Title: Modelling conformational state dynamics and its role on infection for SARS-CoV-2 Spike protein variants

Authors: Natália Teruel[1], Olivier Maihot[1,2] & Rafael Josef Najmanovich[1,*]

Affiliation:

[1] Department of Pharmacology and Physiology, Faculty of Medicine, Université de Montréal

[2] Institute for Research in Immunology and Cancer (IRIC), Faculty of Medicine, Université de Montréal

[*] To whom correspondence should be addressed: rafael.najmanovich@umontreal.ca

Keywords: Normal Mode Analysis; Dynamics, Conformational states; SARS-CoV-2; Spike Protein; Infectivity; mutation surveillance; D614G; N501Y

Abstract

The COVID-19 pandemic has greatly affected us all, from individuals to the world economy. Whereas great advances have been achieved in record time, a lot remains to be learned about the infection mechanisms of its causative agent, the SARS-CoV-2 coronavirus. The Spike protein interacts with the human angiotensin converting enzyme 2 receptor as part of the viral entry mechanism. To do so, the receptor binding domain (RBD) of Spike needs to be in an open state conformation. Here we utilise coarse-grained normal mode analyses to model the dynamics of the SARS-CoV-2 Spike protein and the transition probabilities between open and closed conformations for the wild type, the D614G mutant as well other variants isolated experimentally. We proceed to perform several possible *in silico* single mutations of Spike, 17081 in total, to determine positions and specific Spike mutations that may affect the occupancy of the open and closed states. We estimate transition probabilities between the open and closed states from the calculated normal modes. Transition probabilities are employed in a Markov model to determine conformational state occupancies. Our results correctly model a shift in occupancy of the more infectious D614G strain towards higher occupancy of the open state via an increase of flexibility of the closed state and concomitant decrease of flexibility of the open state. Our results also suggest that the N501Y mutation recently observed, drastically increases the occupancy of the open state. We utilize global vibrational entropy differences to select candidate single point mutations that affect the flexibility of the open and closed states and confirm that these lead to shifts in occupancies for the most critical mutations. Among those, we observe a number of mutations on Glycine residues (404, 416, 504) and G252 in particular accepting a number of mutations. Other residues include K417, D467 and N501. This is, to our knowledge, the first use of normal mode analysis to model conformational state transitions and the effect of mutations thereon. The specific mutations of Spike identified here, while still requiring experimental validation, may guide future studies to increase our understanding of SARS-CoV-2 infection mechanisms as well as guide public health in their surveillance efforts.

1. Introduction

The coronavirus pandemic has emerged as a major and urgent issue affecting individuals, families and societies as a whole. Among all outbreaks of aerosol transmissible diseases in the 21st century, the COVID-19 pandemic, caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) virus [1,2], has the highest infection and death cumulative numbers - 61 million infections and over 1.4 million deaths, according to the World Health Organization (WHO) epidemiological report of December 1, 2020 [3]. Recent WHO reports also show significant weekly increases in the number of infections and deaths as countries start to face upcoming waves of the disease. In 2003 the SARS coronavirus (SARS-CoV) pandemic caused 8,098 infections and 774 deaths before it was brought under control [4,5]. In 2012, the Middle East respiratory syndrome-related coronavirus (MERS-CoV) outbreak caused 2499 infections and 858 deaths, presenting the highest fatality rate [6]. SARS-CoV-2, SARS-CoV and MERS-CoV, as coronaviruses in general, present considerable mutation rates, which may contribute to future outbreaks. For instance, SARS-CoV-2 is estimated to have a mutation rate close to the ones presented by MERS-CoV [7] and by SARS-CoV [8], as well as other RNA viruses, showing a median of 1.12×10^{-3} mutations per site per year [9]. The high mutation rate may in part be responsible for the zoonotic nature of these viruses and points to a clear risk of still-undetected additional members of the *coronavirideae* family of viruses making the jump from their traditional hosts to humans in the future.

The SARS-CoV-2 Spike protein (Uniprot ID P0DTC2) is responsible for anchoring the virus to the host cell. The entry receptor for SARS-CoV-2 and other lineages of human coronaviruses is the human cell-surface protein angiotensin converting enzyme 2 (ACE2) (Uniprot ID Q9BYF1) [10]. Therefore, studying the Spike protein family is essential to understand the evolution of coronaviruses.

SARS-CoV-2 Spike is a homo-trimeric glycoprotein, with each chain built by subunits S1 and S2, delimited by a furin cleavage site at residues 682-685. The S1 subunit comprises the N-terminal Domain (NTD), located in the peripheric part of the extramembrane extreme, and the Receptor Binding Domain

(RBD), the most flexible site, located in the central part of this same extreme. The S2 subunit consists of the fusion peptide (FP), heptad repeat 1 (HR1), heptad repeat 2 (HR2), the transmembrane domain (TM), and the cytoplasmic tail (CT) (Figure 1). The interaction between Spike and ACE2 relies on Spike to be in its open conformation, in which the Receptor Binding Domain (RBD) is extended [11]. The study of the binding properties between Spike and ACE2, although important, cannot explain all the nuances of the infection mechanism. An example of this limitation is the comparison between SARS-CoV and SARS-CoV-2, which have different rates of infection even though they share similar Spike-ACE2 affinities [12]. These facts lead us to consider the contribution of the dynamics of the Spike protein to the infection process.

Computational structural biology methods have grown in both accuracy and usability over the years and are increasingly accepted as part of an integrated approach to tackle problems in molecular biology. Such integration permits to speed up research, decrease needs in infrastructure, reagents, and human resources and allows us to evaluate increasingly larger data sets. Computational approaches are being extensively used in the study of SARS-CoV-2 and its mechanisms of infection [13-15]. Among these, we highlight the study of dynamic properties of the Spike protein as well as in antibody recognition and the search for therapeutic interventions [16-18].

Several aspects of the dynamics of the Spike protein are being currently studied, with a range of particular goals: to evaluate the docking of small molecules to the RBD domain [19], to search for alternative target binding-sites for vaccine development [20], to understand residue-residue interactions and their effects on conformational plasticity [21] and to investigate the flexibility of different domains in particular conformational states [22].

Normal Mode Analysis (NMA) methods are being employed in the study of different conformational states [23] and of different coronavirus variants [24]. These methods, however, are limited with respect to their ability to study the effects of mutations on dynamics due to the fact that such methods are either extremely taxing on computational resources (e.g., molecular dynamics) or agnostic to the nature of

amino acids (e.g., traditional coarse-grained NMA methods). In the past, our group developed a coarse-grained NMA method called ENCoM (for Elastic Network Contact Model) that is more accurate than alternative coarse-grained NMA methods due to the explicit consideration of the chemical nature of amino acids and their interactions and consequently their effect on dynamics [25]. ENCoM performs better than other NMA methods on traditional applications and is the only coarse-grained NMA method capable of predicting the effect of mutations on protein stability and function as a result of dynamic properties [26-28].

In this study, we use the ENCoM method to study the dynamics of the Spike protein, considering different conformational states and several sequence variants observed during the current pandemic, as well as through large-scale analysis of *in silico* mutations. Experimental analysis of the effect of the SARS-CoV-2 Spike mutation D614G and the comparison between SARS-CoV and SARS-CoV-2 Spike proteins show unique dynamic characteristics that correlate with epidemiological and experimental data on infection. The present work shows that we can replicate such results computationally, suggesting that rigidity or flexibility of different Spike conformational states affects infectivity. We present a high throughput analysis of simulated single amino acid mutations on dynamic properties to seek potential hotspots and individual Spike variants that may be more infectious and therefore may guide public health decisions if such variants were to appear in the population. We also introduce a Markov model of occupancy of molecular states with transition probabilities derived from our analysis of dynamics that recapitulates experimental data on conformational state occupancies. This is the first application of an NMA method that derives transition probabilities from normal modes and employs them in a dynamic system to predict the occupancy of different conformational states. We model the occupancy of several variants and highlight those that may be useful in studying future epidemiological trends that can be responsible for new outbreaks.

2. Methods

2.1 Spike protein models

We performed our analyses using the crystallographic models of the SARS-CoV-2 Spike protein in the open (PDB ID 6VYB) and closed (PDB ID 6VXX) states. The open (prefusion) state was designed with an abrogated furin S1/S2 cleavage site and two consecutive proline stabilizing mutations [29]. Despite the mutations, the engineered structures correctly represent the conformational states of Spike, as confirmed by independently solved structures [23,30]. The PDB structures used for the SARS-CoV comparison were 5X58 and 5X5B for closed state and one RBD open state, respectively [31].

Occupancy calculations were performed with reconstructed closed and one-RBD-open structures using as template 6VXX and 6VYB and the reference sequences published with experimental occupancy calculations, as follows (acronyms described in results): S-GSAS/WT: 7KDG,7KDH; S-GSAS/D614G: 7KDI,7KDJ [30]; S-R/x2: 6ZOX; S-R/PP/x1: 6ZOY,6ZOZ; S-R: 6ZP0; S-R/PP: 6ZP1,6ZP2 [32].
We removed heteroatoms, water molecules, and hydrogen atoms from the PDB structures. Missing residues were reconstructed using template-based loop reconstruction and refinement with Modeller [33]. Single amino acid mutants were generated using FoldX4 [34].

2.2 Dynamic analyses

We analysed dynamic properties of the Spike protein with ENCoM [25]. ENCoM employs a potential energy function that includes a pairwise atom-type non-bonded interaction term and thus makes it possible to consider the effect of the specific nature of amino-acids on dynamics. NMA explores protein vibrations around an equilibrium conformation by calculating the eigenvectors and eigenvalues associated with different normal modes [35-37]. Representing each protein residue as a single point, for a given conformation of a protein with N amino acids, we obtain 3N - 6 nontrivial eigenvectors. Each eigenvector represents a linear, harmonic motion of the entire protein in which each amino acid moves along a unique 3-dimensional Euclidean vector. The associated eigenvalues rank the eigenvectors in terms of energetic accessibility, lower values corresponding to global, more easily accessible motions.

