Bifurcations and mutation hot-spots in SARS-CoV-2 spike protein

Xubiao Peng[1] and Antti J. Niemi[2,3,4,*]

[1] *Center for Quantum Technology Research and Key Laboratory of Advanced Optoelectronic Quantum Architecture and Measurements (MOE), School of Physics, Beijing Institute of Technology, Beijing 100081, China*

[2] *Nordita, Stockholm University, Roslagstullsbacken 23, SE-106 91 Stockholm, Sweden*

[3] *School of Physics, Beijing Institute of Technology, Haidian District, Beijing 100081, People's Republic of China*

[4] *Pacific Quantum Center and School of Biomedicine, Far Eastern Federal University, 690950 Sukhanova 8, Vladivostok, Russia*

Abstract

The spike protein is a most promising target for the development of vaccines and therapeutic drugs against the SARS-CoV-2 infection. But the apparently high rate of mutations makes the development of antiviral inhibitors a challenge. Here a methodology is presented to try and predict mutation hot-spot sites, where a small local change in spike protein's structure can lead to a large scale conformational effect, and change the protein's biological function. The methodology starts with a systematic physics based investigation of the spike protein's Cα backbone in terms of its local topology. This topological investigation is then combined with a statistical examination of the pertinent backbone fragments; the statistical analysis builds on a comparison with high resolution Protein Data Bank (PDB) structures. Putative mutation hot-spot sites are identified as proximal sites to bifurcation points that can change the local topology of the Cα backbone in an essential manner. The likely outcome of a mutation, if it indeed occurs, is predicted by a comparison with residues in best-matching PDB fragments together with general stereochemical considerations. The detailed methodology is developed using the already observed D614G mutation as an example. This is a mutation that could have been correctly predicted by the present approach. Several additional examples of potential hot-spot residues are identified and analyzed in detail, some of them are found to be even better candidates for a mutation hot-spot than D614G.

* corresponding author: Antti.Niemi@su.se

Significance statement: A novel approach to predict mutation hot-spots in SARS-CoV-2 spike protein is presented. The approach introduces new topology based techniques to biophysical protein research. For a proof-of-concept the approach is described with the notorious D614G mutation of the spike protein as an example. It is shown that this mutation could have been correctly predicted by the present methods. Several additional mutation hot-spots are then identified and a number of them are shown to be topologically similar to the observed D614G mutation. The methodology can be used to design effective drugs and antibodies against the spike protein. It can also be employed more generally, whenever one needs to search for and identify mutation hot-spots in a protein.

INTRODUCTION

The COVID-19 is a global public health emergency that continues to spread across the world. The disease is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It belongs to the same β genus with the SARS-CoV and MERS-CoV viruses that caused epidemics in 2002 and 2012 respectively. All three viruses are notorious for causing a disease with severe symptoms and even death. Several studies of the SARS-CoV-2 virus have been published including investigations on the source of infection [1–5], the mechanism of transmission [6–10], and the structure and function of its various proteins [11–17].

Of particular interest to vaccine and therapeutic drug development is the transmembrane spike glycoprotein that assembles into a homo-trimer to cover the virion surface, giving the virus its distinctive crown-like look. The spike protein monomer starts with a short signal peptide at the N-terminal, followed by two larger subunits called S1 and S2. The subunits mediate binding to host cell and membrane fusion, respectively. The Figure 1 identifies the major functional domains in a monomeric component of the spike protein [11, 12]:

The S1 subunit consists of residues between the sites 14-685. It starts with a N-terminal domain (NTD, residues 14-305) that participates to conformational changes during attachment. The NTD is followed by the receptor binding domain (RBD, residues 319-541) that initiates the viral entry by recognizing and binding to the host receptor angiotensin-converting enzyme 2 (ACE2). The junction segment between the S1 and S2 subunits includes several cleavage sites that can be hydrolyzed and cleaved by a variety of host proteases including furin and in particular TM protease serine 2, that prime the spike protein for membrane

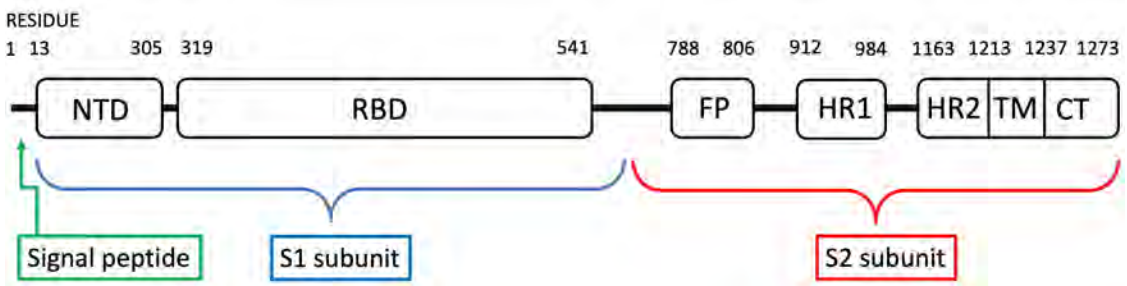

FIG. 1: (Color online) Spike protein subunit S1 consists of the N-terminal domain NTD and the receptor binding domain RBD. The subunit S2 consists of the fusion peptide FP, two heptapeptide repeat sequences HR1 and HR2, the transmembrane domain TM and the cytoplasm domain tail CT.

fusion [18]. The S2 subunit comprises the rest of the protein. It starts with the fusion peptide (FP, residues 788-806) that initiates the fusion. The two heptapeptide repeat sequences HR1 (residues 912-984) and HR2 (residues 1163-1213) form a six-helical bundle in the homotrimer that is essential for the viral fusion and entry. The transmembrane domain (TM with residues 1213-1237) anchors the protein to the viral membrane, and the S2 subunit ends in a cytoplasm domain tail (CT, residues 1237-1273).

There are two principal conformational states in the spike protein, called the closed state and the open state. The major difference between these two states is in the relative positioning of the RBD [13]. The closed state is the native prefusion state, in this state the RBD is surrounded by the NTD. The virus is also glycosylated with polysaccharide molecules that help it evade the host immune system. When the virus interacts with the host cell, the spike protein transits from the closed state to the open state with a major conformational change taking place in the S1 subunit. In the open state the RBD becomes exposed and can bind with the ACE2 of the cell membrane. The proteases of the host cell membrane then search for a cleavage site in the junction region between the S1 and S2 subunits, and hydrolyze the protein by cutting off the S1 subunit. This exposes the FP that resides in the head of the S2 subunit, and the FP undergoes a conformational change inserting itself into the host cell membrane. The fusion core that forms the junction between the HR1 and the HR2 then bends, and the viral envelope and the cell membrane become pulled together for fusion that completes the cell invasion [19].

