

# Neutral evolution test of the spike protein of SARS-CoV-2 and its implications in the binding to ACE2

**Georgina I. López-Cortés**

National Autonomous University of Mexico

**Miryam Palacios-Pérez**

National Autonomous University of Mexico

**Gabriel S. Zamudio**

National Autonomous University of Mexico

**Hannya F. Veledíaz**

National Autonomous University of Mexico

**Enrique Ortega**

National Autonomous University of Mexico

**Marco V. José** (✉ [marcojose@biomedicas.unam.mx](mailto:marcojose@biomedicas.unam.mx))

National Autonomous University of Mexico

---

## Research Article

**Keywords:** Spike evolution, Neutrality test, Phylogenetic analyses, Binding of Spike-ACE2, SARS-CoV-2 variants, Selective pressure

**Posted Date:** April 28th, 2021

**DOI:** <https://doi.org/10.21203/rs.3.rs-453111/v1>

**License:**  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

1 **Neutral evolution test of the spike protein of SARS-CoV-2 and its implications in the**  
2 **binding to ACE2**

3 Georgina I. López-Cortés<sup>1,2</sup>, Miryam Palacios-Pérez<sup>2</sup>, Gabriel S. Zamudio<sup>2</sup>, Hannya F.  
4 Veledíaz<sup>2,3</sup>, Enrique Ortega<sup>1,\*</sup>, Marco V. José<sup>2,\*</sup>

5 <sup>1</sup>*Department of Immunology, Instituto de Investigaciones Biomédicas, Universidad*  
6 *Nacional Autónoma de México, Mexico*

7 ✉ email: [ortsoto@unam.mx](mailto:ortsoto@unam.mx)

8 <sup>2</sup>*Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad*  
9 *Nacional Autónoma de México, 04510 Ciudad Universitaria, Mexico*

10 ✉ email: [marcojose@biomedicas.unam.mx](mailto:marcojose@biomedicas.unam.mx)

11 <sup>3</sup>*Universidad Latinoamericana, Nutrición, Campus Cuernavaca, Morelos*  
12

13 **Abstract**

14 As the SARS-CoV-2 has spread and the pandemic has dragged on, the virus continued to  
15 evolve rapidly resulting in the emergence of new highly transmissible variants that can be  
16 of public health concern. The evolutionary mechanisms that drove this rapid diversity are  
17 not well understood but neutral evolution should open the first insight. The neutral theory  
18 of evolution states that most mutations in the nucleic acid sequences are random and they  
19 can be fixed or disappear by purifying selection. Herein, we performed a neutrality test to  
20 better understand the selective pressures exerted over SARS-CoV-2 Spike protein, as well  
21 as in four of the identified health concern variants. Lys and Thr have higher occurrence rate  
22 on the Receptor Binding Domain (RBD) than in the overall sequence whereas Cys, His, and  
23 importantly Arg have low occurrence rate both in the whole protein and the RBD. Amino  
24 acids that have lower occurrence than the expected neutral control influence in the stability  
25 and or functionality of the protein. Our results show that most unique mutations either for  
26 SARS-CoV-2 or the variants of health concern are under selective pressures, which could  
27 be related either to the evasion of the immune system, increasing the virus' fitness or  
28 altering protein – protein interactions with host proteins. Altogether all these forces have  
29 shaped the Spike protein. Understanding the evolutionary forces that act upon Spike protein  
30 may help designing better treatments and vaccines that target variants of health concern.

31 **Keywords:** Spike evolution, Neutrality test, Phylogenetic analyses, Binding of Spike-  
32 ACE2, SARS-CoV-2 variants, Selective pressure

33

34

## 35 Introduction

36 The ongoing COVID-19 pandemic caused by the rapid global transmission of SARS-CoV-  
37 2<sup>1</sup> illustrates the planetary consequences of recurrent episodes of zoonotic transmission  
38 from animals to human populations. At least seven coronaviruses have been identified to  
39 infect humans causing principally respiratory difficulties but only three of them pose  
40 potential pandemic threats<sup>2-4</sup>. Among the 4 genera of coronavirus, only the  
41 *Alphacoronavirus* and the *Betacoronavirus* can infect humans<sup>5</sup>. These two genera have a  
42 common ancestor that infects bats while the *Gammacoronavirus* and *Deltacoronavirus*  
43 have a bird coronavirus origin<sup>5</sup>. This means that human coronavirus may be directly related  
44 to bat coronavirus or to other mammals as intermediate hosts. Phylogenetic analyses have  
45 revealed that SARS-CoV-2 is a *Betacoronavirus* related to the bat *Rhinolophus affinis*  
46 coronavirus Bat-SL-RaTG13 and the Malayan pangolin (*Manis javanica*) coronavirus, and  
47 that SARS-CoV-2 and SARS-CoV belong to the same B lineage, whereas MERS-CoV  
48 belongs to the C lineage<sup>1,6</sup>.

49 Structural and genomic analysis of viral components are key for understanding the  
50 evolution of the virus and being able to propose therapeutic strategies both to combat the  
51 pandemic and to prevent further spread. As all the coronavirus, SARS-CoV-2 recognizes  
52 and fuses into the host cells membranes through the Spike glycoprotein<sup>7</sup>. The SARS-CoV-2  
53 Spike glycoprotein (SARS2-S) attaches to the human Angiotensin Converting Enzyme 2  
54 (ACE2) expressed on the cell membrane and is then processed by host's proteases<sup>8</sup> which  
55 are necessary for fusion. SARS2-S is made up of the Subunit 1(S1) that contains the  
56 Receptor Binding Domain (RBD) and Subunit 2 (S2), responsible for fusion with the cell  
57 membrane<sup>9</sup>. Given the essential role of this protein in the virus life cycle, it is assumed that  
58 it has undergone strong evolutionary pressures to ensure the propagation of the virus. The  
59 human membrane protease ACE2 has been identified as the viral receptor for several  
60 coronavirus that infect humans, including other bat SARS-like coronavirus, the SARS-CoV  
61 and SARS-CoV-2, as well as the *Alphacoronavirus* hCoV- NL63<sup>6,7</sup>. Different analyses  
62 revealed that the RBD of Spike proteins of SARS-CoV-2, SARS-CoV and MERS-CoV  
63 allow binding to the receptor from various species while staying within a range of possible  
64 mutations<sup>10,11</sup>, although neither the binding affinities nor the effect of such mutations on the  
65 affinity have been quantified. What has been certainly demonstrated through structural  
66 analysis is that the binding affinity of SARS2-S protein to its receptor ACE2 is greater than  
67 the one of SARS-CoV to the same receptor<sup>12,13</sup>. The amino acids most probably responsible  
68 for the increase in affinity that could have resulted in enhancing the spread of the virus  
69 SARS-CoV-2 have already been proposed<sup>4,9,14-16</sup>.

70 Herein, we analyze the sequence of the Spike protein of the SARS-CoV-2 and compare it to  
71 the sequences of Spike proteins from other coronavirus, to best fit an evolution model that  
72 explains the amino acids preferences that have been selected for a higher affinity binding to  
73 the host's receptor. To understand the evolution of the spike protein, we applied an amino

74 acid substitution test (neutrality test) to identify the amino acids that deviate from neutral  
75 mutations. This test is directly related to the degeneracy of the Standard Genetic Code<sup>17</sup> and  
76 is applied to both the whole sequence of the Spike protein and to the sequence of the RBD.

## 77 **Materials and methods**

### 78 **Data sources**

79 The nucleotidic and amino acidic sequences of the Spike protein of 54 coronavirus were  
80 obtained from the GenBank (<https://www.ncbi.nlm.nih.gov/>). There were 6  
81 *Alphacoronavirus* and 48 sequences belonging to *Betacoronavirus* genus. The structures of  
82 SARS2-S (6X6P and 6XR8), the RBDs of SARS2-S in complex with ACE2 (6M0J),  
83 SARS-S RBD bound to ACE2 (2AJF), were downloaded from the Protein Data Bank  
84 (<https://www.rcsb.org/>). Also, the reference structure of SARS2-S was downloaded from  
85 the SARS-CoV-2-dedicated ZhangLab webpage  
86 (<https://zhanglab.ccmb.med.umich.edu/COVID-19/>). SARS-CoV-2 variants' Spike  
87 sequences were retrieved from Situation Reports deposited in the site outbreak.info.

### 88 **Neutral Evolution Model**

89 Forty-eight Spike sequences of *Betacoronavirus* were pairwise aligned. Each pair of protein  
90 sequences was aligned using MUSCLE<sup>18</sup> with default parameters. The protein alignment  
91 was used as template to derive a nucleotide alignment that would not have gaps that could  
92 split codons. From the nucleotide alignment a table of mutations was computed that  
93 account for the total of changes in codons. The table of codon mutations was transformed  
94 into an amino acid mutation matrix by adding up the values of the codons for a given amino  
95 acid. Hence, this matrix considers synonymous and non-synonymous mutations. The amino  
96 acid mutation matrix was computed for every pair of sequences and added up. Then, the  
97 matrix was normalized by rows, so that each row adds up to 1, and yields a probability  
98 transition matrix. The stationary distribution of the probability transition matrix was  
99 derived and compared to the control of neutral evolution as described in<sup>17</sup>. To assess the  
100 statistical robustness of the sample of sequences, a jackknife procedure was applied. The  
101 procedure of deriving the stationary distribution from the probability transition matrix of a  
102 sample of sequences was repeated to all possible subsets of 50 sequences. A confidence  
103 interval of 95% was computed around the stationary distribution derived from the set of 48  
104 sequences. The whole process was also applied to a set of 9 RBD ACE2 sequences and a  
105 confidence interval of 95% for the stationary distribution was computed and compared to  
106 the neutral control of evolution.

### 107 **Phylogenetic analysis**

108 All the evolutionary analyses were conducted with MEGA X software<sup>19</sup>. The multiple  
109 alignments of the spike sequences were performed with MUSCLE algorithm. The test for

110 the best evolutionary method for the *Betacoronavirus* Spike sequences resulted to be  
111 WAG+G+I+F, where the Invariable (I) value was 0.081 and the Gamma (G) value was  
112 1.143. Then a phylogenetic tree was constructed by Maximum Likelihood analysis.  
113 Another evolutionary test model for the Spike sequences that bind to ACE2 was also  
114 WAG+G+I+F, where I is equal to 0.064 and the G value was 6.54. SARS2-S sequence was  
115 compared to both groups i) the most proximal amino acid (a.a) sequences and ii) the ACE2  
116 binding CoVs. Consequently, unique mutations for SARS2-S and conserved residues were  
117 identified manually using both groups.

## 118 **Structural analysis**

119 The structures were cleaned to have the most accurate protein and complexes of the RBDs  
120 to its receptor. The structural analysis was visualized and analyzed with Chimera<sup>20</sup>, and I-  
121 TASSER<sup>21-24</sup>. The complex of SARS2-S RBD with the receptor was used to point out  
122 unique mutations and conserved residues. Distances between the amino acids involved in  
123 the protein – protein interaction were computed. Other parameters like the hydrophobicity  
124 and electrostatic potential were calculated for the a.a. in the interface. The same was  
125 calculated for the complex SARS-S RBD with the same receptor. The number of contacts,  
126 number of hydrogen bonds, hydrophobicity and mean distances were compared.  
127 Glycosylation sites were identified in both the linear and structural model. Besides, Spike  
128 mutations of the most prominent health concern variants of SARS-CoV-2 were identified in  
129 the three-dimensional model of SARS2-S and the structural models were predicted for each  
130 of them with I-TASSER platform. The reference structure was downloaded from the  
131 SARSCoV2-dedicated ZhangLab webpage (<https://zhanglab.ccmb.med.umich.edu/COVID-19/>)  
132 where the accurate predicted and curated structures of all the SARS-CoV-2 proteins  
133 are deposited. For each mutation physicochemical characteristics were discussed.

