Distant Residues Modulate the 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*

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

E-mail: andricio@uci.edu

Keywords

SARS-CoV-2, COVID-19, spike protein, molecular dynamics, tICA, allosteric correlation, network model, mutation

Abstract

Infection of human cells by the novel coronavirus (SARS-Cov-2) involves the attachment of the receptor binding domain (RBD) of the spike protein to the peripheral membrane ACE2 receptors. The process is initiated by a *down* to *up* conformational change in the spike presenting the RBD to the receptor. Early stage computational and experimental studies on potential therapeutics have concentrated on the receptor binding domain, although this region is prone to mutations with the possibility of giving rise to widespread drug resistance. Here, using atomistic molecular dynamics simulation, we study the correlations between the RBD dynamics with physically distant residues in the spike protein, and provide a deeper understanding of their role in the infection, including the prediction of important mutations and of distant allosteric binding sites for therapeutics. Our model, based on time-independent component analysis (tICA) and protein graph connectivity network, was able to identify multiple residues, exhibiting long-distance coupling with the RBD opening dynamics. Mutation on these residues can lead to new strains of coronavirus with different degrees of infectivity and virulence. The most ubiquitous D614G mutation is predicted *ab-initio* from our model. Conversely, broad spectrum therapeutics like drugs and monoclonal antibodies can be generated targeting these key distant regions of the spike protein.

Significance statement

The novel coronavirus SARS-CoV-2 has created the biggest pandemic of 21st century resulting in significant economic and public health crisis. Significant research effort for drug design against COVID-19 is focused on the receptor binding domain of the spike protein, although this region is prone to mutations causing resistance against therapeutics. We applied time-independent component analysis (tICA) and protein connectivity network model, on all-atom molecular dynamics trajectories, to identify key non-RBD residues, playing crucial role in the conformational transition facilitating spike-receptor binding and infection of hu-

man cell. These residues can not only be targeted by broad spectrum antibodies and drugs, mutations in them can generate new strains of coronavirus resulting in future epidemic.

Introduction

Emerged in China last December, the COVID-19 pandemic currently continues its rapid spread across the world, with more than 53 million confirmed cases and about 1.3 million deaths, according to the World Health Organization (WHO) report on 14 November 2020. The etiological agent, SARS-CoV-2, is a member of the *Coronaviridae* family including SARS-CoV-1 (2002-2004) and MERS-CoV (since 2012), with the sequence identity of 79.6% and 50%, respectively.[1,2] Largely expressed on the surface of SARS-CoV-2, the spike (S) protein plays a crucial role in binding to the host angiotensin-converting enzyme 2 (ACE2) through the receptor-binding domain (RBD) and facilitating viral entry,[3,4] which is therefore considered as one of the most preferred targets against SARS-CoV-2. Assessing genomic variability of SARS-CoV-2 presents a moderate mutation rate compared to other RNA viruses (around 1.12×10^{-3} nucleotide substitution/site/year),[5] which is at the same level as SARS-CoV-1.[6] Significantly, the sites in S protein have been demonstrated to be vulnerable to acquire 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.[5] It suggests a strong possibility of generating new strains of this coronavirus and its family with higher virulence and more complicated epidemiology with a D614G mutant as a notable example.[10] This could make therapeutic agents currently being developed ineffective in combating SARS-CoV-2 and other probable SARS epidemics in the future.

Large scale screening of therapeutic molecules and antibodies are underway aiming to target the spike protein and consequently prevent infection. Majority of both experimental[11-14] and computational[15-17] efforts for inhibitor design are focused on the receptor binding domain (RBD) despite the fact that this region is highly mutation prone.[18] The human im-

mune system started generating antibodies specific to residues outside RBD even at the earlier stage of the pandemic. Liu et al. extracted multiple COVID-19 neutralizing antibodies from infected patients, a fraction of which binds to non-RBD epitopes of the S-protein such as the N-terminal domain (NTD).[19] Moreover, a separate group of antibodies, present in the bloodstream of uninfected human (particularly children), were observed to bind specifically to the S2-stem region which has virtually identical sequence in all coronavirus strains, including the ones causing common cold.[20] These results motivated the effort towards designing small molecule drugs and antibodies targeted towards residues which are far from RBD in the 3D structure and are consequently less prone to mutation.

A number of studies explored the possibility of the appearance of druggable hotspots in the spike protein due to the allosteric effect of ACE2 binding.[21,22] Although this is a significant step towards targeting the non-RBD residues, the S-ACE2 binding already initiates the infection process. So inhibitors which bind to the ACE2 bound spike are only partially effective in preventing the viral entry to the cell. 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 conformation. We refer to this process as the RBD opening, which plays a key role in the infection by displaying the RBD to the ACE2 receptor (Fig 1). Therapeutics inhibiting this structural transition will prevent ACE2 binding altogether providing higher degree of barrier towards the infection.

The process of *down* to *up* transition of the S-protein RBD has been subjected to extensive cryogenic electron microscopy (cryo-EM) studies leading to the elucidation of the atomistic resolution structures of the closed (PDB ID: 6VXX,[23] 6XR8[24]), partially open (PDB ID: 6VSB[25]) and fully open (PDB ID: 6VYB[23]) states. Attempts have been made to study the dynamics of RBD opening using force field based classical molecular dynamics (MD) simulation.[26,27] Although each computational study provided key insight on different aspects, an RBD opening transition could not be observed directly using a continuous atomistic MD simulation with a solvated spike. This is primarily because of the large size of the

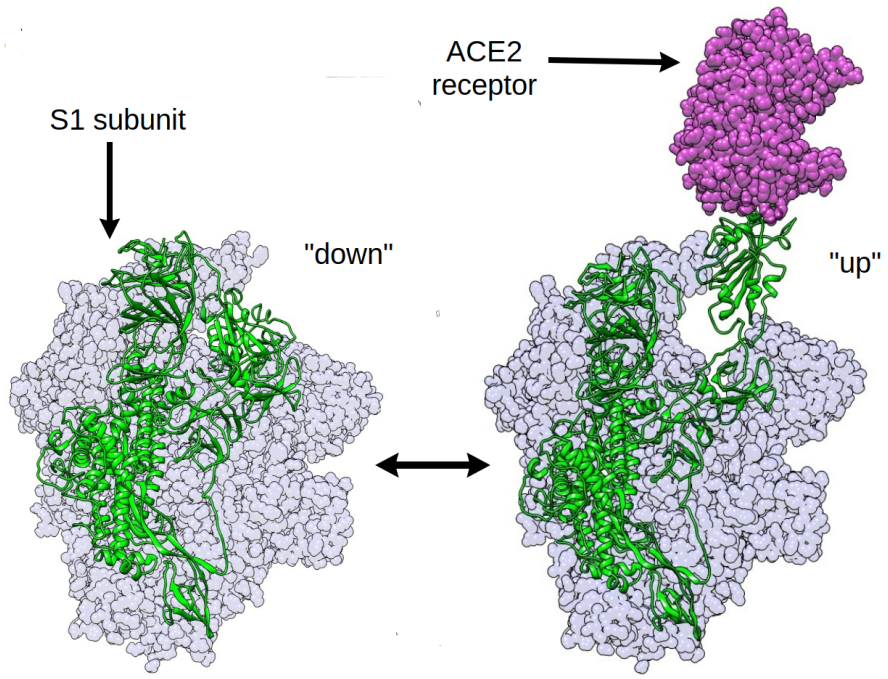

Figure 1: The RBD opening transition in the SARS-CoV-2 spike protein: the RBD of the chain shown in green is undergoing the *down* to *up* transition leading to the binding to the human ACE2 receptor.

system and the slow timescales involved. Roy et. al. used structure based coarse grained modeling (using only the backbone C_α structure of the protein) in implicit solvent system to study the dynamics of RBD opening and closing.[28] This reduced the number of degrees of freedom by many orders of magnitude, and allowed the authors 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). Coarse grained models have also been used by Verkhivker to identify allosteric communication pathways in the spike protein.[18] Although these studies provided considerable qualitative insight, the details of solvent and protein side-chain interactions were missed due the approximate nature of their model. Also the glycan-shield, which has recently been proven to have dynamics effects apart from just shielding,[26] was also not taken into account.

Apart form the fact that direct all atom MD simulation of RBD opening is challenging, the observation of a single conformational transition event hardly provides enough statistical

information to draw meaningful conclusions. So we took a different approach to identify the distant residues which shows correlated motion coupled to the dynamics of RBD opening and closing. First, we obtained the underlying free energy profile of the RBD dynamics and performed multiple short unbiased simulations representing different regions of the configurations space. We proposed a novel approach to identify residues important for the conformational change by quantifying the correlation of the backbone torsion angles of the protein with slowest degrees of freedom representing the down to up transition of RBD. Second, we took a more traditional route to study allosteric connections by constructing a dynamical network model by using the mutual information metric computed from longer trajectories. Both approaches resulted in the prediction of a handful of residues in specific regions of the S protein that may play crucial role in the spike protein RBD dynamics. 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.

Free energy profile of RBD opening

We performed extensive umbrella sampling simulation to obtain the underlying free energy landscape of the *down* to *up* transition of the recepetor binding domain. Roy et al. observed that this transition in SARS-CoV-2 spike is a complex multidimensional process with unique intermediates different from SARS-CoV-1 and MERS.[28] They characterized an unusual Proline-proline interaction (P230-P521) between the RBD and the N-terminal domain which stabilizes the intermediate partially open aka "NTD-assisted up" state. In consistency with their finding, our enhanced sampling simulations could not distinguish between three different states based on single geometric reaction coordinate. Following the work by Folding@Home group,[27] our first set of simulations starting from the RBD down structure (PDB ID: 6VXX) used the distance of the center of mass of the RBD from the same in the closed

state, as the RC. We could obtain only one deep free energy minima which corresponds to the closed state (Fig 2a). The position of the minima is in agreement with the most probable structure predicted in Ref.[27] The small 3 Å separation of the free energy minimum from the cryo-EM structure is the representative of the most stable state for the given force field. The free energy profile is virtually identical to the results of Ref[28] when RBD-NTD interactions were artificially removed. It indicates the all atom enhanced sampling simulation starting from RBD down state is unable to reach the partially open configuration within reasonable amount of computational time. To sample the conformations around the NTD-assisted up state of the RBD, we performed another set of enhanced sampling simulation starting from the experimental structure of the partially open intermediate. The underlying free energy profile qualitatively resembles the coarse grained MD results of Ref[28] but with fewer approximations. Unlike the previous work[28] our free energy profile, along a inter-residue distance based RC, indicates a lower free energy of the partially open state compared to states which are closer to the RBD-down configuration (Fig. 2b). This is not entirely unexpected as the closed-like state obtained from the second set of simulations does not relax to the true closed state, indicated by insufficient increase of the Pro-pro distance. This is not a problem as obtaining the true underlying free energy landscape is not the goal of the present work. We used the umbrella sampling trajectories to seed multiple unbiased simulations at and near the closed, partially open, and fully open states.

