Distant Residues Modulate Conformational Opening in SARS-CoV-2 Spike Protein

Dhiman Ray,[†] Ly Le,[†,‡] and Ioan Andricioaei[*,†,¶]

†*Department of Chemistry, University of California Irvine, Irvine CA 92697*

‡*School of Biotechnology, International University, Vietnam National University, Ho Chi Minh City, Vietnam*

¶*Department of Physics and Astronomy, University of California Irvine, Irvine CA 92697*

E-mail: andricio@uci.edu

Keywords

Molecular biophysics, molecular dynamics, virus structure, statistical mechanics, computer simulation of proteins, COVID-19

Classification: Biophysics and Computational Biology

Abstract

Infection by SARS-CoV-2 involves the attachment of the receptor binding domain (RBD) of its spike proteins to the ACE2 receptors on the peripheral membrane of host cells. Binding is initiated by a *down* to *up* conformational change in the spike protein, an opening which presents the RBD to the receptor. To date, computational and experimental studies for therapeutics have concentrated, for good reason, on the RBD. However, the RBD region is highly prone to mutations, and therefore will possibly arise drug resistance. In contrast, we here focus on the correlations between the RBD and residues distant to it in the spike protein. We thereby provide a deeper understanding of the role of distant residues in the molecular mechanism of infection. Predictions of key mutations in distant allosteric binding sites are provided, with implications for therapeutics. Identifying these emerging mutants can also go a long way towards pre-designing vaccines for future outbreaks. The model we use, based on time-independent component analysis (tICA) and protein graph connectivity network, is able to identify multiple residues that exhibit long-distance coupling with the RBD opening. Mutation on these residues can lead to new strains of coronavirus with different degrees of transmissibility and virulence. The most ubiquitous D614G mutation and the A570D mutation of the highly contagious UK SARS-CoV-2 variant are predicted *ab-initio* from our model. Conversely, broad spectrum therapeutics like drugs and monoclonal antibodies can be generated targeting these key distant but conserved regions of the spike protein.

Significance statement

The novel coronavirus (SARS-CoV-2) pandemic resulted in the largest economic and public health crises in recent times. Significant drug design effort, against SARS-CoV-2, is focused on the receptor binding domain (RBD) of the spike protein, although this region is prone to mutations causing therapeutic resistance. We applied deep data analysis methods on all-atom molecular dynamics simulations to identify key non-RBD residues that play a crucial

role in spike-receptor binding and infection. Because the non-RBD residues are typically conserved across multiple coronaviruses, they can not only be targeted by broad spectrum antibodies and drugs against existing strains, but can also offer predictive insights into how to design adaptable antiviral therapeutic framework against spike of strains that might appear during future epidemics.

Introduction

The COVID-19 pandemic continues its spread, with more than 83 million confirmed cases worldwide and 1.8 million deaths by the end of December 2020, according to the World Health Organization. The etiological agent, SARS-CoV-2, is a member of the *Coronaviridae* family, which includes SARS-CoV-1 (2002-2004) and MERS-CoV (since 2012), viruses with which SARS-CoV-2 has a sequence identity of 79.6% and 50%, respectively.[1,2] Expressed on the surface of SARS-CoV-2, the spike (S) protein plays a crucial role. It binds to the host angiotensin-converting enzyme 2 (ACE2) through the S protein's receptor-binding domain (RBD), thereby facilitating viral entry into host cells.[3,4] Therefore, the spike protein is a preponderant target for inhibitors that impede SARS-CoV-2 infection. When assessing genomic variability of SARS-CoV-2, we are presented with a moderate mutation rate compared to other RNA viruses, around 1.12×10^{-3} nucleotide substitution/site/year,[5] (but much larger than DNA viruses); the SARS-CoV-2 mutation rate is at the same level as SARS-CoV-1.[6] Significantly, the spike protein has been demonstrated to be particularly susceptible to acquiring mutations.[5,7–9] More specifically, a study analyzing 10,022 SARS-CoV-2 genomes from 68 countries revealed 2969 different missense variants, with 427 variants in the S protein alone.[5] This suggests a strong propensity to form new strains with higher virulence and more complicated epidemiology; the dominant D614G mutation and the recent B.1.1.7 mutant are notable examples.[10] Spike protein variability can thus render currently existing therapeutic agents ineffective in combating SARS-CoV-2 and other probable SARS epidemics in the fu-

ture. It therefore is of fundamental value to understand, in microscopic detail, the role of spike mutations to the structural dynamics that triggers infection.

Large scale screening of therapeutic molecules and antibodies are underway, aiming to target the spike protein and consequently to prevent infection. Most of the experimental[11–14] and computational[15–17] efforts for inhibitor design focus on the receptor binding domain (RBD), despite the fact that this region is highly mutation-prone[18] and can cause resistance to therapeutics. For example, the mutations in the RBD observed in the emerging viral lineage in South Africa resulted in up to a ten-fold reduction in the neutralization capacity of conventional antibody therapy.[19]

Domains in the spike other than the RBD are also possible targets for inhibition. The human immune system started generating antibodies specific to residues outside the RBD even at the early stages of the pandemic. Liu et al. extracted from infected patients multiple COVID-19 neutralizing antibodies, a fraction of which bind to non-RBD epitopes of the S-protein, such as the N-terminal domain (NTD).[20] Moreover, a separate group of antibodies, present in the bloodstream of uninfected humans (particularly, and importantly, in children), were observed to bind specifically to the S2-stem region. This structure has virtually identical sequence in all coronavirus strains, including the ones causing common cold.[21] The presence of such universal antibodies, selective to the conserved regions of the spike, is hypothesized to cause the absence of severe infections in children. Such observations motivate an effort towards designing small-molecule drugs and antibodies targeted towards residues far from the RBD in the 3D structure which, consequently, are less prone to mutation.

A number of studies explored druggable hotspots in the spike protein that modulate, via allosteric effects, ACE2 binding.[22,23] Although this is a noteworthy approach towards targeting non-RBD residues, the S-ACE2 bound complex once formed already can initiate infection. So inhibitors which bind to the ACE2-bound spike are only partially effective in preventing viral entry. We, therefore, focus our study on the effect of distant residues on the dynamics of the structural transition in the spike protein leading to the RBD-up confor-

mation, i.e., *before* it becomes posed for binding (see Fig 1). This is because RBD opening plays an obligatory role in the infection by displaying the RBD to the ACE2 receptor. Therapeutics inhibiting this structural transition can therefore prevent ACE2 binding altogether, providing a higher degree of barrier towards the infection.

