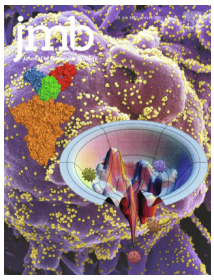

Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Research Article



# An Electrostatically-steered Conformational Selection Mechanism Promotes SARS-CoV-2 Spike Protein Variation

**Marija Sorokina [1,2,3], Jaydeep Belapure [4], Christian Tüting [4], Reinhard Paschke [3,5], Ioannis Papasotiriou [2], João P. G. L. M. Rodrigues [6] and Panagiotis L. Kastritis [1,4,5]***

*1 - **Institute of Biochemistry and Biotechnology,** Martin Luther University Halle-Wittenberg, Kurt-Mothes-Str. 3, 06120 Halle/Saale, Germany*
*2 - **RGCC International GmbH,** Baarerstrasse 95, Zug 6300, Switzerland*
*3 - **BioSolutions GmbH,** Weinbergweg 22, 06120 Halle/Saale, Germany*
*4 - **Interdisciplinary Research Center HALOmem,** Charles Tanford Protein Center, Martin Luther University Halle-Wittenberg, Kurt-Mothes-Str. 3a, 06120 Halle/Saale, Germany*
*5 - **Biozentrum,** Martin Luther University Halle-Wittenberg, Weinbergweg 22, 06120 Halle/Saale, Germany*
*6 - **Department of Structural Biology,** Stanford University, Stanford, CA 94305*

***Correspondence to Panagiotis L. Kastritis:** Institute of Biochemistry and Biotechnology, Martin Luther University Halle-Wittenberg, Kurt-Mothes-Str. 3, 06120 Halle/Saale, Germany. panagiotis.kastritis@bct.uni-halle.de (P.L. Kastritis), @3dstructure (P.L. Kastritis)*
https://doi.org/10.1016/j.jmb.2022.167637
***Edited by Anna Panchenko***

## Abstract

After two years since the outbreak, the COVID-19 pandemic remains a global public health emergency. SARS-CoV-2 variants with substitutions on the spike (S) protein emerge increasing the risk of immune evasion and cross-species transmission. Here, we analyzed the evolution of the S protein as recorded in 276,712 samples collected before the start of vaccination efforts. Our analysis shows that most variants destabilize the S protein trimer, increase its conformational heterogeneity and improve the odds of the recognition by the host cell receptor. Most frequent substitutions promote overall hydrophobicity by replacing charged amino acids, reducing stabilizing local interactions in the unbound S protein trimer. Moreover, our results identify "forbidden" regions that rarely show any sequence variation, and which are related to conformational changes occurring upon fusion. These results are significant for understanding the structure and function of SARS-CoV-2 related proteins which is a critical step in vaccine development and epidemiological surveillance.



## Introduction

The COVID-19 pandemic triggered a worldwide health crisis, claiming over 3.3 million lives within 16 months and causing substantial damage to global economy (https://www.imf.org/en/Publications/WEO/Issues/2021/03/23/world-economic-outlook-april-2021). Although preventive methods such as social distancing, face covering, testing, and tracing can mitigate spread of the virus, only the achievement of herd immunity through vaccination can put an end to the pandemic in a timely way.[1,2] Development of COVID-19 vaccines started as early as January 2020, enabled by rapid breakthroughs in genome sequencing and structural biology, leading to the first mass vaccination programs starting in early December 2020 (https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-(covid-19)-vaccines). The majority of the currently

available vaccines target the Spike protein (S protein) (https://www.nytimes.com/interactive/2020/science/coronavirus-vaccine-tracker.html), which mediates entry into the host cell[3] through diverse mechanisms.[4] Although SARS-CoV-2 sequence diversity is very low,[5] natural selection can lead to the appearance of favorable S protein mutations that confer viral fitness advantages and pose a threat to vaccine effectiveness.[6]

The S protein is a highly glycosylated homotrimeric transmembrane protein[7] that enables binding and fusion of the virus with the host cell.[4] Each S protein monomer consists of nearly 1300 residues divided in two furin-cleavable subunits, S1 and S2.[8] Additionally, SARS-CoV-2 S possesses a second cleavage site - S2′ - for the TMPRSS2 protease, which is needed to mediate membrane fusion.[9] The S1 subunit is responsible for binding to the host cell receptor, the angiotensin converting enzyme 2 (ACE2),[10–12] a zinc-dependent peptidyl dipeptide hydrolase.[13] The recognition happens via the receptor binding motif (RBM),[14] which is part of the receptor binding domain (RBD).[15] In its metastable prefusion state, the S protein exists predominantly in two distinct conformations, "closed" and "open", that have been structurally resolved through cryo-EM.[16] In the "closed" conformation, the S protein cannot bind to the ACE2 receptor. Thus, to enable the binding, S1 must undergo conformational changes that expose the RBD and bring the protein into the "open" conformation. The S2 is responsible for the fusion of the viral and host cell membranes, a mechanism underlined by a complex series of events.[17] In summary, the S protein evolved to compartmentalize functional aspects across the annotated structural domains, including the receptor binding domain, cleavage sites and fusion-related regions (Figure 1(a-e)).[18].

Slight changes in the ACE2 sequence occurring within humans[19] or across species,[20] can have influence not only on disease severity, but also on viral infectivity: large-scale structural modeling efforts targeting ACE2 variants aim to illuminate slight changes in ACE2-S protein interface and comprehend critical interface residues and biophysical properties involved in the recognition.[21–24] Obviously, not only variations of the ACE2 protein, the cell receptor, but also variations across SARS-CoV-2 proteins, especially the S protein,[25–28] can have a vast effect on viral infectivity[6,25,27] and severity.[29,30] Although ongoing efforts are connecting structure–function relationships of S protein variants as compared to the Wuhan strain,[26,31–35] a systematic, large-scale statistical, biophysical, and structural analysis of S protein variants is missing. This is further complicated by the different conformations that S protein acquires before, during and after recognition by the ACE2 receptor.

Here, we systematically analyzed all S protein substitutions at the amino acid residue level. S protein variants were extracted from SARS-CoV-2 sequencing projects deposited in the public database GISAID[36] (Global Initiative on Sharing All Influenza Data) until the 2nd of January 2021. The end date was chosen to select substitutions that occurred before ongoing vaccination campaigns that started in December 2020. Our results show "forbidden" regions for residue substitutions and a common biophysical denominator for their selection, underlined by a "steered" conformational selection mechanism for subsequent biomolecular recognition.[37,38] In this type of recognition, unbound structural ensembles shift their energy distributions towards less favorable energies upon an external trigger to achieve and facilitate protein–protein recognition, lowering energy barriers.[38] Our results have a direct impact on understanding S protein variation as well as on current and future efforts for deriving vaccines and antiviral therapeutics.

## Results

### Three distinct sequence regions within S2 are strictly conserved.

We downloaded all 311,255 SARS-CoV-2 S protein sequences deposited in the public database GISAID before the 2nd of January 2021 and considered those with more than 95% sequence coverage ($N$ = 276,712). These sequences account for a total of 505,403 amino acid changes, of which nearly half ($N$ = 257,552) corresponded to the D614G "mutation", a substitution that became dominant after July 2020.[6] For each substitution present in the dataset, we calculated its percentage of occurrence (*p.occ.*) considering all possible substitutions ($N$ = 24,187) (Figure S1). This analysis revealed that almost every residue across the S protein has been substituted at least once (Figure S1), except for 44 specific residues (Figure 1(e), Figure 2(a), Figure S2(a–d)). These 44 residues are not randomly distributed across the SARS-CoV-2 S protein sequence, but cluster in 3 specific regions in the S2 subunit: (a) at the interval between furin and S2′ cleavage sites, the Upstream Helix domain (UH), (b) within the Heptad Repeat 1 (HR1) and Central Helix (CH) domains and (c) within the Connector Domain (CD), implying high conservation (Figure 1(e), Figure S2(a-d)). Residues C738, S746, C749, G757, F759, and F800 do not exhibit variation and are all localized within the UH, which shields the postfusion helical trimer.[39] Interestingly, the role of F800 is unclear and is not resolved in the postfusion structure.[39] HR1 and Heptad Repeat 2 (HR2) mediate membrane fusion[39] and are known to be conserved within SARS-CoV-2 lineages and other

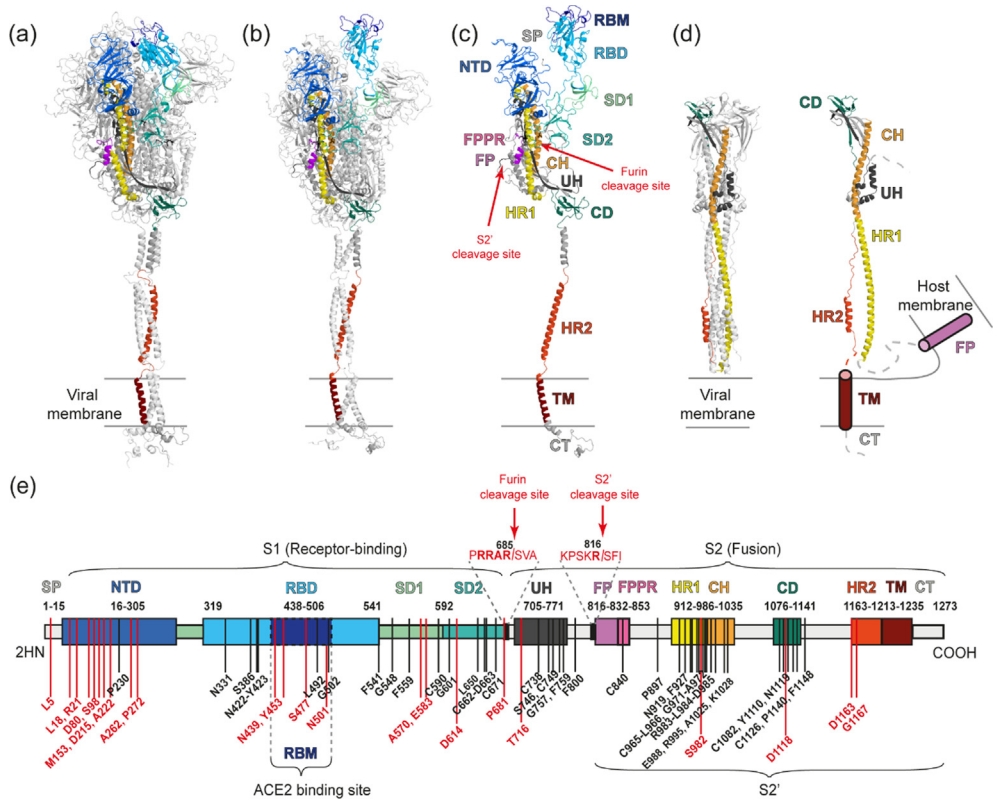



**Figure 1. Overview of the SARS-CoV-2 S protein and identified substitution frequencies for S protein variants.** (*a*) Prefusion conformation closed state. (*b*) Prefusion conformation open state. (*c*) Single chain from the prefusion conformation with an open RBD domain. (*d*) Postfusion conformation. All domain and cleavage sites are indicated in panel e. (*e*) S protein topology diagram with positioning of residues substituted with frequency $\geq$ 0.5% (red) and highly conserved residues which have never been mutated (black) on Spike protein domains. SP: signal peptide; NTD: N-terminal domain; RBD: receptor-binding domain; RBM: receptor-binding motif; SD1: subdomain 1; SD2: subdomain 2; UH: upstream helix; FP: fusion-peptide; FPPR: fusion-peptide proximal region; HR1: heptad repeat 1; CH: central helix region; CD: connector domain; HR2: heptad repeat 2; TM: transmembrane region; CT: C-terminal domain. Furin cleavage site (cleavage between S1 and S2 subunits) and S′ cleavage site are indicated. Numbers correspond to amino acid residues in the protein sequence.

CoVs.[18,40] However, we clearly observe that HR2 is undergoing selection pressure, indicated by the absence of fully conserved residues (Figure 1(e)). This contrasts with the HR1 domain, where various residues were not observed to be substituted (Figure 1(e), Figure S2(a)). The adjacent CH domain has 4 residues that were never observed to be substituted, *i.e.,* E988, R995, A1025, K1028. These residues are embedded in the interface of CH and UH, in very close proximity to the UH conserved residues and co-localize at a specific lateral position across the postfusion conformation (Figure S2(b)). Lastly, CD also includes residues that remain conserved, which are again found near UH and CH, pointing to a critical stabilizing interaction located at the inter-chain interface (Figure S2(b)). Interestingly, C1082 and C1126, both present in the CD, form a cysteine bridge and were never observed to be substituted (Figure S2(b–d), Figure S3(a)). Rare to none substitution events occur for the rest of the cysteine bridges (Figure S3(a)). Glycosylation sites are also very conserved, with the exception of T323 exhibiting slightly higher frequency of substitution (Figure S3(b)).

Functionally, the localization of the UH, HR1/CH and CD regions also correlates with crucial conformation changes connected to (a) the opening of the RDB; and (b) the prefusion to postfusion extensive reorganization of the S2. For (a), molecular dynamics simulations of the closed and open S protein conformations showed that in the prefusion state appears a cavity which is formed by HR1/CH and CD regions.[41] This cavity formed by domains of the S2 subunit was observed