NMA calculations allow us to computationally estimate b-factors associated with the protein structure, as shown in Equation 1 for the ith residue, which in turn are related to local flexibility. Higher predicted b-factors denote more flexible positions. Individually calculated b-factors are combined in a vector for a protein sequence or part thereof and called Dynamic Signature.

$$B_i = \sum_{n=7}^{3N} \frac{E_{n,i,x}^2 + E_{n,i,y}^2 + E_{n,i,z}^2}{\lambda_n} \qquad (1)$$

The eigenvectors and associated eigenvalues can also be used to obtain a variable called vibrational entropy that can be used to compare the relative stability of two states. For example, by measuring the difference of vibrational entropy (ΔS_{vib}) between a mutant and a wild type (WT), one can calculate how much a mutation affects the overall flexibility and stability of the mutant relative to the WT. The ΔS_{vib} value predicted by ENCoM is positive when the mutation makes the protein more flexible and negative when the mutation makes the protein more rigid. The differences between the ΔS_{vib} values for closed and open states were calculated for each mutant ($\Delta\Delta S_{vib} = \Delta S_{vib\ (open)} - \Delta S_{vib\ (closed)}$) in order to evaluate individual mutations according to a single score. Vibrational Entropy calculations are dependent on the thermodynamic β factor, that for pseudo-physical models such as ENCoM serves as a scaling factor. This term was optimized to fit experimental Gibbs free energy differences [38] and established as β = 1. The vibrational contribution of the entropic components of the free energy is calculated as described in Eq. 2 [39] in units of $J.K^{-1}$, where N is the total number of amino acids in the protein, v_i is the vibrational frequency and K_B is the Boltzmann constant. Equation 3 shows the association between eigenvalues and vibrational frequency.

$$S_{vib} = K_B \sum_{n=7}^{3N} \left\{ \frac{\beta v_n}{e^{\beta v_n - 1}} - \ln\left[1 - e^{-\beta v_n}\right] \right\} \quad (2)$$

$$\lambda_n = v_n^2 \qquad (3)$$

The Najmanovich Research Group Toolkit for Elastic Networks (NRGTEN) [38], with the latest implementation of ENCoM, also includes a function to evaluate state occupancies by calculating transition probabilities between different states. A probability P_j of moving along each eigenvector j can be obtained using a Boltzmann distribution given its associated eigenvalue λ_j and a scaling factor γ.

$$P_j = \frac{e^{-\lambda_j/\gamma}}{\sum_{i=7}^{3N} e^{-\lambda_i/\gamma}} \qquad (4)$$

Let's consider two conformations A and B of the same protein and the vector $\mathbf{E}_{A\rightarrow B}$, which represents the conformational change going from conformation A to conformation B. The overlap between each normal mode $\boldsymbol{M}_j$ computed from conformation A and the $\mathbf{E}_{A\rightarrow B}$ vector is a value between 0 and 1 describing how well that normal mode recapitulates the conformational change required to go from one state to the other [40].

$$O\left(\boldsymbol{E}_{A\rightarrow B}, M_j\right) = \frac{\left|\boldsymbol{E}_{A\rightarrow B} \cdot \boldsymbol{M}_j\right|}{\|\boldsymbol{E}_{A\rightarrow B}\|\|\boldsymbol{M}_j\|} \qquad (5)$$

We can then calculate the transition probability of going from conformation A to conformation B as the weighted sum of the Boltzmann probability P_j of each normal mode $\boldsymbol{M}_j$ times the overlap between that normal mode and the conformational change $E_{A\rightarrow B}$.

$$P_{A\rightarrow B} = \sum_{j=7}^{3N} P_j \times O\left(\boldsymbol{E}_{A\rightarrow B}, \boldsymbol{M}_j\right) \qquad (6)$$

The reverse probability $P_{B\rightarrow A}$ can be computed in the same fashion, giving an indication of which conformation is favored between the two.

A simple way of computing the occupancies of these conformations from the transition probabilities is to use a Markov model. Each conformation is represented by a state, and the transition probabilities between states are computed as described above. We add a constant k to all states as the probability of staying in that state. Since all states must have outgoing transition probabilities that sum to 1, we normalize these values after the addition of k. For a two-state Markov chain representing the open and closed states of the Spike protein, we obtain the diagram shown in Figure 2. All transition probabilities

are computed using ENCoM and Eq. 6. The parameters k and γ need to be optimized for the system being studied as they are not directly coupled to physical quantities because of the pseudo-physical, coarse-grained nature of the ENCoM model. Once the parameters are set, there is a unique equilibrium solution that gives the occupancies of the two states. This approach could be easily generalized to a Markov model with more than two states, where the transition between any two states is computed exactly as described above if that transition is deemed possible.

3. Results and Discussion

3.1. Dynamic Signature of different Spike variants

3.1.1. G614 and D614 dynamic comparison

An important event in the progression of the COVID-19 pandemic was the appearance of the D614G variant in mid-February 2020 in Europe. The fast spread of this variant raised the possibility that this mutation conferred advantages relative to other forms of the virus in circulation at the time [41, 42]. Studies revealed that the mutation has indeed greater infectivity, triggering higher viral loads [43, 44]. Several hypotheses have emerged to explain the mechanisms behind this higher infectivity primarily focused on possible effects on the furin cleavage site [30, 45, 46], but recently also considering possible important dynamic differences [44, 47, 48].

In order to test if Dynamic Signatures reveal differences between Spike variants, we analysed the 13741 sequences of the protein available on May 08 in the COVID-19 Viral Genome Analysis Pipeline, enabled by data from GISAID [49, 50]. The mutant Spike proteins harboring mutations (Table S1) were modelled in the open and closed states. Dynamic Signatures were calculated for each mutant in both states and clustered (Figure 3). Mutations in positions that had no occupancy in the original templates used for the open and closed states (positions 5, 8, and 1263) were ignored.

Analysis of the effect of mutations on the Dynamic Signature show that the D614G mutation produces similar dynamic patterns largely independent of the other mutations accumulated, and dynamic patterns that are distinct from that of the wild type and other mutants on both the open and closed states. The dynamic characteristics of D614G are very specific and cannot be obtained with random mutations (Figure S1, Table S2). Performing the clustering using segments of the Dynamic Signature representing lengths of 100 amino acids identifies a section of the Spike protein from around position 250 to around position 750, responsible for the unique characteristics that the mutation D614G confers to the dynamics of the Spike protein (data not shown). This section of Spike includes part of the N-Terminal Domain (NTD) and all of the RBD domain.

When checking the difference between the Dynamic Signatures of the wild type D614 and the mutant G614 we observe that for the closed conformation, the pattern tends towards negative values, indicating that this mutation makes the closed state more flexible, especially around the position of the mutation. On the other hand, for the open B chain conformation the pattern is positive for the open RBD, the same chain NTD and the adjacent chain NTD, indicating that this mutation makes these areas of the open conformation more rigid (Figure 4). This result led us to hypothesize that a more flexible closed state would favor the opening of Spike and that a more rigid open state would disfavor its closing, thus shifting the conformational equilibrium towards the open state and favouring interaction with ACE2, leading to increased cell entry. Mutating position 614 to every other amino acid, we observe a correlation in the closed state between residue size and flexibility. Namely, smaller amino acids tend to make the closed state more flexible. However, we do not observe the opposite effect on the open state. Mutation of D614 to Glutamine, which is similar to Aspartate, barely shows any effect. Nevertheless, we can see that other amino acids have a similar effect as Glycine, such as Proline and Threonine (Figure S2).

3.1.2. Comparison of the Dynamic Signatures of Spike from SARS-CoV and SARS-CoV-2

It has been previously observed that RBD flexibility in SARS-CoV influences binding to ACE2 and facilitates fusion with host cells [51]. Thus, considering the lesser infectivity of SARS-CoV relative to SARS-CoV-2 and our aforementioned results for the D614G mutation, we expected the SARS-CoV Spike to be more rigid in the closed state and more flexible in the open state relative to Spike from SARS-CoV-2. This is indeed the case (Figure 5). The dynamic signature values of SARS-CoV are smaller than those of SARS-CoV-2 in several areas throughout the closed structure, indicating that when in the closed state, the SARS-CoV Spike protein is more rigid. For the open state we can see that SARS-CoV open RBD and adjacent NTD are significantly more flexible than for SARS-CoV-2 Spike.

3.2. Vibrational entropy

It is possible to combine the trend of a Dynamic Signature into a single value to represent the overall flexibility of any given mutation and compare it to the WT. This can be achieved with ΔS_{vib}, calculated with Eq. 2 for each state (see materials and methods). For any given state, positive ΔS_{vib} values represent mutants that relative to the wild type make the protein more rigid, whereas negative values of ΔS_{vib} describe mutations that cause the protein to be more flexible in the given state relative to the wild type. In the case of the mutation D614G, we obtain $\Delta S_{vib\ (open)} = 5.26\text{x}10^{-2}$ $J.K^{-1}$ and $\Delta S_{vib\ (closed)} = -9.27\text{x}10^{-2}$ $J.K^{-1}$ with a $\Delta\Delta S_{vib}$ (calculated as $\Delta S_{vib\ (open)} - \Delta S_{vib\ (closed)}$) of $1.45\text{x}10^{-1}$ $J.K^{-1}$.

We generated *in silico* the 19 possible single mutations in each position from residue 14 to residue 913 and calculated $\Delta S_{vib\ (open)}$, $\Delta S_{vib\ (closed)}$ and $\Delta\Delta S_{vib}$. Other positions were ignored due to uncertainties in modelling or the fact that they are not expected to have a pronounced effect on dynamic [23]. It should be noted that Spike cannot accommodate the vast majority of such single mutations, particularly in its core as these would lead to unstable or misfolded conformations. However, those that occur near the surface are more likely to represent single residue variations of the Spike protein that lead to a stable, correctly folded protein. Therefore, the stability of specific mutations highlighted in this work, unless

otherwise stated (such as those already observed experimentally or within the RBD domain as stated below), needs to be validated experimentally.

The heatmap in Figure 6A shows ΔS_{vib} values associated with mutations on the closed conformational state (left) and open conformational state (right). Lighter colors represent high ΔS_{vib} values, meaning that the specific mutant is more flexible than the WT, and darker colors represent low ΔS_{vib} values, meaning that the specific mutant is more rigid than the WT. The second heatmap (Figure 6B) shows $\Delta\Delta S_{vib}$ values, or Difference Scores, highlighting positions and specific mutations with great contrast between their effect on the open and closed states. In this representation, blue mutants are more rigid in the closed state and more flexible in the open state, therefore candidates for less infectious mutants, and red mutants are more flexible when closed and more rigid when open, candidates for more infectious mutants.

In Figure 7 we map $\Delta\Delta S_{vib}$ values (Figure 6B) on the structure of Spike, colored according to the median value for each position with the same color scheme as the heatmap. From the 17081 single mutations considered, we highlight the top 64 mutants (with $\Delta\Delta S_{vib}>0.3$) and bottom 20 in terms of $\Delta\Delta S_{vib}$ values (Table 1). The Dynamic Signature comparison for 3 of those most infectious candidates (Figure 8A) and 3 of the least infectious candidates (Figure 8B) shows some of the patterns that could lead to a greater or lesser effect on infectivity. For instance, in Figure 8A we can see that high scores can come from a large flexibility of the closed state, a very large rigidity of the open state, or have the contribution of both. We can also observe that these effects can be different in each chain and can affect more the NTD, the RBD, or both. Finally, these single mutants also show how a point mutation can have widespread impacts across the protein.

3.3. Conformational state occupancies

We calculated forward and reverse transition probabilities between the open and closed states (Eq. 4, 5 & 6) from the calculated normal modes and used the Markov model described in Materials and Methods

to calculate the equilibrium occupancies for each state in wild type and mutant Spike proteins. It is unclear if any additional conformational states other than those with either all three RBD domains in the closed state or only one RBD open state are biologically relevant. Specifically, Yurkovetskiy *et al.* [44] observed an occupancy for states with two or three RBD domains in the open conformation, but these were not observed by Gobeil *et al.* [30] and Xiong *et al.* [32] or taken into consideration in several other structural studies [20-24]. As such, we employ the two-state model shown in Figure 2, with one state representing all three RBD domains closed and the second state representing one RBD open. We calculated the robustness of this Markov occupancy model utilizing 60 different reconstructed structures, varying the positions of loops and with minor differences in the core structure, representing the closed state and the open state for each chain. The results are equivalent no matter what specific structural template is used to represent each of the two states above.