## 134 **Results**

135 The neutral theory of molecular evolution assumes that evolution is driven by random  
136 stochastic point mutations that eventually may be fixed by genetic drift or natural selection.  
137 From this point of view, we applied a neutrality evolution model to better understand the  
138 type of selective pressure acting upon a.a. present in the Spike protein<sup>17</sup>. This analysis  
139 revealed that in the *Betacoronavirus* genus Trp, Cys, His, Gly, Pro, Leu, Ser, and Arg  
140 underwent negative selective pressures, as the number of changes in these a.a. are lower  
141 than the expected by neutral evolution. In contrast, Tyr, Lys, Gln, Phe, Asn, Asp, Thr and  
142 Val, displayed positive selection (**Figure 1**). Mutations giving Met, Glu, Ile and Ala  
143 exhibited neutral or nearly neutral forces. To accurately analyze the great adaptation of  
144 SARS-CoV-2 to its receptor, we tested the neutrality of mutations in the RBD of the ACE2  
145 binding sequences, which are crucial for specific receptor recognition, and thus for  
146 infection. Importantly, Cys, His, Gly, Pro, Val, Ala, Leu, and Arg showed negative  
147 selection, whereas Lys, Gln, Phe, Asn, Asp, Ile, Thr, and Ser manifested positive selection

148 **(Figure 1)**. This means that it is less likely to find an Arg that appeared by random  
149 mutation than a Lys or a Thr, because the hexa-codonic Arg is under high negative  
150 selective pressure.

151 For the phylogenetic analysis, we carried out a multiple alignment of the a.a. sequences of  
152 Spike proteins of coronavirus including *Alphacoronavirus* and *Betacoronavirus*.  
153 Consequently, we identified the most related Spike sequences to SARS-CoV-2 **(Figure 2)**.  
154 In agreement with previous analysis, the phylogenetic tree computed shows that the bat  
155 coronavirus RaTG13 Spike protein exhibits the highest similarity with the SARS2-S  
156 followed by the pangolin coronavirus (PnCoV) Spike protein<sup>6</sup>. As expected, the S2  
157 subdomains had a high degree of similarity, so, we focused in the RBD sequence of the S1  
158 subdomain to identify mutations that could be advantageous for SARS2-S binding to  
159 ACE2. Therefore, a multiple alignment was performed including the Spikes of SARS-CoV-  
160 2, RaTG13, PnCoV, and several Spikes known to bind to human ACE2<sup>8</sup>. We discarded the  
161 *Alphacoronavirus* HCoV NL63 because even though it binds to the human ACE2, the  
162 orientation of the RBD is completely different. We identified point mutations in the RBD  
163 of SARS2-S that could be responsible for binding. Most a.a. were conserved among all the  
164 sequences, but there are few mutations that are unique for SARS2-S **(Table S1)**. Some of  
165 these mutations are present in CoV RaTG13 and PnCoV Spikes, what would suggest that  
166 these spike proteins could bind to the human ACE2. Interestingly, most mutations are  
167 located at the interface with ACE2 **(Figure S1)**. **Figure 3a** shows the interaction between  
168 SARS2-S RBD and ACE2 and **Figure 3b** show a close-up where the side chains of the  
169 amino acids involved in the protein-protein interaction are shown with sticks. The  
170 conserved residues are shown in pale pink as the rest of the structure (*i.e.*, Tyr 449, Tyr  
171 453, Asn 487, Tyr 489, Thr 500, Gly 502, Tyr 505) **(Table 1)** while the a.a. that are unique  
172 for SARS2-S, are shown in red **(Figure 3b)**. Among the 17 a.a. involved in the interaction,  
173 10 are unique for SARS2-S, including Lys 417, Gly 446, Leu 455, Phe 456, Ala 475, Phe  
174 486, Gln 493, Gly 495, Gln 498 and Asn 501 **(Table 2)**. Compared to SARS-S, SARS2-S  
175 forms more hydrogen bonds with the receptor, 8 and 11, respectively. This is because all  
176 a.a. involved in forming hydrogen bonds are shared, except for 2 mutations that lead to the  
177 formation of new hydrogen bonds with the receptor's surface (*i.e.*, Gly 446 and Lys 417).  
178 Apart from the identification of a.a. that are unique for SARS-CoV-2, our analysis revealed  
179 that a.a. important for maintaining the structure of both the domain and the complete  
180 protein, such as Cys residues as well as the glycosylated a.a. (*i.e.* Asn) and most Gly and  
181 Pro, are highly conserved.

182 Moreover, we measured the distances between the amino acids at the interface of SARS2-S  
183 with ACE2 and compared them with the distances of SARS-S with ACE2 (Supplementary  
184 Information). Seventeen a.a. of SARS2-S contact 17 a.a. of the receptor, with a mean  
185 distance of 3.563 Å. In contrast, 15 a.a. of SARS-S contact 18 a.a. of the receptor, at a mean  
186 distance of 3.605 Å. We also compared the hydrophobicity and electrostatic potential of the

187 interface of the spike proteins, and we observed that both have similar values (SARS2-S  
188 interface: minimum -27.2, mean -4.743 and maximum 22.83 of hydrophobicity potential  
189 while SARS-S's values are -26.41, mean -3.47 and maximum 23.19) (**Figure 4**). Both Phe  
190 456 and Phe 486 in SARS2-S generate more hydrophobic contacts than the Leu at the same  
191 positions present in SARS-S. Also, there is slightly bigger hydrophilic surface at the other  
192 edge of the interface in SARS2-S. All these factors may contribute to the higher binding  
193 affinity of SARS2-S to ACE2 that has been reported<sup>12,13</sup>.

194 Then we concentrated in the mutations of spikes proteins of SARS-CoV-2 variants to  
195 unravel the evolutionary behavior of the virus. The variants that have received most  
196 attention due to their importance for public health are those identified in the United  
197 Kingdom (B.1.1.7/ UK), in South Africa (B.1.135/ SA), in Brazil (B.1.1.248 / P.1/ BR),  
198 and in California USA (B.1.429/ CAL.20C/ CL). Interestingly, three variants share two  
199 mutations N501Y and D614G (UK, SA, BR variants). Tyr is a positive selected a.a. for the  
200 complete spike protein of *Betacoronavirus*, but neutral for the RBDs of ACE2 binding  
201 sequences. This means that in this domain, Tyr is not subjected to any selective pressure  
202 and therefore this change occurred randomly. In contrast, these same variants have a Gly  
203 which is clearly negatively selected at position 614 (**Figure 1**). It is important to note that  
204 the first reported variant was the one isolated from UK which has the highest percentage of  
205 neutral mutations (*circa* 30%), probably because selective pressures had not shaped the  
206 variant yet. The structure of the spike protein of SARS-CoV-2 and the variants of concern  
207 are illustrated in **Figure 5**. The structure of SARS2-S is shown with a zoom of the RBD  
208 painted in green (**Figure 5A**). Cys of the RBD are shadowed in yellow and the two  
209 glycosylated Asn are magenta. All sites of point mutations in the variants are shadowed in  
210 cyan and deletions in grey. Predicted structures of four SARS-CoV-2 variants (UK, BR,  
211 SA, and CL) with mutations shown in cyan are shown in **Figure 5B**, as well as the  
212 comparison with the reference structure. At the center, the reference structure overlapped  
213 with the predicted structure of variants. To note, the predicted structure of BR and CL  
214 variants protrude the RBD at a different position than that in the reference structure and the  
215 UK and SA variants. In total, the UK's variant had 1 positively selected mutation, 3  
216 negatively selected, and 3 neutral mutations, besides three deletions (**Table 3**). SA variant  
217 has 2 a.a. positively selected, 1 a.a. negatively selected and 2 neutral mutations.  
218 Additionally, the BR variant shares a third mutation with the one first isolated from South  
219 Africa, E484K; Lys is under positive selective pressure. In total, the BR variant has the  
220 higher number of mutations: 6 positively selected, 3 negatively selected and 2 neutral.  
221 Lastly, the CL variant has only 2 negatively selected and 1 neutral mutation.

222 Finally, we looked at residues that are potential glycosylation sites. For SARS2-S the  
223 reported glycosylation sites of two of the structures 6X6P and 6XR8 revealed that there are  
224 14 Asn forming glycoside bonds (**Table 4**). Compared to the other ACE2 binding  
225 sequences, three Asn (17, 149 and 657) are unique to SARS2-S, while the rest (11) are

226 shared. To note, PgCoV and bat CoV RaTG13 express all potential glycosylation sites  
227 identified in the causal agent of COVID-19. These three unique Asn could have been  
228 important for the spread of the viruses previously mentioned but not directly affecting  
229 binding to the receptor ACE2. However, it seems that most Asn with the capability to form  
230 a glycosidic bond are unlikely to mutate.

## 231 **Discussion**

232 The neutral theory of evolution states that most mutations in nucleic acid sequences are  
233 random, and these can be fixed by different evolutionary mechanisms. We exploited a  
234 neutrality test to interrogate the molecular evolution of the spike protein. Herein, we  
235 constructed the neutral evolution model for spike proteins of *Betacoronavirus*, focusing on  
236 the evolution of the RBD and the possible implications for binding to its receptor. Positive  
237 selective pressures cause a.a. to be fixed in higher frequencies than neutral mutations, while  
238 negative selective pressures cause a.a. to appear in lower frequencies than a neutral variant.  
239 Therefore, most fixed mutations under negative selective pressures remain because they are  
240 advantageous for the protein either increasing stability, affinity to a ligand, or others. These  
241 pressures are similar among homologous proteins of similar organisms and, importantly,  
242 they are the major driving forces for adaptation.

243 Applying the neutrality test, we identified the amino acids that had suffered negative  
244 selection through evolution of spike proteins of *Betacoronavirus*. One of these is Trp, that  
245 is encoded by only one codon and thus its frequency of occurrence is expected to be the  
246 lowest. However, Arg, Pro, Gly, His and Cys that can be coded by 6, 4, 4, 2 and 2 codons  
247 respectively, appeared at frequencies significantly lower than the expected by neutral  
248 mutations. Of note the polybasic motif (RRAR) in SARS2-S which is a tremendously  
249 important site for infection, resulted from the insertion of the motif PRRA (4 a.a. under  
250 negative selection). This insertion has been a crucial virulence factor that enables the  
251 cleavage by furin protease which generates a neuropilin-1 binding motif that enhances  
252 internalization<sup>25,26</sup>. Within the context of the neutral theory of evolution, the probability for  
253 inserting these four a.a. was very low. However the insertion of a furin cleavage site  
254 (RXXR) is not new in CoVs<sup>27</sup>. This motif implies the insertion of 12 nucleotides, but since  
255 Arg is hexa-codonic and Ala is tetra-codonic, the probability of appearance of this sequence  
256 was not that low. Once presented, it may have remained because this motif has provided a  
257 high increase in virulence to the etiological agent of COVID-19. Other *Betacoronavirus*  
258 present similar polybasic motifs but they were achieved by point mutations rather than by  
259 an insertion of four amino acids, such as, the Spikes of the Murine Hepatitis Virus  
260 (YP009824982.1), Murine CoV RA59/R13 (ACN89689.1), Murine CoV RA59/SJHM  
261 (ACN89705.1), Rat CoV (YP003029848.1), HCoV HKU1 (YP173238.1), HCoV OC43  
262 (YP009555241.1), Rabbit CoV (YP005454245.1), Canine CoV (AQT26498.1), Human  
263 enteric CoV (ACJ35486.1), Bovine CoV (NP150077.1) and the Sambar deer CoV US/OH-  
264 WD388TC/1994 (ACJ67012.1).