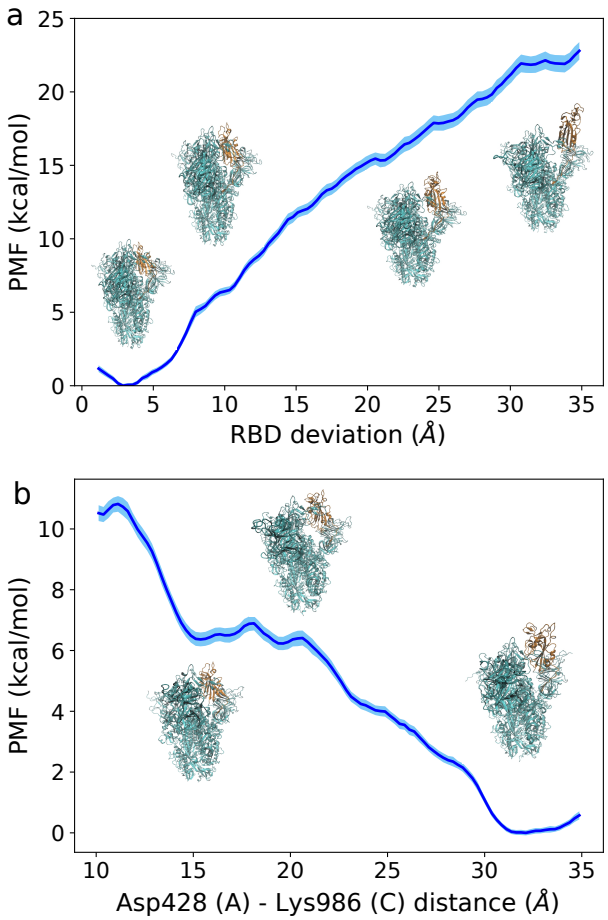

Figure 2: Free energy profile along the RBD opening coordinate for (a) Set 1 (starting from PDB ID: 6VXX) and (b) Set 2 (starting from PDB ID: 6VSB). For further details see Methods section and SI text. Representative structures are shown along on the PMF. The transiting RBD has been colored in orange whereas the rest of the spike is colored light blue.

Correlation between tIC and backbone dihedral angle

Multiple unbiased trajectories were propagated at different regions of the conformational space of RBD-opening. 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 (details in SI Appendix). The stability of these three conformations were ensured by manual inspection of the trajectory. 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 RBD (details in SI Appendix). The projected trajectory data was subjected to principal component analysis (PCA)[29] and time-independent component analysis (tICA).[30,31] The former method calculates the degrees of freedom with the highest variance in the system while latter obtains the ones with the longest timescale. These methods are widely used in the literature for quantifying large conformation changes in complex bio-molecules. The goal of performing PCA and tICA is to find out one or two coordinates which best describes the RBD opening motion. As one long trajectory hardly samples the transition event, multiple short trajectories spanning a large range of the configuration space were used.

The first principal component (PC) and the first time-independent component (tIC), obtained from PCA and tICA analysis respectively, could both distinguish the closed and open states, although the partially open and fully open states could not be distinguished with the first two PCs. Yet, the projection of the open and closed state trajectories along the first two principal components is in agreement with the results of long multi-microsecond spike protein simulation.[26] Because of the clear distinction between the closed, open and the intermediate partially open states (Fig 3b), we chose the first two time independent components (tICs) for our subsequent analysis.

Complex conformational changes in protein are, at a very fundamental level, characterized by complex combinations of transitions between various states in the backbone torsion angles ϕ and ψ. Much earlier, Levinthal used the backbone torsion angles to approximately gauge the number of conformational states accessible to a protein.[32] Consorted transition between one state to another, in the backbone torsional angle space, often leads to large scale motions in proteins leading to folding or change in secondary structure. So we hypothesize that there are specific residues in the spike protein, for which the transition in backbone dihedral angle states result the opening of the RBD. 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 correlation is found to be

significantly large only for a handful of torsion angles, whereas the majority show near zero correlation (Fig 3c).

We chose ten largest negative and largest positively correlated torsion angles (both sine and cosine) for tIC 1 and tIC 2. These provided us with 80 dihedral angles with multiple redundancies. Table 1 shows some of the residues which makes 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. It suggests that two consecutive torsion angles in certain regions of the protein are highest correlated with RBD opening motion. This correlation can also mean causation as change in the conformational state for two subsequent torsion angles create kink or bending in the backbone which propagate along the chains leading to change in the secondary and tertiary 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 3d. Similar plots for all 80 angles are provided in SI Appendix. Most of them belonged to the residues in the loop structure joining the RBD with the S2 stem as this region can work as a hinge for the opening of the RBD domain. The angles could distinguish one or both the open states from the RBD down configuration. Primarily, the closed and fully open state shared a similar region in the dihedral angle space different from the partially open structure. But we should refrain from over-interpreting the fully open state as this structure, which was generated from the closed state using steered molecular dynamics, can be different from the exact experimental RBD up structure that binds the ACE2 receptor. The correlated torsion angles span over all the chains, namely A (the one showing RBD transition), B and C.

The virus will also mutate these residues in a way to increase its virulence. A D614G mutation is already observed in numerous strains of the SARS-CoV-2 all over the world.[9] The residue D614 is one of the top ranked residues in our model for the potential to play a crucial role in RBD opening. A glysine residue has the least barrier for conformational transition in the ϕ-ψ space due to the absence of a side chain. Replacing an Asp residue,

which has higher barriers to such transition, with a glysine 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 mutation became so widespread among COVID-19 infected patients.

Researchers have characterized a wide range of spike protein mutant sequences observing different mutations with varying degrees of abundance. A relatively rare mutation was observed in the Ala570 (converting it to Val) although resulting in a decrease of the overall stability of the spike protein in all three state based on FoldX empirical force field.[9,33] The free energy values in Ref[9] were obtained from only structural data and no dynamical information has been considered. Yet it is worth noting that the change in total and solvation free energy, due to this mutation, were substantially different for the closed and open states, resulting in a change in ΔG of RBD opening. But, as the side chains of Ala and Val are similar in terms of steric bulk, this mutation is unlikely to have significant impact on the RBD dynamics. As it does not increase the evolutionary advantage of the virus by increasing infectivity, this mutation has only occurred in one strain[9] and did not become as prevalent as D614G. These results clearly indicate that mutations in the highest correlated residues (Table 1) can in fact have significant physiological impact and change the course of the pandemic.

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)

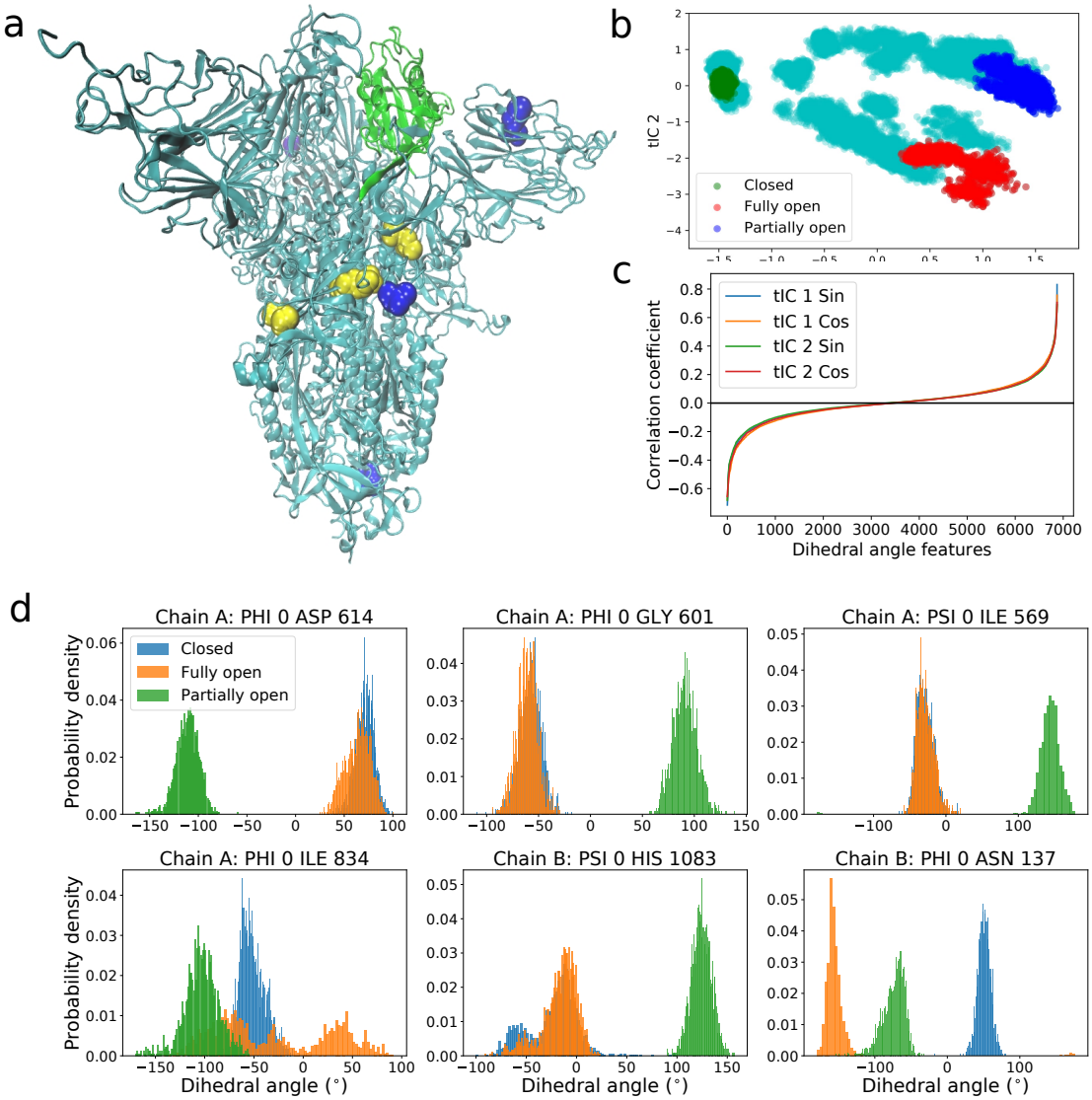

Figure 3: (a) A structure of the spike protein with the residues in RBD shown in green color. The non-RBD residues, strongly correlated with the RBD opening motion, are represented using spheres. Residues in chain A are colored yellow and those in chain B are colored blue. The RBD of chain A is performing *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.