The Markov model calculation of occupancies requires two parameters (see Materials and Methods) that were optimized based on experimental data for six Spike variants. These variants were: S-GSAS/D614, an engineered Spike with the sequence GSAS in the furin cleavage site and no 614 mutation; S-GSAS/G614, with the same furin site modifications and the D614G mutation [30]; S-R, the Spike protein with original furin site RRAR; S-R/x2, with added S383C, D985C mutations inducing a disulfide bond; S-R/PP, engineered with two prolines in positions 986 and 987; S-R/PP/x1, in which from the double prolines sequence the mutations G413C, V987C were performed to induce a disulfide bond [32]. It is worth stressing that all 6 variants used to calibrate the two parameters affecting the occupancy were modelled on the same open and closed state conformations. All differences in observed occupancies and the agreement with experimental occupancy data came about as a consequence of the effect of the mutations on the normal modes and derived transition probabilities and not as a result of structural differences between variants. We obtained a good fitting to the experimental results with k and γ of 0.5 and 0.001, respectively (Pearson correlation = 0.89, p-value = 1.94×10^{-2}). Predicted occupancies of the open and closed states for each of the six variants above, as well as the experimental data, are

presented in Table 2. We utilized this data to calculate occupancy differences for each variant (Figure 9). The range of variation of our predicted occupancies is small compared to that of experimental values. We believe that given the limitations of our coarse-grained model as well as additional phenomena that ultimately affect occupancy, our predictions reflect only a fraction of the myriad of factors contributing to the occupancy. Nonetheless, our predictions correctly capture the pattern of relative variations of occupancy observed in the experimental data. Furthermore, we simulated random sets of occupancies for the 6 sequence variants and calculated simulated correlations for the 110 different combinations of k and γ to determine if the observed correlations represent an actual signal in the data or could be randomly obtained with different values for the parameters k and γ. We observed a marked shift with higher correlations for the data representing our predicted occupancies when compared to the gaussian noise data (Figure S3), suggesting that the predicted occupancies are not due to chance.

The computational resources needed for the calculation of occupancies for all 8250 mutations with $\Delta\Delta S_{vib}>0$ are beyond our current capabilities. We set a threshold of $\Delta\Delta S_{vib}>0.3$ to select candidates for the calculation of occupancies. This threshold corresponds to 64 mutations (Table 1, in red). Using the parameters k and γ obtained above, we calculated occupancies for these 64 mutants as well as the 20 mutants with lowest $\Delta\Delta S_{vib}$ values (Table 1, in blue). In Figure 10A we show the difference in occupancy between the open and closed states using a non-linear scale adapted to better show the results around the wild type occupancy. Whereas $\Delta\Delta S_{vib}$ values for particular mutations may hint at a more flexible closed state and more rigid open state, this is a global measure that may not reflect the necessary pattern of flexibility across the structure that leads to effective transition probabilities between the open and closed states. Yet, for the most part, $\Delta\Delta S_{vib}$ can predict the shifts in occupancy, showing a clear distinction between the 64 mutants predicted using $\Delta\Delta S_{vib}$ as shifting occupancy towards the open state and the 20 mutants predicted to shift the equilibrium towards the closed state (p-value=2.04×10^{-6}). Figure 10B shows the location in the structure of the mutants in Table 1. We can see that the least infectious

candidates (blue) are positioned in the interfaces between NTD and RBD domains, while the most infectious candidates, especially the ones validated by the occupancy prediction (dark red), are more concentrated in the interfaces between different RBD domains.

Residue G252 stands out as capable of accommodating a large number of mutations (C, D, E, H, M, P, Q, S, T, W) that shift the occupancy in favour of the open state. The fact that variants in this position do not seem to be prevalent in outbreaks to date, points to the possibility that this position may be under additional functional constraints that prevent the emergence of variants. A number of other Glycine residues also could accept mutations that we predict to increase the occupancy of the open state: G72W; G404W; G413M; G416E,W; and G404I. In fact, three of the top four mutations are mutations on Glycine. A number of other potential mutations are adjacent to Glycine residues above. Namely, R403S and K417D,E,G,P. Additionally, D467P,W and I468T are also positions that are adjacent to others that can accommodate mutations that may lead to a conformational shift favouring the open state. The mutation that favours the open state the most in our calculations is N501W with $\Delta S_{vib\ (open)} = 6.02x10^{-1}$ $J.K^{-1}$ and $\Delta S_{vib\ (closed)} = 2.30x10^{-1}$ $J.K^{-1}$ and a resulting $\Delta\Delta S_{vib}$ value of $3.72x10^{-1}$ $J.K^{-1}$ leading to occupancies compared to those of the wild type (in parenthesis) of 62.7% (25.8%) and 37.3% (74.2%) for the open and closed states respectively.

The COG-UK consortium (https://www.cogconsortium.uk/about/) monitors the appearance and spread of new strains of SARS-CoV-2. COG-UK recently detected a strain containing the mutation N501Y that has been observed to be spreading rapidly at the time of writing. We believe that shifts in occupancy may be in part responsible for its emergence. According to our calculations, the N501Y mutant shows $\Delta S_{vib\ (open)} = -1.60x10^{-2}$ $J.K^{-1}$ and $\Delta S_{vib\ (closed)} = 2.37x10^{-1}$ $J.K^{-1}$, with $\Delta\Delta S_{vib} = 2.53x10^{-1}$ $J.K^{-1}$. The predicted occupancies for the N501Y mutant compared to those of the wild type (in parenthesis) are 54.3% (25.8%) and 45.7% (74.2%) for the open and closed states, respectively. Therefore, the N501Y mutant shows a doubling of the occupancy of the open state with the concomitant decrease of the occupancy of the closed state. Additionally, this mutation was shown to also increase

binding affinity to the ACE2 receptor relative to the wild type with a $\Delta\log_{10}(K_{D,app})$ of 0.24 [53]. Therefore, we predict that N501Y has a strong potential to contribute to increased transmission. The calculations above were performed in the context of D614. However, the double mutant representing the N501Y mutation in the context of G614 has equally shifted occupancies with values of 35.06% and 64.04% for the open and closed state respectively. The recently observed A222V mutation on the other hand [54], does not show in our analysis any propensity of altering the occupancy of states with a negative $\Delta\Delta S_{vib}$ of $-1.64\text{x}10^{-2}$ $J.K^{-1}$. Predicted occupancies for A222 and V222 are nearly identical either in the context of D614 (WT) or the mutant containing G614.

Notice that N501Y has a $\Delta\Delta S_{vib}$ value of $2.53\text{x}10^{-1}$ $J.K^{-1}$ that is slightly below the $3.00\text{x}10^{-1}$ $J.K^{-1}$ threshold. Furthermore, inspection of Table 1 and Figure 10A shows at least one mutation with negative $\Delta\Delta S_{vib}$ (G413M) with increased predicted occupancy of the open state. These observations suggest that there may be many other mutations with $\Delta\Delta S_{vib}$ values below our set threshold that turn out to have augmented occupancies for the open state.

D614G shows that changes in the occupancy of conformational states can impact infectivity despite no changes or even weaker binding affinities [44]. A recent study [53] on binding and expression of Spike mutations within the RBD domain (positions 331 to 531) shows that several (but not all, see below) of the mutations that we predicted to have increased occupancy of the open state are associated with a decrease of binding affinity with ACE2. Incidentally, the data also shows that the mutations in Table 1 within the RBD produce stable and properly folded Spike proteins. As shown for D614G, infection does not rely on binding affinity alone, and even a strain with higher dissociation from ACE2 can bring about fitness advantages.

The mutation N501W is predicted to have the largest effect in augmenting the occupancy of the open state. This mutation is associated with stronger binding to ACE2 ($\Delta\log_{10}(K_{D,app})$=0.11) [53] relative to the wild type Spike (but lower than N501Y). Furthermore, N501W appears to have increased expression relative to the wild type with a $\Delta\log(MFI)$ of 0.1 compared to decrease in relative expression of -0.14 for

N501Y [53]. The authors note that changes in expression correlate with folding stability [53]. However, even with a Δlog(MFI) of -0.14, N501Y is viable and spreading. Therefore, N501W might be even more stable and infective.

We consider all mutations with increased predicted occupancy of the open state in Table 1 as good candidates for further experimental validation to better understand the role of binding and dynamics of Spike and their role in SARS-CoV-2 infectivity. Furthermore, we suggest that their appearance in outbreaks should be closely monitored.

3.4. Data Availability

Raw data and structures used to build the images presented here are available in a Github repository (https://github.com/nataliateruel/data_Spike). All vibrational entropy results are available for visualisation and analysis through a link to the dms-view open-access tool, available on GitHub [52] through the same URL above. On dms-view, it is possible to visualise the effects of different mutations for each residue of the Spike protein and visualise these on the 3D structure of Spike. Each site has 20 $\Delta\Delta S_{vib}$ values, one of them being zero (corresponding to the amino acid found in the wild type). The option max will show the top $\Delta\Delta S_{vib}$ score for each position. Therefore, it shows which mutation for that specific position represents the candidate with the highest predicted infectivity as defined here in terms of a propensity to higher occupancy of the open state. The option min will show the lowest score for each position and the mutation associated with the least predicted infectious candidate. The option median returns the median score, presenting a general trend for any given position, and var shows the variance between the results for each position, highlighting sites in which mutations to different residues lead to a broader range of $\Delta\Delta S_{vib}$ values. Furthermore, for the mutations for which occupancy was calculated, the data can be accessed through the same menu. As new occupancy data is calculated, it will be added to this resource. Readers interested on the occupancy of particular mutations not yet available are invited to contact the authors via email or through the GitHub repository. When selecting each specific point on

the first panel, it is possible to access all $\Delta\Delta S_{vib}$ values on the second panel and see the highlighted position in 3D on the structural representation.

The Najmanovich Research Group Toolkit for Elastic Networks (NRGTEN) including the latest ENCoM implementation is freely available at (https://github.com/gregorpatof/nrgten_package).

4. Conclusions

SARS-CoV-2 mutations are still arising and spreading around the world - the A222V mutation, reportedly responsible for many infections, emerged in Spain during the Summer of 2020 and since then has spread to neighbor countries [54]; In Denmark, new strains related to SARS-CoV-2 transmission in mink farms were confirmed in early October by the WHO and shown to be caused by specific mutations not previously observed with the novelty of back-and-forth transmission between minks and humans [55]. In the UK, a new strain containing N501Y is on the rise at the time of writing. Such occurrences point to the possibility that new mutations in SARS-CoV-2 may bring about more infectious strains.

Using the methods described in this paper, it is possible to predict potential variants that might have an advantage over the wild type virus insofar as these are the result of changes in occupancy of states and with the limitations of the simplified coarse-grained model employed here. In our analyses, flexibility properties and conformational state occupancy probabilities contribute to the infectivity of a SARS-CoV-2. Our results explain the behaviour of the D614G strain, the increased infectivity of SARS-CoV-2 relative to SARS-CoV as well as offers a possible explanation for the rise of new strains such as those harboring the N501Y mutation.

