The Protein Journal (2021) 40:260–295
https://doi.org/10.1007/s10930-021-09967-8

A Biochemical Perspective of the Nonstructural Proteins (NSPs) and the Spike Protein of SARS CoV-2

Francis K. Yoshimoto[1]

Accepted: 30 January 2021 / Published online: 24 February 2021

Abstract
The global pandemic that shut down the world in 2020 was caused by the virus, SARS CoV-2. The chemistry of the various nonstructural proteins (NSP3, NSP5, NSP12, NSP13, NSP14, NSP15, NSP16) of SARS CoV-2 is discussed. Secondly, a recent major focus of this pandemic is the variant strains of SARS CoV-2 that are increasingly occurring and more transmissible. One strain, called "D614G", possesses a glycine (G) instead of an aspartate (D) at position 614 of the spike protein. Additionally, other emerging strains called "501Y.V1" and "501Y.V2" have several differences in the receptor binding domain of the spike protein (N501Y) as well as other locations. These structural changes may enhance the interaction between the spike protein and the ACE2 receptor of the host, increasing infectivity. The global pandemic caused by SARS CoV-2 is a rapidly evolving situation, emphasizing the importance of continuing the efforts to interrogate and understand this virus.

Keywords SARS CoV-2 · Enzymes · Proteases · Viral proteins

1 Introduction

This article aims to follow up a recent report that discussed the proteins of severe acute respiratory coronavirus-2 (SARS CoV-2) [1], the cause of coronavirus disease-2019 (COVID19). Although the virus first emerged in December 2019, the effects of this pandemic are still increasingly evident. With promise of a new vaccine against SARS CoV-2 in the coming months, there is hope that we may soon return to our normal lives. However, this return is threatened by the recent emergence of new SARS CoV-2 strains that are more transmissible from the original strain [2]. This article will discuss the chemistry of the various nonstructural proteins (NSPs) and the spike protein found in SARS CoV-2 to serve as a resource to help understand this virus that caused this global pandemic. Besides the structural proteins: S protein, E protein, M protein, and N protein (or ORF2, ORF4, ORF5, and ORF9) and accessory proteins, there are specific enzymes expressed by the ORF1ab gene of SARS CoV-2 that catalyze essential reactions. Furthermore, Part 2 of this manuscript discusses the research progress towards understanding the spike protein, the basis of the recent vaccines against SARS CoV-2 [3–5]. The discussion of the recent investigations of the spike protein will provide insight into how the frequently occurring strains of SARS CoV-2 are manifesting. Possible explanations to rationalize the increased prevalence of these strains that have changes in the spike protein sequence are described. The structures of the protein presented in the text were made using UCSF Chimera software (version 1.12) [6]. In general, the figures with superimposed structures were generated using the default parameters in the Chimera software: Needleman–Wunsch alignment algorithm with a BLOSUM-62 matrix.

2 Discussion

2.1 Part 1: The Enzymes That are Nonstructural Proteins (NSPs) from the ORF1ab Gene

The ORF1ab gene of SARS CoV-2 results in the expression of a polypeptide that is cleaved into 16 nonstructural proteins [1]. From the ORF1ab gene, SARS CoV-2 has two protease enzymes: NSP3 (papain like protease) and NSP5 (3C-like protease), an RNA polymerase that copies viral RNA: NSP12, a 5′-RNA triphosphatase enzyme: NSP13,

✉ Francis K. Yoshimoto
francis.yoshimoto@utsa.edu

[1] The University of Texas at San Antonio (UTSA), Department of Chemistry, San Antonio, TX 78249-0698, USA

guanosine N7-methyltransferase: NSP14, an endoribonuclease: NSP15, and a 2′-*O*-ribose-methyltransferase (NSP16). This group of enzymes can generally be classified into different subgroups: (i) proteases (NSP3 and NSP5), (ii) enzymes involved in the 5′-capping modification of viral RNA—a posttranslational modification of viral RNA (to allow the viral RNA to escape the host innate immune system) (NSP13, NSP14, NSP15, and NSP16), (iii) RNA replication (NSP12), and (iv) other RNA modifying activities such as posttranslational modification of proteins (ADP ribose phosphatase activity of NSP5) and exoribonuclease/endoribonuclease activity (NSP14/NSP15 activity).

2.1.1 NSP3 (Papain-Like Protease)

The NSP3 (papain-like protease) is a multifunctional protein with 1,945 amino acid residues. The papain like domain catalyzes the reaction that cleaves the peptide bonds between: (i) NSP1 and NSP2, (ii) NSP2 and NSP3 [7], and (iii) NSP3 and NSP4 (Table 1) [8]. This enzyme cleaves at the consensus sequence LXGG [8]. NSP3 from SARS CoV-2 has been crystallized and biochemically characterized [9]. The catalytic triad of NSP3 of SARS CoV-2 are found in the following residues: D286, H272, and C111. This recent group expressed the PLpro-Ubl domain of NSP3 (amino acid residues 746-1060) [9]. SARS CoV-2 NSP3 has been shown to preferentially cleave the ubiquitin-like interferon-stimulated gene 15 (ISG15) protein. This cleavage of ISG15 from interferon factor 3 (IRF3) weakens the type I interferon response. Another group reported the crystal structure (PDB ID: 7CMD)—the structure is shown in Fig. 1 [10]. There are many efforts in studying the crystal structure of NSP3 and designing effective inhibitors for this protease—for instance Michael acceptors have been designed to form covalent thioether bonds with the active site of the cysteine residue [11]. Table 1 shows the sequences that NSP3 cleaves with its catalytic triad. The reaction catalyzed by NSP3 is shown in Scheme 1.

In addition to its protease activity, NSP3 also has other domains that confer other activities. For instance, there is a ribose phosphatase domain. The crystal structure of the ADP ribose phosphatase domain of NSP3 has been elucidated (PDB ID: 6w02) [12]. The structure of the ADP ribose phosphatase domain is shown in Fig. 2. SARS CoV has an essential asparagine-41 for ADP ribose-1″-phosphate phosphatase activity [7]. This ADP deribosylating activity is related to avoiding the host's immune system [7].

2.1.2 NSP5 (3C-Like Protease)

NSP5 cleaves at 11 distinct sites in the ORF1ab polyprotein with 306 amino acids after excising itself from the polyprotein [13]. The active site of NSP5 (3C-like main protease) has a catalytic dyad of a cysteine-145 residue and histidine-41 residue. The crystal structure of NSP5 has been reported (PDB ID: 6Y2E) [14]. The structure of NSP5 is shown in Fig. 3. Other structures with inhibitors bound to NSP5 [15, 16] have been reported [17]. Figure 3 also shows the structure of NSP5 with the inhibitor, GC376 bound (PDB ID: 6WTT) [18]. Table 2 shows the sequences that are cleaved by NSP5.

2.1.3 NSP12—RNA Polymerase

NSP12 is the RNA polymerase with 932 amino acids that copies viral RNA. The structure of NSP12 has been reported [19]. The structure of NSP12 that is complexed with an RNA template and NSP8 is shown in Fig. 4. Interestingly, NSP8 is shown stabilizing the RNA template with its positively charged residues coordinating to the negatively charged phosphate backbone of the RNA template (Fig. 5, expanded view of Fig. 4). Remdesivir is the current antiviral drug used to treat SARS CoV-2, and this drug is a prodrug, which is metabolized to the active form and is incorporated by NSP12 into the RNA template to stall replication [20]. In the inhibition assays between SARS CoV-2 RdRp complex with remdesivir triphosphate, the investigators used 100 nM concentration of remdesivir to show inhibition of RNA polymerase activity [20]. For comparison remdesivir triphosphate against the RNA-dependent RNA polymerase activity of the Ebola virus, concentrations at 33 μM had showed effects of inhibition [21]. In Vero E6 cells, remdesivir blocked SARS CoV-2 infection with a half maximum

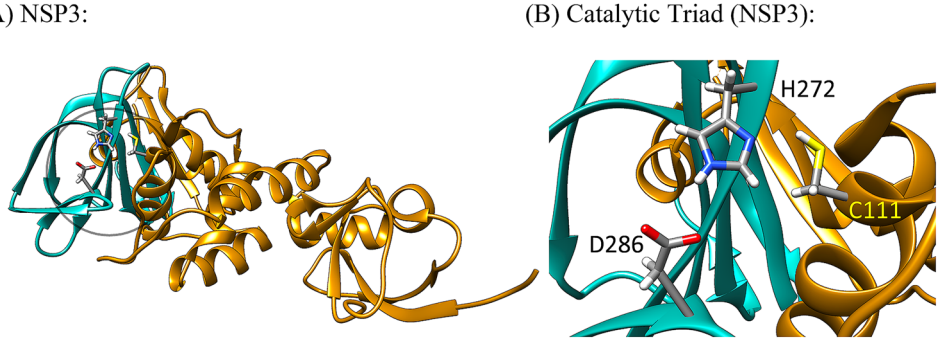
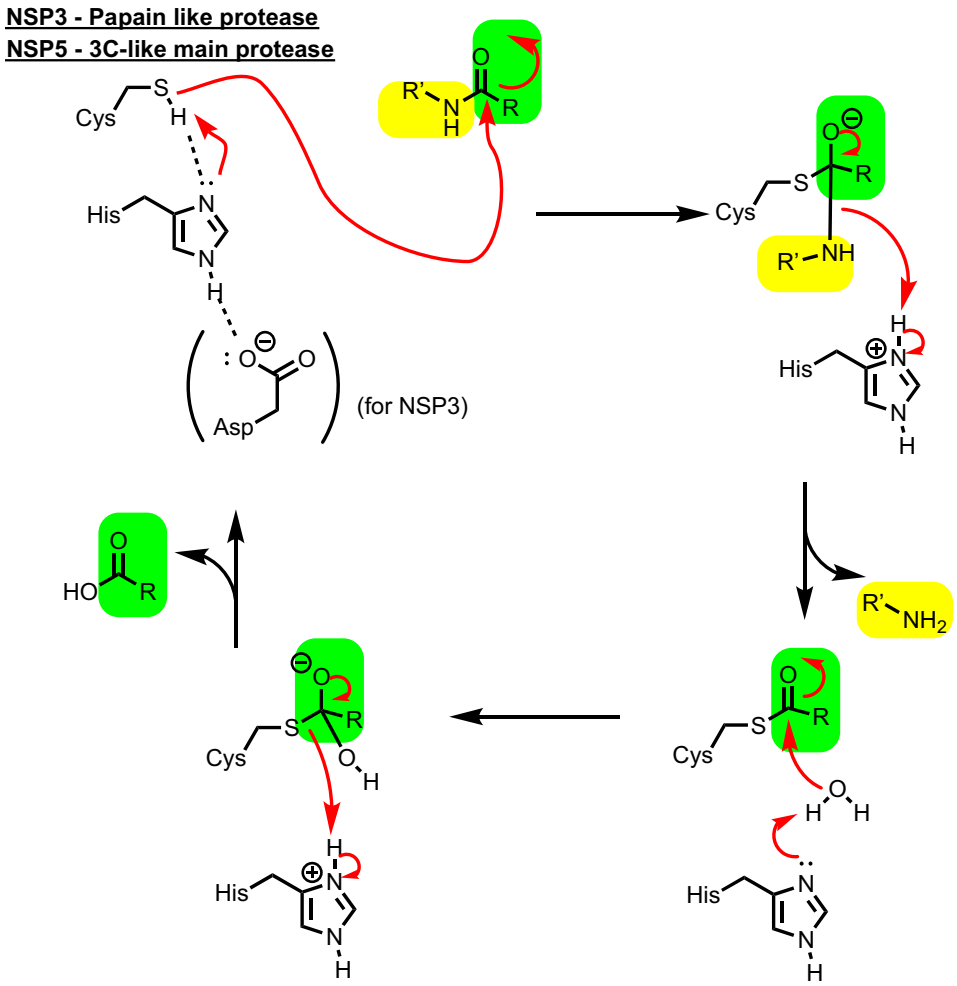


Fig. 1 The crystal structure of the papain like protase domain of NSP3 (PDB ID: 7CMD). Zoomed in region of the catalytic triad of NSP3 (aspartate-286, histidine-272, and cysteine-111)



Scheme 1 The chemical reaction catalyzed by NSP3 (papain like protease or PLpro) and NSP5 (3C-like protease or 3CLpro). Also see Ref. [105] for more information. NSP3 has a catalytic triad (cysteine–histidine–aspartate) while NSP5 has a catalytic dyad (cysteine–histidine) [106]. For NSP3 the key residues are: D286, H272, and C111, and for NSP5, the active site residues are: H41 and C145

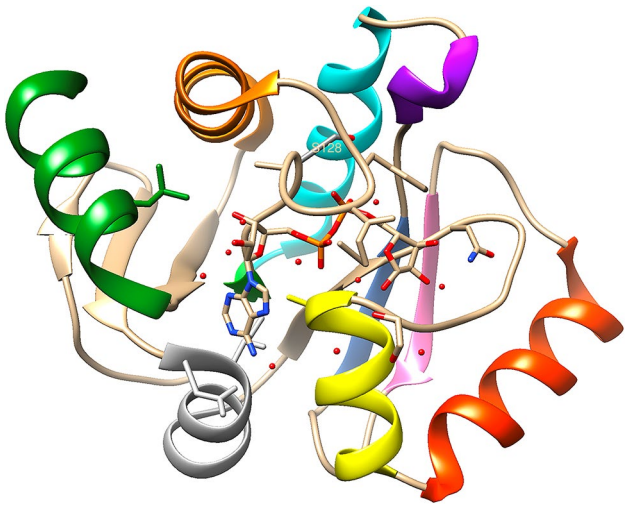


Fig. 2 Crystal structure of the ADP ribose phosphatase domain of NSP3 (PDB ID: 6W02) [12]. The red spheres are water molecules (Color figure online)

effective concentration (EC_{50}) of 0.77 µM [22]. Remdesivir triphosphate inhibits ebola virus replication in HMVEC/TERT (human microvascular endothelial) cells with a half maximum effective concentration (EC_{50}) of 0.06 µM [23]. Scheme 2 shows the generic mechanism of how RNA polymerase replicates the viral genome.

A structure of NSP12 with remdesivir in the active site is available (PDB ID: 7BV2) [24]. This structure is a complex between NSP12 with NSP7 and NSP8 (Fig. 6 and see Fig. 7 for expanded view of the active site of NSP12 with remdesivir bound).

The mode of action of remdesivir is worth discussing as more nucleoside based antiviral drugs could be developed based on this drug and favipiravir [25, 26]. Remdesivir is a prodrug—it is metabolized into its active form after it enters the cell membrane (Fig. 8). The enzymes, cathepsin A (CatA) and carboxyesterase 1 (CES1), convert remdesivir to its alanine metabolite, which then undergoes hydrolysis by the enzyme, histidine triad nucleotide binding protein 1 (HINT1), to the monophosphate [27]. The monophosphate is finally modified by kinases to form remdesivir triphosphate

Table 1 The catalytic triad of NSP3 is known to cleave sites between NSP1–NSP2, NSP2–NSP3, and NSP3–NSP4

Entry	NSP3—papain-like protease site	Protein sequence
1	NSP1 ↓ NSP2:	…SGVTRELMRELNGG ↓ AYTRYVDNNFCG…
2	NSP2 ↓ NSP3:	…NMMVTNNTFTLKGG ↓ APTKVTFGD-DTV…
3	NSP3 ↓ NSP4:	…VVNVVTTKIALKGG ↓ KIVNNWLKQLIK…

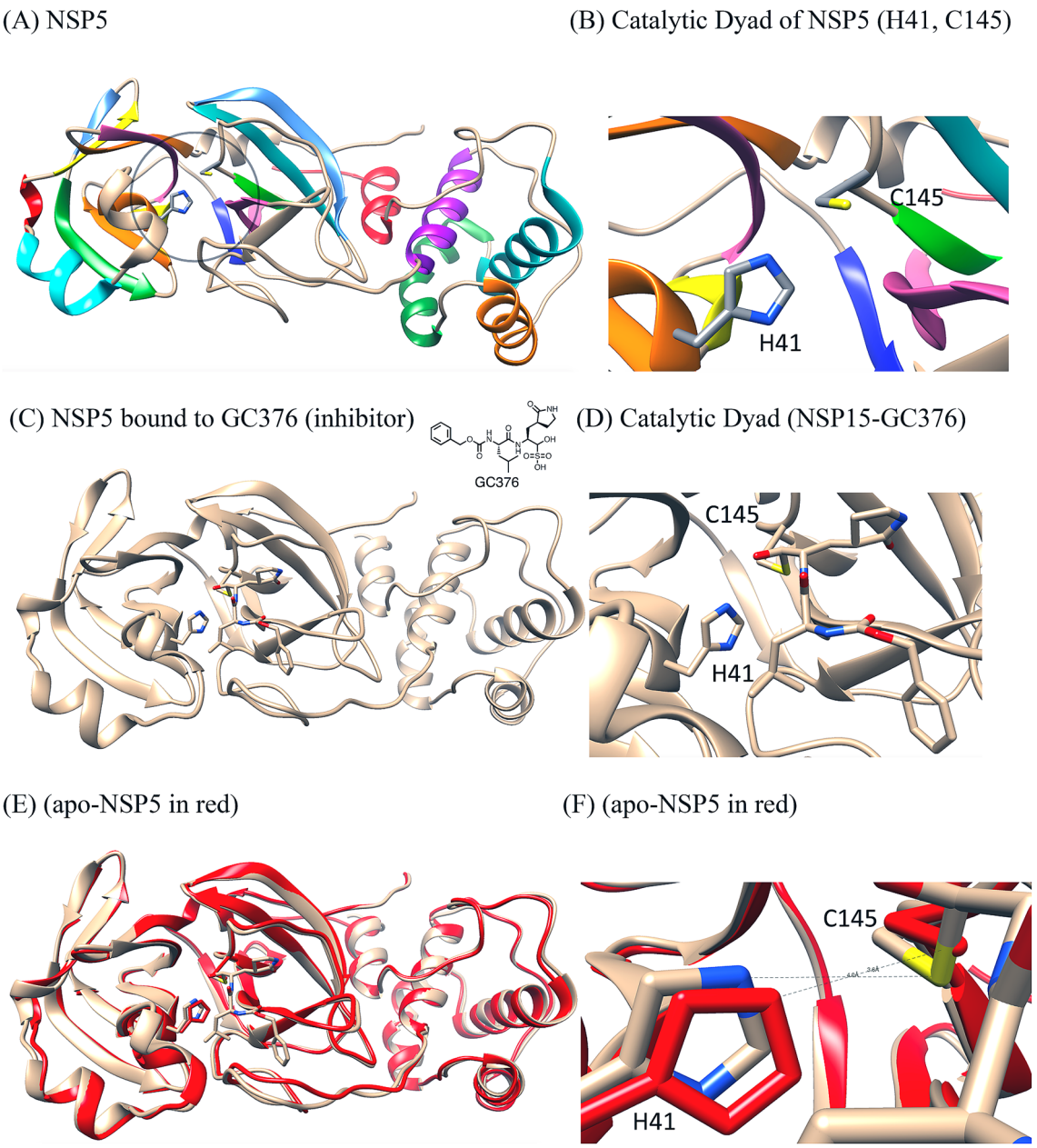


Fig. 3 **a** Crystal structure of NSP5 (PDB ID: 6Y2E). **b** Zoomed in view of the catalytic dyad (cysteine-145 and histidine-41) on the right (PDB ID: 6Y2E) [14]. **c** Crystal structure of NSP5 bound to inhibitor, GC-376 (PDB ID: 6WTT) [18]. **d** Zoomed in view of the catalytic dyad with inhibitor bound to the cysteine residue (C145). **e** Superimposed structures of (**a**) and (**b**). (apo protein is in red). **f** Zoomed in view of the superimposed structures. The distances between the histidine and the sulfur of the cysteine in the apo protein and inhibitor bound forms are 3.6 and 4.0 angstroms, respectively

Table 2 Sites of cleavage of NSP5—the 3C-like protease [97]

Entry	NSP5—3C-like protease site	Protein sequence
1	NSP4 ↓ NSP5:	…DVLYQPPQTSITSAVLQ ↓ SGFRKMAFPSGK…
2	NSP5 ↓ NSP6:	…EFTPFDVVRQCSGVTFQ ↓ SAVKRTIKGTHH…
3	NSP6 ↓ NSP7:	…KLLGVGGKPCIKVATVQ ↓ SKMSDVKCTSVV…
4	NSP7 ↓ NSP8:	…DINKLCEEMLDNRATLQ ↓ AIASEFSSLPSY…
5	NSP8 ↓ NSP9:	…WPLIVTALRANSAVKLQ ↓ NNELSPVALRQM…
6	NSP9 ↓ NSP10:	…LNRGMVLGSLAATVRLQ ↓ AGNATEVPANST…
7	NSP10 ↓ NSP11:	…WKGYGCSCDQLREPMLQ ↓ SADAQSFLNGFA…
8	NSP12 ↓ NSP13:	…WEPEFYEAMYTPHTVLQ ↓ AVGACVLCNSQT…
9	NSP13 ↓ NSP14:	…LQFTSLEIPRRNVATLQ ↓ AENVTGLFKDCS…
10	NSP14 ↓ NSP15:	…YKQFDTYNLWNTFTRLQ ↓ SLENVAFNVVNK…
11	NSP15 ↓ NSP16:	…MLWCKDGHVETFYPKLQ ↓ SSQAWQPG-VAMP…

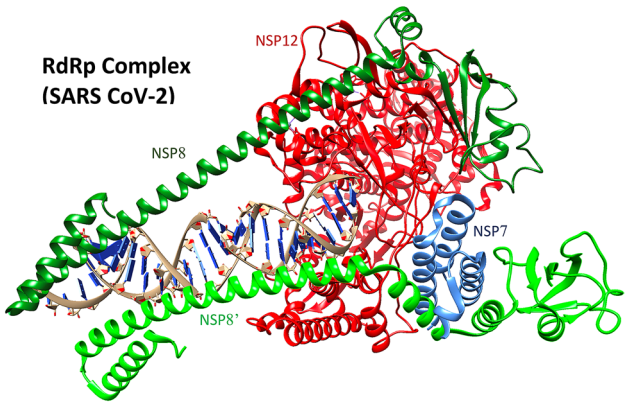


Fig. 4 Structure of the RNA-dependent RNA polymerase (RdRp) complex: NSP12 RNA polymerase (red) from SARS CoV-2 (PDB ID: 6YYT) [19]. The green proteins are the two NSP8 proteins (NSP8 and NSP8′) that are believed to interact and stabilize the RNA. NSP7 is shown in blue (Color figure online)

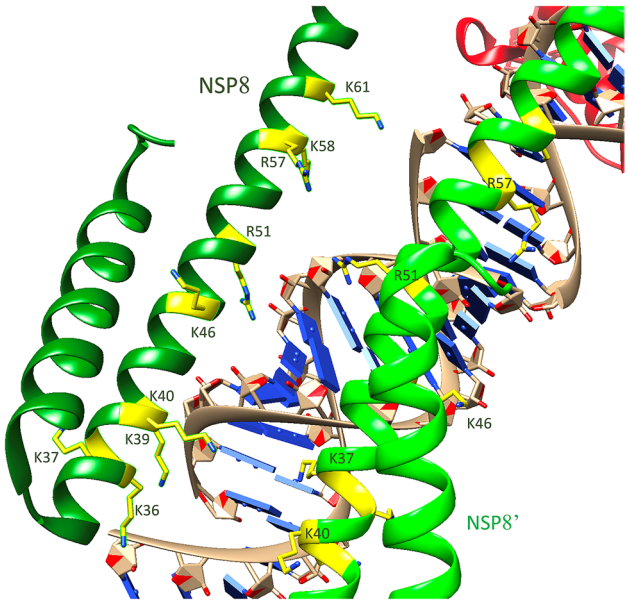


Fig. 5 There are positively charged amino acid residues on NSP8 and NSP8′ (K37, K36, K40, K46, R51, R57, K58, K61) that stabilize the negatively charged phosphate groups in the RNA template (PDB ID: 6YYT) [19]

(also referred to as GS441326 or RDV-TP) [20], the substrate for the RNA polymerase (NSP12) complex for incorporation into the primer strand [28].