The spike protein has a key role in processes that range from receptor recognition to viral attachment and entry into host cell. Therefore it has become a prevalent target in

both vaccine and therapeutic research that combat the SARS-CoV-2 virus. In particular the RBD domain and the HR1/HR2 domain are in the prime foci of these endeavors. It appears that the RBD is part of a highly mutable region, and a mutation in the RBD can change the interactions between the spike protein and the ACE2 of the host cell membrane. With limited understanding of mutation effects, the RBD can be a demanding target for the development of broad-spectrum antiviral inhibitors. The HR1/HR2 domain plays an important role in the infection itself. It is assumed to be relatively conserved, which makes the HR1/HR2 domain an attractive target for the development of fusion inhibitors to prevent a SARS-CoV-2 infection [20].

Apparently, there are many putative mutations that can take place in the SARS-CoV-2 virus. But most of them are likely biologically insignificant. This is because many mutations can only minimally alter the shape of the affected protein. Therefore they tend to have a limited impact on its biological functionality. Many mutations are also right out harmful to the virus and thus do not prevail. Probably the most prominent mutation that has taken place in the spike protein with documented epidemiological consequences, is the D→G substitution that occurred at its site 614, near the junction between the subunits S1 and S2. Apparently this mutation converted SARS-CoV-2 into a more transmissible form, and the D614G mutated virus now dominates in the global COVID-19 pandemic [21]: Since the G amino acid has a smaller size than the D amino acid, it is plausible that the D→G substitution increased the conformational flexibility of the spike protein around the S1-S2 junction domain. This may have enhanced the efficiency of the RBD to bind with the ACE2, increasing the exposure of the S1-S2 junction for cleavages by proteases of the host cell.

Here the goal is to introduce methodology that can help to identify those sites along the spike protein backbone, where a small localized change can cause a large conformational effect. If a mutation takes place at such a conformational hot-spot site, it can substantially alter the protein's biological function.

The approach builds on a geometric scrutiny of the spike protein's Cα backbone. As a piecewise linear polygonal chain the Cα backbone is subject to the universal rules that govern the shape of all space curves. In the case of a regular space curve the shape is determined by an interplay of its curvature and torsion; these are the two geometric quantities that describe how the curve bends and how it twists. An essential change in the local topology of a curve can only take place at a bifurcation point, where either the curvature or the torsion

vanish. In the case of a protein, when a mutation occurs at a residue that is proximal to such a bifurcation point, it has an increased potential to affect the protein's conformation and hence also change its biological activity. Thus the identification and analysis of these mutation hot-spot sites in the spike protein can help to predict its future evolution. It can aid the development of structure based therapeutic drugs and vaccines.

The Methods section first summarizes the known results that govern the geometry and the local topology of a smooth space curve. These results are then adapted to the special case of a discrete, piecewise linear chain such as the protein Cα backbone. In particular, the two curve specific bifurcations that can alter the local topology of a curve [22–24] called the inflection point perestroika and the bi-flattening perestroika, are adjusted to the Cα backbone.

The Results section applies the methodology to the SARS-CoV-2 spike protein; the Protein Data Bank structures 6VXX (closed state) 6VYB (open state)[13] and 6XS6 (D614G mutation)[25] are used. The potential mutation hot-spots are first identified by a tabulation of all those sites that are proximal to a flattening point. The details of the methodology are then worked out in the case of the known D614G mutation; the D→G substitution is correctly predicted by the present methodology. Additional examples are then analyzed in detail, including the identification of the likely amino acid substitution that may take place if a mutation occurs at a hot-spot.

MATERIALS AND METHODS

Regular space curves and the Frenet Equation

The geometry of a regular, analytic space curve that is not a straight line is governed by the Frenet equation [26]. To describe this equation, consider a parametric representation $\mathbf{x}(s) \in \mathbb{R}^3$ of a space curve, with $s \in [0, L]$ the arc-length parameter and L its fixed length; the curve can be open or closed. The curve is self-avoiding so that $\mathbf{x}(s_1) \neq \mathbf{x}(s_2)$ for all $s_1 \neq s_2$. The unit length tangent vector is

$$\mathbf{t} = \frac{d\mathbf{x}(s)}{ds} \equiv \mathbf{x}_s \tag{1}$$

The unit length binormal vector is

$$\mathbf{b} = \frac{\mathbf{x}_s \times \mathbf{x}_{ss}}{|\mathbf{x}_s \times \mathbf{x}_{ss}|} \tag{2}$$

and the unit length normal vector is

$$\mathbf{n} = \mathbf{b} \times \mathbf{t} \tag{3}$$

Together, the three vectors $(\mathbf{n}, \mathbf{b}, \mathbf{t})$ define the right-handed orthonormal Frenet frame at the regular point $\mathbf{x}(s)$ of the curve. The Frenet equation relates the frames along the curve,

$$\frac{d}{ds}\begin{pmatrix}\mathbf{n}\\ \mathbf{b}\\ \mathbf{t}\end{pmatrix} = \begin{pmatrix}0 & \tau & -\kappa\\ -\tau & 0 & 0\\ \kappa & 0 & 0\end{pmatrix}\begin{pmatrix}\mathbf{n}\\ \mathbf{b}\\ \mathbf{t}\end{pmatrix} \tag{4}$$

The curvature $\kappa(s)$ is a measure how the curve bends on the osculating plane that is spanned by $\mathbf{t}(s)$ and $\mathbf{n}(s)$, and the torsion $\tau(s)$ measures how the curve deviates from this osculating plane.

The Frenet frames can be introduced whenever the curvature $\kappa(s)$ is non-vanishing. In the limit where the curvature is very small in comparison to the torsion, but does not vanish

$$\left|\frac{\kappa(s)}{\tau(s)}\right| \to 0$$

the Frenet equation (4) becomes

$$\begin{aligned}\frac{d}{ds}(\mathbf{n} + i\mathbf{b}) &\approx -\, i\tau(\mathbf{n} + i\mathbf{b})\\ \frac{d}{ds}\mathbf{t} &\sim 0\end{aligned} \tag{5}$$

The limit is a framed, straight line. The framing rotates around the line at a rate and direction that is determined by $\tau(s)$.