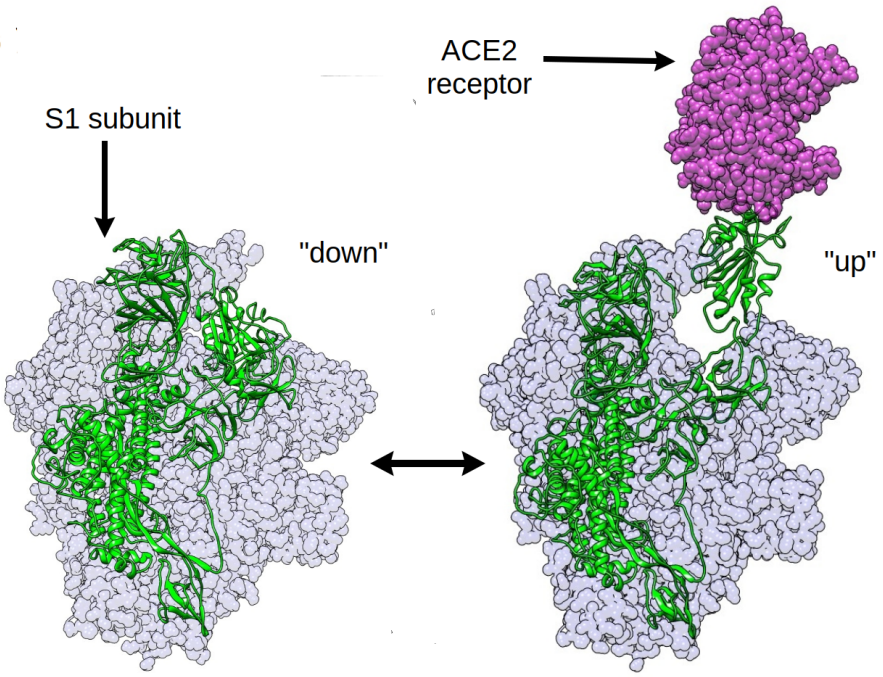

Figure 1: The RBD opening transition in the SARS-CoV-2 spike protein: the RBD of the chain shown in green is undergoing the *down* to *up* transition leading to the binding to the human ACE2 receptor.

The down-to-up transition of the S-protein RBD has been subjected to extensive cryogenic electron microscopy (cryo-EM) studies that elucidated the atomic resolution structures of the closed (PDB ID: 6VXX,[24] 6XR8[25]), partially open (PDB ID: 6VSB[26]) and fully open (PDB ID: 6VYB[24]) states. Attempts have been made to study the dynamics of RBD opening using force-field based classical molecular dynamics (MD) simulation.[27,28] But the direct observation of an RBD opening transition is beyond the scope of atomistic MD simulation, primarily because of the large size of the system and the long timescales involved. Yet, combining multiple short trajectories totaling up to 1.03 ms, the Folding@Home project could resolve various intermediates of the RBD opening transition.[28] Moreover, structure-based

coarse-grained modeling could be employed to explore the conformational landscape of multiple configurations of the spike trimer (1-up-2-down, 2-up-1-down, 3-up and 3-down),[29] and, in a separate study, to identify allosteric communication pathways in the spike protein.[18] Despite providing a qualitative insight, the absence of explicit solvent and atomistic detail, and the exclusion of the glycan-shield (which plays dynamic roles beyond shielding itself[27]), obscured a quantitative mechanistic picture of the conformational opening transition.

We took a three-pronged approach to identify the distant residues that show correlated motion coupled to RBD opening and closing. Firstly, we obtained the underlying free energy profile of the RBD dynamics and performed multiple short unbiased simulations representing different regions of the configuration space. Secondly, we designed and employed a novel way to identify residues important for the conformational change by quantifying the correlation of the backbone torsion angles of the protein with the slowest degrees of freedom representing the down-to-up transition of the RBD. Thirdly, we took an alternative route to study allosteric connections by constructing a dynamical network model that uses the mutual information metric computed from MD trajectories. Taken together, the three approaches resulted in the prediction of a handful of residues in specific regions of the S protein that play a crucial role in the conformational rearrangements of the spike. These residues suggest possible future mutational hotspots, as well as targets for designing inhibitors that can reduce the flexibility of the RBD, leading to reduced receptor binding capability. In fact, the most ubiquitous spike protein mutation to date, the D614G,[30] and the A570D mutation in the recently emerged highly contagious UK SARS-CoV-2 strain,[31] both appear among our predicted set of residues, the latter being discovered after the completion of our calculations.

Results

Correlation between RBD-opening and backbone dihedral angle

Multiple unbiased trajectories were propagated from different regions of the S-protein RBD-opening conformational space, priorly explored by steered MD and umbrella sampling (details in Supporting Information (SI) Appendix). Three of those trajectories were assigned as the closed, partially open, and fully open state, based on the position of free energy minima along the umbrella sampling reaction coordinate (cf. SI Appendix). The stability of these three conformations were ensured by inspection of the trajectories. The cumulative simulation data was projected onto a feature space composed of pairwise distances between residues from RBD and from other parts of the spike near the RBD (details in SI Appendix); these distances increase during the down to up transition. The projected trajectory data was subjected to principal component analysis (PCA)[32] and time-independent component analysis (tICA).[33,34] The former method calculates the degrees of freedom with the highest variance in the system, while the latter obtains the ones with the longest timescale. These methods quantify large conformation changes in complex bio-molecules. The goal of performing PCA and tICA is to find out one or two coordinates which best describe the RBD opening motion. As one continuous trajectory hardly samples the transition event, multiple short trajectories spanning a large range of the configuration space were used.[35]

The first principal component (PC) and the first time-independent component (tIC), obtained from PCA and, respectively, from tICA analysis could both distinguish the closed and open states, although the partially open and fully open states could not be distinguished within the first two PCs. Yet, the projections of the open- and closed-state trajectories along the first two principal components are in agreement with the results of previous long multi-microsecond spike protein simulations.[27] Because of the clear distinction between the closed, open and the intermediate, partially open states (Fig. 2b), we chose the first two time-independent components (tICs) for our subsequent analysis.

Large scale conformational changes in proteins are, at a fundamental level, stemming from complex combinations of transitions between various states of the protein-backbone torsional angles ϕ and ψ. These combinations add up to global displacements, which set the timescale for internal friction[36] and gauge the paradoxically large number of conformational states accessible to a protein as it folds.[37] On one hand, concerted transitions of the backbone torsions typically lead to large scale motion. On the other hand, exponential divergence in nonlinear dynamical systems[38] is such that only certain dihedrals are likely to predominantly effect the conformational changes. We therefore hypothesize that there exist specific residues in the spike protein, for which the transition in backbone dihedral states result in the opening of the RBD. Moreover, we conjecture that, given the significant mutation rate, and because of the selection pressure on the virus, the residues with large impact in collective motions that facilitate infectivity are likely to be selected in spike mutants. To test this hypothesis we calculated the Pearson correlation coefficients of the sines and cosines of all the ϕ and ψ backbone torsion angles of all the residues with the first two tICs. The magnitude of the correlation is found to be significantly large only for a handful of torsion angles, whereas the majority show near-zero correlation (Fig. 2c).

We sorted the sines and cosines of the torsion angles in increasing order of the correlation coefficient for both the tICA components, this resulted in four ordered lists, depicted in Fig. 2 c. (See data availability statement to access the complete lists). We then selected the ten largest positively and negatively correlated residues from each list, adding up to 80 residues with multiple repetitions. Table 1 shows residues with the highest number of occurrences in the list of the top 80 angles. The list is dominated by pairs of consecutive residues, with ψ of the first and the ϕ of the second residue. This suggests that two consecutive torsion angles in certain regions of the protein are highest correlated with the RBD opening motion. This correlation does not exclude causation, since change in the conformational state for two subsequent torsion angles can induce crankshaft motion[39] in the backbone which, propagating along the chains, leads to a change in protein structure.

The distribution of some of these dihedral angles for the highest correlated residues, in the closed, open and partially open state, are depicted in Fig. 2d. Similar plots for all 80 angles are in the SI Appendix. Most dihedrals belonged to residues in the loop structure joining the RBD with the S2 stem, as this region is a hinge for the opening of the RBD. The angles could distinguish one or both of the open states from the RBD-down configuration. Primarily, the closed and fully open states shared a similar region in the dihedral angle space different from the partially open structure. We do consider the caveat that the fully open state, generated from the closed state using steered molecular dynamics and umbrella sampling (see SI Appendix), is an approximation for a more exact RBD up structure that binds the ACE2 receptor. Yet, the fact, that the highly correlated residues follow a distinctly different distribution in the backbone torsion angle space, indicates that a handful of non-RBD residues can play a pertinent role in the conformational change of the spike and, consequently, in the viral infection. Interestingly, the correlated torsion angles span over all three chains of the spike trimer, namely A (the one undergoing the RBD transition), B and C, hinting at the potential role of inter-residue couplings ranging over long distances in presenting the RBD to the ACE2 receptor.