265 The neutral evolution model applied to the RBD sequences that bind to ACE2, shows that  
266 there are amino acids that have similar selective pressure as within the whole protein. The  
267 negatively selected a.a. such as Cys, His, Gly, Pro and Arg are more likely to affect the  
268 thermodynamic stability of the protein, which could impact the structure, the function, or  
269 protein- protein interactions. Thus, they are conserved in certain positions and it is unlikely  
270 that they appear by mutation in other positions. For example, eight Cys establish disulphide  
271 bonds that preserve the structure of the RBD (C336-C361, C379-C432, C391-C525, and  
272 C480-C488). In contrast, Asn has been positively selected such that the probability of  
273 finding an Asn that appeared by mutation is higher. This does not mean that Asn in certain  
274 positions are not conserved; in the RBD, the two glycosylated Asn (N331, N343) are  
275 conserved and mutations in these positions are not favored. The alignment with sequences  
276 of other RBD suggests that those Asn residues are probably glycosylated as well. Like Asn,  
277 the amino acids Glu, Lys, Gln, Phe, Asp and Thr have higher occurrence than predicted by  
278 neutral mutations; this means that there must be a selective pressure that favors them.

279 Interestingly, in the RBD, Lys and Thr are the two a.a. which show the highest deviation  
280 from the neutral mutation model; their chemical characteristics may favor interaction with  
281 the receptor or may be important for maintaining the domain's structure. This is not related  
282 to the predicted probability of occurrence, which is relatively high for Thr but relatively  
283 low for Lys (**Figure 1**). Lys is a basic a.a. coded by 2 distinct triplets (AAA and AAG) and  
284 the side chain consists of four carbons ending with an amino group which gives it a positive  
285 charge. Thr is a small and polar amino acid which has a hydroxyl group; and in contrast to  
286 Lys it is coded by four different codons (ACU, ACC, ACA and ACG). By neutral mutation,  
287 the probability of occurrence of Thr is higher because any mutation in the third position  
288 maintains Thr, therefore a mutation in any of the first two positions, or an insertion or  
289 deletion that moves the reading frame are required for a non- synonymous mutation.

290 The non-polar Phe has been positively selected in the RBD of SARS2-S where there are  
291 two Phe involved in the interaction with ACE2, Phe 456 and Phe 486. Both substitute for  
292 Leu expressed in other ACE2 binding sequences which contribute to generate more  
293 hydrophobic interactions with atoms from ACE2: SARS2-S Phe456 with Asp 30, Thr 27,  
294 Lys 31 from ACE2 and SARS2-S Phe486 with Leu 79 and Met 82 and Tyr 83 from ACE2.  
295 These hydrophobic interactions were probably the driving force for maintaining the  
296 mutations. In comparison to the neutral model of the whole spike protein, Tyr has neutral  
297 evolution in the RBD, meaning that there is no selective pressure that favors or disfavors  
298 the mutations towards Tyr. However, there are four Tyr in the interface with the receptor,  
299 and three variants of SARS-CoV-2 also express a fifth one. This may suggest a great  
300 importance of Tyr's properties in protein – protein interactions.

301 The variants of SARS-CoV-2 were analyzed with the same neutrality test. The UK variant  
302 shows several mutations that could contribute to its higher transmissibility. The P681H  
303 substitution may cause slight differences in the secondary structure immediately before the

304 furin cleavage site. Pro introduces slight bends to protein structures because the amino  
305 group is binding both the  $\alpha$ C and the lateral R group in a cyclic form. The resulting  
306 structure could change the susceptibility for furin cleavage, but further analyses are needed  
307 to confirm this. Interestingly, three deletions of amino acids in this variant do not change  
308 the structure and function of the protein, but probably contribute to the decrease of  
309 recognition by patients' serum.

310 The mutation D1118H of UK variant introduced another His in S2 that may alter the  
311 structure and function because the opposite charge. However, other mutations may not  
312 influence the structure and function of the protein. In this same variant, UK, the substitution  
313 S982A transduces into an a.a. under neutrality but in the same group according to its polar  
314 requirement<sup>28,29</sup>. In the SA variant, the mutation A701V turns to an amino acid with  
315 positive selection through a mutation in the second position of the triplet. The Brazil variant  
316 is the variant with more mutations all along the Spike protein so far described, and  
317 furthermore most of these mutations imply changes towards a.a. with positive selection,  
318 such as L18F, T20N, D138Y and H655Y. Also, P26S and R190S changed to Ser (slightly  
319 negative selected) and T1027I is a neutral mutation. The California variant has only one  
320 mutation drove by neutrality S13I, and two mutations negatively selected W152C and  
321 L452R, the first resulting in the change of a.a. with similar polar requirements, and the  
322 second mutation involving a.a with different polar requirements.

323 The reference sequence of SARS2-S has Lys 417 which enables the formation of a  
324 hydrogen bond with the receptor in comparison to other ACE2 binding sequences, however  
325 SA and BR variants substitutes this into an Asp (K417D). It remains to confirm whether if  
326 these Spikes could be able to maintain the hydrogen bond as the N atom of the R chain  
327 which was donating the electron, is absent. Similarly, the mutation E484K, shared with BR  
328 variant too, occurs towards an a.a. with almost the same polar requirement even though the  
329 charge changes. This mutation leads to the a.a. with the highest positive selection. Other  
330 shared mutation is D614G substitution which is present in UK, SA and BR. The mutation  
331 translates to a small a.a. without charge which is negatively selected. It has been proven  
332 that this substitution increases binding to ACE2 in comparison with the ancestral virus,  
333 therefore infectivity and transmission of the variant also has raised<sup>30-32</sup>. D614G alters the  
334 affinity to the receptor due to a conformational change that causes the RBDs to turn into the  
335 up position, which is necessary for receptor recognition<sup>[33]</sup>. Besides, N501Y substitution  
336 shared by the same variants probably alters more the binding to the receptor rather than  
337 favoring evading the immune response. Actually this substitution enables infection to mice  
338 cells through interaction with mouse ACE2<sup>34</sup>. It is probable that the strong selective  
339 pressure exerted on both mutations, has already drove them to fixation in human  
340 populations by genetic drift. Shared mutations among different health concern variants may  
341 arose by convergent evolution coming from a strong selection.

342 Mutations in the S2 subdomain would probably be less frequent to reach fixation because  
343 the sequence is crucial for the function, as consequence, there is a high degree of  
344 conservation. This subdomain is involved in fusing with the membrane, so non- polar  
345 amino acids are required. On the contrary, S1 is very important for receptor recognition,  
346 specifically the Carboxyl Terminal Domain (CTD) where the RBD is contained, so the  
347 sequence of this domain is crucial for adapting to the actual host or for the ability to infect  
348 other species. Interestingly, the first reported variants had more mutations in S2, and this  
349 number seems to progressively decrease in the latter variants. The exact biological  
350 significance of this observation is not clear. It is probable that the last variant reported  
351 could either i) had been mutating before and by the time it was reported it may have had  
352 enough time to fix or disappear neutral mutations at S2, or ii) it may have appeared recently  
353 and have mutated only in S1. Either way, the fact that Cal.20 variant had few mutations  
354 draws attention. Now, we cannot sustain any hypothesis because we discussed only four  
355 variants and we lack the accurate mutation rate of each variant and the transmission rate at  
356 different places.

357 Also, the accumulation of mutations is linked to the capability to correct errors.  
358 Coronavirus have RNA-dependent RNA polymerases which are prone to mistake, unlike  
359 other RNA virus, they have also a 3' to 5' exoribonuclease (nsp14-ExoN) that proofreads  
360 the new sequence<sup>[35]</sup>. Nsp14-ExoN is one of the major factors enabling long and stable  
361 RNA genomes. Therefore, the accumulation of errors slows down, and synonymous  
362 mutations become the most frequent. However, the neutrality test constructed here  
363 considers both synonymous and non-synonymous mutations, and this allowed us to obtain  
364 information of the types of selective pressures that influenced deviation from neutral  
365 mutations.

366 In other Coronavirus the highly glycosylated Amino Terminal Domain (NTD), play an  
367 important role in attachment to the host cell and immune response evasion. For influenza C  
368 virus and some coronavirus (HCoV HKU1 and HCoV OC43), attachment through 9-O-  
369 acetylated sialic acid receptors is crucial and constitute another species barrier<sup>36,37</sup>. The  
370 interaction with specific hosts' proteins facilitates the approaching of the fusion machinery  
371 to the cell membrane. Furthermore, saccharides mask potential epitopes recognized by  
372 antibodies, making it difficult to the immune system to eliminate the virus. It has been  
373 observed that the Spike glycoprotein must be shielded by the protective glycans from the  
374 immune system<sup>38</sup>. It remains to be determined whether the mutations near glycosylation  
375 sites interfere with the formation of glycosidic bonds in the variants of concern. Although  
376 L18F, T20N, P26S and D138Y substitutions from the Brazil variant and del144 from the  
377 UK variant are not precisely replacing glycosylation sites, they have been reported reduce  
378 recognition by antibodies<sup>39</sup>. Besides probably affecting glycosylation, these mutations  
379 could create new epitopes exposed on the protein surface making previous antibodies  
380 unable to recognize them. Hence, it is important to undercover whether if the glycosylation

381 of SARS-CoV-2 spike protein influence receptor recognition, membrane fusion, or immune  
382 evasion. Knowing the role of saccharides bound to the spike protein could help us  
383 understand more of the physiopathology of the virus and to develop better prophylactic or  
384 therapeutic strategies, effective against all variants.

## 385 **Conclusion**

386 The long-lasting pandemic, the wide geographic distribution, and the rapid contagiousness  
387 mainly during epidemic waves, have influenced the generation of variants of SARS-CoV-2.  
388 At global or local scale, the evolution of this virus can be appreciated. Vaccines and drugs  
389 have been developed and tested aiming to stop transmission which would also result in  
390 preventing the virus from mutating and developing new variants that cannot be recognized  
391 by newly developed treatments. Therefore, evolutionary studies play an important role in  
392 the prevention of epidemiological catastrophes and in the development of better treatments  
393 that covers most viral variants. The first evolutionary mechanism to consider is the fact that  
394 most mutations occur stochastically, meaning that neutral mutations are not the result of a  
395 selective pressure and do not respond to conferring any advantage or disadvantage to the  
396 virus. Some of them will be maintained until another substitution occurs; however, there  
397 are selective pressures that influence the fixation or disappearance of mutations. Selective  
398 pressures are similar among the viral particles in different hosts, and in the long term they  
399 shape proteins. Important selective pressures for SARS-CoV-2 evolution are related to the  
400 pathogenesis as mutations that improve fitness, alter interactions with host's proteins or  
401 evade the immune response. Consequently, they select mutations and conduct to adaptation.  
402 Shared mutations with different geographical origins may have been subjected to common  
403 selective pressures over specific residues, meaning that each mutation must have a given a  
404 significantly advantage.

405 Comparison of homologue proteins enables to construct an evolutionary model for  
406 stochastic mutations. The neutrality test computed shows the type of selective pressure for  
407 each amino acid in the spike protein of *Betacoronavirus*. This evolutionary study enables to  
408 understand and describe changes in SARS2-S sequence that affects its stability, structure,  
409 or function. One of the most relevant mutations is the insertion of a four- a.a. motif that  
410 allows the cleavage of a protease, that despite not being favored, a clear advantage in terms  
411 of the virus transmissibility won over neutral evolution mechanisms. Furthermore, selective  
412 pressures in the RBD favored the ability of SARS-CoV-2 to infect humans and chemical  
413 features were gained (*i.e.* increasing the number of hydrogen bonds and forming more  
414 hydrophobic contacts with the receptor). To note, SARS-CoV-2 continues to evolve rapidly  
415 throughout the globe, generating lineages with accumulated mutations. Here it has been  
416 shown that most of these mutations have been selected by selective pressures. Yet, most  
417 mutations in the interface and the ones in the variants of interest were favored implying  
418 other evolutionary mechanisms such as selection as important driving forces in the spike

419 glycoprotein that have enhanced the viral transmission. In the end mutations have allowed  
420 SARS-CoV-2 to become a threat to mankind, on the scale of a pandemic.

