of Blood Glucose Control and Outcomes in Patients with COVID-19 and Pre-existing Type 2 Diabetes," *Cell Metab.*, vol. 31, no. 6, pp. 1068-1077.e3, Jun. 2020, doi:

10.1016/j.cmet.2020.04.021.

[8] W. Guan *et al.*, “Clinical Characteristics of Coronavirus Disease 2019 in China,” *N. Engl. J. Med.*, vol. 382, no. 18, pp. 1708–1720, Apr. 2020, doi: 10.1056/nejmoa2002032.

[9] N. Chen *et al.*, “Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study,” *Lancet*, vol. 395, no. 10223, pp. 507–513, Feb. 2020, doi: 10.1016/S0140-6736(20)30211-7.

[10] C. Huang *et al.*, “Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China,” *Lancet*, vol. 395, no. 10223, pp. 497–506, Feb. 2020, doi: 10.1016/S0140-6736(20)30183-5.

[11] D. Wang *et al.*, “Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China,” *JAMA - J. Am. Med. Assoc.*, vol. 323, no. 11, pp. 1061–1069, Mar. 2020, doi: 10.1001/jama.2020.1585.

[12] “WHO Director-General’s opening remarks at the media briefing on COVID-19 - 3 March 2020.” https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---3-march-2020 (accessed Nov. 09, 2020).

[13] D. D. Rajgor, M. H. Lee, S. Archuleta, N. Bagdasarian, and S. C. Quek, “The many estimates of the COVID-19 case fatality rate,” *The Lancet Infectious Diseases*, vol. 20, no. 7. Lancet Publishing Group, pp. 776–777, Jul. 01, 2020, doi: 10.1016/S1473-3099(20)30244-9.

[14] R. Lorenzo-Redondo *et al.*, “A Unique Clade of SARS-CoV-2 Viruses is Associated with Lower Viral Loads in Patient Upper Airways.,” *medRxiv Prepr. Serv. Heal. Sci.*, 2020, doi: 10.1101/2020.05.19.20107144.

[15] A. D. Iuliano *et al.*, “Estimates of global seasonal influenza-associated respiratory mortality: a

modelling study," *Lancet*, vol. 391, no. 10127, pp. 1285–1300, Mar. 2018, doi: 10.1016/S0140-6736(17)33293-2.

[16] "Up to 650 000 people die of respiratory diseases linked to seasonal flu each year." https://www.who.int/news/item/14-12-2017-up-to-650-000-people-die-of-respiratory-diseases-linked-to-seasonal-flu-each-year (accessed Nov. 09, 2020).

[17] V. M. Corman, D. Muth, D. Niemeyer, and C. Drosten, "Hosts and Sources of Endemic Human Coronaviruses," in *Advances in Virus Research*, vol. 100, Academic Press Inc., 2018, pp. 163–188.

[18] C. Drosten *et al.*, "Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome," *N. Engl. J. Med.*, vol. 348, no. 20, pp. 1967–1976, May 2003, doi: 10.1056/nejmoa030747.

[19] A. Moh Zaki, S. Van Boheemen, T. M. Bestebroer, A. D. M. E. Osterhaus, and R. A. M. Fouchier, "Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia," *NEJM.org. N Engl J Med*, vol. 367, pp. 1814–1834, 2012, doi: 10.1056/NEJMoa1211721.

[20] D. X. Liu, J. Q. Liang, and T. S. Fung, "Human Coronavirus-229E, -OC43, -NL63, and -HKU1 (Coronaviridae)," in *Encyclopedia of Virology*, Elsevier, 2021, pp. 428–440.

[21] J. W. LeDuc and M. A. Barry, "SARS, the First Pandemic of the 21st Century1," *Emerg. Infect. Dis.*, vol. 10, no. 11, pp. e26–e26, Nov. 2004, doi: 10.3201/eid1011.040797_02.

[22] "WHO EMRO | MERS situation update, January 2019 | MERS-CoV | Epidemic and pandemic diseases." http://www.emro.who.int/pandemic-epidemic-diseases/mers-cov/mers-situation-update-january-2019.html (accessed Nov. 05, 2020).

[23] A. Assiri *et al.*, "Hospital Outbreak of Middle East Respiratory Syndrome Coronavirus," *N. Engl.*

J. Med., vol. 369, no. 5, pp. 407–416, Aug. 2013, doi: 10.1056/nejmoa1306742.

[24] A. I. Zumla and Z. A. Memish, "Middle East respiratory syndrome coronavirus: Epidemic potential or a storm in a teacup?," *European Respiratory Journal*, vol. 43, no. 5. European Respiratory Society, pp. 1243–1248, May 01, 2014, doi: 10.1183/09031936.00227213.

[25] A. R. Fehr and S. Perlman, "Coronaviruses: An overview of their replication and pathogenesis," in *Coronaviruses: Methods and Protocols*, vol. 1282, Springer New York, 2015, pp. 1–23.

[26] Y. Gao *et al.*, "Structure of the RNA-dependent RNA polymerase from COVID-19 virus," *Science (80-.).*, vol. 368, no. 6492, pp. 779–782, May 2020, doi: 10.1126/science.abb7498.

[27] Y. M. Báez-Santos, S. E. St. John, and A. D. Mesecar, "The SARS-coronavirus papain-like protease: Structure, function and inhibition by designed antiviral compounds," *Antiviral Research*, vol. 115. Elsevier B.V., pp. 21–38, Mar. 01, 2015, doi: 10.1016/j.antiviral.2014.12.015.

[28] V. Mody *et al.*, "Identification of 3-chymotrypsin like protease (3CLPro) inhibitors as potential anti-SARS-CoV-2 agents," *Commun. Biol.*, vol. 4, no. 1, pp. 1–10, Dec. 2021, doi: 10.1038/s42003-020-01577-x.

[29] E. C. Smith and M. R. Denison, "Coronaviruses as DNA Wannabes: A New Model for the Regulation of RNA Virus Replication Fidelity," *PLoS Pathog.*, vol. 9, no. 12, p. e1003760, Dec. 2013, doi: 10.1371/journal.ppat.1003760.

[30] J. Gribble *et al.*, "The coronavirus proofreading exoribonuclease mediates extensive viral recombination," *PLoS Pathog.*, vol. 17, no. 1, p. e1009226, Jan. 2021, doi: 10.1371/journal.ppat.1009226.

[31] D. Kim, J. Y. Lee, J. S. Yang, J. W. Kim, V. N. Kim, and H. Chang, "The Architecture of SARS-CoV-2 Transcriptome," *Cell*, vol. 181, no. 4, pp. 914-921.e10, May 2020, doi:

10.1016/j.cell.2020.04.011.

[32] C. J. Michel, C. Mayer, O. Poch, and J. D. Thompson, “Characterization of accessory genes in coronavirus genomes,” *Virol. J.*, vol. 17, no. 1, Aug. 2020, doi: 10.1186/s12985-020-01402-1.

[33] A. Wu *et al.*, “Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China,” *Cell Host Microbe*, vol. 27, no. 3, pp. 325–328, Mar. 2020, doi: 10.1016/j.chom.2020.02.001.