The results we present on SARS-CoV-2 Spike mutations have several limitations. First and foremost, some of the *in silico* mutation discussed may not be thermodynamically stable, may affect expression, cleavage, or binding to ACE2, and our approach does not consider that Spike is, in fact, a glycoprotein and the sugar molecules may have an effect on dynamics. However, the remarkable agreement between our model and experimental observations shows that the simplified model of Spike and the coarse-

grained methods used here allow us to calculate dynamic properties of Spike that are relevant to understand infection and epidemiological behavior. It is important to keep in mind that all of the mutations that we discuss in Table 1 that lay within positions 331 and 531 within the RBD domain were already experimentally validated and are viable [53]. However, we highlight the need for experimental validation of our predictions particularly for those candidates that we believe would help elucidate the extent of the effect of the conformational dynamics of Spike on infectivity. Beyond *in vitro* biophysical studies, experimental alternatives exist such as using pseudo-type viruses or virus-like-particles that would not require studying gain-of-function mutations using intact viruses. Alternatively, loss-of-function mutations can be created with intact viruses and compared to the wild type SARS-CoV-2 virus to validate the role of dynamics on infectivity.

To the best of our knowledge, this is the first time that a Normal Mode Analysis method is used to model the effect of mutations on the occupancy of conformational states opening a new opportunity in computational biophysics to create dynamic models of transitions between conformational states of proteins based on physical properties and sensitive to sequence variations. We hope that our results help public health surveillance programs decide on the risk posed by new strains, contribute to inform the research community in understanding SARS-CoV-2 infection mechanisms and open new possibilities in computational biophysics to study protein dynamics.

Acknowledgements

OM is the recipient of a PhD fellowship from the Fonds de Recherche du Quebec Nature et Technologie (FRQ-NT). RN is a Fonds de Recherche du Québec en Santé (FRQ-S) Senior Fellow. The authors would like to dedicate this work to the memory of Mordechai Najmanovich, Z"L, father of RN, who passed away from complications due to COVID-19 on November 26, 2020. RN would like to thank all healthcare workers, particularly ICU nurses and physicians at the Avista Adventist Hospital in Louisville, Colorado, for their efforts.

References

[1] Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N. and Bi, Y., 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet, 395(10224), pp.565-574.

[2] Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L. and Chen, H.D., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. nature, 579(7798), pp.270-273.

[3] World Health Organization, 2020. COVID-19 weekly epidemiological update, 1 December 2020. https://apps.who.int/iris/handle/10665/337341

[4] Centers for Disease Control and Prevention, 2012. SARS basics fact sheet. https://www.cdc.gov/sars/about/fs-sars.html.

[5] Wang, L.F., Shi, Z., Zhang, S., Field, H., Daszak, P. and Eaton, B.T., 2006. Review of bats and SARS. Emerging infectious diseases, 12(12), p.1834.

[6] Memish, Z.A., Perlman, S., Van Kerkhove, M.D. and Zumla, A., 2020. Middle East respiratory syndrome. The Lancet 395(10229): 1063-1077.

[7] Haagmans, B.L., Al Dhahiry, S.H., Reusken, C.B., Raj, V.S., Galiano, M., Myers, R., Godeke, G.J., Jonges, M., Farag, E., Diab, A. and Ghobashy, H., 2014. Middle East respiratory syndrome coronavirus in dromedary camels: an outbreak investigation. The Lancet infectious diseases, 14(2), pp.140-145.

[8] Zhao, Z., Li, H., Wu, X., Zhong, Y., Zhang, K., Zhang, Y.P., Boerwinkle, E. and Fu, Y.X., 2004. Moderate mutation rate in the SARS coronavirus genome and its implications. BMC evolutionary biology, 4(1), p.21.

[9] Koyama, T., Platt, D. and Parida, L., 2020. Variant analysis of SARS-CoV-2 genomes. Bulletin of the World Health Organization, 98(7), p.495.

[10] Letko, M., Marzi, A. and Munster, V., 2020. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. Nature microbiology, 5(4), pp.562-569.

[11] Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y. and Zhou, Q., 2020. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science, 367(6485), pp.1444-1448.

[12] Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A. and Li, F., 2020. Cell entry mechanisms of SARS-CoV-2. Proceedings of the National Academy of Sciences, 117(21), pp.11727-11734.

[13] Selvaraj, C., Dinesh, D.C., Panwar, U., Abhirami, R., Boura, E. and Singh, S.K., 2020. Structure-based virtual screening and molecular dynamics simulation of SARS-CoV-2 Guanine-N7

methyltransferase (nsp14) for identifying antiviral inhibitors against COVID-19. Journal of Biomolecular Structure and Dynamics (*in press*).

[14] Ali, A. and Vijayan, R., 2020. Dynamics of the ACE2–SARS-CoV-2/SARS-CoV spike protein interface reveal unique mechanisms. Scientific reports, 10(1), pp.1-12.

[15] Suárez, D. and Díaz, N., 2020. SARS-CoV-2 Main Protease: A Molecular Dynamic Study. *Journal of Chemical Information and Modelling* (*in press*).

[16] Pinto, D., Park, Y.J., Beltramello, M., Walls, A.C., Tortorici, M.A., Bianchi, S., Jaconi, S., Culap, K., Zatta, F., De Marco, A. and Peter, A., 2020. Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. Nature, pp.1-6.

[17] Rogers, T.F., Zhao, F., Huang, D., Beutler, N., Burns, A., He, W.T., Limbo, O., Smith, C., Song, G., Woehl, J. and Yang, L., 2020. Isolation of potent SARS-CoV-2 neutralizing antibodies and protection from disease in a small animal model. Science, 369(6506), pp.956-963.

[18] Cao, Y., Su, B., Guo, X., Sun, W., Deng, Y., Bao, L., Zhu, Q., Zhang, X., Zheng, Y., Geng, C. and Chai, X., 2020. Potent neutralizing antibodies against SARS-CoV-2 identified by high-throughput single-cell sequencing of convalescent patients' B cells. Cell182(1): 73-84.

[19] Deganutti, G., Prischi, F. and Reynolds, C.A., 2020. Supervised molecular dynamics for exploring the druggability of the SARS-CoV-2 spike protein. Journal of computer-aided molecular design (*in press*).

[20] Arantes, P.R., Saha, A. and Palermo, G., 2020. Fighting COVID-19 Using Molecular Dynamics Simulations. *ACS Cent. Sci.* 2020, 6, 10, 1654–1656.

[21] Karathanou, K., Lazaratos, M., Bertalan, É., Siemers, M., Buzar, K., Schertler, G.F., Del Val, C. and Bondar, A.N., 2020. A graph-based approach identifies dynamic H-bond communication networks in spike protein S of SARS-CoV-2. Journal of structural biology, 212(2), p.107617.

[22] Melero, R., Sorzano, C.O.S., Foster, B., Vilas, J.L., Martínez, M., Marabini, R., Ramírez-Aportela, E., Sanchez-Garcia, R., Herreros, D., Del Caño, L. and Losana, P., 2020. Continuous flexibility analysis of SARS-CoV-2 Spike prefusion structures. IUCrJ, 7(6):1059-1069.

[23] Verkhivker, G.M., 2020. Molecular Simulations and Network Modeling Reveal an Allosteric Signaling in the SARS-CoV-2 Spike Proteins. Journal of proteome research, 19(11): 4587-4608.

[24] Majumder, S., Chaudhuri, D., Datta, J. and Giri, K., 2020. Exploring the intrinsic dynamics of SARS-CoV-2, SARS-CoV and MERS-CoV spike glycoprotein through normal mode analysis using anisotropic network model. Journal of Molecular Graphics and Modelling, 102: 107778.

[25] Frappier, V. and Najmanovich, R.J., 2014. A coarse-grained elastic network atom contact model and its use in the simulation of protein dynamics and the prediction of the effect of mutations. PLoS Comput Biol, 10(4), p.e1003569.

[26] Frappier, V., Chartier, M. and Najmanovich, R.J., 2015. ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability. Nucleic acids research, 43(W1), pp.W395-W400.

[27] Frappier, V., Chartier, M. and Najmanovich, R., 2017. Applications of normal mode analysis methods in computational protein design. In Computational Protein Design (pp. 203-214). Humana Press, New York, NY.

[28] Frappier, V. and Najmanovich, R., 2015. Vibrational entropy differences between mesophile and thermophile proteins and their use in protein engineering. Protein Science, 24(4), pp.474-483.

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

[30] Gobeil, S., Janowska, K., McDowell, S., Mansouri, K., Parks, R., Manne, K., Stalls, V., Kopp, M., Henderson, R., Edwards, R.J. and Haynes, B.F., 2020. D614G mutation alters SARS-CoV-2 spike conformational dynamics and protease cleavage susceptibility at the S1/S2 junction. bioRxiv.

[31] Yuan, Y., Cao, D., Zhang, Y., Ma, J., Qi, J., Wang, Q., Lu, G., Wu, Y., Yan, J., Shi, Y. and Zhang, X., 2017. Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. Nature communications, 8(1), pp.1-9.

[32] Xiong, X., Qu, K., Ciazynska, K.A., Hosmillo, M., Carter, A.P., Ebrahimi, S., Ke, Z., Scheres, S.H., Bergamaschi, L., Grice, G.L. and Zhang, Y., 2020. A thermostable, closed SARS-CoV-2 spike protein trimer. Nature Structural & Molecular Biology, 27(10), pp.934-941.

[33] Šali, A. and Blundell, T.L., 1993. Comparative protein modelling by satisfaction of spatial restraints. Journal of molecular biology, 234(3), pp.779-815.

[34] Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. and Serrano, L., 2005. The FoldX web server: an online force field. Nucleic acids research, 33(suppl_2), pp.W382-W388.

[35] Case, D.A., 1994. Normal mode analysis of protein dynamics. Current Opinion in Structural Biology, 4(2), pp.285-290.

[36] Wako, H. and Endo, S., 2017. Normal mode analysis as a method to derive protein dynamics information from the Protein Data Bank. Biophysical reviews, 9(6), pp.877-893.

[37] Tang, Q.Y. and Kaneko, K., 2020. Long-range correlation in protein dynamics: Confirmation by structural data and normal mode analysis. PLoS computational biology, 16(2), p.e1007670.

[38] Mailhot, O. and Najmanovich, RJ. The NRGTEN Python package: an extensible toolkit for coarse-grained normal mode analysis of proteins, nucleic acids, small molecules and their interactions (in preparation).

[39] Xu, B., Shen, H., Zhu, X. & Li, G, 2011. Fast and accurate computation schemes for evaluating vibrational entropy of proteins. Journal of Computational Chemistry 32(15): 3188-3193.

[40] Marques, O. and Sanejouand, Y.H., 1995. Hinge-bending motion in citrate synthase arising from normal mode calculations. Proteins: Structure, Function, and Bioinformatics, 23(4), pp.557-560.

[41] Volz, E., Hill, V., McCrone, J.T., Price, A., Jorgensen, D., O'Toole, Á., Southgate, J., Johnson, R., Jackson, B., Nascimento, F.F. and Rey, S.M., 2020. Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity. Cell (*in press*).

[42] Li, Q., Wu, J., Nie, J., Zhang, L., Hao, H., Liu, S., Zhao, C., Zhang, Q., Liu, H., Nie, L. and Qin, H., 2020. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell, 182(5), pp.1284-1294.

[43] Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E.E., Bhattacharya, T., Foley, B. and Hastie, K.M., 2020. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell, 182(4), pp.812-827.

[44] Yurkovetskiy, L., Wang, X., Pascal, K.E., Tomkins-Tinch, C., Nyalile, T.P., Wang, Y., Baum, A., Diehl, W.E., Dauphin, A., Carbone, C. and Veinotte, K., 2020. Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell*, *183*(3), pp.739-751.