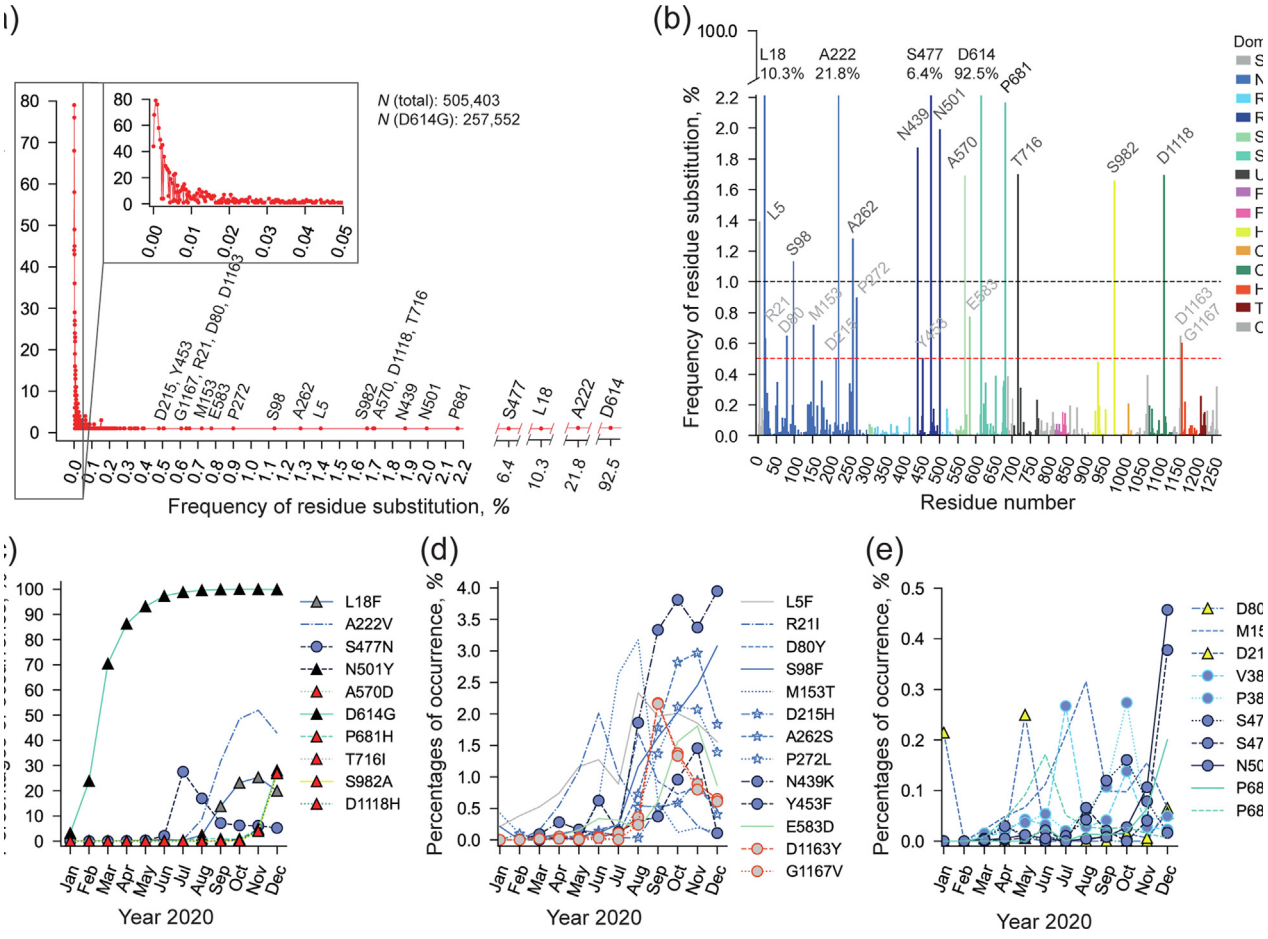



**Figure 2. Statistical analysis of substitutions in the S protein up to December 2020.** *(a)* Number of residues with corresponding frequency of residue substitution. For the calculation, all S protein genomes available for the year 2020 were considered. Residues with substitution frequency $\geq$ 0.5% are labelled. *(b)* Distribution of S protein substitutions across domains. Residues with substitution frequency $\geq$ 0.5% are labelled. *(c–e)* Appearance and development of individual residue substitutions with total substitution frequency $\geq$ 0.5% tracked over the year 2020. *(c)* Substitutions which at some point reached percentage of occurrence (*p.occ.*) $\geq$ 25%, *(d)* 0.5% $\leq$ *p.occ.* $\leq$ 4.0%, and *(e) p.occ.* $\leq$ 0.5%. Individual substitutions which belong to one of the lineages (B.1.1.7: red, B.1.351: yellow) or to multiple lineages (B.1.1.7; B.1.351; P1: black) are depicted as triangles. The remaining substitutions (circles) are color-coded according to their specific location (blue: RBD domain; grey: S2 subunit) or are present in the rest of the S1 subunit.

to be more rigid than regions of the S1 subunit[41] and is suggested to allosterically modulate the opening of the RBD domain through a "bouncing spring" mechanism.[41] For (b), conserved residues are initially in proximity to each other in a condensed coiled coil formed by CH-HR1, but after fusion, the coiled-coil is extended (Figure S2(b)). This is observed also in other coronaviruses,[42] which implies active role of the conserved residues in the conformational rearrangement. These findings are now further corroborated by our conservation analysis, particularly since mutation rates of RNA viruses are up to a million times higher than that of their hosts.[43] Since the virus can escape neutralizing antibodies by just a single amino acid substitution,[44,45] identifying protein regions with extensive, localized conservation is extremely important for designing drugs against SARS-CoV-2 and for vaccine development.

## Identification and characterization of frequent S protein variants and residing substitutions

Although these highly conserved sites are critical for drug design, frequent substitutions are also equally important since they enable viral immune and drug escape[46,47] or increase viral fitness. Viral genetic diversity has to be steadily monitored to swiftly adapt vaccine and therapeutic drug development to emerging SARS-CoV-2 variants which appear through accumulation of nucleotide variations.[48] Our frequency analysis shows that most of the substitutions are not persisting, having very low frequency of residue substitution, most being below 0.5% (Figure 2(a)), probably partly due to CoVs proofreading function,[49] and partly because many of the substitutions do not confer substantial fitness advantages and therefore do not persist.

Only 23 residues are substituted with frequency $\geq$ 0.5% and are annotated in Figure 2(a, b). Interestingly, 18 of these residues are located on the S1 subunit, formed by residues M1-R685 with 4 residues (N439, Y453, S477, N501) located in the receptor binding motif (RBM) of the RBD (Figure 1(e), Figure 2(b)). We also observe residue clustering ($N$ = 9) in the NTD (Figure 1(e), Figure 2(b)). One of the mutations of interest, which is placed at the NTD, is A222V. This mutation first appeared in July 2020[50] and was present in over half of the sequences by November 2020 (Figure 2(c)). Despite becoming widespread, it remains mysterious for its structural or functional impact, although 2 reports suggested that it does not affect antigenicity[51] and has minimal impact on viral entry.[50,51] The high frequency substitutions present in S1 should affect the unbound and bound states of the S protein and not the postfusion state. Residues with frequency of substitution $\geq$ 0.5% ($N$ = 23) are not only participating in characterized substitutions across lineages, but also in variants not extensively described to date, which comprise a large portion of the retrieved data (Figure 2(c–e), Figure S3(c), Figure S4). Multiple substitutions per residue are also observed (Figure S3(c), Table S1). Many identified substitutions influence monoclonal antibody (mAb) escape or alter binding affinity for ACE2, especially those located on the RBD and include:

*(a) S477N*. S477 is located within a flexible loop (residues A475-G485)[52] and exhibits the highest local flexibility.[52] Most often S477 substitutes to asparagine (S477N)[50] and its relative population peaked in July 2020 (*p.occ.* ~27.5%), remaining still high (*p.occ.* ~5.2%) by the end of 2020 (Figure 2(c)). This substitution was shown to strengthen binding of S protein with the human ACE2 receptor[52,53] and to be resistant against multiple mAbs.[54] Additionally, S477 was observed to mutate to arginine (S477R) and isoleucine (S477I) (Figure 2(e)). As such, S477 substitutions may promote viral transmission[55] and possibly increase affinity to ACE2,[56,57] although the biophysical basis of these effects is unknown and warrants further structural characterization.

*(b) Y453F* and *(c) N439K*. Y453F[58] is known as "cluster five", which arose among farmed minks in Denmark and had its frequency peak in November 2020 (*p.occ.* ~1.4%, Figure 2(d)). Y453F is located in the interaction interface with ACE2 and increases binding affinity by 4-fold but does not alter inhibition potency in convalescent sera.[53,58] Both Y453F and N439K may potentially be escape substitutions.[59]

*(d) N501Y*. The most prominent substitution of N501 is N501Y (Figure 2(c)), appearing in the B.1.1.7, B.1.351 and P.1 lineages. This substitution strongly increases binding affinity to ACE2,[53] *e.g.*, for B.1.1.7, a 7-fold affinity increase was measured.[33] This effect can be explained via a destabilization effect caused by N501Y, increasing S protein open state population.[60] Another frequent substitution is N501T (Figure 2(e)) which was identified as one of the dominating mutations in minks in the USA.[61]

## Hydrophobicity as a driving force for S protein variation at the primary structure level

We hypothesized that variation in amino acid substitutions across the S protein sequence may be driven by certain physical–chemical properties, since certain physical–chemical rules exist for rationalizing binding affinity of protein–protein interactions.[62,63] To discover possible global effects, we first compared hydrophobicity between the Wuhan strain and 148 amino acid substitutions with a spectrum of measured frequencies. These substitutions were selected based on their Jan-Dec 2020 time series profiles as shown in Figure 2(c–e) and Figure S4. The overall distribution of hydrophobicity across the S protein shows 3 prominent sets, at low, near-neutral and high hydrophobicity values (Figure S5(a)). These 3 distinct sets are recapitulated for a subset of the residues from the Wuhan strain that corresponds to the wild-type residues for the selected 148 substitutions (Figure 3(a)). In short, an obvious shift towards hydrophobicity is detectable upon mutation (Figure 3(a–c)). The shift is underlined by a disappearance of set 1, decrease of set 2, prominent increase of set 3 (Figure 3(a–c)) and is substantial when either all 148 selected residues are considered (Figure 3(a)) or relatively low frequency substitutions (*p.occ.* < 0.5%) (Figure 3(b)). For the group of high frequency substitutions (*p.occ.* $\geq$ 0.5%) a similar trend is observed, with the prominent appearance of set 3, but shift was not substantial (Figure 3(c)). Certain percentages are calculated for the S protein sequence of the Wuhan strain when considering the type of amino acid (Figure S5(b)). Percentages are also derived in a similar range for the Wuhan strain residues of the 148 selected residues and their substitututions considered in the different groups (Figure 3(d–f)). Their mutated equivalents exhibit a substantial increase of non-polar amino acids, reducing the proportion of charged and polar residues. This result holds true for both low- and high-frequency groups of the substitutions (Figure 3(e–f)). Overall, there is a clear shift toward hydrophobicity for all considered groups of substitutions that points to a possible denominating biophysical effect of specific origin.

## Structure-based analysis of S protein variation reveals suboptimal scoring across conformational states for the Wuhan strain

Protein variation and selection is underlined by a plethora of factors,[64] some of which can be derived from structural data.[65] Therefore, considering the key observation of the global increase in hydrophobicity across the selected substitutions, we asked

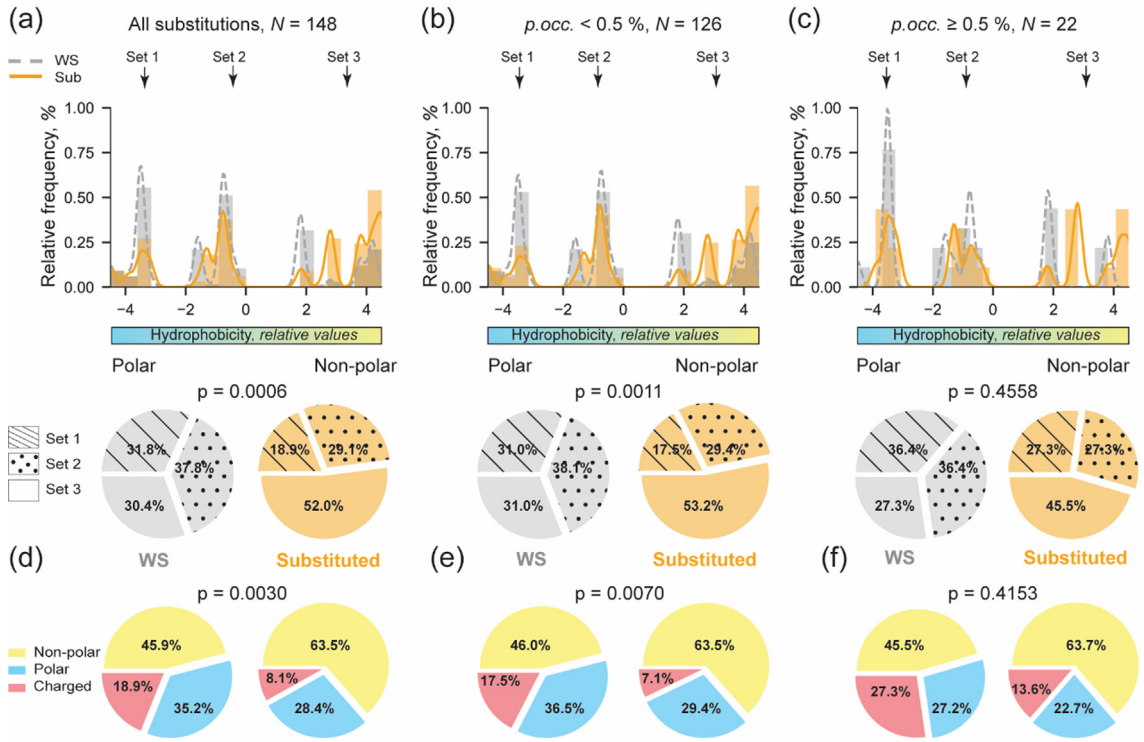



**Figure 3. Substitutions promote increased hydrophobicity at the primary sequence level.** (*a-c*) Distribution of sequence-based hydrophobicity scores for the Wuhan strain (WS) (grey) and all considered substitutions (orange) ($N$ = 148) (*a*), substitutions with *p.occ.* < 0.5% ($N$ = 126) (*b*), and *p.occ.* ≥ 0.5% ($N$ = 22) (*c*). Significance of the variation for each subset was tested via chi-square test comparing the Wuhan strain (grey) and substitution (orange). The *p-values* are depicted above of each subset. WS and Sub denotes Wuhan strain and substitution. (*d–f*) Distribution of nonpolar, polar, and charged residues for the subsets of residues (WS and Sub) corresponding to the plots directly above them *(a–c)*.

what these mutations could bring on a structural level. We constructed 475 structural models of S protein corresponding to all 148 substitutions in 4 known conformational states which are involved in the recognition, namely unbound ("closed"[66,67] ($N$ = 148), "open"[66,67] ($N$ = 148)), bound[68] ($N$ = 148) and postfusion[39] ($N$ = 31) (Figure 1(a-d)). As a control, the Wuhan strain S protein was included. We refined all structural models via short molecular dynamics simulations using the HADDOCK server[69] and HADDOCK scores, along with its components, were retrieved for all 4 known states and 5 lineages (B.1.1.7, B.1.325, P.1, B.1.617.2, B.1.1.529), whereas B.1.617.2, B.1.1.529 were modelled only in closed, open and bound state (Figure 4(a–d)). All scores are included in Tables S2–S6. Details on the performed calculations by HADDOCK as well as considered limitations of docking methods and their scoring functions are described in the Materials and Methods section.

Results show that distribution of scoring components calculated for the residing interfaces are relatively narrow, and most mutations only change score values by few relative units (Figure 4, Figure 5, Figure S6). This result shows that an introduced variation in the S protein sequence is not expected to majorly impact the S protein structure, but rather finely regulate interface non-covalent interactions related to HADDOCK scoring components. Interestingly, calculated scores of Wuhan strain S protein in four conformational states most of the times falls within the average values of the calculated distributions (Figure 4), but in two cases it considerably shifts towards lower or higher values (Figure 4(b, c)). One might argue that slight score shifts are close to noise, but, within these unit shifts, HADDOCK scoring can recapitulate binding affinities of protein–protein interactions within experimental inaccuracies.[62,63,70] This observation for the Wuhan strain S protein indicates that its initial sequence was suboptimal in terms of either stability or conformation if scores are to be interpreted as stability or affinity proxies; this means that additional mutations show better scores, overall stabilizing the unbound states, binding to the human ACE2 host receptor and also its fusion.