When the virus mutates these particular residues in a way that increases its virulence, this increase stems from the propensity of the RBD to "flap" open and thereby increase ACE2 binding. As evidence, we highlight the example of the D614G mutation, which is already observed in numerous strains of the SARS-CoV-2 all over the world.[9,30] D614 is one of the top ranked residues in our model for the potential to play a crucial role in RBD opening. A glycine residue has the least backbone-torsion barrier for conformational transition in ϕ-ψ space due to the absence of a side chain. Replacing an Asp residue, which has higher barriers to such transitions, with a glycine can increase the flexibility of the backbone, significantly impacting the probability of observing an RBD-up conformation. We speculate that this can be the reason why this particular mutation was selected to become so widespread among COVID-19 infected patients.

A wide range of spike protein mutant sequences have been characterized, each with varying degrees of abundance. A relatively rare mutation, Ala-570-Val, resulted in a decrease of the overall stability of the spike protein in all three states, based on the FoldX empirical force field.[9,40] Free energy values[9] were obtained from only structural data and no dynamical information was considered. Yet it is worth noting that the change in total and solvation free energies, due to this mutation, were substantially different for the closed and open states, resulting in a change in ΔG for RBD opening. But, as the side chains of Ala and Val are similar in terms of steric bulk, this mutation is unlikely to significant impact RBD dynamics. As it likely did not increase the evolutionary advantage of the virus by increasing infectivity, this mutation, only occurring in one strain[9] so far, did not become as prevalent as D614G.

On the contrary, a A570D mutation is observed in the same residue in the newly emerged and highly infectious B.1.1.7 strain in the UK.[31,41,42] This mutation is likely to play a pertinent role in infection as it replaces a hydrophobic amino acid with a charged one. This leads to a significant difference in the conformational dynamics of the 570 residue and consequently impacts the large scale RBD opening motion. The B.1.1.7 also shows a P681H mutation close to the highly correlated N679 residue predicted from our model (see SI Appendix). As this mutation replaces a structurally rigid proline residue, it can possible impact the conformational space accessible to nearby residues, including N679.

Giving pause for thought, these results indicate that mutations in the highest correlated residues (Table 1) can in fact have significant physiological impact in changing the course of the pandemic. Therefore, we provide a list of residues (Table 1 and Fig. S3-S10 in SI Appendix), future mutations of which could impact RBD dynamics and consequently change the transmissibility or virulence of SARS-CoV-2. (See data availability statement for access to the raw correlation coefficient data for all residues.)

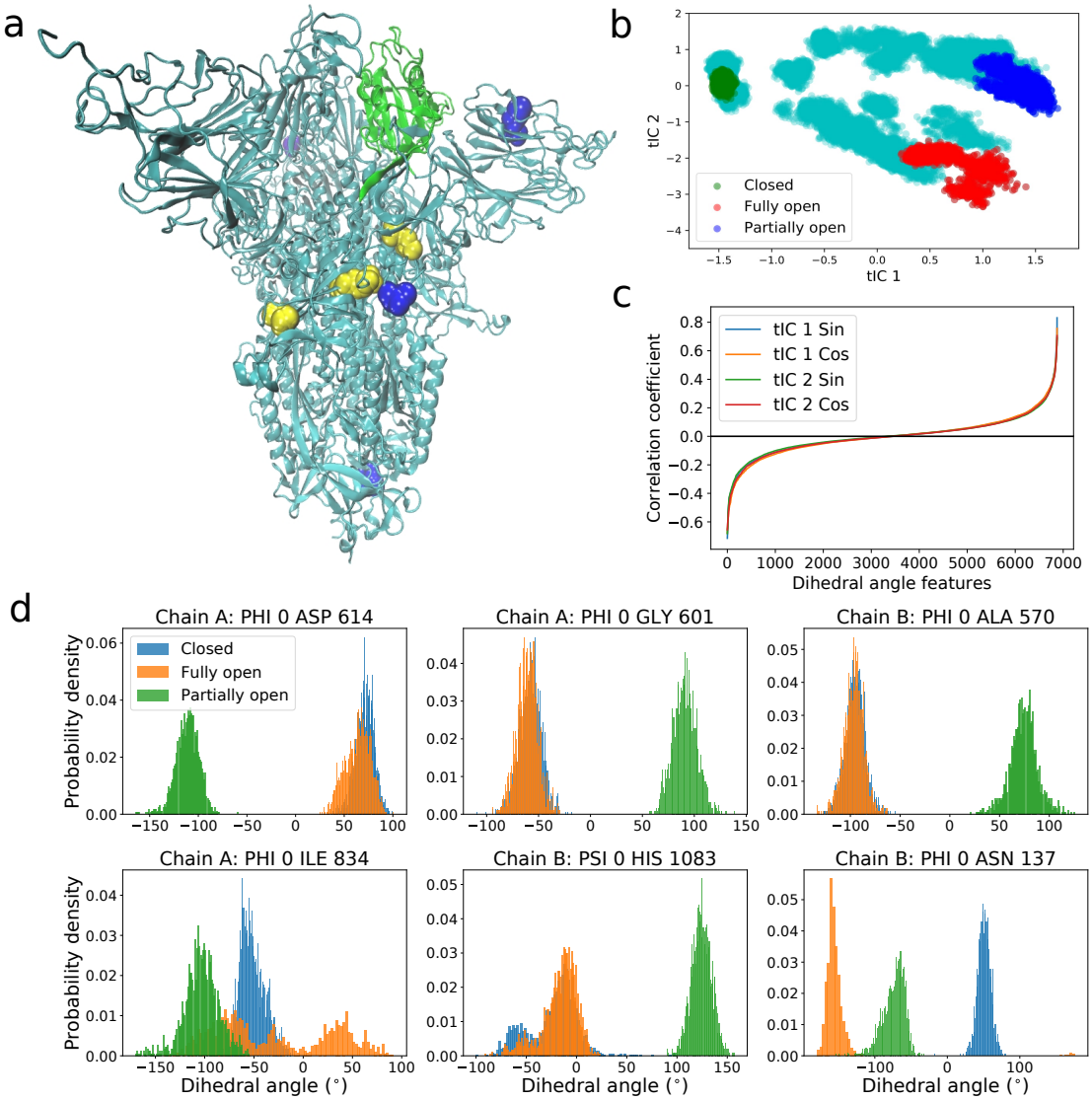

Figure 2: (a) Structure of the spike protein with the residues in RBD shown in green color. The non-RBD residues, strongly correlated with the RBD opening motion, are represented by spheres. Residues in chain A are colored yellow and those in chain B blue. The RBD of chain A is performing *down* to *up* conformational change. (b) The projection of all unbiased trajectories along the two slowest degrees of freedom (tICs) obtained from tICA analysis. (c) Pearson correlation coefficient of the sines and cosines of the backbone dihedral angles with the tIC 1 and tIC 2. The correlation values are sorted from most negative to most positive. There are 6876 dihedral angles for 3438 residues. The x axis represents the rank of each dihedral angle based on its correlation value. (d) Normalized distribution of representative backbone dihedral angles which are strongly correlated with tIC 1 and tIC 2. The distributions are calculated from closed, partially open and fully open state trajectories.