421

422 **References**

- 423 1. Zhao, X., Ding, Y., Du, J. & Fan, Y. 2020 update on human coronaviruses: One  
424 health, one world. *Med. Nov. Technol. Devices* **8**, 100043 (2020).
- 425 2. Peiris, J. S. M. *et al.* Coronavirus as a possible cause of severe acute respiratory  
426 syndrome. *Lancet* **361**, 1319–1325 (2003).
- 427 3. Zaki, A. M., van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. M. E. &  
428 Fouchier, R. A. M. Isolation of a Novel Coronavirus from a Man with Pneumonia in  
429 Saudi Arabia. *N. Engl. J. Med.* **367**, 1814–1820 (2012).
- 430 4. Wang, C., Horby, P. W., Hayden, F. G. & Gao, G. F. A novel coronavirus outbreak  
431 of global health concern. *Lancet* **395**, 470–473 (2020).
- 432 5. Woo, P. C. Y. *et al.* Discovery of Seven Novel Mammalian and Avian  
433 Coronaviruses in the Genus Deltacoronavirus Supports Bat Coronaviruses as the  
434 Gene Source of Alphacoronavirus and Betacoronavirus and Avian Coronaviruses as  
435 the Gene Source of Gammacoronavirus and Deltacoronavirus. *J. Virol.* **86**, 3995–  
436 4008 (2012).
- 437 6. Jaimes, J. A., André, N. M., Chappie, J. S., Millet, J. K. & Whittaker, G. R.  
438 Phylogenetic Analysis and Structural Modeling of SARS-CoV-2 Spike Protein  
439 Reveals an Evolutionary Distinct and Proteolytically Sensitive Activation Loop. *J.*  
440 *Mol. Biol.* **432**, 3309–3325 (2020).
- 441 7. Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor  
442 usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat. Microbiol.* **5**,  
443 562–569 (2020).
- 444 8. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and  
445 Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).
- 446 9. Huang, Y., Yang, C., Xu, X. feng, Xu, W. & Liu, S. wen. Structural and functional  
447 properties of SARS-CoV-2 spike protein: potential antivirus drug development for  
448 COVID-19. *Acta Pharmacol. Sin.* **41**, 1141–1149 (2020).
- 449 10. Lu, G., Wang, Q. & Gao, G. F. Bat-to-human: Spike features determining ‘host  
450 jump’ of coronaviruses SARS-CoV, MERS-CoV, and beyond. *Trends Microbiol.* **23**,  
451 468–478 (2015).
- 452 11. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable  
453 bat origin. *Nature* **579**, 270–273 (2020).
- 454 12. Wang, Q. *et al.* Structural and Functional Basis of SARS-CoV-2 Entry by Using  
455 Human ACE2. *Cell* **181**, 894-904.e9 (2020).
- 456 13. Yan, R. *et al.* Structural basis for the recognition of SARS-CoV-2 by full-length  
457 human ACE2. *Science (80-. ).* **367**, 1444–1448 (2020).

- 458 14. Benton, D. J. *et al.* Receptor binding and priming of the spike protein of SARS-  
459 CoV-2 for membrane fusion. *Nature* (2020) doi:10.1038/s41586-020-2772-0.
- 460 15. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to  
461 the ACE2 receptor. *Nature* **581**, 215–220 (2020).
- 462 16. Li, F., Li, W., Farzan, M. & Harrison, S. C. Structural biology: Structure of SARS  
463 coronavirus spike receptor-binding domain complexed with receptor. *Science* (80-. ).  
464 **309**, 1864–1868 (2005).
- 465 17. Zamudio, G. S., Prosdocimi, F., de Farias, S. T. & José, M. V. A neutral evolution  
466 test derived from a theoretical amino acid substitution model. *J. Theor. Biol.* **467**,  
467 31–38 (2019).
- 468 18. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high  
469 throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
- 470 19. Kumar, Stecher, Li, Knyaz & Tamura. MEGA version X. (2018).
- 471 20. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory  
472 research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
- 473 21. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: A unified platform for automated  
474 protein structure and function prediction. *Nat. Protoc.* **5**, 725–738 (2010).
- 475 22. Yang, J. *et al.* The I-TASSER suite: Protein structure and function prediction.  
476 *Nature Methods* vol. 12 7–8 (2015).
- 477 23. Yang, J. & Zhang, Y. I-TASSER server: New development for protein structure and  
478 function predictions. *Nucleic Acids Res.* **43**, W174–W181 (2015).
- 479 24. Zhang, C. *et al.* I-TASSER Genome wide structure and function modeling of SARS-  
480 CoV2. vol. 19 <https://zhanglab.ccmb.med.umich.edu/COVID-19/> (2020).
- 481 25. Cantuti-Castelvetri, L. *et al.* Neuropilin-1 facilitates SARS-CoV-2 cell entry and  
482 infectivity. *Science* (80-. ). **2985**, eabd2985 (2020).
- 483 26. Daly, J. L. *et al.* Neuropilin-1 is a host factor for SARS-CoV-2 infection. *Science*  
484 (80-. ). **3072**, eabd3072 (2020).
- 485 27. Wu, Y. & Zhao, S. Furin cleavage sites naturally occur in coronaviruses. *Stem Cell*  
486 *Res.* **50**, 102115 (2021).
- 487 28. Mathew, D. C. & Luthey-Schulten, Z. On the physical basis of the amino acid polar  
488 requirement. *J. Mol. Evol.* **66**, 519–528 (2008).
- 489 29. Woese, C. R., Dugre, D. H., Saxinger, W. C. & Dugre, S. A. The molecular basis for  
490 the genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **55**, 966–974 (1966).
- 491 30. Hou, Y. J. *et al.* SARS-CoV-2 D614G variant exhibits efficient replication ex vivo  
492 and transmission in vivo. *Science* (80-. ). **370**, 1464–1468 (2021).

- 493 31. Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* 1–6  
494 (2020) doi:10.1038/s41586-020-2895-3.
- 495 32. Zhou, B. *et al.* SARS-CoV-2 spike D614G variant confers enhanced replication and  
496 transmissibility. *bioRxiv* (2020) doi:10.1101/2020.10.27.357558.
- 497 33. Yurkovetskiy, L. *et al.* Structural and Functional Analysis of the D614G SARS-  
498 CoV-2 Spike Protein Variant. *Cell* **183**, 739–751.e8 (2020).
- 499 34. Gu, H. *et al.* Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine  
500 efficacy. *Science* (80-. ). **369**, 1603–1607 (2020).
- 501 35. Smith, E. C. & Denison, M. R. Coronaviruses as DNA Wannabes: A New Model for  
502 the Regulation of RNA Virus Replication Fidelity. *PLoS Pathog.* **9**, e1003760  
503 (2013).
- 504 36. Alejandra Tortorici, M. *et al.* Structural basis for human coronavirus attachment to  
505 sialic acid receptors. *Nat. Struct. Mol. Biol.* **26**, 481–489 (2019).
- 506 37. Hulswit, R. J. G. *et al.* Human coronaviruses OC43 and HKU1 bind to 9-O-  
507 acetylated sialic acids via a conserved receptor-binding site in spike protein domain  
508 A. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 2681–2690 (2019).
- 509 38. Sikora, M. *et al.* Computational epitope map of SARS-CoV-2 spike protein. *PLoS*  
510 *Comput. Biol.* **17**, e1008790 (2021).
- 511 39. Plante, J. A. *et al.* The Variant Gambit: COVID’s Next Move. *Cell Host Microbe*  
512 104743 (2021) doi:10.1016/j.chom.2021.02.020.

513 **Author Contributions:** Conceptualization: GL-C, MVJ, methodology: GL-C, GSZ, MP-P,  
514 MVJ; software GSZ, MP-P; validation: GL-C, GSZ, MP-P; formal analysis: GL-C,  
515 GSZ, MP-P, MVJ; investigation: GL-C, GSZ, MP-P, HVF, MVJ; resources MVJ; data  
516 curation: GSZ, MP-P; writing- original draft preparation GL-C, MVJ; review and  
517 editing of ms: GL-C, MVJ and EO; visualization GL-C, GZS, MP-P; supervision  
518 MVJ; project administration: EO and MVJ; funding acquisition MVJ. All authors have  
519 read and agreed to the published version of the manuscript.

520 **Conflicts of Interest:** The authors declare no conflict of interest.

## 521 **Funding**

522 GL-C is a doctoral student from the Programa Maestría y Doctorado de Ciencias  
523 Bioquímicas, Universidad Nacional Autónoma de México (UNAM), and received  
524 fellowship (699886) from CONACyT. GSZ is a doctoral student from Programa de  
525 Doctorado en Ciencias Biomédicas, UNAM, and received doctoral fellowship from  
526 CONACyT, number 737920. E.O. was supported by DGAPA-PAPIIT IN208320 and MVJ  
527 was financially supported by DGAPA-PAPIIT-IN201019 UNAM, México.

528

529

530 **Figures and tables legends**

531 **Figure 1.** Neutral evolution test of the a.a. of the spike protein and the RBD. The computed  
532 frequency of occurrence of individual a.a. substitution by neutral mutations (black line), the  
533 a.a. of the Spike protein of *Betacoronavirus* (green) and of the RBD (blue) of the ACE2  
534 binding CoVs are graphed. A.a. with higher occurrence than that predicted by purely  
535 stochastic changes refer to the a.a. is under positive selection pressure, while frequencies  
536 lower than the neutral prediction are amino acids that underwent negative selection  
537 pressures. A Jackknife procedure was performed with 95% of confidence interval.

538 **Figure 2.** Evolutionary analysis by Maximum Likelihood Method. Phylogenetic tree of the  
539 Spike protein of the *Betacoronavirus* genus. Representatives of the 4 lineages are shown.  
540 CoVs that bind to hACE2 are marked with an orange dot, whereas the green marker marks  
541 the CoVs that infect humans and use other receptors. Evolutionary analyses were conducted  
542 using MEGA X software.

543  
544 **Figure 3.** Interaction between the RBD of the spike protein of SARS-CoV-2 with ACE2.  
545 **A)** Interaction between the RBD of SARS2-S (pale pink) and the human receptor ACE2  
546 (blue). **B)** A close-up of the interface shows the R side chains of the a.a. of the RBD  
547 involved in the binding with the human receptor. Unique a.a. for SARS-CoV-2 are colored  
548 in red.

549 **Figure 4.** Chemical characteristics of spikes' RBD interface with the receptor. The surface  
550 of the amino acids involved in protein- protein interaction with the receptor is shown. **A)**  
551 The hydrophobic potential is colored from blue (hydrophilic), to white (neutral) and to gold  
552 (hydrophobic) to compare the RBDs of SARS-CoV (left) and SARS-CoV-2 (right). Head  
553 arrows point towards important changes in hydrophobicity potentials. **B)** The electrostatic  
554 potential of the surface of both interfaces shows slight differences. Scale goes from red  
555 (negative), to white (neutral) and to blue (positive) charged.

556 **Figure 5.** Structure of the spike protein of SARS-CoV-2 and the variants of concern. **A)**  
557 The structure of SARS2-S is shown with a zoom of the RBD painted in green. Cys of the  
558 RBD are shadowed in yellow and the two glycosylated Asn are magenta. All sites of point  
559 mutations in the variants are shadowed in cyan and deletions in grey. **B)** Predicted  
560 structures of four SARS-CoV-2 variants (UK, BR, SA, and CL) with mutations shown in  
561 cyan. At the center, the reference structure overlapped with the predicted structure of  
562 variants is shown.