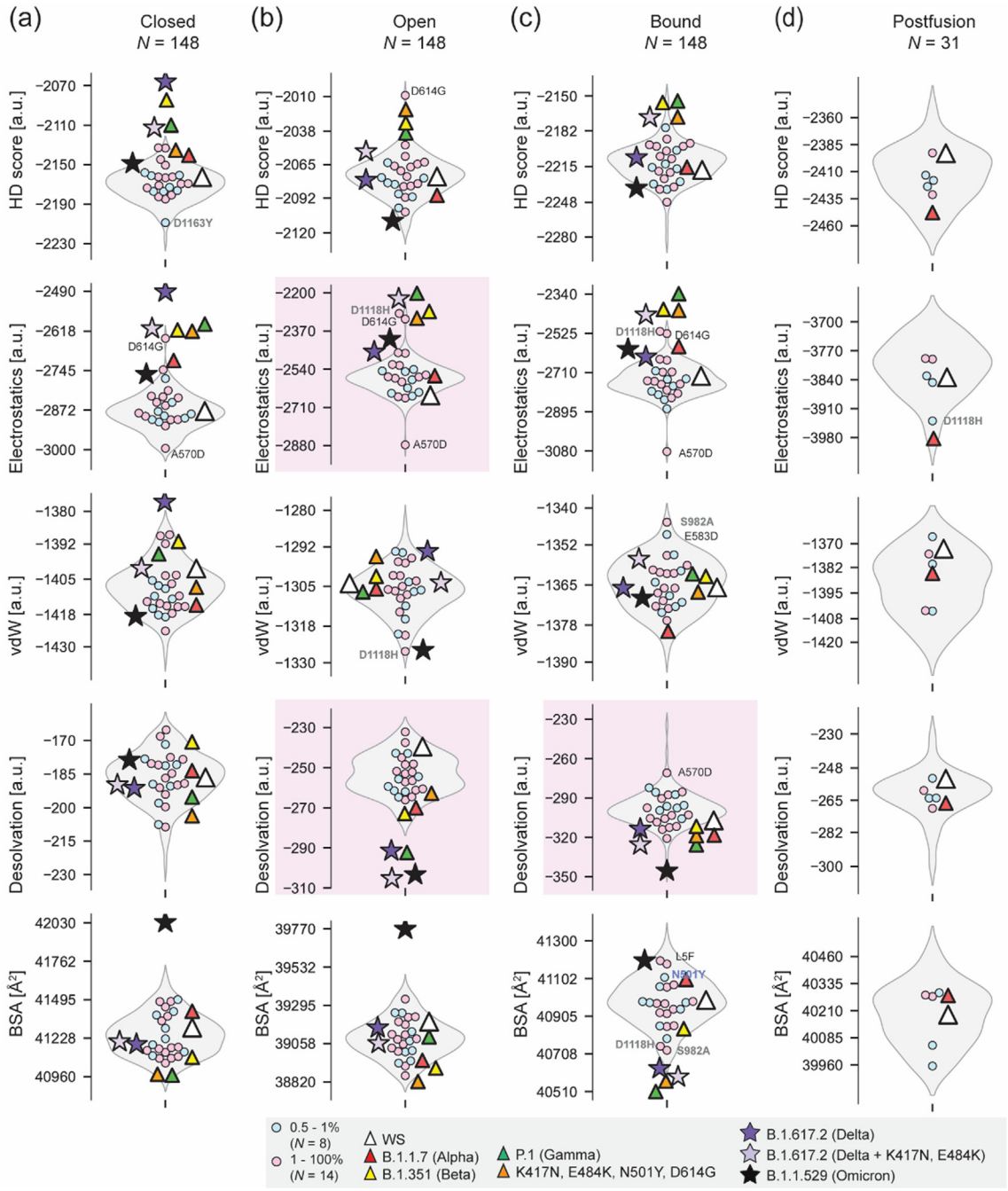



**Figure 4. Distributions of all calculated scores (HADDOCK and its components) for all considered substitutions (*N* = 148) in comparison to the Wuhan strain (WS) values (white triangle).** Violin plots represent scoring values distribution for (*a*) closed, (*b*) open, (*c*) bound states and (*d*) postfusion conformation (*N* = 31). Individual substitutions (circles) with the *p.occ.* $\leq$ 0.5% are colour coded (0.5% $\leq$ *p.occ.* < 1.0%: light blue; *p. occ.* $\geq$ 1.0%: pink). Wuhan strain (WS: white), considered lineages (B.1.1.7: red; B.1.351: yellow; P1: green), and combination of frequent variants (K417N, E484K, N501Y, D614G) are depicted as triangles. Lineages which appeared after start of vaccination efforts are depicted as stars (B.1.617.2: dark violet; B.1.617.2 (+K417N, E484K): violet; B.1.1.529: black). Plots with pink background highlight calculated distributions with the shift of average values towards lower or higher values.

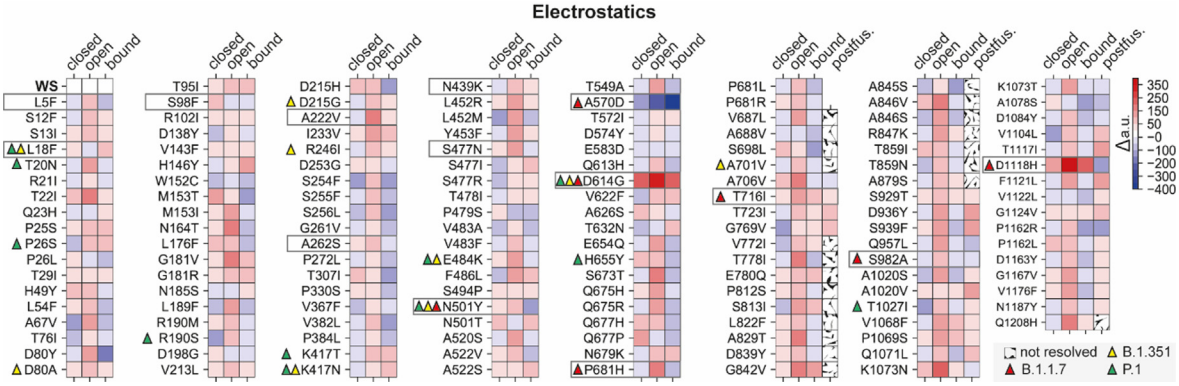



**Figure 5. Difference in electrostatics score between WS and corresponding substitutions across all S protein conformational states**. The change of electrostatics score compared to Wuhan strain (WS) for individual substitutions in closed, open, bound states and postfusion conformation is shown. Substitutions which appear in SARS-CoV-2 lineages (B.1.1.7: red; B.1.351: yellow; P1: green) are depicted as triangles and substitutions with *p. occ.* $\geq$ 0.5% are framed.

## Scoring of S protein unbound states with HADDOCK and Rosetta uncovers electrostatics as a robust scoring component

Our observations rely on the HADDOCK score and its components that are not absolute energetic values, but rather indicative of effects of different scoring components that each corresponds to a physical (*e.g.*, van der Waals, electrostatics, desolvation energy) or structural (*e.g.*, buried surface area: BSA (Å$^2$)) contribution. To corroborate our analysis, we implemented similar approach using Rosetta.[71] Energy scores for each HADDOCK refined model with single substitution were re-calculated with Rosetta and its own scoring function. As the scores from these two methods are in different scales, rather than comparing the absolute values 1-to-1 we can only compare if they both follow a similar trend (positive, negative or zero slope). A Bayesian-Ridge linear regression model is fitted for every pair of similar type of scores from the two methods. Correlation coefficient $R^2$ and *p-value* (for null hypothesis that the slope is zero, using Wald Test with t-distribution) are calculated and quoted in respective subplots (Figure S7). The results show the strongest correlation between the scores from the two methods for buried surface area (BSA, Å$^2$), followed by the electrostatics score for closed and open conformational states (Figure S7(a, b)). However, for ACE2-bound and postfusion states, the strongest correlation is seen for BSA and desolvation (Figure S7(c, d)). The results from this assignment indicate that both methods robustly capture the electrostatics, desolvation, and BSA scores in closed and open states corroborating that performed HADDOCK scoring is not biased towards its own scoring.

## The electrostatics scoring component of HADDOCK as a common denominator for S protein variation selection

To disentangle possible scoring components that may be important for the studied substitutions, structures were modelled by introducing corresponding substitutions, followed by a refinement procedure as detailed in the Materials and Methods. All models were ranked corresponding to their increasing probability of occurrence, and split into two groups, low- and high- frequency, using a variable percentage of occurrence (*p.occ.*) value. The *p.occ.* value was discretely changed between a range of 0.04% $\leq$ *p.occ.* $\leq$ 2.0% in incremental step size of 0.02%. At every step, the two groups thus formed, were compared with each other, and statistically treated by applying (a) a t-test (Figures S8, S9) and (b) Bayesian methods (Figure S10). For the Bayesian analysis the groups were formed at slightly broader step size of 0.1% in the range of 0.1% $\leq$ *p.occ.* $\leq$ 2.0%. Number of mutations per step is reported in Table S7. Interestingly, the HADDOCK score, which is a combination of reported non-covalent forces (van der Waals, desolvation, electrostatics) shows substantial decreased values with increased *p.occ.* thresholds towards lower *p-values*, specifically for unbound closed and open states (Figure S8). This behavior is derived from the electrostatics scoring component. Analysis via the t-test shows that electrostatics scores calculated for the closed and open states (Figure 6(a), Figure S8) exhibit substantial decrease and not those calculated for the bound (Figure S8) and postfusion states (Figure S9, Figure S11(a)). In addition, the difference in average electrostatics scoring

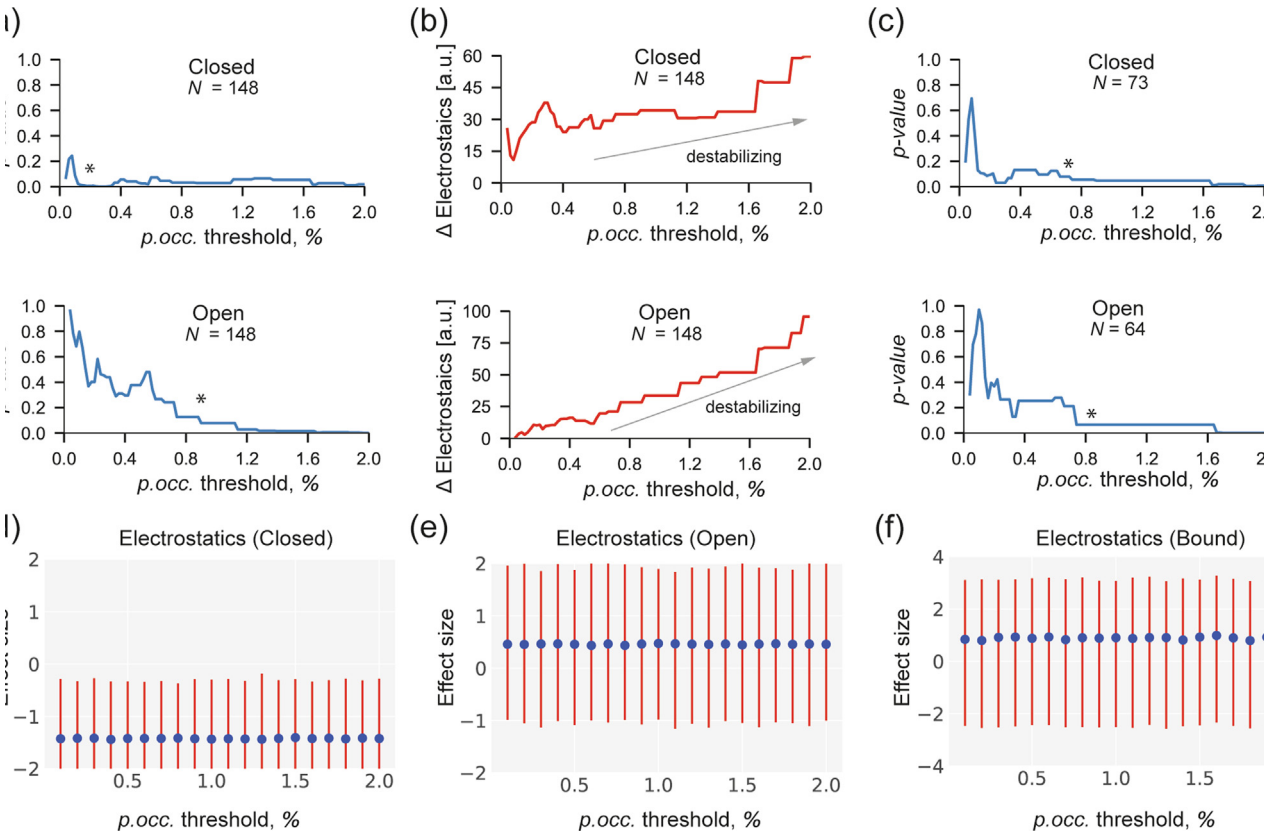



**Figure 6. Classical (a–c) and Bayesian (d–f) statistical analysis reveals importance of decreased electrostatics scores across *p.occ.* values, especially for the unbound S protein conformational states.** Here, the consistently significant contributions of destabilizing electrostatics govern increasingly frequent substitutions on closed *(a–d)* and open *(a–c, e)* S protein states, but not ACE2-bound *(f)*. (*a*) *p-value* distribution of all considered substitutions ($N = 148$) in closed and open conformations separated into low- and high- frequency groups as function of threshold ($0.04\% \leq p.occ. \leq 2.0\%$ with 0.02% step). (*b*) Electrostatics score differences between the groups showing that increasingly frequent substitutions have weaker electrostatics (*c*) *p-value* distribution of substitutions inside interfaces and rim regions of the trimeric S protein ($d_{C\alpha\text{-}C\alpha} \leq 15.0$ Å) in closed ($N = 73$) and open ($N = 64$) conformations separated into low- and high- frequency groups as function of threshold ($0.04\% \leq p.occ. \leq 2.0\%$ with 0.02% step). Asterisk in (*a–c*) panels indicate position starting from which the calculated *p-values* are indicative of a consistent trend (*p-value* < 0.1). (*d–f*) Calculated effect size (and 95% high probability density interval, HDI) using Bayesian parameter estimation for closed, open and bound states. The variants were separated into low- and high- frequency groups as function of threshold ($0.1\% \leq p.occ. \leq 2.0\%$ with 0.1% step). The farther away from 0 the effect size (and the 95% HDI) is, the higher is the effect. The effect sizes (and 95% HDI) for the electrostatics scores for only the closed conformation show significant differences.

components between consecutive groups is negative. This translates to higher electrostatics scores in the high-frequency groups, pointing to, possibly, a global, destabilizing effect underlying selection of the substitutions (Figure 6(b)). This effect is not observed when the bound state or the postfusion conformation are taken into account, although increasing destabilizing effects are observed for the bound state (Figure S11(b)).

Electrostatics scoring components may capture long-range charge interactions in a protein complex.[62,72] Therefore, we calculated all distances between Cα atoms of residues in the models and grouped the substitutions according to their inter-chain proximity. Results show that residue variation is not localized in the interface but is well-distributed across the S protein, namely inside interfaces and rim regions ($d_{C\alpha\text{-}C\alpha} \leq 15.0$ Å) and non-interacting surfaces ($d_{C\alpha\text{-}C\alpha} > 15.0$ Å), see number ($N$) of substitutions calculated per state (Figure S12-S13). Statistical analysis for each of the classes ($N_{interface}$, $N_{non\text{-}interface}$) as function of increased frequency (as performed before, *e.g.*, for Figure S10-S11) shows that substitutions proximal to the interfaces are involved ($N_{interface}$), (Figure 6(c), Figure S12), and not the residues localizing further (Figure S13). It is of note that, again, electrostatics is the single reliable scoring component for the group of residues proximal to the inter-trimeric S protein interface with increasing *p.occ.* thresholds (Figure 6(c)). To critically assess our observation from the above *p-value* trends, we applied a thorough Bayesian

parameter estimation method to calculate the effect size/s as described in Materials and Methods. Based on the Bayesian analysis, we see that in the closed state the electrostatics score predominantly shows a significant effect size irrespective of the *p.occ* value. In other words, electrostatics is a significant contributor but only in the closed conformational state of the S protein (Figure 6(d–f)) and is calculated not to dependent on interface proximity (Figure S10).