Consider a curve that is subject to continuous local deformations and can change its shape freely; in the case of an open curve it is assumed that the end points remain fixed. The shape changes are subject to the following rules [22–24]:

At a point where the curvature $\kappa(s)$ vanishes, the Frenet frame can not be defined; here it is assumed that $\kappa(s)$ has only isolated simple zeroes. Those points where $\kappa(s) = 0$ are called the inflection points of the curve. A local deformation that retains the osculating plane can move an inflection point along the curve. But if a deformation lifts a point with $\kappa(s) = 0$ off the osculating plane the inflection point becomes removed. This implies that the co-dimension of an inflection point is two: An inflection point is not a local topological invariant of the curve, and a generic curve does not have any inflection points.

A point where the torsion vanishes $\tau(s) = 0$ is a flattening point. Unlike an inflection point, at least one single simple flattening point is generically present in a curve. A flattening point is a local topological invariant that can not be removed by any continuous local deformation of the curve, it can only be moved along the curve.

A solution of the Frenet equation determines a framing of the curve, and either $\mathbf{b}(s)$ or $\mathbf{n}(s)$ or their linear combination can be chosen as the framing vector. The self-linking of the curve describes how it links with a nearby curve that is obtained by pushing points of the original curve along the framing vector. The self-linking number is a local topological invariant of the curve, in the absence of an inflection point the self-linking number can not change.

When the shape of a curve changes freely, an isolated inflection point generically occurs at some instance. When an inflection point appears the curve undergoes a bifurcation that is called an inflection point perestroika [22]. This bifurcation can change the number of flattening points: Since the torsion $\tau(s)$ changes its sign at a simple flattening point, and since the curvature $\kappa(s)$ is generically not zero, an inflection point perestroika commonly changes the number of simple flattening points by two.

When the shape of a curve changes so that a pair of flattening points comes together they combine into a single bi-flattening point. A bi-flattening point can then be removed by a further, generic local deformation of the curve. Similarly, a bi-flattening point can first be created by a proper local deformation of a curve, and when the curve is further deformed the bi-flattening point can resolve into two separate simple flattening points. When either of these occur the curve undergoes a bifurcation that is called a bi-flattening perestroika [22].

Inflection point perestroika and bi-flattening perestroika are the only two bifurcations where the number of flattening points can change. Furthermore, the number of flattening points and the self-linking number that is determined by the Frenet framing are the only two curve specific local topological invariants that can be assigned to a curve [22–24].

The number of flattening points and the self-linking number are independent topological invariants, but in the presence of an inflection point they can interfere with each other. For example, when a curve is deformed so that two simple flattening points become combined and disappear in a bi-flattening perestroika, the self-linking number in general does not change. However, if the bi-flattening perestroika occurs in conjunction of an inflection point perestroika the self-linking number can change: If the torsion is initially positive and two

flattening points combine and disappear with the passage of an inflection point, the self-linking number increases by one. But if the torsion is initially negative the self-linking number decreases by one.

The protein Cα backbone and the discrete Frenet equation

The protein Cα backbone determines a piecewise linear polygonal chain, with Cα atoms at the vertices. When the protein folds and unfolds the backbone chain bends and twists. Various shape changes also commonly occur in a biologically active protein. Since a polygonal chain such as the Cα backbone can always be thought as a limiting case of a regular, analytic space curve, the changes in the shape of the Cα backbone are subject to same rules that govern the local topology of any regular space curve. In particular, the three essential shape deformations *i.e.* discrete variants of inflection point perestroikas, bi-flattening perestroikas, and changes in a self-linking number should all have a profound role in physiological processes.

The discrete Frenet frame formalism [27] describes the geometry of a piecewise linear chain with vertices $\mathbf{r}_i$ $(i = 1, ..., N)$ that in the case of a protein backbone are the space coordinates of the Cα atoms. A line segment defines the unit tangent vector

$$\mathbf{t}_i = \frac{\mathbf{r}_{i+1} - \mathbf{r}_i}{|\mathbf{r}_{i+1} - \mathbf{r}_i|} \tag{6}$$

It points from the center of the ith Cα atom towards the center of the $(i+1)^{st}$ Cα atom, in the case of a protein backbone. The unit binormal vector is

$$\mathbf{b}_i = \frac{\mathbf{t}_{i-1} \times \mathbf{t}_i}{|\mathbf{t}_{i-1} \times \mathbf{t}_i|} \tag{7}$$

and the unit normal vector is

$$\mathbf{n}_i = \mathbf{b}_i \times \mathbf{t}_i \tag{8}$$

and in the case of a protein these two vectors are virtual in that they do not point towards any particular atom. The orthonormal triplet $(\mathbf{n}_i, \mathbf{b}_i, \mathbf{t}_i)$ defines a discrete version of the Frenet frames (1)-(3) at each position $\mathbf{r}_i$ along the chain. In lieu of the curvature $\kappa(s)$ and the torsion $\tau(s)$ there are now their discrete versions, the bond angles κ_i and the torsion angles τ_i. The values of these angles can be computed from the discrete Frenet frames. The bond angles are

$$\kappa_{i+1,i} \;=\; \arccos\left(\mathbf{t}_{i+1} \cdot \mathbf{t}_i\right) \;\equiv\; \kappa_{i+1}(\mathbf{r}_i, \mathbf{r}_{i+1}, \mathbf{r}_{i+2}) \tag{9}$$

and the torsion angles are

$$\tau_{i+1,i} \;=\; \mathrm{sign}\{\mathbf{b}_i \times \mathbf{b}_{i+1} \cdot \mathbf{t}_i\} \cdot \arccos\left(\mathbf{b}_{i+1} \cdot \mathbf{b}_i\right) \;\equiv\; \tau_{i+1}(\mathbf{r}_{i-1}, \mathbf{r}_i, \mathbf{r}_{i+1}, \mathbf{r}_{i+2}) \tag{10}$$

It is notable that the value of the bond angle κ_i is evaluated from three, and the value of the torsion angle τ_i is evaluated from four consecutive vertices.