563 **Figure S1** Protein- protein interaction between RBD of SARS2-S and ACE2. **A)**  
564 Tridimensional structure of the RBD (red) of the Spike of SARS-CoV-2 interacting with  
565 ACE2 (blue) and **B)** the linear representation of Spike protein showing the location of the  
566 RBD. Each loop in contact with the receptor is colored as in Table S1.

567 **Table 1.** Conserved residues involved in protein- protein interaction with ACE2 among the  
568 ACE2 binding Spikes. The selection type according to the neutrality test are indicated.

569 **Table 2.** Unique residues for SARS2-S involved in protein-protein interaction with ACE2.  
570 The selection type according to the neutrality test are mentioned specifically for SARS2-S.  
571 Other ACE2 binding Spike proteins expressed different a.a. Here SARS-S is shown as an  
572 example. The empty spaces in SARS-S are a.a. that do not interact with ACE2.

573 **Table 3.** Type of selective pressure for mutations in health concern variants. The mutations  
574 colored by physicochemical properties are enlisted for each variant and the type of selective  
575 pressure applied for each a.a. is shown. The type of pressure is specified either for the  
576 whole protein or for amino acids positioned in the RBD.

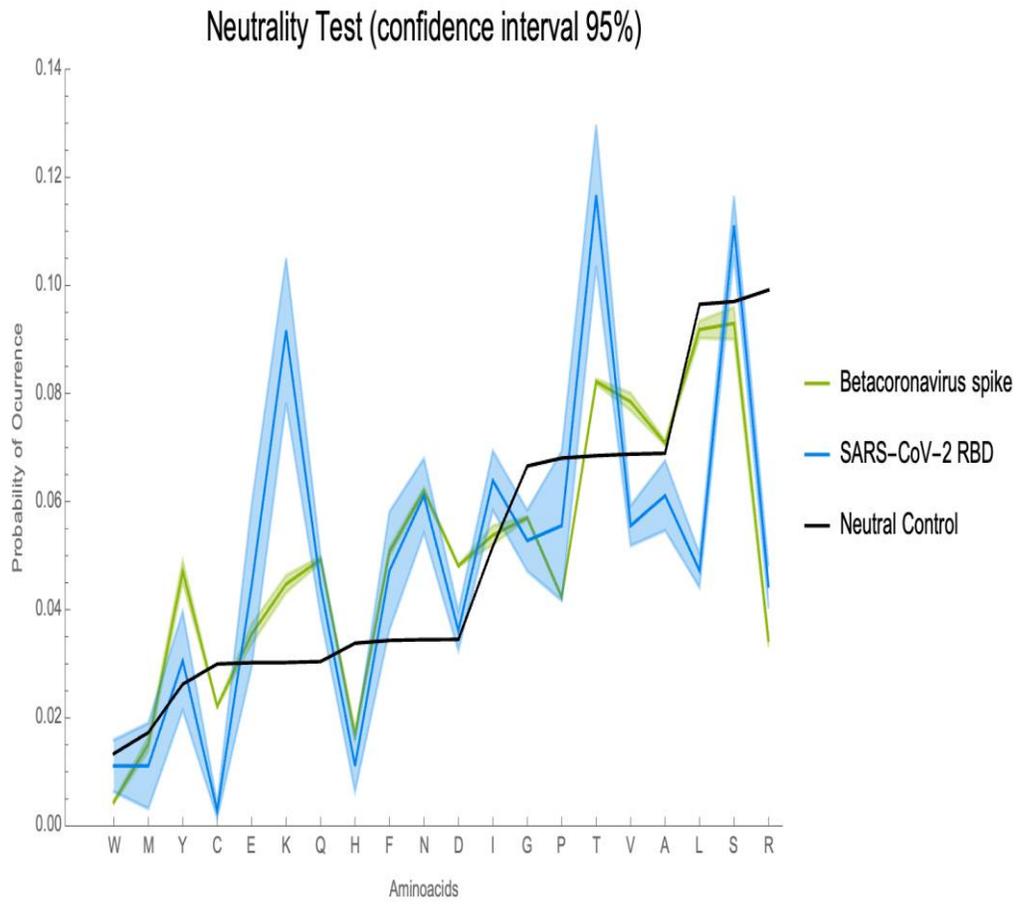
577 **Table 4.** Potential glycosylation sites. Comparison of the glycosylation sites reported from  
578 the structures of SARS2-S with all the ACE2 binding sequences.

579 **Table S1.** Unique mutations in the RBD for SARS-CoV-2. Compared to the other ACE2  
580 binding Spikes, SARS-CoV-2 Spike protein has 49 mutations in the RBD, the majority  
581 selected by a positive pressure. The point mutations of the SARS-CoV-2 Spike RBD are  
582 listed with the corresponding amino acid expressed in the rest of the Spike proteins that  
583 bind to ACE2; the amino acids are highlighted depending on the chemical nature: in yellow  
584 the non-polar, in green the polar and neutral amino acids, in blue the positively and in red  
585 the negatively charged amino acids. Amino acids in contact with ACE2 are in bold type  
586 letters.