[34] D. Bestle *et al.*, “TMPRSS2 and furin are both essential for proteolytic activation of SARS-CoV-2 in human airway cells,” *Life Sci. Alliance*, vol. 3, no. 9, Sep. 2020, doi: 10.26508/LSA.202000786.

[35] A. C. Walls, Y. J. Park, M. A. Tortorici, A. Wall, A. T. McGuire, and D. Veesler, “Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein,” *Cell*, vol. 181, no. 2, pp. 281-292.e6, Apr. 2020, doi: 10.1016/j.cell.2020.02.058.

[36] M. Letko, A. Marzi, and V. Munster, “Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses,” *Nat. Microbiol.*, vol. 5, no. 4, pp. 562–569, 2020, doi: 10.1038/s41564-020-0688-y.

[37] K. E. Follis, J. York, and J. H. Nunberg, “Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell-cell fusion but does not affect virion entry,” *Virology*, vol. 350, no. 2, pp. 358–369, Jul. 2006, doi: 10.1016/j.virol.2006.02.003.

[38] S. Belouzard, V. C. Chu, and G. R. Whittaker, “Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 14, pp. 5871–5876, Apr. 2009, doi: 10.1073/pnas.0809524106.

[39] G. Khelashvili, A. Plante, M. Doktorova, and H. Weinstein, “Ca2+-dependent mechanism of membrane insertion and destabilization by the SARS-CoV-2 fusion peptide,” *Biophys. J.*, vol. 120,

no. 6, pp. 1105–1119, Mar. 2021, doi: 10.1016/J.BPJ.2021.02.023.

[40] M. A. Tortorici and D. Veesler, “Structural insights into coronavirus entry,” *Adv. Virus Res.*, vol. 105, p. 93, Jan. 2019, doi: 10.1016/BS.AIVIR.2019.08.002.

[41] S. Liu *et al.*, “Interaction between heptad repeat 1 and 2 regions in spike protein of SARS-associated coronavirus: implications for virus fusogenic mechanism and identification of fusion inhibitors,” *Lancet*, vol. 363, no. 9413, pp. 938–947, Mar. 2004, doi: 10.1016/S0140-6736(04)15788-7.

[42] Y. Zhu, D. Yu, H. Yan, H. Chong, and Y. He, “Design of Potent Membrane Fusion Inhibitors against SARS-CoV-2, an Emerging Coronavirus with High Fusogenic Activity,” *J. Virol.*, vol. 94, no. 14, Jul. 2020, doi: 10.1128/JVI.00635-20.

[43] Y. Cai *et al.*, “Distinct conformational states of SARS-CoV-2 spike protein,” *Science (80-.).*, vol. 369, no. 6511, pp. 1586–1592, 2020, doi: 10.1126/science.abd4251.

[44] G. Duart, M. J. García-Murria, B. Grau, J. M. Acosta-Cáceres, L. Martínez-Gil, and I. Mingarro, “SARS-CoV-2 envelope protein topology in eukaryotic membranes: SARS-CoV-2 E protein topology,” *Open Biol.*, vol. 10, no. 9, Sep. 2020, doi: 10.1098/rsob.200209.

[45] L. Wilson, C. Mckinlay, P. Gage, and G. Ewart, “SARS coronavirus E protein forms cation-selective ion channels,” *Virology*, vol. 330, no. 1, pp. 322–331, Dec. 2004, doi: 10.1016/j.virol.2004.09.033.

[46] B. Boson *et al.*, “The SARS-CoV-2 envelope and membrane proteins modulate maturation and retention of the spike protein, allowing assembly of virus-like particles,” *J. Biol. Chem.*, vol. 296, Jan. 2021, doi: 10.1074/jbc.RA120.016175.

[47] Y. Cong *et al.*, “Nucleocapsid Protein Recruitment to Replication-Transcription Complexes Plays

a Crucial Role in Coronaviral Life Cycle," *J. Virol.*, vol. 94, no. 4, Jan. 2020, doi: 10.1128/JVI.01925-19.

[48] L. X, P. J, T. J, and G. D, "SARS-CoV nucleocapsid protein antagonizes IFN-β response by targeting initial step of IFN-β induction pathway, and its C-terminal region is critical for the antagonism," *Virus Genes*, vol. 42, no. 1, pp. 37–45, Feb. 2011, doi: 10.1007/S11262-010-0544-X.

[49] M. Surjit, B. Liu, S. Jameel, V. T. K. Chow, and S. K. Lal, "The SARS coronavirus nucleocapsid protein induces actin reorganization and apoptosis in COS-1 cells in the absence of growth factors," *Biochem. J.*, vol. 383, no. 1, pp. 13–18, Oct. 2004, doi: 10.1042/BJ20040984.

[50] B. Korber *et al.*, "Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus," *Cell*, vol. 182, no. 4, pp. 812-827.e19, Aug. 2020, doi: 10.1016/j.cell.2020.06.043.

[51] L. Zhang *et al.*, "The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity," *bioRxiv Prepr. Serv. Biol.*, 2020, doi: 10.1101/2020.06.12.148726.

[52] L. Yurkovetskiy *et al.*, "SARS-CoV-2 Spike protein variant D614G increases infectivity and retains sensitivity to antibodies that target the receptor binding domain.," *bioRxiv Prepr. Serv. Biol.*, vol. 13, p. 14, 2020, doi: 10.1101/2020.07.04.187757.

[53] P.-Y. Shi *et al.*, "Spike mutation D614G alters SARS-CoV-2 fitness and neutralization susceptibility.," *Res. Sq.*, 2020, doi: 10.21203/rs.3.rs-70482/v1.

[54] Y. J. Hou *et al.*, "SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo," *Science (80-.).*, p. eabe8499, Nov. 2020, doi: 10.1126/science.abe8499.

[55] M. Becerra-Flores and T. Cardozo, "SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate," *Int. J. Clin. Pract.*, vol. 74, no. 8, pp. 0–2, 2020, doi: 10.1111/ijcp.13525.

[56] D. Weissman *et al.*, “D614G Spike Mutation Increases SARS CoV-2 Susceptibility to Neutralization.,” *medRxiv*, p. 2020.07.22.20159905, Sep. 2020, doi: 10.1101/2020.07.22.20159905.

[57] R. Mansbach, S. Chakraborty, K. Nguyen, D. Montefiori, and B. Korber, “The SARS-CoV-2 Spike Variant D614G Favors an Open Conformational State,” *bioRxiv Prepr. Serv. Biol.*, p. 2020.07.26.219741, Jul. 2020, doi: 10.1101/2020.07.26.219741.

[58] Y. Shu and J. McCauley, “GISAID: Global initiative on sharing all influenza data – from vision to reality,” *Eurosurveillance*, vol. 22, no. 13, p. 30494, Mar. 2017, doi: 10.2807/1560-7917.ES.2017.22.13.30494.

[59] WHO, “Tracking SARS-CoV-2 variants,” *Who*, 2021. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/ (accessed Oct. 02, 2021).

[60] A. Rambaut *et al.*, “Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology - Virological,” *Virological.org*, 2020. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563 (accessed Feb. 04, 2021).