Conversely, when the values of the bond and torsion angles are all known, the discrete Frenet equation

$$\begin{pmatrix} \mathbf{n}_{i+1} \\ \mathbf{b}_{i+1} \\ \mathbf{t}_{i+1} \end{pmatrix} = \begin{pmatrix} \cos\kappa\cos\tau & \cos\kappa\sin\tau & -\sin\kappa \\ -\sin\tau & \cos\tau & 0 \\ \sin\kappa\cos\tau & \sin\kappa\sin\tau & \cos\kappa \end{pmatrix}_{i+1,i} \begin{pmatrix} \mathbf{n}_i \\ \mathbf{b}_i \\ \mathbf{t}_i \end{pmatrix} \tag{11}$$

computes the Frenet frame at the vertex $\mathbf{r}_{i+i}$ from the frame at the preceding vertex $\mathbf{r}_i$. A continuum limit of the discrete Frenet equation coincides with the continuum Frenet equation (4) [27].

It should be noted that besides the Frenet framing, a protein backbone can also be framed in other ways, for example using the peptide planes: A peptide plane frame is built starting with the vector $\mathbf{t}_i$ and the unit normal vector of the ensuing peptide plane. Unlike the Frenet frames that can not be introduced if a bond angle vanishes, the peptide plane frames can always be introduced along a protein Cα backbone. When both frames exist they are closely related [27].

The fundamental range of the bond angles is $\kappa_i \in [0, \pi]$ and in the case of the torsion angles $\tau_i \in [-\pi, \pi)$. For visualization purposes, the bond angles κ_i can be identified with the latitude angle of a two-sphere that is centered at the ith Cα atom; the north pole coincides with the inflection point $\kappa_i = 0$. The torsion angles $\tau_i \in [-\pi, \pi)$ correspond to the longitudinal angle on the sphere, it increases in the counterclockwise direction around the tangent vector and the value $\tau_i = 0$ of flattening points coincides with the great semi circle that starts from north pole and passes through the tip of the normal vector $\mathbf{n}$ to the south pole. The sphere can be stereographically projected onto the complex (x, y) plane. A projection from the south pole is

$$z = x + iy \;\equiv\; \sqrt{x^2 + y^2}\, e^{i\tau} \;=\; \tan\left(\kappa/2\right)\, e^{i\tau} \tag{12}$$

as shown in figure 2: The north pole *i.e.* the point of inflection with $\kappa = 0$ becomes mapped to the origin (x, y)=$(0, 0)$ and the south pole $\kappa = \pi$ is sent to infinity.

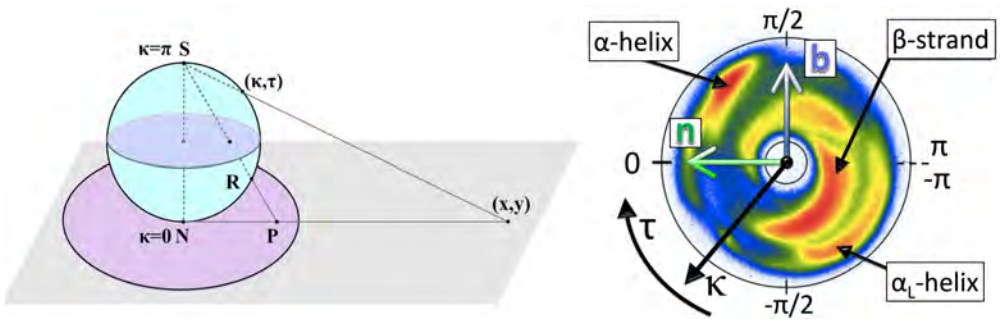

FIG. 2: (Color online) a) Stereographic projection of two sphere from the north pole (N) with latitude κ and longitude τ. b) The stereographically projected Frenet frame map of backbone Cα atoms, with major secondary structures identified. Also shown is the directions of the Frenet frame normal vector $\mathbf{n}$ and bi-normal vector $\mathbf{b}$; the vector $\mathbf{t}$ points upwards from the figure. Colour coding corresponds to the number of PDB entries with red large, blue small and white none.

The Cα backbone can be visualized on the stereographically projected two sphere as follows [28]: At each Cα atom, introduce the corresponding discrete Frenet frames (6)-(8). The base of the ith tangent vector $\mathbf{t}_i$ that is located at the position $\mathbf{r}_i$ of the ith Cα atom coincides with the centre of a two-sphere with the vector $\mathbf{t}_i$ pointing towards its north pole. Now, translate the sphere from the location of the ith Cα atom to the location of the $(i+1)^{th}$ Cα atom, without any rotation of the sphere with respect to the ith Frenet frames. Identify the direction of $\mathbf{t}_{i+1}$, *i.e.* the direction towards the Cα atom at site $\mathbf{r}_{i+2}$ from the site $\mathbf{r}_{i+1}$, on the surface of the sphere in terms of the ensuing spherical coordinates (κ_i, τ_i). When this construction is repeated for all the protein structures in Protein Data Bank that have been measured with better than 2.0 Å resolution, the result can be summarized by the map that is shown in figure 2 b). The color intensity correlates directly with the statistical distribution of the (κ_i, τ_i); red is large, blue is small and white is none. The map describes the direction of the Cα carbon at $\mathbf{r}_{i+2}$ as it is seen at the vertex $\mathbf{r}_{i+1}$, in terms of the Frenet frames at $\mathbf{r}_i$.

Approximatively, the statistical distribution in figure 2 b) is concentrated within an annulus that corresponds to the latitude angle values (in radians) $\kappa = 0.57$ and $\kappa = 1.82$ shown in the Figure. The exterior of the annulus is a sterically excluded region. The entire interior is sterically allowed, but there are very few entries in this region. The four major secondary structure regions, α-helices, β-strands, left-handed α-helices and loops, are identified according to their PDB classification. For example, (κ, τ) values (in radians) for

which

$$\begin{cases} \kappa_i \approx \frac{\pi}{2} \\ \tau_i \approx 1 \end{cases} \tag{13}$$

describes a right-handed α-helix, and values for which

$$\begin{cases} \kappa_i \approx 1 \\ \tau_i \approx \pm\pi \end{cases} \tag{14}$$

describes a β-strand.

In the case of a regular space curve both $\kappa(s)$ and $\tau(s)$ are smooth functions and the inflection points and the flattening points are easily identified as the points where $\kappa(s) = 0$ and $\tau(s) = 0$. In a crystallographic protein structure where the Cα positions are experimentally determined, an inflection point is detectable as a very small value of the bond angle κ_i at the proximal Cα vertex. Similarly, the presence of a simple flattening point can be deduced from a very small torsion angle value at the proximal vertex, accompanied by a sign change in τ between two neighboring vertices; if the sign of τ does not change the proximal vertex has the character of a bi-flattening point. Accordingly, when searching for Cα atoms where essential shape changes such as inflection point or bi-flattening perestroikas can take place, the natural points to start are the neighborhoods of vertices where either $\kappa_i \approx 0$ or $\tau_i \approx 0$. These are the likely locations where a small change in the shape of the backbone can change its local topology, with a potentially substantial change in the protein's biological function.