587

588

589 **Figure 1**



590

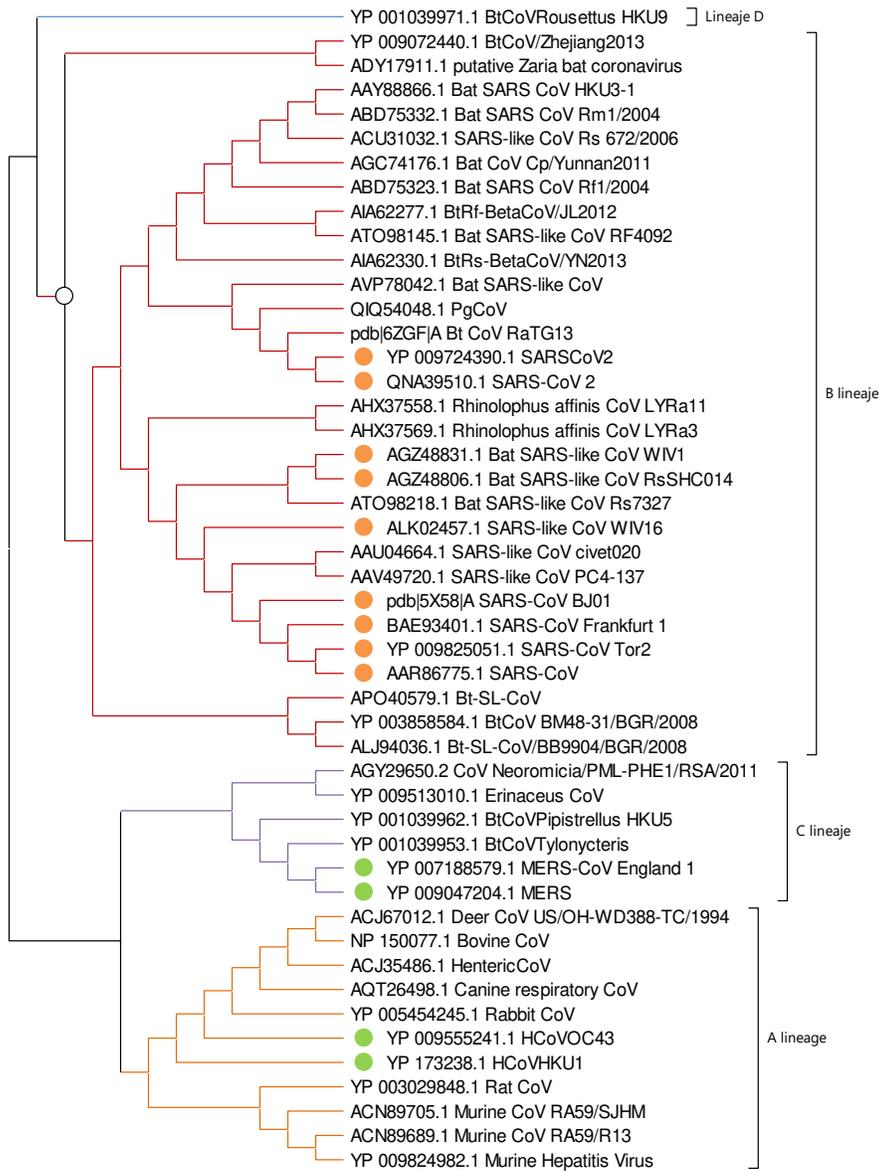
591

592

593

594

595 **Figure 2**



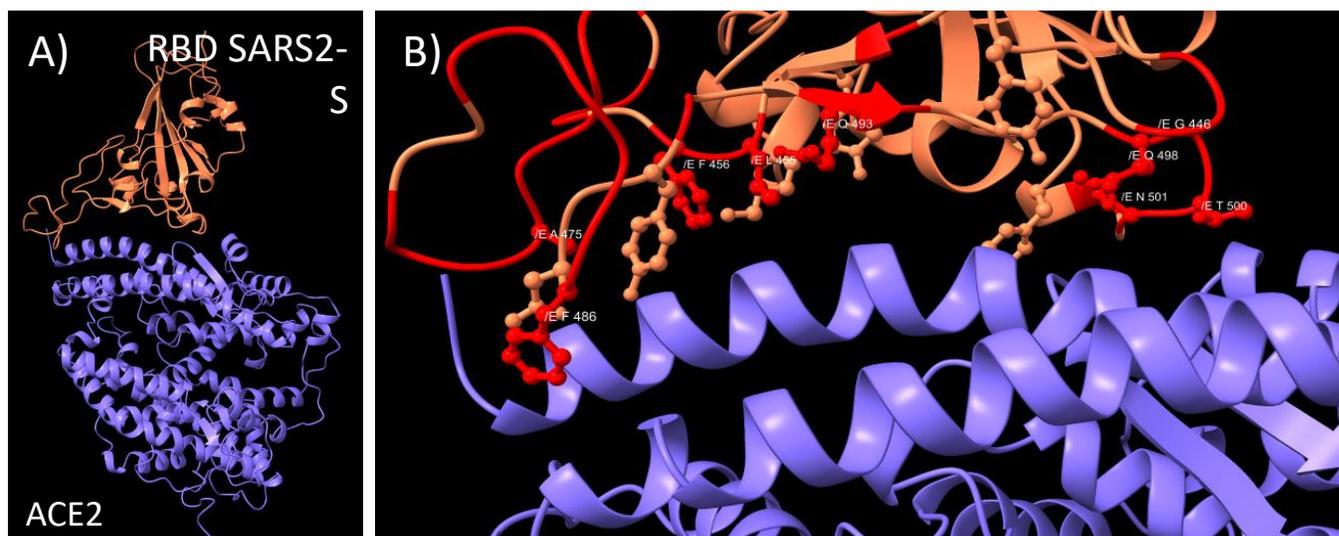
596

597

598

599

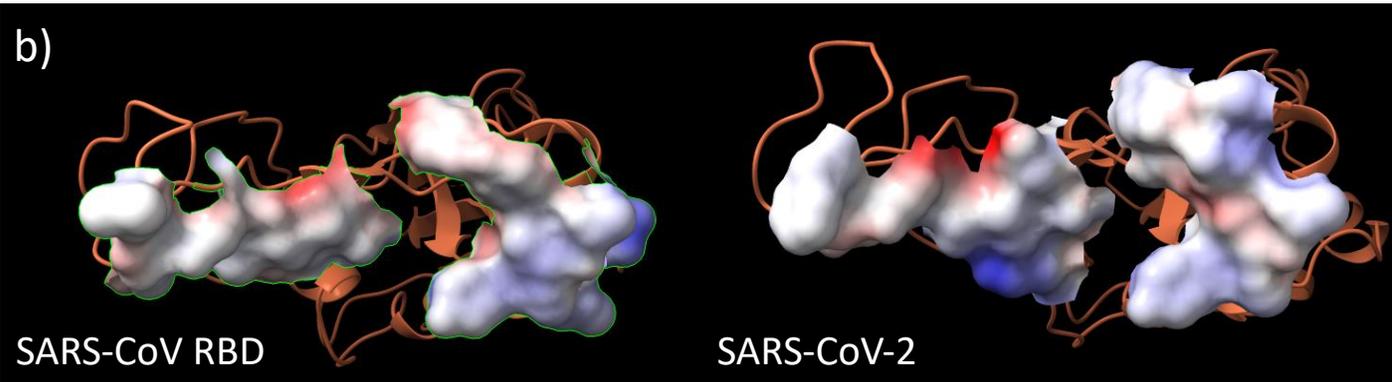
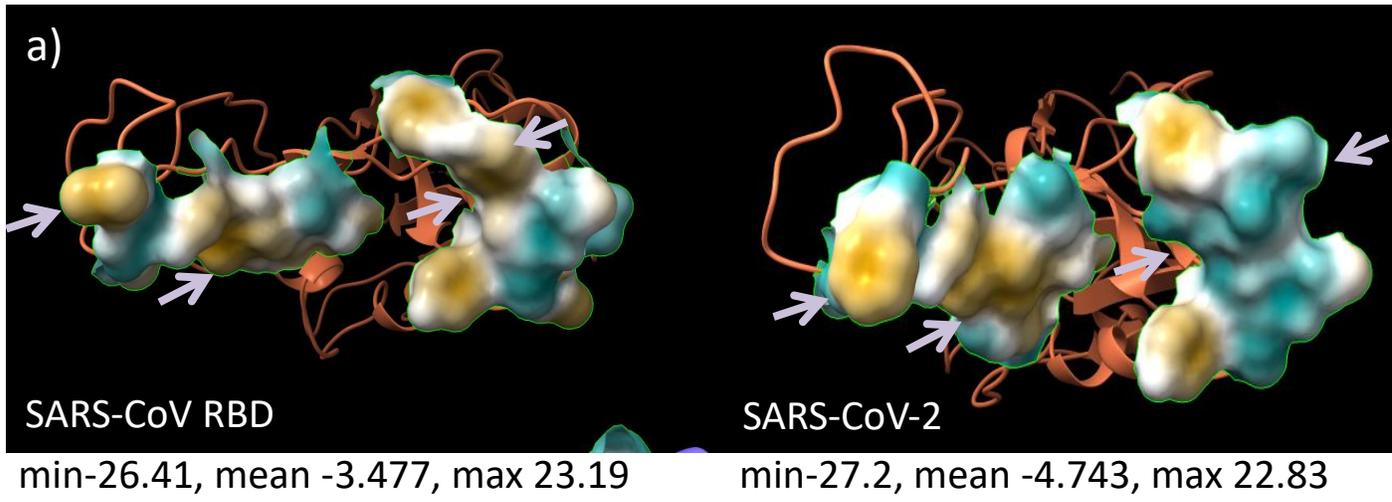
600 **Figure 3**



601

602 **Figure 4**

603



604

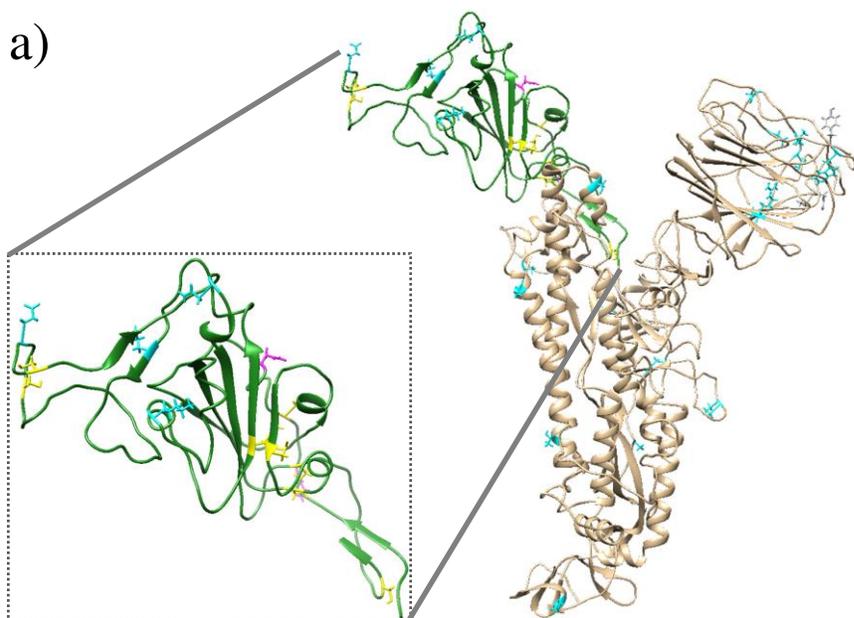
605

606 **Figure 5**

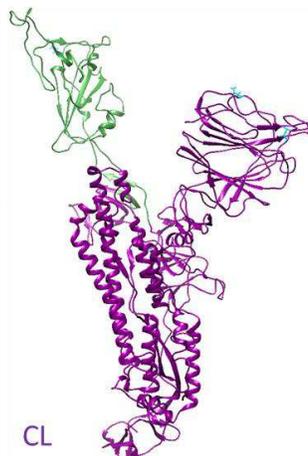
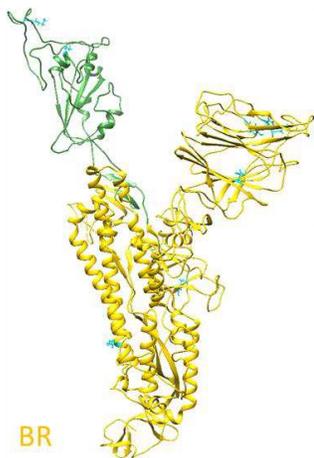
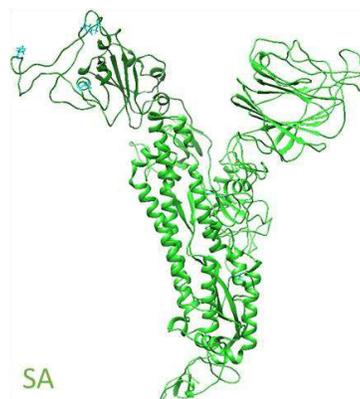
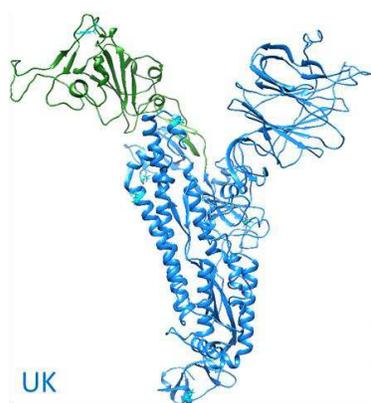
607 **a)**

608

609



616 **b)**



617

618 **Table 1**

Conserved a.a.		Selection according to neutrality
449	Y	<i>Neutral</i>
453	Y	<i>Neutral</i>
487	N	<i>Positive</i>
489	Y	<i>Neutral</i>
500	T	<i>Positive</i>
502	G	<i>Negative</i>
505	Y	<i>Neutral</i>

a.a. code Polar

619

620 **Table 2.**

SARS2-S	SARS-S		Selection according to neutrality
417	K		<i>Positive</i>
446	G	433 T	<i>Negative</i>
455	L	442 Y	<i>Negative</i>
456	F	443 L	<i>Positive</i>
475	A		<i>Negative</i>
486	F	472 L	<i>Positive</i>
493	Q	479 N	<i>Positive</i>
496	G		<i>Negative</i>
498	Q	484 Y	<i>Positive</i>
501	N	487 T	<i>Positive</i>

a.a. code Non-polar Polar Polar positive

621

622

623

624

625

626 **Table 3.**

Position	SARS2	UK/ B.1.1.7	SA/ B.1.351	BR/ P.1	CL/ CAL.20C	Selective pressure	
						Whole protein	RBD
13	S				I	Neutral	
18	L			F		Positive	
20	T			N		Positive	
26	P			S		Negative	
69	H	*					
70	V	*					
138	D			Y		Positive	
144	Y	*					
152	W				C	Negative	
190	R			S		Negative	
417	K		N	T			Positive
452	L				R		Negative
484	E		K	K			Positive
501	N	Y	Y	Y			Neutral
570	A	D				Positive	
614	D	G	G	G		Negative	
655	H			Y		Positive	
681	P	H				Negative	
701	A		V			Positive	
761	T	I				Neutral	
982	S	A				Neutral	
1027	T			I		Neutral	
1118	D	H				Negative	

a.a. code	Non-polar	Polar	Polar positive	Polar negative
-----------	-----------	-------	----------------	----------------

\* Deletions

627

628

629

630

631

632 **Table 4.**

SARS2-S	Conserved with ACE2 binding sequences	a.a. substitution
17	No. Only SARS2-S branch	T
61	Yes	
149	No. Only SARS2-S branch	T
165	Yes	
234	Yes	
282	Yes	
331	Yes	
343	Yes	
603	Yes	
616	Yes	
657	No. Only SARS2-S branch	D
709	yes	
1098	yes	
1134	yes	

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648 **Supplementary information**649 **Table S1.**

ACE2 binding CoVs		SARS- CoV2			Chemical changes/ characteristics	
Position	a.a.	Position	a.a.	Selective pressure		
335	P	348	A	<i>Negative</i>	Polarity	
341	E	354	N	<i>Positive</i>	Charge	
359	T	372	A	<i>Negative</i>	Polarity	
360	F	373	S	<i>Positive</i>	Polarity	
371	A	384	P	<i>Negative</i>	Polarity	
380	S	393	T	<i>Positive</i>	Polar	
389	V	402	I	<i>Positive</i>	Non- polar	
390	K	403	R	<i>Negative</i>	Same charge (+)	
393	D	406	E	<i>Neutral</i>	Same charge (-)	
404	V	<b>417</b>	K	<i>Positive</i>	Polarity and charge	
417	M	430	T	<i>Positive</i>	Polarity	
421	L	434	I	<i>Positive</i>	Non- polar	
425	T	438	S	<i>Positive</i>	Polar	
426	R	439	N	<i>Positive</i>	Charge	
loop 1	428	I	441	L	<i>Negative</i>	Non- polar
	430	A	443	S	<i>Positive</i>	Polarity
	431	T	444	K	<i>Positive</i>	Charge
	432	S/Q	445	V	<i>Negative</i>	Polarity
	<b>433</b>	T	<b>446</b>	G	<i>Negative</i>	Polar
	439	K	452	L	<i>Negative</i>	Polarity and charge
	<b>442</b>	Y/S	<b>455</b>	L	<i>Negative</i>	Non- polar
	<b>443</b>	L	<b>456</b>	F	<i>Positive</i>	Non- polar
	445	H	458	K	<i>Positive</i>	Same charge (+)
	446	G	459	S	<i>Positive</i>	Polar
	447	K	460	N	<i>Positive</i>	Charge
	449	R	462	K	<i>Positive</i>	Same charge (+)
	457	N	470	T	<i>Positive</i>	Polar
	458	V	471	E	<i>Neutral</i>	Polarity and charge
	459	P	472	I	<i>Positive</i>	Polarity
	460	F	473	Y	<i>Neutral</i>	Polarity
461	S	474	Q	<i>Positive</i>	Polar	
loop 2	462	P	<b>475</b>	A	<i>Negative</i>	Polarity
	463	D	476	G	<i>Negative</i>	Charge
	464	G	477	S	<i>Positive</i>	Polar