Mutual information and network model

A more traditional approach to study the coupling between distant regions in a protein is to obtain the cross correlations between the positions of different residues in the 3D space. This method is widely used for studying allosteric effects in proteins due to ligand binding.[34-39] A conventional approach is to compute the dynamic cross correlation map (DCCM) of the position vectors of the C_α atoms.[40] But DCCM ignores correlated motions in the orthogonal directions.[39] The problem can be avoided by using a linear mutual information (LMI) based cross correlation metric which we use in the current study.[34,41] 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 as a Shannon type entropy function:

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

x_i, x_j are the Cartesian coordinates of the atoms whereas $p(x_i)$ and $p(x_i, x_j)$ indicates probability density of x_i and joint probability density of x_i and x_j.[34,41]

Strong correlations between distant residues is observed for the trajectories for all the three states (Fig 4). The change in the values of cross-correlation between apo and holo states of a protein is often utilized to trace the allosteric communication in proteins.[34] The RBD opening (or closing) is not strictly categorized as an allosteric process as there is no ligand binding involved. But change in the cross-correlation between RBD down and up states (similar to holo and apo structures) provides information about the long distance

residue connectivity which often plays an important role in facilitating the conformational change. That overall change in LMI correlation is clearly larger for the fully open state in comparison to the partially open state (Fig 4). 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. Residues in the range 524-600 show a gain in correlation in the partially open state unlike the fully open state. This result is consistent with the dihedral angle correlation, described in previous section, as the residues in the post RBD loop region exhibits 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 RBD's and NTD's of chain B and C which are in close proximity to the chain A RBD in the closed state. Additionally, residues 600-1000 of chain C show significant gain in correlation upon the opening transition. Majority of these residues are situated at the opposite end (S2 region) of the spike, indicating the presence of long range correlated motion.

To get a more detailed picture, we built a network model based on protein graph connectivity, considering the C_α atom of each amino acid as a node and the correlation between them as the edges between nodes. The total number or residues in our system is N=3438, higher than almost all other systems studied previously using this method.[34,39] The betweenness centrality (BC) is graph theoretical measure that provides a way to quantify the amount of information flow between nodes or edges of a network. If a node i is working as a bridge between two other nodes along the shortest path joining them, 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.[35] The change of BC in the dynamics of spike protein has been observed using coarse grained simulation.[18] Despite the limitation of the approximate

model, a handful of residues participating in the information propagation 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 in the allosteric information pathway. The difference in the normalized BC is measured by comparing the numbers 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 glycan shield.

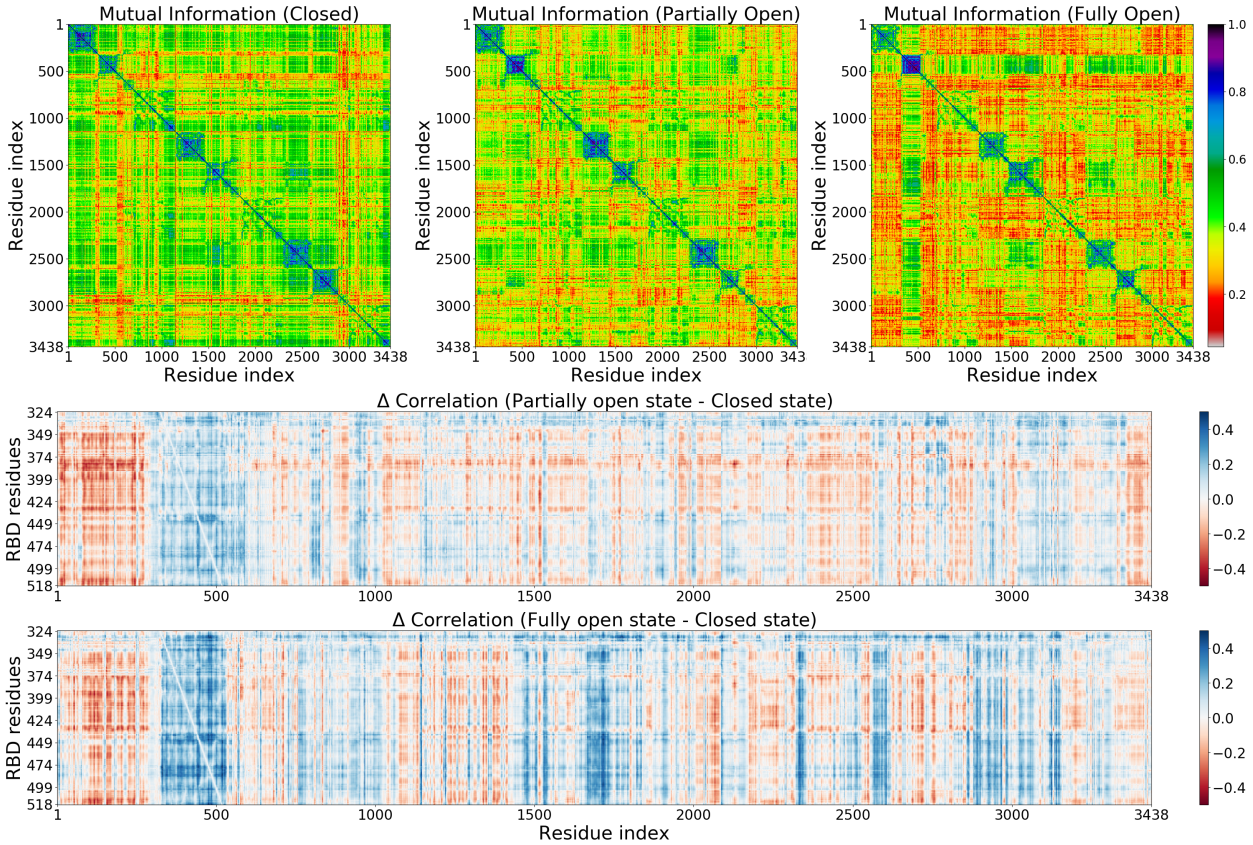

Figure 4: Upper panel: Cross correlation matrices for all three states computed using linear mutual information. RBD regions form highly correlated blocks indicating that the motion of these residues are largely decoupled from the rest of the protein. Still signatures of long distance correlated motion is detectable. Lower panel: Difference in correlation matrix elements for the partially and fully open states with respect to the closed state.

For both open and partially open state the residues which shows significant (>0.1) change in BC are from the NTD region or RBD region of the B and C chain. It shows that the

allosteric information flows through the nearby NTD's and RBDs, and mutations in this region can break the allosteric network[35] and affect the functionality of the spike protein. SARS-CoV-2 neutralizing antibodies were indeed observed to bind in these regions of the spike.[19] The identified residues are in proximity of the opening RBD of chain A, so the connectivity captured in the BC data is primarily due to the short range coupled fluctuations. Such coupling are broken when the RBD and NTD moves 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 it's internal vibrations becomes more independent of the rest of the protein.

But, most interesting aspect is the strikingly large change in BC of the residues which are distant from the RBD. 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 RBD dynamics as we already established from the dihedral angle analysis. The allosteric network analysis reinforces the same idea. Moreover, the large change of BC in S2 domain indicates a complex long range information flow connecting the RBD with the core residues of the protein. It has a major pharmaceutical implication as mutations inside S2 domain can potentially impact the receptor binding propensity of the viral spike. These results also suggest that therapeutics targeted to S2 and RBD-S2 linker can have significantly effect in preventing COVID-19 infection.

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.

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*
B	454	A	472
C	187	C	509
C	438	A	495
C	383	A	494
C	382	B	75
C	381	A	897
A	549	C	713
C	380	A	329
B	1134	A	471
B	1135	A	493
A	227	A	330
A	548	C	8
B	453	A	328
C	98*	A	327
		B	575
		A	491
		B	76

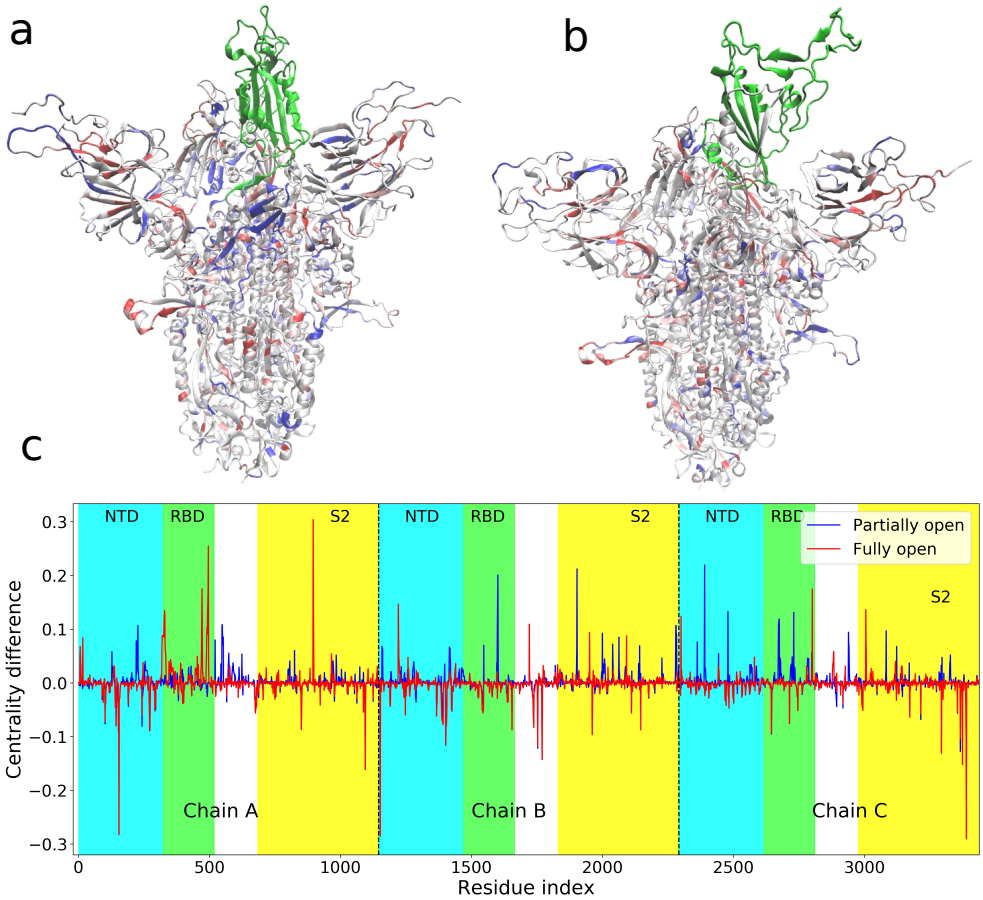

Figure 5: Structures of the (a) partially open and the (b) fully open state of the spike protein with residues colored according to the difference of betweenness centrality (BC) with respect to the closed conformation. Red indicates most negative and blue indicates most positive. The RBD of chain A, 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 bioinformatic analysis

To understand the mutability of the most important dynamically coupled regions, we performed sequence alignment of the 67 sequences of SARS-CoV2 spike protein from Protein Data Bank. Analysis of the alignment using the ConSurf server[42] indicates most residues of the spike protein are highly conserved. In agreement with previous study,[18,43] the variable residues are primarily concentrated in the receptor binding domain (Fig. 6). Most glycosilations[44] 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 of these residues 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.