From Figure 2 b) one observes that inflection points *i.e.* small κ_i values are extremely rare in crystallographic protein structures. Indeed, a generic space curve does not have any inflection points. At the same time general arguments state that generically at least one flattening point can be expected to be present. As shown in Figure 2 b) flattening points where $\tau_i \approx 0$ do appear even though they are relatively rare in protein structures. Moreover, it is observed from the Figure that at a flattening point the bond angle values are mostly either $\kappa_i \approx 1$ or $\kappa_i \approx \pi/2$.

Since the torsion angles are defined $mod(2\pi)$, in the case of discrete Frenet frames there is an additional structure: There is the line $\tau_i = \pm\pi$ in Figure 2 b) where the torsion angle has a 2π discontinuity, hence it can change sign by crossing the line. This multivaluedness is absent in regular space curves, with $\tau(s)$ a single-valued continuous function. But the limits $\tau_i \to \pm\pi$ can be thought of as the small curvature and large torsion limits of the equations

(5). Notably, $\kappa_i = 1$ and $\tau_i \to \pm\pi$ corresponds to an ideal, straight β-strand. Therefore a point on the line $\tau_i \to \pm\pi$ (or in its immediate vicinity) will be called a *β-point* in the sequel, even when the bond angle has a value that is different from its β-strand value $\kappa \approx 1$.

The multi-valuedness of the torsion angle τ_i affects the local topological invariants, in the case of a discrete chain. This is exemplified in the Figures 3. These Figures depict the three characteristic examples of a protein loop structure that interpolates between two right-handed α-helices; similar considerations apply when the interpolation is between any two generic points.

For clarity: Whenever a Figure similar to those in Figures 3 is drawn in the sequel, the two neighboring vertices *i.e.* values (κ_i, τ_i) and $(\kappa_{i+1}, \tau_{i+1})$ are always to be connected by a virtual segment that is a straight line on the disk. Even though in reality those values of (κ, τ) that are on the straight line may not correspond to any actual atomic position along the piecewise linear Cα backbone.

In all Figures 3 the initial points (A) and the final points (E) are fixed and located near the α-helical region where the torsion angle has a positive value close to $\tau \approx 1$. At the points (B) and (C) the torsion angles are always negative in the Figures 3. At the point (D) the torsion angle is positive.

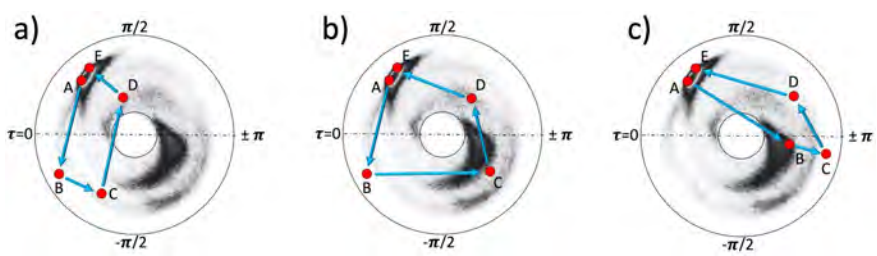

FIG. 3: (Color online) A chain with five vertices. a) With two crossings of the flattening point line $\tau = 0$. b) With one crossing of the flattening point line $\tau = 0$ and one crossing of the line of β-points $\tau = \pm\pi$. c) With two crossings of the line of β-points $\tau = \pm\pi$. In both a) and c) the folding index Eq(15) vanishes but in b) the folding index has value -1 since the trajectory encircles the disk center once in counterclockwise direction.

In the Figure 3 a) the chain proceeds from point (A) to point (B) by crossing the line of flattening points $\tau = 0$. It then continues to point (C). From there it proceeds and crosses the line of flattening points a second time, to arrive at point (D) where the torsion angle returns to a positive value. The chain then continues and ends at point (E). In this case the local topology is fully analogous to that of a regular curve. In particular, the two flattening

points along the chain can be removed by a bi-flattening perestroika that lifts both points (B) and (C) above the line $\tau = 0$ to positive τ_i values, without the chain crossing the inflection point $\kappa = 0$ at the center of the disk.

In the Figure 3 b) the chain proceeds to the point (B), again by crossing the line of flattening points $\tau = 0$. From there it proceeds to point (C) that is located near the β-stranded region. The chain then crosses the line of β-points $\tau = \pm\pi$ as it proceeds to point (D) where the torsion angle is positive. The chain finally ends at point (E) in the α-helical region. In this case there is only one flattening point along the chain, since the second time the torsion angle changes its sign at a β-point. The chain encircles the inflection point once in the counterclockwise direction and neither the flattening point nor the β-point can be removed without encountering an inflection point perestroika along the chain.

The Figure 3 b) motivates to introduce a winding number termed folding index Ind_f [29] for a backbone chain segment between sites n_1 and n_2. The folding index classifies loop structures and entire folded proteins by counting the number of times the chain encircles the inflection point. Its value can be obtained from the equation (15)

$$Ind_f = \left[\frac{\Gamma}{2\pi}\right] \quad \text{with} \quad \Gamma = \sum_{i=n_1+2}^{n_2-2} \begin{cases} \tau_{i+1} - \tau_i - 2\pi & \text{if } \tau_{i+1} - \tau_i > \pi \\ \tau_{i+1} - \tau_i + 2\pi & \text{if } \tau_{i+1} - \tau_i < -\pi \\ \tau_{i+1} - \tau_i & \text{otherwise} \end{cases} \tag{15}$$

with $[x]$ the integer part of x. Here Γ is the total rotation angle (in radians) that the chain makes around the inflection point in Figures such as 3. The folding index is positive when the rotation is clockwise, and negative when the rotation is counterclockwise. In the Figure 3 b) the folding index has the value $Ind_f = -1$ since the chain encircles the inflection point once in counterclockwise direction.

In the Figure 3 c) the torsion angle changes its sign between points (A) and (B) and between points (C) and (D). Now the sign change occurs at β-points. The general rules of local topology for the sign change at a β-point are like those at a flattening point. In particular, a pair of β-points can either be created or removed in a bifurcation called β-perestroika that is akin a bi-flattening perestroika. In the Figure 3 c) a β-perestroika occurs if the vertices (B) and (C) are both moved upwards across $\tau = \pm\pi$ line without the chain crossing the inflection point $\kappa = 0$ at the center of the disk.