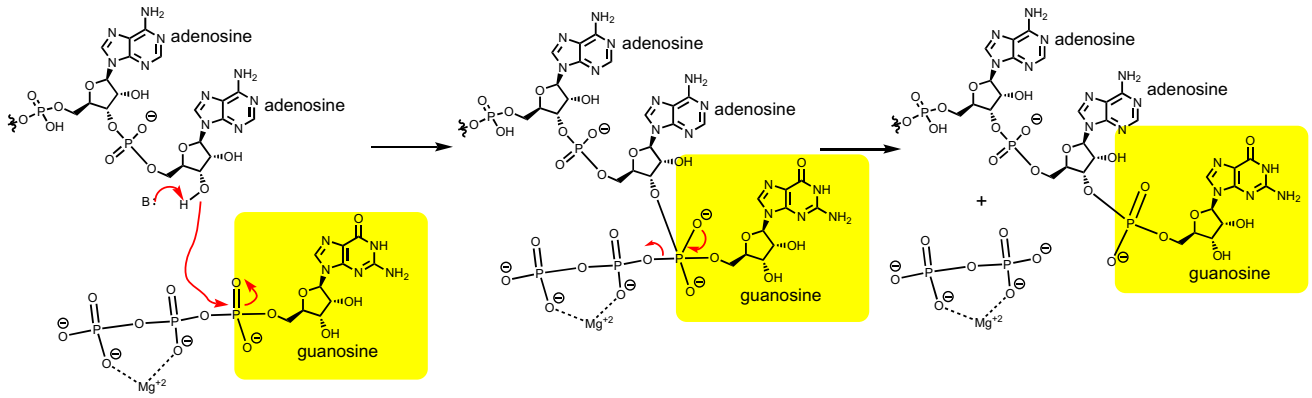
After remdesivir triphosphate (RTP) is incorporated into the RNA primer—inhibition of the NSP12 complex occurs through chain termination as shown in Fig. 9. Remdesivir takes the place of adenosine and is incorporated opposite of uridine from the template strand. Serine-861 from NSP12 is suspected to have a steric clash with the C1′-nitrile moiety of remdesivir unit only after three additional nucleotides are incorporated. This steric interaction was determined from a model [20]. Moreover, although there was no experimental evidence of a nucleophilic addition (i.e. covalent adduction) of the serine hydroxy group (S861) onto the carbon of the nitrile moiety—this may be a reasonable possibility as well as an electrostatic interaction between the O–H of the serine residue and the terminal nitrogen lone pair of the nitrile.

In the clinic, remdesivir was better than the placebo in treating adults, who were hospitalized with COVID-19. These patients received 200 mg of remdesivir on day 1 followed by 100 mg each day for up to 9 additional days [29].

2.1.4 NSP13—Helicase and RNA 5′-Triphosphatase

NSP13 with 601 amino acids has multiple enzymatic activities—helicase, RNA 5′-triphosphatase [30], and NTPase [31] activities.

2.1.4.1 RNA Capping RNA capping and methylation is a process that post-translationally modifies viral RNA to help the viral RNA hide from the recognition of the h ost's innate immune system [32]. This capping also ensures binding to the host ribosome for translation of the proteins. Figure 10



Scheme 2 RNA polymerase reaction mechanism incorporating a new RNA (RNTP) into the primer strand

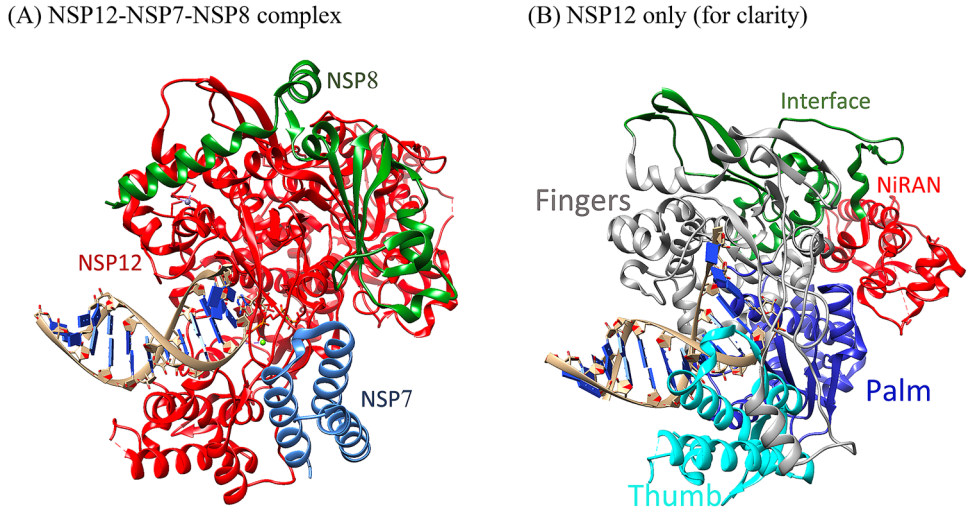


Fig. 6 **a** Structure of NSP12 (red), also called: RNA-dependent RNA polymerase (RdRp) in complex with NSP7 (green) and NSP8 (blue) (PDB ID: 7BV2). **b** NSP12 alone with the RNA template (NSP7 and NSP8 are hidden for clarity). The different domains of NSP12 [98]—Nidovirus RdRp-associated nucleotidyltransferase (NiRAN): 51–249 (red), Interface: 250–365 (green), Fingers: 366–581 and 621–679 (grey), Palm: 582–620 and 680–815 (blue), and Thumb: 816–932 (cyan) (Color figure online)

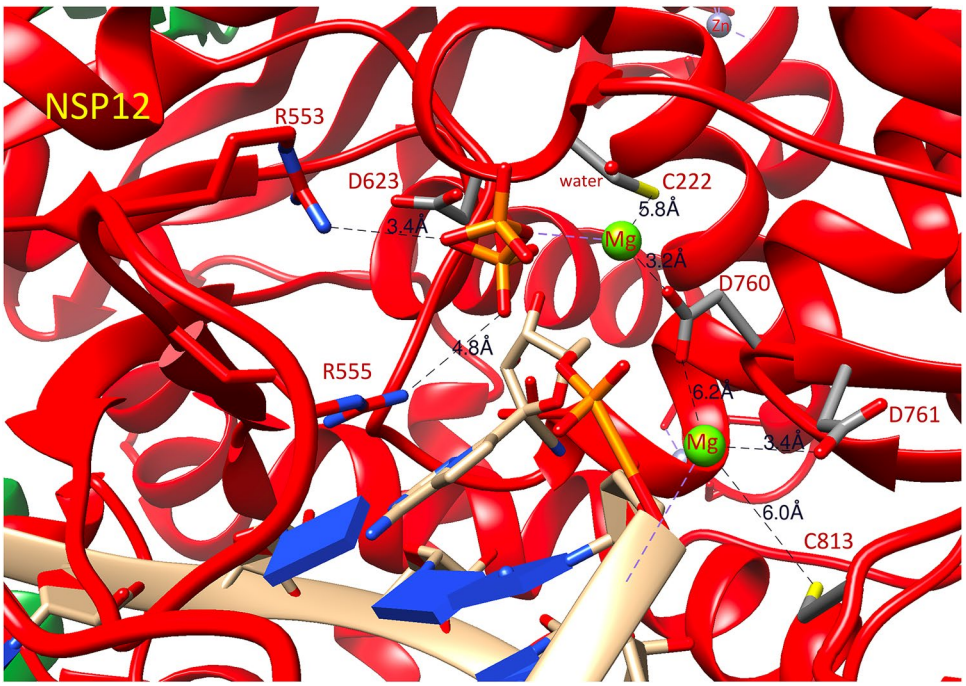


Fig. 7 The active site of NSP12 with remdesivir incorporated (PDB ID: 7BV2). The red sphere by C222 is a water molecule (Color figure online)

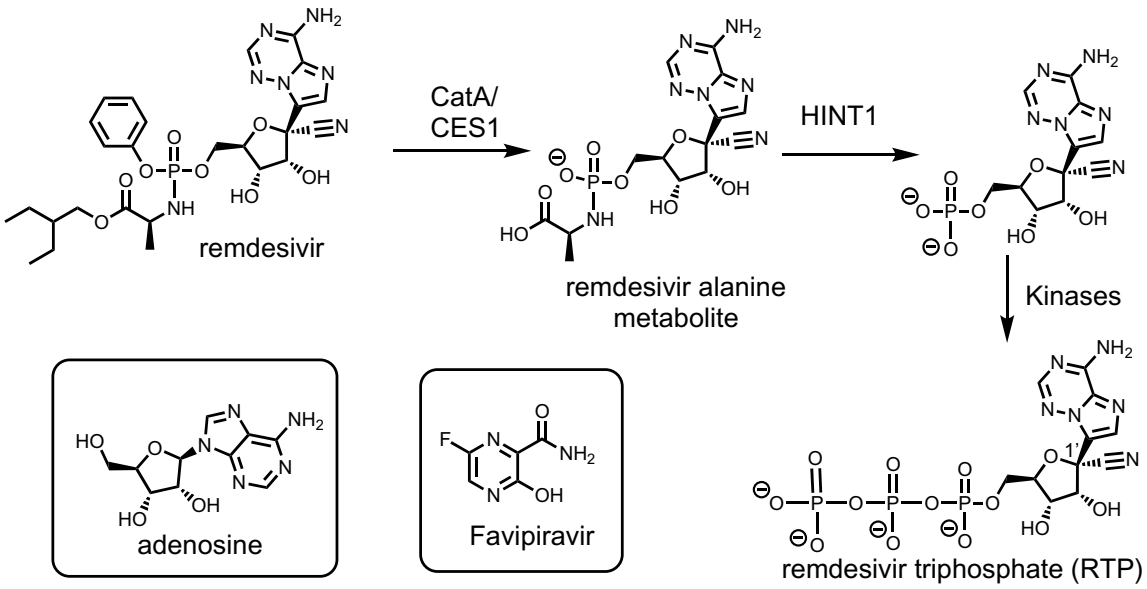


Fig. 8 The metabolism of remdesivir into its triphosphate metabolite, the substrate of NSP12. Also shown is the structure of adenosine for comparison. The structure of favipiravir, another antiviral prodrug is shown

shows the general chemical structure of the 5′-cap modification of RNA. This process in coronaviruses involves 4 steps [33].

(i) RNA triphosphatase activity by NSP13—which involves the removal of the 5′-gamma-phosphate group of the mRNA.

(ii) Guanylyltransferase activity—involving the transfer of a GMP group on the remaining 5′-diphosphate end (the enzyme that transfers this GMP group is still unknown).

(iii) N7-methyltransferase activity of NSP14—this activity caps the N7-nitrogen of the guanosine at the 5′-end (making the "cap-0" structure—^{7Me}GpppN).

(iv) 2′-O-methyltransferase activity of NSP16.

The RNA 5′-triphosphatase activity is important for initiating the RNA capping process. This activity of NSP13 is regioselective for hydrolyzing the γ-phosphate group of the 5′-terminus of the viral RNA (Scheme 3). The subsequent guanylyl transferase step (step (ii) in the 5′-capping process) introduces a guanosine-monophosphate (GMP) group at this resulting diphosphate end. However, the enzyme that catalyzes the GMP incorporation reaction is unknown. An interesting note is that a different protein, baculovirus LEF-4 (late expression factor-4) protein, is known to possess multiple activities including RNA 5′-triphosphatase, nucleoside triphosphatase, and guanylyltransferase activities [34].

In terms of the 5′-triphosphatase activity for NSP13 [30], the key residues in the active site have been identified in SARS CoV-1 [35]. This active site was suggested to be the same site for NTPase activity as well. The key amino acid residues in the active site were determined to be K288, S289, D374, Q404, and R567 [35]. When any of these amino acid residues were changed to alanine residues, the activity was shut down [35]. These amino acid residues are conserved in SARS CoV-2. Based on the sequence alignment between the NSP13 of SARS CoV-1 and SARS CoV-2, only one amino acid out of 601 is different (position-570 is an isoleucine in SARS CoV-1 and a valine in SARS CoV-2) [1]. Scheme 3 shows a proposed mechanism of how NSP13 may regioselectively hydrolyze the γ-phosphate group of its substrate with the help of key amino acid residues (K288 and D374), which is supported by the structure shown in Fig. 11.

The structure of NSP13 for SARS CoV-2 has been reported as a complex with NSP12–NSP7–NSP8 (PDB ID: 6XEZ) [36]. In order to focus on the 5′-triphosphatase active site of the structure of NSP13, the NSP13–NSP12–NSP7–NSP8 complex (PDB ID: 6XEZ) was taken and only NSP13 was shown (i.e. NSP12, NSP7, and NSP8 are hidden) in Fig. 11. This structure (6XEZ) contains an ADP moiety bound to the active site. The available structure of the apo form of NSP13 of SARS CoV-1 is also available (PDB ID: 6YJT). Using the NSP13 structure from SARS CoV-1 and the knowledge of the active site residues (K288, S289, D374, Q404, and R567), the two structures were superimposed to show the active site of NSP13 for SARS CoV-2.

Figure 12 shows the structure of NSP13 of SARS CoV-2 (PDB ID: 6XEZ) as a complex with RNA polymerase, which is relevant for its helicase activity [36]. This structure was used to show the active site of the RNA-5′-triphosphatase active site of NSP13 in Fig. 11. The figure that follows (Fig. 13) shows the NSP13 in complex with the RNA-dependent RNA polymerase complex (RdRp complex:

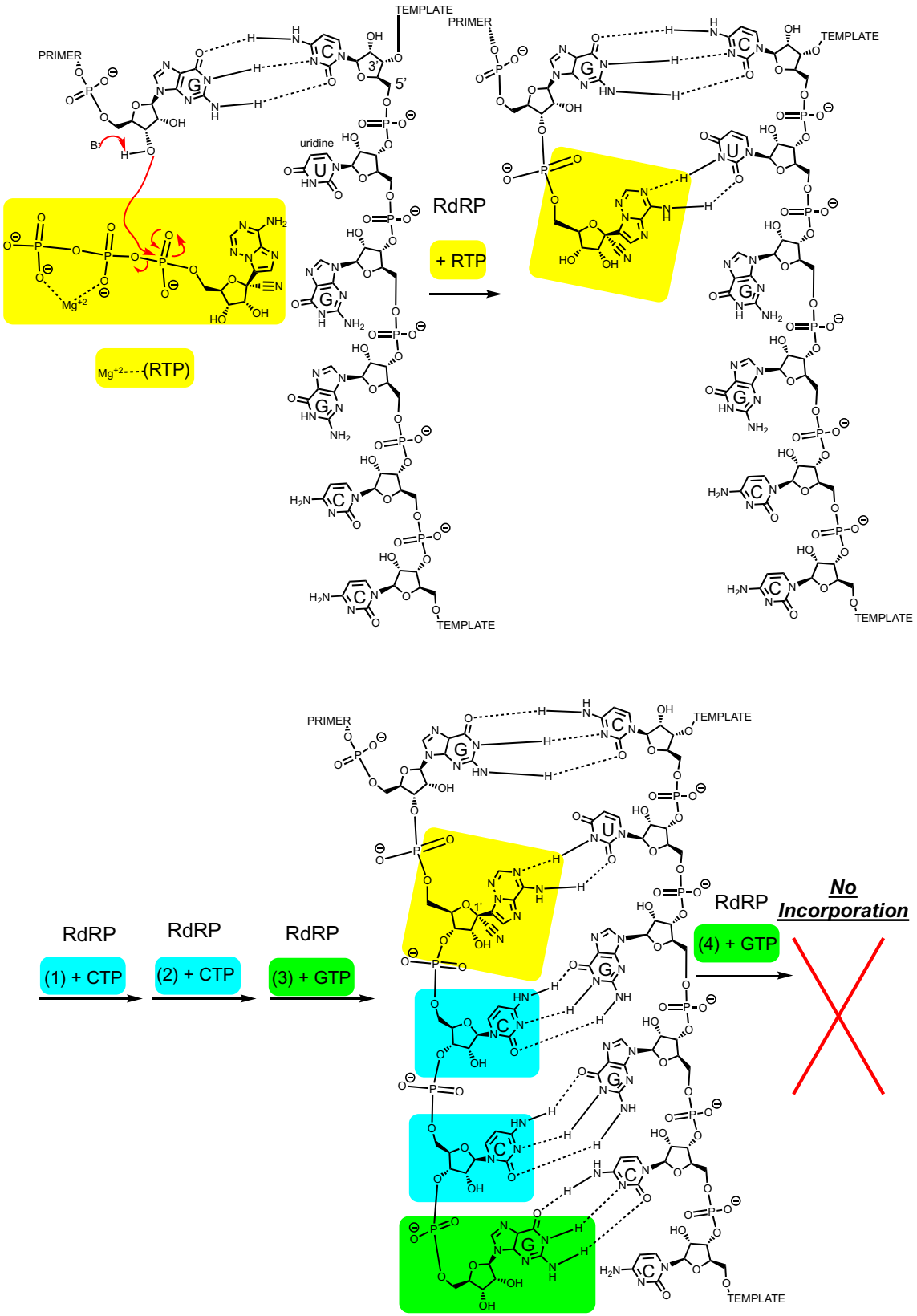


Fig. 9 How remdesivir incorporation into the RNA primer inhibits RNA-dependent RNA polymerase activity (NSP12–NSP7–NSP8 complex) through chain termination. After incorporation of remdesivir into the primer strand, the RNA polymerase complex incorporates three more nucleotides before stalling. A hypothetical sequence for the template is shown above to illustrate that three NTPs are incorporated after remdesivir incorporation into the primer while the fourth NTP is not incorporated [20]

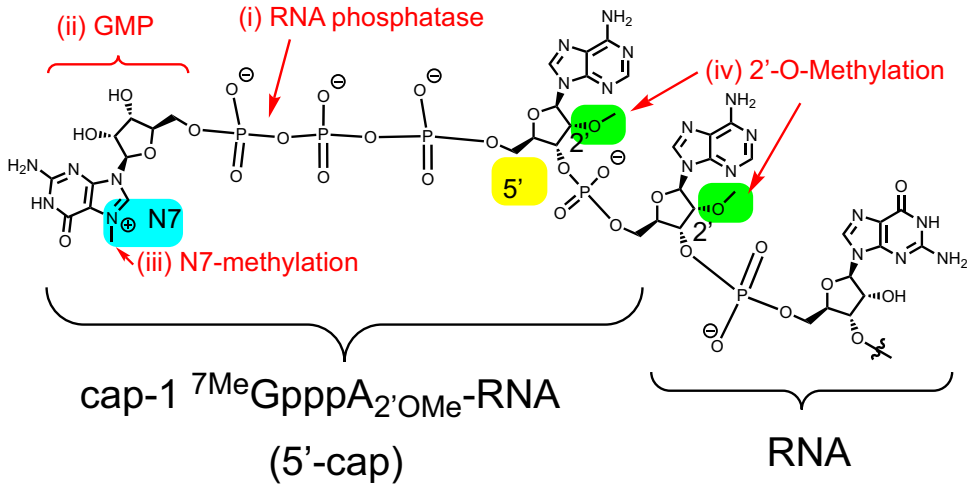
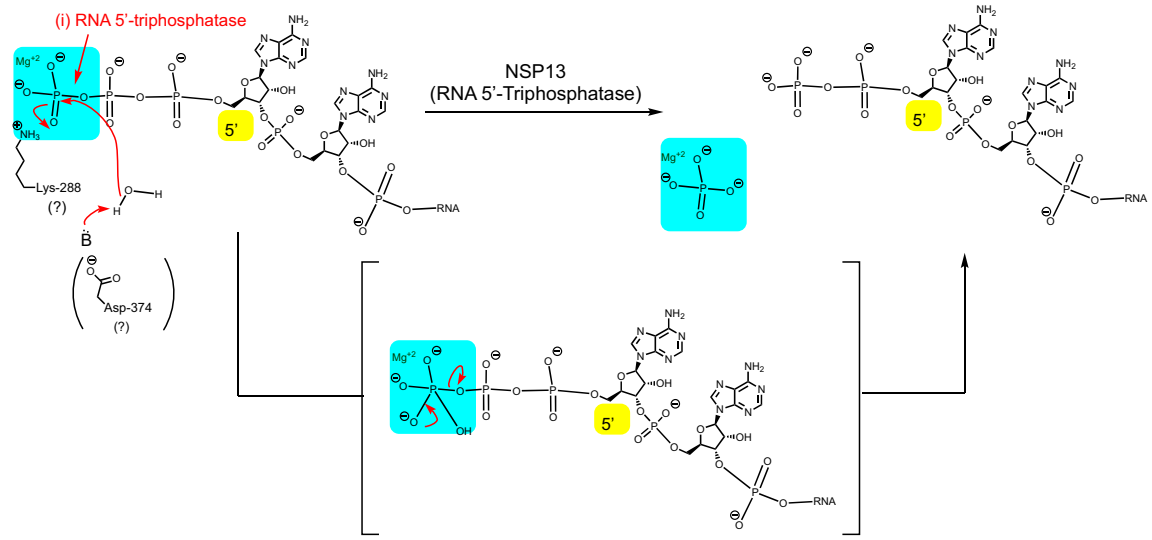


Fig. 10 The structure of the 5′-cap of RNA, processed by the viral proteins of SARS CoV-2. The 5′-cap of viral RNA prevents recognition by the host innate immune system and promotes translation by the ribosome



Scheme 3 The reaction catalyzed by NSP13 involving the RNA-5′-phosphatase activity to initiate the 5′-capping of mRNA. The amino acid residues K288 and D374 are proposed to play roles in promoting the terminal phosphate to leave and deprotonating the hydrolyzing water molecule, respectively. The support for this hypothesis is shown with the structure analysis in Fig. 11 (PDB ID: 6XEZ and 6YJT)

NSP12–NSP7–NSP8) with a strand of RNA embedded in the NSP13′ unit (PDB ID: 7XCM) [37], which presumably corresponds to the RNA template that the helicase is "unwinding" for replication to occur.

2.1.5 NSP14—Guanosine N7-Methyltransferase and Exoribonuclease

NSP14 comprised of 527 amino acids has been shown to have two activities: guanosine N7-methyltransferase activity as well as exoribonuclease activity. In the former, the methyl group of *S*-adenosylmethionine is transferred to the N7-group of the terminal guanosine [33]. Scheme 4 shows how NSP14 methylates the N7-position of the guanosine at the 5′-cap of RNA.

Although a structure of NSP14 has not been reported, the structure of NSP14 in complex with NSP10 for SARS CoV-1 has been reported (PDB ID: 5C8T) [38]. Fig. 14 shows the crystal structure of NSP14 for SARS CoV-1.

NSP14 also is reported to have exoribonuclease activity. Inactivating the exoribonuclease (ExoN) activity has been shown to be lethal for SARS CoV-2 [39].