[45] Mohammad, A., Alshawaf, E., Marafie, S.K., Abu-Farha, M., Abubaker, J. and Al-Mulla, F., 2020. Higher binding affinity of Furin to SARS-CoV-2 spike (S) protein D614G could be associated with higher SARS-CoV-2 infectivity. *International Journal of Infectious Diseases (in press)*.

[46] Tang, L., Schulkins, A., Chen, C.N., Deshayes, K. and Kenney, J.S., 2020. The SARS-CoV-2 Spike Protein D614G Mutation Shows Increasing Dominance and May Confer a Structural Advantage to the Furin Cleavage Domain. *Preprints Preprints* **2020**, 2020050407

[47] Zhang, L., Jackson, C.B., Mou, H., Ojha, A., Rangarajan, E.S., Izard, T., Farzan, M. and Choe, H., 2020. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *bioRxiv* 2020.06.12.148726.

[48] Berger, I. and Schaffitzel, C., 2020. The SARS-CoV-2 spike protein: balancing stability and infectivity. *Cell research*, pp.1-2.

[49] Shu, Y. and McCauley, J., 2017. GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, *22*(13), p.30494.

[50] Elbe, S. and Buckland-Merrett, G., 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, *1*(1), pp.33-46.

[51] Kirchdoerfer, R.N., Wang, N., Pallesen, J., Wrapp, D., Turner, H.L., Cottrell, C.A., Corbett, K.S., Graham, B.S., McLellan, J.S. and Ward, A.B., 2018. Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. *Scientific reports*, *8*(1), pp.1-11.

[52] Hilton, S.K., Huddleston, J.L., Black, A., North, K., Dingens, A.S., Bedford, T. and Bloom, J.D., 2020. dms-view: Interactive visualization tool for deep mutational scanning data. *Journal of Open Source Software* 5:2353.

[53] Starr, T.N., Greaney, A.J., Hilton, S.K., Ellis, D., Crawford, K.H., Dingens, A.S., Navarro, M.J., Bowen, J.E., Tortorici, M.A., Walls, A.C., King, N.P., Veesler, D. and Bloom, J.D., 2020. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*, *182*(5), pp.1295-1310.

[54] Hodcroft, E.B., Zuber, M., Nadeau, S., Comas, I., Candelas, F.G., Stadler, T. and Neher, R.A., 2020. Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *medRxiv* 10.25.20219063.

[55] Munnink, B.B.O., Sikkema, R.S., Nieuwenhuijse, D.F., Molenaar, R.J., Munger, E., Molenkamp, R., Van Der Spek, A., Tolsma, P., Rietveld, A., Brouwer, M. and Bouwmeester-Vincken, N., 2020. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* (*in press)*.

Variant	$\Delta S_{vib\ (open)} - \Delta S_{vib\ (close)}$	Predicted Occupancy		
		Open state	Closed State	Difference (Open – Closed)
N501W	0.372	62.705%	37.295%	25.410%
G504I	0.349	40.491%	59.509%	-19.019%
G416E	0.314	30.111%	69.889%	-39.778%
G416Y	0.403	28.478%	71.522%	-43.045%
R403S	0.323	26.794%	73.206%	-46.412%
D467W	0.431	26.379%	73.621%	-47.243%
G252W	0.383	26.302%	73.698%	-47.396%
G404W	0.387	26.302%	73.698%	-47.396%
G252Q	0.394	26.295%	73.705%	-47.410%
G252H	0.353	26.268%	73.732%	-47.464%
G252E	0.365	26.224%	73.776%	-47.551%
K417G	0.315	26.179%	73.821%	-47.641%
G252C	0.401	26.164%	73.836%	-47.672%
G252S	0.447	26.159%	73.841%	-47.682%
G252T	0.440	26.157%	73.843%	-47.686%
G252D	0.390	26.149%	73.851%	-47.702%
G413M	-0.521	26.147%	73.853%	-47.705%
G252M	0.481	26.134%	73.866%	-47.731%
G252P	0.403	26.120%	73.880%	-47.759%
K417D	0.518	26.114%	73.886%	-47.773%
D467Y	0.399	26.106%	73.894%	-47.788%
S161F	0.395	26.056%	73.944%	-47.888%
K417P	0.502	26.035%	73.965%	-47.930%
S161Y	0.303	25.980%	74.020%	-48.040%
K417E	0.494	25.971%	74.029%	-48.058%
D467P	0.308	25.954%	74.046%	-48.093%
R34Y	0.380	25.908%	74.092%	-48.184%
I468T	0.350	25.907%	74.093%	-48.186%
R355F	-0.697	25.889%	74.111%	-48.222%
S161I	0.391	25.855%	74.145%	-48.290%
G72W	0.404	25.851%	74.149%	-48.298%
T73F	0.376	25.844%	74.156%	-48.313%
Q14C	0.312	25.839%	74.161%	-48.322%
WT	0.000	25.837%	74.163%	-48.327%
Q14H	0.345	25.836%	74.164%	-48.328%
N61P	0.324	25.835%	74.165%	-48.330%
K417I	0.467	25.806%	74.194%	-48.389%
P491H	0.374	25.804%	74.196%	-48.391%

R355Y	-0.632	25.772%	74.228%	-48.456%
Q14F	0.350	25.765%	74.235%	-48.470%
Y369N	0.379	25.731%	74.269%	-48.538%
K417C	0.500	25.725%	74.275%	-48.549%
Q409A	0.403	25.719%	74.281%	-48.562%
F486M	0.310	25.658%	74.342%	-48.684%
R355W	-0.541	25.648%	74.352%	-48.705%
I231P	-0.566	25.622%	74.378%	-48.756%
G416S	0.331	25.503%	74.497%	-48.994%
K417M	0.310	25.467%	74.533%	-49.066%
G416N	0.306	25.319%	74.681%	-49.362%
P230Y	-0.547	25.248%	74.752%	-49.505%
Y489M	0.422	25.168%	74.832%	-49.663%
D111W	-0.547	25.162%	74.838%	-49.677%
F464L	-0.575	25.030%	74.970%	-49.941%
E465W	0.361	24.977%	75.023%	-50.046%
E465K	0.322	24.896%	75.104%	-50.208%
E465Y	0.391	24.798%	75.202%	-50.405%
D111F	-0.656	24.782%	75.218%	-50.436%
Y369V	0.382	24.711%	75.289%	-50.578%
L368N	0.339	24.650%	75.350%	-50.700%
L368G	0.330	24.581%	75.419%	-50.838%
P230D	-0.683	24.515%	75.485%	-50.969%
L368S	0.337	24.336%	75.664%	-51.329%
L368C	0.358	24.173%	75.827%	-51.654%
S469W	-0.523	23.995%	76.005%	-52.011%
L368A	0.383	23.954%	76.046%	-52.091%
L368P	0.331	23.829%	76.171%	-52.342%
Y369E	0.501	23.744%	76.256%	-52.513%
E465A	0.360	23.694%	76.306%	-52.612%
E465S	0.336	23.051%	76.949%	-53.897%
G404Y	0.414	22.941%	77.059%	-54.117%
V503D	0.303	22.614%	77.386%	-54.773%
G404H	0.341	22.321%	77.679%	-55.359%
R403N	0.331	21.822%	78.178%	-56.356%
R403P	0.308	21.716%	78.284%	-56.568%
N394K	-0.524	21.586%	78.414%	-56.828%
G404N	0.411	21.544%	78.456%	-56.911%
Y421G	0.455	14.552%	85.448%	-70.896%

G232E	-0.707	11.578%	88.422%	-76.844%
G232S	-0.686	8.996%	91.004%	-82.008%
G232V	-0.713	5.508%	94.492%	-88.983%
G232P	-0.647	5.042%	94.958%	-89.917%
G232C	-0.664	3.065%	96.935%	-93.870%
G232M	-0.609	3.033%	96.967%	-93.935%
G232Q	-0.688	2.773%	97.227%	-94.454%
G232T	-0.602	1.350%	98.650%	-97.300%

Table 1. Table 1. Putative mutations and associated $\Delta\Delta S_{vib}$ (in units of $J.K^{-1}$) and predicted occupancies for the open and closed states for the 64 mutants with $\Delta\Delta S_{vib}>0.3$ and the 20 mutants with lowest $\Delta\Delta S_{vib}$ scores. Their occupancy calculation results are shown for the open conformation, the closed conformation, and the difference between the two (closed – open). Mutations in dark red represent The colours are the same as those used to denote the different classes of effects in Figure 10A.

Variant	Experimental			Predicted		
	Open occupancy	Closed occupancy	Occupancy difference	Open occupancy	Closed occupancy	Occupancy difference
S-GSAS/D614	25.160%	74.840%	-49.680%	25.781%	74.219%	-48.439%
S-GSAS/G614	43.938%	56.062%	-12.123%	25.812%	74.188%	-48.376%
S-R/x2	0.000%	100.000%	-100.000%	25.578%	74.422%	-48.843%
S-R/PP/x1	0.000%	100.000%	-100.000%	25.532%	74.468%	-48.935%
S-R	18.646%	81.354%	-62.707%	25.599%	74.401%	-48.801%
S-R/PP	77.967%	22.033%	55.935%	25.848%	74.152%	-48.305%

Table 2. Experimental and predicted occupancies for the open and closed states and their difference for multiple SARS-CoV-2 variants. Experimental values obtained from Gobeil *et al.* [30] and Xiong *et al.* [32].