[61] Meera Chand *et al.*, “Investigation of novel SARS-COV-2 variant: Variant of Concern 202012/01,” Dec. 2020.

[62] E. Volz *et al.*, “Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data,” *medRxiv*, p. 2020.12.30.20249034, 2021, [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.12.30.20249034v2%0Ahttps://www.medrxiv.org/content/10.1101/2020.12.30.20249034v2.abstract.

[63] F. Grabowski, G. Preibisch, M. Kochańczyk, and T. Lipniacki, “SARS-CoV-2 Variant Under

Investigation 202012/01 has more than twofold replicative advantage," doi: 10.1101/2020.12.28.20248906.

[64] R. Challen, E. Brooks-Pollock, J. M. Read, L. Dyson, K. Tsaneva-Atanasova, and L. Danon, "Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study," *BMJ*, vol. 372, p. n579, Mar. 2021, doi: 10.1136/bmj.n579.

[65] C. K. Wibmer *et al.*, "SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma," *Nat. Med.*, pp. 1–4, Mar. 2021, doi: 10.1038/s41591-021-01285-x.

[66] P. Wang *et al.*, "Antibody Resistance of SARS-CoV-2 Variants B.1.351 and B.1.1.7," *Nature*, 2021, doi: 10.1038/s41586-021-03398-2.

[67] T. N. Starr *et al.*, "Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding," *bioRxiv*. bioRxiv, Jun. 17, 2020, doi: 10.1101/2020.06.17.157982.

[68] J. Zahradník *et al.*, "SARS-CoV-2 RBD in vitro evolution follows contagious mutation spread, yet generates an able infection inhibitor," *bioRxiv*. Cold Spring Harbor Laboratory, p. 2021.01.06.425392, Jan. 08, 2021, doi: 10.1101/2021.01.06.425392.

[69] F. Tian *et al.*, "Mutation N501Y in RBD of Spike Protein Strengthens the Inter-action between COVID-19 and its Receptor ACE2," *bioRxiv*, p. 2021.02.14.431117, Feb. 2021, doi: 10.1101/2021.02.14.431117.

[70] E. Kudo *et al.*, "Detection of SARS-CoV-2 RNA by multiplex RT-qPCR," *bioRxiv*. bioRxiv, Jun. 17, 2020, doi: 10.1101/2020.06.16.155887.

[71] S. Kemp *et al.*, "Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion ΔH69/ΔV70," *bioRxiv*, p. 2020.12.14.422555, Dec. 2020, doi: 10.1101/2020.12.14.422555.

[72] M. Pappalardo, M. Julia, M. J. Howard, J. S. Rossman, M. Michaelis, and M. N. Wass, "Conserved differences in protein sequence determine the human pathogenicity of Ebolaviruses," *Sci. Rep.*, vol. 6, no. 1, pp. 1–11, Mar. 2016, doi: 10.1038/srep23743.

[73] D. Bojkova *et al.*, "SARS-CoV-2 and SARS-CoV differ in their cell tropism and drug sensitivity profiles," *bioRxiv*, p. 2020.04.03.024257, Apr. 2020, doi: 10.1101/2020.04.03.024257.

[74] J. A. Capra and M. Singh, "Predicting functionally important residues from sequence conservation," *Bioinformatics*, vol. 23, no. 15, pp. 1875–1882, Aug. 2007, doi: 10.1093/bioinformatics/btm270.

[75] F. Sievers *et al.*, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega," *Mol. Syst. Biol.*, vol. 7, 2011, doi: 10.1038/msb.2011.75.

[76] F. Sievers and D. G. Higgins, "Clustal Omega for making accurate alignments of many protein sequences," *Protein Sci.*, vol. 27, no. 1, pp. 135–145, Jan. 2018, doi: 10.1002/pro.3290.

[77] J. Söding, "Protein homology detection by HMM-HMM comparison," *Bioinformatics*, vol. 21, no. 7, pp. 951–960, Apr. 2005, doi: 10.1093/bioinformatics/bti125.

[78] E. Jurrus *et al.*, "Improvements to the APBS biomolecular solvation software suite," *Protein Sci.*, vol. 27, no. 1, pp. 112–128, Jan. 2018, doi: 10.1002/PRO.3280.

[79] C. H. Rodrigues, D. E. Pires, D. B. Ascher, C. B. David Ascher, and unimelb eduau, "DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations," 2020, doi: 10.1002/pro.3942.

[80] K. Lewandowski *et al.*, "Metagenomic nanopore sequencing of influenza virus direct from clinical respiratory samples," *J. Clin. Microbiol.*, vol. 58, no. 1, Jan. 2020, doi: 10.1128/JCM.00963-19.

[81] D. Wrapp *et al.*, "Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation,"

2019. Accessed: Oct. 30, 2020. [Online]. Available: http://science.sciencemag.org/.

[82] L. Hanke *et al.*, “An alpaca nanobody neutralizes SARS-CoV-2 by blocking receptor interaction,” *Nat. Commun.*, vol. 11, no. 1, pp. 1–9, Dec. 2020, doi: 10.1038/s41467-020-18174-5.

[83] F. Grabowski, M. Kochańczyk, and T. Lipniacki, “L18F substrain of SARS-CoV-2 VOC-202012/01 is rapidly spreading in England,” *medRxiv*, p. 2021.02.07.21251262, Feb. 2021, doi: 10.1101/2021.02.07.21251262.

[84] M. McCallum *et al.*, “N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2,” *Cell*, vol. 184, no. 9, pp. 2332-2347.e16, Apr. 2021, doi: 10.1016/j.cell.2021.03.028.

[85] S. Cele *et al.*, “Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma,” *Nature*, vol. 593, no. 7857, pp. 142–146, May 2021, doi: 10.1038/s41586-021-03471-w.

[86] B. zhong Zhang *et al.*, “Mining of epitopes on spike protein of SARS-CoV-2 from COVID-19 patients,” *Cell Research*, vol. 30, no. 8. Springer Nature, pp. 702–704, Aug. 01, 2020, doi: 10.1038/s41422-020-0366-x.

[87] B. Korber *et al.*, “Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2,” *bioRxiv*, p. 2020.04.29.069054, Apr. 2020, doi: 10.1101/2020.04.29.069054.

[88] S. Vilar and D. G. Isom, “One Year of SARS-CoV-2: How Much Has the Virus Changed?,” *bioRxiv*, p. 2020.12.16.423071, Dec. 2020, doi: 10.1101/2020.12.16.423071.

[89] S. Xia *et al.*, “The role of furin cleavage site in SARS-CoV-2 spike protein-mediated membrane fusion in the presence or absence of trypsin,” *Signal Transduction and Targeted Therapy*, vol. 5, no. 1. Springer Nature, pp. 1–3, Dec. 01, 2020, doi: 10.1038/s41392-020-0184-0.

[90] M. Hoffmann, H. Kleine-Weber, and S. Pöhlmann, “A Multibasic Cleavage Site in the Spike

Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells," *Mol. Cell*, vol. 78, no. 4, pp. 779-784.e5, May 2020, doi: 10.1016/j.molcel.2020.04.022.