Table 1: A list of residues for which the backbone torsional angles are strongly correlated with the first two tIC components. The number of occurrences of each of them among the 80 angles (see text and SI Appendix) is also reported.

Residue	Chain	Predominant angle	Occurrences (tIC1,tIC2)
Gln613, Asp614	ABC,ABC	ψ,ϕ	(5,4),(4,4)
Pro600, Gly601	ABC,ABC	ψ,ϕ	(3,2),(3,2)
Phe833, Ile834	BC,ABC	ψ,ϕ/ψ	(0,2),(2,3)
Ile569, Ala570	ABC,BC	ψ,ϕ	(1,3),(1,2)
His1083, Asp1084	BC,B	ψ,ϕ	(1,2),(1,1)
Cys136, Asn137	B,B	ψ,ϕ	(0,1),(1,1)
Ser112	C	ψ/ϕ	(2,1)

Mutual information and network model

A different approach to characterize the coupling between distant regions in a protein is to calculate the cross correlations between the positions of different residues in 3D space. This method is often used to study allosteric effects upon ligand binding.[43–48] Conventional implementations compute the dynamic cross correlation map (DCCM) of the position vectors of the C_α atoms.[49] However, DCCM ignores correlated motions in orthogonal directions.[48] This problem can be avoided by using a linear mutual information (LMI) based cross-correlation metric, which we use in the current study.[43,50] The cross-correlation matrix elements, C_{ij}, are given by

$$C_{ij} = (1 - e^{-(2/d)I_{ij}})^{-1/2}, \tag{1}$$

where I_{ij} is the linear mutual information computed as

$$I_{ij} = H(\mathbf{x}_i) + H(\mathbf{x}_j) - H(\mathbf{x}_i, \mathbf{x}_j) \tag{2}$$

with H the Shannon entropy function:

$$
\begin{aligned}
H(\mathbf{x}_i) &= -\int p(\mathbf{x}_i) \ln p(\mathbf{x}_i) d\mathbf{x}_i \\
H(\mathbf{x}_i, \mathbf{x}_j) &= -\int\int p(\mathbf{x}_i, \mathbf{x}_j) \ln p(\mathbf{x}_i, \mathbf{x}_j) d\mathbf{x}_i d\mathbf{x}_j,
\end{aligned}
\tag{3}
$$

where $\mathbf{x}_i, \mathbf{x}_j$ are the Cartesian vectors of atomic coordinates, whereas $p(\mathbf{x}_i)$ and $p(\mathbf{x}_i, \mathbf{x}_j)$ indicate the probability density for $\mathbf{x}_i$ and, respectively, the joint probability density of $\mathbf{x}_i$ and $\mathbf{x}_j$.[43,50]

Strong correlations between distant residues are observed for the trajectories in all the three states (Fig. 3). The *change* in cross-correlation between apo and holo states of a protein is a gauge that traces allosteric communication in proteins.[43] While the RBD opening (or closing) is not strictly an allosteric process (as there is no ligand binding involved), the change in the cross-correlation between RBD down and up states (similar to holo and apo structures) provides information about long-distance residue communication, which often facilitates conformational change. That overall change in LMI correlation is clearly larger for the fully open state in comparison to the partially open state (Fig. 3). Unsurprisingly, the change is largest for the RBD and proximal residues encompassing the N-terminal domain region (residue 100-300) in all chains, as they loose contact during the opening motion. Residues in the range 524-600 show a correlation gain in the partially open state, unlike in the fully open state. This result is consistent with the dihedral angle correlation, described in the previous section: the residues in the loop region next to the RBD exhibit a change in the values of the backbone dihedral angles upon the *down* to *up* transition. The change in the correlation coefficient (ΔCorrelation) is also large for the RBDs and NTDs of chain B and C, which are in close proximity to chain A of the RBD in the closed state. Additionally, residues 600-1000 of chain C show significant gain in correlation upon the opening transition. The majority of these residues are situated at the opposite end (S2 region) of the spike, indicating the presence of long range correlated motion.

For a more profound insight, we built a protein connectivity graph network-model . In it, the C_α atoms of each amino acid are the nodes and the correlation between them are the edges connecting the nodes (the number of nodes in our system, N=3438, is to our knowledge higher than almost all other systems studied previously with this method[43–48]). We then calculated the betweenness centrality (BC), a graph theoretical measure that provides a way to quantify the amount of information that flows via the nodes and edges of a network. If a node i is working as a bridge between two other nodes along the shortest path joining them, then the BC of node i is given by

$$BC(i) = \sum_{st} \frac{n^i_{st}}{g_{st}}, \tag{4}$$

where g_{st} is the total number of geodesics (shortest paths) joining nodes s and t, out of which n^i_{st} paths pass through node i.[44] The change of BC in the dynamics of the spike protein has been recently observed using coarse-grained simulation methods.[18] Despite the coarseness of the model, a handful of residues participating in the information-propagating pathway could be identified directly from the centrality values. In the current work, we used the difference in BC as a metric to identify key residues which gain or loose relative importance along the allosteric information pathway. The difference in the normalized BC is measured by comparing the number for the partially open and fully open states with the closed conformation (i.e. $BC^{\text{partially open}} - BC^{\text{closed}}$ and $BC^{\text{fully open}} - BC^{\text{closed}}$ for every residue in the spike protein) from our all atom trajectories with explicit solvation. Importantly, our model includes the highly relevant glycan shield, particularly the N-glycans at the NTD sites N165 and N234, which were shown to modulate the conformational dynamics of the RBD by favoring a *down* conformation.[27]