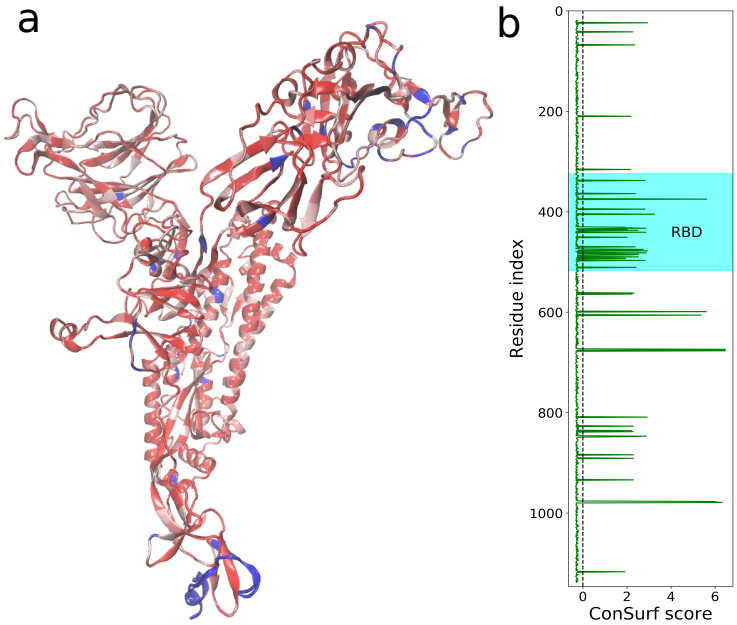

Figure 6: (a) The structure of the spike monomer with residues colored according to their degrees of variability over different strains. The red color suggests most conserved and blue indicates most variable regions. (b) The ConSurf score for all residues. The region corresponding to the RBD is denoted in cyan. The lower (more negative) the ConSurf score, the more conserved are the residues, and vice versa.

Concluding discussions

To control the pandemic raging across the world, we need a deeper understanding of the fundamental principles of the functioning of the spike protein, the key infection machinery of SARS-CoV-2 virus, so that we can expect new mutations and can customize treatments accordingly. This work focuses on the allosteric effect of protein residues of the spike on the

conformational change increasing the ability to infect human cells.

We performed molecular dynamics simulation and 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 proposed a novel approach of correlating backbone dihedral angles with slowest independent component, which could identify small number of non-RBD residues strongly influencing the conformational change of the spike. Such residues in the linker between RBD and the S2 stem can work as a hinge by driving the down to up transition by changing their backbone torsion angles. Out of the most correlated residues, D614 ranks close to the top. Indeed a D614G mutation is observed in SARS-CoV-2, which became widespread among the infected patients throughout the world. Our model predicts the D614 residue as a key player in RBD dynamics from physics based atomistic simulations without prior knowledge about the mutation profile of the spike. As glysine has more flexibility in its backbone torsion angles compared to Aspartate, this mutation, according to our hypothesis, will facilitate the attainment of the partially open state transiting from the closed structure. Another rare mutation A570V was observed within our predicted residues, but did not become as widespread as D614G, for not having substantial evolutionary advantage. The consistency with the genomic profile confirms that our dihedral angle based analysis can not only find out distant residues impacting RBD dynamics, but also predicts future mutations that can increase the infection capability of the virus.

A more conventional approach of using mutual information (LMI) based cross correlation metric 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 some long distance effects could still be observed. Betweenness centrality (BC) of each residue of the spike was computed from a protein C_α based graph network model for all three conformational states. The residues showing largest change 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 S2 domain. It reinforces the idea that RBD dynamics is

dominated by long distance allosteric effects in the spike protein. As our analysis is based entirely on correlation, the causality relations remain unknown. But it can be understood from common knowledge that the RBD opening motion is the result of the internal fluctuations of the protein like most protein conformational transitions.

From the point of view of immediate therapeutic application, this study opens up the possibility of designing inhibitors that bind to the regions outside RBD and can prevent infection by freezing the RBD dynamics by applying steric restrictions on the distant residues. Such treatments are unlikely to be affected by the evolutionary adaptations in RBD sequence that the virus performs to evade the immune response. On a broader context, future mutations in these key residues can change the infection rate and virulence, significantly altering the course of the pandemic. Our study and follow up 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 supporting information (SI) Appendix. A brief outline is included below.

Umbrella sampling

Glycosylated and solvated structures of closed (PDB ID: 6VXX[23]) and partially open (PDB ID: 6VSB[25]) spike head trimers were obtained from the CHARMM-GUI COVID-19 archive.[45] All simulations were performed using CHARMM36m force field.[46] 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 coor-

dinates (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 *closed* and the *partially open* set respectively. Free energy profiles were computed using weighted histogram analysis method (WHAM).[47]

Unbiased simulations

US 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[48] 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.[34] 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.[49]

Bioinformatics analysis

Iterative sequence alignment of the 67 strains of SARS-CoV-2 spike protein sequences from the RCSB PDB database has been performed using MAFFT-DASH program[50] using the G-INS-i algorithm. The sequence of PDB ID: 6VXX has been used as the template. The alignment was analyzed by ConSurf server[42] to derive conservation score 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 bioinformatics analysis. DR, LL and IA wrote the paper.

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 work has benefited from the computational resources of the UC Irvine High Performance Computing (HPC) cluster. The authors declare no competing financial interest.

Supporting Information Available

Please see the SI Appendix for simulation details and figures S1-S10

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**,

(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**,

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

(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**,

(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**,

(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) 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**,

(11) TF, R. Isolation of potent SARS-CoV-2 neutralizing antibodies and protection from disease in a small animal model. *Science* **2020**, *369*, 956.

(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) SJ, Z. Potently neutralizing and protective human antibodies against SARS-CoV-2. *Nature* **2020**, *584*, 443.

(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. **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. **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) 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.

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

(21) 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.

(22) 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.

(23) 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.

(24) 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.

(25) 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.

(26) 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.

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

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

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

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

(31) 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.

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

(33) 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.

(34) 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.

(35) 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.

(36) 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.

(37) 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.

(38) 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.

(39) Bowerman, S.; Wereszczynski, J. *Methods in enzymology*; Elsevier, 2016; Vol. 578; pp 429–447.

(40) 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.

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

(42) 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.

(43) 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.

(44) 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**, cwaa042.

(45) 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.

(46) 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.

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

(48) 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.

(49) 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.

(50) 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.

Supporting Information for: Distant Residues Modulate the Conformational Opening in SARS-CoV-2 Spike Protein

Dhiman Ray[1], Ly Le[1], and Ioan Andricioaei*[1,2]

[1] Department of Chemistry, University of California, Irvine, California 92697, USA

[2] Department of Physics and Astronomy, University of California, Irvine, California 92697, USA

*Correspondence: andricio@uci.edu

1 Computational Methods

1.1 System preparation and equilibration

The cryo-EM structures of the SARS-CoV-2 spike (S) protein in closed (PDB ID: 6VXX [1]) and partially open state (PDB ID: 6VSB [2]) were used in this work. Fully glycosylated and structurally complete (missing residues rectified) S protein head-only models (residue 1-1146) were obtained from the CHARMM-GUI COVID archive [3]. The two protein structures were obtained as pre-solvated in cubic water box of edge length 201 Å and 202 Å respectively. The ion concentration was maintained at 150 mM by including appropriate number of K^+ and Cl^- ions in the system. The total number of atoms in the closed and partially open S-protein systems were ∼762,000 and ∼773,000 respectively. All proteins and glycans were modelled using CHARMM36m [4] force field and a TIP3P model was used for the water.

Molecular dynamics simulations were performed using NAMD 2.14b2 package with CUDA acceleration [5]. Each structure was first minimized using conjugate gradient algorithm for 10000 steps which was followed by a short equilibration in NPT ensemble for 250 ps with 2 fs time. The temperature were controlled using a velocity rescaling thermostat and a langevin barostat was used to control the temperature and pressure at 310.15 K and 1 atm respectively. The systems were further equilibrated at the same temperature and pressure for 10 ns with hydrogen mass re-partitioning (HMR) [6] where the mass of each hydrogen atom was made

3 a.m.u. and the mass of the connecting heavy atoms were adjusted to keep total mass unchanged. This allowed for the use of 4 fs time-step to integrate the equations of motion. The thermostat was switched to a Langevin thermostat with 1/ps damping coefficient. The equilibrated structures obtained at the end were subjected to further study. The simulation protocol for the rest of the work was identical to the second round of equilibration (except for the application of harmonic biases in umbrella sampling which is mentioned in the respective section). Up to this point, all simulation input files were obtained from CHARMM GUI COVID-19 archive and were used without modification. They are freely available in the CHARMM-GUI web-server (`http://www.charmm-gui.org/?doc=archive&lib=covid19`) and can be accessed for further simulation details.

1.2 Steered molecular dynamics and umbrella sampling

The open structure and multiple intermediate structures for the spike protein were generated by steered molecular dynamics. The RBD deviation coordinate [7] (the distance of the center of mass of RBD residues in a given from of simulation from the RBD center of mass in the crystal structure) was used as a reaction coordinate (RC) to perform SMD on the closed structure (6VXX). For the other system (6VSB), the RC was chosen to be the distance between Asp428 residue of chain A (the one whose RBD is going from "down" to "up" conformation) and the Lys986 residue of chain C. These residues are in proximity in the closed form and are far apart in the partially open and open conformation. For the SMD trajectory starting from the closed (PDB ID: 6VXX) system, the value of the RC was varied from 0 to 35 Å at constant velocity over 8 ns of simulation. A moving harmonic restraint was employed on the RC with a force constant of 1 kcal mol^{-1} Å^{-2}. The system was kept at the final state with a static restrain for additional 2 ns. All protein C_α atoms, except for the ones pertaining to residues 300 to 600 in chain A, were harmonically restrained with 0.5 kcal mol^{-1} Å^{-2} force constant. This prevents any spurious conformational change of the rest of the spike protein during the application of strong steering force on the RBD of Chain A. For the system generated from the partially open form (PDB ID: 6VSB), SMD simulation was performed with identical condition except the RC was varied from 35 Å to 10 Å over the first 8 ns.

Structures, which were ∼1 Å apart in RC space, were sampled from the SMD trajectories to perform umbrella sampling. For the first set (originating from PDB ID: 6VXX) 35 structures were sampled between RC value 1Å to 35Å. For each structure, a 17 ns simulation was performed in NPT ensemble. The restraint on the RBD-deviation coordinate was gradually increased to 1 kcal mol^{-1} Å^{-2} over the first 100 ps, and was held constant for the rest of the

simulation. The trajectory data of the last 11 ns was used for the analysis. For the second set 26 structures were sampled for the RC (Asp428(A)-Lys986(C) distance) values between 10Å to 35Å. Each trajectory was subjected to 15 ns of simulation with harmonic restraint on the RC, and last 10 ns of each trajectory was used for the analysis. Rest of the protocol is identical to the first set. Weighted Histogram Analysis Method (WHAM) [8] was used to calculate the 1D PMF from both set of trajectories. Error bars were computed from the Monte Carlo estimator implemented in WHAM program [9].