2.1.6 NSP15—Endoribonuclease

NSP15 (also called EndoU) with 346 amino acids is the endoribonuclease enzyme that cleaves the 5′-polyuridine

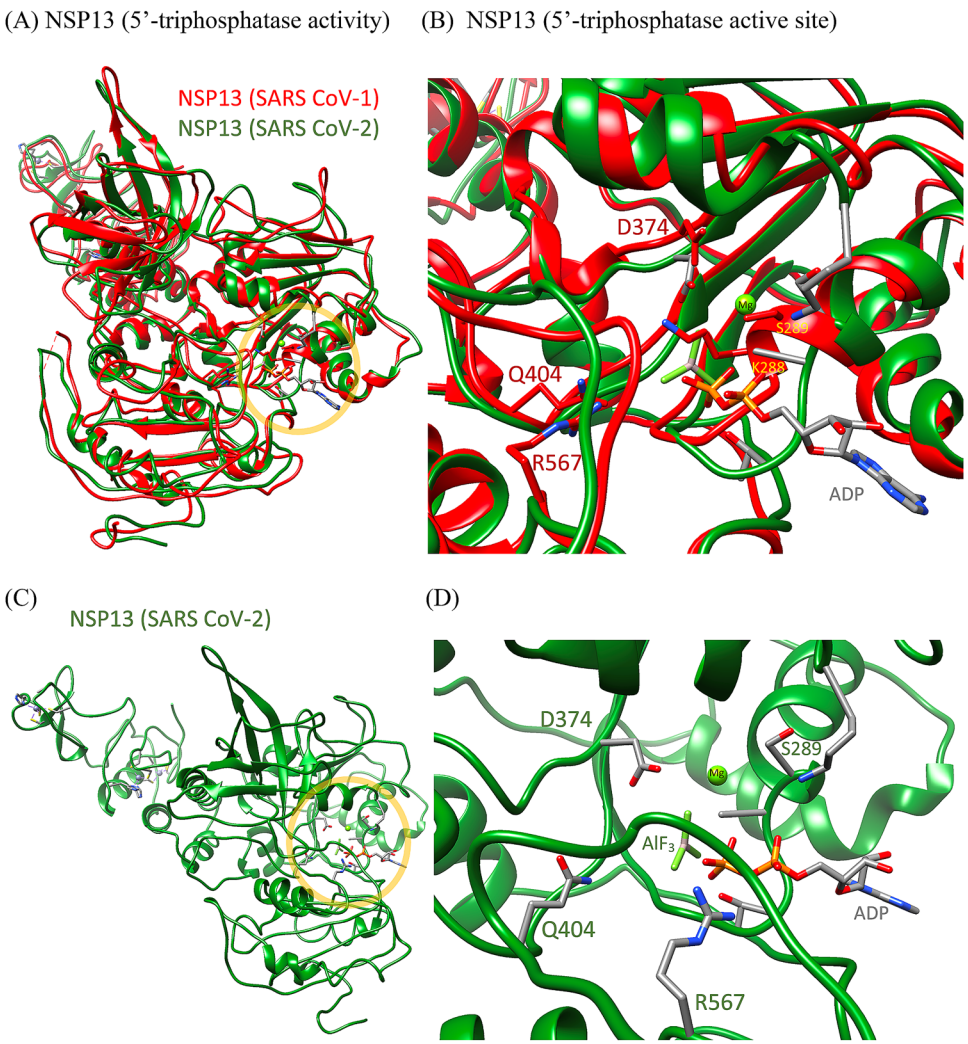


Fig. 11 **a** Structural superposition between NSP13 of SARS CoV-2 and SARS CoV-1 (PDB ID: 6XEZ and 6YJT). Under the Matchmaker option in Chimera software, the reference chain was set to chain F (green) of 6XEZ (NSP13 complex of SARS CoV-2 PDB ID), and the chain to match was set to chain A (red) of 6YJT (PDB ID for NSP13 of SARS CoV-1, apo protein). **b** Focused view of the 5′-triphosphatase active site. **c** A different angle of SARS CoV-2 NSP13 (green, alone, PDB ID: 6XEZ) for clarity. **d** Expanded view of the active site of SARS CoV-2 NSP13 (green)—an AlF_3 molecule is shown, which mimics the terminal monophosphate. The green spheres in **b** and **d** are Mg^{2+} ions (they are identical) (Color figure online)

motif of negative sense viral RNA [40]. The polyuridine tail arises from the polyA-templated processing that occurs for messenger RNA (mRNA) [41]. Histidine residues are hypothesized to be involved in the catalysis of the active site of endoribonucleases [42]. Scheme 5 shows the endoribonuclease activity of NSP15.

The structure of NSP15 has been reported [43]. The key active site residues of NSP15 are: His235, His250, Lys290, and Thr341 [43]. Fig. 15 shows the structure of NSP15 (PDB ID: 6VWW). Moreover, a structure of NSP15 with uridine-3′,5′-diphosphate is available (PDB ID: 7K1O) (Figs. 15c, d).

2.1.7 NSP16—2′-*O*-Ribose-methyltransferase

The final 5′-capping enzyme that this review covers is NSP16 containing 298 amino acids. This enzyme methylates the 2′-position of the ribose of the first transcribed nucleotide with *S*-adenosylmethionine [44]. The structure of NSP16 has been reported (PDB ID: 6YZ1) and the structure is shown in Fig. 16 [45]. The reaction catalyzed by NSP16 is shown in Scheme 6.

2.2 Part 2: The Spike Protein of SARS CoV-2

The spike protein is a trimeric glycoprotein expressed by ORF2 in the viral genome. It has been recently suggested that a SARS CoV-2 strain that possesses the D614G variant of the spike protein is more widely spread than the original strain with the aspartate residue at position-614 [46]. This 614G variant is not associated with an increased severity of infection, but it has been suggested that the 614G variant has increased infectivity relative to the D variant [47]. In order to explain the enhanced viral loads of the D614G mutant strain [46], a thorough structural analysis of the spike protein was performed.

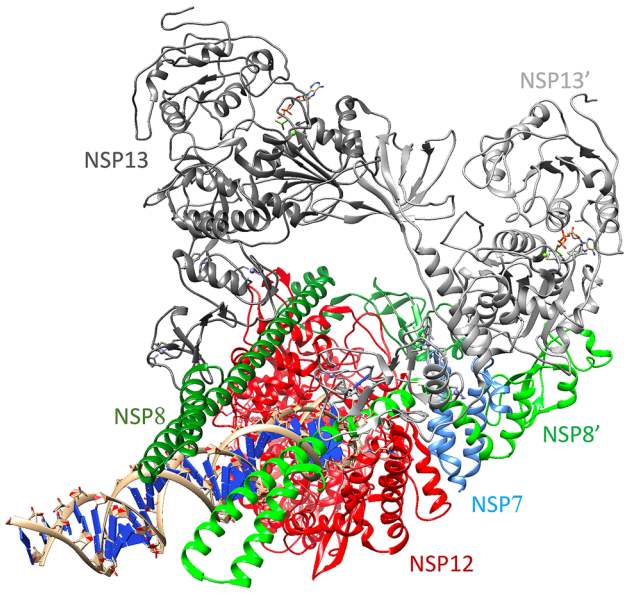


Fig. 12 The structure of NSP13 (grey) in complex with NSP7 (blue), NSP8 (green), and NSP12 (red) bound to an RNA template (PDB ID: 6XEZ). (There are two NSP13 units (NSP13 and NSP13′), two NSP8 units (NSP8 and NSP8′), one NSP12 unit, and one NSP7 unit) (Color figure online)

2.3 Structural Comparisons of the SARS CoV-2 Spike Protein with Other Known Coronaviruses that Infect Humans (SARS CoV-1, HCoV-299E, MERS CoV, HCoV-OC43, HCoV-HKU1, HCoV-NL63)

Figures 17, 18, 19, 20, 21, and 22 show the individual primary sequence alignments between the SARS CoV-2 spike protein and the 6 spike proteins from other human coronaviruses: (i) SARS CoV-1 (β-coronavirus), (ii) Human coronavirus 229E or HCoV-299E (α-coronavirus), (iii) Middle East respiratory syndrome coronavirus or MERS CoV (β-coronavirus), (iv) Human coronavirus OC43 or HCoV-OC43 (β-coronavirus) [48], (v) Human coronavirus HKU1 [49] or HCoV-HKU1 (lineage A β-coronavirus), and (vi) Human coronavirus-NL63 or HCoV-NL63 (α-coronavirus). All sequence identities and similarities shown in Table 3 were determined using LALIGN software (See Supporting Information for alignments) [50]. SARS CoV-1 and HCoV-NL63 [51] are known to interact with the angiotensin-converting enzyme 2 (ACE2) receptor while HCoV-229E [52] and MERS CoV [53] interact with the aminopeptidase N (APN) and dipeptidyl peptidase 4 (DPP4) receptors, respectively. In fact, it has been suggested that DPP4 can also act as a receptor for the spike protein of SARS CoV-2 [53]. Moreover, structural alignments of the available cryo-EM structures from the protein data bank (PDB) of the spike proteins have been performed to gain insight between similarities and differences between some of these viruses (virus/PDB ID: SARS CoV-2/7JJJ [54], SARS CoV-1/5XLR [55], HCoV 229-E/6U7H [52], MERS CoV/5X5U [56], HCoV OC43/6OHW [52], HCoV HKU1/5I08 [57], HCoV NL63/5SZS [58]).

2.4 A Structural Analysis of SARS CoV-2 Spike Protein

The spike protein exists as a trimer. Figures 23 and 24 show the three protomers that come together to form the trimer of the spike protein (PDB ID: 7JJJ) [54]. Figure 24 shows the different domain organizations of the spike protein (receptor binding domain, S1 unit, and S2 unit). The S1-S2 site is where the protease, furin, cleaves. The S1 unit binds to the ACE2 receptor [59] and the S2 unit mediates fusion of the viral and cellular membranes [60]. A more detailed discussion is presented in the next section.

2.5 Spike Protein Role in Viral Entry into the Host Cell

With many of the conformations of the spike proteins elucidated by cryo electron microscopy [61, 62], a better understanding of the dynamic process of viral entry is gained. Initially, the receptor binding domain (RBD) opens (step (i) in Figure 28) to readily bind to the ACE2 protein (step (ii)) on the host cell. The resulting spike protein binds to two more ACE2 proteins to form the spike protein bound to three ACE2 proteins (steps (iv) and (v)). Finally, the spike protein complex is cleaved by furin and TMPRSS2 to release the ACE2-S1 fragments (step (vi))—furin cleaves at the S1/S2 site (see Fig. 25, position: 685–686) and TMPRSS2 cleaves at the S2′ site (See Fig. 25, position: 816) [63, 64]. After cleavage, the S2 domain of the spike protein remains, which is now primed for viral entry into the host cell [63].

2.6 Glycosylated residues on the spike protein:

The spike protein has 22 distinct glycosylation sites [65] on each protomer: N17, N61, N74, N122, N149, N165, N234, N282, N331, N343, N603, N616, N657, N709, N717, N801, N1074, N1098, N1134, N1158, N1173, N1194 (positions 1158, 1173, and 1194 are not available from the structure, PDB ID: 7JJJ). The glycans play an important role in protein folding and evading the host immune system. The sequence of the spike protein was formatted using ProtParam and is shown in Fig. 25 [66]. Furthermore, the sites of glycosylation are highlighted in yellow in one of the protomers in Fig. 26 (PDB ID: 7JJJ). Understanding the sites of glycosylation is important because the spike protein produced by the vaccine and the spike protein from the virus have been shown to have different glycosylation patterns [67]. In

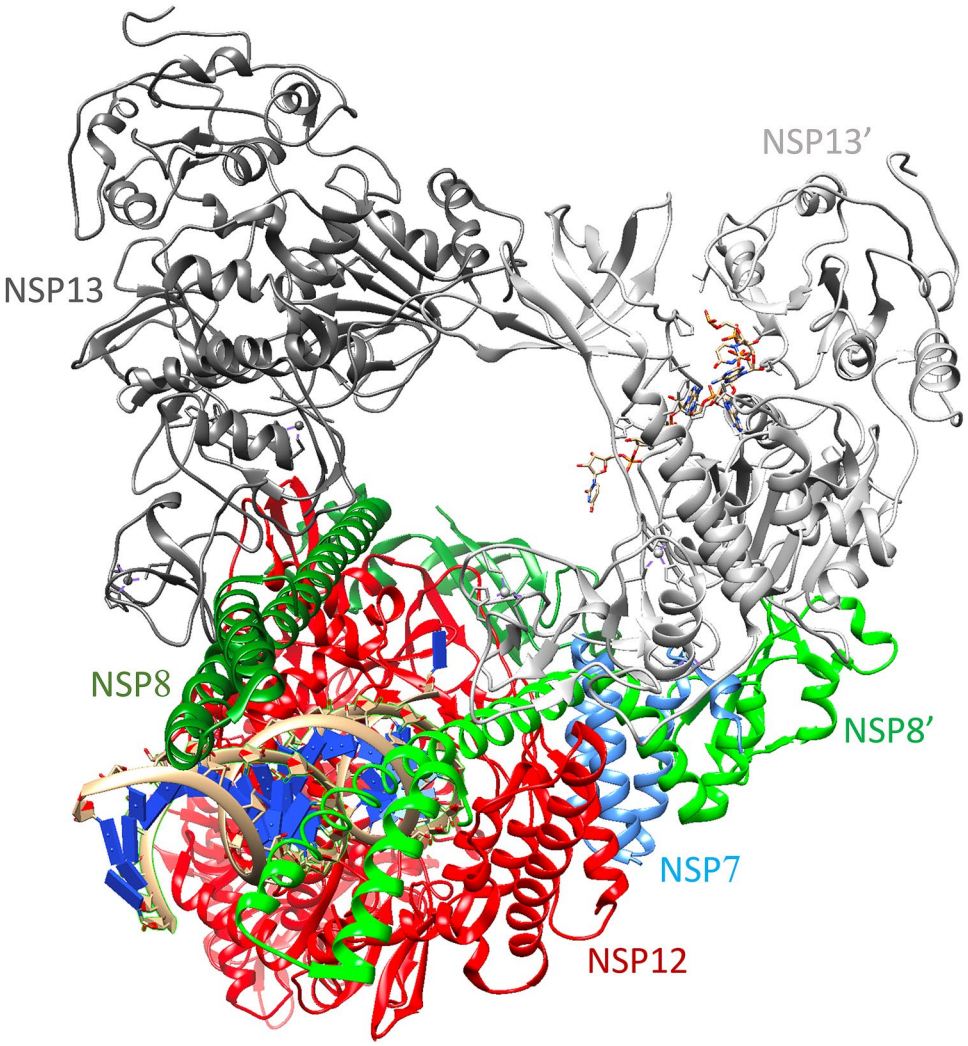


Fig. 13 Structure of NSP13 (grey, helicase) in complex with NSP12 (red), NSP7 (blue), NSP8 (green), and RNA (PDB ID: 7XCM) [37]. (There are two NSP13 units (NSP13 and NSP13′), two NSP8 units (NSP8 and NSP8′), one NSP12 unit, and one NSP7 unit). NSP13′ has part of the RNA template bound, which shows the helicase activity of this protein (Color figure online)

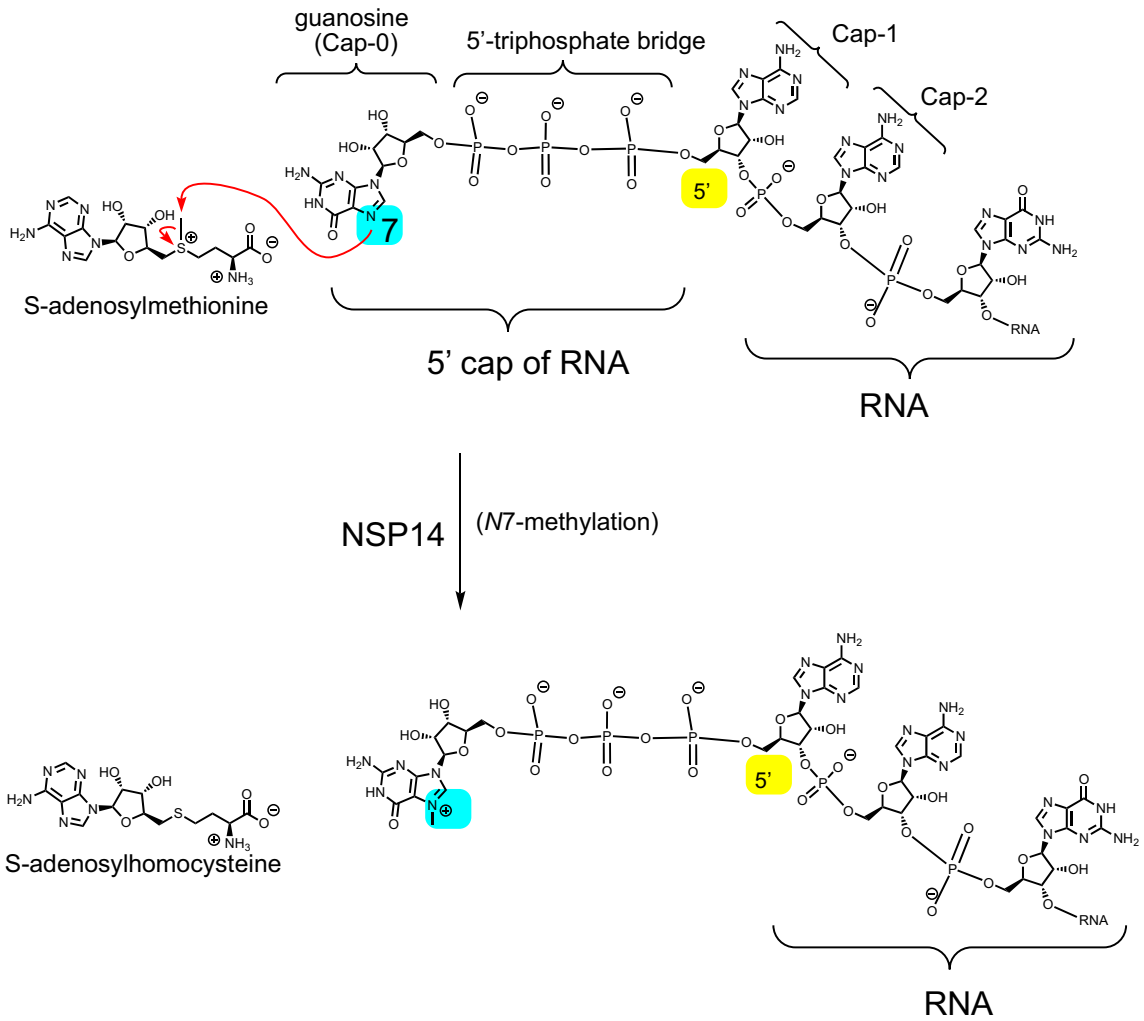
addition to the asparagine residues, O-linked glycosylation of spike proteins have also been observed at residues S325 and T323, which was determined by mass spectrometry [68].

2.6.1 Spike Protein Bound to the Angiotensin Converting Enzyme-2 (ACE2)

Angiotensin converting enzyme-2 (ACE2) is the proposed receptor for SARS CoV-2, and a human recombinant soluble ACE2 has recently been shown to block SARS CoV-2 infection in engineered human tissue [69]. A structure of the spike protein with ACE2 bound has also been reported (PDB ID: 7A98) [70]. For comparison of the ACE2-spike protein complex (PDB ID: 7A98) and spike protein (PDB ID: 7JJJ), a structural overlay is shown in Fig. 27 [60]. Furthermore, the various conformations that the spike protein undergoes upon ACE2 binding have been detected using cryo-electron microscopy [70]. These sequential steps of the trimer include: (i) closed conformation, (ii) open conformation (only one receptor binding domain (RBD) points "up" the other two RBD remain "closed") but unbound to ACE2 (similar to the MERS CoV spike structure provided with PDB ID: 5X5U), (iii) one ACE2 bound to the RBD of one protomer, (iv) two ACE2 bound to two protomers at the RBDs, (v) three ACE2 bound to three protomers at the RBDs, and (vi) release of the monomeric S1-ACE2 complex (S1 unit includes residues 14–685). The S1 unit is first cleaved by the protease, furin [59, 64], from the host cell [63, 71]. The serine protease, transmembrane protease serine 2 (TMPRSS2), is also known to prime the spike protein for cell entry by cleaving at the S2′ site (position 816) [61]. The use of a TMPRSS2 inhibitor, camostat mesylate, blocked SARS-2-S-driven entry into Caco-2 and Vero-TMPRSS2 cells (Fig. 28) [72].

2.6.2 Spike Protein Bound to Antibodies

There have been structures of the spike protein bound to antibodies. These antibodies bind to the RBD of the spike protein. One study showed the antigen-binding fragment



Scheme 4 The reaction catalyzed by NSP14 involving the *N*7-methylation of the guanosine residue of the 5′-cap of viral RNA. The methylating substrate is *S*-adenosylmethionine (SAM), which converts to *S*-adenosylhomocysteine (SAH)

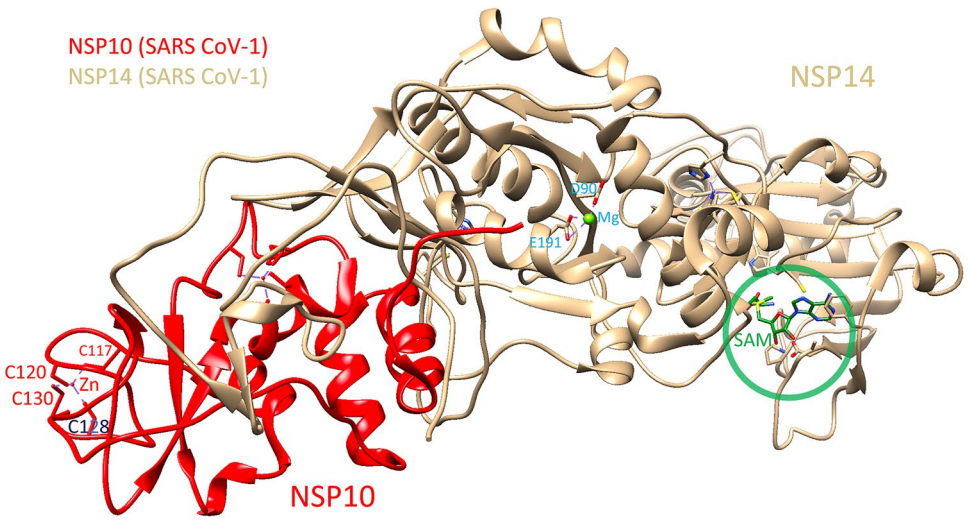
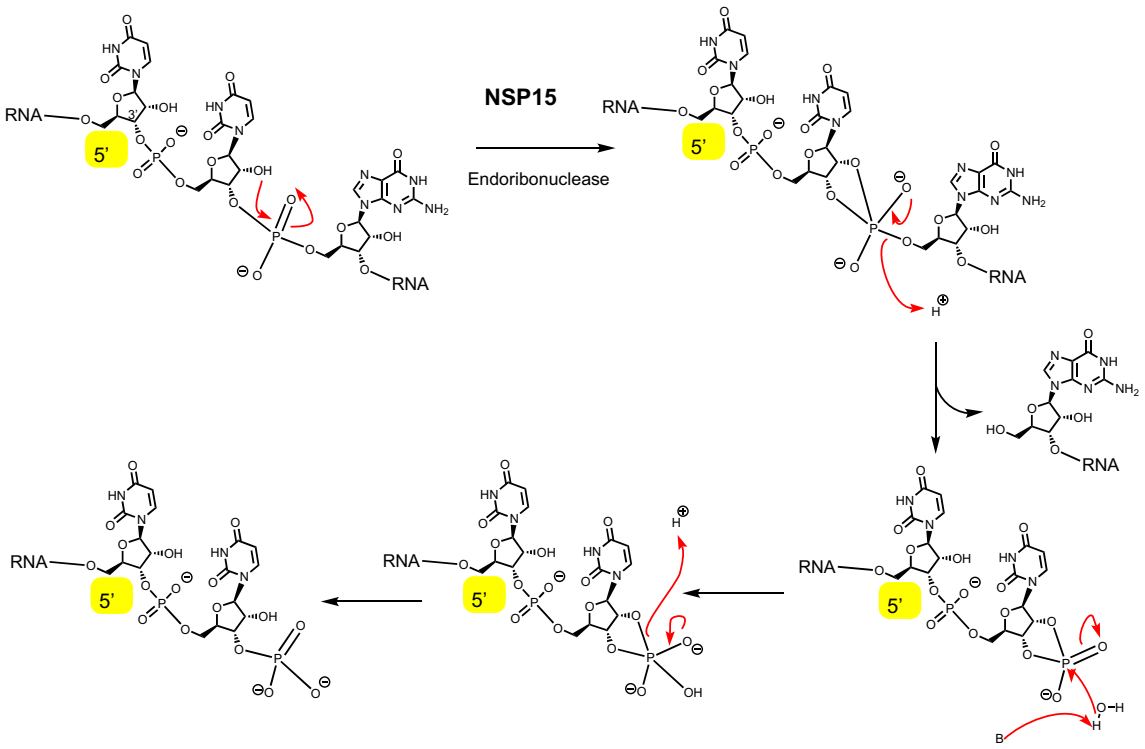


Fig. 14 Structure of NSP14 for SARS CoV-1 in complex with NSP10 (PDB ID: 5C8T). NSP14 in tan (right) and NSP10 is in red (left). An S-adenosylmethionine (SAM) ligand (green) is shown in the complex (circled). The green sphere is a magnesium (II) ion coordinated to the residues, D90 and E191 of NSP14 (Color figure online)