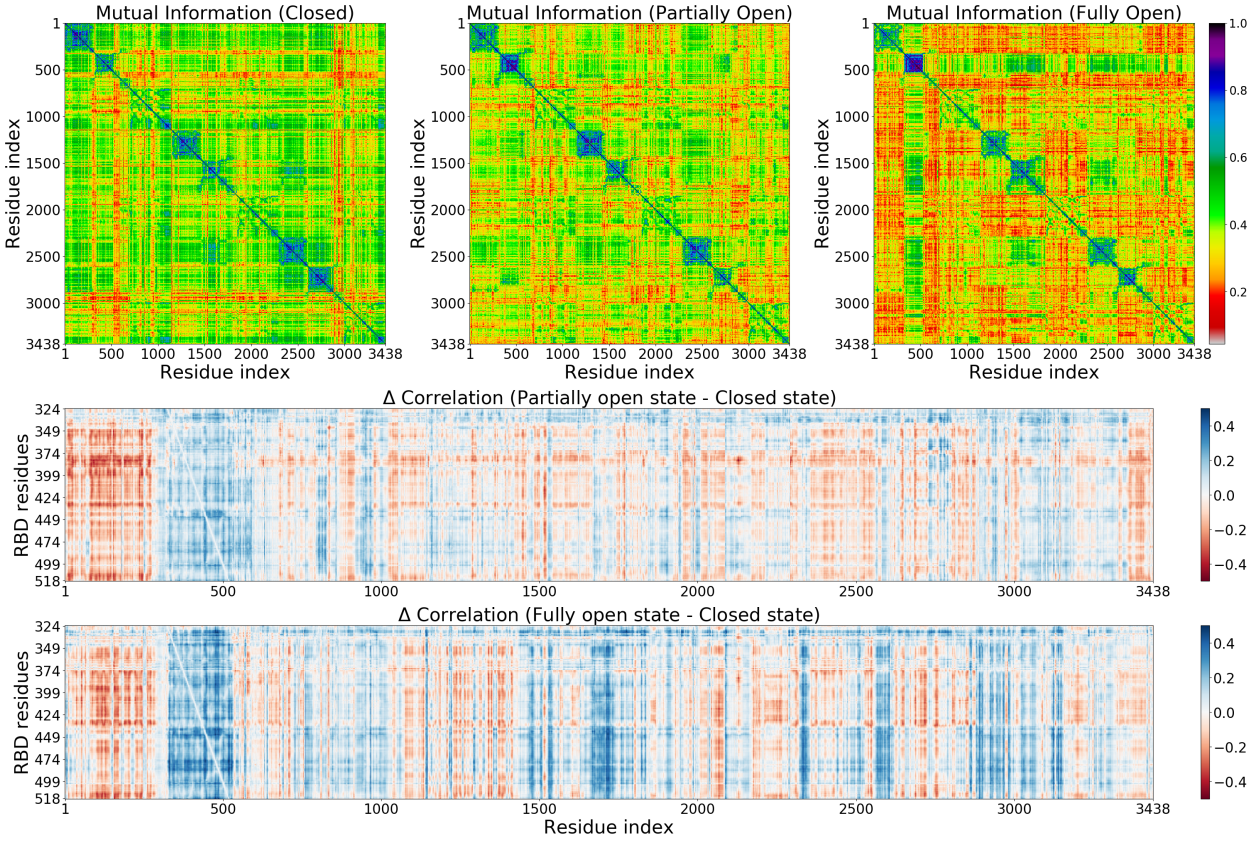

Figure 3: Upper panel: Cross correlation matrices for all three states computed using linear mutual information. RBD regions form highly correlated blocks indicating that these residues are largely decoupled from the rest of the protein. Still signatures of long-distance correlated motion is detectable. Lower panel: Difference in correlation matrix elements for the partially and fully open states with respect to the closed state. Residue index are assigned in the following way: Chain A: 1-1146; Chain B: 1147-2292; Chain C: 2293-3438.

For both the open and partially open states, the residues with significant (i.e., >0.1) change in BC are mostly from the NTD region or RBD region of the B and C chain. This suggests that the allosteric information flows through the nearby NTD's and RBDs, and mutations in this region can break the allosteric network[44] and affect the functionality of the spike protein. SARS-CoV-2 neutralizing antibodies were indeed observed to bind these regions of the spike.[20] The identified residues are in close proximity to the RBD of chain A, the one undergoing the *down* to *up* transition.

So the connectivity captured in the BC data is primarily due to short-range coupled fluctuations. Such couplings are broken when the RBD and NTD move apart, leading to

the change in BC. For the same reason, the BC of the RBD of chain A increases in the fully open state as its internal vibrations become more independent of the rest of the protein.

In a culmination of the above, the most interesting aspect is the strikingly large change in BC of the residues which are *distant* from the RBD in 3D structure. Significant gain or loss of BC is observed in residues 575, 607, 624, 757 , 896, and 897. The first three residues are present in the linker region joining the RBD with the S2 domain, while the other three are in the S2 itself. The linker region has a strong impact on the dynamics of the RBD as we already established from the dihedral angle analysis. The allosteric network analysis reinforces this conclusion. Moreover, the large change of BC in the S2 domain indicates a complex long-range information flow connecting the RBD with the core residues of the protein. This has substantial implications for pharmaceutical design, as mutations within the S2 domain can impact the receptor-binding propensity of the viral spike. These results also suggest that therapeutics targeted towards the S2 and towards the RBD-S2 linker can be effective in preventing COVID-19 infection, without complications stemming from the high rate of mutations in the RBD.

Table 2: Residues with large (> 0.1) changes in normalized betweenness centrality (BC) due to RBD opening. Residues marked with (*) show very high (> 0.2) centrality difference with respect to the closed structure. Positive and negative values are separated by horizontal line.

Partially open		Fully open	
Chain	Residue	Chain	Residue
C	1097*	C	1097*
A	155*	B	6*
B	6	A	155*
A	1095	A	1095
B	624	C	1082
C	1074	B	624
B	607	C	1002
C	1002	B	607
B	256	B	256
C	1082	C	1074
		C	1096
B	757*	A	896*
B	455*	A	496*
C	98*	A	472
B	454	C	509
C	187	A	495
C	438	A	494
C	383	B	75
C	382	A	897
C	381	C	713
A	549	A	329
C	380	A	471
B	1134	A	493
B	1135	A	330
A	227	C	8
A	548	A	328
B	453	A	327
		B	575
		A	491
		B	76

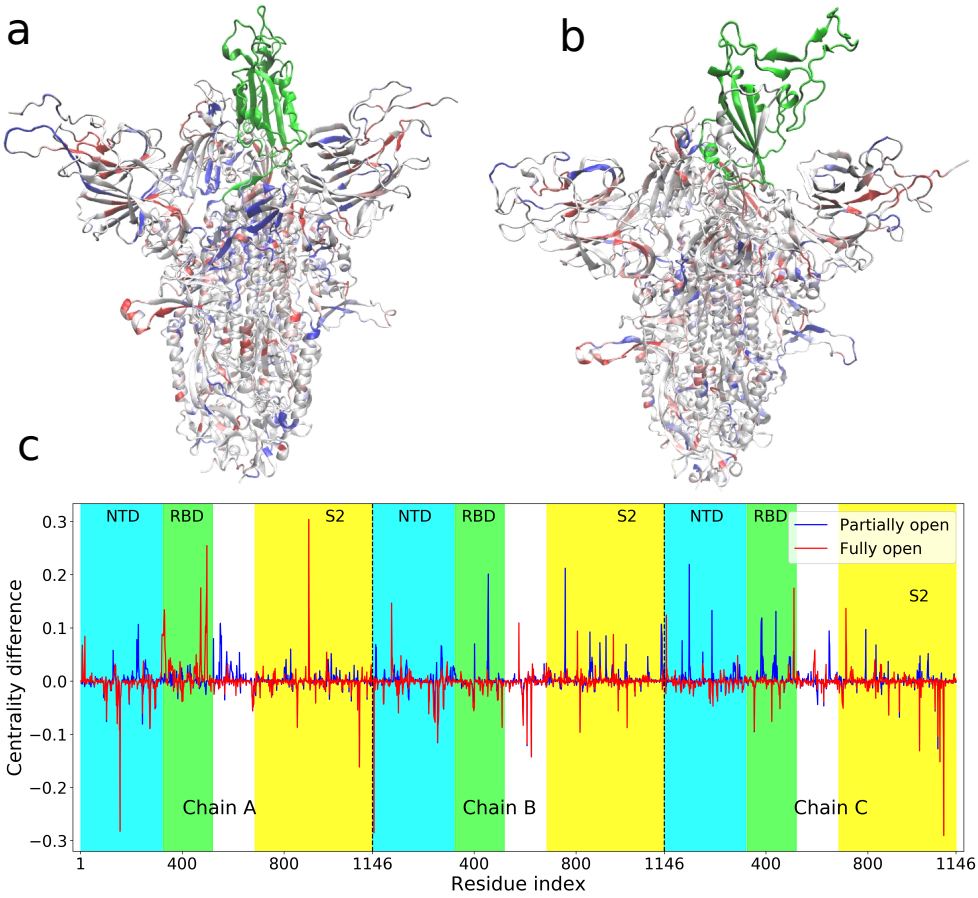

Figure 4: Structures of the (a) partially open and the (b) fully open state of the spike protein with residues colored according to the difference of betweenness centrality (BC) with respect to the closed conformation. Red indicates most negative and blue indicates most positive. The RBD of chain A, that undergoes opening motion, is colored in green. (c) The values of the difference in centrality with respect to closed state plotted as a function of residue indices.