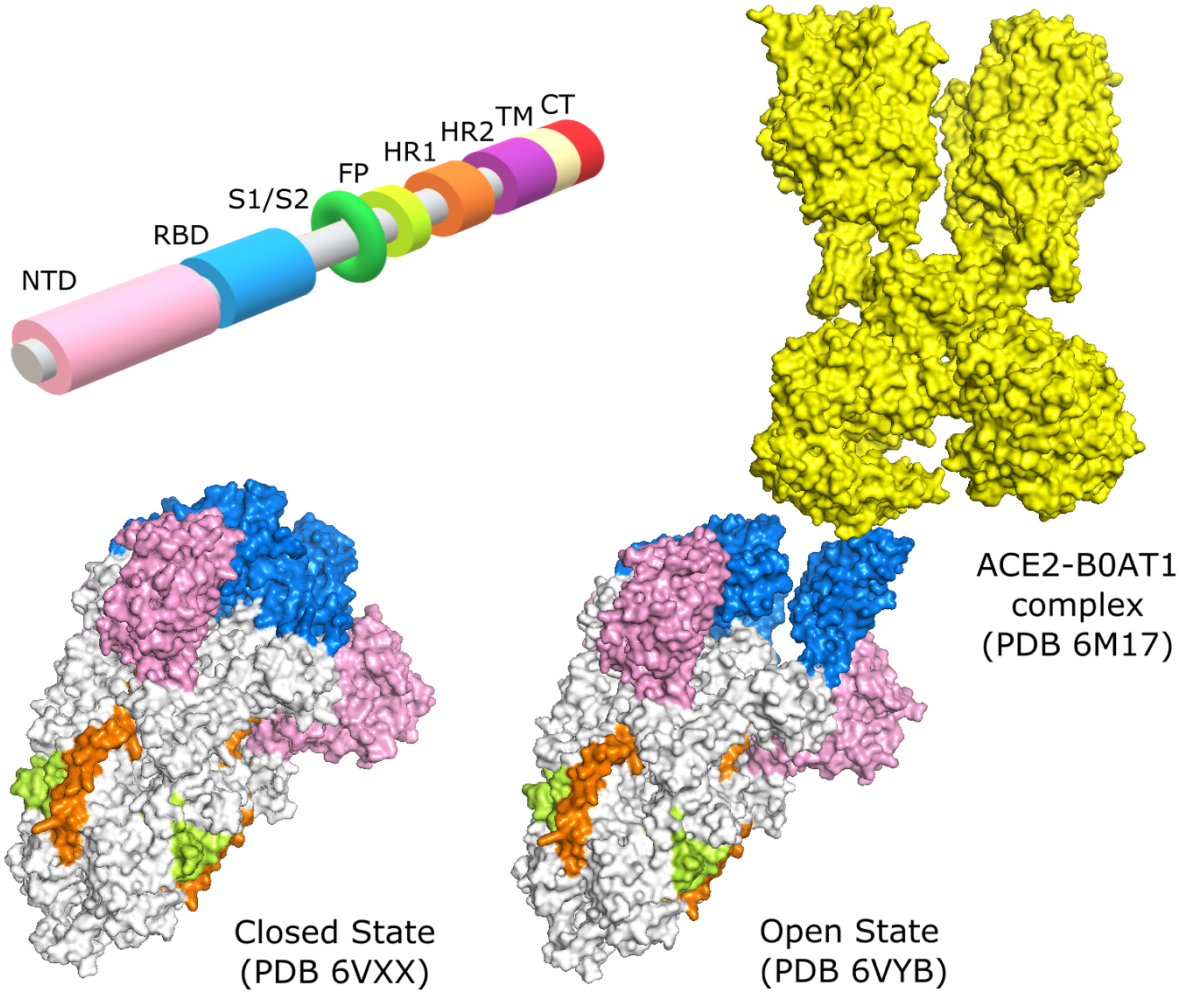


Figure 1. Domains of the Spike protein. N-Terminal Domain (NTD), Receptor Binding Domain (RBD), Subunit 1/Subunit 2 junction (S1/S2), Fusion Peptide (FP), Heptad Repeat 1 (HR1), Heptad Repeat 2 (HR2), Transmembrane Domain (TM), and the Cytoplasmic Tail (CT). Crystallography structure in the conformational state of all 3 RBD domains closed (PDB 6VXX) and of 1 RBD open (PDB 6VYB), binding to ACE2 (PDB 6M17).

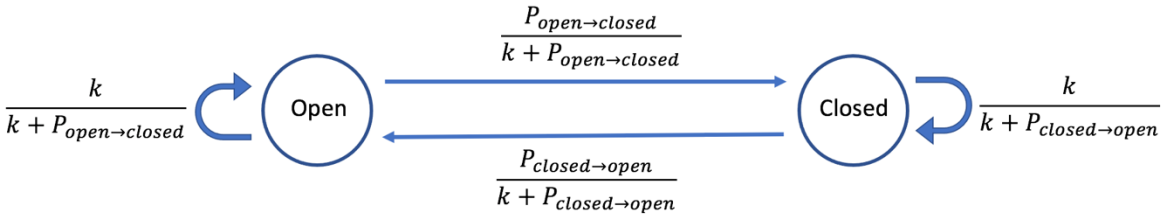
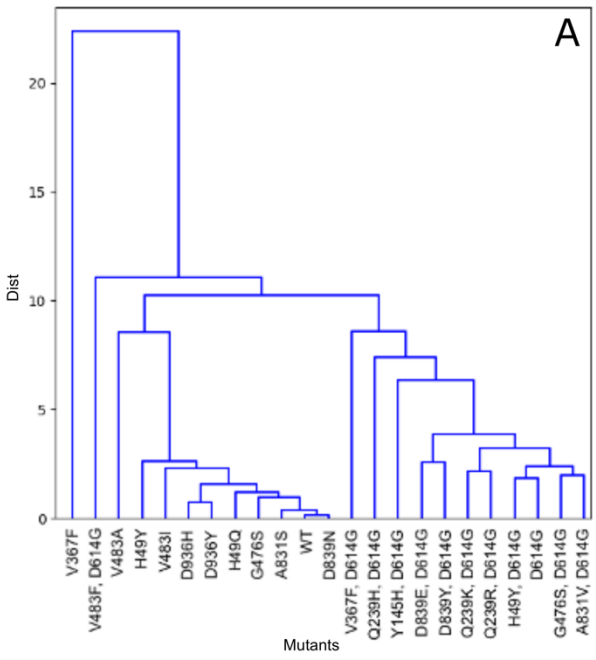


Figure 2. 2-state Markov chain of spike protein conformations.



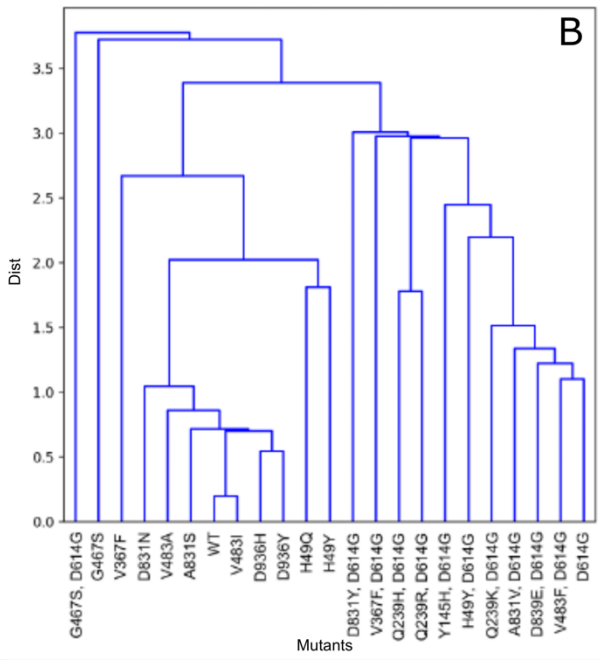


Figure 3. Dynamic Signature clustering for the closed (A) and open (B) state structures for WT and 22 mutants from GISAID (Table S1).

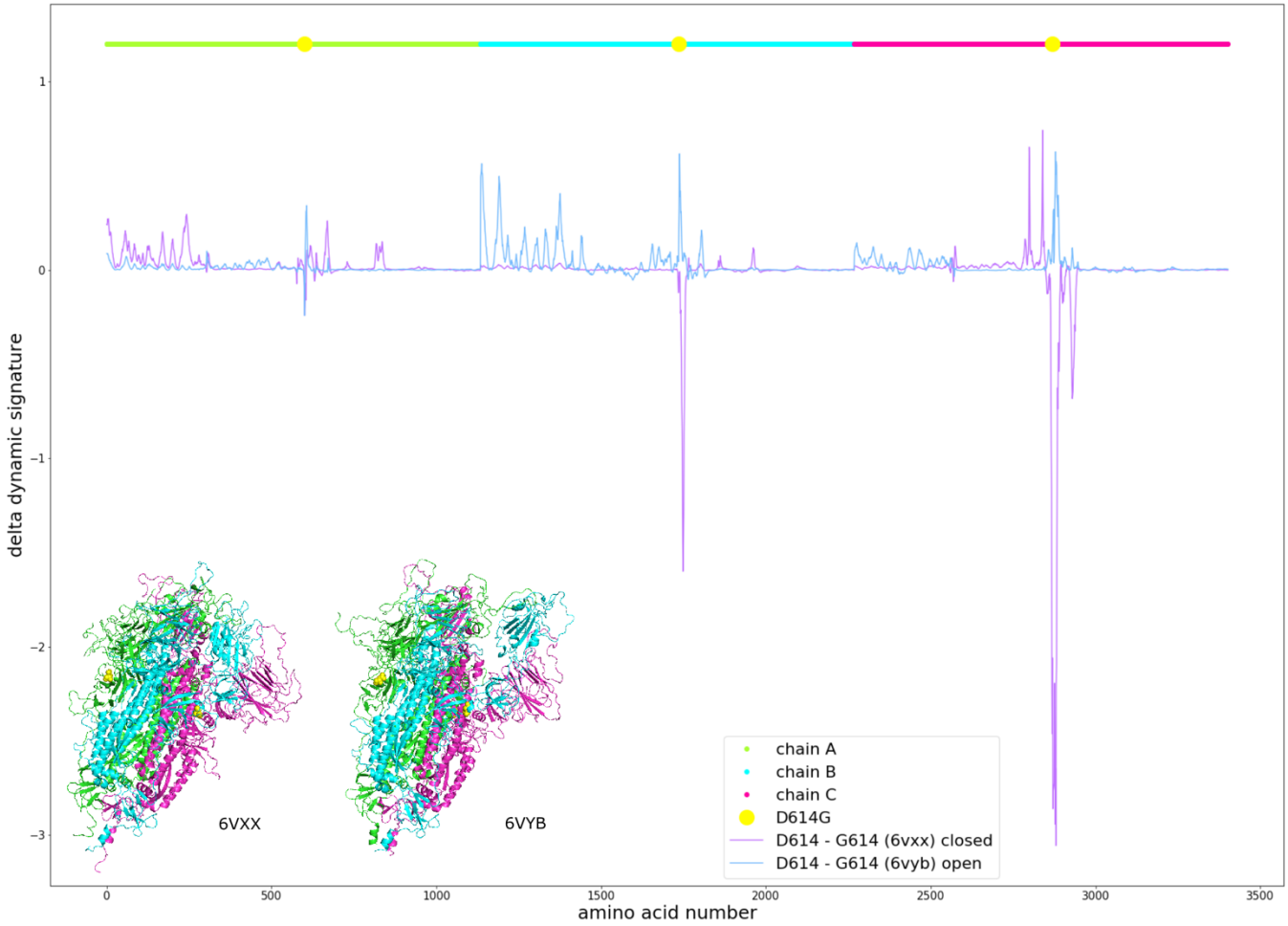


Figure 4. Effects of the D614G mutation on the Dynamic Signature of the closed (purple) and open B chain RBD (blue) structures, measured by the difference between the calculated b-factors of D614 and G614. Chains are represented in different colours and the position of the mutation is marked in yellow, using the same colours as for different regions of the structure as represented in the colours of the structures.