1.3 Unbiased simulation and time independent component analysis

Multiple unbiased trajectories, each of length ∼40 ns were initiated from the last frames of specific umbrella sampling simulations. Particular care was practiced in the choice of the starting structures. The US windows corresponding to 1Å to 10Å can better represent the vicinity of the closed state, while RBD-deviation of 22Å to 30Å corresponds to fully open structures, as realized from manual inspection of the trajectories. The partially open state is supposed to be the intermediate between closed and fully open states. So we chose umbrella windows 16Å to 35Å from the second set (starting from PDB ID: 6VSB) of US simulations to represent the conformational space in and around the partially open state. As all of the trajectories, intended to sample the fully open conformations, showed gradual RBD closing motion, we performed an additional simulation starting from the 33Å window of the first set. In total 38 unbiased simulations were performed. The trajectories corresponding the umbrella window 3Å and 33Å of the first set and the 32Å from the second set were identified as the closed, fully open and partially open state judging from the free energy profile. These unbiased trajectories were extended to ∼80 ns. A cumulative ~ 1.6 μs of unbiased simulation data was generated.

The unbiased trajectories were subjected to principal component analysis (PCA) [10] and time-independent component analysis (tICA) [11, 12]. First, all trajectories were projected onto a feature space consisting of contact distances between pairs of residues from the RBD and the rest of the spike protein. Particularly, we focused on the residue pairs which are close to each other in the closed state and far in the open state. We identified residues within the region 250-550 of chain A, we identified residues whose α-carbon is within 8 Å of the α-carbon of any other residue in the rest of the spike. The $C_\alpha - C_\alpha$ distances of such pairs (a total of 173 distances) were used as the feature space for PCA and tICA analysis.

As we found that the tICA analysis can better distinguish between all three states, only tICA results were used for subsequent analysis. The projection of the trajectories in the

space of first two principal components has been depicted in Fig. S1. We computed the sines and cosines of the backbone dihedral angles (ϕ and ψ) of all residues in the spike and their Pearson correlation coefficients with the two slowest degrees of obtained from tICA analysis. The residues were ranked based on highest (positive or negative) values of the correlation coefficients.

1.4 Mutual information and cross correlation

The three 80 ns trajectories corresponding to the closed, partially open and fully open states were used to calculate mutual information and inter-residue cross correlation. This protocol is traditionally used for calculating allosteric connections in Apo and Holo protein ligand systems. We followed this protocol except we used three conformational states of the spike protein.

The theoretical details of this method are described in Ref [13–15]. In information theory the mutual information metric (I_{ij}) between pairs of atoms can be computed as

$$I_{ij} = \int\int p(\mathbf{x}_i, \mathbf{x}_j) \ln \left(\frac{p(\mathbf{x}_i, \mathbf{x}_j)}{p(\mathbf{x}_i)p(\mathbf{x}_j)} \right) d\mathbf{x}_i d\mathbf{x}_j, \tag{1}$$

where the x_i, x_j are the Cartesian coordinates of the atoms, whereas $p(x_i)$ and $p(x_j)$ are their marginal distribution and $p(x_i, x_j)$ is the observed joint distribution [15]. A Pearson like correlation coefficient C_{ij} can be obtained from I_{ij} using the following expression:

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

We used linear mutual information in our analysis. It is given by

$$I_{ij} = H(\mathbf{x}_i) + H(\mathbf{x}_j) - H(\mathbf{x}_i, \mathbf{x}_j) \tag{3}$$

where H is a Shannon type entropy function:

$$\begin{aligned} H(\mathbf{x}_i) &= -\int p(\mathbf{x}_i) \ln p(\mathbf{x}_i) d\mathbf{x}_i \\ H(\mathbf{x}_i, \mathbf{x}_j) &= -\int\int p(\mathbf{x}_i, \mathbf{x}_j) \ln p(\mathbf{x}_i, \mathbf{x}_j) d\mathbf{x}_i d\mathbf{x}_j \end{aligned} \tag{4}$$

x_i, x_j are the Cartesian coordinates of the C_α atoms.

The absolute difference between the cross correlation heat map is computed to highlight the difference of long distance interactions between the RBD and the rest of the protein in the two open states in comparison to the closed state.

We further constructed a protein dynamic correlation network from the mutual information correlation matrix for each of the three states. We calculated and compared the

between centrality (BC) for each residue. The BC provides a way to quantify the amount of information flow between nodes or edges of a network. As the α-carbon of each residue is considered to be a node in our network construction, the value of BC provides the functional relevance of a residue in the total information flow through the protein. If a node i is working as a bridge between two other nodes along the shortest path joining them, the BC of node i is given by

$$BC(i) = \sum_{st} \frac{n_{st}^{i}}{g_{st}}, \tag{5}$$

where g_{st} is the total number of geodesics (shortest paths) joining nodes s and t, out of which n_{st}^{i} paths pass through node i [16, 17]. We measured the change in BC for each residue by quantifying the difference their values between two open states and the closed state.

1.5 Bioinformatics analysis

We performed sequence alignment for 67 sequences of SARS-CoV2 Spike Protein from Protein Data Bank. The software for constructing the alignment is MAFFT-DASH [18]. MAFFT utilises FFT algorithm to accelerate iterative sequence alignment without sacrificing too much accuracy. The web-based DASH (Database of Aligned Structural Homologs) provides structural alignments at the domain and chain levels for all proteins in the Protein Data Bank (PDB), resulting in improvement for alignment involving multiple sequences with weak similarity. The G-INS-i algorithm was used with maximum iteration cyles of 1000. This algorithm assumes that entire region can be aligned and tries to align them globally using the Needleman-Wunsch algorithm; that is, a set of sequences of one domain must be extracted by truncating flanking sequences. The result of sequence alignment is shown in figure 1, with sequence of PDB ID: 6VXX as the template. The alignment was analyzed by ConSurf server [19] to derive conservation score for each residue position in the alignment. The input for ConSurf analysis includes the sequence alignment by MAFFT above, atom coordinate of 6VXX and the sequence of 6VXX as the template. The residues will be sorted by a scale of 9 levels of conservation, starting from very high variability to very high conservation. The distribution of conserved and variable amino acids on the structure of S protein can be visualized by mapping the conserved and variable residues are also mapped onto the structure of 6VXX taken from PDB (Fig. S10). The mapping tool was also provided on ConSurf server. The map shows that conserved residues are predominant throughout the structure of S protein whereas most variable residues cluster around the tip of the receptor binding domain.

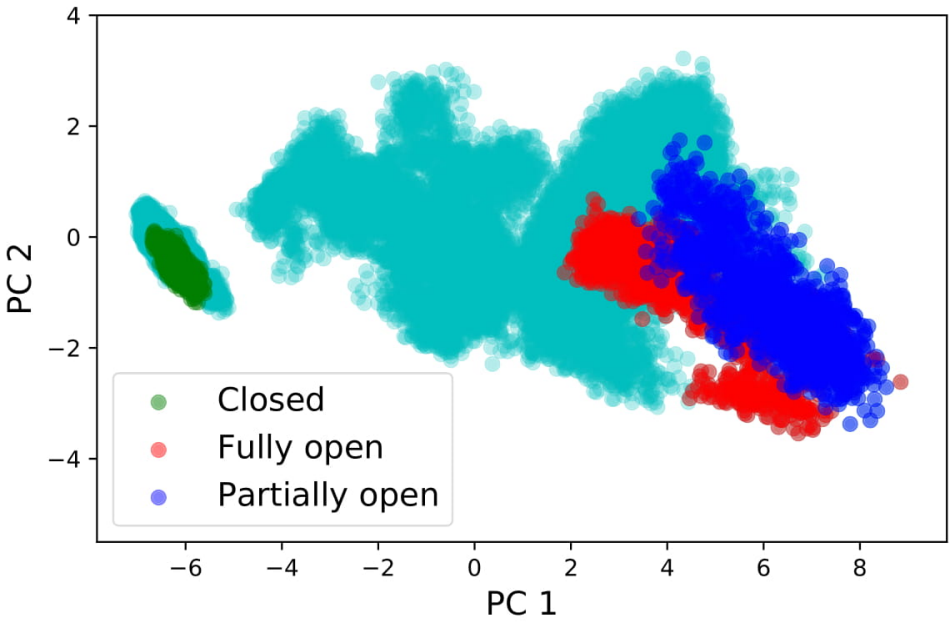

Figure S1: The projection of all unbiased trajectories along the first two principal components obtained from PCA.

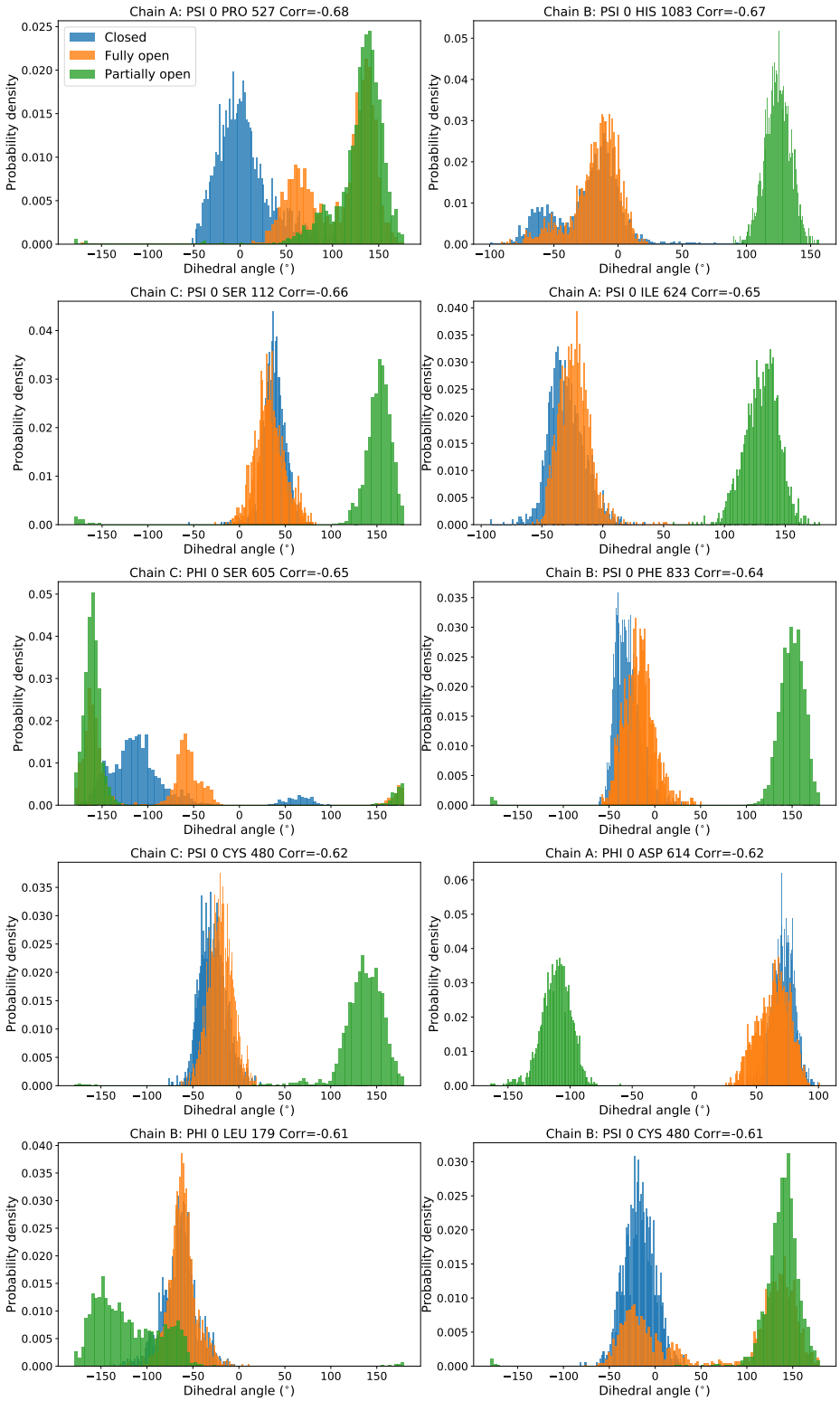

Figure S2: Distribution of the dihedral angles, the **cosine** functions of which are most **negatively** correlated with **tIC 1**. The distribution is plotted for closed, fully open and partially open state trajectories.