Sequence alignment and analysis

To understand the mutability of the most important dynamically-coupled regions, we performed sequence alignment of the 67 sequences of the SARS-CoV2 spike protein from the Protein Data Bank. Analysis of the alignment using the ConSurf server[51] indicates that most residues of the spike protein are highly conserved. In agreement with previous studies,[18,52] the variable residues are primarily concentrated in the receptor binding domain (Fig. 5). Most glycosylations[53] are found in the conserved regions of the protein, except for N1134. All

of the strongly correlated residues predicted from our dihedral angle analysis are highly conserved (Gln613, Pro600, Gly601, Asn137, Ser112, Cys136, Phe833, Ile834, Ile569, His1083, Asp1084), except for Asp614. Most of the non-RBD residues involved in the allosteric pathway of RBD opening are also highly conserved. The high level of conservation reinforces their important role in the functioning of the spike. Possible mutations in these residues can potentially result in new strains of higher infectivity or virulence in the near future.

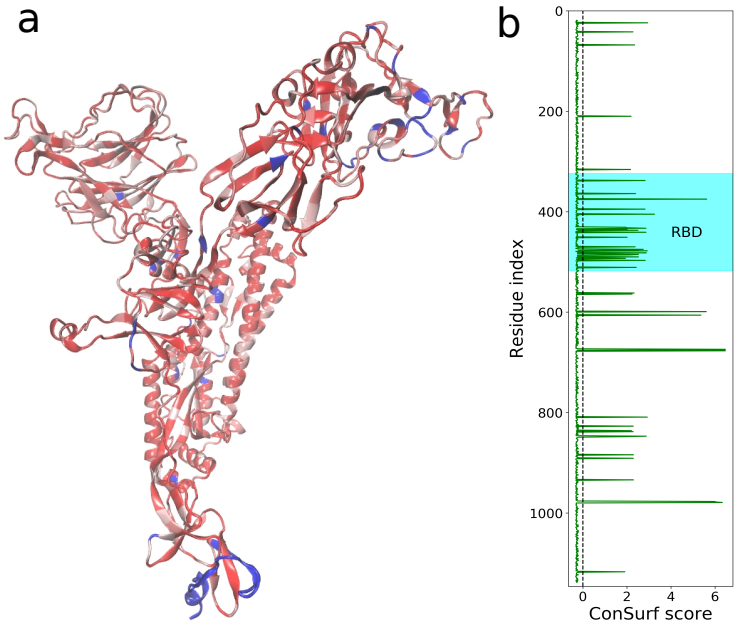

Figure 5: (a) The structure of the spike monomer with residues colored according to their degrees of variability over different strains. The red color suggests most conserved and blue indicates most variable regions. (b) The ConSurf score for all residues. The region corresponding to the RBD is denoted in cyan. The lower (more negative) the ConSurf score, the more conserved are the residues, and vice versa.

Concluding Discussion

To tame this raging pandemic, we need to able to control the fundamental dynamics of the spike protein. Its motion is key to the infection machinery of the SARS-CoV-2 virus. By understanding the role of residues in its structure, we can anticipate new mutations and customize treatments ahead of time. To curb the negative impact of mutations, we here

focused on the allosteric effect of protein residues that are *away* from its rapidly-mutating epitope (the RBD) on the conformational change needed for infection.

We performed molecular dynamics simulations, with unsupervised machine learning (tICA) and graph theory-based analysis to identify the role of physically distant residues in the dynamics of the receptor binding domain in the SARS-CoV-2 spike protein. We correlated the protein backbone torsion angles with the slowest degrees of freedom encompassing the structural transition of the receptor binding domain. With this approach we were able to elucidate a small number of distant non-RBD residues which strongly influence the conformational change of the spike, change that in turn leads to binding to the ACE2 receptor and then to infection. Residues in the linker between the RBD and the S2 stem work as a hinge by driving the down-to-up RBD transition via backbone torsional changes. Out of the most correlated residues, D614 ranks close to the top. The D614G mutation is currently observed in SARS-CoV-2, and is becoming widespread among infected patients throughout the world. Our model predicted the D614 residue as a key player in RBD dynamics from physics-based atomistic simulations without any prior knowledge of the mutation profile of the spike. With glycine being more flexible in its backbone torsion compared to aspartate, this mutation, according to our hypothesis, will facilitate the attainment of the partially open state transiting from the closed structure. Another mutation, A570V, was observed within our predicted residues, but did not yet become as widespread as D614G, likely because it did not have substantial evolutionary advantage. However, a different mutation at residue 570 (A570D), has indeed appeared in the recently-emerged more-contagious B.1.1.7 strain of SARS-CoV-2.[31] The consistency with the mutation profile confirms that our dihedral-angle based analysis can not only find out distant residues impacting RBD dynamics, but can also predict future mutations that can increase infection capability.

A cross-correlation metric based on linear mutual information (LMI) was also employed to understand the long-distance coupled motion between RBD and non-RBD residues. The change in LMI correlation primarily takes place in the residues adjacent to the RBD, but

we could also distinctly observe long-distance effects. Betweenness centrality (BC) of each residue of the spike was computed from a C_α-based graph network model for all three conformational states. The residues showing largest changes in the BC are concentrated in NTD, RBD and also in the linker regions joining the RBD with the rest of the protein and also in the S2 domain. This reinforces the conclusion that RBD dynamics is in fact *dominated* by long-distance allosteric effects in the spike protein, emmerging as a result of the collective internal fluctuations of the amino-acid residues.

From the point of view of immediate therapeutic interventions, this study opens up the possibility of designing inhibitors that bind to the regions outside the RBD, thereby preventing infection by freezing RBD dynamics via steric restrictions on specific distant residues. Such treatments are less likely to be affected by the evolutionary adaptations in RBD sequence that the virus performs frequently to evade the immune response. In a broader context, we pinpoint to future mutations in these key residues that can change the infection rate and virulence, giving rise to new strains and significantly altering the course of the pandemic. Our study and future work in this direction can make the scientific community better prepared for such scenarios and can help in efficient prevention of future outbreaks.

Methods

The details of molecular dynamics simulations, tICA analysis, and mutual information based network analysis are provided in the SI Appendix . A brief outline is included below.

System preparation and simulation details

Glycosylated and solvated structures of closed (PDB ID: 6VXX[24]) and partially open (PDB ID: 6VSB[26]) spike head trimers were obtained from the CHARMM-GUI COVID-19 archive.[54] All simulations were performed using the CHARMM36m force field.[55] For the purpose of the current work we considered residues 324-518 as the RBD. After minimization and short

equilibration, steered molecular dynamics (SMD) simulation were performed to induce the opening of the closed state and closing of the partially open state of the RBD. Reaction coordinates (RC) were chosen to represent the distance of the RBD from its closed state position. Multiple structures were chosen from the two SMD trajectory and two independent umbrella sampling (US) simulations were performed for PDB ID: 6VXX and 6VSB. The reaction coordinate was restrained by a harmonic potential with force constant of 1 kcal/mol/Å^2 for 35 windows and 26 windows for the two sets respectively. The colvars module[56] was used for SMD and US calculations. Free energy profiles were computed using weighted histogram analysis method (WHAM).[57]

Unbiased simulations

Umbrella sampling trajectory frames were sampled from the regions near the open, closed and the partially open intermediate state judging the free energy value. Unbiased simulations were performed starting from these frames, resulting in 38 trajectories, each 40 ns long. Three of these trajectories were identified as stable conformations corresponding to the closed, partially open and fully open structure. These three trajectories were extended to 80 ns. A cumulative $\sim$1.6 μs unbiased simulation data were generated and used in subsequent analysis.