Scheme 5 Endoribonuclease activity of NSP15

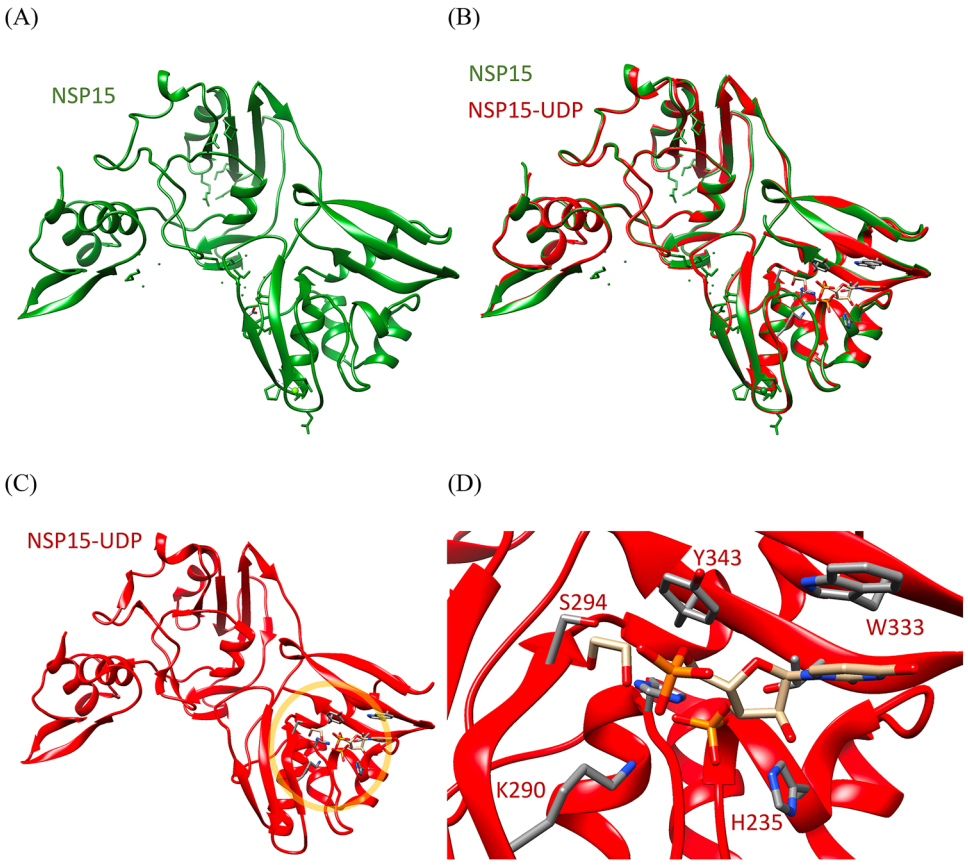


Fig. 15 **a** Structure of apo NSP15 (green, PDB ID: 6VWW) [43]. **b** Structural alignment of NSP15 apo form (green, PDB ID: 6VWWL: 6VWW) and form bound to uridine diphosphate (red, PDB ID: 7K1O) [99]. **c** Structure of NSP15 from SARS CoV-2 bound to uridine diphosphate (red, PDB ID: 7K1O). **d** The expanded view of the active site of NSP15 with uridine diphosphate bound (PDB ID: 7K1O). The green spheres in **a** and **b** are water molecules (Color figure online)

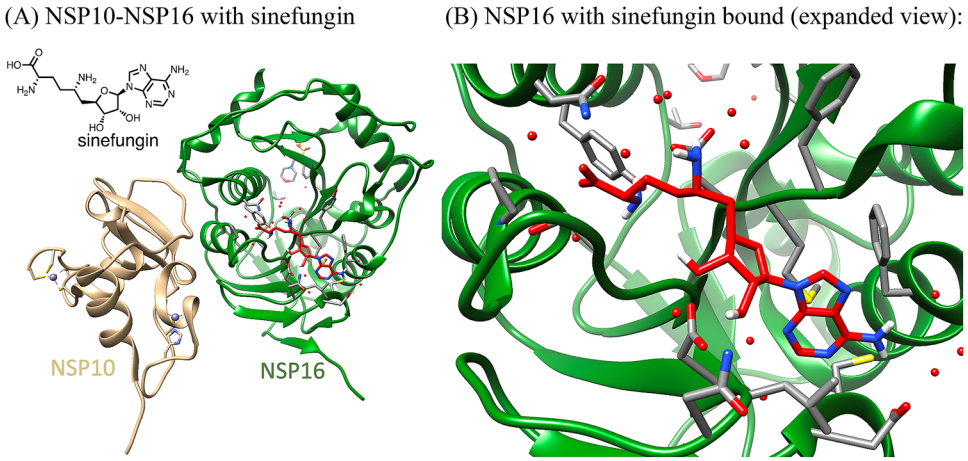
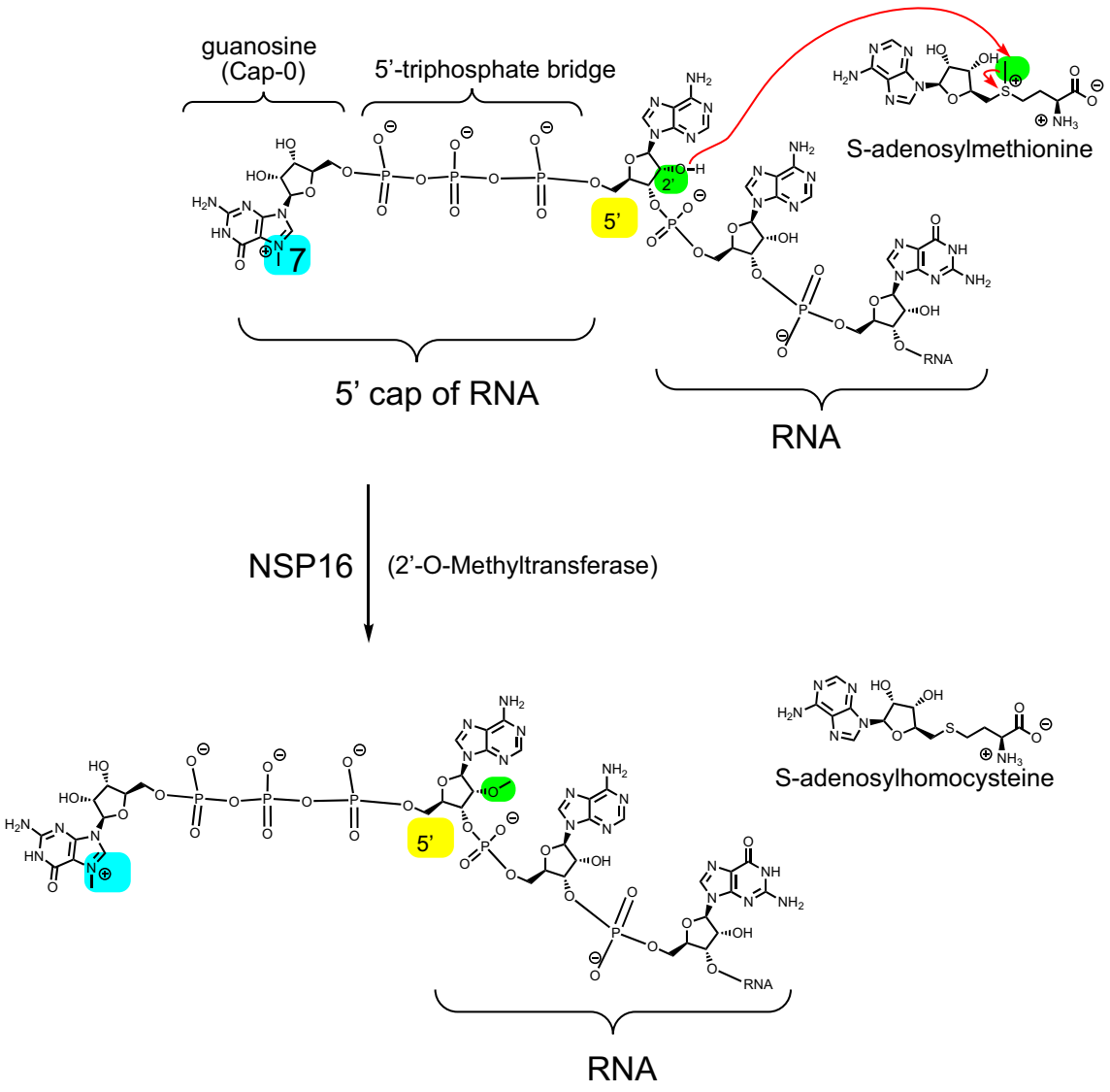


Fig. 16 **a** Structure of NSP10–NSP16 complex with sinefungin bound (PDB ID: 6YZ1). NSP16 is in green. NSP10 is tan. The structure of sinefungin is shown in the top left. **b** shows expanded view of the active site of NSP16 (green) with sinefungin (red) bound. The red spheres are water molecules (Color figure online)

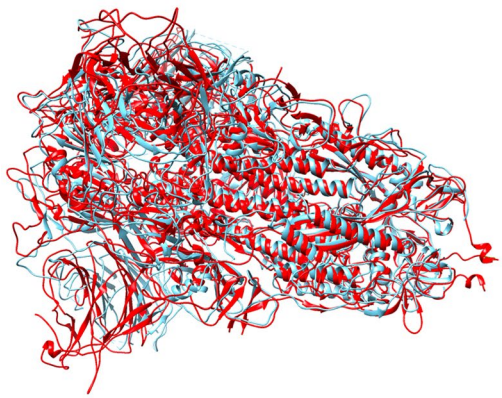


Scheme 6 The reaction catalyzed by NSP16. NSP16 transfers the methyl from *S*-adenosylmethionine to the 2′-*O* position in the 5′-cap of viral RNA

(A)

```
BCA87361.1   1     MFVFLVLLPLVSSQCVNLTTRTQ--LPPAYTN--SFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGT  76
AAP13441.1   1     MFIFLLFLTLTSGSDLDRCTTFDDVQAPNYTQHTSSMRGVYYPDEIFRSDTLYLTQDLFLPFYSNVTGFHTINHT-----  75

BCA87361.1   77    KRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESE  156
AAP13441.1   76    --FGNPVIPFKDGIYFAATEKSNVVRGWVFGSTMNNKSQSVIIINNSTNVVIRACNFELCDNPFFAV----SKPMGTQTH  149

BCA87361.1   157   FRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINIT  236
AAP13441.1   150   TMIFDNAFNCTFEYISDAFSLDVSEKSGNFKHLREFVFKNKDGFLYVYKGYQPIDVVRDLPSGFNTLKPIFKLPLGINIT  229

BCA87361.1   237   RFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTS  316
AAP13441.1   230   NFRAILTA----FSPAQDI--WGTSAAAYFVGYLKPTTFMLKYDENGTITDAVDCSQNPLAELKCSVKSFEIDKGIYQTS  303

BCA87361.1   317   NFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVY  396
AAP13441.1   304   NFRVVPSGDVVRFPNITNLCPFGEVFNATKFPSVYAWERKKISNCVADYSVLYNSTFFSTFKCYGVSATKLNDLCFSNVY  383

BCA87361.1   397   ADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAG  476
AAP13441.1   384   ADSFVVKGDDVRQIAPGQTGVIADYNYKLPDDFMGCVLAWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPD  463

BCA87361.1   477   STPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESN  556
AAP13441.1   464   GKPCTP-PALNCYWPLNDYGFYTTTGIGYQPYRVVVLSFELLNAPATVCGPKLSTDLIKNQCVNFNFNGLTGTGVLTPSS  542

BCA87361.1   557   KKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVY  636
AAP13441.1   543   KRFQPFQQFGRDVSDFTDSVRDPKTSEILDISPCSFGGVSVITPGTNASSEVAVLYQDVNCTDVSTAIHADQLTPAWRIY  622

BCA87361.1   637   STGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPT  716
AAP13441.1   623   STGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVS----LLRSTSQKSIVAYTMSLGADSSIAYSNNTIAIPT  698

BCA87361.1   717   NFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKD  796
AAP13441.1   699   NFSISITTEVMPVSMAKTSVDCNMYICGDSTECANLLLQYGSFCTQLNRALSGIAAEQDRNTREVFAQVKQMYKTPTLKY  778

BCA87361.1   797   FGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSA  876
AAP13441.1   779   FGGFNFSQILPDPLKPTKRSFIEDLLFNKVTLADAGFMKQYGECLGDINARDLICAQKFNGLTVLPPLLTDDMIAAYTAA  858

BCA87361.1   877   LLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNA  956
AAP13441.1   859   LVSGTATAGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKQIANQFNKAISQIQESLTTTSTALGKLQDVVNQNA  938

BCA87361.1   957   QALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQ  1036
AAP13441.1   939   QALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQ  1018

BCA87361.1   1037  SKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIIT  1116
AAP13441.1   1019  SKRVDFCGKGYHLMSFPQAAPHGVVFLHVTYVPSQERNFTTAPAICHEGKAYFPREGVFVFNGTSWFITQRNFFSPQIIT  1098

BCA87361.1   1117  TDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNES  1196
AAP13441.1   1099  TDNTFVSGNCDVVIGIINNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNES  1178

BCA87361.1   1197  LIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT  1273 SARS CoV-2
AAP13441.1   1179  LIDLQELGKYEQYIKWPWYVWLGFIAGLIAIVMVTILLCCMTSCCSCLKGACSCGSCCKFDEDDSEPVLKGVKLHYT  1255 SARS CoV-1
```

(B) SARS-CoV-1 spike (light blue)



(C) SARS-CoV-1 spike (light blue)

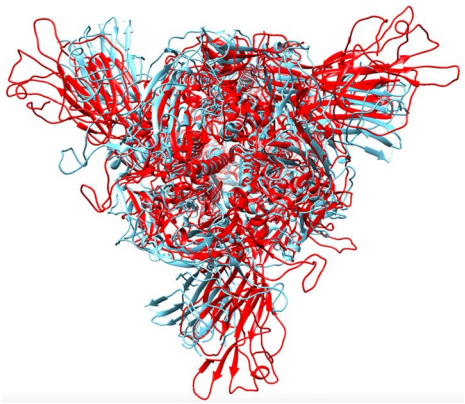


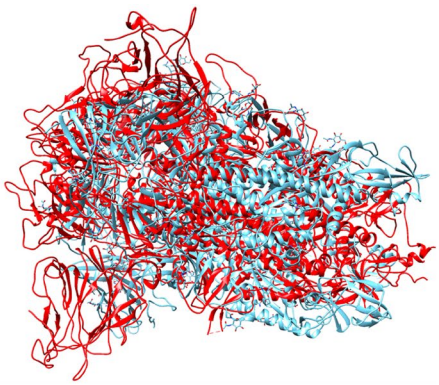
Fig. 17 **a** Primary sequence alignment of SARS CoV-2 (GenBank: BCA87361.1) and SARS CoV-1 (GenBank: AAP13441.1). **b** The structural comparison of the spike proteins from SARS CoV-2 (red, PDB ID: 7JJJ) and SARS CoV-1 (light blue, PDB ID: 5XLR) [55]. AAP13441.1 (now obsolete but previously used: NP_828851.1 [1], where position S577A). **c** Rotated view (Color figure online)

(Fab) fragment of the neutralizing antibody (C105) complexed with the receptor binding domain (RBD) of the spike protein (PDB ID: 6XCM) [73]. Another study reported the spike protein complexed with the S2A4 neutralizing antibody Fab fragment (PDB ID: 7JVC). These structures are available as determined by cryo-electron microscopy (cryo-EM) and their superpositions with the unbound spike protein (PDB ID: 7JJJ) are shown in Fig. 29.

(A)

```
BCA87361.1   1     MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFT----------RGVYYPDK---------VFRSSVLHSTQDLFLPFFSN  61
QOP39313.1   1     MFVLLVAYALLHIAGCQTTNGTNTSHSVCNGCVGHSENVFDVESGGYIPSNFAFNNWFLLTNTSSVVDGVVRSFQPLLLN  80

BCA87361.1   62    VTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFL  141
QOP39313.1   81    CLW----SVSGSQVITGF----------VYFNGTGRG-ACKGFYSNASSDVIRYNINFEENLRRGTI------LFKTSYG  139

BCA87361.1   142   GVYYHKNNKSWMESEFRV---------YSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINL  212
QOP39313.1   140   AVVFYCTNNTLVSGDAHIPSGTVLGNFYCFVNTTIGNETTSAFVGALP------KTVREFVISRTGHFYINGYRYFSLGD  213

BCA87361.1   213   VRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCA  292
QOP39313.1   214   VEA---------------VNFNVTNAAT--------------TDFCTVALASY--------ADVLVNVSQTAIANIIYCN  256

BCA87361.1   293   LDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSA  372
QOP39313.1   257   -SVINRLRCDQLSFDVPDVFYSTS--------------------PIQPVELPESIVSLPVYHKHTFIVLHVKFE-----H  310

BCA87361.1   373   SFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYL  452
QOP39313.1   311   GPGPGKCYNCRPAVIN-ITLANFNETKGPLCVDTSHFTTKYVAVYANVGRWSASINTGNCPFSFGKVNNFVKFG------  383

BCA87361.1   453   YRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGF--NCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKS  530
QOP39313.1   384   -------------------------SVCFSLKAIPGGCAMPIMANLVNYKSHNIGSLYVSWSDGDVITGVPKPVEGVSSF  438

BCA87361.1   531   TNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVA-  609
QOP39313.1   439   MNVTLNKCTKYNIYDVSGVGVIRISNDTFLNGITYTSTSGNLL-GFKDVTNGTIYSITPCNPPDQLVVYQQAVVGAMLSE  517

BCA87361.1   610   --VLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPI----GAGICASYQTQTNSPRR  683
QOP39313.1   518   NFTSYGFSNVVEMPKFFYAS----------------------------NGTYNCTDAVLTYSSFGVCADGSIIAVQPRN  568

BCA87361.1   684   ARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQL  763
QOP39313.1   569   VSYDSVSAIVT---------------ANLSIPSNWTTSVQVEYLQITSTPIVVDCSTYVCNGNVRCVELLKQYTSACKTI  633

BCA87361.1   764   NRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILP-DPSKPS---KRSFIEDLLFNKVTLADAGFIK-QYG  838
QOP39313.1   634   EDALRNSAMLESADVSEMLTFDKKAFTLANVSSFGDYNLSSVIPSLPRSGSRVAGRSAIEDILFSKLVTSGLGTVDADYK  713

BCA87361.1   839   DCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYE  918
QOP39313.1   714   KCTKGLSIADLACAQYYNGIMVLPGVADAERMAMYTGSLIGGIALGGLT----SAASIPFSLAIQSRLNYVALQTDVLQE  789

BCA87361.1   919   NQKLIANQFNSAIGKI--------------QDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRL  984
QOP39313.1   790   NQRILAASFNKAMTNIVDAFTGVNDAITQTSQALQTVATALNKIQDVVNQQGNSLSHLTSQLRQNFQAISSSIQAIYDRL  869

BCA87361.1   985   DKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLH  1064
QOP39313.1   870   DIIQADQQVDRLITGRLAALNVFVSHTLTKYTEVRASRQLAQQKVNECVKSQSKRYGFCGNGTHIFSLVNAAPEGLVFLH  949

BCA87361.1   1065  VTYVPAQEKNFTTAPAICHDGKAHFPREG---VFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPL  1141
QOP39313.1   950   TVLLPTQYKDVEAWSGLCVDGINGYVLRQPNLALYKEGNYYRITSRIMFEPRIPTIADFVQIENCNVTFVNISRSELQTI  1029

BCA87361.1   1142  QPELDSFKEELDKY---FKNHTSPDVDLGDISGINASVVNIQKEIDRLNE--------------VAKNLNESLIDLQELG  1204
QOP39313.1   1030  VPEYIDVNKTLQELSYKLPNYTVPDLVVEQ---YNQTILNLTSEISTLENKSAELNYTVQKLQTLIDNINSTLVDLKWLN  1106

BCA87361.1   1205  KYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCS--CGSCCKFDEDDSEPVLKGVKLHYT  1273 SARS CoV-2
QOP39313.1   1107  RVETYIKWPWWVWLCISVVLIFVVSMLLLCCCSTGCCGFF-SCFASSIRGCCE---STKLPYYDVEKIHIQ  1173 HCoV-229E
```

(B) HCoV-229E spike (light blue)



(C) HCoV-229E spike (light blue)

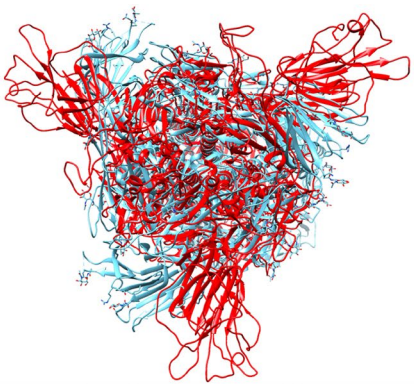


Fig. 18 **a** Primary sequence alignment of SARS CoV-2 (GenBank: BCA87361.1) and HCoV-229E (GenBank: QOP39313.1). **b** The structural comparison of the spike proteins from SARS CoV-2 (red, PDB ID: 7JJJ) and HCoV-229E (PDB ID: 6U7H) [52]. **c** Rotated view (Color figure online)

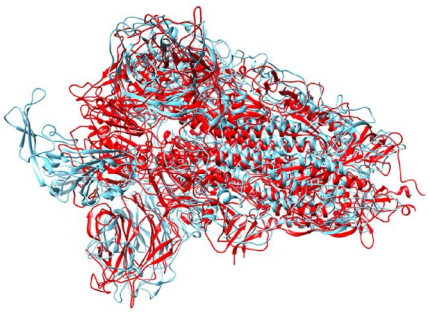
2.6.3 The D614G Variant of the Spike Protein

From the primary sequence alignment of the spike proteins, the D614 residue is conserved in both SARS CoV-1 and SARS CoV-2. However, a SARS CoV-2 strain containing a glycine residue (G) at position 614 has been suggested to be more globally widespread. A structure of this variant without the receptor binding domain (RBD) is available and it