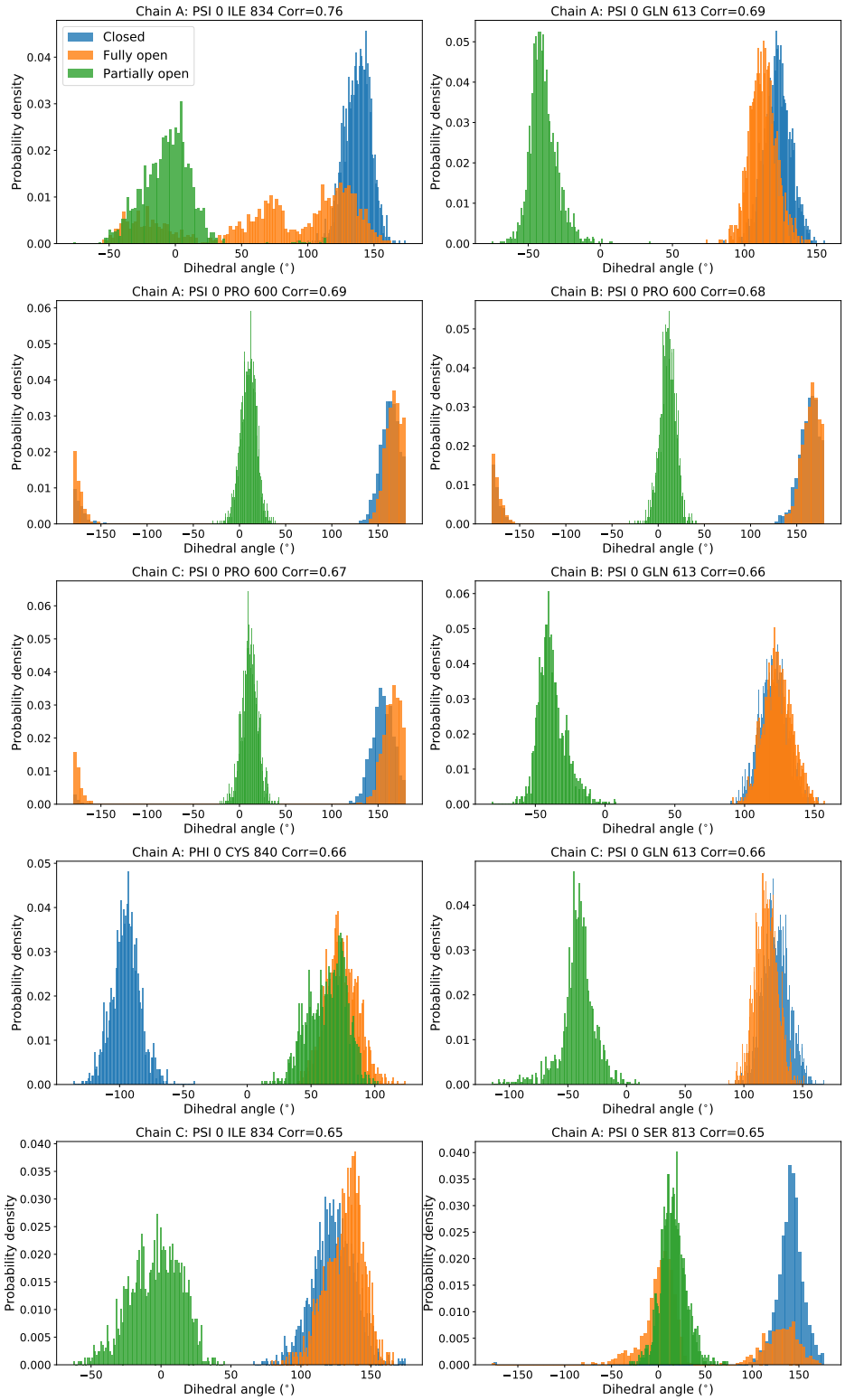

Figure S3: Same as Fig. S2 except for **positive** correlation

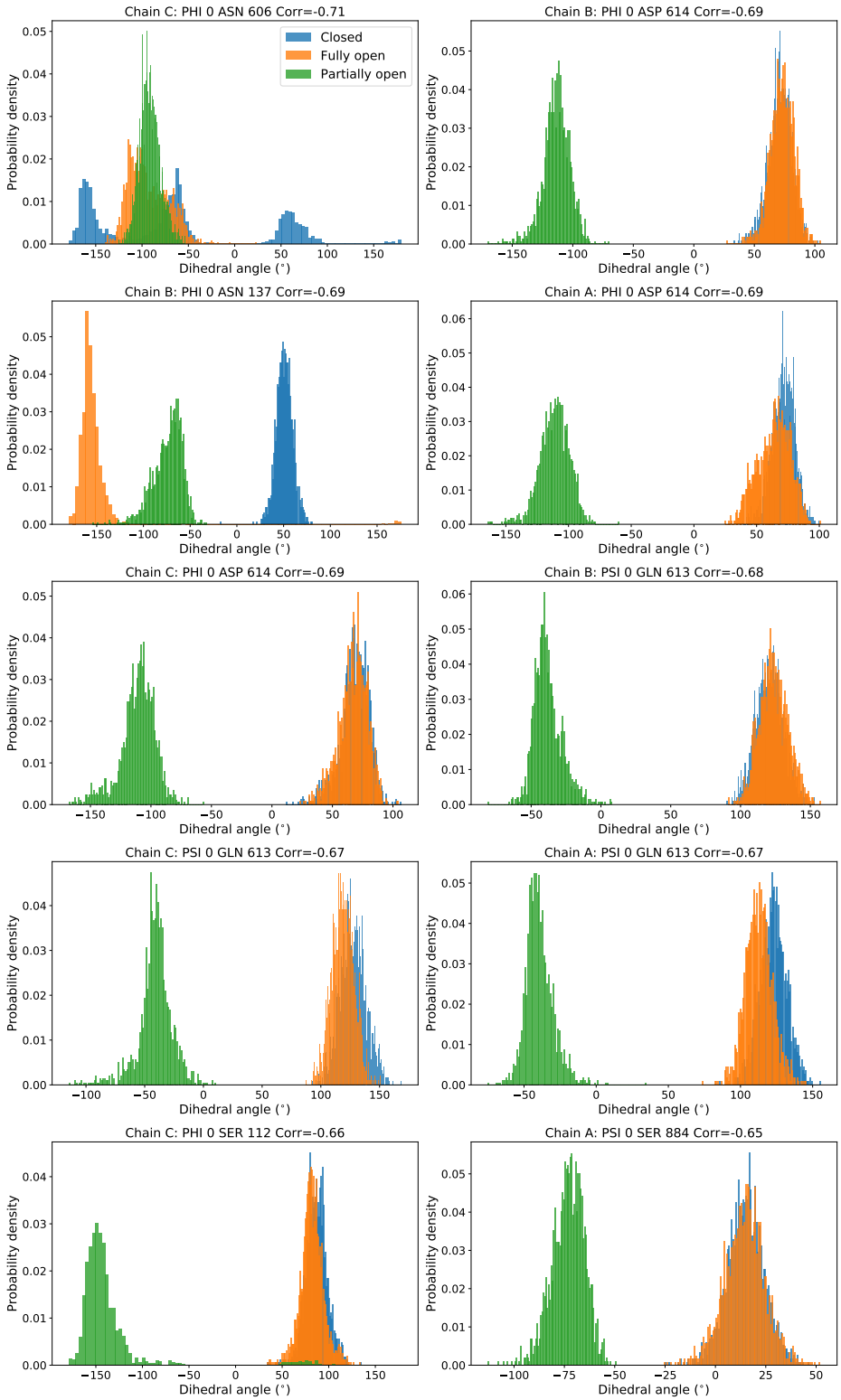

Figure S4: Distribution of the dihedral angles, the **sine** functions of which are most **negatively** correlated with **tIC 1**. The distribution is plotted for closed, fully open and partially open state trajectories.