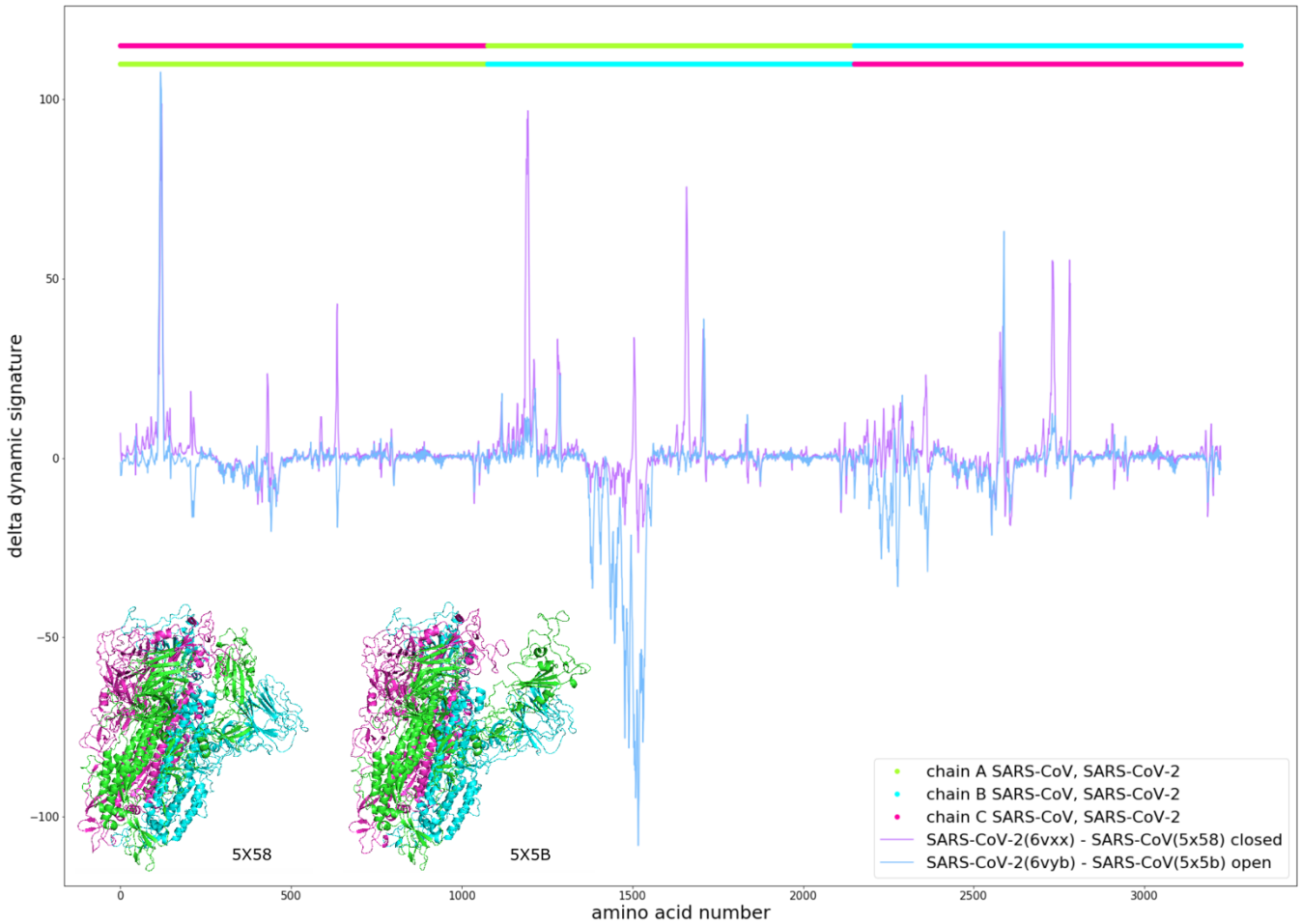


Figure 5. Comparison between SARS-CoV-2 and SARS-CoV. Dynamic Signature difference of the closed (purple) and open B chain RBD (blue) between aligned residues of the Spike protein from SARS-CoV-2 and SARS-CoV, with SARS-CoV chains represented in the top bar and equivalent colors in the structures and SARS-CoV-2 chains represented in the bar just below.

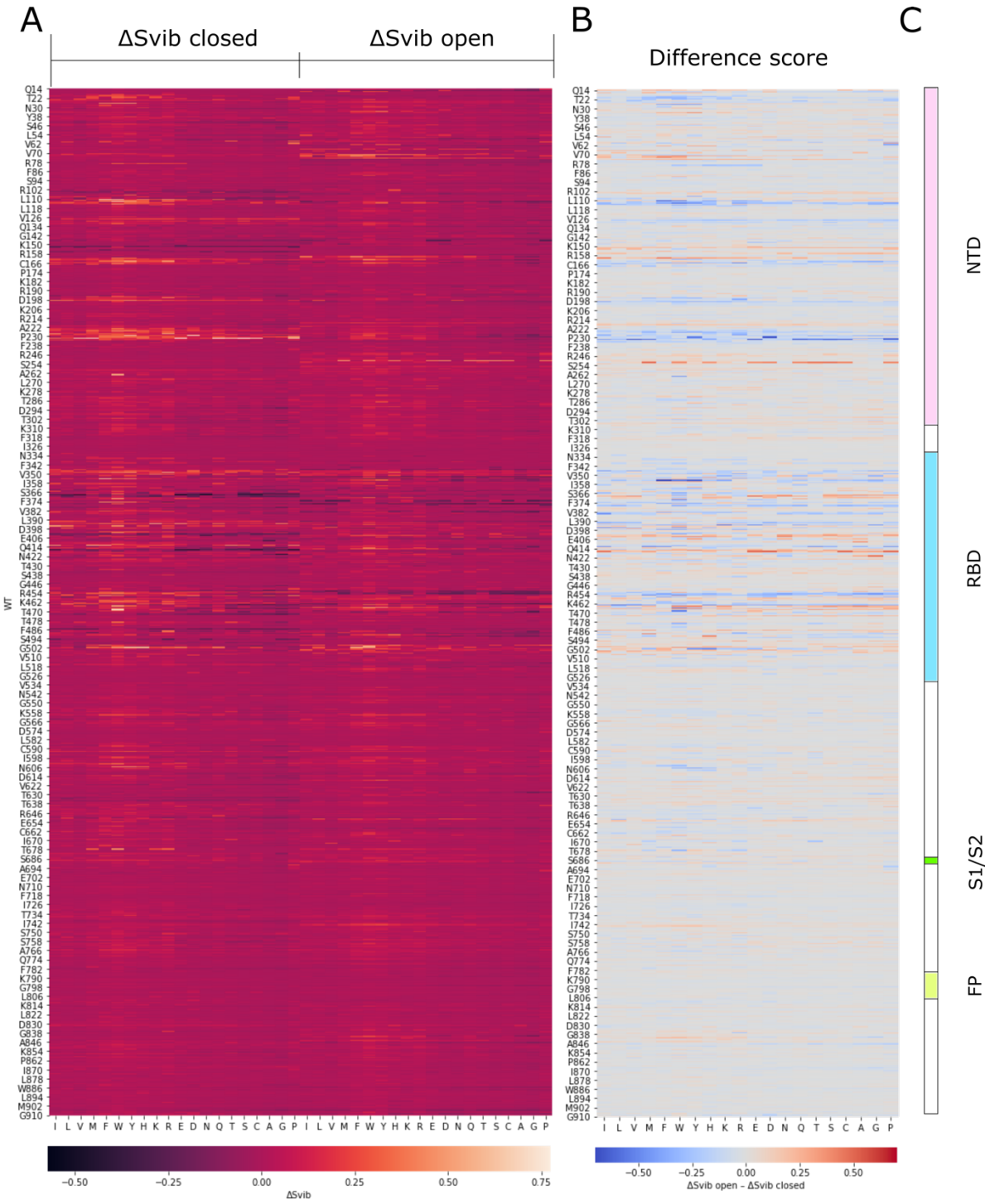


Figure 6. Heatmaps representing the values of ΔS_{vib} for the closed structure (A, left-hand section) and for the open structure (A, right-hand section), and the values of $\Delta\Delta S_{vib}$ (B) for every possible mutant in Spike from positions 14 to 913. Each column represents one of the 20 amino acids (repeated in the left heatmap). Notice that for each position (represented in a row), one particular column represents the value of the WT amino acid found at that position. Higher values of ΔS_{vib} are represented in yellow and lower values in dark purple. Higher values of $\Delta\Delta S_{vib}$ are represented in red and lower values in blue. The domain structure of Spike is represented in (C) for reference purposes.

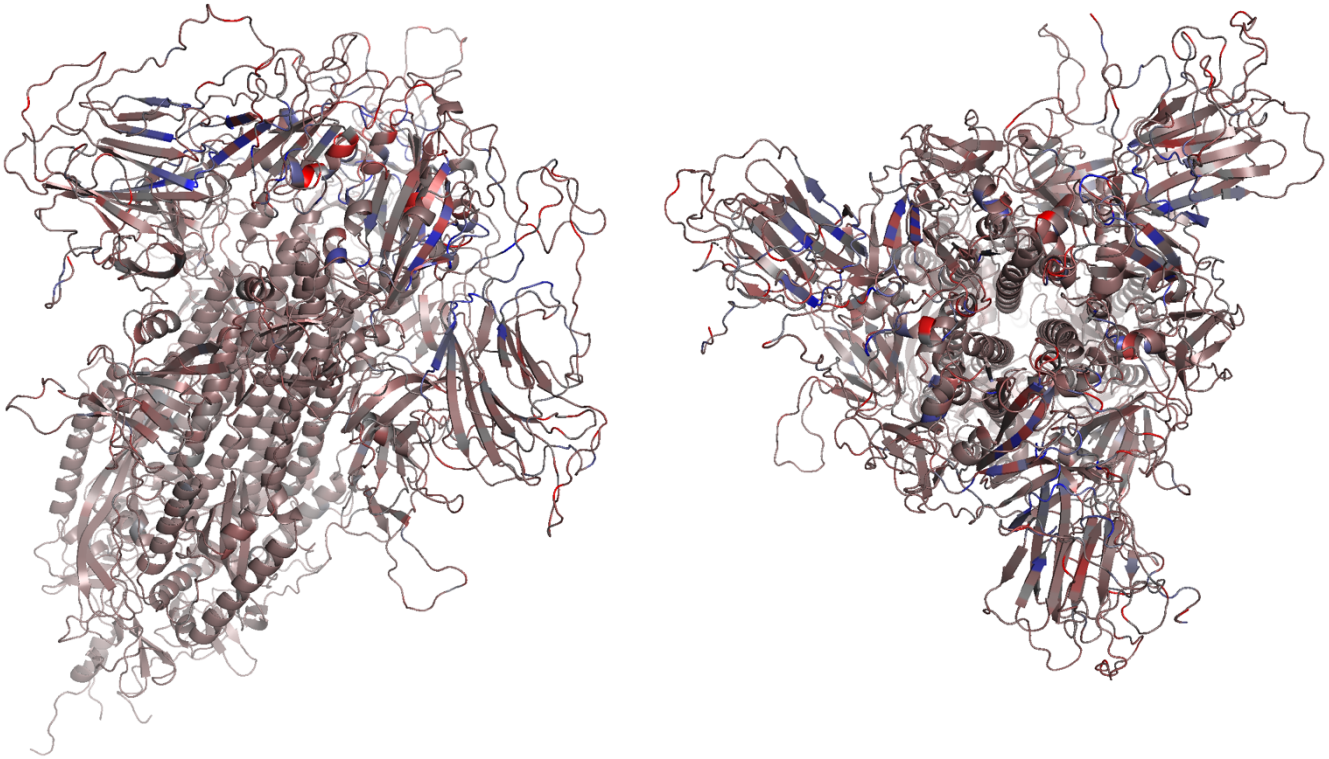


Figure 7. $\Delta\Delta S_{vib}$ scores represented in the structure of Spike from two angles according to the median value for each position and the same color scheme as in the Difference Score Heatmap (Fig 6B).