The folding index is a local topological invariant. For example, in Figure 3 b) its value

does not change unless the chain is deformed so that either the flattening point between (A) and (B) or the β-point between (C) and (D) becomes removed by an inflection point perestroika, that either converts the β-point into a flattening point by a deformation that takes the chain in Figure 3 b) into the chain in Figure 3 a) or converts the flattening point into a β-point by a deformation that takes the chain in Figure 3 b) into the chain in Figure 3 c). In both cases the final folding index vanishes.

The three examples in Figures 3 summarize all essential aspects of local topology that are encountered in the case of a discrete chain. In particular, the line of flattening points and the line of β-points have a very similar character in terms of local topology. They can be interchanged by an inflection point perestroika that also changes the folding index.

Finally, the Figure 2 b) is akin the Newman projection of stereochemistry. The vector $\mathbf{t}_i$ which is denoted by the red dot at the center of the figure, points along the backbone from the proximal Cα at $\mathbf{r}_i$ towards the distal Cα at $\mathbf{r}_{i+1}$, and the colour intensity displays the statistical distribution of the $\mathbf{r}_{i+2}$ direction. Moreover, unlike the Ramachandran map the figure 2 b) provides non-local information on the backbone geometry. The Ramachandran map can only provide localized information in the immediate vicinity of a single Cα carbon, but the information content in the figure 2 b) map extends over several peptide units. As shown in [30], the Cα backbone bond and torsion angles (κ_i, τ_i) are sufficient to reconstruct the entire backbone, but the Ramachandran angles are not.

.

RESULTS

The Figure 2 b) shows that there are relatively many vertices in the vicinity of β-points $\tau \approx \pm\pi$. These vertices include the regular β-strands that are often stabilized by interatomic forces such as hydrogen bonds. At the same time there are relatively few vertices near flattening points and there are no stable regular secondary structures with $\tau \approx 0$. The proximity to a flattening point appears to make a Cα position relatively unstable in a protein structure.

The Table I lists all those SARS-CoV-2 spike protein Cα-sites that are proximal to a flattening point in the PDB structures 6VXX (closed state) and 6VYB (open state). Here a torsion angle value is determined to be proximal to a flattening point when $|\tau_i| < 0.2$. In

terms of distance, this is less than the radius of a carbon atom. The histogram in Figure 4

Residue	CL/P	RMSD/P	MC/P	CL/A	RMSD/A	MC/A	CL/F	RMSD/F	MC/F	closed/open
40D	24	0.69	A	I	0.69	V/S/N	X	0.58	L/A/S/T	c / o
46S	24	0.92	A	16	0.82	G	10	0.63	G	c / o
59F	7	0.44	G	26	0.18	G	XI	0.24	G	c / o
98S	XII	0.84	A	14	0.78	V	X	0.3	A/T	o-A
103G	9	0.77	I/T	VI	0.21	A	25	0.38	A/V/T	c / o-A
109T	II	0.74	G	6	0.54	G	X	0.55	A/L	c / o
114T	26	0.87	A	XI	0.71	A/G/S	12	0.35	S/V	o-C
124T	17	0.74	G	III	0.5	G	21	0.5	G	o-BC
138D	XII	0.39	S/A	6	0.39	G	25/X	0.72	A/V	o-C
287D	24	0.17	A	VI	0.18	A/G	25	0.23	T/L/A/V	c / o
316S	24	0.21	A	VI	0.45	A/G	25	0.41	T/L/A/V	o
324E	24	0.27	A	I	0.39	A	25	0.34	T/L/A/V	c / o-B
376T	VII	0.83	A/G	I	0.36	A	25	0.34	A/V	c / o-BC
382V	26	1.18	A/T	IX	0.72	A/T	25	0.86	A/T	c
394N	24	0.73	A	I	0.36	A	25	0.13	T/A	c / o
422N	IX	1.06	G	IX	1.13	T	20	0.87	E/V/D	c
423Y	IX	1.13	G	20	0.87	L/H	25	0.27	T/A/L/V	o-AB
459S	XII	1.10	A	14	0.58	V	12	0.82	V	o-A
464F	17	0.21	G	26	0.13	G	XI	0.23	G	c / o
545G	17	0.69	G	26	0.51	G	XI	0.6	G	o-B
565F	V	1.15	S/T/D/E	VI	0.21	A/G	25	0.28	A/T/V/L	o-C
575A	24	0.22	A	VI	0.19	A/G	25	0.33	A/T/V	o-AB
605S	IX	0.83	G	20	1.0	L	V	1.14	L/A/T/V	o
614D	XII	0.76	S/A/N/G	26	0.75	G	10	0.42	G	c / o-A
758S	V	0.76	T	13	0.94	S	IV	0.61	L	c / o
892A	X	0.71	V	VIII	1.145	V	25/12	0.73	V	o-A
974S	19	0.74	H	I	0.78	A	V	0.59	A/V/L	c / o-C
1035G	IX	0.93	G	IX	0.51	T	12	0.62	S/V	c / o-BC
1041D	VII	0.67	A/T/G	26	0.54	G	IV	0.67	L/N	c / o
1058H	17	0.64	G/A/S	III	0.64	G	21	0.52	G	o-B
1080A	24	0.87	A	I	0.26	A	25	0.26	A/T/V	c / o
1084D	24	0.78	A/S	14	1.33	N/V/S	III	0.3	N	c / o
1092E	24	0.48	A	14	0.87	V/S	21	0.56	G	c / o
1093G	14	0.87	L	21	0.56	V	28	0.76	A/V/L	o-B
1131G	VII	0.53	A/G/T	6	0.43	G	XI	0.39	G	o-AB

TABLE I: Flattening points in spike protein PDB structures 6VXX and 6VYB. Columns are organized as follows: Residue number with present amino acid. CL stands for cluster with smallest RMSD (in Å); clusters are taken from [31]. MC stands for most common mutagenic amino acid substitutions in the cluster. P stands for preceding, A stands for adjacent and F stands for following fragment. Closed (c) and open (o) denotes the state where flattening point is observed. A,B,C denotes the monomer of spike protein.

shows the distribution of the sites in relation to spike protein backbone. There are relatively many entries in the NTD and RBD domains, and in the fusion core between the HR1 and HR2 domains. Notably the number of proximal sites is also different in the closed and open states of the spike protein, there are more proximal states in the open state. Thus a transition between the two states involves changes in local topology that affect in particular the NTD and RBD domains, and the HR1-HR2 junction.