Given the fact that in the high-frequency group (*p. occ.* $\geq$ 0.5%) only around a quarter of the residues (27%) (Figure 3(f)) are charged and the rest are either polar (27%) or non-polar (46%), the global result is striking. Electrostatics is the only consistent contributor that underlies the selection of higher frequency substitutions in a global manner. The difference in electrostatics scores between high and low frequency substitutions, although mild (Figure S14 and S15), shows a trend for higher electrostatics scores underlying high-frequency substitutions – especially D614G and D1118H. Higher electrostatics scores translate into a global destabilizing effect.

To compare the calculated scores of the models of S protein wild type (WT) and variants (Tables S8–S10) with the scores of experimentally resolved structures of unbound S protein in closed and open conformations and ACE2-bound conformational state we performed HADDOCK refinements of all available S protein structures of the relevant variants (see Materials and Methods section "Refinement of experimental structures"). To assess correlation, the scoring components for the models and experimental structures were plotted against each other (Figure S16(a–e)). Due to extremely small number of experimentally resolved structures of S protein variants, the number of datapoints is very limited but the highest correlation scores are observed for electrostatics score for ACE2-bound structures (Figure S16(d, e)) pointing out, once again, the importance of the electrostatics scoring component for this system.

## Lineage-related variation exhibits higher electrostatics contributions with destabilizing effects across S-protein conformational states

To further look into the scoring data, substitutions were categorized into two distinct classes, namely those with low (*p.occ.* < 1.0%) and high (*p. occ.* $\geq$ 1.0%) *p.occ*. The categorization was performed for 148 considered substitutions (Figure S14, S15, and S17) and lineages (B.1.1.7, B.1.351; P.1; a combination of substitutions: K417N, E484K, N501Y, D614G; B.1.617.2; B.1.617.2 with additional substitutions at the RBD: K417N, E484K; B.1.1.529) as well as for substitutions inside interfaces and rim regions ($d_{C\alpha\text{-}C\alpha} \leq 15.0$ Å) and lineages (Figure S17). Looking into the distributions of electrostatics for those substitutions (Figure 5, Figure 6(a-f)), a shift towards destabilizing electrostatics is derived as expected. Interestingly, variants that belong to specific lineages, such as B.1.1.7, B.1.351, P.1, B.1.617.2 and B.1.1.529 (Figure 5) have increasingly destabilizing electrostatic contributions as compared to Wuhan strain, not only for the unbound closed and open states, but also for the ACE2-bound S protein state. This implies a possibility that lineage-related variants could affect the bound state electrostatics by further destabilizing the S protein in addition to the global effects derived for destabilization of the unbound closed and open states. These effects could be important only in proximity to the S protein interfaces (Figure 6(c), Figure S12, S17). To quantify the change in the electrostatics scoring component as compared to the Wuhan strain, all 148 substitutions were considered. Changes for all substitutions show a mixed effect on the closed state, having both increased or decreased electrostatics scores (Figure 5). However, most substitutions show considerable and systematic decrease in the electrostatics score when the unbound state is considered; again, mixed effects are observed for bound and postfusion states (Figure 5). These results demonstrate that most of the destabilization effects should occur in the unbound closed state of the S protein, with contributions from both closed and open states.

## Examples of decreased electrostatic interactions within S protein inter-trimeric interfaces

To understand the structural effects of electrostatics in closed and open unbound states, substitutions were visualized after refinement and compared to the Wuhan strain S protein trimer (Figure 7(a–d)). As expected, our calculations directly show the loss of an intertrimeric interface salt bridge when D614 is mutated to a G (Figure 7 (a)). In addition, K854 forms a salt bridge with D614 of the neighboring chain.[39] This salt bridge is abolished through the most common D614G mutation which is known to increase infectivity of SARS-CoV-2.[6] K854, in the fusion peptide proximal region (FPPR), supports the closed conformation,[39] further indicating importance of the FPPR in regulating the opening and closing of the RBD.

Cryo-EM data have shown that such an effect leads to a destabilization of the closed state[73] as also shown by the presented calculations (Figure 6 (a, b, d)). In addition, calculations show a drastic effect on the electrostatics scoring component, being one of the most destabilizing substitutions together with D1118 (Figure 5). Interestingly, D1118H removes a negative charge from the side chain, therefore removing a formed salt bridge with R1091, and replaces it with a protonatable group (Figure 7(b)). The newly introduced H1118 still

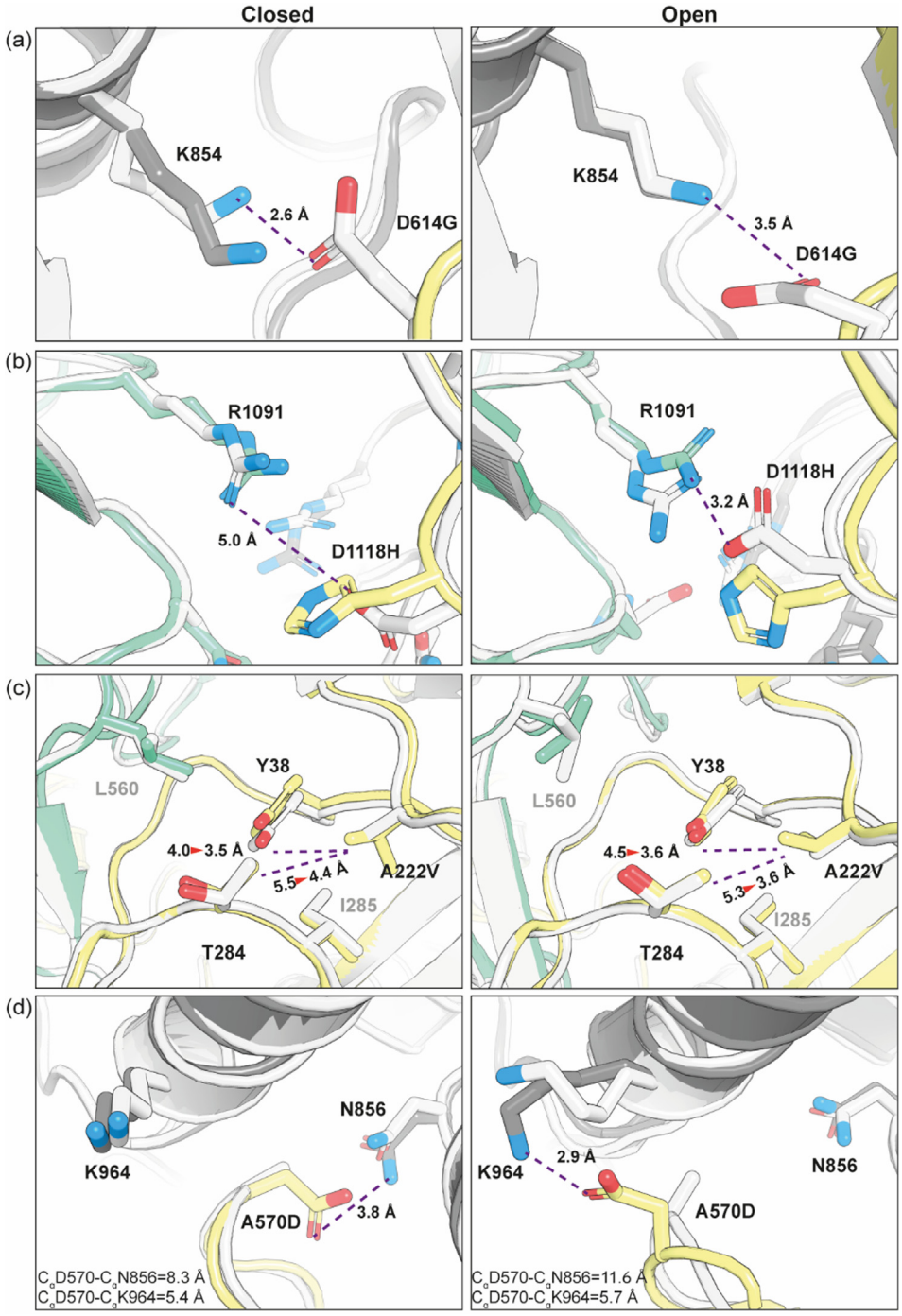



**Figure 7. Illustration of calculated effects of frequent (*p.occ.* > 0.5%) amino acid substitutions observed in S protein variants in the closed and open conformational states.** The Wuhan strain is colored white; each chain of the trimeric S protein with the introduced substitution (*a*) D614G; (*b*) D1118H; (*c*) A222V; (*d*) A570D is colored differently (yellow, grey, green). Distances between interacting atoms are drawn, and changes in distances are indicated with a red triangle.

interacts with R1091, but the polar interaction formed is weaker due to possible protonation effects of His and, therefore, introduction of similar charge that would lead to severe repulsion.

The A222V variant remains however, structurally unclear and subtle localized electrostatic effects are influenced. This is because when closed or open states of S protein are investigated, the localized interaction network of A222 does not considerably change. However, due to the marginally longer side-chain of V and its higher hydrophobicity, hydrophobic atoms across the localized region tend to come closer as indicated by the presented atom–atom distances of side chains which upon

mutation, overall, decrease (Figure 7(c)). This effect may well simulate a localized atom–atom hydrophobic attraction that is translated into decreased polarity, and therefore, electrostatics score in the corresponding calculations (Figure 5).

An interesting outlier is the A570D variant, which provides contradictory results as compared to all other for both open and closed state (Figure 5). Introduction of D570 is highly stabilizing, owing to a large conformational change of the Asp side chain, where D570 forms a salt bridge with N856, but in the open state forms another salt bridge with K964 (Figure 7(d)). K946 and N856 are 10.5 Å apart, indicating that the introduced mutation can alter interaction partners and contribute to localized flexibility and plasticity and, consequently conformational selection of either closed or open state.

To better understand these intriguing results concerning D570 pairing, we performed additional analysis of Cα-Cα distance distributions of both its interactions, namely D570-N856 and D570-K964 in closed and open conformational states, considering that the A570D substitution occurs in all three S protein chains (A, B, C). Analysis of the 20 final structures after the HADDOCK water refinement step shows that the Cα-Cα distance distribution for D570-N856 does not overlap in closed and open conformations (Figure 8(a)). For closed conformation it fluctuates between 7.5 and 9.0 Å, whereas in open conformation it spreads between 10.0 and 12.0 Å (Figure 8(a)). In the case of D570-K964, the distance distribution in closed (5.0–6.0 Å) and open (5.3–6.3 Å) conformations majorly overlap (Figure 8(b)). These observations indicate that the FPPR domain which contains the N856 residue is very flexible and is displaced upon opening of the RBD. Interestingly, each distribution for the open conformation, where RBD of chain A is turned upwards, is relatively narrow and has a clear maximum (Figure 8(a–b)), whereas all distributions of the closed state have many peaks. Therefore, we can assume, based on the calculations and derived distributions that there is a preferred, most populated state in the open conformation but not in the closed one.

Next, we performed molecular dynamics simulations of the A570D S protein variant in closed and open conformational states in

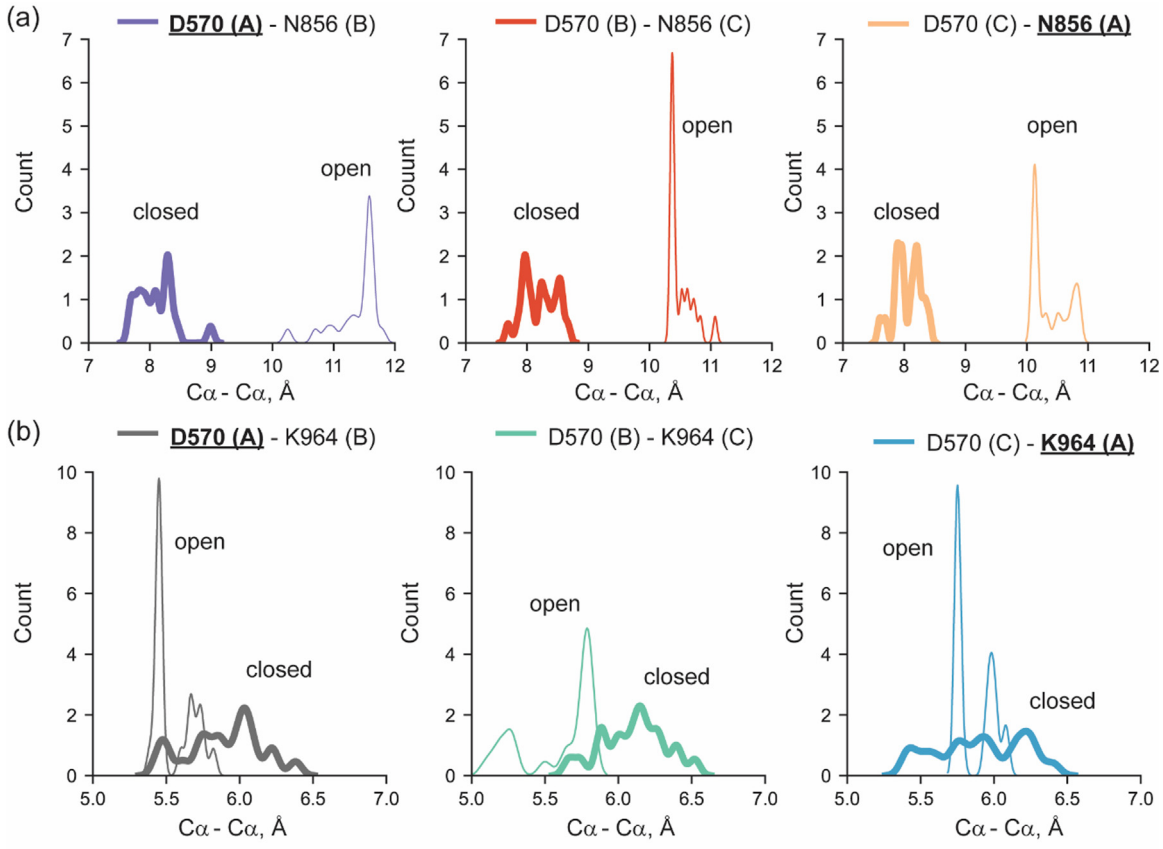



**Figure 8. Analysis of A570D-N856 and A570D-K964 interactions – statistical analysis of HADDOCK models.** *(a–b)* Distribution of Cα-Cα distances of (*a*) D570-N856 and (*b*) D570-K964 for S protein with A570D substitution in closed (thick line) and open (thin line) conformational states across all three chains (A, B, C). The distances were extracted from 20 final *.pdb* files after HADDOCK water refinement. The residue of the chain which RBD domain is turned upwards (chain A) is underlined and marked bold.