(A)

```
BCA87361.1  1     MF--VFLVLLPLVSSQ-------------CVNLTTRTQL-----PPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPF--  58
ASU91305.1  1     MIHSVFLLMFLLTPTESYVDVGPDSVKPACIEVDIQQTFFDKTWPRPIDVSKADGIIYPQGRTYSNITITYQGLF-PYQG  79

BCA87361.1  59    -FSNVTWFHAIHVSGTNGTKRF----DNPVLPF-------------NDGVYFASTEKSNIIR--------GWIFGTTLDS  112
ASU91305.1  80    DHGDMYVYSAGHATGTTPQKLFVANYSQDVKQFANGFVVRIGAAANSTGTVIISPSTSATIRKIYPAFMLGSSVGNFSDG  159

BCA87361.1  113   K-----TQSLLIVNNATNVVIKVCEFQFCNDPFLGVY------------YH---------KNNKSWMESEFRVYSSANNC  166
ASU91305.1  160   KMGRFFNHTLVLLPDGCGTLLRA--FYCILEPRSGNHCPAGNSYTSFATYHTPATDCSDGNYNRNASLNSFKEYFNLRNC  237

BCA87361.1  167   TFEY---VSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLA  243
ASU91305.1  238   TFMYTYNITEDEILEWFGITQTAQGVHLFSSRYVDLY-----------------GGNMFQ-FATLPVYDTIKYYSIIPH  298

BCA87361.1  244   LHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPT  323
ASU91305.1  299   SIRSIQS---DRKAW----AAFYVYKLQPLTFLLDFSVDGYIRRAIDCGFNDLSQLHCSYESFDVESGVYSVSSFEAKPS  371

BCA87361.1  324   ESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIR  403
ASU91305.1  372   GSVVEQAEGVE-CDFSPLLSGTP-PQVYNFKRLVFTNCNYNLTKLLSLFSVNDFTCSQISPAAIASNCYSSLILDYFSYP  449

BCA87361.1  404   GDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGS-TPC--  480
ASU91305.1  450   LSMKSDLSVTSAGPISQFNYKQSFSNPTCLILATVPHNLTTITKPLKYSY-INKCSRLLSDDRTEVPQLVNANQYSPCVS  528

BCA87361.1  481   --------NGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPAT----VCGPKKSTNLVK-----NKCVNFNF  543
ASU91305.1  529   IVPSTVWEDGDYYRKQLSPLEGGGWLVASGSTVAMTEQLQMGFGITVQYGTDTNSVCPKLEFANDTKIASQLGNCVEYSL  608

BCA87361.1  544   NGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVA  623
ASU91305.1  609   YGVSGRGVFQNCTAVGVRQQRFVYDAYQNLVGYYSDDG-NYYCLRACVSVPVSVIYDKETKTH--ATLFGSVACEHISST  685

BCA87361.1  624   IHADQLTPTWRVYSTGSNV--FQTRAGCLIGAEHVNNSY---ECDIPIGAGICASYQT-QTNSPRRARSVASQSIIAYTM  697
ASU91305.1  686   MSQYSRSTRSMLKRRDSTYGPLQTSVGCVLGL--VNSSLFVEDCKLPLGQSLCALPDTPSTLTPRSVRSVPGEMRLA-SI  762

BCA87361.1  698   SLGAENSVAYSNNS---IAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQ  774
ASU91305.1  763   AFNHPIQVDQLNSSYFKLSIPTNFSFGVTQEYIQTTIQKVTVDCKQYVCNGFQKCEQLLREYGQFCSKINQALHGANLRQ  842

BCA87361.1  775   DKNTQEVFAQVKQIYKTPPIKDFGG-FNFSQILP-DPSKPS--KRSFIEDLLFNKVTLADAGFIKQYGDCL--GDIAARD  848
ASU91305.1  843   DDSVRNLFASVKSSQSSPIIPGFGGDFNLTLLEPVSISTGSRSARSAIEDLLFDKVTIADPGYMQGYDDCMQQGPASARD  922

BCA87361.1  849   LICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFN  928
ASU91305.1  923   LICAQYVAGYKVLPPLMDVNMEAAYTSSLLGSIAGVGWTAGLSSFAAIPFAQSIFYRLNGVGITQQVLSENQKLIANKFN  1002

BCA87361.1  929   SAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYV  1008
ASU91305.1  1003  QALGAMQTGFTTTNEAFRKVQDAVNNNAQALSKLASELSNTFGAISASIGDIIQRLDVLEQDAQIDRLINGRLTTLNAFV  1082

BCA87361.1  1009  TQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAIC---HDG  1085
ASU91305.1  1083  AQQLVRSESAALSAQLAKDKVNECVKAQSKRSGFCGQGTHIVSFVVNAPNGLYFMHVGYYPSNHIEVVSAYGLCDAANPT  1162

BCA87361.1  1086  KAHFPREGVFVSNGTHWFV-----TQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQP--ELDSFKEELDKYFKN  1158
ASU91305.1  1163  NCIAPVNGYFIKTNNTRIVDEWSYTGSSFYAPEPITSLNT-KYVAPQVTYQNISTNLPPLLGNSTGIDFQDELDEFFKN  1241

BCA87361.1  1159  HTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMT  1238
ASU91305.1  1242  VSTSIPNFGSLTQINTTLLDLTYEMLSLQQVVKALNESYIDLKELGNYTYYNKWPWYIWLGFIAGLVALALCVFFILCCT  1321

BCA87361.1  1239  SCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT  1273 SARS CoV-2
ASU91305.1  1322  GCGTNCMGKLKCNRCCDRYEEYD---LEPHKVHVH  1353 MERS CoV
```

(B) MERS CoV spike (light blue)



(C) MERS CoV spike (light blue)

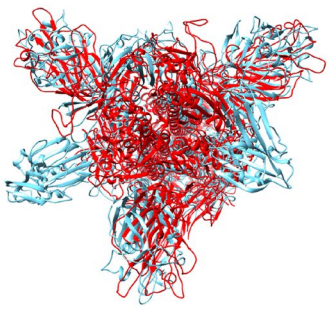


Fig. 19 **a** Primary sequence alignment of SARS CoV-2 (GenBank: BCA87361.1) and MERS-CoV (GenBank: ASU91305.1). **b** The structural comparison of the spike proteins from SARS CoV-2 (red, PDB ID: 7JJJ) and MERS CoV (PDB ID: 5X5U, open RBD conformation) [56]. **c** Rotated view of (**a**) (Color figure online)

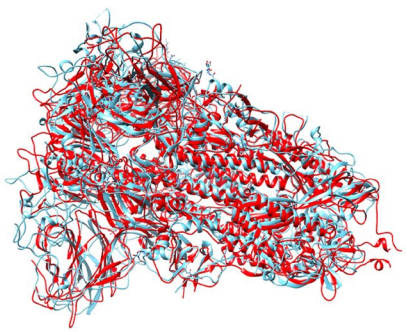
has been suggested that the D614G variant (PDB ID: 6XS6) of the spike protein more frequently adopts an open conformation compared to the D614 variant [74]. The structure of the G614 variant of the spike protein that lacks the RBD (PDB ID: 6XS6) is superimposed with the D614 version of the spike protein in Fig. 30.

From a careful look at the spike protein structure, D614 forms a salt bridge with R634 (2.9 angstroms, cf. Fig. 31b). Furthermore, a lysine residue (K854) is located on a separate protomer that appears to interact with D614 (7.4 angstroms away). In the G614 variant, these salt bridges are absent, which suggests that the spike protein trimer is held less tightly together when the aspartate (D) is a glycine (G). This "looser" conformation with the G614 variant could possibly explain how the D614G strain is more infectious than the wild type. Furthermore, in the complex of ACE2 and spike (PDB ID: 7A98), the ionic interaction between D614 and K854 is enhanced when the distance was measured to be 4.1 angstroms (Fig. 31d). In the structure of the ACE2-spike complex (PDB ID: 7A98), the residue R634 is not

(A)

```
BCA87361.1  1     MFVFLVLL-------------------------PLVSSQCVNLT----TRTQLPPAYTNS--FTRGVY------YPDKVF  43
AAA03055.1  1     MFLILLISLPTAFAVIGDLKCTSDNINDKDTGPPPISTDTVDVTNGLGTYYVLDRVYLNTTLFLNGYYPTSGSTYRNMAL  80

BCA87361.1  44    RSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVL-----PFNDGVYFASTEKSNIIRGWIFGTTLD--SKTQS  116
AAA03055.1  81    KGSVLLSRLWFKPPFLSD--FINGIFAKVKN-TKVIKDRVMYSEFPAITIGSTFVNTSYSVVVQPRTINSTQDGDNKLQG  157

BCA87361.1  117   LLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKN  196
AAA03055.1  158   LLEVS--------VCQYNMCEYP--QTICHPNLGNHRKELWHLDTGVVSCLYK---RNFTYDVNA------DYLYFHFYQ  218

BCA87361.1  197   IDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFL  276
AAA03055.1  219   EGGTFYAYFTDTGV----------VTKFLFNVYLG---------MALSHYYVMPLTCNSKLTL---EYWVTPLTSRQYL  275

BCA87361.1  277   LKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRF-PNITNLCPFGEVFNATRFASVYAWNR  355
AAA03055.1  276   LAFNQDGIIFNAEDCMSDFMSEIKCKTQSIAPPTGVYELNGYTVQPIADVYRRKPNLPN-CNIEAWLNDKSVPSPLNWER  354

BCA87361.1  356   KRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIA  435
AAA03055.1  355   KTFSNCNFNMSSLMSFIQADSFTCNNIDAAKIYGMCFSSITIDKFAIPNGRKVDLQLGNLGYLQSFNYRIDTTATSCQLY  434

BCA87361.1  436   WNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSF  515
AAA03055.1  435   YNLPAANVSVSR-------------------------FNP----STWNKRFGFIEDSVFKPRPAGVLTNHDVVYAQ  481

BCA87361.1  516   ELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILD--------  586
AAA03055.1  482   HCFKAPKNFCPCKLN-----GSCVGSGPGKNNGIGTCPAGTNYL-----------TCDNLCTPDPITFTGTYKCPQTKS  544

BCA87361.1                  ----------------------------------------------------------------------
AAA03055.1  545   LVGIGEHCSGLAVKSDYCGGNSCTCRPQAFLGWSADSCLQGDKCNIFANFILHDVNSGLTCSTDLQKANTDIILGVCVNY  624

BCA87361.1  587   ----------------------------------------------ITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEV  620
AAA03055.1  625   DLYGILGQGIFVEVNATYYNSWQNLLYDSNGNLYGFRDYIINRTFMIRSCYSGRVSAAFHAN--SSEPALLFRNIKCNYV  702

BCA87361.1  621   PVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVN--NSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMS  698
AAA03055.1  703   FNNSLTRQLQPI--------NYFDSYLGCVVNAYNSTAISVQTCDLTVGSGYCVDYSKN----RRSRGAITTGYRFTNFE  770

BCA87361.1  699   LGAENSVAYSNN------SIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAV  772
AAA03055.1  771   PFTVNSVNDSLEPVGGLYEIQIPSEFTIGNMVEFIQTSSPKVTIDCAAFVCGDYAACKSQLVEYGSFCDNINAILTEVNE  850

BCA87361.1  773   EQDKNTQEVFAQVKQIYKTPPI------KDFGGFNFSQILP----DPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLG  842
AAA03055.1  851   LLDTTQLQVANSLMNGVTLSTKLKDGVNFNVDDINFSPVLGCLGSECSKASSRSAIEDLLFDKVKLSDVGFVEAYNNCTG  930

BCA87361.1  843   DIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKL  922
AAA03055.1  931   GAEIRDLICVQSYKGIKVLPPLLSENQISGYTLAATSASLFPPWT----AAAGVPFYLNVQYRINGLGVTMDVLSQNQKL  1006

BCA87361.1  923   IANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQ  1002
AAA03055.1  1007  IANAFNNALYAIQEGFDATNSALVKIQAVVNANAEALNNLLQQLSNRFGAISASLQEILSRLDALEAEAQIDRLINGRLT  1086

BCA87361.1  1003  SLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAIC  1082
AAA03055.1  1087  ALNAYVSQQLSDSTLVKFSAAQAMEKVNECVKSQSSRINFCGNGNHIISLVQNAPYGLYFIHFSYVPTKYVTARVSPGLC  1166

BCA87361.1  1083  HDG-KAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHT-  1160
AAA03055.1  1167  IAGDRGIAPKSGYFVNVNNTWMYTGSGYYYPEPITENNVVVMSTCAVNYTKAPYVMLNTSIPNLPDFKEELDQWFKNQTS  1246

BCA87361.1  1161  -SPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTS  1239
AAA03055.1  1247  VAPDLSLDY---INVTFLDLQVEMNRLQEAIKVLNQSYINLKDIGTYEYYVKWPWYVWLLICLAGVAMLVLFFICCCTG  1323

BCA87361.1  1240  CCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT  1273 SARS CoV-2
AAA03055.1  1324  CGTSCFKKC--GGCCDDYTGYQELVIK--TSHDD  1353 HCoV-OC43
```

(B) HCoV OC43 spike (light blue)



(C) HCoV OC43 spike (light blue)

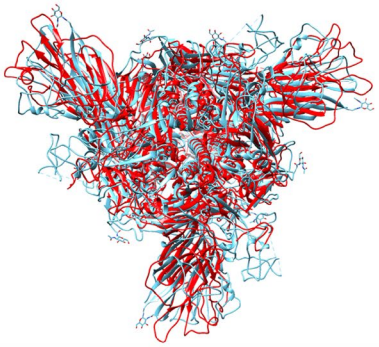


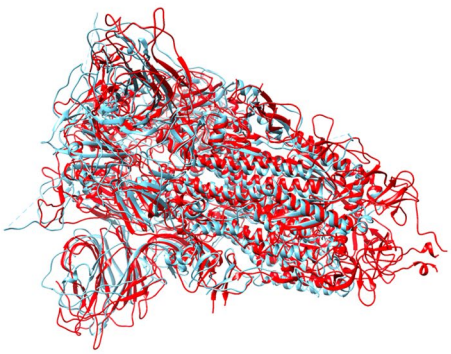
Fig. 20 **a** Primary sequence alignment of SARS CoV-2 (GenBank: BCA87361.1) and HCoV OC43 (GenBank: AAA03055.1). **b** The structural comparison of the spike proteins from SARS CoV-2 (red, PDB ID: 7JJJ) and HCoV OC43 (PDB ID: 6OHW) [52]. **c** Rotated view of (**a**) (Color figure online)

included. This lack of interaction between D614 and R634 in the ACE2-spike complex confirms that the D614 residue plays a role in keeping the spike protein more "compact" so that the receptor binding domain (RBD) is less likely to reach out to bind to the ACE2 protein. In contrast, with the G614 variant, which lacks these interhelical salt bridge interactions, the spike protein is more free to be in the "open" conformation to interact with the ACE2 receptor.

(A)

```
BCA87361.1  1     MFVFLVLLPLVSS-----QCVNL------TTRTQLPPAYTN-SFTRGVYYP-DKVFRSSVLHSTQDLFLPFF-SNVTWFH  66
ADN03339.1  1     MLLIIFILPTTLAVIGDFNCTNFAINDLNTTIPRISEYVVDVSYGLGTYYILDRVYLNTTI-----LFTGYFPKSGANFR  75

BCA87361.1  67    AIHVSGTN--GTKRFDNPVLP-FNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVV--------IKVCEFQF  135
ADN03339.1  76    DLSLKGTTKLSTLWYQKPFLSDFNNGIFSRVKNTKLYVNKTLYSEFSTIVIGSVFINNSYTIVVQPHNGVLEITACQYTM  155

BCA87361.1  136   CNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFK-NLREFVFKNIDGYFKIYSKHTPINLVR  214
ADN03339.1  156   CEYP------HTICKS-------IGSSRNESWHFDKSEPLCLFKKNFTYNVSTDWLYFHFYQERGTFYAYYADS------  216

BCA87361.1  215   DLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALD  294
ADN03339.1  217   GMPTTF-----LFSLYLGTLLSHYYVLP------LTCNAISSNTDNETLQYWVTPLSKRQYLLKFDDRGVITNAVDCSSS  285

BCA87361.1  295   PLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVR-FPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSAS  373
ADN03339.1  286   FFSEIQCKTKSLLPNTGVYDLSGFTVKPVATVHRRIPDLPD-CDIDKWLNNFNVPSPLNWERKIFSNCNFNLSTLLRLVH  364

BCA87361.1  374   FSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLY  453
ADN03339.1  365   TDSFSCNNFDESKIYGSCFKSIVLDKFAIPNSRRSDLQLGSSGFLQSSNYKIDTTSSSCQLYYSLPAINVTIN-NYN---  440

BCA87361.1  454   RLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNL  533
ADN03339.1  441   ----------------------------PSSWNRR-----YGFNN----FNLSSHSVVYSRYCFSVNNTFCPCAKPS--  480

BCA87361.1  534   VKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCS--------------FG------  593
ADN03339.1  481   FASSCKSHK-PPSASCPIGTNYRSCESTTVL---DHTDWCRCSCLPDPITAYDPRSCSQKKSLVGVGEHCAGFGVDEEKC  556

BCA87361.1  594   GV---------------------------------------SVITPGTNTSNQV----------------------------  608
ADN03339.1  557   GVLDGSYNVSCLCSTDAFLGWSYDTCVSNNRCNIFSNFILNGINSGTTCSNDLLQPNTEVFTDVCVDYDLYGITGQGIFK  636

BCA87361.1  609   ---AVLYQ-------DVN----------------------CTEVPVAIH--ADQLTPTWRVYSTG---SNV-------FQT  645
ADN03339.1  637   EVSAVYYNSWQNLLYDFNGNIIGFKDFVTNKTYNIFPCYAGRVSAAFHQNASSLALLYRNLKCSYVLNNISLATQPYFDS  716

BCA87361.1  646   RAGCLIGAEHVNNSY--ECDIPIGAGICASYQTQTNSP--RRARSVASQSIIAYTMSLGAENSVAYSNNS---------I  712
ADN03339.1  717   YLGCVFNADNLTDYSVSSCALRMGSGFCVDYNSPSSSSSRRKRRSISA----SYRFVTFEPFNVSFVNDSIESVGGLYEI  792

BCA87361.1  713   AIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQI----  788
ADN03339.1  793   KIPTNFTIVGQEEFIQTNSPKVTIDCSLFVCSNYAACHDLLSEYGTFCDNINSILDEVNGLLDTTQLHVADTLMQGVTLS  872

BCA87361.1  789   --YKTPPIKDFGGFNFSQI---LPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPP  863
ADN03339.1  873   SNLNTNLHFDVDNINFKSLVGCLGPHCGSSSRSFFEDLLFDKVKLSDVGFVEAYNNCTGGSEIRDLLCVQSFNGIKVLPP  952

BCA87361.1  864   LLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTAS  943
ADN03339.1  953   ILSESQISGYTTAATVAAMFPPWS----AAAGIPFSLNVQYRINGLGVTMDVLNKNQKLIATAFNNALLSIQNGFSATNS  1028

BCA87361.1  944   ALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASAN  1023
ADN03339.1  1029  ALAKIQSVVNSNAQALNSLLQQLFNKFGAISSSLQEILSRLDALEAQVQIDRLINGRLTALNAYVSQQLSDISLVKLGAA  1108

BCA87361.1  1024  LAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHF-PREGVFVSNGTHW  1102
ADN03339.1  1109  LAMEKVNECVKSQSPRINFCGNGNHILSLVQNAPYGLLFMHFSYKPISFKTVLVSPGLCISGDVGIAPKQGYFIKHNDHW  1188

BCA87361.1  1103  FVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLG-DISGINASVVNIQK  1181
ADN03339.1  1189  MFTGSSYYYPEPISDKNVVFMNTCSVNFTKAPLVYLNHSVPKLSDFESELSHWFKNQTSIAPNLTLNLHTINATFLDLYY  1268

BCA87361.1  1182  EIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDS  1261
ADN03339.1  1269  EMNLIQESIKSLNNSYINLKDIGTYEMYVKWPWYVWLLISFSFIIFLVLLFFICCCTGCGSAC--FSKCHNCCDEYGGHH  1346

BCA87361.1  1262  EPVLKGVKLHYT  1273 SARS CoV-2
ADN03339.1  1347  DFVIK--TSHDD  1356 HCoV HKU1
```

(B) HCoV HKU1 spike (light blue)



(C) HCoV HKU1 spike (light blue)

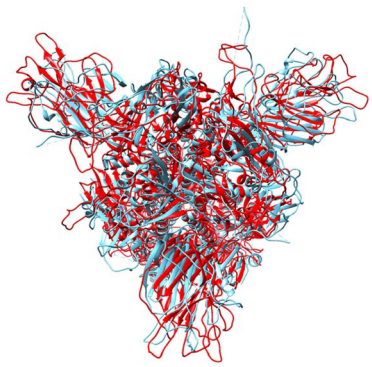


Fig. 21 **a** Primary sequence alignment of SARS CoV-2 (GenBank: BCA87361.1) and HCoV HKU1 (GenBank: ADN03339.1). **b** The structural comparison of the spike proteins from SARS CoV-2 (red, PDB ID: 7JJJ) and HCoV HKU1 (PDB ID: 5I08) [57]. **c** Rotated view (Color figure online)

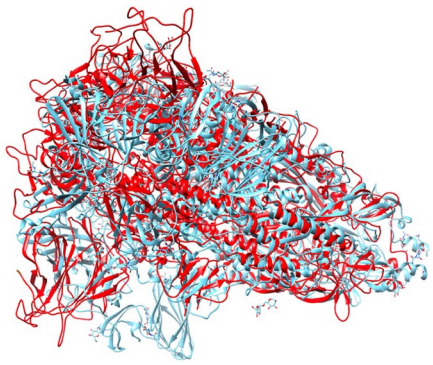
2.6.4 The 501Y.V1 Variant of the Spike Protein

In addition to the strain that contained a D614G change in the spike protein, another SARS CoV-2 strain possessing an N501Y (asparagine to tyrosine change at position-501) was reported in the United Kingdom as 10% more transmissible than the original strain (501N lineage). In fact, an even more transmissible strain, which was 75% more transmissible than

(A)

```
BCA87361.1  1     MFVFLVLLPLVSSQCVNL------------------TTRTQLPPAY---TNSFT-----RGVYYPDKVFRSSV--LHST  51
AGT51394.1  1     MKLFLILLVLPLASCFSTCNSNANISMLQLGVPDNSSTIVTGLLPTHWICANQSTSVYSAKGFFYIDVGNHRSAFALHTG  80

BCA87361.1  52    QDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIR-------GWIFGTTLDSKTQSLLIVN---  121
AGT51394.1  81    YYDANQYYIYVTNEIGLNASVTLKICKFRNTTFDFLSN---ASSSFDCIVNLLFTEQLGAPLGITISGETVRLHLYNVTR  157

BCA87361.1  122   ----------------------------NATNVVI------KVCEFQFCND---------------------PFLGVYYH  146
AGT51394.1  158   TFYVPAAYKLTKLSVKCYFNYSCVFSVVNATVTVNVTTHNGRVVNYTVCDDCNGYTDNIFSVQQDGRIPNGFPFNNWFLL  237

BCA87361.1  147   KNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVF-------KNIDGYFKIYSKHTPINLVRDLPQG  219
AGT51394.1  238   TNGSTLVDGVSRLY------------QPLRLTCLWPVPGLKSSTGFVYFNATGSDVNCNGY-----QHNSVADVMRYNLN  300

BCA87361.1  220   FSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSET  299
AGT51394.1  301   FSA-NSVDNLKSGVIV--FKTLQYDVLFYCS--NSSSG-VLDTTIPFGPSSQPYYCFINSTINTTHVSTFVGILPP----  370

BCA87361.1  300   KCTLKSFTVEK-GIYQTSNFRVQPT---ESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFS  375
AGT51394.1  371   --TVREIVVARTGQFYINGFKYFDLGFIEAVNFNVTTASATDFWTVAFATFVDVLVNVSATNIQN------LLYCDSPFE  442

BCA87361.1  376   TFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRL  455
AGT51394.1  443   KLQCEHLQFGLQDGFYSANFLDDNVLPETYVALPIYYQHT----DINFTATASFGGSCYVCKPRQVNISLNGNTSVCVRT  518

BCA87361.1  456   FRKSNLKPFER-------DISTEIY-QAGSTP-----CNGVEGF-----------NCYFPLQSYGFQPTNGVGYQPYRV  510
AGT51394.1  519   SHFSIRYIYNRVKSGSPGDSSWHIYLKSGTCPFSFSKLNNFQKFKTICFSTVEVPGSCNFPLEATWHYTSYTIVGALYVT  598

BCA87361.1  511   VVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPC  590
AGT51394.1  599   WSEGNSITGVPYPVSGIREFSNLVLNNCTKYNIYDYVGTGIIRSSNQSLAGGITYVSNSGN-----------LLGFKNV  666

BCA87361.1  591   SFGGVSVITPGTNTSNQVAVLYQDV--NCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGA  668
AGT51394.1  667   STGNIFIVTP-CNQPDQVAVYQQSIIGAMTAVNESRYGLQ-----------NLLQLPNFYYVSNGGNNCTTAVMTYSNF  733

BCA87361.1  669   GICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTE  748
AGT51394.1  734   GICADGSLIPVRPRNSSDNGISAIIT--------------ANLSIPSNWTTSVQVEYLQITSTPIVVDCATYVCNGNPR  798

BCA87361.1  749   CSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPD----PSKPSKRSFIEDLLFN  824
AGT51394.1  799   CKNLLKQYTSACKTIEDALRLSAHLETNDVSSMLTFDSNAFSLANVTSFGDYNLSSVLPQRNIHSSRIAGRSALEDLLFS  878

BCA87361.1  825   KVTLADAGFIK-QYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMA  903
AGT51394.1  879   KVVTSGLGTVDVDYKSCTKGLSIADLACAQYYNGIMVLPGVADAERMAMYTGSLIGGMVLGGLT----SAAAIPFSLALQ  954

BCA87361.1  904   YRFNGIGVTQNVLYENQKLIANQFNSAIGKI--------------QDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSN  969
AGT51394.1  955   ARLNYVALQTDVLQENQKILAASFNKAINNIVASFSSVNDAITQTAEAIHTVTIALNKIQDVVNQQGSALNHLTSQLRHN  1034

BCA87361.1  970   FGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHL  1049
AGT51394.1  1035  FQAISNSIQAIYDRLDSIQADQQVDRLITGRLAALNAFVSQVLNKYTEVRGSRRLAQQKINECVKSQSNRYGFCGNGTHI  1114

BCA87361.1  1050  MSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREG---VFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNC  1126
AGT51394.1  1115  FSIVNSAPDGLLFLHTVLLPTDYKNVKAWSGICVDGIYGYVLRQPNLVLYSDNGVFRVTSRVMFQPRLPVLSDFVQIYNC  1194

BCA87361.1  1127  DVVIGIVNNTVYDPLQPELDSFKEELDKY---FKNHTSPDVDLGDISGINASVVNIQKEIDRLNE--------------V  1189
AGT51394.1  1195  NVTFVNISRVELHTVIPDYVDVNKTLQEFAQNLPKYVKPNFDLTP---FNLTYLNLSSELKQLEAKTASLFQTTVELQGL  1271

BCA87361.1  1190  AKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCS--CGSCCKFDEDDSEPVLKG  1267
AGT51394.1  1272  IDQINSTYVDLKLLNRFENYIKWPWWVWLIISVVFVVLLSLLVFCCLSTGCCGCC-NCLTSSMRGCCDCGST-KLPYYEF  1349

BCA87361.1  1268  VKLHYT  1273 SARS CoV-2
AGT51394.1  1350  EKVHVQ  1355 HCoV NL63
```

(B) HCoV NL63 spike (light blue)



(C) HCoV NL63 spike (light blue)

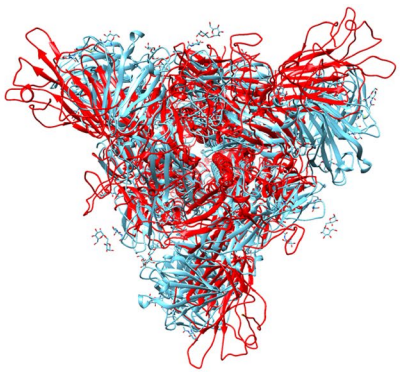


Fig. 22 **a** Primary sequence alignment of SARS CoV-2 (GenBank: BCA87361.1) and HCoV NL63 (GenBank: AGT51394.1). **b** The structural comparison of the spike proteins from SARS CoV-2 (red, PDB ID: 7JJJ) and HCoV NL63 (PDB ID: 5SZS) [58]. **c** Rotated view of (**a**) (Color figure online)

the 501N lineage, was reported with additional changes in the spike protein sequence besides N501Y: H69 and V70 deletion, Y144 deletion, N501Y, A570D, P681H,* T716I, S982A, and D1118H [75]. These positions for this variant are highlighted in Fig. 32. This particular strain was called 501Y.V1 and is also referred to as 501.V1, B.1.1 [76], or 20B (Nexstrain nomenclature [77, 78]) [79]. One can hypothesize that the deletion of the amino acid residues H69

Table 3 Summary of sequence identities and similarities between SARS CoV-2 (GenBank ID: BCA87361.1) and other human coronaviruses [50]. AA overlap: amino acid overlap

Entry	Name	GenBank ID	Sequence identity (%)	Sequence similarity (%)	AA overlap
1	SARS CoV-1	AAP13441.1	76.0	91.5	1277
2	HCoV 229-E	QOP39313.1	31.3	61.6	777
3	MERS CoV	ASU91305.1	34.8	65.6	1049
4	HCoV OC43	AAA03055.1	30.2	57.9	1344
5	HCoV HKU1	ADN03339.1	29.2	59.0	1358
6	HCoV NL63	AGT51394.1	29.8	60.0	861

and V70 could potentially truncate the S1 domain, which could in turn cause the spike protein to remain in the open conformation more frequently. This open conformation is more likely to bind to the ACE2 protein on the surface of the host cell. Interestingly, the P681 position is the location of the furin cleavage site of the spike protein and it is not clear whether the change from proline to histidine in this strain may play a role in affecting this activity [59].