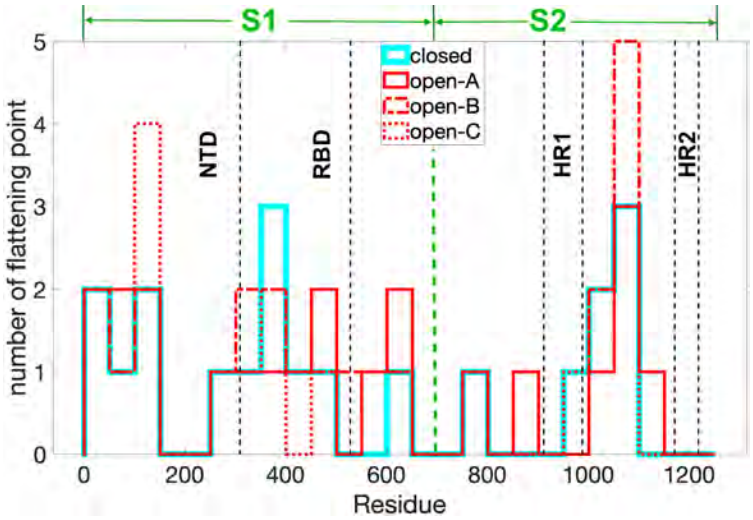

FIG. 4: (Color online) The histogram shows the distribution of sites that are proximal to a flattening point along the spike protein backbone, in blocks of 40 sites. The proximal sites accumulate in the NTD and RBD domains, and in the fusion core between the HR1 and HR2 domains. The PDB structures are 6VXX for closed state and 6VYB for open state.

In the case of a bond angle, the value is considered small when $\kappa_i < 0.5$. General arguments state that inflection points are not generic, and there are indeed very few sites that are close to an inflection point. The smallest value is $\kappa_i = 0.39$ and it is located at the site 103 that also appears in Table I. Thus sites that are proximal to an inflection point are not considered separately.

The local geometry of all the residues that are proximal to a flattening point has been studied and the potential for a mutation hot-spot at a residue that is proximal to a flattening point has been estimated. For this, a combination of statistical analysis and stereochemical constraints has been utilized. The Table I also summarizes these findings. The statistical analysis employs the classification scheme of Protein Data Bank structures described in [31]. This scheme decomposes a Cα backbone into fragments that consist of backbone segments with six successive sites; in accordance with Eq. (9), (10) a fragment with six sites determines three pairs of bond and torsion angles. In [31] the fragments that appear in high resolution PDB structures have been organized into disjoint clusters. To assign a cluster to a fragment, there must be at least one other fragment in the same cluster within a prescribed RMS cut-off distance; in [31] the cut-off is 0.2 Å. Two clusters are then disjoint, when the RMSD between any fragment in the first cluster and any fragment in the second cluster exceeds this RMS cut-off distance. It was found that around 38% of protein loops in

the high resolution PDB structures can be decomposed into fragments that belong to twelve disjoint clusters, labeled I-XII in [31]. When fragments from an additional set of 30 disjoint clusters are included, the coverage increases to $\sim 52\%$ [31].

In the Table I both cluster sets I-XII and 1-30 appear; the notation of [31] is used throughout in the following. Beyond these two sets, the clusters become increasingly smaller, and in the present study those smaller clusters have not been considered. The somewhat low resolution of the available spike protein structures in comparison to the very high resolution structures used in [31] does not justify a more detailed scrutiny.

The clusters in the Table I have been identified as follows: A pair of bond and torsion angles (κ_i, τ_i) at the ith site of the spike protein that is a proximal site to a flattening point can be assigned to three different clusters. The first cluster describes the bond and torsion angles for the sites $(i-2, i-1, i)$; this cluster is labelled P (for Preceding) in the Table I. The second cluster that is labelled A (for Adjacent) in the Table I describes the angles for sites $(i-1, i, i+1)$. The third cluster is labelled F (for Following) and it describes the angles at spike protein sites $(i, i+1, i+2)$. The cluster that provides the best match to the spike protein is listed in the Table I and it is determined as follows: Let (x_a, y_a, z_a) denote the six space coordinates of a segment that corresponds to three consecutive pairs of (κ, τ) values in the spike protein. Let $(x_{k,a}, y_{k,a}, z_{k,a})$ be the corresponding six space coordinates of a kth fragment in each of the clusters of [31]. The best matching cluster in Table I is the one that contains a fragment with the minimal root-mean-square distance (RMSD) to the given spike protein segment (in units of Ångström):

$$\Delta_{min} = \min_{clusters} \sqrt{\min_k \sum_{a=1}^{6} (x_a - x_{k,a})^2 + (y_a - y_{k,a})^2 + (z_a - z_{k,a})^2} \tag{16}$$

Once the best matching cluster is determined, the corresponding statistical distribution of amino acids found in [31] is used to identify those amino acids that are most probable to appear at the ith site of the spike protein, in case a mutation occurs. If the size of this statistically most probable amino acid is smaller than the size of the present amino acid at the ith site, a substitution by mutation is considered to be sterically possible. But if the size is larger, a substitution likely requires an extended rearrangement of the spike protein conformation and this can be energetically costly. The Table I lists the most probable substitutions that are predicted in this manner.

Example: The D614G mutation of spike protein

The Figures 5 exemplify how the methodology works, in the case of the known D→G mutation at the site 614 of the spike protein. Each of the three Figures depicts the three bond and torsion angle pairs for the three backbone segments (P, A and F respectively) that include the angles of the site 614.

In the sequel, in Figures such as 5, the vertices that mark the Cα sites of spike protein are always identified by stars. These stars are color-coded and organized according to increasing site number following the Cyan-Magenta-Yellow (CMY) color table. The dotted background comprises the Cα sites of all the fragments in the corresponding best matching cluster. The background sites are always ordered similarly using the Red-Green-Blue (RGB) color table. The adjacent histogram displays the amino acid distribution in the best matching cluster, using the data that is obtained from [31].