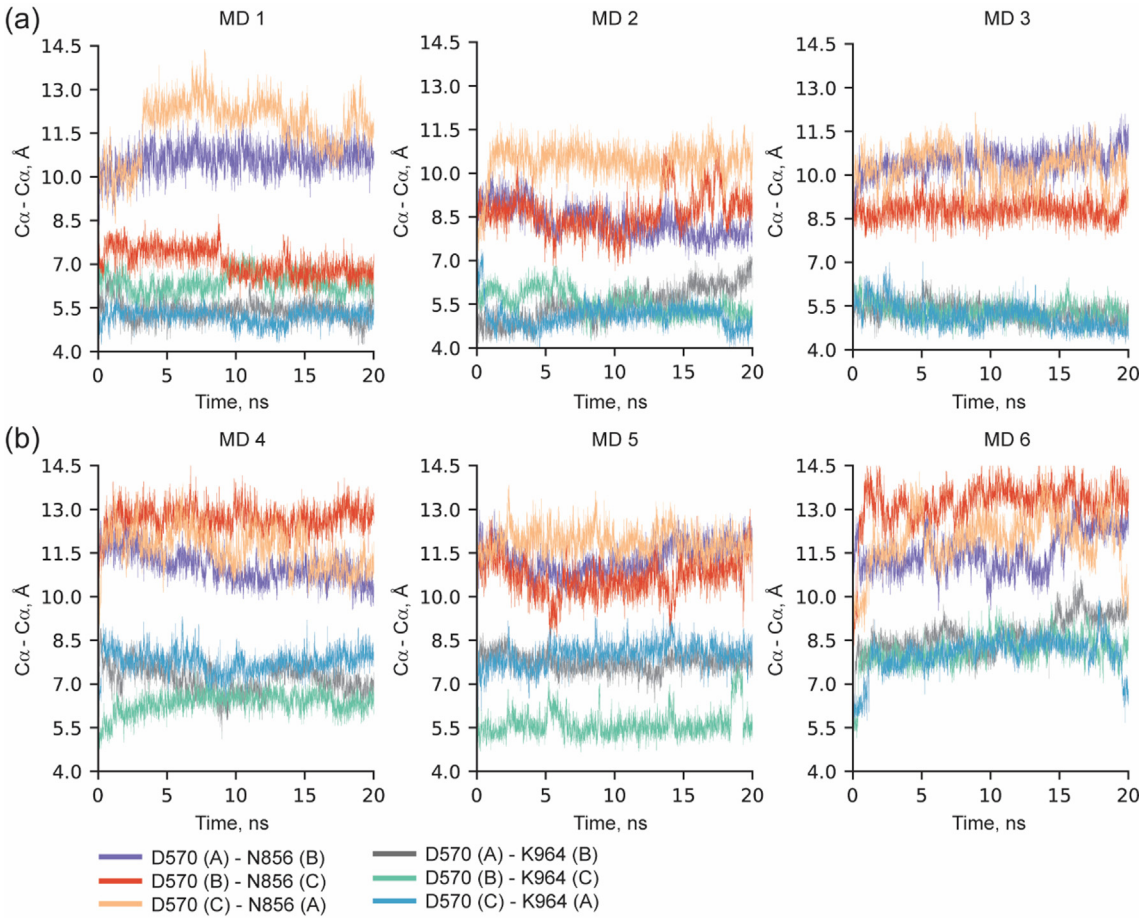



**Figure 9. Analysis of A570D-N856 and A570D-K964 interactions – *molecular dynamics simulations.*** Fluctuations of D570-N856 and D570-K964 Cα-Cα distances over the 20 ns of *triplicate* MD simulations for (*a*) closed (MD1, MD2, MD3) and (*b*) open (MD4, MD5, MD6) conformational states. Purple, red, and yellow line correspond to the D570-N856 pair per chain in the S trimer, and, correspondingly, grey, green, and cyan illustrate the distance fluctuations of the D570-K964 pair.

triplicates (Figure 9(a, b), Figure S18(a, b)). These MD simulations confirmed the observation that N856 is placed in a very flexible region, since the Cα-Cα distance between D570-N856 fluctuates between 6 and 14 Å. On the other hand, the distance between D570-K964 is stable in all simulations over 20 ns and fluctuates in a range between 4 and 7 Å. Despite the increased electrostatic component score that we have shown (Figure 5), the interaction network of D570 changes as function of S protein state, indicating an additional contributor for supporting or inhibiting opening of the RBD.

## Discussion

In this work we analyzed all SARS-CoV-2 genomes deposited in GISAID[36] up to January 2021. This date is critical to investigate only substitutions representing vaccine-unhindered virus adaptation. The S protein sequence was specifically investigated and its naturally occurring sequence variation. We have first observed that there are “hot” and “cold” spots for variation and showed that 44 positions across the S protein sequence have never undergone selection pressure during this timeline. Such “cold” spots, which are independent of viral adaptation, are mainly clustered in 3 regions (UH, HR1/CH and CD) of the S2 subunit and are correlated to important structure-based regions for biomolecular folding, function, and recognition. On the contrary, the most frequently substituted residues, the “hot” spots, are mainly located in the S1 subunit and especially in the NTD region. Therefore, the “cold” regions are of uttermost importance for development of therapeutics which might cover a broader spectrum of SARS-CoV-2 variants, whereas monitoring of “hot” spots and their structural characterization is of uttermost importance for vaccine adaptation.

Overall, our primary structure-based analysis showed that the hydrophobicity is a driving force for a selection of low-frequency substitutions. The same trend for selection of more hydrophobic

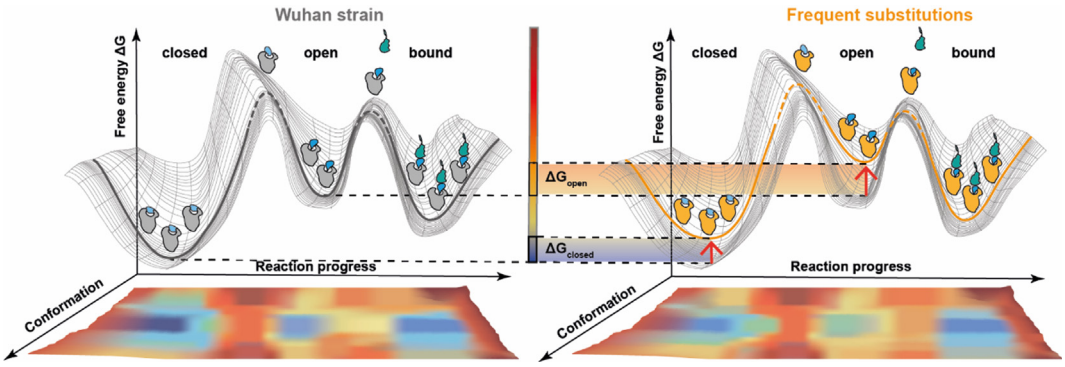



**Figure 10. Representation of the observed electrostatically-steered destabilization effects on the unbound states of S protein by frequent amino acid substitutions.** Effects of frequent substitutions (right) are shown to weaken the unbound closed and open state stability, forcing the S protein to explore a variability of conformational states that may be energetically more potent to overcome the energy barrier (dashed line) and reach the ACE2-bound state.

residues was observed for the high-frequency substitutions. This tendency may affect structural features of the S protein such as *e.g.*, promotion of the S protein RBD domain opening and subsequent binding to ACE2. To analyze influence of the substitutions on the S protein's physical–chemical properties, we constructed models of 148 substitutions as well as 3 lineages (B.1.1.7, B.1.325, P.1) in all 4 main conformational states of the protein and models of two additional lineages which appeared after start of mass-vaccination (B.1.617.2, B.1.1.529) in closed, open and bound states, and performed short molecular dynamics simulations utilizing the HADDOCK software. Thus, we observed that all high-frequency substitutions of S protein, which are also included in multiple SARS-CoV-2 lineages, destabilize the electrostatics of the unbound S protein, both in the closed, but also in the open state. We have concluded that a mechanistic model can explain this type of adaptation (Figure 10). In this model, the ground state of the closed S protein conformation is destabilized upon introduction of substitutions that would be subsequently selected; those that are not selected have variable effects. This destabilization majorly affects S protein interfaces in the open conformation. We hypothesize that such destabilization may raise the free energy of the unbound state, and therefore, a conformational ensemble that could recognize the ACE2 receptor is easier sampled. This mechanism is a reminiscent of a "steered" conformational selection mechanism for biomolecular recognition.[37,38] It is rather surprising that the "steered" conformational selection is majorly and globally underlined by electrostatics and is not strongly influenced by any other type of non-covalent interactions.

Globally, when looking into the structural data in PDB, just few structurally resolved S protein variants are reported; most of the variants included in our study are not experimentally structurally determined (compare *N = 476* vs few experimental data in Figure S14). Moreover, according to number of structurally resolved variants in closed and open conformational states (Tables S8-S10), it becomes clear that the S protein variants are appearing mostly in the open (*N* = 49) than in the closed (*N* = 21) state, whereas the number of Wuhan strain structures in closed (*N* = 20) conformation is exceeding the number of those in the open (*N* = 17). Of course, biochemical manipulation of the constructs, image processing or crystal lattice preferences complicates direct correlation to our observations, but such data groups possibly point to a preference towards an open state in determined variants.

Furthermore, the S protein has not yet been fully resolved at high resolution due to the sheer flexibility and complexity of its structure. Most structures in PDB do not contain the *C-ter* domains as well as flexible loops of various protein domains. In addition, most experimentally resolved structures contain additional mutations which are stabilizing S protein prefusion conformation. A lot of these limitations are minimized in computational studies of the S protein variants, which makes computational studies where dozens of variants are studied in large scale very valuable for the scientific community.

Our conformational selection model for SARS-CoV-2 S protein adaptation (Figure 10) now rationalizes multiple biochemical and structure-based observations from x-ray and cryo-EM structures. Conformational changes have been observed in the unbound states of the S protein when a variant is introduced (*e.g.*, for D614G,[31,32,73,74] N501Y[27] and lineages B.1.1.7, B.1.351, B.1.1.28[28]). Although it is arguable that the RBD up/down conformations in the structures that have been solved may be an artifact of the introduced mutations to study the Spike with struc-

tural methods, cryo-electron tomography data showed that the different Spike conformations that we studied here, resolved by single-particle cryo-EM,[66–68] are relevant in the context of intact SARS-CoV-2 viruses.[75,76]

All above-mentioned structural data collectively suggest that introduced mutations may increase the representation of open S protein states, in agreement with the mechanistic model that we present (Figure 10). Structural data for variants for which their bound states with ACE2 are available indicate intricate and opposing effects on interface stabilization, without pointing to a common denominator.[27,34,77] These differential contributors are also captured in our study (*e.g.*, shown in Figure 5, Figure S6), but our calculations suggest that S protein adaptation mainly occurs in the unbound and not the ACE2-bound ensembles, and, if the size effect is considered, only in the closed state. Conclusively, our mechanistic model on "steered" conformational selection for high-frequency substitutions provides an explanation of a global underlying denominator for SARS-CoV-2 adaptation and might be relevant to describe adaptations of other protein complexes with implications to biotechnology, structure-based design, and deeper understanding of molecular-level biophysical adaptation.

## Materials and Methods

### Models of closed, open, ACE2 bound states and postfusion conformation

Initial structures for prefusion S protein conformation in closed and open states were downloaded from the CHARMM-GUI[78] COVID-19 Archive.[79] Out of eight existing structural models for each of the states, which were built as described by Woo *et al.*,[80] the "1_1_1" models were chosen. As an initial structure for the S protein postfusion conformation, the partially resolved cryo-EM structure (PDB ID: 6XRA)[39] was considered.

***Closed and open states.*** Only ectodomains of the S protein without glycans were kept (residues M1-P1213), since substitutions occurring in the extracellular protein domains can diminish vaccine efficiency. For the closed and open states, the residues K825-V860 (FPPR region) were optimized with MODELLER[81] v.9.24 using the cryo-EM structure resolved at 2.90 Å global resolution (FSC = 0.143, PDB ID: 6XR8)[39] as a template, where this functionally important region was resolved for the first time. The reason for the additional modeling was the fact that the reconstruction was missing important structural bonds (C840-C851, K835-D848) which stabilize the FPPR region.[39]

***ACE2 bound state.*** The bound conformation was optimized using the prearranged open conformation. To maintain the RBD-ACE2 interface, the RBD domain of the only open chain (residues V320-A520) was remodeled with MODELLER v.9.24 using the x-ray structure of ACE2 receptor bound to RBD solved at 2.45 Å resolution (PDB ID: 6M0J)[14] as a template. The divalent metal ion zinc(2+) was kept in the catalytic center of ACE2 protein since the ion is placed in the vicinity of the interface (~20–25 Å) and may have influence on ACE2 folding.

***Postfusion conformation.*** The postfusion conformation was optimized using a cryo-EM structure of SARS-CoV-2 S2 in the postfusion conformation solved at 3.00 Å average resolution (PDB ID: 6XRA).[39] The partly resolved segments of each chain (residues A1174-E1202) were modelled using the x-ray structure solved at 2.90 Å resolution (PDB ID: 6LXT)[15] as a template. The final structure contains residues N703-I770 and T912-E1202.

***Refinement of experimental structures.*** The PDB was queried on the 8th of March 2022 and all experimentally resolved structures of S protein wild type and all relevant variants in closed ($N$ = 42), open ($N$ = 67), ACE-bound full S protein ($N$ = 38) as well as ACE2-RBD bound ($N$ = 23) states were downloaded (Tables S8–S10), prepared in accordance with HADDOCK submission requirements using pdb-tools[82] and subjected to HADDOCK refinement. Thereafter, if there were multiple structures of the same variant (Table S11), the average score was calculated per variant. For some conformational states variants are listed twice in the table (*e.g.*, WT) because some of the expressed proteins are containing S protein prefusion conformation stabilizing mutations (K986P, V987P)[83] or (F817P, A892P, A899P, A942P, K968P, V969P)[84] (Tables S8–S10).

***Docking, scoring and MD software strengths and limitations.*** One should be aware of molecular docking algorithms as well as molecular dynamics simulations strengths as well as weaknesses. Molecular docking is a combination of a conformational sampling algorithm and a scoring function.[85] Conformational sampling algorithms are developed and trained on limited number of samples in a test set, thus depending on the system of study and may vary in performance.[85] There are many different categories of sampling algorithms based on molecular dynamics (MD) simulations or stochastic methods such as Monte Carlo and genetic algorithms.[86] MD simulations are simulating atomic motions using simple approximations based on Newtonian physics.[87] The forces which arise from bonded and non-bonded interactions and are described in the force fields, which are limited by two principal challenges: force fields which require further refinements[88] as well as limitations in com-

putational power. It is known that, post-docking refinements provide higher hit rates and higher correlation with experimental data.[89] For example, for HADDDOCK it was shown that solvated docking/refinement do not always improve the docking results but improve the scoring,[90–92] which is highly important for our study. Water refinement procedure of models via HADDOCK is described in the following section.