[91] M. Hoffmann *et al.*, "SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor," *Cell*, vol. 181, no. 2, pp. 271-280.e8, Apr. 2020, doi: 10.1016/j.cell.2020.02.052.

[92] R. Yan, Y. Zhang, Y. Li, L. Xia, Y. Guo, and Q. Zhou, "Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2," *Science (80-.).*, vol. 367, no. 6485, pp. 1444–1448, Mar. 2020, doi: 10.1126/science.abb2762.

[93] T.-J. Yang *et al.*, "Impacts on the structure-function relationship of SARS-CoV-2 spike by B.1.1.7 mutations," *bioRxiv*, p. 2021.05.11.443686, May 2021, doi: 10.1101/2021.05.11.443686.

[94] J. L. Daly *et al.*, "Neuropilin-1 is a host factor for SARS-CoV-2 infection," *Science (80-.).*, vol. 370, no. 6518, pp. 861–865, Nov. 2020, doi: 10.1126/science.abd3072.

[95] M. Calcagnile, P. Forgez, M. Alifano, and P. Alifano, "The lethal triad: SARS-CoV-2 Spike, ACE2 and TMPRSS2. Mutations in host and pathogen may affect the course of pandemic," *bioRxiv*, p. 2021.01.12.426365, Jan. 2021, doi: 10.1101/2021.01.12.426365.

[96] P. E. Oluniyi, C. Happi, C. Ihekweazu, J. Nkengasong, and I. Olawoye, "Detection of SARS-CoV-2 P681H Spike Protein Variant in Nigeria," *Virological.org*, Dec. 2020. https://virological.org/t/detection-of-sars-cov-2-p681h-spike-protein-variant-in-nigeria/567 (accessed Jun. 07, 2021).

[97] N. S. Zuckerman *et al.*, "A unique SARS-CoV-2 spike protein P681H strain detected in Israel Israel National Consortium for SARS-CoV-2 sequencing," *medRxiv*, p. 2021.03.25.21253908, Mar. 2021, doi: 10.1101/2021.03.25.21253908.

[98] D. P. Maison, L. L. Ching, C. M. Shikuma, and V. R. Nerurkar, “Genetic Characteristics and Phylogeny of 969-bp S Gene Sequence of SARS-CoV-2 from Hawai’i Reveals the Worldwide Emerging P681H Mutation,” *Hawai’i J. Heal. Soc. Welf.*, vol. 80, no. 3, pp. 52–61, Mar. 2021, doi: 10.1101/2021.01.06.425497.

[99] E. Lasek-Nesselquist, J. Pata, E. Schneider, and K. S. George, “A tale of three SARS-CoV-2 variants with independently acquired P681H mutations in New York State,” *medRxiv*, p. 2021.03.10.21253285, Mar. 2021, doi: 10.1101/2021.03.10.21253285.

[100] B. Lubinski, T. Tang, S. Daniel, J. A. Jaimes, and G. R. Whittaker, “Functional evaluation of proteolytic activation for the SARS-CoV-2 variant B.1.1.7: role of the P681H mutation.,” *bioRxiv Prepr. Serv. Biol.*, p. 2021.04.06.438731, Apr. 2021, doi: 10.1101/2021.04.06.438731.

[101] V. V. Edara *et al.*, “Infection and mRNA-1273 vaccine antibodies neutralize SARS-CoV-2 UK variant.,” *medRxiv Prepr. Serv. Heal. Sci.*, p. 2021.02.02.21250799, Feb. 2021, doi: 10.1101/2021.02.02.21250799.

[102] T. Tada *et al.*, “Neutralization of viruses with European, South African, and United States SARS-CoV-2 variant spike proteins by convalescent sera and BNT162b2 mRNA vaccine-elicited antibodies.,” *bioRxiv Prepr. Serv. Biol.*, p. 2021.02.05.430003, Feb. 2021, doi: 10.1101/2021.02.05.430003.

[103] B. Turoňová *et al.*, “In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges,” *Science (80-.).*, vol. 370, no. 6513, pp. 203–208, Oct. 2020, doi: 10.1126/science.abd5223.

[104] D. A. Ostrov, “Structural Consequences of Variation in SARS-CoV-2 B.1.1.7,” *J. Cell. Immunol.*, vol. 3, no. 2, p. 103, Apr. 2021, doi: 10.33696/immunology.3.085.

[105] E. C. Thomson *et al.*, “Circulating SARS-CoV-2 spike N439K variants maintain fitness while

evading antibody-mediated immunity," *Cell*, vol. 184, no. 5, pp. 1171-1187.e20, Mar. 2021, doi: 10.1016/j.cell.2021.01.037.

[106] B. Brejová *et al.*, "B.1.258Δ, a SARS-CoV-2 variant with ΔH69/ΔV70 in the Spike protein circulating in the Czech Republic and Slovakia," *arXiv Prepr. arXiv2102.04689*, 2021, Accessed: Jun. 23, 2021. [Online]. Available: https://virological.org/t/b-1-258-a-sars-cov-2-variant-with-h69-v70-in-the-spike-protein-circulating-in-the-czech-republic-and-slovakia/613.

[107] M. Hoffmann *et al.*, "SARS-CoV-2 mutations acquired in mink reduce antibody-mediated neutralization ll SARS-CoV-2 mutations acquired in mink reduce antibody-mediated neutralization," 2021, doi: 10.1016/j.celrep.2021.109017.

[108] K. R. McCarthy *et al.*, "Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape," *Science (80-.).*, vol. 371, no. 6534, pp. 1139–1142, Mar. 2021, doi: 10.1126/science.abf6950.

[109] M. Mukherjee and S. Goswami, "Global cataloguing of variations in untranslated regions of viral genome and prediction of key host RNA binding protein-microRNA interactions modulating genome stability in SARS-CoV-2," *PLoS One*, vol. 15, no. 8, p. e0237559, Aug. 2020, doi: 10.1371/JOURNAL.PONE.0237559.

[110] J. Zhang *et al.*, "Multi-site co-mutations and 5'UTR CpG immunity escape drive the evolution of SARS-CoV-2," *bioRxiv*, p. 2020.07.21.213405, Jul. 2020, doi: 10.1101/2020.07.21.213405.

[111] Y. Liang *et al.*, "Comprehensive Antibody Epitope Mapping of the Nucleocapsid Protein of Severe Acute Respiratory Syndrome (SARS) Coronavirus: Insight into the Humoral Immunity of SARS," *Clin. Chem.*, vol. 51, no. 8, pp. 1382–1396, Aug. 2005, doi: 10.1373/CLINCHEM.2005.051045.

[112] L. P. P. Patro, C. Sathyaseelan, P. P. Uttamrao, and T. Rathinavelan, "Global variation in the

SARS-CoV-2 proteome reveals the mutational hotspots in the drug and vaccine candidates," *bioRxiv*, p. 2020.07.31.230987, Aug. 2020, doi: 10.1101/2020.07.31.230987.

[113] L. P. P. Patro, C. Sathyaseelan, P. P. Uttamrao, and T. Rathinavelan, "Global variation in SARS-CoV-2 proteome and its implication in pre-lockdown emergence and dissemination of 5 dominant SARS-CoV-2 clades," *Infect. Genet. Evol.*, vol. 93, p. 104973, Sep. 2021, doi: 10.1016/J.MEEGID.2021.104973.