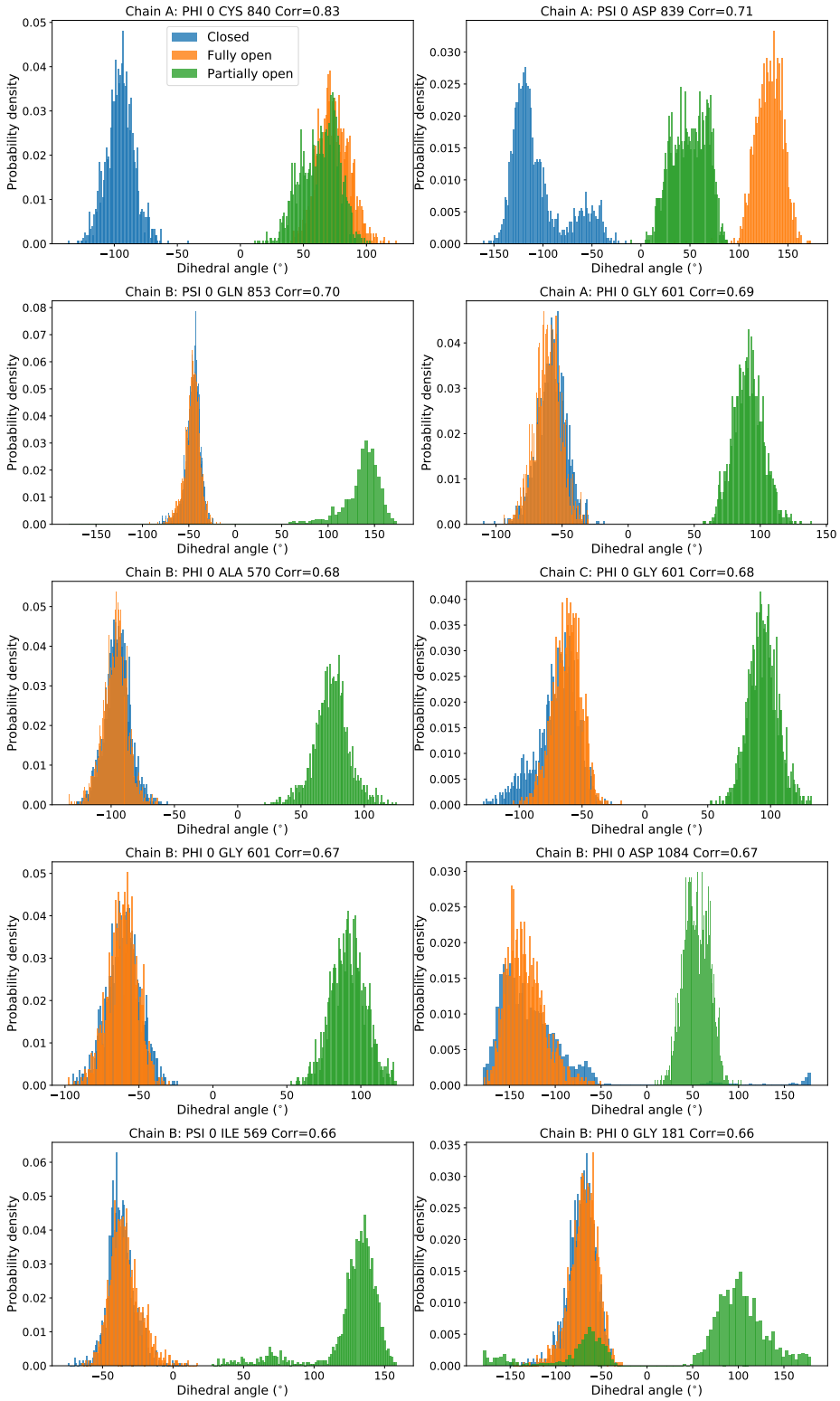

Figure S5: Same as Fig. S4 except for **positive** correlation

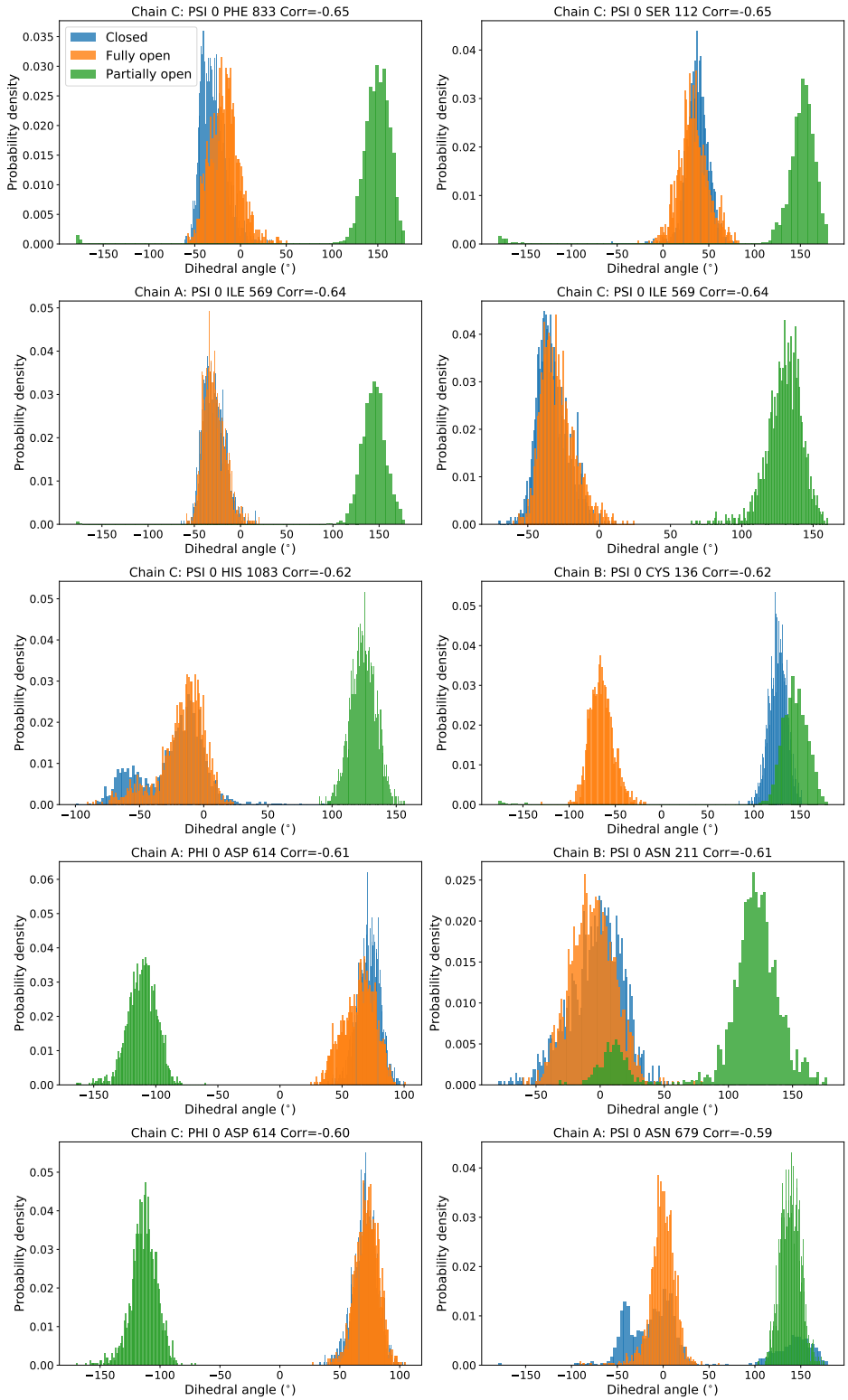

Figure S6: Distribution of the dihedral angles, the **cosine** functions of which are most **negatively** correlated with **tIC 2**. The distribution is plotted for closed, fully open and partially open state trajectories.