***HADDOCK refinement.*** HADDOCK (High Ambiguity Driven protein–protein DOCKing) is a semiflexible docking method for biomolecular research which is based on bioinformatical predictions and experimental knowledge of biochemical/biophysical interactions. The HADDOCK approach is using python scripts derived from ARIA.[93] For structure calculations it uses CNS[94] (Crystallography and NMR system). The method can be used not only for docking of biomolecules but also for structure-refinement purposes.[69] HADDOCK uses nonbonded electrostatic and van der Waals energy terms of the OPLS force field with the cutoff distance of 8.5 Å from a modified version of the parallhdg5.22.pro parameter file[95] for the evaluation of inter- and intramolecular energies. The usual docking protocol consists of three following stages: a rigid-body energy minimization, three steps of a semi-flexible simulated annealing refinements in torsion angle space and a final refinement in Cartesian space with explicit solvent or DMSO (in our case water refinement was applied). The HADDOCK refinement interface, which was utilized in our study, includes only the last stage.[69] Here the protein is solvated in an 8 Å shell of TIP3P water molecules. At this stage all MD simulations are performed with 2 fs time step for the integration of the equation of motion. First, the system is heated to 300 K performing 100 steps of MD simulation at 100, 200 and 300 K keeping position restraints ($k_{pos}$ = 5 kcal $mol^{-1}$ $Å^{-2}$) on all atoms. Next, 1250 steps of an MD are performed at 300 K with position restraints ($k_{pos}$ = 1 kcal $mol^{-1}$ $A^{-2}$) only on heavy atoms. During the final cooling step, only the backbone atoms are restrained. At this step an MD simulation is performed at 300, 200 and 100 K with 500 MD steps at each temperature.[96]

The user can decide how many times the initial structure will be refined, in this work we run 20 refinements per model. After each of the refinement a HADDOCK score ($HS_{water}$) is derived as a weighted sum of van der Waals ($E_{vdW}$), electrostatics ($E_{elec}$) and desolvation ($E_{desolv}$) scoring components. The resulting value for each scoring component is calculated as an average for the top four best-scoring models with the smallest weighted sum value[69] (https://alcazar.science.uu.nl/services/HADDOCK2.2/).

***Rosetta calculations.*** For energetic calculations with Rosetta, the latest weekly release (Rosetta 2021.38)[71] was used. The best ranked HADDOCK refined model was used as input, taking only models of variants with single substitutions. Zinc entries in the.*pdb* files of ACE2-bound were modified according to Rosetta atom names. The altered. *pdb* file was scored according to the Rosetta tutorial "Analyzing Interface Quality" (https://www.rosettacommons.org/demos/latest/public/analyzing_interface_quality/README, accessed 10. March 2022).

To assess any correlation between the resulting scoring values from Rosetta and HADDOCK, scoring components of HADDOCK output (Electrostatics, vdW, Desolvation, BSA) and Rosetta scores (fa_elec, fa_atr, fa_sol, dSASA_int) are paired up (x,y) (Figure S7). For each pair (*e.g.*, an attribute from Rosetta vs that corresponding from HADDOCK) a Bayesian ridge linear regression model is fitted using Scikit-Learn library (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.BayesianRidge.html#). The $R^2$ score is calculated per pair (as quoted in the subplots) and is defined as (1-u/v), where 'u' is the residual sum of squares $\sum((y\_true - y\_pred)^2)$ and 'v' is the total sum of squares $\sum((y\_true - mean(y\_true))^2$, where, y_true is the actual data and y_pred is the linear regression model value. The best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). Moreover, the *p-value* (for null hypothesis that the slope is zero, using Wald Test with t-distribution) was calculated and quoted in the subplots (Figure S7).

## Analysis of SARS-CoV-2 S protein sequence variation and substitutions selection

For this work we queried the public database GISAID[36] for SARS-CoV-2 S protein sequences uploaded over period of time starting from July 2020 with the last query on the 2nd of January 2021. From the total number of 311,255 S protein sequences, sequences with more than 95% sequence coverage (276,712) were considered for the subsequent analysis. These sequences were aligned against the reference SARS-CoV-2 spike protein sequence (UniProt[97] P0DTC2) using MAFFT[98] v7.471. Frequency of mutations for each aligned site was calculated ignoring undefined residues (X) and gaps (–).

## Evaluation of relevant substitutions for structural analysis

In this work, percentage of occurrence (*p.occ.*) of all appearing substitutions was continuously monitored starting from July 2020 until the 2nd of January 2021 and substitutions appearing at the ectodomain of S protein were carefully selected (Table S12). Initially, in the dataset all substitutions appearing in July with *p.occ.* $\geq$ 0.2% (*e.g.,* T76I, M153I, V213L, S254F, V382L, T778I, T1117I, G1124V, V1176F) were included. After

July, considering the exponential increase in deposited data, substitutions appearing with *p.occ.* $\geq$ 0.5% at any considered month were immediately selected. Also, substitutions of the same residues *(e.g.,* G181R, N501T, A1020V) or of the residues next or in immediate vicinity from the residue with substitution *p.occ.* $\geq$ 0.5% (*e.g.,* S254F, T572I, G842V) were selected starting from the *p.occ.* $\geq$ 0.1%. Moreover, substitutions appearing at or near such important regions as RBD/RBM (R319-F541/S438-Q506) (*e.g.,* L452M, S477I, (PMID: 33270653: *e.g.*, T478I, P479S, V483A, F486L, S494P), T549A,), cleavage sites (Furin cleavage: P681-A688; S2′ cleavage: K812-I819) (*e.g.*, P681H, P681L, P681R, V687L, V688V, S698L, P812S, S813I; also, Q677P as it is very close to cleavage site and substitution is proline) were selected even if they appeared at any of the considered months with *p.occ.* $\sim$ 0.1%. Furthermore, substitutions which were observed to appear continuously over the period of at least three months with *p.occ.* $\geq$ 0.1% were selected (*e.g.,* S12F, P25S, P26L, T29I, T95I, H146Y, M153I, R190M, D198G, Q675R, V687L, V772I, R847K, T859I, A879S, P1069S, Q1071L, A1078S, V1104L, V1122L). Additionally, substitutions which periodically appeared with *p.occ.* $\geq$ 0.1% (T307I, A845S, T859I) and/or were described in the literature (T572I[99] and A706V[100] were also included. Also, we considered three lineages which are named according to Pangolin[101] nomenclature. Lineage B.1.1.7 (Alpha variant) and B.1.351 (Beta variant) were reported in December 2020 in United Kingdom (https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html) and in South Africa[102] respectively. In early January 2021, lineage P.1 (Gamma variant) was reported to appear in Manaus, Brazil.[103] All substitutions appearing in three considered lineages (B.1.1.7; B.1.351; P.1) were added to the dataset. Some of these substitutions were not observed in GISAID database (*e.g.,* T20N) or appeared with *p.occ.* $\ll$ 0.1% for any of the considered months (*e.g.,* D80A, R190S, R246I, K417T/N, E484K, T1027I) pointing that coverage of the S protein sequences in the database is not sufficient for identification of all relevant substitutions. Additionally, two variants which appeared after initiation of vaccination efforts were added to the dataset (B.1.617.2 (Delta variant)[104]; B.1.617.2 with two additional substitutions K417N, E484K (https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/); B.1.1.529 (Omicron variant: https://www.ecdc.europa.eu/en/publications-data/threat-assessment-brief-emergence-sars-cov-2-variant-b.1.1.529). Moreover, additional substitutions of the lineage-specific residues (*e.g.,* P681L, P681R) or of the residues next or in immediate vicinity from those (V483F, E654Q, T1117I, F1121L) were selected even if the substitution did not appear with *p.occ.* $\geq$ 0.1% over longer period of time. Finally, substitutions appearing with *p.occ.* $\geq$ 0.3% in November or December 2020 (*e.g.,* N185S, I233V, Q613H, K1073T, P1162R) were added to the dataset. The rest to complete the dataset (T22I, A67V, R102I, N1187Y, Q1208H) appeared to increase in frequency at the time of monitoring, for most of the cases reaching *p.occ.* $\geq$ 0.1%, during at least one of the considered months. Monthly *p.occ.* values calculated for each of the selected substitutions ($N$ = 148) considering sequences of the corresponding month are shown in Figure 2(c–e), Figure S4 and in detail in Table S12. For the rest of the performed statistics the frequency of residue substitution and the percentage of occurrence (*p.occ.*) for each of the substitutions was calculated considering all S protein sequences and are shown in Figure 2(a, b), Figure S1–S4.

## Introduction of substitutions and model refinement

Altogether 148 substitutions were selected and introduced as point mutations into all three chains of the trimeric S protein of the models of closed, open, ACE2 bound states ($N$ = 148) and postfusion conformation ($N$ = 31), altogether 475 structural models were created. In addition, the Wuhan strain "wild type" residue was also introduced as a control. Moreover, structural models for lineages B.1.1.7 (Δ H69/ Δ V70, Δ Y144, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H); B.1.351 (L18F, D80A, D215G, R246I, K417N, E484K, N501Y, D614G, A701V); P.1 (L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, D614G, H655Y, T1027I); model with mutations from RBD domain of B.1.351 lineage with D614G mutation (K417N, E484K, N501Y, D614G); B.1.617.2 (T19R, E156G, Δ 157/ Δ 158, L452R, T478K, D614G, P681R, D950N); B.1.617.2 with two additional substitutions K417N, E484K; B.1.1.529 (A67V, Δ H69/ Δ V70, T95I, G142D, Δ V143/ Δ Y144/ Δ Y145, Δ N211, L212I, ins214EPE, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493K, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K, L981F) were created. All created models were submitted to water refinement using HADDOCK webserver v2.2[69] as previously described.[62,92]

## Computational analysis of hydropathy

Sequence-based hydropathy calculations were performed using the Kyte-Doolittle hydrophobicity scale.[105] Distribution of the Kyte-Doolittle hydrophobicity scores (K-Dhs) was visualized for the complete ectodomain sequence (M1-P1213) of the Wuhan strain (Figure S5), for the "wild type" residues and corresponding substitutions for the set of all considered substitutions ($N$ = 148), substitu-

tions with *p.occ.* < 0.5% ($N$ = 126), and substitutions with *p.occ.* $\geq$ 0.5% ($N$ = 22) (Figure 3). To assess statistical significance of the observed shifts appeared for the substitution's distribution of each of the sets, chi-square test of independence of variables in a contingency table was performed (significance threshold: *p = 0.1*). The residues were divided into three sets: with low (K-Dhs < -2.3), near-neutral ($-2.3 \leq$ K-Dhs < 1) and high ($1 \leq$ K-Dhs) hydrophobicity values for the case of distribution according to K-Dhs (Figure 3(a–c)), or the amino acid polarity: non-polar (A, G, I, L, M, F, P, W, V), polar (N, Q, S, T, H, Y, C) and charged (R, D, E, K) (Figure 3(d–f)).

## Computational analysis of scoring components contributions

Scores (HD (HADDOCK) score, electrostatics, van der Waals, desolvation, BSA (buried surface area)) derived from structure-refinement calculations for each of the considered substitution ($N$ = 148), substitutions appearing inside interfaces and rim regions ($d_{C\alpha\text{-}C\alpha} \leq$ 15.0 Å) ($N_{closed}$ = 73, $N_{open}$ = 64, $N_{bound}$ = 80) and substitutions appearing at the non-interacting surfaces ($d_{C\alpha\text{-}C\alpha}$ > 15.0 Å) ($N_{closed}$ = 75, $N_{open}$ = 84, $N_{bound}$ = 68) were systematically classified in low- and high- frequency groups with variable thresholds (0.04% $\leq$ *p.occ.* $\leq$ 2.0% with 0.02% step, number of mutations per step is reported in Table S7). To identify statistically significant contribution of the calculated scoring components, *p-value* for each step was calculated via unpaired t-test (Figure S8, S9, Figure S12, S13).

To observe destabilization profiles of the electrostatics scoring component, the only significant contributor (Figure 6(a, c), Figure S8, S9), a mean scoring value for each of the groups (0.04% $\leq$ *p.occ.* $\leq$ 2.0% with 0.02% step) was calculated, and difference in the electrostatics scoring component ($\Delta$Electrostatics) between the group of substitutions with higher and lower *p.occ.* was derived for each step (Figure 6(b), Figure S11 (b)). Since *p-values* provide evidence for rejecting the underlying $H_0$, we consider groups of (a) strong (*p-value* < 0.01, *e.g.*, in the case of hydrophobicity), (b) moderate (*p-value < 0.05*), (c) weak evidence or trend (*p-value < 0.1*) and (d) no evidence (*p-value* $\geq$ 0.1).

## Bayesian parameter estimation

We use PyMC3,[106] a probabilistic programming package in Python, that fits Bayesian models using notably Markov chain Monte Carlo (MCMC) methods. To quantitatively assess how different any given two groups of data are from the other, we perform a rigorous Bayesian parameter estimation, using the module - Bayesian Estimation Supersedes the T-test (BEST) under PyMC3 based on Kruschke, 2013.[107] Driven by Bayesian probability, this is a comprehensive and more solid approach than the testing approaches that involve expressing a null hypothesis. Moreover, we estimate the uncertainty associated with the estimated parameter that accounts for our lack of knowledge of the model parameters. For a given (group 1, group 2) data, we calculate two parameters, namely, (a) the effect size, and (b) a high-density probability interval around the effect size. Farther away from 0 the effect size (and the 95% HDI) is, the better. Technical details of the procedure:

A student t-distribution is used to describe the attributes of each group. A t-distribution is a robust choice for our data, since it is less sensitive to outliers compared to a normal distribution. A t-distribution is represented by three parameters, mean ($\mu$), standard deviation ($\sigma$), and degrees-of-normality parameter ($\nu$) (which is assumed to be same for both groups). The prior distribution corresponding to each model parameter is set to be a very broad Uniform distribution, of width equal to 10 times the standard derivation. The prior distribution for parameter $\nu$ is assumed to be a very wide exponential distribution, of mean of 30. These wide prior distributions make the estimates of output posterior distributions less sensitive to the input prior distributions.

The posterior distributions of all parameters are estimated by the process of MCMC sampling. The MCMC process generates a large (up to 100,000) representative sample of credible parameter values that better represents the underlying posterior distribution. Note, the MCMC process generates sample of parameter values and not that of the actual data. For each credible parameter estimate ($\mu1$, $\mu2$, $\sigma1$, $\sigma2$) the effect size if computed as $(\mu1 - \mu2)/\sqrt{(\sigma1^2 + \sigma2^2)/2}$. A distribution of effect size (also of 100,000 samples) is computed along with a 95% credible interval, a High-Density probability Interval (HDI). If the means of two groups are not significantly different, then the effect size would tend towards 0. Therefore, a higher effect size indicates a significant difference in two groups. Unlike a single-point value of $p \leq 0.05$ in standard t-test, the interpretation of Bayesian estimation is not black-and-white, it uses an entire distribution of parameters for calculating the effect size. The conclusions are probabilistic in nature, and therefore, we observe if the estimated 95% HDI of the distribution of effect size does or does not include 0. Moreover, a Region of Practical Equivalence (ROPE) of $-0.1$ to 0.1 around the null value (0) is considered as 0, such that, the effect size indicates a significant difference in two groups only if the ROPE is completely outside the 95% HDI.