[114] L. Guruprasad, "Human SARS CoV-2 spike protein mutations," *Proteins Struct. Funct. Bioinforma.*, vol. 89, no. 5, pp. 569–576, May 2021, doi: 10.1002/PROT.26042.

[115] S. Kang *et al.*, "Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites," *bioRxiv*, p. 2020.03.06.977876, Mar. 2020, doi: 10.1101/2020.03.06.977876.

[116] A. . C. for S. G. of I. D. (CSGID) Chang, C.; Michalska, K.; Jedrzejczak, R.; Maltseva, N.; Endres, M.; Godzik, A.; Kim, Y.; Joachimiak, "RCSB PDB - 6VYO: Crystal structure of RNA binding domain of nucleocapsid phosphoprotein from SARS coronavirus 2," 2020. https://www.rcsb.org/structure/6VYO (accessed Aug. 02, 2021).

[117] E. C. Rouchka, J. H. Chariker, and D. Chung, "Variant analysis of 1,040 SARS-CoV-2 genomes," *PLoS One*, vol. 15, no. 11, p. e0241535, Nov. 2020, doi: 10.1371/JOURNAL.PONE.0241535.

[118] I. Saha, N. Ghosh, D. Maity, N. Sharma, J. P. Sarkar, and K. Mitra, "Genome-wide analysis of Indian SARS-CoV-2 genomes for the identification of genetic mutation and SNP," *Infect. Genet. Evol.*, vol. 85, p. 104457, Nov. 2020, doi: 10.1016/J.MEEGID.2020.104457.

[119] Gunadi *et al.*, "Molecular epidemiology of SARS-CoV-2 isolated from COVID-19 family clusters," *BMC Med. Genomics 2021 141*, vol. 14, no. 1, pp. 1–14, Jun. 2021, doi: 10.1186/S12920-021-00990-3.

[120] A. Oulas *et al.*, “Generalized linear models provide a measure of virulence for specific mutations in SARS-CoV-2 strains,” *PLoS One*, vol. 16, no. 1, p. e0238665, Jan. 2021, doi: 10.1371/JOURNAL.PONE.0238665.

[121] H. Jiang *et al.*, “SARS-CoV-2 Orf9b suppresses type I interferon responses by targeting TOM70,” *Cell. Mol. Immunol. 2020 179*, vol. 17, no. 9, pp. 998–1000, Jul. 2020, doi: 10.1038/s41423-020-0514-8.

[122] L. Cavallo and R. Oliva, “D936Y and Other Mutations in the Fusion Core of the SARS-Cov-2 Spike Protein Heptad Repeat 1 Undermine the Post-Fusion Assembly,” *bioRxiv*, p. 2020.06.08.140152, Jun. 2020, doi: 10.1101/2020.06.08.140152.

[123] R. Y. Aljindan, A. M. Al-Subaie, A. I. Al-Ohali, T. Kumar D, G. P. Doss C, and B. Kamaraj, “Investigation of nonsynonymous mutations in the spike protein of SARS-CoV-2 and its interaction with the ACE2 receptor by molecular docking and MM/GBSA approach,” *Comput. Biol. Med.*, vol. 135, p. 104654, Aug. 2021, doi: 10.1016/J.COMPBIOMED.2021.104654.

[124] L. K. Clark, T. J. Green, and C. M. Petit, “Structure of Nonstructural Protein 1 from SARS-CoV-2,” *J. Virol.*, vol. 95, no. 4, Jan. 2021, doi: 10.1128/JVI.02019-20.

[125] J. wen Lin *et al.*, “Genomic monitoring of SARS-CoV-2 uncovers an Nsp1 deletion variant that modulates type I interferon response,” *Cell Host Microbe*, vol. 29, no. 3, pp. 489-502.e8, Mar. 2021, doi: 10.1016/J.CHOM.2021.01.015.

[126] A. R. N. Zekri *et al.*, “Genome sequencing of SARS-CoV-2 in a cohort of Egyptian patients revealed mutation hotspots that are related to clinical outcomes,” *Biochim. Biophys. Acta - Mol. Basis Dis.*, vol. 1867, no. 8, p. 166154, Aug. 2021, doi: 10.1016/J.BBADIS.2021.166154.

[127] B. Liu, W. Shi, and Y. Yang, “RCSB PDB - 6XQB: SARS-CoV-2 RdRp/RNA complex,” 2020. https://www.rcsb.org/structure/6XQB (accessed Aug. 16, 2021).

[128] M. Bianchi, A. Borsetti, M. Ciccozzi, and S. Pascarella, “SARS-Cov-2 ORF3a: Mutability and function,” *Int. J. Biol. Macromol.*, vol. 170, pp. 820–826, Feb. 2021, doi: 10.1016/J.IJBIOMAC.2020.12.142.

[129] E. Issa, G. Merhi, B. Panossian, T. Salloum, and S. Tokajian, “SARS-CoV-2 and ORF3a: Nonsynonymous Mutations, Functional Domains, and Viral Pathogenesis,” *mSystems*, vol. 5, no. 3, Jun. 2020, doi: 10.1128/MSYSTEMS.00266-20.

[130] W. M. Chan *et al.*, “Phylogenomic analysis of COVID-19 summer and winter outbreaks in Hong Kong: An observational study,” *Lancet Reg. Heal. - West. Pacific*, vol. 10, p. 100130, May 2021, doi: 10.1016/J.LANWPC.2021.100130.

[131] L. Yan *et al.*, “Architecture of a SARS-CoV-2 mini replication and transcription complex,” *Nat. Commun.*, vol. 11, no. 1, Dec. 2020, doi: 10.1038/S41467-020-19770-1.

[132] F. Pereira, “SARS-CoV-2 variants combining spike mutations and the absence of ORF8 may be more transmissible and require close monitoring,” *Biochem. Biophys. Res. Commun.*, vol. 550, pp. 8–14, Apr. 2021, doi: 10.1016/J.BBRC.2021.02.080.

[133] Y. Zhang *et al.*, “The ORF8 Protein of SARS-CoV-2 Mediates Immune Evasion through Potently Downregulating MHC-I,” *bioRxiv*, p. 2020.05.24.111823, May 2020, doi: 10.1101/2020.05.24.111823.

[134] B. E. Young *et al.*, “Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study,” *Lancet*, vol. 396, no. 10251, pp. 603–611, Aug. 2020, doi: 10.1016/S0140-6736(20)31757-8.

[135] J. M. Velasco *et al.*, “Coding-Complete Genome Sequences of 11 SARS-CoV-2 B.1.1.7 and B.1.351 Variants from Metro Manila, Philippines,” *Microbiol. Resour. Announc.*, vol. 10, no. 28, Jul. 2021, doi: 10.1128/MRA.00498-21.

[136] S. N. Slavov *et al.*, “Genomic monitoring unveil the early detection of the SARS-CoV-2 B.1.351 lineage (20H/501Y.V2) in Brazil,” *medRxiv*, p. 2021.03.30.21254591, Apr. 2021, doi: 10.1101/2021.03.30.21254591.