Time-independent component analysis and mutual information

Time-independent component analysis (tICA) and principal component analysis (PCA) were performed using pyEMMA package[58] on the entire unbiased trajectory data. The feature space for PCA and tICA consisted of pairwise distances between specific residues in and around RBD of chain A and nearby static regions.

The linear mutual information (LMI) based correlation was computed for the three 80 ns trajectories. A graph theory based network model was constructed with the C_α atom of each residue as node. The edge length between nodes were computed from the cross correlation

values using previously described procedure.[43] Betweenness centrality of each residue was computed for each of the three trajectories and compared. All the LMI and network analysis were performed using bio3D package.[59]

Sequence alignment

Iterative sequence alignment of the 67 strains of SARS-CoV-2 spike protein sequences from the RCSB PDB database was performed using the MAFFT-DASH program[60] using the G-INS-i algorithm. The sequence of PDB ID: 6VXX was used as the template. The alignment was analyzed with the ConSurf server[51] to derive conservation scores for each residue position in the alignment.

Author contribution

DR designed research with inputs from LL and IA. DR performed MD simulations and analyzed results. LL performed and analyzed sequence alignment. DR, LL and IA wrote the paper.

Competing Interest

The authors declare no competing financial interest.

Data availability

The molecular dynamics trajectories are available from `https://doi.org/10.5281/zenodo.4310268`. The codes and the residue correlation data used in this study are available from `https://github.com/dhimanray/COVID-19-correlation-work.git`. All further details about the methods and the data are available within the article and the SI Appendix.

Acknowledgement

This work was supported by the National Science Foundation (NSF) via grant MCB 2028443. DR acknowledges support by the Molecular Science Software Institute (MolSSI) seed fellowship funded by NSF via grant number OAC-1547580. The authors thank Trevor Gokey for a stimulating discussion. The work has benefited from the computational resources of the UC Irvine High Performance Computing (HPC) cluster.

Supporting Information Available

Please see the SI Appendix for simulation details, results of steered MD and umbrella sampling, and figures S1-S11

References

(1) Zhou, P.; Yang, X.-L.; Wang, X.-G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H.-R.; Zhu, Y.; Li, B.; Huang, C.-L., et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **2020**, 1–4.

(2) Lu, R.; Zhao, X.; Li, J.; Niu, P.; Yang, B.; Wu, H.; Wang, W.; Song, H.; Huang, B.; Zhu, N., et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* **2020**, *395*, 565–574.

(3) Ou, X. et al. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nature Communications* **2020**, *11*, 1–12.

(4) Tai, W.; He, L.; Zhang, X.; Pu, J.; Voronin, D.; Jiang, S.; Zhou, Y.; Du, L. Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. *Cellular and Molecular Immunology* **2020**, *17*, 613–620.

(5) Koyama, T.; Platt, D.; Parida, L. Variant analysis of SARS-cov-2 genomes. *Bulletin of the World Health Organization* **2020**, *98*, 495.

(6) Zhao, Z.; Li, H.; Wu, X.; Zhong, Y.; Zhang, K.; Zhang, Y. P.; Boerwinkle, E.; Fu, Y. X. Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC Evolutionary Biology* **2004**, *4*, 21.

(7) van Dorp, L.; Acman, M.; Richard, D.; Shaw, L. P.; Ford, C. E.; Ormond, L.; Owen, C. J.; Pang, J.; Tan, C. C.; Boshier, F. A.; Ortiz, A. T.; Balloux, F. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infection, Genetics and Evolution* **2020**, 104351.

(8) Kaushal, N.; Gupta, Y.; Goyal, M.; Khaiboullina, S. F.; Baranwal, M.; Verma, S. C. Mutational frequencies of SARS-CoV-2 genome during the beginning months of the outbreak in USA. *Pathogens* **2020**, *9*, 565.

(9) Laha, S.; Chakraborty, J.; Das, S.; Manna, S. K.; Biswas, S.; Chatterjee, R. Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission. *Infection, Genetics and Evolution* **2020**, *85*, 104445.

(10) Isabel, S.; Graña-Miraglia, L.; Gutierrez, J. M.; Bundalovic-Torma, C.; Groves, H. E.; Isabel, M. R.; Eshaghi, A. R.; Patel, S. N.; Gubbay, J. B.; Poutanen, T.; Guttman, D. S.; Poutanen, S. M. Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide. *Scientific Reports* **2020**, *10*, 14031.

(11) Rogers, T. F. et al. Isolation of potent SARS-CoV-2 neutralizing antibodies and protection from disease in a small animal model. *Science* **2020**, *369*, 956–963.

(12) Li, W.; Chen, C.; Drelich, A.; Martinez, D. R.; Gralinski, L. E.; Sun, Z.; Schäfer, A.; Kulkarni, S. S.; Liu, X.; Leist, S. R., et al. Rapid identification of a human antibody with high prophylactic and therapeutic efficacy in three animal models of SARS-CoV-2 infection. *Proceedings of the National Academy of Sciences* **2020**, *117*, 29832–29838.

(13) Zost, S. J. et al. Potently neutralizing and protective human antibodies against SARS-CoV-2. *Nature* **2020**, *584*, 443–449.

(14) Jiang, S.; Zhang, X.; Du, L. Therapeutic antibodies and fusion inhibitors targeting the spike protein of SARS-CoV-2. *Expert Opinion on Therapeutic Targets* **2020**, *0*, 1–7, PMID: 32941780.

(15) Smith, M.; Smith, J. C. Repurposing Therapeutics for COVID-19: Supercomputer-Based Docking to the SARS-CoV-2 Viral Spike Protein and Viral Spike Protein-Human ACE2 Interface. *ChemRxiv* **2020**,

(16) Chowdhury, R.; Adury, V. S. S.; Vijay, A.; Singh, R. K.; Mukherjee, A. Generate, Repurpose, Validate: A Receptor-Mediated Atom-by-Atom Drug Generation for SARS-Cov-2 Spike Protein and Similarity-Mapped Drug Repurposing for COVID-19 with Rigorous Free Energy Validation Using Well-Tempered Metadynamics. *ChemRxiv* **2020**,

(17) Han, Y.; Král, P. Computational Design of ACE2-Based Peptide Inhibitors of SARS-CoV-2. *ACS Nano* **2020**, *14*, 5143–5147, PMID: 32286790.

(18) Verkhivker, G. M. Molecular Simulations and Network Modeling Reveal an Allosteric Signaling in the SARS-CoV-2 Spike Proteins. *Journal of proteome research* **2020**, *19*, 4587–4608.

(19) Greaney, A. J.; Loes, A. N.; Crawford, K. H.; Starr, T. N.; Malone, K. D.; Chu, H. Y.; Bloom, J. D. Comprehensive mapping of mutations to the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human serum antibodies. *bioRxiv* **2021**,

(20) Liu, L.; Wang, P.; Nair, M. S.; Yu, J.; Rapp, M.; Wang, Q.; Luo, Y.; Chan, J. F.-W.; Sahi, V.; Figueroa, A., et al. Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature* **2020**, *584*, 450–456.

(21) Ng, K. W. et al. Preexisting and de novo humoral immunity to SARS-CoV-2 in humans. *Science* **2020**, *370*, 1339–1343.

(22) Di Paola, L.; Hadi-Alijanvand, H.; Song, X.; Hu, G.; Giuliani, A. The discovery of a putative allosteric site in the SARS-CoV-2 spike protein using an integrated structural/dynamic approach. *Journal of proteome research* **2020**, *19*, 4576–4586.