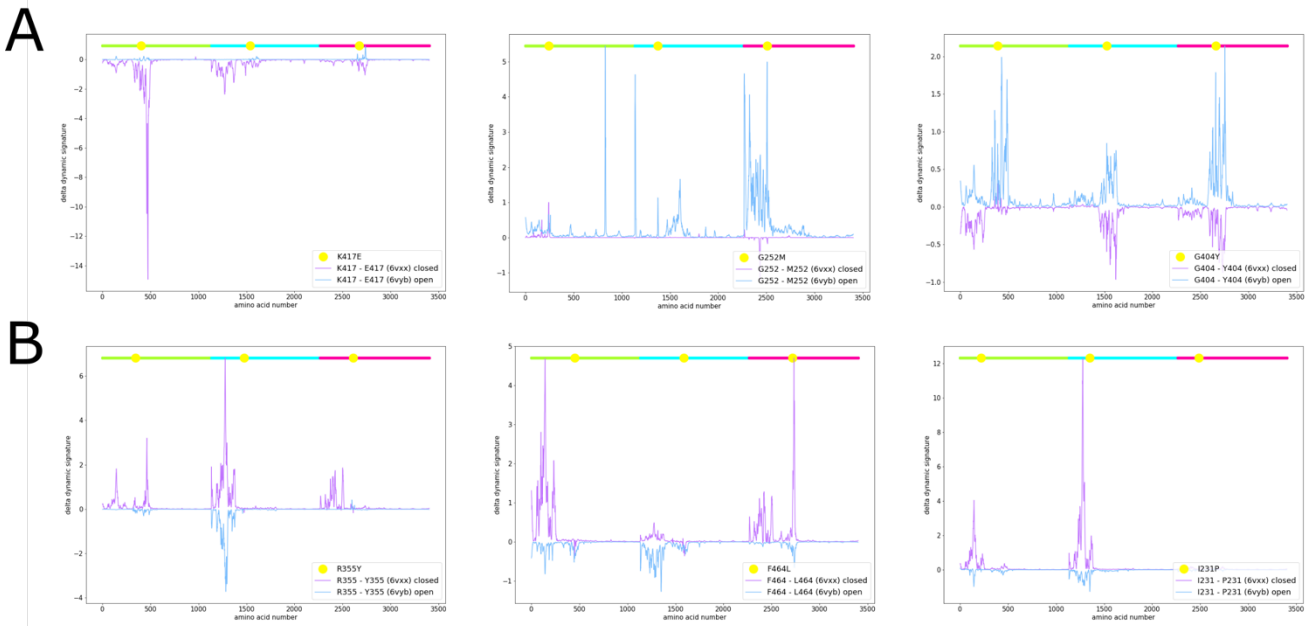


Figure 8. Dynamic Signature differences for three mutations among the top $\Delta\Delta S_{vib}$ scores (A) and bottom $\Delta\Delta S_{vib}$ scores (B) for closed (purple) and open (blue) B chain RBD structures.

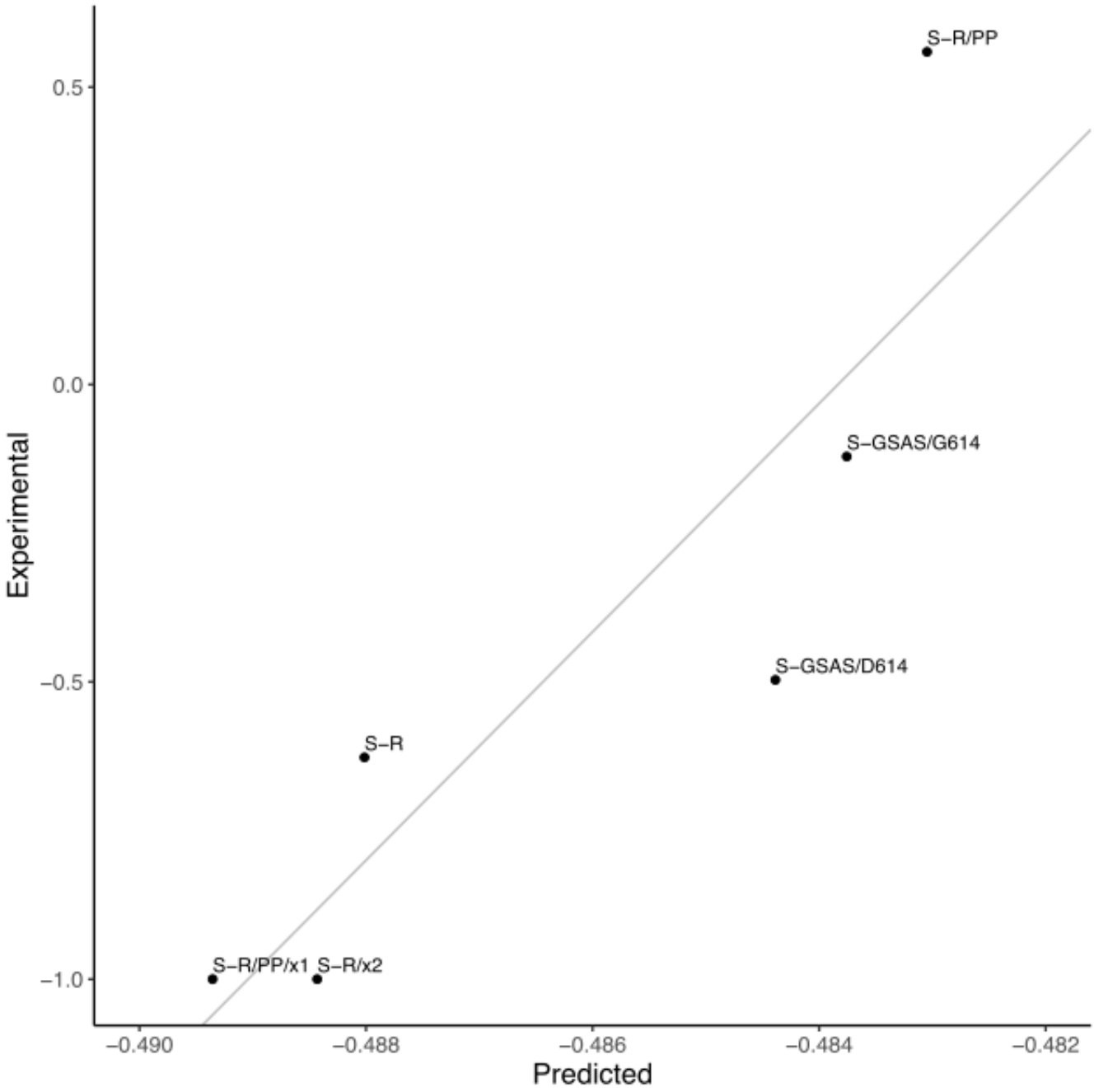


Figure 9. Difference in the occupancies for the open and the closed states (open – closed) for six variants of the Spike protein. Experimental values are represented on the Y-axis and the predicted values in the scale on the X-axis. Predicted values for the parameters k = 0.5, γ = 0.001. Represented linear fit of *Experimental = 192.011*Predicted + 92.9013*. Errors on the experimental measurements are not known.

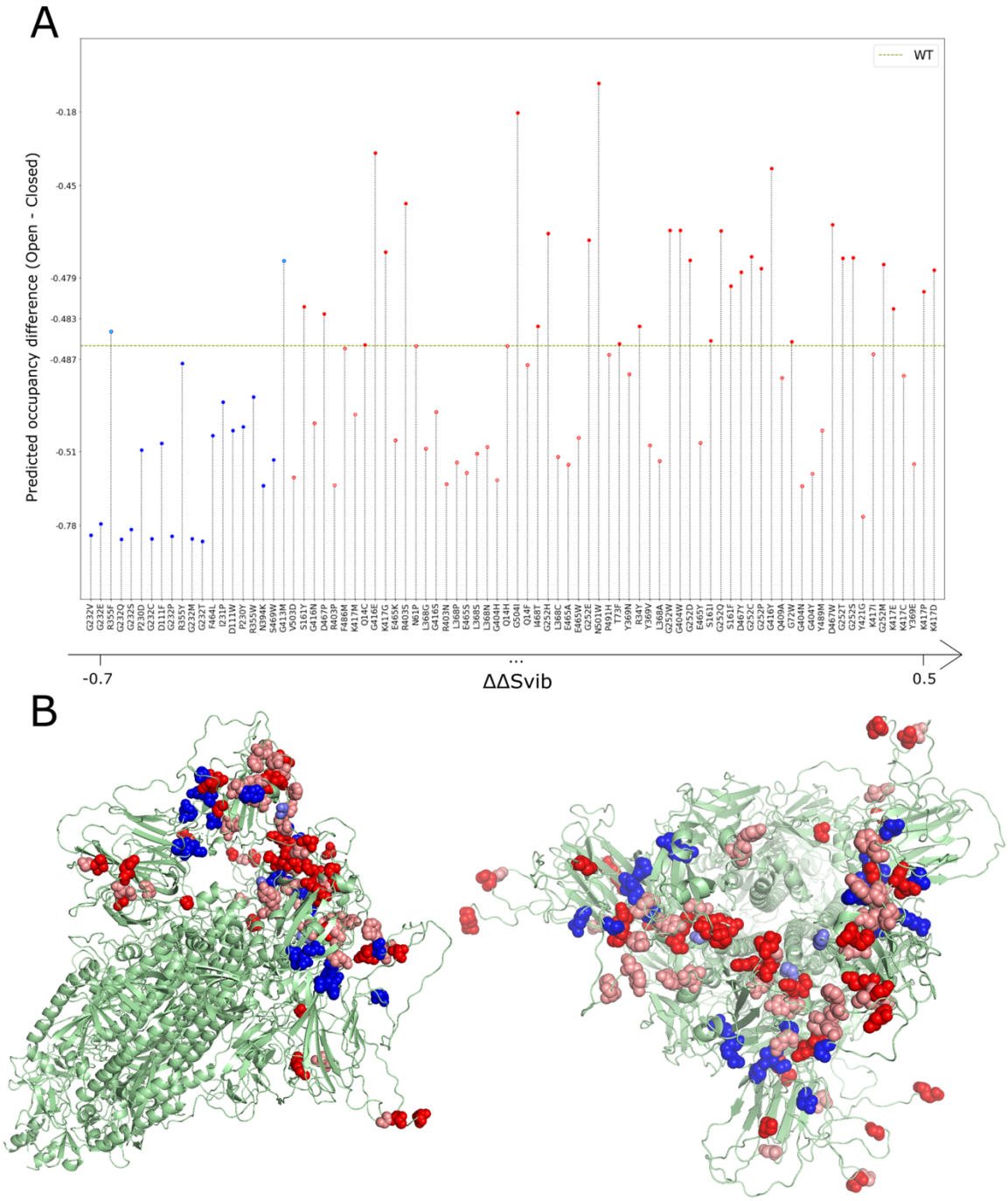


Figure 10. (A) Difference in the occupancies for the open and the closed states for the top 64 mutants with $\Delta\Delta S_{vib}$>0.3 (red) and the 20 mutants with lower $\Delta\Delta S_{vib}$ scores (blue). Occupancy difference for the WT is represented by the dashed green line. Y-axis based on the transformation of a symmetric logarithmic scale. (B) Two visualizations of the 6VYB structure highlighting the mutations. The bottom 20 mutant positions are marked in two shades of blue, with the darker shade indicating positions in which at least one mutant had an (open – closed) occupancy value smaller than wild type. The top 64 mutant positions are marked in two shades of red, with the darker shade indicating positions in which at least one mutant had an (open – closed) occupancy value higher than wild type.