[137] C. Del Vecchio *et al.*, “Emergence of N antigen SARS-CoV-2 genetic variants escaping detection of antigenic tests,” *medRxiv*, p. 2021.03.25.21253802, Mar. 2021, doi: 10.1101/2021.03.25.21253802.

[138] S. Calvignac-Spencer *et al.*, “Emergence of SARS-CoV-2 lineage A.27 in Germany, expressing viral spike proteins with several amino acid replacements of interest, including L18F, L452R, and N501Y in the absence of D614G - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology - Virolog.” https://virological.org/t/emergence-of-sars-cov-2-lineage-a-27-in-germany-expressing-viral-spike-proteins-with-several-amino-acid-replacements-of-interest-including-l18f-l452r-and-n501y-in-the-absence-of-d614g/693 (accessed Aug. 25, 2021).

[139] E. B. Hodcroft *et al.*, “Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020,” doi: 10.1101/2020.10.25.20219063.

[140] A. D. Andres *et al.*, “SARS-CoV-2 ORF9c Is a Membrane-Associated Protein that Suppresses Antiviral Responses in Cells,” *bioRxiv*, p. 2020.08.18.256776, Aug. 2020, doi: 10.1101/2020.08.18.256776.

[141] F. Amanat and F. Krammer, “SARS-CoV-2 Vaccines: Status Report,” *Immunity*, vol. 52, no. 4, pp. 583–589, Apr. 2020, doi: 10.1016/J.IMMUNI.2020.03.007.

[142] S. M. Hamed, W. F. Elkhatib, A. S. Khairalla, and A. M. Noreddin, “Global dynamics of SARS-CoV-2 clades and their relation to COVID-19 epidemiology,” *Sci. Reports 2021 111*, vol. 11, no. 1, pp. 1–8, Apr. 2021, doi: 10.1038/s41598-021-87713-x.

[143] C. Farkas, A. Mella, and J. J. Haigh, “Large-scale population analysis of SARS-CoV-2 whole

genome sequences reveals host-mediated viral evolution with emergence of mutations in the viral Spike protein associated with elevated mortality rates," *medRxiv*, p. 2020.10.23.20218511, Mar. 2021, doi: 10.1101/2020.10.23.20218511.

[144] B. Bartolini *et al.*, "The newly introduced SARS-CoV-2 variant A222V is rapidly spreading in Lazio region, Italy Authors," *medRxiv*, p. 2020.11.28.20237016, Nov. 2020, doi: 10.1101/2020.11.28.20237016.

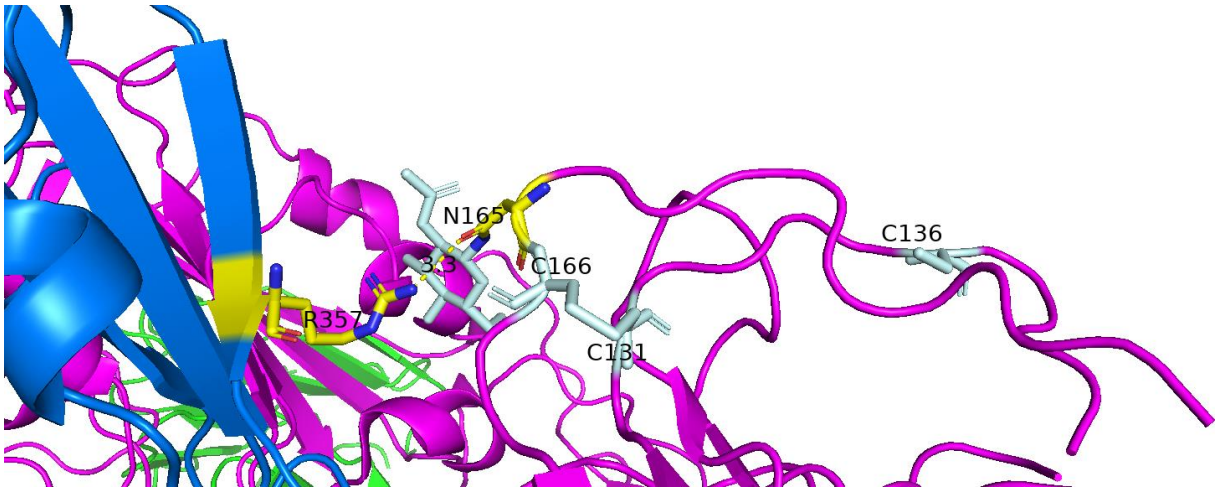
[145] D. Bojkova, M. Widera, S. Ciesek, M. N. Wass, M. Michaelis, and J. Cinatl, "Reduced interferon antagonism but similar drug sensitivity in Omicron variant compared to Delta variant of SARS-CoV-2 isolates," *Cell Res.*, vol. 32, no. 3, pp. 319–321, 2022, doi: 10.1038/s41422-022-00619-9.

[146] "COVID-19 Response: Living with COVID-19 - GOV.UK." https://www.gov.uk/government/publications/covid-19-response-living-with-covid-19 (accessed Mar. 12, 2022).