2.6.5 The 501Y.V2 Variant of the Spike Protein

Recently another strain of SARS CoV-2 termed "501Y.V2" containing changes in the amino acid residues of the spike protein have been reported to be widespread [80]. The label 501Y.V2 is interchangeable with 501.V2 as it appears in the literature. The changes of this variant are located in the N-terminal domain: L18F, D80A, D215G, R246I, receptor binding domain (RBD): K417N, E484K, and N501Y, and at position A701V. These specific positions are highlighted in Fig. 33. At the RBD interface with the ACE2 protein, three residues are highlighted (K417N, E484K, and N501Y). Interestingly, the change at position 501 from the asparagine to the tyrosine (N501Y) introduces a possible cation–pi interaction between the aromatic tyrosine-501 residue on the spike protein with the positively charged lysine-353 residue on the ACE2 protein (Fig. 34a). The K417 residue is 9.3 angstroms away from K26 in the ACE2 protein, suggesting that converting the K417 residue to an asparagine (K417N) will likely introduce a hydrogen bond between the asparagine's (N417) carbonyl oxygen on the spike protein and the lysine's (K26) proton of ACE2. Furthermore, the E484K change would potentially introduce a salt bridge between the K484 residue and E35 of the ACE2 protein. As shown in Fig. 31b, E484 is 8.8 angstroms away from E35 of the ACE2 protein. Therefore, the change to a positive charge at position 484 as the lysine residue (K484), should enhance the interaction between the spike protein and the ACE2 protein. Because of the stronger interaction between the RBD (receptor binding domain) of the spike protein and the ACE2, this new SARS CoV-2 strain (501Y.V2) potentially improves the virus's ability to enter the host cells. Although the only similarity in the spike protein sequence between the 501Y.V1 strain and the 501Y.V2 strain is the N501Y residue, it is likely that the other amino acid changes promote the enhanced infectivity of each strain. For instance, the E484K change in 501Y.V2 (which 501Y.V1 lacks) introduces a potential salt bridge (K484 of the spike protein with E35 of the ACE2 protein). On the other hand, 501Y.V1 has a deletion of H69, V70, and Y144, which potentially can truncate the NTD of the spike protein, forcing the spike protein to adopt the open conformation more frequently. The facts that (i) these emerging SARS CoV-2 strains (501Y.V1, 501Y.V2, and D614G) have changes in the spike protein sequence and (ii) the recent vaccines are developed against the spike protein of the original strain, gives reason to focus efforts on understanding how these new strains are different from the original strain. It is interesting to note that a study was performed where the sera of 20 patients with the COVID19 mRNA-based vaccine, BNT162b2, successfully neutralized a SARS CoV-2 N501Y spike mutant [81]. However, this particular SARS CoV-2 mutant only possessed the N501Y mutation and not the other changes present in either 501Y.V1 or 501Y.V2. Therefore, further studies focused on the entire set of amino acid variations belonging to these more transmissible strains are necessary to confidently assess the effects of the changes.

2.6.6 Selected Updates on SARS CoV-2 Research—23 New SARS CoV-2 Proteins

One notable report recently identified 23 previously unidentified viral open reading frames (ORFs) (Table 4), suggesting that there are many other unknown features of this virus that have yet to be explored [82]. The original approach to identify the genes of SARS CoV-2 was based on comparing the sequences of other known betacoronaviruses—especially with SARS CoV. These genes were identified by locating the sequential open reading frames (ORFs) that begin with the start codon sequence: AUG (also, please see Supporting Information, section II, for the AUG start codons that were highlighted in the SARS CoV-2 genome) [83]. However, it is likely that some genes were missed in the original characterization for two reasons: (i) the fact that some AUG start codons can be embedded within the originally identified ORFs (i.e. overlapping ORFs) and (ii)

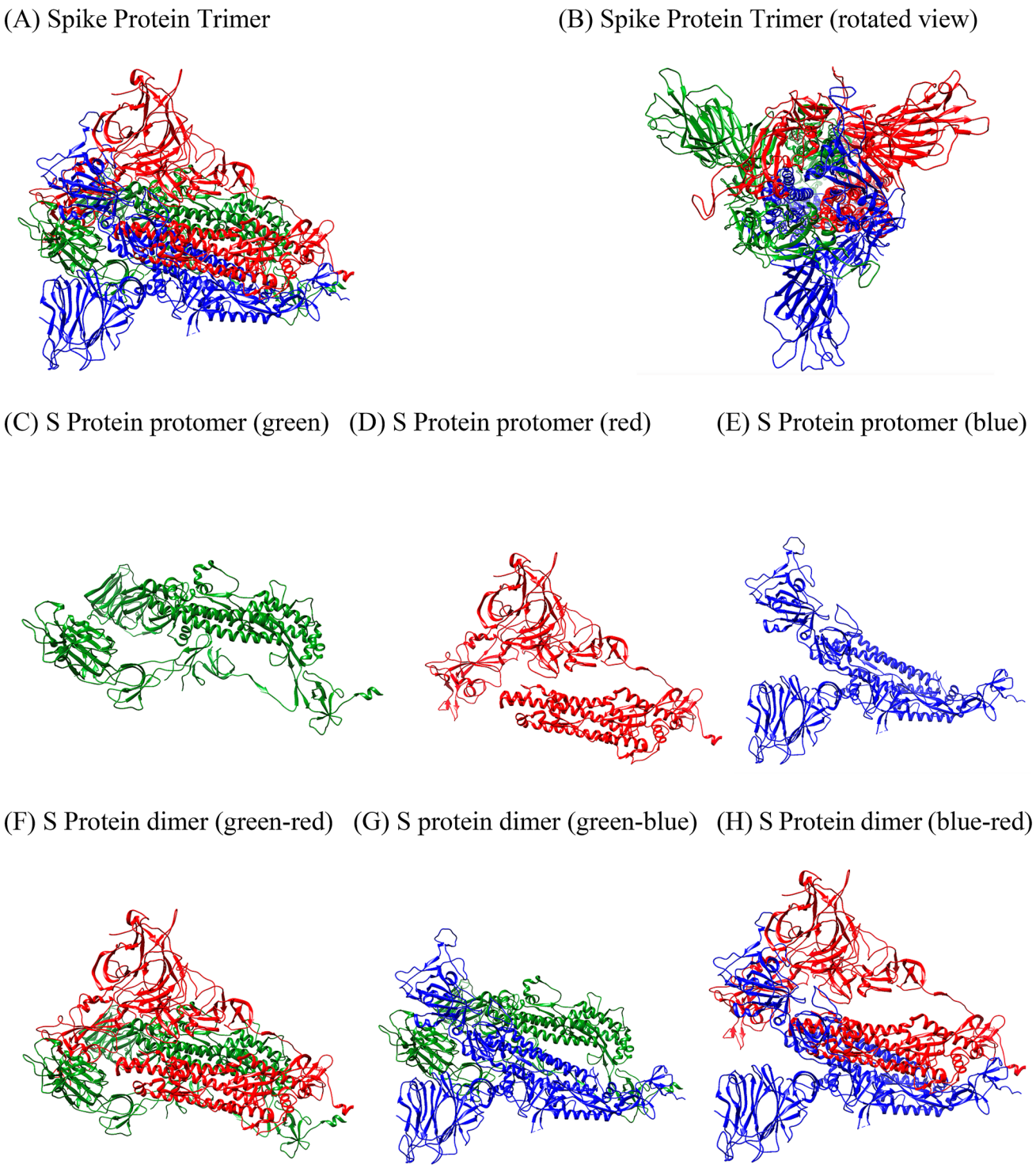


Fig. 23 Structure of the spike protein trimer from SARS CoV-2 (PDB ID: 7JJJ). Each protomer is a different color (i.e. red, green, or blue) (Color figure online)

some start codons have a different sequence besides the canonical AUG sequence [82] (for instance, CUG, ACG, AUU, AUC, UUG, and AUG—see Tables 4 and 5). Therefore, in order to identify the novel proteins, the researchers performed ribosome profiling experiments [84], where Vero E6 cells (African green monkey kidney cells) and Calu-3 cells (human lung cancer cells) were infected with SARS CoV-2. After a certain amount of time, the cells were treated

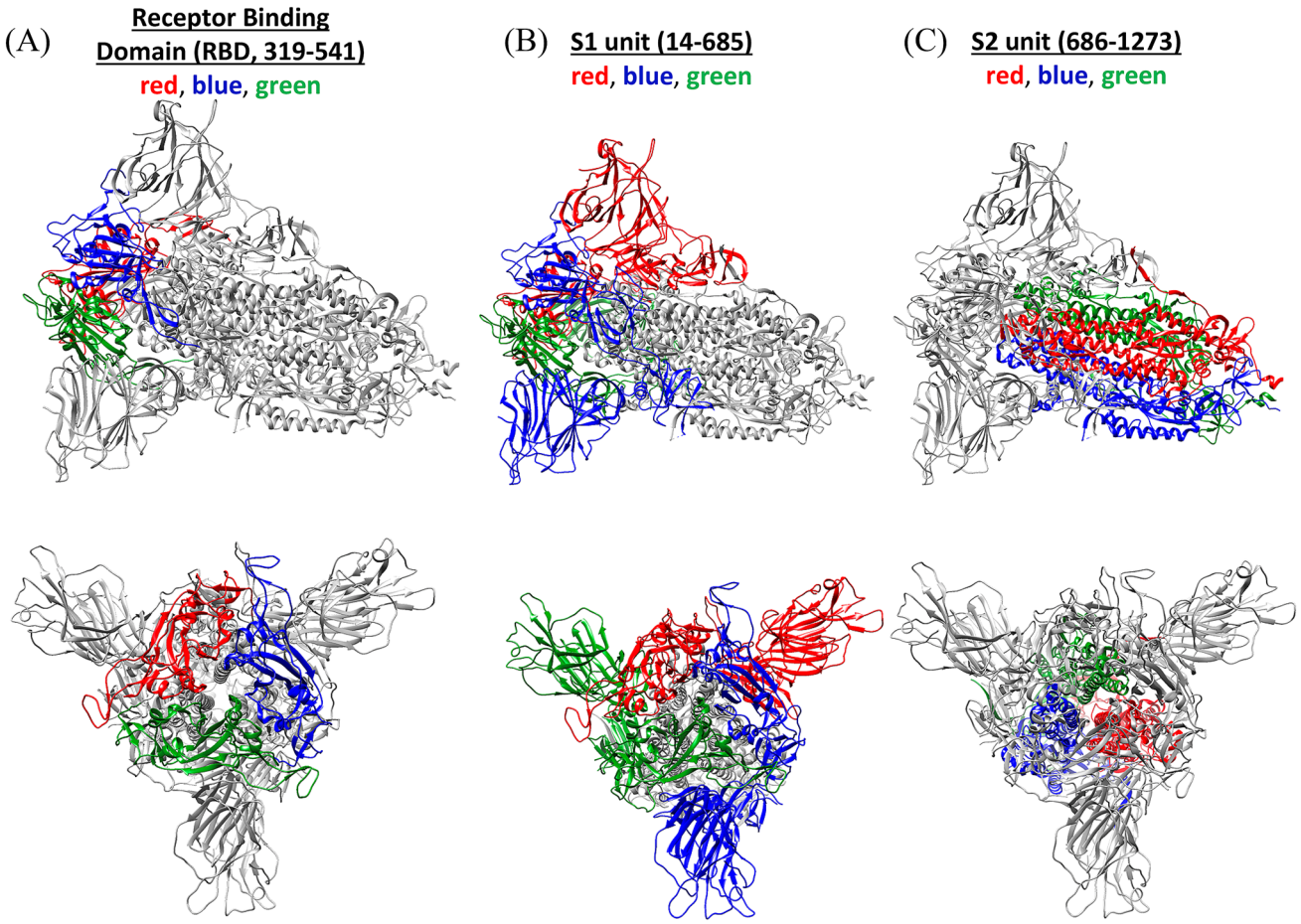


Fig. 24 Structure of the spike protein (PDB ID: 7JJJ, chain **a**: red, chain **b**: blue, chain **c**: green) and highlighted are the different receptor binding domains (RBDs, position 319–541) for each protomer (green, blue, and red). **a** Side view (top) and rotated view (bottom) of the RBD of the spike protein. **b** The S1 fragment (position 14–685) of the spike protein is highlighted for each protomer (red, blue, and green) where the protease, furin, cleaves (side view, top) (rotated view, bottom), and **c** S2 fragment (686–1273) of the spike protein (side view, top) (rotated view, bottom) (Color figure online)

with harringtonine or lactimidomycin, which halt ribosomes at initiating codons on the mRNA and provide translation initiation libraries. Alternatively, the cells were treated with cycloheximide, which would generate translation elongation libraries (instead of translation initiation libraries). In particular, lactimidomycin binds to the empty E-site of the large 80S ribosome allowing for isolation of the ribosome at strictly the start codon [85]. On the other hand, cycloheximide binding to the ribosome is reversible and when cycloheximide dissociates, ribosomes continue translation of the mRNA, which enables the trapping of ribosome downstream of the initiation codon [86]. The mRNA that was embedded in the ribosome was isolated and sequenced to reveal the new proteins. The combination of the translation initiation libraries and translation elongation libraries enabled the identification of the new protein sequences. Tables 4 and 5 show the 23 new proteins identified from this study. Among the newly identified proteins were in-frame internal ORFs (iORFs) located within known ORFs (e.g. S.iORF1 (Table 4, entry 6) is an internal ORF located downstream of

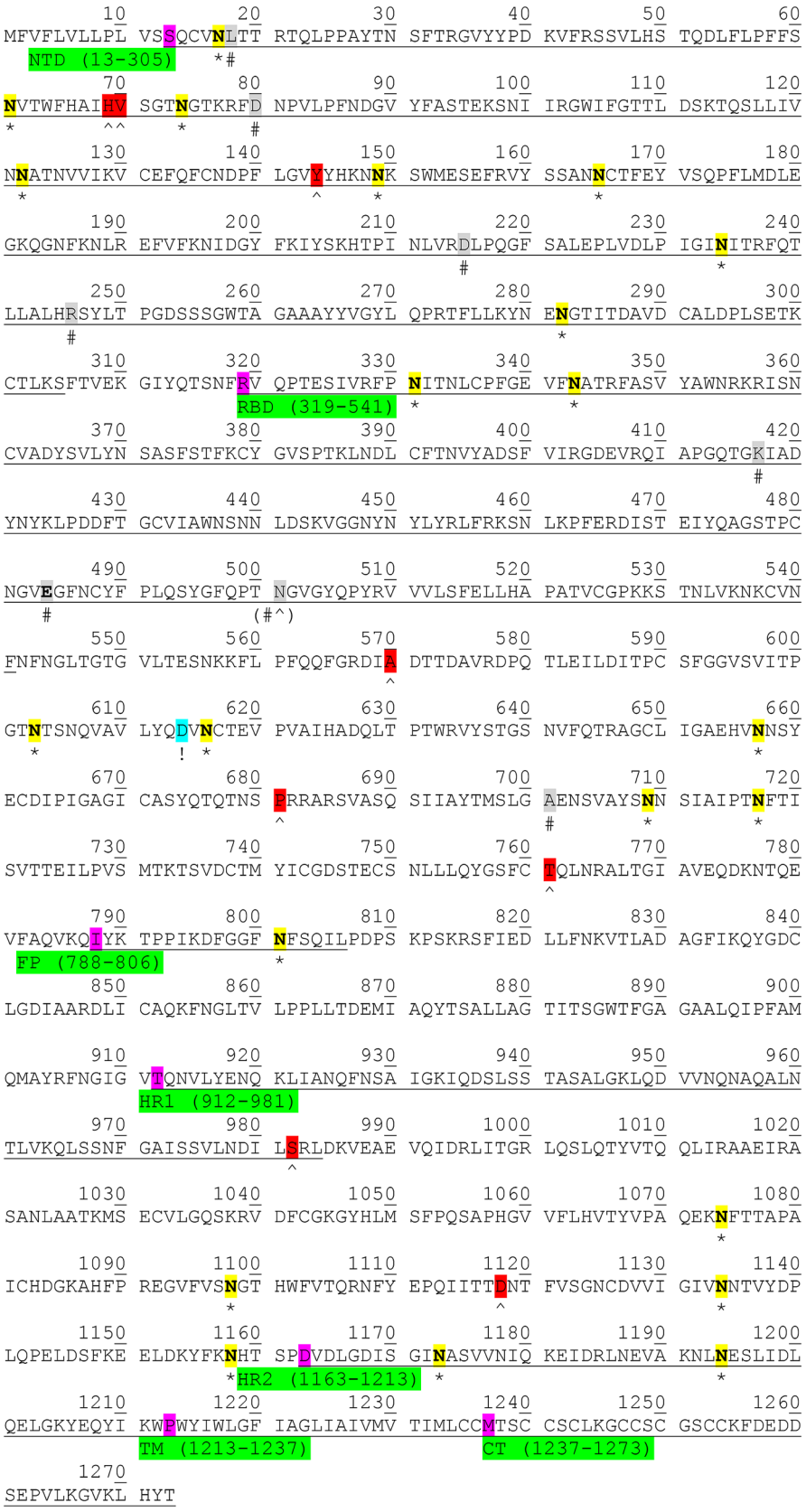


Fig. 25 The sequence of the spike protein of SARS CoV-2. The 22 asparagine (N) residues that undergo glycosylation are highlighted. S1 subunit: 14–685 [100], S2 subunit: 686–1273, S2′ cleavage [61] site: 816. *NTD* N-terminal domain, *RBD* receptor binding domain, *FP* fusion peptide, *HR1* heptapeptide repeat (or heptad repeat) sequence 1, *HR2* heptad repeat 2, *TM* transmembrane domain, *CT* cytoplasm domain [101]. "*" indicates glycosylation sites. "!" Marks the location of the D614G variant (cyan) [47, 74]. "#" Marks the locations of the 501Y.V2 variant in grey ([i] NTD region: L18F, D80A, D215G, R246I, [ii] RBD region: K417N, E484K, N501Y, and [iii] A701V) [80]. "^" Marks the locations of the 501Y.V1 variant in red (H69, V70, Y144, (N501Y), A570D, P681H, T761I, S982A, D1118H–N501Y is already marked in grey with "#" for the 501Y.V2 variant) (Color figure online)

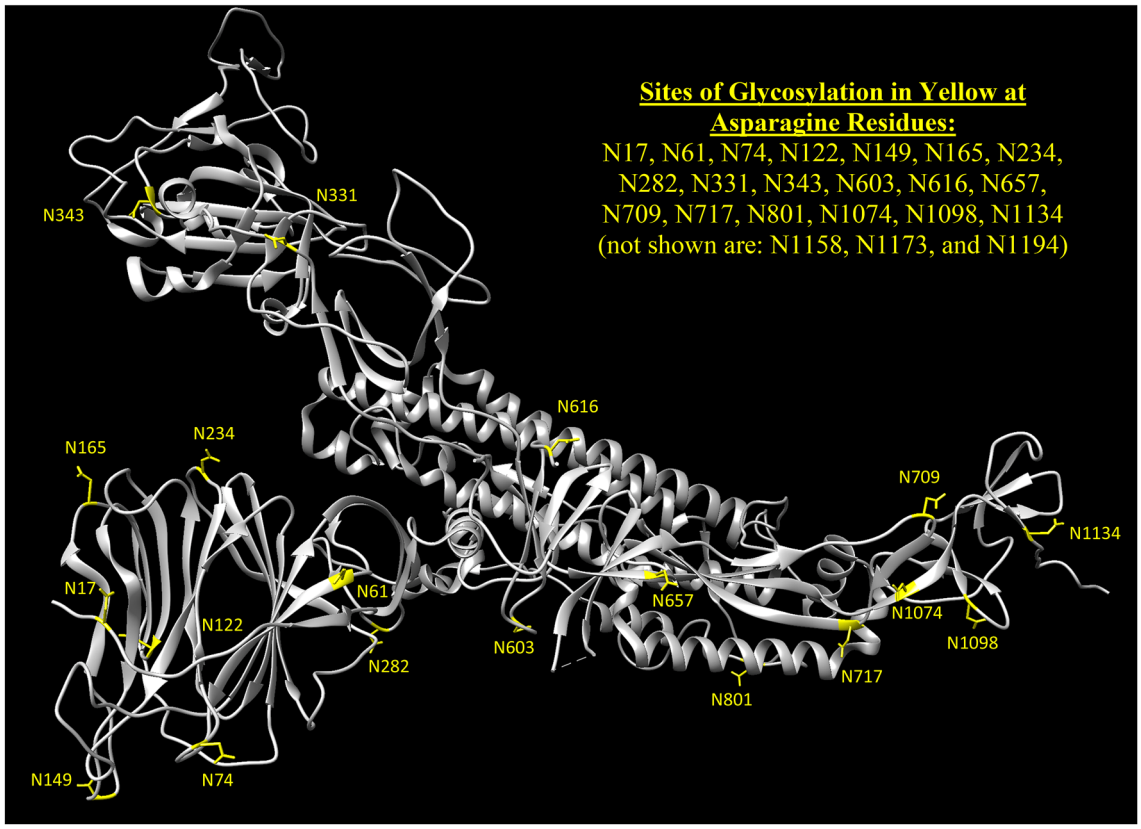


Fig. 26 Structure of the spike protein (protomer) with asparagine residues that undergo glycosylation are highlighted in yellow (PDB ID: 7JJJ) (Color figure online)