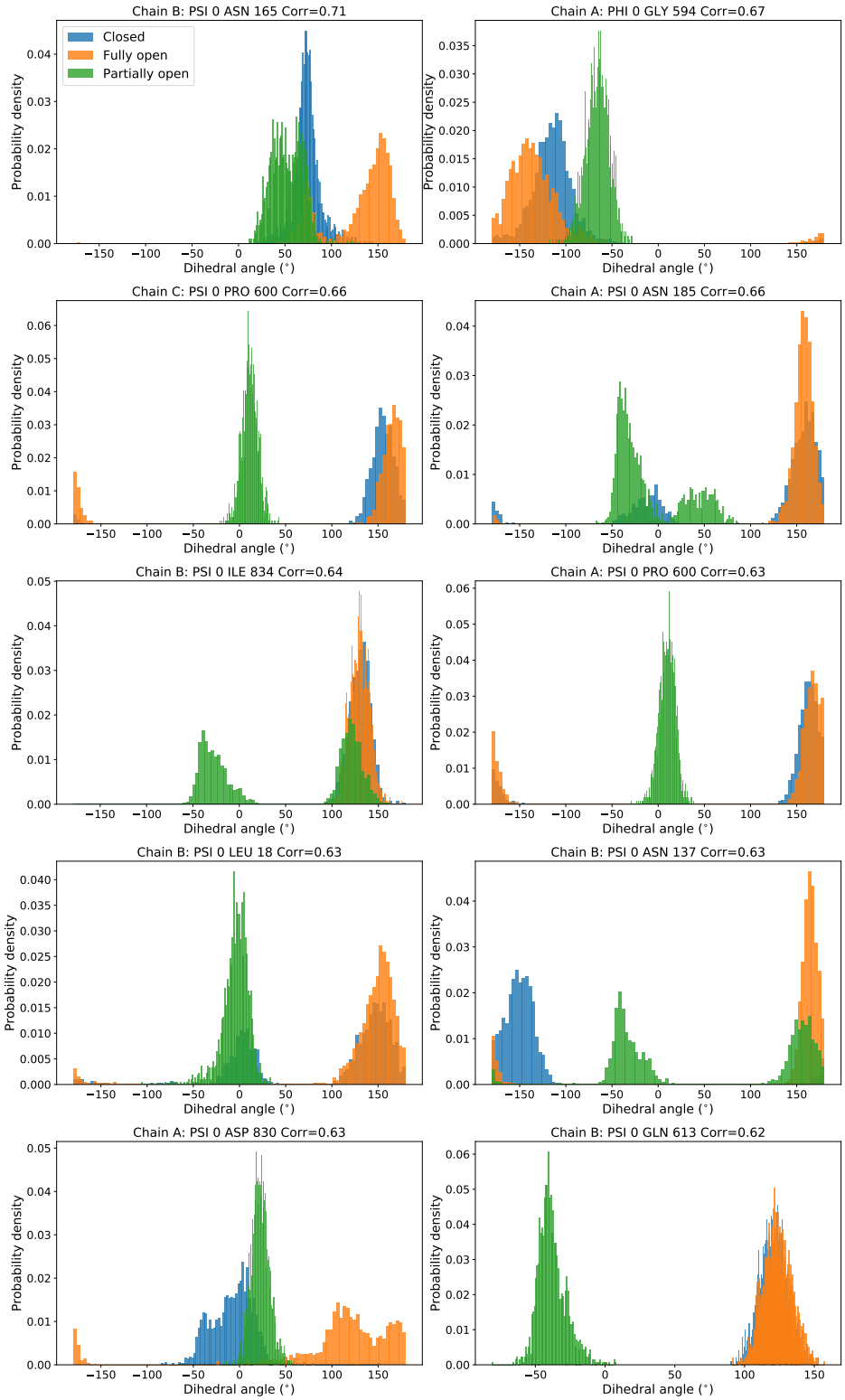

Figure S7: Same as Fig. S6 except for **positive** correlation

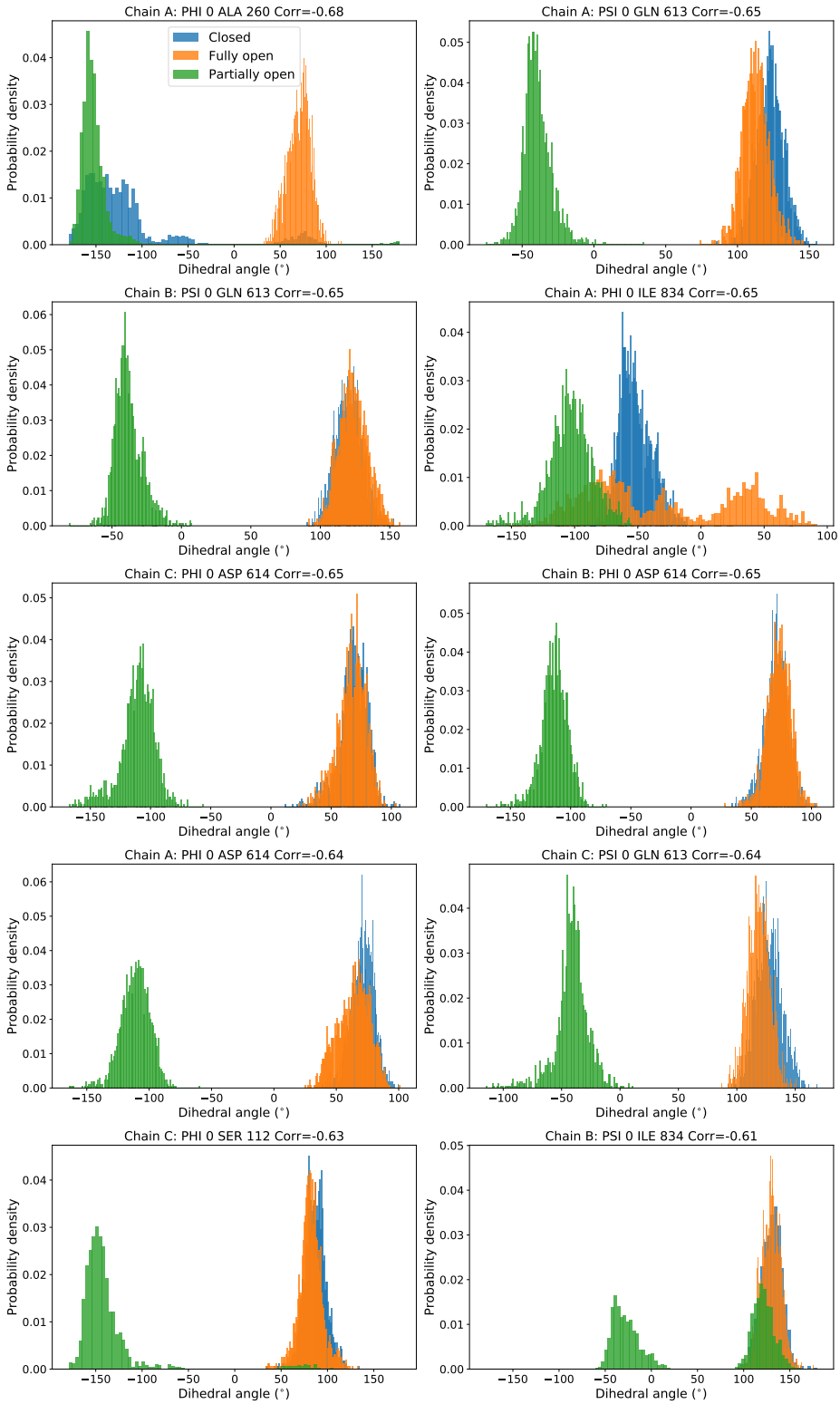

Figure S8: Distribution of the dihedral angles, the **sine** functions of which are most **negatively** correlated with **tIC 2**. The distribution is plotted for closed, fully open and partially open state trajectories.

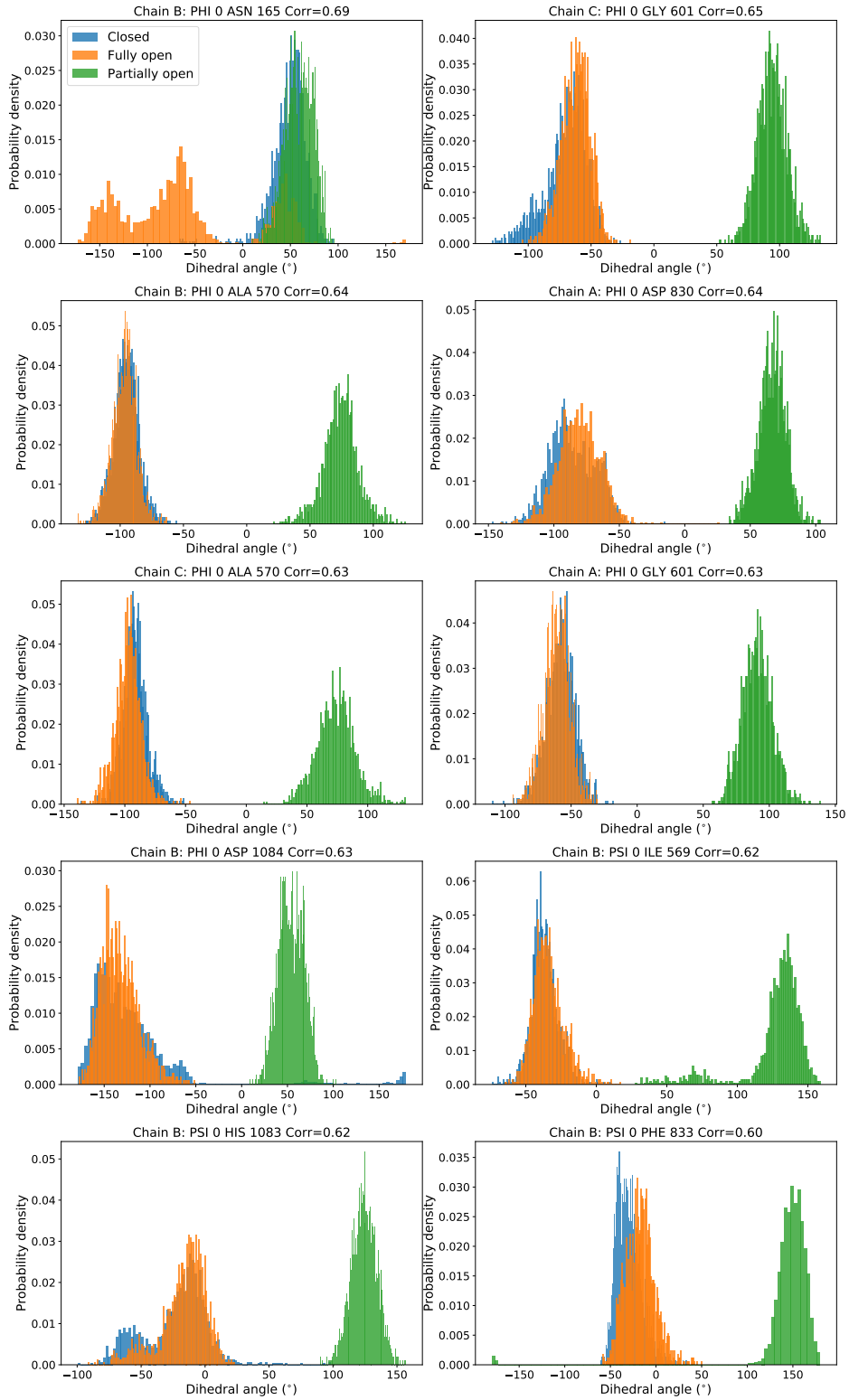

Figure S9: Same as Fig. S8 except for **positive** correlation

Figure S10: Results of sequence alignment study showing the most and least conserved residues in PDB ID: 6VXX over multiple mutant strains of SARS-CoV-2.

References

[1] Alexandra C. Walls, Young-Jun Park, M. Alejandra Tortorici, Abigail Wall, Andrew T. McGuire, and David Veesler. Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell*, 181(2):281 – 292.e6, 2020.

[2] Daniel Wrapp, Nianshuang Wang, Kizzmekia S Corbett, Jory A Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S Graham, and Jason S McLellan. Cryo-em structure of the 2019-ncov spike in the prefusion conformation. *Science*, 367(6483):1260–1263, 2020.

[3] Hyeonuk Woo, Sang-Jun Park, Yeol Kyo Choi, Taeyong Park, Maham Tanveer, Yiwei Cao, Nathan R. Kern, Jumin Lee, Min Sun Yeom, Tristan I. Croll, Chaok Seok, and Wonpil Im. Developing a fully glycosylated full-length sars-cov-2 spike protein model in a viral membrane. *The Journal of Physical Chemistry B*, 124(33):7128–7137, 2020. PMID: 32559081.

[4] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L de Groot, Helmut Grubmüller, and Alexander D MacKerell. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, 14(1):71–73, 1 2017.

[5] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, dec 2005.

[6] Chad W. Hopkins, Scott Le Grand, Ross C. Walker, and Adrian E. Roitberg. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *Journal of Chemical Theory and Computation*, 11(4):1864–1874, apr 2015.

[7] Maxwell I. Zimmerman, Justin R. Porter, Michael D. Ward, Sukrit Singh, Neha Vithani, Artur Meller, Upasana L. Mallimadugula, Catherine E. Kuhn, Jonathan H. Borowsky, Rafal P. Wiewiora, Matthew F. D. Hurley, Aoife M Harbison, Carl A Fogarty, Joseph E. Coffland, Elisa Fadda, Vincent A. Voelz, John D. Chodera, and Gregory R. Bowman. Sars-cov-2 simulations go exascale to capture spike opening and reveal cryptic pockets across the proteome. *bioRxiv*, 2020.

[8] Benoît Roux. The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, 91(1-3):275–282, 9 1995.

[9] Alan Grossfield. Wham: the weighted histogram analysis method, version 2.0.10. *http://membrane.urmc.rochester.edu/wordpress/?page_id=126*.

[10] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 2 edition, 2002.

[11] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72(23):3634–3637, 6 1994.

[12] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for Markov model construction. *The Journal of Chemical Physics*, 139(1):015102, 7 2013.

[13] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[14] Oliver F Lange and Helmut Grubmüller. Generalized correlation for biomolecular dynamics. *Proteins: Structure, Function, and Bioinformatics*, 62(4):1053–1061, 2006.

[15] S Bowerman and J Wereszczynski. Detecting allosteric networks using molecular dynamics simulation. In *Methods in enzymology*, volume 578, pages 429–447. Elsevier, 2016.

[16] Christian FA Negre, Uriel N Morzan, Heidi P Hendrickson, Rhitankar Pal, George P Lisi, J Patrick Loria, Ivan Rivalta, Junming Ho, and Victor S Batista. Eigenvector centrality for characterization of protein allosteric pathways. *Proceedings of the National Academy of Sciences*, 115(52):E12201–E12208, 2018.

[17] Ivan Rivalta, Mohammad M. Sultan, Ning-Shiuan Lee, Gregory A. Manley, J. Patrick Loria, and Victor S. Batista. Allosteric pathways in imidazole glycerol phosphate synthase. *Proceedings of the National Academy of Sciences*, 109(22):E1428–E1436, 2012.

[18] John Rozewicki, Songling Li, Karlou Mar Amada, Daron M Standley, and Kazutaka Katoh. Mafft-dash: integrated protein sequence and structural alignment. *Nucleic acids research*, 47(W1):W5–W10, 2019.

[19] Haim Ashkenazy, Shiran Abadi, Eric Martz, Ofer Chay, Itay Mayrose, Tal Pupko, and Nir Ben-Tal. Consurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic acids research*, 44(W1):W344–W350, 2016.