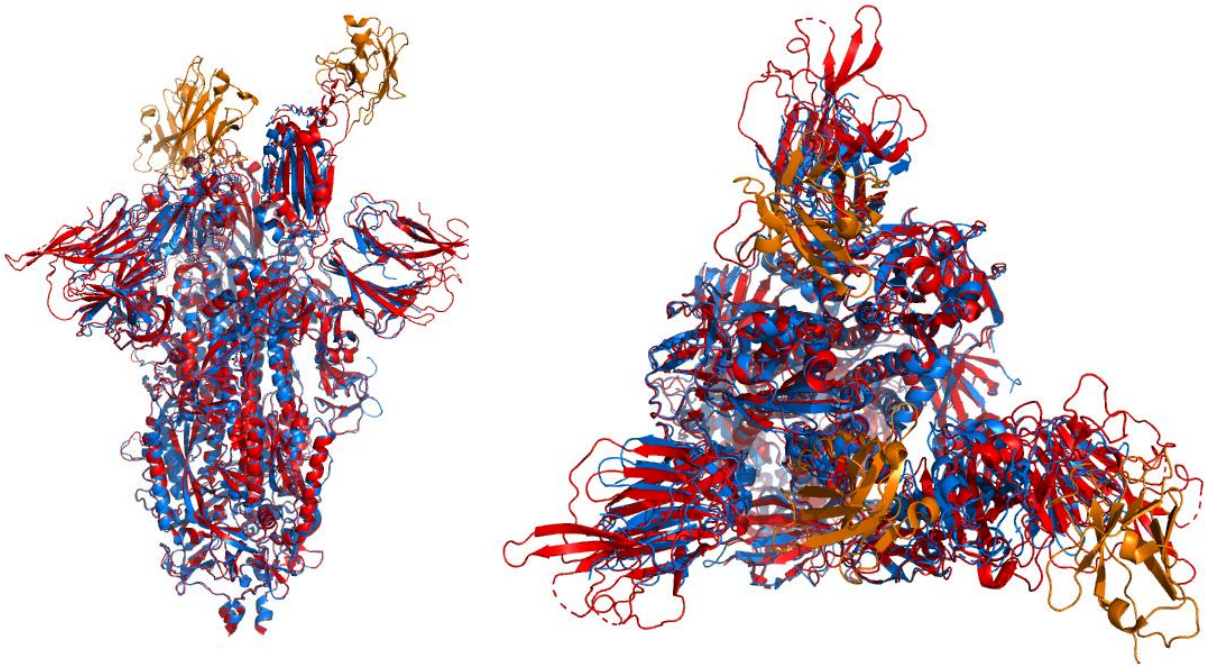
7.0 Supplementary Figures

Protein	Mutation	PDB Structure	Dynamut2 ΔΔG (kcal/mol)
Spike	**L18F**	6XR8	A-C: -0.55, -0.62, -0.6
		6ZXN 6ZXN (-Ty1)	A-C: -0.52, -0.72, -0.64 A-C: -0.52, -0.70, -0.64
	T716I	6ZXN 6ZXN (-Ty1)	A-C: -0.02, -0.04, -0.44 A-C: -0.2, -0.1, -0.44
	F157S	6XR8	A-C: -2.07, -2.09, -2.09
		6VSB	A-C: -0.36, -0.11, -1.10
	A222V	6XR8	A-C: 0.33, 0.06, 0.31
		6VSB	A-C: 0.24, 0.3, 0.28
	N501Y	6XR8	A-C: 0.0, -0.02, -0.07
		6M17	E-F: -0.42, -0.44
	A570D	6XR8	A-C: -1.28, -1.27, -1.29
		6VSB	A-C: -0.28, -0.69, -0.53
	D614G	6XR8	A-C: -0.51, -0.78, -0.63
		6VSB	A-C: -0.3, -0.23, -0.29
	T716I	6XR8	A-C: -0.04, 0.38, 0.42
		6VSB	A-C: 0.54, 0.45, 0.55
	S982A	6XR8	A-C: -0.56, -0.57, -0.61
		6VSB	A-C: -0.15, 0.00, -0.37
	D1118H	6XR8	A-C: -0.37, -0.35, -0.35
		6VSB	A-C: -0.37, -0.37, -0.29
NSP13	**T481M**	7CXM	E-F: 0.31, 0.35
Nucleoprotein	**N126Y**	6VYO	A-D: -0.35, -0.34, -0.34, -0.31

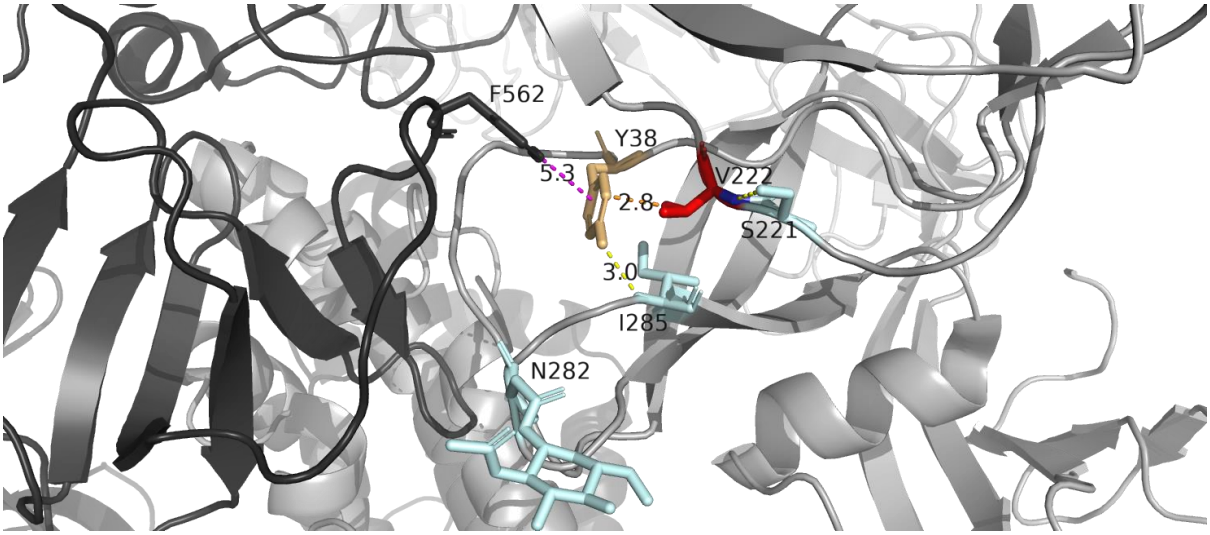
(S1) Table 7 – Dynamut2 ΔΔG values (kcal/mol) for discussed SARS-CoV protein mutations*, 6XR8 refers to the "closed" S protein conformation PDB, 6ZXN/6VSB refer to "one-up"S protein conformation and 6M17 is an S protein RBD:ACE2 structure. 6ZXN has been inputted twice with (-Ty1) referring to removal of nano-Abs. Values in blue are stabilising whereas red values are destabilising.*



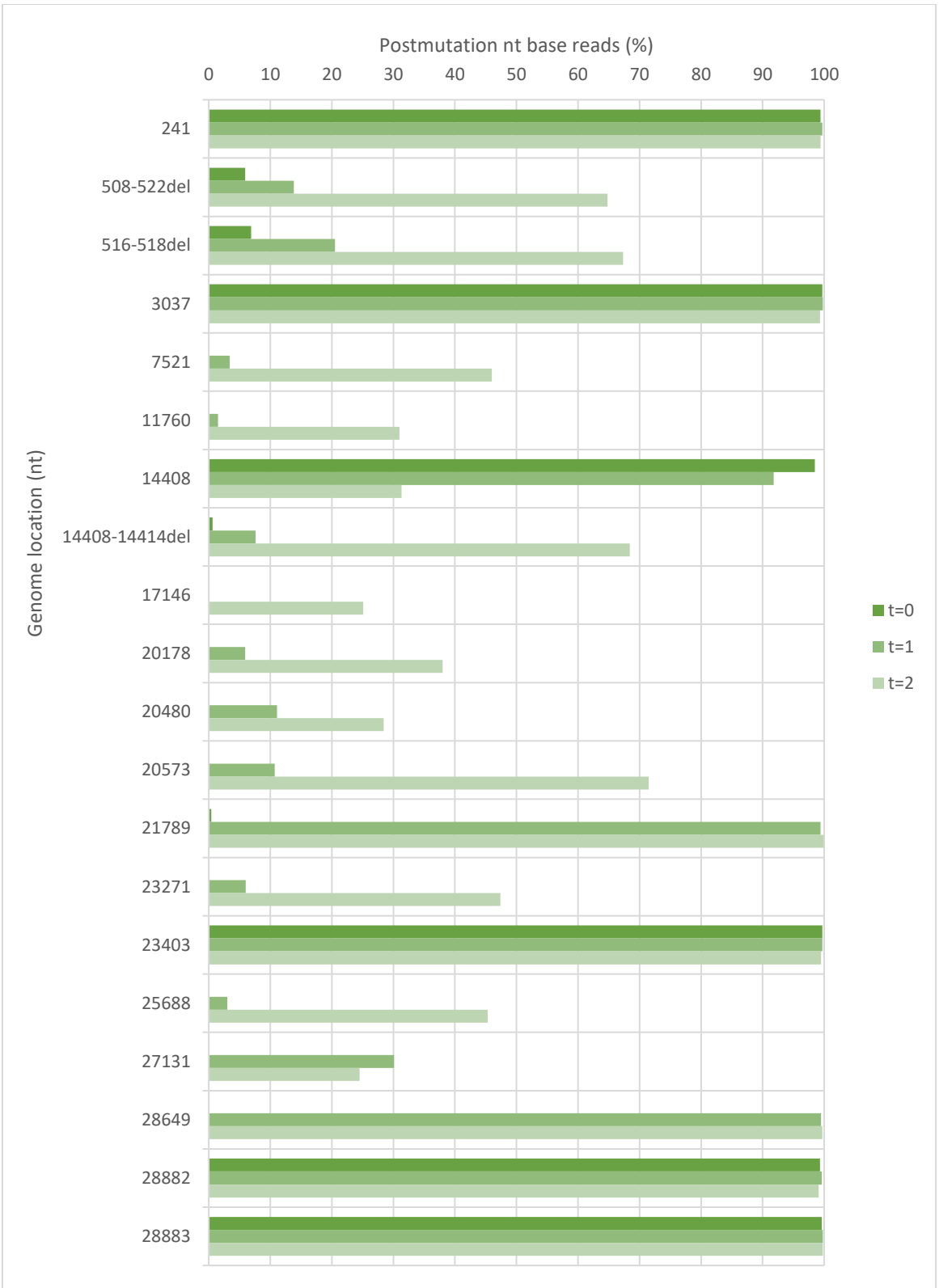
(S2) Fig. 38 - Spike "one-up" conformation: interaction of the RBD of the "up" monomer (Chain A) with the NTD of Chain B with nearby disulphide bridges (C131-C166 and C136) PDB: 6VSB. *Chains are coloured as in Fig. 1A-B, interchain interaction (Chain A's RBD to Chain C's NTD) - R357-N165 coloured elemental yellow, features such as N-glycans/disulphide bridges are light cyan.*