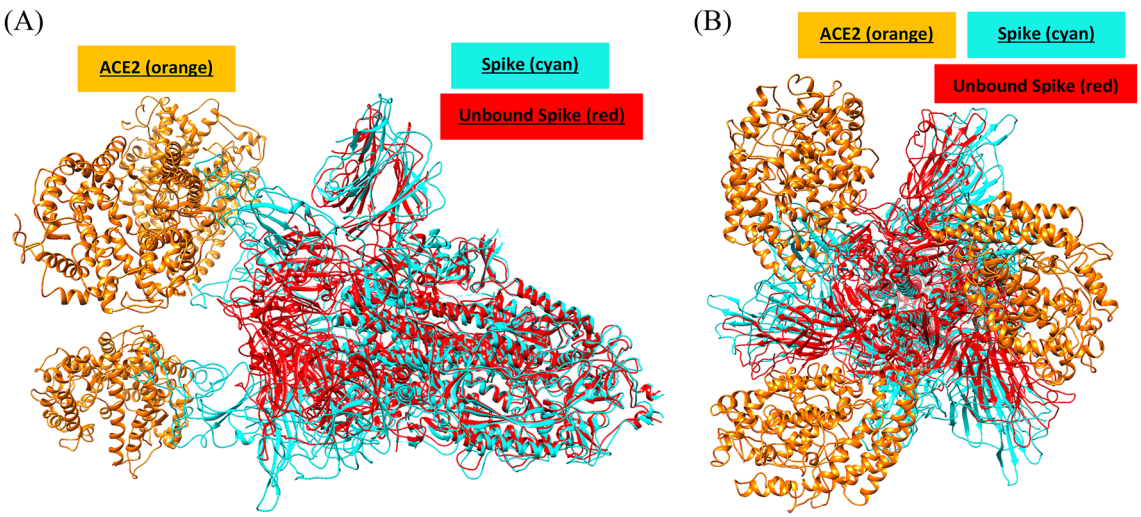


Fig. 27 Superimposed structures of (i) the spike protein with ACE2 bound (spike protein is cyan and ACE2 protein is orange, PDB ID: 7A98) and (ii) the unbound spike protein (red, PDB ID: 7JJJ). The open state of the spike protein can be seen when the ACE2 protein (orange) binds at the RBD of the spike protein (cyan). **a** Side view and **b** rotated view of (**a**) (Color figure online)

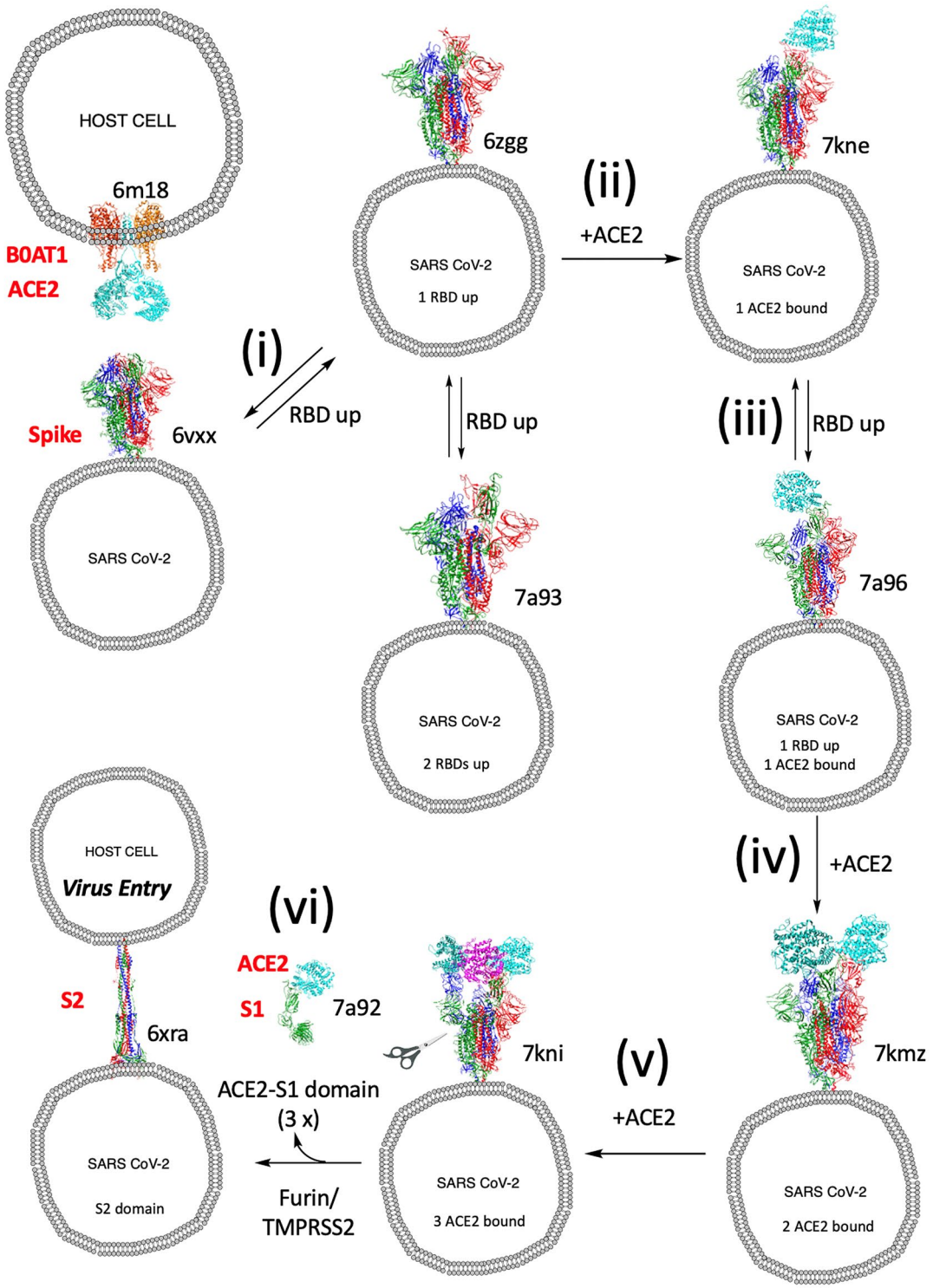


Fig. 28 Illustration of SARS CoV-2 entry into the host cell via ACE2-B0AT1 (B0AT1 is also called SLC6A19, solute carrier family 6 member 19, sodium dependent neutral amino acid transporter) [102] (PDB ID: 6M18) [102]. (i) One receptor binding domain (RBD) (or two RBDs – PDB ID: 7A93) [70] of the spike protein orients in the open conformation (PDB ID: 6ZGG) [103] from the closed conformation (PDB ID: 6VXX) [60], (ii) ACE2 protein binds to the RBD of the spike protein (PDB ID: 7KNE) [104], (iii) a second RBD "opens" up (PDB ID: 7A96) [70], (iv) a second ACE2 protein binds to the second RBD of the spike protein (PDB ID: 7KMZ) [104], (v) a third ACE2 protein binds to the final RBD (PDB ID: 7KNI) [104], (vi) furin and TMPRSS2 cleave at the S1-S2 site and S2′ site of the spike protein releasing the ACE2-S1 complex (PDB ID: 7A92) [70] and in turn, leaving behind the S2 domain (PDB ID: 6XRA) [61], which is primed for entry into the host cell. Shown in the S2 domain trimer (PDB ID: 6XRA) is the spike protein sequence from T912-N1173 and Q1180-L1197. Interestingly, the spike protein with two RBD units in the open conformation has also been observed through cryo-EM (PDB ID: 7A93) [70] (Color figure online)

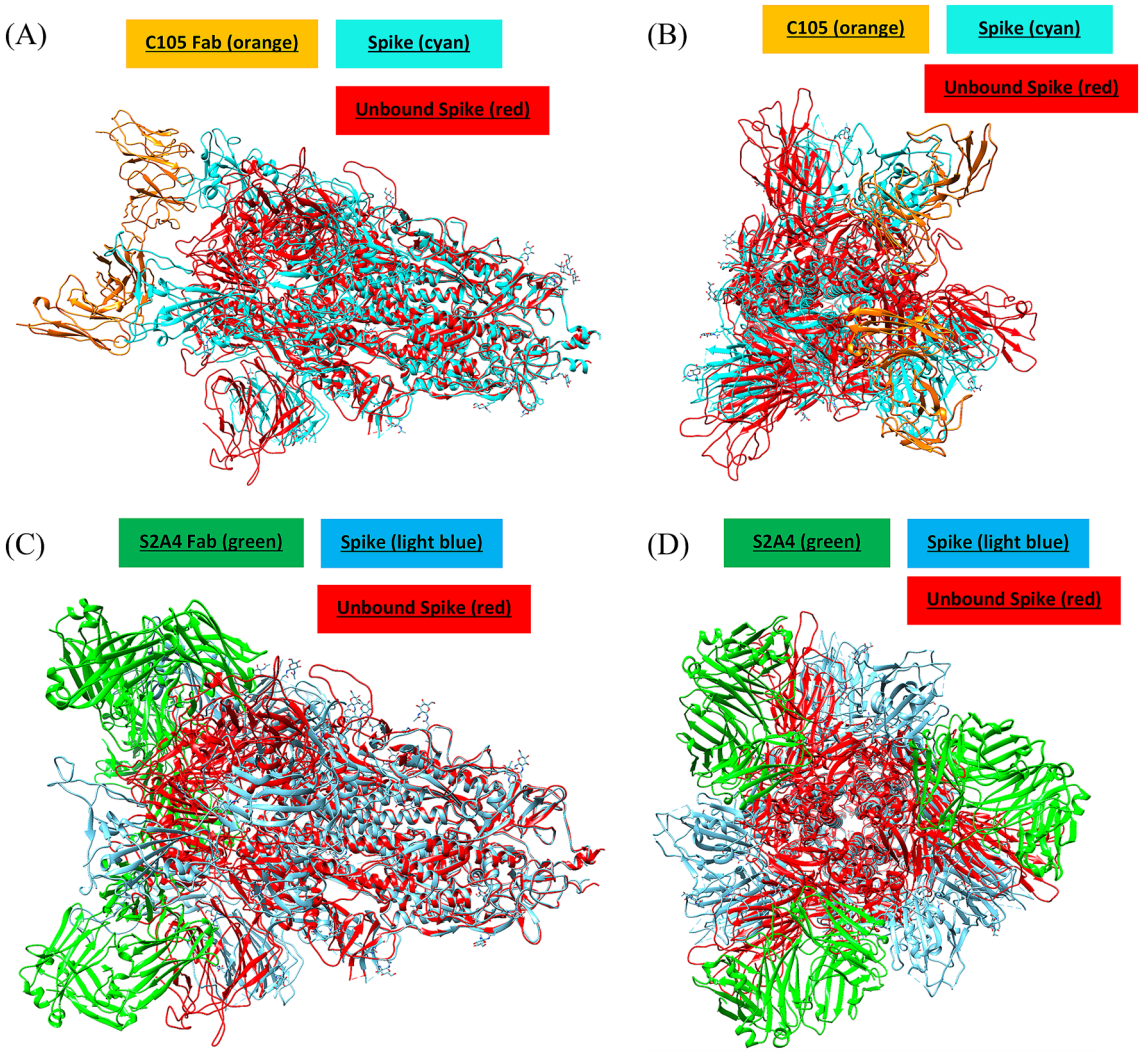


Fig. 29 **a** Spike protein (cyan) bound to C105 neutralizing antibody Fab fragment (orange) (PDB ID: 6XCM) superimposed with spike protein (red, PDB ID: 7JJJ). **b** Rotated view of (**a**). **c** Spike protein (light blue) bound to S2A4 neutralizing antibody Fab fragment (green) (PDB ID: 7JVC) superimposed with spike protein (red, PDB ID: 7JJJ) (Color figure online)

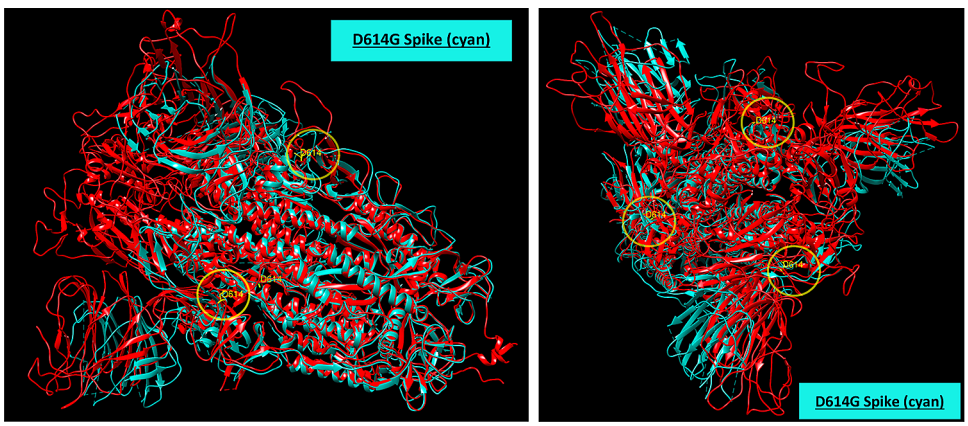


Fig. 30 Structural overlay of the spike protein D614 (red, PDB ID: 7JJJ) and G614 (cyan, PDB ID: 6XS6). Circled in yellow is the location of D614. The cryo-EM structure of the G614 variant has no RBD but is in a slightly more "open" conformation. **a** is the side view of the spike protein and **b** is the rotated view (Color figure online)

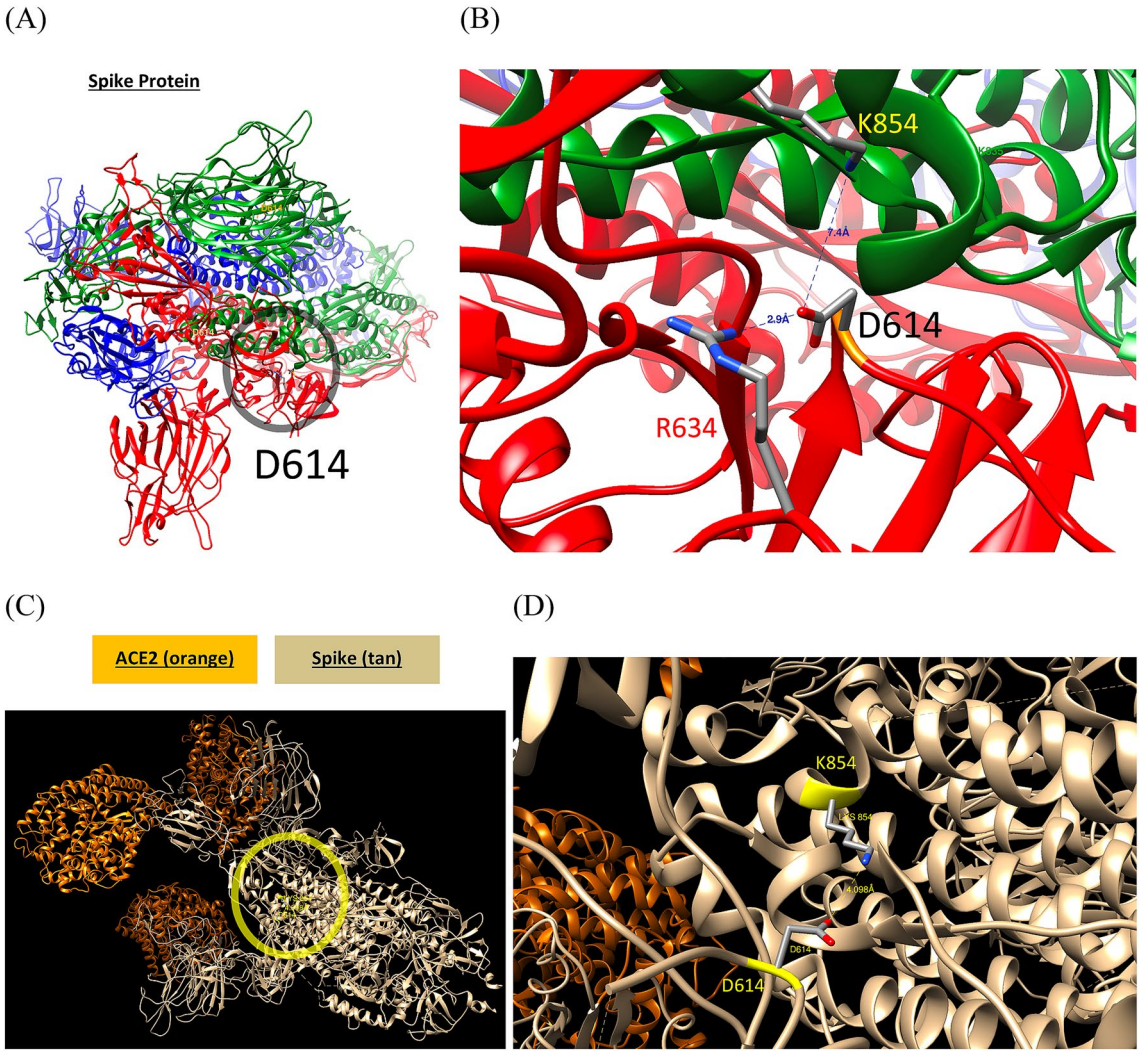


Fig. 31 **a** A salt bridge between D614 and R634 within the same protomer (red) of the spike protein is shown. Another amino acid K854 from a different protomer (green) interacts with D614 (7.4 angstroms away) (PDB ID: 7JJJ). **b** is the zoomed in region of the salt bridge interactions (D614-R634 and D614-K854). The red color is one protomer and the green color is the second protomer (K854 is on a separate protomer from the D614 residue on the red protomer—they are 7.4 angstroms apart). Each protomer in the trimer is a different color: red, green, blue. **c** In the ACE2 bound-spike protein, D614, the K854 is shown closer to D614 (4.1 angstroms) suggesting a stabilizing role for this salt bridge (PDB ID: 7A98). **d** Zoomed in image of the salt bridge between D614 and K854 in the spike-ACE2 complex (4.1 angstroms apart) (Color figure online)

the AUG start codon of S-ORF), upstream ORFs (uORFs), internal out-of-frame translations, and extended ORFs (such as M.ext, a 13 amino acid extension of ORF-M, cf. Table 4, entry 11). Furthermore, this study reported that virus translation dominates host translation due to increased levels of viral transcripts. Identification of the proteins may be important in the development of new vaccines and elucidating new biochemical properties of SARS CoV-2.

2.6.7 Vaccines

The most promising news has been the success of a SARS CoV-2 vaccine (BNT162b1) that has been shown to be 95% [87] effective [4]. This vaccine is an RNA based vaccine. BNT162b1 is a lipid nanoparticle formulated RNA vaccine that encodes a trimerized SARS CoV-2 receptor binding domain (RBD) [3]. A second mRNA vaccine (mRNA-1273), which encodes the prefusion spike protein of SARS CoV-2, is also showing effectiveness in producing antibody responses [5, 88]. And even another vaccine (ChAdOx1 nCoV-19) is also undergoing phase 2/3 clinical trials with promising results [89]. ChAdOx1 is a chimpanzee adenovirus vector, and the researchers have designed the vaccine to deliver the codon-optimized full-length spike protein of SARS CoV-2 [90]. From the clinical trial studies, ChAdOx1 nCoV-19 had an acceptable

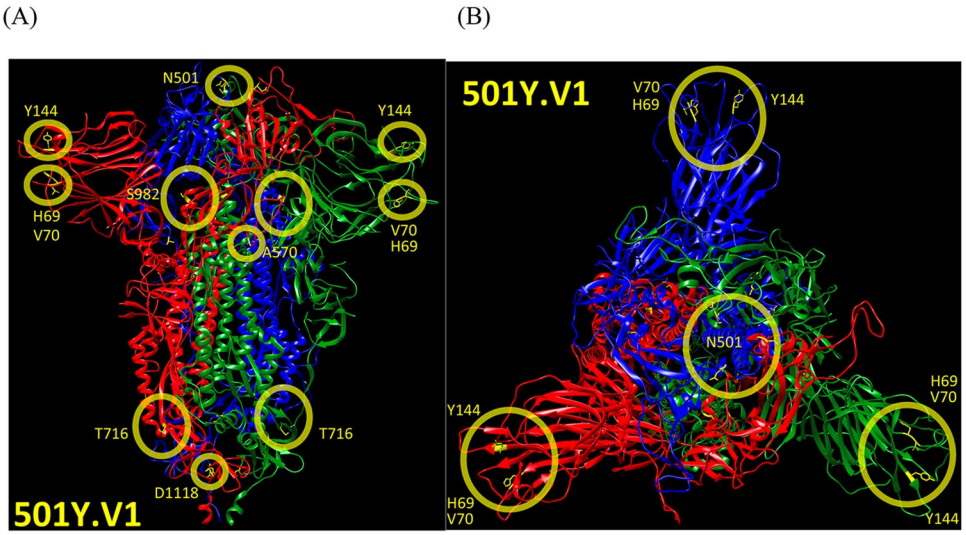


Fig. 32 **a** The spike protein (PDB ID: 7JJJ) with the protomers colored differently (red, green, and blue). The amino acid residues that are changed in the 501Y.V1 variant are highlighted in yellow. **b** Rotated view. Highlighted in yellow are: H69 and V70 (deletion), Y144 (deletion), N501(Y), A570(D), T716(I), S982(A), and D1118(H)—position P681(H) is not included in the structure (Color figure online)

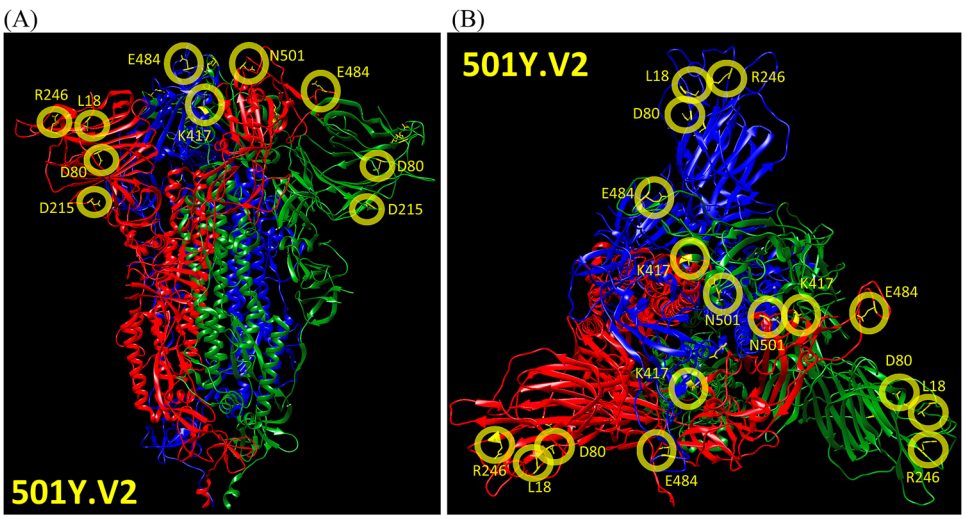
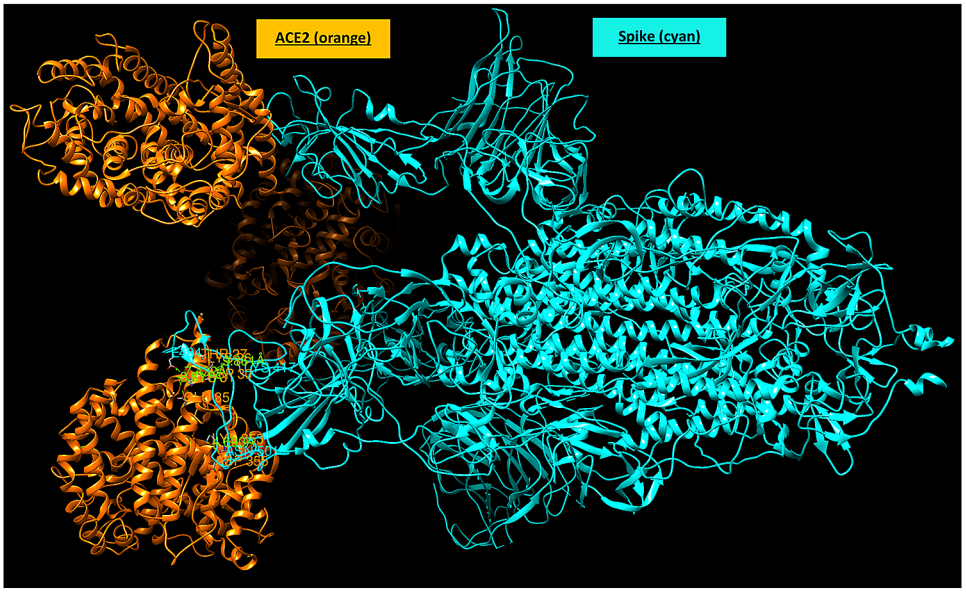


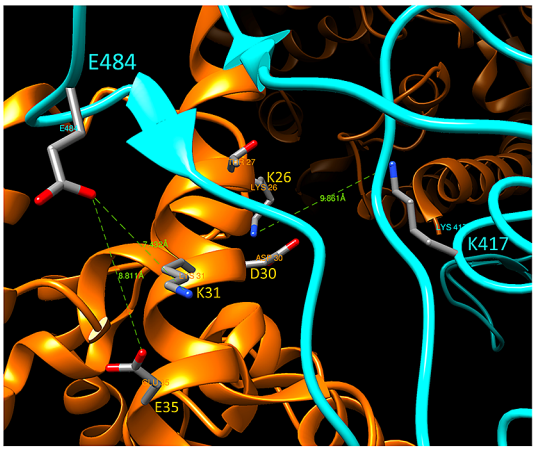
Fig. 33 **a** Structure of the spike protein with the amino acid residue changes of the 501Y.V2 variant highlighted: L18F, D80A, D215G, R246I, K417N, E484K, N501Y, and A701V. **b** Rotated view (Color figure online)

safety profile and was effective against symptomatic COVID-19 [91]. The new vaccines are made available [92] at the time the emergence of other variants of SARS CoV-2 are appearing. These recent strains (D614G and 501Y.V2) possibly have enhanced infectivity and stability of virions compared to the originally identified strain [93, 94]. Although these new mutations of SARS CoV-2 strains have been suggested to not have increased transmissibility [95], this research area remains a high priority worldwide. Investigating the biochemistry of the proteins found in SARS CoV-2 enables a better understanding to prevent another global pandemic.