For every attribute (*e.g.*, HADDOCK score, HADDOCK components (Electrostatics score, vdW score, desolvation, BSA *etc.*)), the mutations

are divided in to two groups (*p.occ.*, %, below and above a threshold), while changing the threshold value incrementally from 0.1% to 2%. For each pair of groups, we calculate a distribution of effect size and plot the mean effect size along with the 95% HDI as a function of *p.occ.* threshold (%). The process is repeated for closed, open, and bound structures.

## Molecular dynamics simulations

The MD simulations of the S protein in "closed" and "open" conformational states were performed using NVIDIA CUDA version 9010 acceleration modules of NAMD 2.13[108] for Linux-x86_64-multicore-CUDA employing CHARMM36[109] force field. The long-range electrostatics were treated using the particle-mesh Ewald approach.[110] The S protein included M1-L1141 residues of each chain of HADDOCK refined model of A570D variant. The system was solvated in a water box of TIP3P water (with a water layer of 10 Å) and neutralized with 150 mM NaCl, eventually comprising 523,769 atoms in "closed" and 552,368 atoms in "open" conformation. System preparation as well as MD analysis were performed using VMD 1.9.3.[111]

Note that lipids and sugars were not included in the simulation because the MD simulations were performed using "hard-restrained" approach where only atoms of protein 15 Å around protein residues D570, N856 and K964 in all three chains and water-ion environment were allowed to move without harmonical restrains. On all other atoms of the system was applied harmonic force constant of 5 kcal $mol^{-1}$ $Å^{-2}$. Each approach was repeated three times for 10 ns per replicate after observing system equilibration. Prior to each of the MD simulations, two all-atom minimization-relaxation cycles were performed. At the beginning of each cycle, the system was minimized for 10,000 steps. During the first cycle, the water-ion environment and hydrogens were allowed to relax, keeping protein harmonically restrained. Subsequently, the temperature was incrementally changed from 0 to 310 K, relaxing the system for 1 ps per increment of 10 K with a final relaxation for 0.1 ns at 310 K. In the second cycle, side chains within 15 Å around (and including) D570, N856, K964 residues in all three S protein chains, as well all hydrogens, water molecules and ions were set free while all other atoms were kept restrained. The system was incrementally heated as described for the first cycle with a final relaxation for 0.1 ns at 310 K. Finally, the system with unrestricted atoms 15 Å around (and including) D570, N856, K964 residues in all three S protein chains was incrementally heated as previously described and the MD simulation was performed for 10 ns at 310 K, 1.01325 bar. During MD simulation an integration time step was set to 2 fs applying SHAKE algorithm on all bonds involving hydrogen atoms.

## Data availability section

Structure files and associated data of S protein variants in closed, open, ACE2-bound and postfusion conformational states generated in this work have been deposited in SBGrid.[112] Four directories are shared: "S_closed", "S_open", "S_ACE2" and "S_postfus". In each folder an initial *.pdb* file for each of the conformational states and two subdirectories ("param_index" and "structures") are placed. The "param_index" directory includes two files for each of the considered variants: the results file after the HADDOCK2.2 refinement[69] (*.html* file) and the parameter file that was used for the structure calculation (*.web*).

The user can reproduce any run by uploading the *.web* file using the HADDOCK2.2 webserver (https://alcazar.science.uu.nl/services/HADDOCK2.2/haddockserver-file.html).

The "structures" directory contains the top scoring refined structure file (*.pdb*) for each of the variants. The nomenclature of each file in subdirectories "param_index" and "structures" corresponds to *DIRNAME_R1NUMR2*. *DIRNAME* stands for the name of the main directory ("S_closed", "S_open", "S_ACE2" and "S_postfus"), *R1* stands for the one-letter residue code of the "wild type" residue of the S protein, *NUM* stands for the residue number according to the Uniprot sequence of SARS-CoV-2 Spike protein and *R2* stands for the one-letter residue code of the variant to which the "wild type" residue *R1* was changed. Results of the HADDOCK score and its components calculations performed with HADDOCK2.2 for each generated variant are summarized in Tables S2–S6.

## Code availability

All unpublished code and scripts used in this study are available upon request.

***CRediT authorship contribution statement.*** **Marija Sorokina:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Jaydeep Belapure:** Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – review & editing. **Christian Tüting:** Data curation, Formal analysis, Methodology, Validation, Writing – review & editing. **Reinhard Paschke:** Funding acquisition, Project administration, Resources. **Ioannis Papasotiriou:** Funding acquisition, Project administration, Resources, Supervision. **João P.G.L.M Rodrigues:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft. **Panagiotis L. Kastritis:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration,

Funding acquisition, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

### DATA AVAILABILITY

All data is available at https://data.sbgrid.org/dataset/851/.

## Acknowledgements

The authors thank the members of the Kastritis Laboratory, RGCC and BioSolutions for valuable discussions. This work was supported by the Federal Ministry for Education and Research (BMBF, ZIK program) (Grant nos. 03Z22HI2, 03Z22HN23 and 03COV04 to PLK), the European Regional Development Funds for Saxony-Anhalt (grant no. EFRE: ZS/2016/04/78115 to PLK), funding by Deutsche Forschungsgemeinschaft (DFG) (project number 391498659, RTG 2467), and the Martin-Luther University of Halle-Wittenberg.

## Author contributions

P.L.K. conceived the project; P.L.K. and M.S. designed research; M.S. performed research with contributions from P.L.K., C.T. and J.B.; M.S. and P.L.K. analyzed data with contributions from R.P., I.P., J.P.G.L.M.R, J.B. and C.T.; and P.L.K. and M.S. wrote the paper with contributions from all authors.

## Competing Interest Statement

I.P. is the founder and director of R.G.C.C. International GmbH. RP is the founder and director of BioSolutions Halle GmbH. MS is supported by both R.G.C.C. International GmbH and BioSolutions GmbH. The authors declare that they have no competing interests.

## Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2022.167637.

*Received 18 January 2022;*
*Accepted 6 May 2022;*
*Available online 17 May 2022*

***Keywords***:
S protein;
ACE2;
HADDOCK;
COVID-19;
protein–protein interactions

***Abbreviations***:
ACE2, Angiotensin-converting enzyme 2; ARIA, Ambiguous refinement for interface assessment; BEST, Bayesian estimation supersedes the t-test; BSA, Buried surface area; CD, Connector domain; CH, Central helix region; CHARMM, Chemistry at Harvard macromolecular mechanics; CNS, Crystallographic and NMR system; CT, C-terminal domain; COVID-19, Coronavirus disease 2019; CoVs, Coronaviruses; Desolv, HADDOCK's desolvation energy component/score; Elec, HADDOCK's electrostatics energy component/score; FP, Fusion peptide; FPPR, Fusion peptide proximal region; FSC, Fourier shell correlation; HADDOCK, Highly ambiguous data-driven docking; HD, HADDOCK score; HDI, High-density probability interval; HR1, Heptad repeat 1 domain; HR2, Heptad repeat 2 domain; mAb, Monoclonal antibody; MCMC, Markov chain Monte Carlo; MD, Molecular dynamics simulation; NAMD, Nanoscale molecular dynamics; NTD, N-terminal domain; OPLS, Optimized potential for liquid simulations; PDB, Protein data bank; p.occ, Percentage of occurrence; RBD, Receptor-binding domain; RBM, Receptor-binding motif; rmsd, root-mean-square deviation; ROPE, Region of practical equivalence; SARS-CoV-2, Severe acute respiratory syndrome coronavirus 2; SD1, Subdomain 1; SD2, Subdomain 2; S protein, Spike protein; TM, Transmembrane region; UH, Upstream helix; vdW, HADDOCK's van der Waals energy component/score

## References

1. Zhang, R., Li, Y., Zhang, A.L., Wang, Y., Molina, M.J., (2020). Identifying airborne transmission as the dominant route for the spread of COVID-19. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 14857–14863.
2. Gumel, A.B., Iboi, E.A., Ngonghala, C.N., Ngwa, G.A., (2021). Toward Achieving a Vaccine-Derived Herd Immunity Threshold for COVID-19 in the U.S.. *Front Public Health.* **9**, 709369
3. Li, F., (2016). Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annu. Rev. Virol.* **3**, 237–261.
4. Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A., Li, F., (2020). Cell entry mechanisms of SARS-CoV-2. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 11727–11734.
5. Fauver, J.R., Petrone, M.E., Hodcroft, E.B., Shioda, K., Ehrlich, H.Y., Watts, A.G., Vogels, C.B.F., Brito, A.F., et al., (2020). Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell* **181** 990–996 e995.
6. Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E.E., et al., (2020). Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182** 812–827 e819.
7. Tortorici, M.A., Veesler, D., (2019). Structural insights into coronavirus entry. *Adv. Virus Res.* **105**, 93–116.
8. 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 e775.

9. Peacock, T.P., Goldhill, D.H., Zhou, J., Baillon, L., Frise, R., Swann, O.C., Kugathasan, R., Penn, R., et al., (2021). The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets. *Nature Microbiol.*.
10. Li, W., Moore, M.J., Vasilieva, N., Sui, J., Wong, S.K., Berne, M.A., Somasundaran, M., Sullivan, J.L., et al., (2003). Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* **426**, 450–454.
11. Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., et al., (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273.
12. Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., Zhou, Q., (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* **367**, 1444–1448.
13. Vickers, C., Hales, P., Kaushik, V., Dick, L., Gavin, J., Tang, J., Godbout, K., Parsons, T., et al., (2002). Hydrolysis of biological peptides by human angiotensin-converting enzyme-related carboxypeptidase. *J. Biol. Chem.* **277**, 14838–14843.
14. Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., et al., (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220.
15. Xia, S., Liu, M., Wang, C., Xu, W., Lan, Q., Feng, S., Qi, F., Bao, L., et al., (2020). Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. *Cell Res.* **30**, 343–355.
16. Walls, A.C., Park, Y.J., Tortorici, M.A., Wall, A., McGuire, A.T., Veesler, D., (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181** 281–292 e286.
17. Hoffmann, M., Kleine-Weber, H., Schroeder, S., Kruger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., 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 e278.
18. Xia, X., (2021). Domains and Functions of Spike Protein in Sars-Cov-2 in the Context of Vaccine Design. *Viruses* **13**
19. Benetti, E., Tita, R., Spiga, O., Ciolfi, A., Birolo, G., Bruselles, A., Doddato, G., Giliberti, A., et al., (2020). ACE2 gene variants may underlie interindividual variability and susceptibility to COVID-19 in the Italian population. *Eur. J. Hum. Genet.* **28**, 1602–1614.
20. Latinne, A., Hu, B., Olival, K.J., Zhu, G., Zhang, L., Li, H., Chmura, A.A., Field, H.E., et al., (2020). Origin and cross-species transmission of bat coronaviruses in China. *Nature Commun.* **11**, 4235.
21. Sorokina, M., J, M.C.T., Barrera-Vilarmau, S., Paschke, R., Papasotiriou, I., Rodrigues, J., Kastritis, P.L., (2020). Structural models of human ACE2 variants with SARS-CoV-2 Spike protein for structure-based drug design. *Sci. Data* **7**, 309.
22. Rodrigues, J., Barrera-Vilarmau, S., J, M.C.T., Sorokina, M., Seckel, E., Kastritis, P.L., Levitt, M., (2020). Insights on cross-species transmission of SARS-CoV-2 from structural modeling. *PLoS Comput. Biol.* **16**, e1008449.
23. Lam, S.D., Bordin, N., Waman, V.P., Scholes, H.M., Ashford, P., Sen, N., van Dorp, L., Rauer, C., et al., (2020). SARS-CoV-2 spike protein predicted to form complexes with host receptor protein orthologues from a broad range of mammals. *Sci. Rep.* **10**, 16471.
24. Huang, X., Zhang, C., Pearce, R., Omenn, G.S., Zhang, Y., (2020). Identifying the Zoonotic Origin of SARS-CoV-2 by Modeling the Binding Affinity between the Spike Receptor-Binding Domain and Host ACE2. *J. Proteome Res.* **19**, 4844–4856.
25. Li, Q., Wu, J., Nie, J., Zhang, L., Hao, H., Liu, S., Zhao, C., Zhang, Q., et al., (2020). The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* **182** 1284–1294 e1289.
26. Dejnirattisai, W., Zhou, D., Supasa, P., Liu, C., Mentzer, A.J., Ginn, H.M., Zhao, Y., Duyvesteyn, H.M.E., et al., (2021). Antibody evasion by the P.1 strain of SARS-CoV-2. *Cell* **184**, 2939–2954 e2939.
27. Zhu, X., Mannar, D., Srivastava, S.S., Berezuk, A.M., Demers, J.P., Saville, J.W., Leopold, K., Li, W., et al., (2021). Cryo-electron microscopy structures of the N501Y SARS-CoV-2 spike protein in complex with ACE2 and 2 potent neutralizing antibodies. *PLoS Biol.* **19**, e3001237
28. Gobeil, S.M., Janowska, K., McDowell, S., Mansouri, K., Parks, R., Stalls, V., Kopp, M.F., Manne, K., et al., (2021). Effect of natural mutations of SARS-CoV-2 on spike structure, conformation, and antigenicity. *Science* **373** Article eabi6226.
29. Groves, D.C., Rowland-Jones, S.L., Angyal, A., (2021). The D614G mutations in the SARS-CoV-2 spike protein: Implications for viral infectivity, disease severity and vaccine design. *Biochem. Biophys. Res. Commun.* **538**, 104–107.
30. Biswas, S.K., Mudi, S.R., (2020). Spike protein D614G and RdRp P323L: the SARS-CoV-2 mutations associated with severity of COVID-19. *Genomics Inform.* **18**, e44
31. Gobeil, S.M., Janowska, K., McDowell, S., Mansouri, K., Parks, R., Manne, K., Stalls, V., Kopp, M.F., et al., (2021). D614G Mutation Alters SARS-CoV-2 Spike Conformation and Enhances Protease Cleavage at the S1/S2 Junction. *Cell Rep.* **34**, 108630
32. Zhang, J., Cai, Y., Xiao, T., Lu, J., Peng, H., Sterling, S. M., Walsh Jr., R.M., Rits-Volloch, S., et al., (2021). Structural impact on SARS-CoV-2 spike protein by D614G substitution. *Science* **372**, 525–530.
33. Supasa, P., Zhou, D., Dejnirattisai, W., Liu, C., Mentzer, A.J., Ginn, H.M., Zhao, Y., Duyvesteyn, H.M.E., et al., (2021). Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by convalescent and vaccine sera. *Cell* **184**, 2201–2211 e2207.
34. Thomson, E.C., Rosen, L.E., Shepherd, J.G., Spreafico, R., da Silva Filipe, A., Wojcechowskyj, J.A., Davis, C., Piccoli, L., et al., (2021). Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. *Cell* **184** 1171–1187 e1120.
35. Laurini, E., Marson, D., Aulic, S., Fermeglia, A., Pricl, S., (2021). Computational Mutagenesis at the SARS-CoV-2 Spike Protein/Angiotensin-Converting Enzyme 2 Binding Interface: Comparison with Experimental Evidence. *ACS Nano* **15**, 6929–6948.
36. Shu, Y., McCauley, J., (2017). GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro. Surveill.* **22**
37. Nussinov, R., Ma, B., Tsai, C.J., (2014). Multiple conformational selection and induced fit events take place in allosteric propagation. *Biophys. Chem.* **186**, 22–30.