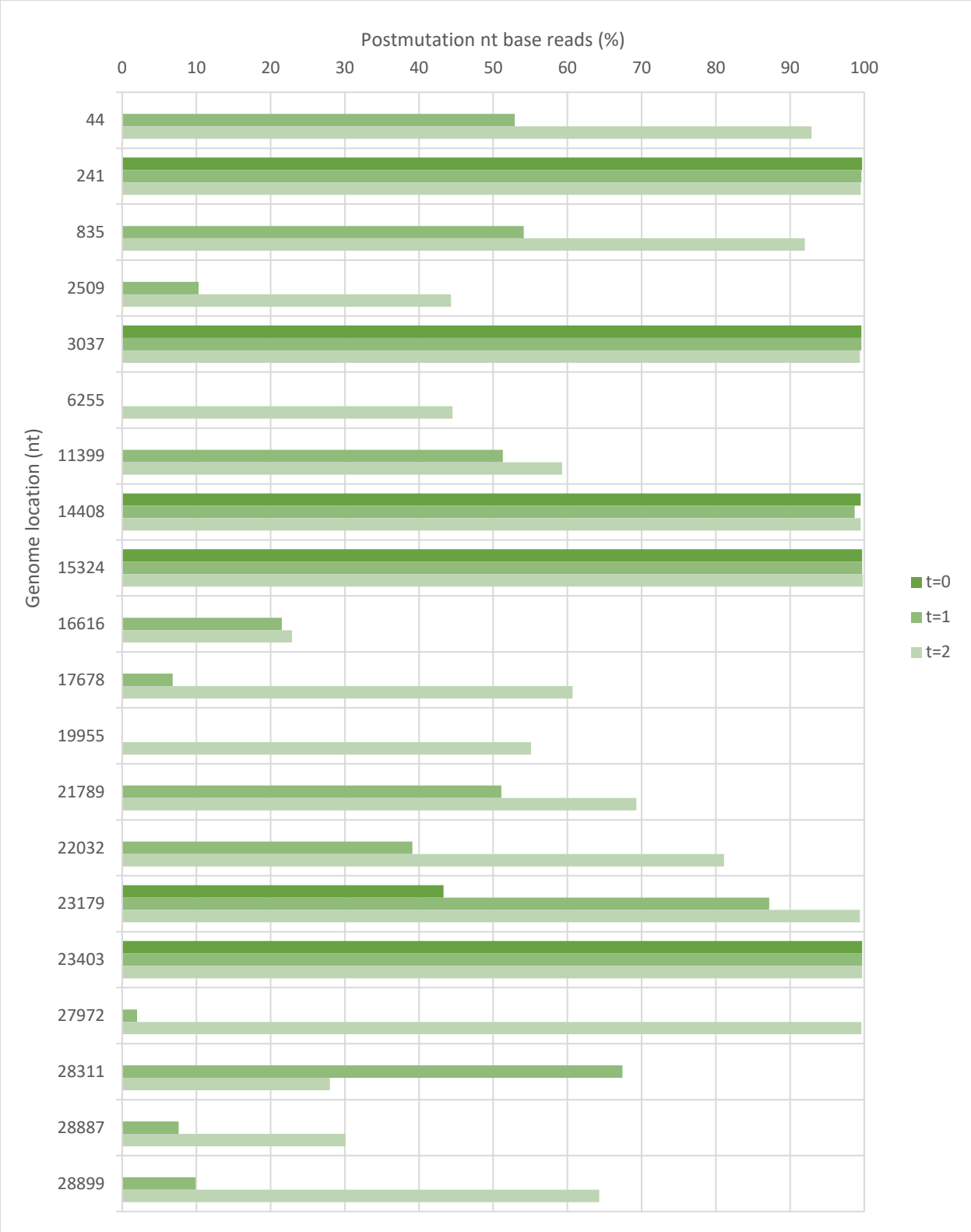
(S3) Fig. 39 – S protein alignment between 6VSB (blue) and 6ZXN (red) from the side and top view *with three RBD-bound Ty1 nano-Abs of 6ZXN coloured in orange. If Ty1 atoms are included the RMSD = 5.587, when Ty1 atoms are removed the RMSD = 0.590.*



(S4) Fig. 40 – Close-up of the Spike A222V DCP from Chain C of the "one-up" prefusion trimer (PDB:6VSB) with the lowest strain rotamer of Y38 selected. *Y38's polar bond distance to I285 is reduced by 0.1Å. The upper image shows pre-mutation (A222) and the bottom image shows the post-mutation (V222). Colour scheme – same as in Fig. 2. except grey cartoon represents the Chain C backbone and black represents Chain B, magenta dashes represent aromatic π- π interactions.*



(S5) Fig. 41 – Compound bar chart of post-mutation frequencies across time for FFM3, *legend refers across the three timescales t=0, t=1 and t=2 (start, middle and end of cultivation respectively)*



(S6) Fig. 42 – Compound bar chart of post-mutation frequencies across time for FFM7, *legend refers across the three timescales t=0, t=1 and t=2 (start, middle and end of cultivation respectively)*