	465	K	478	T	Positive	Charge
	467	T	481	N	Positive	Polar
	469	P	482	G	Negative	Polar
	470	P	483	V	Negative	Polarity
			484	E	Neutral	Insertion
	471	A	485	G	Negative	Polarity
	<b>472</b>	<b>L</b>	<b>486</b>	<b>F</b>	<b>Positive</b>	Non- polar
	loop 3	476	W	490	F	Positive
<b>479</b>		<b>N</b>	<b>493</b>	<b>Q</b>	<b>Positive</b>	Polar
480		D	494	S	Positive	Charge
<b>484</b>		<b>Y</b>	<b>498</b>	<b>Q</b>	<b>Positive</b>	Polarity
loop 4	485	T	499	P	Negative	Polar
	<b>487</b>	<b>T</b>	<b>501</b>	<b>N</b>	<b>Positive</b>	Polar
	489	I	503	V	Negative	Non- polar
	505	N	519	H	Negative	Charge

650

651

652

653

654

655

656

657

658

659

660

661

662

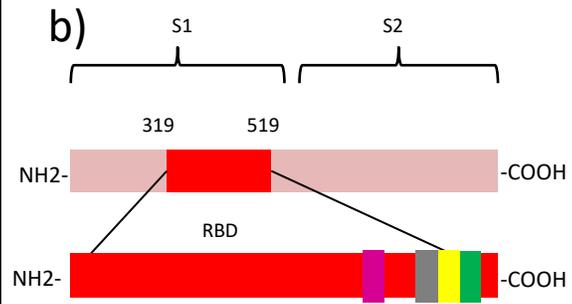
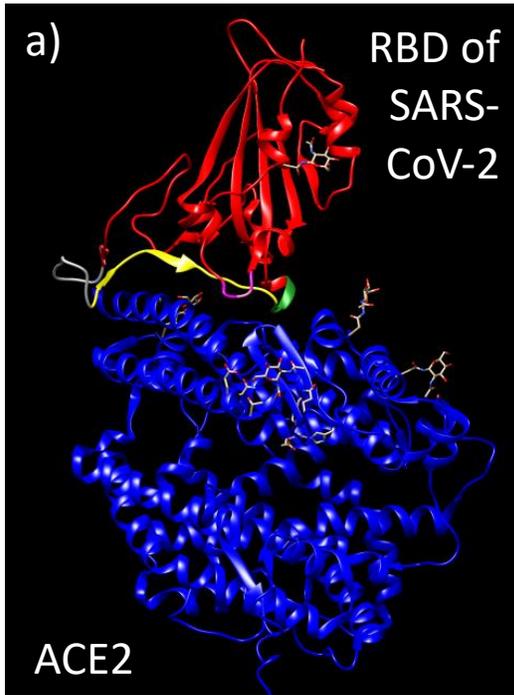
663