38. Boehr, D.D., Nussinov, R., Wright, P.E., (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chem. Biol.* **5**, 789–796.
39. Cai, Y., Zhang, J., Xiao, T., Peng, H., Sterling, S.M., Walsh Jr., R.M., Rawson, S., Rits-Volloch, S., et al., (2020). Distinct conformational states of SARS-CoV-2 spike protein. *Science* **369**, 1586–1592.
40. Trigueiro-Louro, J., Correia, V., Figueiredo-Nunes, I., Giria, M., Rebelo-de-Andrade, H., (2020). Unlocking COVID therapeutic targets: A structure-based rationale against SARS-CoV-2, SARS-CoV and MERS-CoV Spike. *Comput. Struct. Biotechnol. J.* **18**, 2117–2131.
41. Kalathiya, U., Padariya, M., Mayordomo, M., Lisowska, M., Nicholson, J., Singh, A., Baginski, M., Fahraeus, R., et al., (2020). Highly Conserved Homotrimer Cavity Formed by the SARS-CoV-2 Spike Glycoprotein: A Novel Binding Site. *J. Clin. Med.* **9**
42. Walls, A.C., Tortorici, M.A., Snijder, J., Xiong, X., Bosch, B.J., Rey, F.A., Veesler, D., (2017). Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 11157–11162.
43. Duffy, S., (2018). Why are RNA virus mutation rates so damn high? *PLoS Biol.* **16**, e3000003
44. Mitsuki, Y.Y., Ohnishi, K., Takagi, H., Oshima, M., Yamamoto, T., Mizukoshi, F., Terahara, K., Kobayashi, K., et al., (2008). A single amino acid substitution in the S1 and S2 Spike protein domains determines the neutralization escape phenotype of SARS-CoV. *Microbes Infect.* **10**, 908–915.
45. He, Y., Li, J., Jiang, S., (2006). A single amino acid substitution (R441A) in the receptor-binding domain of SARS coronavirus spike protein disrupts the antigenic structure and binding activity. *Biochem. Biophys. Res. Commun.* **344**, 106–113.
46. Bloom, J.D., Gong, L.I., Baltimore, D., (2010). Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science* **328**, 1272–1275.
47. Zhang, T.H., Dai, L., Barton, J.P., Du, Y., Tan, Y., Pang, W., Chakraborty, A.K., Lloyd-Smith, J.O., et al., (2020). Predominance of positive epistasis among drug resistance-associated mutations in HIV-1 protease. *PLoS Genet.* **16**, e1009009
48. Tasakis, R.N., Samaras, G., Jamison, A., Lee, M., Paulus, A., Whitehouse, G., Verkoczy, L., Papavasiliou, F.N., et al., (2021). SARS-CoV-2 variant evolution in the United States: High accumulation of viral mutations over time likely through serial Founder Events and mutational bursts. *PLoS ONE* **16**, e0255169
49. Robson, F., Khan, K.S., Le, T.K., Paris, C., Demirbag, S., Barfuss, P., Rocchi, P., Ng, W.L., (2020). Coronavirus RNA Proofreading: Molecular Basis and Therapeutic Targeting. *Mol. Cell* **80**, 1136–1138.
50. Hodcroft, E.B., Zuber, M., Nadeau, S., Vaughan, T.G., Crawford, K.H.D., Althaus, C.L., Reichmuth, M.L., Bowen, J.E., et al., (2020). Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *medRxiv*.
51. McCallum, M., De Marco, A., Lempp, F.A., Tortorici, M.A., Pinto, D., Walls, A.C., Beltramello, M., Chen, A., et al., (2021). N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184** 2332–2347 e2316.
52. Singh, A., Steinkellner, G., Kochl, K., Gruber, K., Gruber, C.C., (2021). Serine 477 plays a crucial role in the interaction of the SARS-CoV-2 spike protein with the human receptor ACE2. *Sci. Rep.* **11**, 4320.
53. Starr, T.N., Greaney, A.J., Hilton, S.K., Ellis, D., Crawford, K.H.D., Dingens, A.S., Navarro, M.J., Bowen, J.E., et al., (2020). Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182** 1295–1310 e1220.
54. Liu, Z., VanBlargan, L.A., Bloyet, L.M., Rothlauf, P.W., Chen, R.E., Stumpf, S., Zhao, H., Errico, J.M., et al., (2021). Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *Cell Host Microbe* **29** 477–488 e474.
55. Chen, J., Wang, R., Wang, M., Wei, G.W., (2020). Mutations Strengthened SARS-CoV-2 Infectivity. *J. Mol. Biol.* **432**, 5212–5226.
56. Wang, R., Chen, J., Gao, K., Wei, G.W., (2021). Vaccine-escape and fast-growing mutations in the United Kingdom, the United States, Singapore, Spain, India, and other COVID-19-devastated countries. *Genomics* **113**, 2158–2170.
57. Chaturvedi, P., Han, Y., Kral, P., Vukovic, L., (2020). Adaptive Evolution of Peptide Inhibitors for Mutating SARS-CoV-2. *ChemRxiv*.
58. Bayarri-Olmos, R., Rosbjerg, A., Johnsen, L.B., Helgstrand, C., Bak-Thomsen, T., Garred, P., Skjoedt, M.O., (2021). The SARS-CoV-2 Y453F mink variant displays a pronounced increase in ACE-2 affinity but does not challenge antibody neutralization. *J. Biol. Chem.*, 100536.
59. Starr, T.N., Greaney, A.J., Addetia, A., Hannon, W.W., Choudhary, M.C., Dingens, A.S., Li, J.Z., Bloom, J.D., (2021). Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science* **371**, 850–854.
60. Teruel, N., Mailhot, O., Najmanovich, R.J., (2021). Modelling conformational state dynamics and its role on infection for SARS-CoV-2 Spike protein variants. *PLoS Comput. Biol.* **17**, e1009286
61. Cai, H.Y., Cai, A., (2021). SARS-CoV2 spike protein gene variants with N501T and G142D mutation-dominated infections in mink in the United States. *J. Vet. Diagn. Invest.* **33**, 939–942.
62. Kastritis, P.L., Rodrigues, J.P., Folkers, G.E., Boelens, R., Bonvin, A.M., (2014). Proteins feel more than they see: fine-tuning of binding affinity by properties of the non-interacting surface. *J. Mol. Biol.* **426**, 2632–2652.
63. Kastritis, P.L., Bonvin, A.M., (2013). On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *J. R. Soc. Interface* **10**, 20120835.
64. Vihinen, M., (2015). Types and effects of protein variations. *Hum. Genet.* **134**, 405–421.
65. Lee, D., Redfern, O., Orengo, C., (2007). Predicting protein function from sequence and structure. *Nature Rev. Mol. Cell Biol.* **8**, 995–1005.
66. Gur, M., Taka, E., Yilmaz, S.Z., Kilinc, C., Aktas, U., Golcuk, M., (2020). Conformational transition of SARS-CoV-2 spike glycoprotein between its closed and open states. *J. Chem. Phys.* **153**, 075101
67. Khare, S., Azevedo, M., Parajuli, P., Gokulan, K., (2021). Conformational Changes of the Receptor Binding Domain of SARS-CoV-2 Spike Protein and Prediction of a B-Cell

Antigenic Epitope Using Structural Data. *Front. Artif. Intell.* **4**, 630955
68. Lu, M., Uchil, P.D., Li, W., Zheng, D., Terry, D.S., Gorman, J., Shi, W., Zhang, B., et al., (2020). Real-Time Conformational Dynamics of SARS-CoV-2 Spikes on Virus Particles. *Cell Host Microbe* **28** 880–891 e888.
69. van Zundert, G.C.P., Rodrigues, J., Trellet, M., Schmitz, C., Kastritis, P.L., Karaca, E., Melquiond, A.S.J., van Dijk, M., et al., (2016). The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J. Mol. Biol.* **428**, 720–725.
70. Geng, C., Vangone, A., Bonvin, A., (2016). Exploring the interplay between experimental methods and the performance of predictors of binding affinity change upon mutations in protein complexes. *Protein Eng. Des. Sel.* **29**, 291–299.
71. Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R., O'Meara, M. J., DiMaio, F.P., Park, H., Shapovalov, M.V., Renfrew, P. D., et al., (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048.
72. Zhang, Z., Witham, S., Alexov, E., (2011). On the role of electrostatics in protein-protein interactions. *Phys. Biol.* **8**, 035001
73. Benton, D.J., Wrobel, A.G., Roustan, C., Borg, A., Xu, P., Martin, S.R., Rosenthal, P.B., Skehel, J.J., Gamblin, S.J., (2021). The effect of the D614G substitution on the structure of the spike glycoprotein of SARS-CoV-2. *Proc. Natl. Acad. Sci. U. S. A.* **118**
74. Yurkovetskiy, L., Wang, X., Pascal, K.E., Tomkins-Tinch, C., Nyalile, T.P., Wang, Y., Baum, A., Diehl, W.E., et al., (2020). Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. *Cell* **183** 739–751 e738.
75. Turonova, B., Sikora, M., Schurmann, C., Hagen, W.J.H., Welsch, S., Blanc, F.E.C., von Bulow, S., Gecht, M., et al., (2020). In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges. *Science* **370**, 203–208.
76. Ke, Z., Oton, J., Qu, K., Cortese, M., Zila, V., McKeane, L., Nakane, T., Zivanov, J., et al., (2020). Structures and distributions of SARS-CoV-2 spike proteins on intact virions. *Nature* **588**, 498–502.
77. Weekley, C.M., Purcell, D.F.J., Parker, M.W., (2021). SARS-CoV-2 Spike receptor-binding domain with a G485R mutation in complex with human ACE2. *bioRxiv*. 2021.2003.2016.434488.
78. Jo, S., Kim, T., Iyer, V.G., Im, W., (2008). CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865.
79. Choi, Y.K., Cao, Y., Frank, M., Woo, H., Park, S.J., Yeom, M.S., Croll, T.I., Seok, C., et al., (2021). Structure, Dynamics, Receptor Binding, and Antibody Binding of the Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein in a Viral Membrane. *J. Chem. Theory Comput.* **17**, 2479–2487.
80. Woo, H., Park, S.J., Choi, Y.K., Park, T., Tanveer, M., Cao, Y., Kern, N.R., Lee, J., et al., (2020). Developing a Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein Model in a Viral Membrane. *J. Phys. Chem. B* **124**, 7128–7137.
81. Sali, A., Blundell, T.L., (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815.
82. Rodrigues, J., Teixeira, J.M.C., Trellet, M., Bonvin, A., (2018). pdb-tools: a swiss army knife for molecular structures. *F1000Res* **7**, 1961.
83. Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.L., Abiona, O., Graham, B.S., McLellan, J. S., (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263.
84. Hsieh, C.L., Goldsmith, J.A., Schaub, J.M., DiVenere, A. M., Kuo, H.C., Javanmardi, K., Le, K.C., Wrapp, D., et al., (2020). Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science* **369**, 1501–1505.
85. Chen, H., Lyne, P.D., Giordanetto, F., Lovell, T., Li, J., (2006). On evaluating molecular-docking methods for pose prediction and enrichment factors. *J. Chem. Inf. Model.* **46**, 401–415.
86. Meng, X.Y., Zhang, H.X., Mezei, M., Cui, M., (2011). Molecular docking: a powerful approach for structure-based drug discovery. *Curr. Comput. Aided Drug Des.* **7**, 146–157.
87. Durrant, J.D., McCammon, J.A., (2011). Molecular dynamics simulations and drug discovery. *BMC Biol.* **9**, 71.
88. Lopes, P.E., Guvench, O., MacKerell Jr., A.D., (2015). Current status of protein force fields for molecular dynamics simulations. *Methods Mol. Biol.* **1215**, 47–71.
89. Rastelli, G., Pinzi, L., (2019). Refinement and Rescoring of Virtual Screening Results. *Front. Chem.* **7**, 498.
90. Kastritis, P.L., van Dijk, A.D., Bonvin, A.M., (2012). Explicit treatment of water molecules in data-driven protein-protein docking: the solvated HADDOCKing approach. *Methods Mol. Biol.* **819**, 355–374.
91. Kastritis, P.L., Visscher, K.M., van Dijk, A.D., Bonvin, A. M., (2013). Solvated protein-protein docking using Kyte-Doolittle-based water preferences. *Proteins* **81**, 510–518.
92. Kastritis, P.L., Bonvin, A.M., (2010). Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. *J. Proteome Res.* **9**, 2216–2225.
93. Linge, J.P., O'Donoghue, S.I., Nilges, M., (2001). Automated assignment of ambiguous nuclear overhauser effects with ARIA. *Methods Enzymol.* **339**, 71–90.
94. Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., et al., (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54**, 905–921.
95. Linge, J.P., Nilges, M., (1999). Influence of non-bonded parameters on the quality of NMR structures: a new force field for NMR structure calculation. *J. Biomol. NMR* **13**, 51–59.
96. Dominguez, C., Boelens, R., Bonvin, A.M., (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737.
97. UniProt, C., (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515.
98. Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., et al., (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**, W636–W641.

99. Juraszek, J., Rutten, L., Blokland, S., Bouchier, P., Voorzaat, R., Ritschel, T., Bakkers, M.J.G., Renault, L.L. R., et al., (2021). Stabilizing the closed SARS-CoV-2 spike trimer. *Nature Commun.* **12**, 244.
100. Xu, W., Wang, M., Yu, D., Zhang, X., (2020). Variations in SARS-CoV-2 Spike Protein Cell Epitopes and Glycosylation Profiles During Global Transmission Course of COVID-19. *Front. Immunol.* **11**, 565278
101. Rambaut, A., Holmes, E.C., O'Toole, A., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., Pybus, O.G., (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiol.* **5**, 1403–1407.
102. Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., et al., (2020). Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv.* 2020.2012.2021.20248640.
103. Faria, N.R., Mellan, T.A., Whittaker, C., Claro, I.M., Candido, D.D.S., Mishra, S., Crispim, M.A.E., Sales, F. C.S., et al., (2021). Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **372**, 815.
104. Mlcochova, P., Kemp, S.A., Dhar, M.S., Papa, G., Meng, B., Ferreira, I., Datir, R., Collier, D.A., et al., (2021). SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* **599**, 114–119.
105. Kyte, J., Doolittle, R.F., (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132.
106. Salvatier, J., Wiecki, T.V., Christopher, F., (2016). Probabilistic programming in Python using PyMC3. *PeerJ Comput. Sci.* **2**, e55
107. Kruschke, J.K., (2013). Bayesian estimation supersedes the t test. *J. Exp. Psychol. Gen.* **142**, 573–603.
108. Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., et al., (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802.
109. Huang, J., MacKerell Jr., A.D., (2013). CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J. Comput. Chem.* **34**, 2135–2145.
110. Essmann, U., Petera, L., Berkowitz, M.L., Darden, T., Lee, H., Lee, G.P., (1995). A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593.
111. Humphrey, W., Dalke, A., Schulten, K., (1996). VMD: visual molecular dynamics. *J. Mol. Graph.* **14** (33–38), 27–38.
112. Sorokina, M., Kastritis, P.L., Rodrigues, J.P., (2021). Varinats of SARS-CoV-2 Spike protein in closed, open, ACE2-bound and postfusion conformational state. *SBGrid* **851** https://datasbgrid.org/dataset/851/.