(A)



(B) E484 and K417



(C) N501

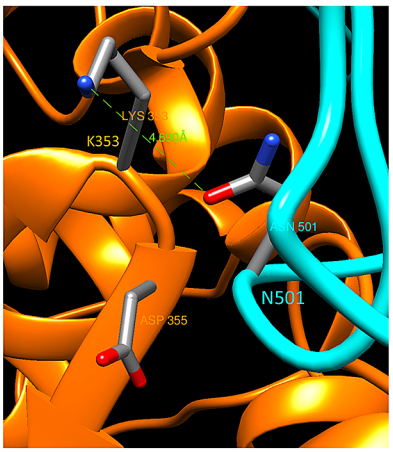


Fig. 34 **a** Spike protein-ACE2 protein complex (PDB ID: 7A98). Focus on the residues of the recently reported variant: K417N, E484K, and N501Y are shown in **b** and **c**. **b** E484 (spike) with K31 (ACE2) (the distance is also measured to E35 of the ACE2). **c** N501 (spike) with K353 (ACE2) is shown

Table 4 The 23 new proteins identified in SARS CoV-2 from the ribosome profiling study [82] of cells infected with SARS CoV-2 (continued to Table 5)

Entry (#)	Name	Location	Amino acid sequence
1(25)	1a.uORF1.ext [a]	59–136	10 20 LFSKRTLKSV WLSLGCMLSA LTQYN
2(9)	1a.uORF1[b]	107–136	MLSALTQYN
3(33)	1a.uORF2.ext [c]	163–264	10 20 30 TSNSSIFCRL LTVSSVLQPI ISTSRFCPGV TER
4(13)	1a.uORF2[d]	223–264	10 ISTSRFCPGV TER
5(5)	1a.iORF[b]	408–425	MALVA
6(39)	S.iORF1[b]	21,744–21,863	10 20 30 MLLGSMLYMS LGPMVLRGLI TLSYHLMMVF ILLPLRSLT
7(31)	S.iORF2[b]	21,768–21,863	10 20 30 MSLGPMVLRG LITLSYHLMM VFILLPLRSL T
8(41)	3a.iORF1 (ORF3c)[b]	25,457–25,582	10 20 30 40 MLLLQILFAL LQRYRYKPHS LSDGLLLALH FLLFFRALPK S
9(33)	3a.iORF2[b]	25,596–25,697	10 20 30 MATSTLQGCS LCLQLAVVVC NSLLTPFARC CWP
10(11)	E.iORF[a]	26,437–26,472	10 LNSSRVPDLL V
11(235)	M.ext[e]	26,484–27,191	10 20 30 40 50 60 IILVFLFGTL ILAMADSNGT ITVEELKKLL EQWNLVIGFL FLTWICLLQF AYANRNRFLY 70 80 90 100 110 120 IIKLIFLWLL WPVTLACFVL AAVYRINWIT GGIAIAMACL VGLMWLSYFI ASFRLFARTR 130 140 150 160 170 180 SMWSFNPETN ILLNVPLHGT ILTRPLLESE LVIGAVILRG HLRIAGHHLG RCDIKDLPKE 190 200 210 220 230 ITVATSRTLS YYKLGASQRV AGDSGFAAYS RYRIGNYKLN TDHSSSSDNI ALLVQ
12(14)	M.iORF[e]	27,151–27,195	10 IPVAVTILLC LYSK
13(43)	6.iORF[b]	27,256–27,387	10 20 30 40 MRTFKVSIWN LDYIINLIIK NLSKSLTENK YSQLDEEQPM EID
14(119)	7a.iORF1[e]	27,400–27,759	10 20 30 40 50 60 IILFLALITL ATCELYHYQE CVRGTTVLLK EPCSSGTYEG NSPFHPLADN KFALTCFSTQ 70 80 90 100 110 FAFACPDGVK HVYQLRARSV SPKLFIRQEE VQELYSPIFL IVAAIVFITL CFTLKRKTE

(#): number of amino acids. Superscript indicates the start codon for each new protein—a: CUG, b: AUG, C: ACG, d: AUC, e: AUU, f: UUG, g: AUA. "uORF": upstream ORF, "iORF": internal ORF, "ORF.ext": extended ORF

Table 5 The 23 new proteins identified in SARS CoV-2 from the ribosome profiling study [82] of cells infected with SARS CoV-2 (continued from Table 4)

Entry(#)	Name	Location	Amino acid sequence
15(8)	7a.iORF2[f]	27,581–27,607	LLLLVLTA
16(15)	7a.iORF3[g]	27,712–27,759	10 IVFITLCFTL KRKTE
17(20)	7b.iORF1[b]	27,825–27,887	10 20 MLIIFWFSLE LQDHNETCHA
18(11)	7b.iORF2[d]	27,862–27,897	10 IIMKLVTPKR T
19(9)	8.iORF[b]	27,965–27,994	MYSTSTICS
20(97)	N.iORF1 (ORF9b)[b]	28,284–28,577	10 20 30 40 50 60 MDPKISEMHP ALRLVDPQIQ LAVTRMENAV GRDQNNVGPK VYPIILRLGS PLSLNMARKT 70 80 90 LNSLEDKAFQ LTPIAVQMTK LATTEELPDE FVVVTVK
21(90)	N.iORF2[b]	28,305–28,577	10 20 30 40 50 60 MHPALRLVDP QIQLAVTRME NAVGRDQNNV GPKVYPIILR LGSPLSLNMA RKTLNSLEDK 70 80 90 AFQLTPIAVQ MTKLATTEEL PDEFVVVTVK
22(10)	10.uORF[b]	28,538–29,570	10 MQTTQGRWAI
23(18)	10.iORF[b]	29,618–29,674	10 MNSRNYIAQV DVVNFNLT

(#): number of amino acids. Superscript indicates the start codon for each new protein – a: CUG, b: AUG, C: ACG, d: AUC, e: AUU, f: UUG, g: AUA

3 Conclusion

A comprehensive review of the current literature of SARS CoV-2 is impossible due to the explosion in publications on this urgent topic. Despite the increase in publications, more experimental work is desperately needed to gain better insights into how SARS CoV-2 caused this global pandemic and to develop new strategies to combat this virus [96]. The catalyst for deeper knowledge towards the cause of COVID-19 is the emergence of the recent variant strains that may be more infectious [2] than the original SARS CoV-2 strain. This pandemic has reminded humanity how fragile our existence is and how important science and research is and will continue to play a vital role in our lives.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10930-021-09967-8.

Acknowledgements The author is grateful for the insightful comments and detailed advice provided by the expert reviewers of this manuscript to significantly enhance the quality of this paper.

Supporting Information Supporting information is available showing the alignment of the SARS CoV-2 spike proteins with the spike proteins from other coronaviruses using LALIGN software [50], RNA genome of SARS CoV-2, and the protein sequences of the 23 newly identified proteins.

Compliance with Ethical Standards

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Yoshimoto FK (2020) The proteins of severe acute respiratory syndrome coronavirus-2 (SARS CoV-2 or n-COV19), the cause of COVID-19. Protein J 39:198–216
2. Korber B et al (2020) Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 182:812–827
3. Walsh EE et al (2020) Safety and immunogenicity of two RNA-based Covid-19 vaccine candidates. NEJM 383:2439
4. Mulligan MJ et al (2020) Phase I/II study of COVID-19 RNA vaccine BNT162b1 in adults. Nature 585:589–593

5. Jackson LA et al (2020) An mRNA vaccine against SARS-CoV-2 - perliminary report. NEJM 383:1920–1931
6. Pettersen EF et al (2004) UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612
7. Lei J, Kusov Y, Hilgenfeld R (2017) Nsp3 of coronaviruses: structures and functions of a large multi-domain protein. Antiviral Res 149:58–74
8. Barretto N et al (2005) The papain-like protease of severe acute respiratory syndrome coronavirus has deubiquitinating activity. J Virol 79:15189–15198
9. Shin D et al (2020) Papain-like protease regulates SARS-CoV-2 viral spread and innate immunity. Nature 587:657
10. Gao X et al (2020) Crystal structure of SARS-CoV-2 papain-like protease. Acta Pharmac Sin B 11:237
11. Rut W et al (2020) Activity profiling and crystal structures of inhibitor-bound SARS-CoV-2 papain-like protease: a framework for anti-COVID-19 drug design. Sci Adv 6:4596
12. Michalska K et al (2020) Crystal structures of SARS-CoV-2 ADP-ribose phosphatase: from the apo form to ligand complexes. IUCrJ 7:814–824
13. Muramatsu T et al (2013) Autoprocessing mechanism of severe acute respiratory syndrome coronavirus 3C-like protease (SARS-CoV 3CL pro) from its polyproteins. FEBS J 280:2002–2013
14. Kneller DW et al (2020) Structural plasticity of SARS-CoV-2 3CL Mpro active site cavity revealed by room temperature X-ray crystallography. Nat Commun 11:3202
15. Jin Z et al (2020) Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. Nature 582:289–293
16. Zhang L et al (2020) Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved alpha-ketoamide inhibitors. Science 368:409–412
17. Kneller DW et al (2020) Malleability of the SARS-CoV-2 3CL Mpro active-site cavity facilitates binding of clinical antivirals. Structure 28:1–8
18. Ma C et al (2020) Boceprevir, GC-376, and calpain inhibitors II, XII inhibit SARS-CoV-2 viral replication by targeting the viral main protease. Cell Res 30:678–692
19. Hillen HS et al (2020) Structure of replicating SARS-CoV-2 polymerase. Nature 584:154–156
20. Gordon CJ et al (2020) Remdesivir is a direct-acting antiviral that inhibits RNA-dependent RNA polymerase from severe acute respiratory syndrome coronavirus 2 with high potency. J Biol Chem 295:6785–6797
21. Tchesnokov EP, Feng JY, Porter DP, Gotte M (2019) Mechanism of inhibition of ebola virus RNA-dependent RNA polymerase by remdesivir. Viruses 11:326
22. Wang M et al (2020) Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. Cell Res 30:269–271
23. Warren TK et al (2016) Therapeutic efficacy of the small molecule GS-5734 against Ebola virus in rhesus monkeys. Nature 531:381–385
24. Yin W et al (2020) Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. Science 368:1499–1504
25. Doi Y et al (2020) A prospective, randomized, open-label trial of early versus late favipiravir therapy in hospitalized patients with COVID-19. Antimicrob Agents Chemother 64:e01897
26. Shannon A et al (2020) Rapid incorporation of Favipiravir by the fast and permissive viral RNA polymerase complex results in SARS-CoV-2 lethal mutagenesis. Nat Commun 11:4682
27. Murakami E et al (2014) Metabolism and pharmacokinetics of the anti-hepatitis C virus nucleotide prodrug GS-6620. Antimicrob Agents Chemother 58:1943–1951
28. Yan VC, Muller FL (2020) Advantages of the parent nucleoside GS-441524 over remdesivir for Covid-19 treatment. ACS Med Chem Lett 11:1361–1366
29. Beigel JH et al (2020) Remdesivir for the treatment of Covid-19: final report. NEJM 383:1813–1826
30. Ivanov KA et al (2004) Multiple enzymatic activities associated with severe acute respiratory syndrome coronavirus helicase. J Virol 78:5619–5632
31. Shu T et al (2020) SARS-coronavirus-2 Nsp13 possesses NTPase and RNA helicase activities that can be inhibited by bismuth salts. Virol Sin 35:321–329
32. Chen Y, Guo D (2016) Molecular mechanisms of coronavirus RNA capping and methylation. Virol Sin 31:3–11
33. Bouvet M et al (2010) *In vitro* reconstitution of SARS-coronavirus mRNA cap methylation. PLOS Pathog 6:10
34. Gross CH, Shuman S (1998) RNA 5'-triphosphatase, nucleoside triphosphatase, and guanylyltransferase activities of baculovirus LEF-4 protein. J Virol 72:10020–10028
35. Jia Z et al (2019) Delicate structural coordination of the severe acute respiratory syndrome coronavirus Nsp13 upon ATP hydrolysis. Nucleic Acids Res 47:6548–6550
36. Chen J et al (2020) Structural basis for helicase-polymerase coupling in the SARS-CoV-2 replication-transcription complex. Cell 182:1560–1573
37. Yan L et al (2020) Architecture of a SARS-CoV-2 mini replication and transcription complex. Nat Commun 11:5874
38. Ma YY, Wu LJ, Zhang RG, Rao ZH (2015) Crystal structure of the SARS coronavirus nsp14-nsp10 complex. PNAS 112:9436–9441
39. Ogando NS et al (2020) The enzymatic activity of the nsp14 exoribonuclease is critical for replication of MERS-CoV and SARS-CoV-2. J Virol 94:e01246-e11220
40. Hackbart M, Deng X, Baker SC (2020) Coronavirus endoribonuclease targets viral polyuridine sequences to evade activating host sensors. PNAS 117:8094–8103
41. Colgan DF, Manley JL (1997) Mechanism and regulation of mRNA polyadenylation. Genes Dev 11:2755–2766
42. Nedialkova DD et al (2009) Biochemical characterization of arterivirus nonstructural protein 11 reveals the nidovirus-wide conservation of a replicative endoribonuclease. J Virol 83:5671–5682
43. Kim Y et al (2020) Crystal structure of Nsp15 endoribonuclease NendoU from SARS-CoV-2. Protein Sci 29:1596–1605
44. Decroly E et al (2008) Coronavirus nonstructural protein 16 Is a Cap-0 binding enzyme possessing (nucleoside-2′O)-methyltransferase activity. J Virol 82:8071–8084
45. Krafcikova P, Silhan J, Nencka R, Boura E (2020) Structural analysis of the SARS-CoV-2 methyltransferase complex involved in RNA cap creation bound to sinefungin. Nat Commun 11:3717
46. Volz E et al (2021) Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. Cell 184:1–12
47. Zhang L et al (2020) SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. Nat Commun 11:6013
48. Hulswit RJG et al (2019) Human coronaviruses OC43 and HKU1 bind to 9-O-acetylated sialic acids via a conserved receptor-binding site in spike protein domain A. PNAS 116:2681–2690
49. Ou X et al (2017) Crystal structure of the receptor binding domain of the spike glycoprotein of human betacoronavirus HKU1. Nat Commun 8:15216
50. Madeira F et al (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 47:W636–W641
51. Hofmann H et al (2005) Human coronavirus NL63 employs the severe acute respiratory syndrome coronavirus receptor for cellular entry. PNAS 102:7988–7993

52. Li Z et al (2019) The human coronavirus HCoV-229E S-protein structure and receptor binding. Elife 8:172
53. Li Y et al (2020) The MERS-CoV receptor DPP4 as a candidate binding target of the SARS-CoV-2 Spike. iScience 23:101160
54. Bangaru S et al (2020) Structural analysis of full-length SARS-CoV-2 spike protein from an advanced vaccine candidate. Science 370:1089–1094
55. Gui M et al (2017) Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. Cell Res 27:119–129
56. Yuan Y et al (2017) Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. Nat Commun 8:15092
57. Kirchdoerfer RN et al (2016) Pre-fusion structure of a human coronavirus spike protein. Nature 531:118–121
58. Walls AC et al (2016) Glycan shield and epitope masking of a coronavirus spike protein observed by cryo-electron microscopy. Nat Struct Mol Biol 23:899–905
59. Ord M, Faustova I, Loog M (2020) The sequence at Spike S1/S2 site enables cleavage by furin and phospho-regulation in SARS-CoV2 but not in SARS-CoV1 or MERS-CoV. Sci Rep 10:16944
60. Walls AC et al (2020) Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 181:281–292
61. Cai Y et al (2020) Distinct conformational states of SARS-CoV-2 spike protein. Science 369:1586–1592
62. Xu C et al (2021) Conformational dynamics of SARS-CoV-2 trimeric spike glycoprotein in complex with receptor ACE2 revealed by cryo-EM. Sci Adv 7:eabe5575
63. Bestle D et al (2020) TMPRSS2 and furin are both essential for proteolytic activation of SARS CoV-2 in human airway cells. Life Sci Alliance 3:1–14
64. Shang J et al (2020) Cell entry mechanisms of SARS-CoV-2. PNAS 117:11727–11734
65. Watanabe Y, Allen JD, Wrapp D, McLellan JS, Crispin M (2020) Site-specific glycan analysis of the SARS-CoV-2 spike. Science 369:330–333
66. Gasteigher E et al (2005) Protein identification and analysis tools on the ExPASy server. In: Walker JM (ed) The proteomics protocols handbook. Humana Press, Totowa
67. Brun J et al (2020) Analysis of SARS-CoV-2 spike glycosylation reveals shedding of a vaccine candidate. bioRxiv
68. Shajahan A, Supekar NT, Gleinich AS, Azadi P (2020) Deducing the N- and O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. Glycobiology 30:981–988
69. Monteil V et al (2020) Inhibition of SARS-CoV-2 infections in engineered human tissues using clinical-grade soluble human ACE2. Cell 181:905–913
70. Benton DJ et al (2020) Receptor binding and priming of the spike protein of SARS-CoV-2 for membrane fusion. Nature 588:327–330
71. Hoffmann M, Kleine-Weber H, Pohlmann S (2020) A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. Mol Cell 78:779–784
72. Hoffmann M et al (2020) SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 181:271–280
73. Barnes CO et al (2020) Structures of human antibodies bound to SARS-CoV-2 spike reveal common epitopes and recurrent features of antibodies. Cell 182:828–842
74. Yurkovestkiy L et al (2020) Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. Cell 183:739–751
75. Leung K, Shum MHH, Leung GM, Lam TTY, Wu JT (2021) Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom, October to November 2020. Eurosurveillance 26:2002106
76. Rambaut A et al (2020) A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 5:1403–1407
77. Lo SW, Jamrozy D (2020) Genomics and epidemiological surveillance. Nat Rev Microbiol 18:478
78. Turakhia Y et al (2020) Stability of SARS-CoV-2 phylogenies. PLOS Genet 181:1009175
79. Alm E et al (2020) Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. Eurosurveillance 25:2001410
80. Tegally H et al (2020) Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with muiltiple spike mutations in South Africa. medRxiv
81. Xie X et al (2020) Neutralization of N501Y mutant SARS-CoV-2 by BNT162b2 vaccine-elicited sera. bioRxiv
82. Finkel Y et al (2020) The coding capacity of SARS CoV-2. Nature 589:125
83. Kim D et al (2020) The architecture of SARS-CoV-2 transcriptome. Cell 181:1–8
84. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat Protoc 7:1534–1550
85. Lee S et al (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. PNAS 109:E2424–E2432
86. Mohammad F, Green R, Buskirk AR (2019) A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. eLife 8:e42591
87. Wadman M (2020) Fever, aches from Pfizer, Moderna jabs aren't dangerous but may be intense for some. Science 371:6529
88. Anderson EJ et al (2020) Safety and Immunogenicity of SARS-CoV-2 mRNA-1273 vaccine in older adults. NEJM 383:2427
89. Ramasamy MN et al (2020) Safety and immunogenicity of ChAdOx1 nCoV-19 vaccine administered in a prime-boost regimen in young and old adults (COV002): a single-blind, randomised, controlled, phase 2/3/ trial. Lancet 396:1979
90. van Doremalen N et al (2020) ChAdOx1nCoV-19 vaccine prevents SARS-CoV_2 pneumonia in rhesus macaques. Nature 586:578–582
91. Voysey M et al (2020) Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. Lancet 589:125
92. Krammer F (2020) SARS-CoV-2 vaccines in development. Nature 586:516–527
93. Plante JA et al (2020) Spike mutation D614G alters SARS-CoV-2 fitness. Nature 579:270
94. Hou YJ et al (2020) SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. Science 370:1464–1468
95. van Drop L et al (2020) No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. Nat Commun 11:5986
96. Dong Y et al (2020) A systematic review of SARS-CoV-2 vaccine candidates. Signal Transduct Target Therapy 5:237
97. Muramatsu T et al (2016) SARS-CoV 3CL protease cleaves its C-terminal autoprocessing site by novel subsite cooperativity. PNAS 113:12997–13002
98. Gao Y et al (2020) Structure of the RNA-dependent RNA polymerase from COVID-19 virus. Science 368:779–782
99. Kim Y, Maltseva N, Jedrzejczak R, Endres M, Welk L, Chang C, Michalska K, Joachimiak A (2020) CSGID. Crystal structure of NSP15 endoribonuclease from SARS CoV-2 in the complex

with uridine-3',5'-diphosphate. 7K10. https://doi.org/10.2210/pdb7K1O/pdb
100. Xia S et al (2020) Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1 domain in spike protein. Cell Mol Immunol 17:765–767
101. Huang Y, Yang C, Xu X-F, Xu W, Liu S-W (2020) Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol Sin B 41:1141–1149
102. Yan R et al (2020) Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science 367:1444–1448
103. Wrobel AG et al (2020) SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. Nat Struct Mol Biol 27:763–767
104. Zhou T et al (2020) Cryo-EM structures of SARS-CoV-2 spike without and with ACE2 reveal a pH-dependent switch to mediate endosomal positioning of receptor-binding domains. Cell Hoste Microbe 28:867
105. Baez-Santos YM, St. John SE, Mesecar AD (2015) The SARS-coronavirus papain-like protease: Structure, function and inhibition by designed antiviral compounds. Antiviral Res 115:21–38
106. Wang H et al (2020) Comprehensive insights into the catalytic mechanism of middle east respiratory syndrome 3C-like protease and severe acute respiratory syndrome 3C-like protease. ACS Catal 10:5871–5890