664

665

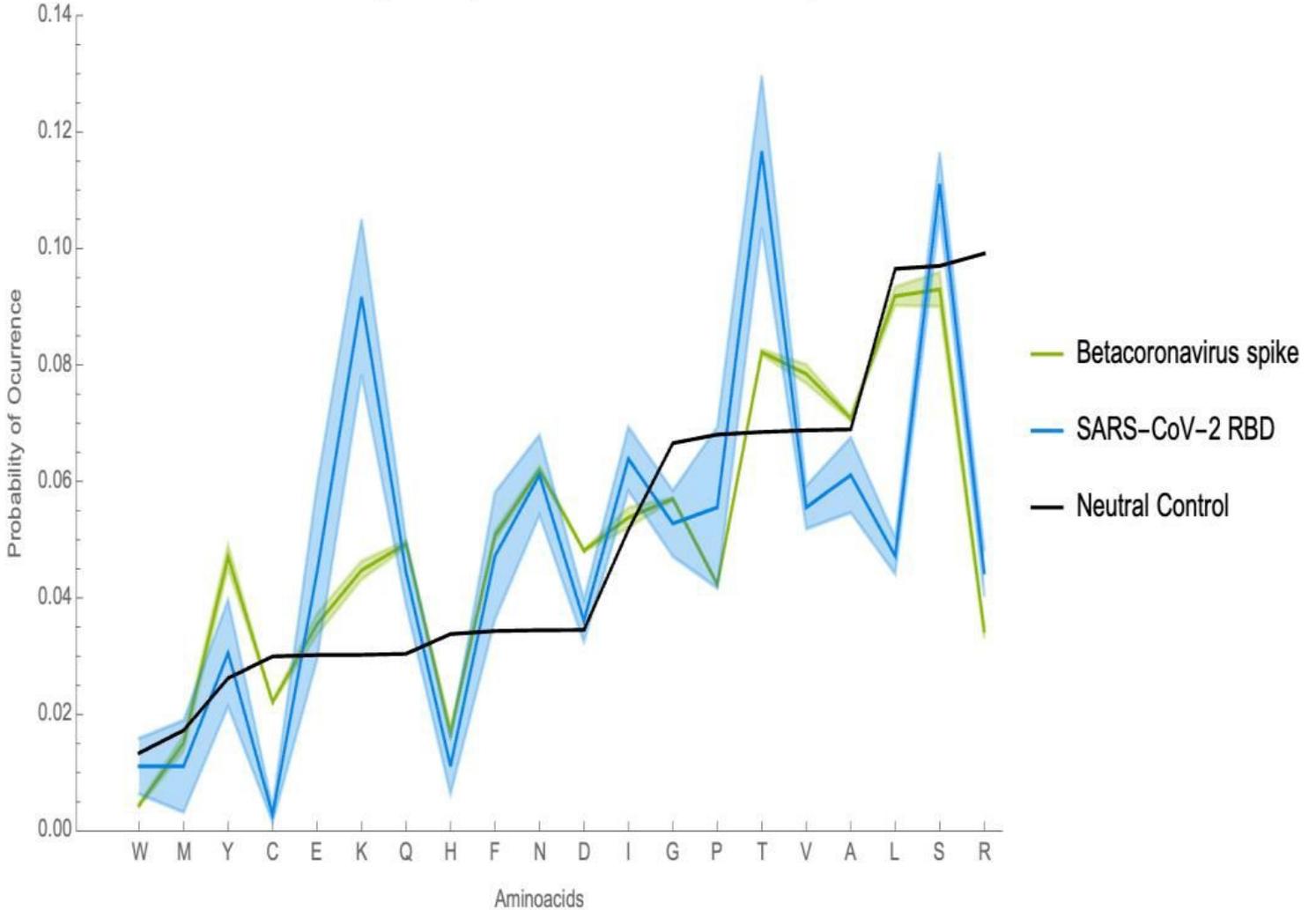
666 **Figure S1**

667



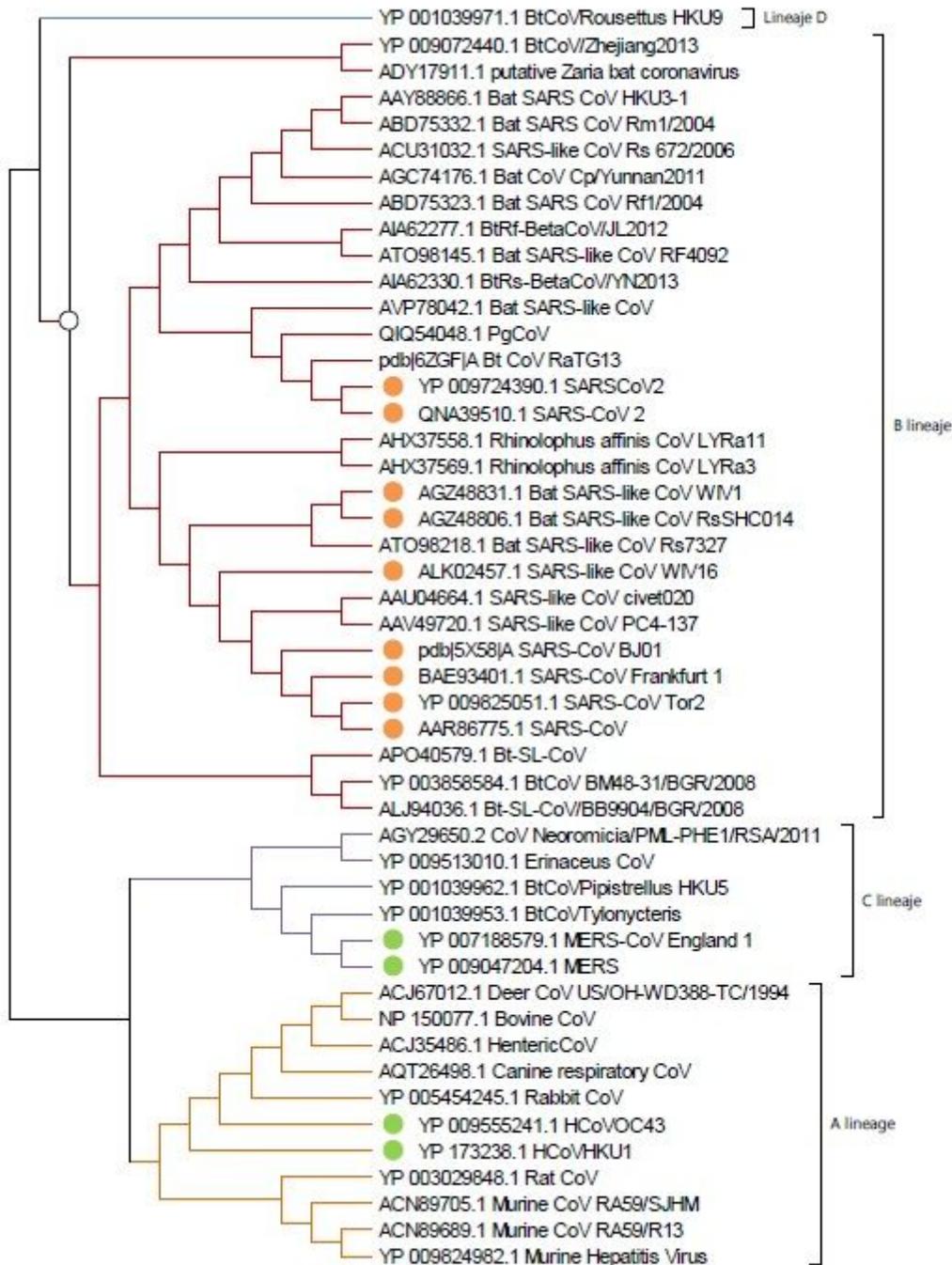
# Figures

## Neutrality Test (confidence interval 95%)



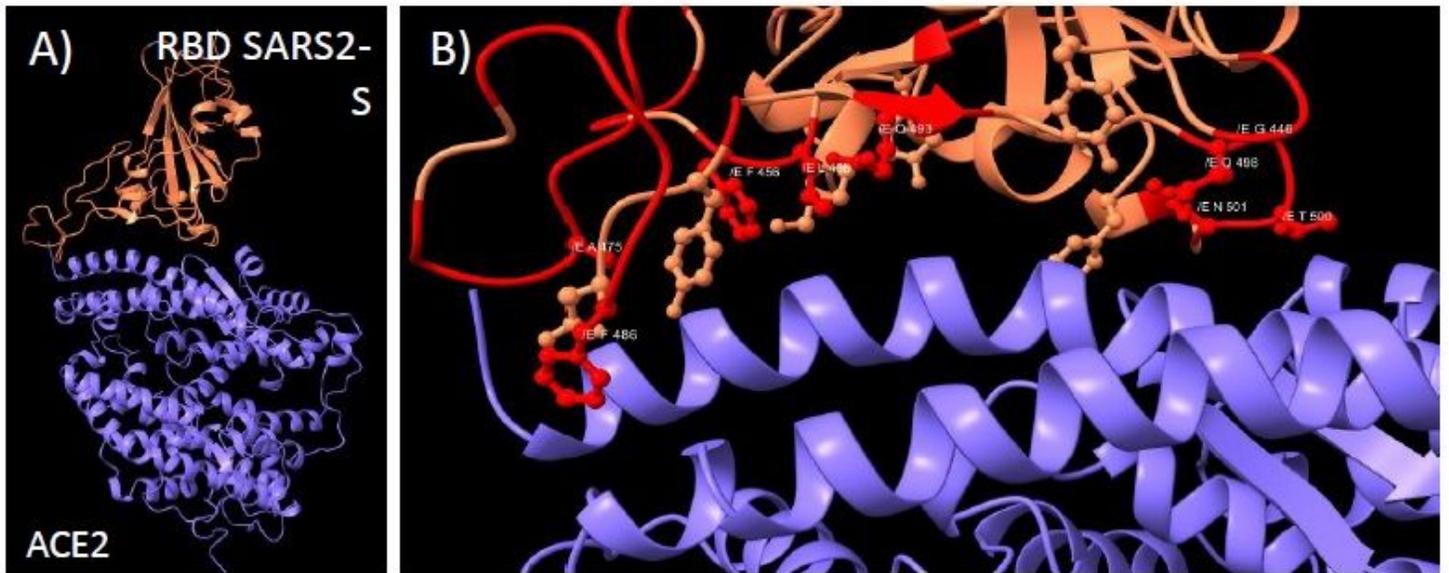
**Figure 1**

Neutral evolution test of the a.a. of the spike protein and the RBD. The computed frequency of occurrence of individual a.a. substitution by neutral mutations (black line), the a.a. of the Spike protein of Betacoronavirus (green) and of the RBD (blue) of the ACE2 binding CoVs are graphed. A.a. with higher occurrence than that predicted by purely stochastic changes refer to the a.a. is under positive selection pressure, while frequencies lower than the neutral prediction are amino acids that underwent negative selection pressures. A Jackknife procedure was performed with 95% of confidence interval.



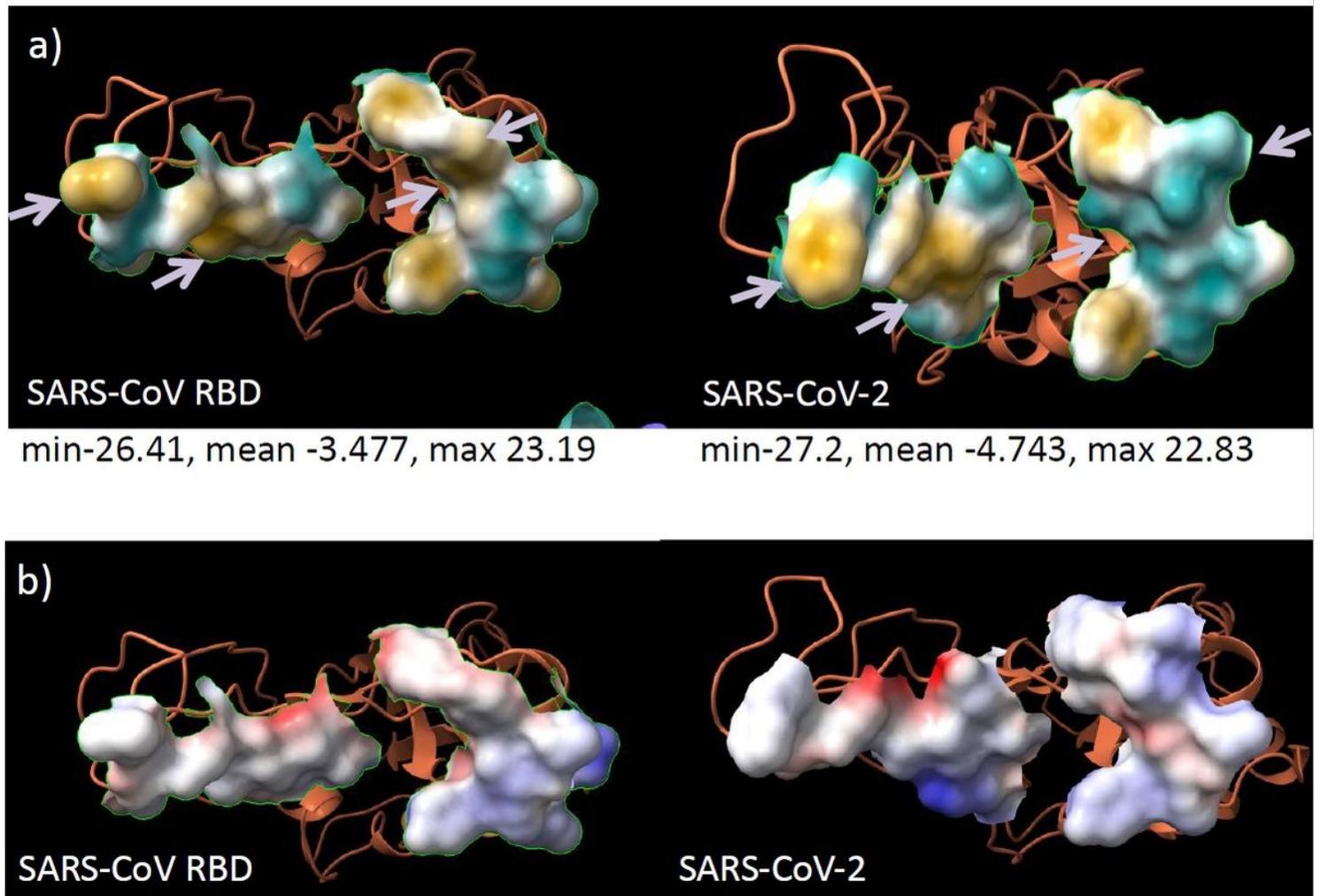
**Figure 2**

Evolutionary analysis by Maximum Likelihood Method. Phylogenetic tree of the Spike protein of the Betacoronavirus genus. Representatives of the 4 lineages are shown. CoVs that bind to hACE2 are marked with an orange dot, whereas the green marker marks the CoVs that infect humans and use other receptors. Evolutionary analyses were conducted using MEGA X software.



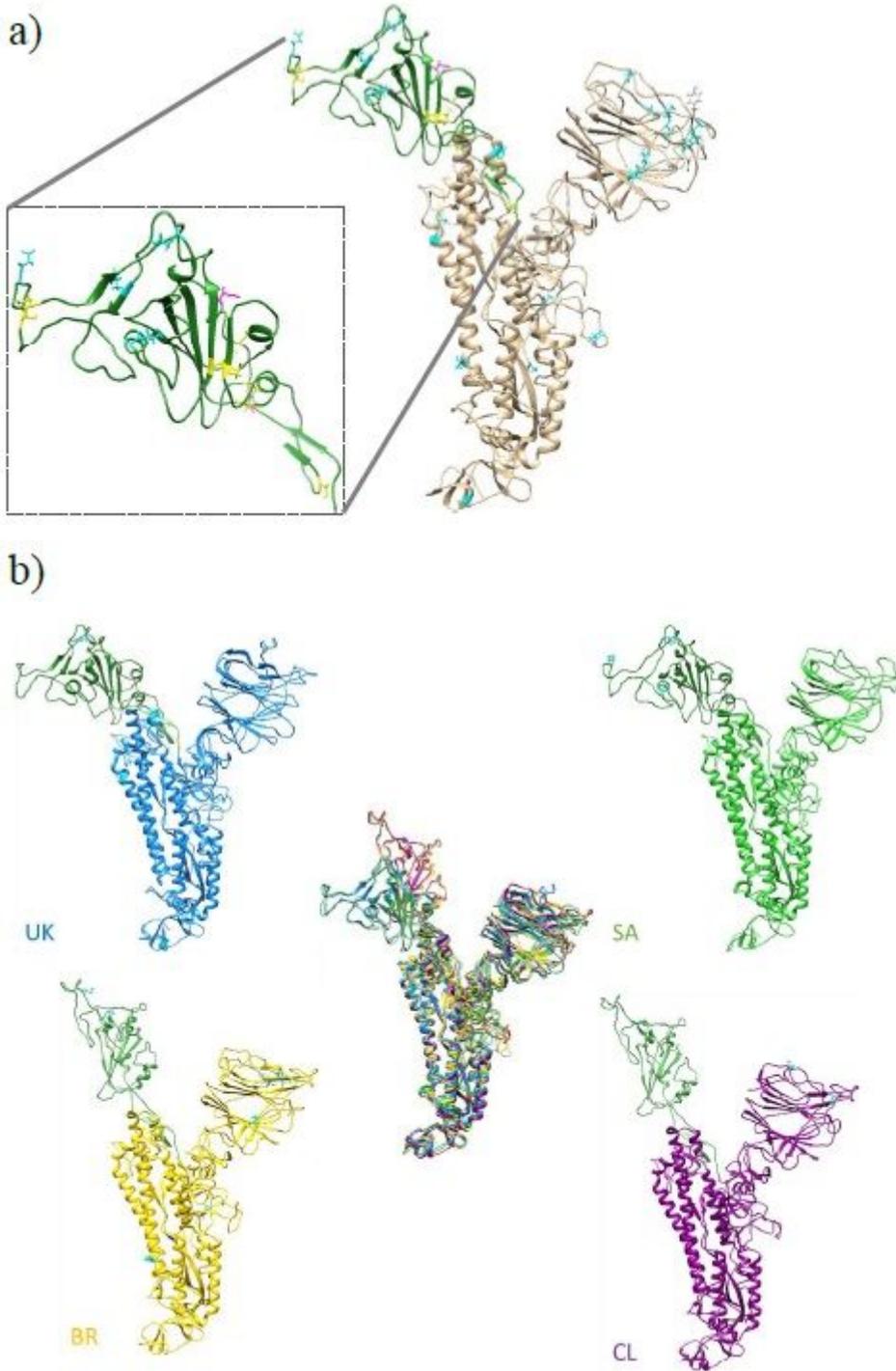
**Figure 3**

Interaction between the RBD of the spike protein of SARS-CoV-2 with ACE2. A) Interaction between the RBD of SARS2-S (pale pink) and the human receptor ACE2 (blue). B) A close-up of the interface shows the R side chains of the a.a. of the RBD involved in the binding with the human receptor. Unique a.a. for SARS-CoV-2 are colored in red.



**Figure 4**

Chemical characteristics of spikes' RBD interface with the receptor. The surface of the amino acids involved in protein- protein interaction with the receptor is shown. A) The hydrophobic potential is colored from blue (hydrophilic), to white (neutral) and to gold (hydrophobic) to compare the RBDs of SARS-CoV (left) and SARS-CoV-2 (right). Head arrows point towards important changes in hydrophobicity potentials. B) The electrostatic potential of the surface of both interfaces shows slight differences. Scale goes from red (negative), to white (neutral) and to blue (positive) charged.



**Figure 5**

Structure of the spike protein of SARS-CoV-2 and the variants of concern. A) The structure of SARS2-S is shown with a zoom of the RBD painted in green. Cys of the RBD are shadowed in yellow and the two glycosylated Asn are magenta. All sites of point mutations in the variants are shadowed in cyan and deletions in grey. B) Predicted structures of four SARS-CoV-2 variants (UK, BR, SA, and CL) with mutations shown in cyan. At the center, the reference structure overlapped with the predicted structure of variants is shown.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [FIGURES1.tiff](#)