(23) Olotu, F. A.; Omolabi, K. F.; Soliman, M. E. Leaving no stone unturned: Allosteric targeting of SARS-CoV-2 spike protein at putative druggable sites disrupts human angiotensin-converting enzyme interactions at the receptor binding domain. *Informatics in medicine unlocked* **2020**, *21*, 100451.

(24) Walls, A. C.; Park, Y.-J.; Tortorici, M. A.; Wall, A.; McGuire, A. T.; Veesler, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **2020**, *181*, 281 – 292.e6.

(25) Cai, Y.; Zhang, J.; Xiao, T.; Peng, H.; Sterling, S. M.; Walsh, R. M.; Rawson, S.; Rits-Volloch, S.; Chen, B. Distinct conformational states of SARS-CoV-2 spike protein. *Science* **2020**, *369*, 1586–1592.

(26) Wrapp, D.; Wang, N.; Corbett, K. S.; Goldsmith, J. A.; Hsieh, C.-L.; Abiona, O.; Graham, B. S.; McLellan, J. S. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **2020**, *367*, 1260–1263.

(27) Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S.; Fadda, E.; Amaro, R. E. Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Central Science* **2020**, *6*, 1722–1734.

(28) Zimmerman, M. I. et al. SARS-CoV-2 Simulations Go Exascale to Capture Spike Opening and Reveal Cryptic Pockets Across the Proteome. *bioRxiv* **2020**,

(29) Roy, S.; Jaiswar, A.; Sarkar, R. Dynamic Asymmetry Exposes 2019-nCoV Prefusion Spike. *The journal of physical chemistry letters* **2020**, *11*, 7021–7027.

(30) Korber, B. et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **2020**,

(31) Rambaut, A.; Loman, N.; Pybus, O.; Barclay, W.; Barrett, J.; Carabelli, A.; Connor, T.; Peacock, T.; Robertson, D. L.; Volz, E. *Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations*; 2020.

(32) Jolliffe, I. *Principal Component Analysis*, 2nd ed.; Springer-Verlag: New York, 2002; p 488.

(33) Molgedey, L.; Schuster, H. G. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters* **1994**, *72*, 3634–3637.

(34) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *The Journal of Chemical Physics* **2013**, *139*, 015102.

(35) Caves, L. S.; Evanseck, J. D.; Karplus, M. Locally accessible conformations of proteins: Multiple molecular dynamics simulations of crambin. *Protein Science* **1998**, *7*, 649–666.

(36) Echeverria, I.; Makarov, D. E.; Papoian, G. A. Concerted dihedral rotations give rise to internal friction in unfolded proteins. *Journal of the American Chemical Society* **2014**, *136*, 8708–8713.

(37) Levinthal, C. How to fold graciously. *Mossbauer spectroscopy in biological systems* **1969**, *67*, 22–24.

(38) Eckmann, J.-P.; Ruelle, D. Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics* **1985**, *57*, 617–656.

(39) Fadel, A. R.; Jin, D. Q.; Montelione, G. T.; Levy, R. M. Crankshaft motions of the polypeptide backbone in molecular dynamics simulations of human type-α transforming growth factor. *Journal of biomolecular NMR* **1995**, *6*, 221–226.

(40) Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX web server: an online force field. *Nucleic Acids Research* **2005**, *33*, W382–W388.

(41) Rapid increase of a SARS-CoV-2 variant with multiple spike protein mutations observed in the United Kingdom. `https://www.ecdc.europa.eu/sites/default/files/documents/SARS-CoV-2-variant-multiple-spike-protein-mutations-United-Kingdom.pdf`.

(42) Kemp, S.; Datir, R.; Collier, D.; Ferreira, I.; Carabelli, A.; Harvey, W.; Robertson, D.; Gupta, R. Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion ΔH69/ΔV70. *bioRxiv* **2020**, 2020.12.14.422555.

(43) Rivalta, I.; Sultan, M. M.; Lee, N.-S.; Manley, G. A.; Loria, J. P.; Batista, V. S. Allosteric pathways in imidazole glycerol phosphate synthase. *Proceedings of the National Academy of Sciences* **2012**, *109*, E1428–E1436.

(44) Negre, C. F.; Morzan, U. N.; Hendrickson, H. P.; Pal, R.; Lisi, G. P.; Loria, J. P.; Rivalta, I.; Ho, J.; Batista, V. S. Eigenvector centrality for characterization of protein allosteric pathways. *Proceedings of the National Academy of Sciences* **2018**, *115*, E12201–E12208.

(45) McClendon, C. L.; Friedland, G.; Mobley, D. L.; Amirkhani, H.; Jacobson, M. P. Quantifying correlations between allosteric sites in thermodynamic ensembles. *Journal of chemical theory and computation* **2009**, *5*, 2486–2502.

(46) McClendon, C. L.; Kornev, A. P.; Gilson, M. K.; Taylor, S. S. Dynamic architecture of a protein kinase. *Proceedings of the National Academy of Sciences* **2014**, *111*, E4623–E4631.

(47) Ahuja, L. G.; Aoto, P. C.; Kornev, A. P.; Veglia, G.; Taylor, S. S. Dynamic allostery-based molecular workings of kinase: peptide complexes. *Proceedings of the National Academy of Sciences* **2019**, *116*, 15052–15061.

(48) Bowerman, S.; Wereszczynski, J. Detecting allosteric networks using molecular dynamics simulation. *Methods in Enzymology* **2016**, *578*, 429–447.

(49) Ichiye, T.; Karplus, M. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Structure, Function, and Bioinformatics* **1991**, *11*, 205–217.

(50) Lange, O. F.; Grubmüller, H. Generalized correlation for biomolecular dynamics. *Proteins: Structure, Function, and Bioinformatics* **2006**, *62*, 1053–1061.

(51) Ashkenazy, H.; Abadi, S.; Martz, E.; Chay, O.; Mayrose, I.; Pupko, T.; Ben-Tal, N. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic acids research* **2016**, *44*, W344–W350.

(52) Flores-Alanis, A.; Sandner-Miranda, L.; Delgado, G.; Cravioto, A.; Morales-Espinosa, R. The receptor binding domain of SARS-CoV-2 spike protein is the result of an ancestral recombination between the bat-CoV RaTG13 and the pangolin-CoV MP789. *BMC research notes* **2020**, *13*, 1–6.

(53) Shajahan, A.; Supekar, N. T.; Gleinich, A. S.; Azadi, P. Deducing the N- and O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. *Glycobiology* **2020**, *30*, 981–988.

(54) Woo, H.; Park, S.-J.; Choi, Y. K.; Park, T.; Tanveer, M.; Cao, Y.; Kern, N. R.; Lee, J.; Yeom, M. S.; Croll, T. I.; Seok, C.; Im, W. Developing a Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein Model in a Viral Membrane. *The Journal of Physical Chemistry B* **2020**, *124*, 7128–7137, PMID: 32559081.

(55) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods* **2017**, *14*, 71–73.

(56) Fiorin, G.; Klein, M. L.; Hénin, J. Using collective variables to drive molecular dynamics simulations. *Molecular Physics* **2013**, *111*, 3345–3362.

(57) Roux, B. The calculation of the potential of mean force using computer simulations. *Computer Physics Communications* **1995**, *91*, 275–282.

(58) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *Journal of Chemical Theory and Computation* **2015**, *11*, 5525–5542.

(59) Grant, B. J.; Rodrigues, A. P.; ElSawy, K. M.; McCammon, J. A.; Caves, L. S. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* **2006**, *22*, 2695–2696.

(60) Rozewicki, J.; Li, S.; Amada, K. M.; Standley, D. M.; Katoh, K. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic acids research* **2019**, *47*, W5–W10.