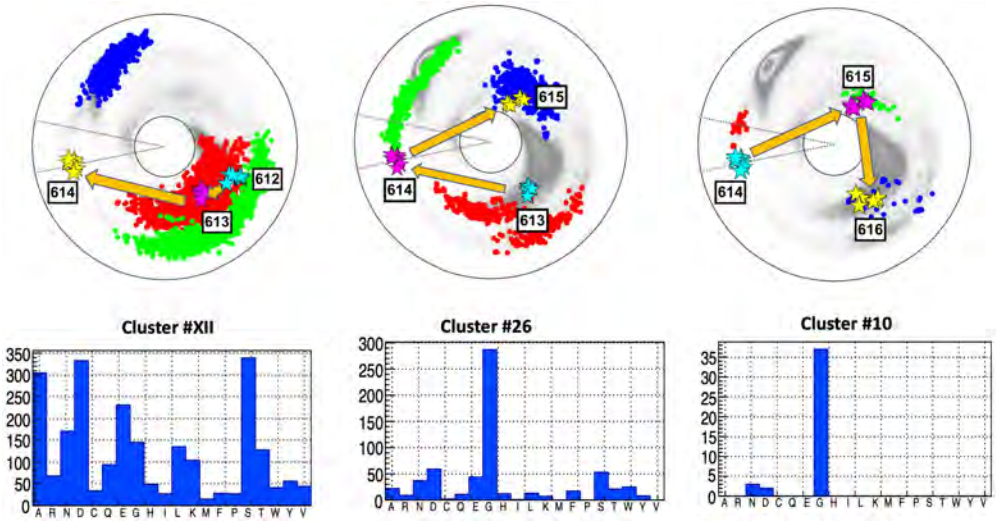

FIG. 5: (Color online) The (κ_i, τ_i) spectrum for $i = 612 - 616$; this engages sites $610 - 617$ along the Cα backbone. a) The spectrum for preceding sites P:$(612, 613, 614)$, in the background of the best matching cluster #XII in [31]. The most common amino acids in the cluster are S, D, A, E, N, G. b) The spectrum for adjacent sites A:$(613, 614, 615)$, in the background of the best matching cluster #26 in [31]. The most common amino acid in the cluster is G. c) The spectrum for following sites F:$(614, 615, 616)$, in the background of the best matching cluster #10 in [31]. The most common amino acid in the cluster is G.

The Figure 5 a) shows the bond and torsion angles for the sites P:(612,613,614) of the spike protein. The best matching preceding (P) cluster is also shown, it is the cluster #XII of [31]. From Table I the RMSD (16) between the spike protein segment and the cluster is $\Delta_{min} \approx 0.76$Å which is clearly larger than the 0.2 Å cut-off used in [31]. But the resolution

at which the spike protein PDB structures (6VXX, 6VYB and 6XS6) have been measured is also clearly larger than the resolution of the structures used in [31]. The statistical analysis [31] shown in Figure 5 a) proposes that the most probable mutations at site 614D are to S, A, N and G; the amino acid E can be excluded on sterical grounds since it has a larger size than D.

The best matching cluster for the adjacent sites A:(613,614,615) shown in Figure 5 b) is the cluster #26 of [31], with $\Delta_{min} \approx 0.75$Å. The statistical analysis [31] now proposes that the most probable mutation at site 614 is a D→G substitution; the probability for any other amino acid substitution is very low. In particular the probability for D itself is low, suggesting instability.

Finally, the best matching cluster for the following sites F:(614,615,616) shown in Figure 5 c) is #10. The RMSD has now a somewhat lower value $\Delta_{min} \approx 0.41$Å. The statistical analysis [31] proposes that the most probable substitution at site 614 is again D→G. The probability for any other substitution is very low. In particular the probability for D itself is low, again suggesting instability of the residue.

The combined spike protein chain shown in the three Figures 5, starting from site 612 in 5 a) and ending at 616 in 5 c), encircles the inflection point once in clockwise direction. Thus the folding index has value +1, and the topology of the trajectory is similar to that in Figure 3 b) (except for the direction and the location of the end points). In particular, a flattening point and a β-point occur at neighboring segments along the chain. This proposes that the site 614 is a potential mutation hot-spot, it is prone to a change in the local topology by an inflection point bifurcation. The mutation can change the local topology from that resembling Figure 3 b) to one that resembles either Figure 3 a) with a bi-flattening point or Figure 3 b) with a pair of β-points. Since G is the only amino acid that consistently appears in all three clusters and since there is no obvious steric hindrance for a D→G mutation, the prediction of the present analysis is that a D→G mutation is probable at the site 614 of spike protein. This is the notorious D614G mutation that has already been observed.

A comparison of the three PDB structures 6VXX, 6VYB with 6XS6 shows that apparently the mutation has not caused any change in local topology, at least according to available structures. But since the site 614 is located near the junction between S1 and S2 subunits, the D→G mutation in combination with the proximity of a flattening point has probably increased the chain flexibility in the junction segment so that it is now more prone to a

cleavage with enhanced infectiousness as a consequence.

A survey of potential mutation hot-spots in spike protein

The Table I proposes that there is quite a large number of potential mutation hot-spots in the spike protein, and in the sequel a selection of examples is analyzed. The examples are representative, but not necessarily the most probable hot-spot sites. There are three examples from the NTD domain, with site numbers 59, 103 and 287. The example at the site 103 is added since this is the site with the lowest bond angle value along the entire spike protein backbone. There is one example that is located in the junction between NTD and RBD domains, with site number 316. There is one example in the RBD domain, with site number 464. Finally, an example from the fusion core between HR1 and HR2 domains with site number 1080 is presented.

The residue 59F

The Figures 6 show the neighborhood of the site 59F, located in the NTD domain of subunit S1. The Figures reveal that the topology of the trajectory from site 57 to 61 is very similar to that in the case of site 614, shown in Figures 5: There is a residue that is proximal to a flattening point and right after it there is a residue that is proximal to a β-point. The folding index has value +1 since the trajectory encircles the inflection point in clockwise direction. Thus, as in the case of 614, the site 59 is prone to an inflection point bifurcation such as those described in Figures 3. The RMSD values (16) are all quite small, indeed clearly smaller than in the case of 614D, so that the three clusters that are identified in the Figures 6 are a very good match. The statistical analysis of all three clusters show that G has a very high probability at the site 59; both S and D have some propensity albeit much smaller than G while the probability of the existing amino acid F is very small. Thus the site 59 is a very good candidate for a F→G mutation hot-spot.

The case of 103

The Figures 7 show the neighborhood of the site 103G, located in the NTD domain of subunit S1. Here the situation is somewhat exceptional, since the residue 103 is already the smallest amino acid G.

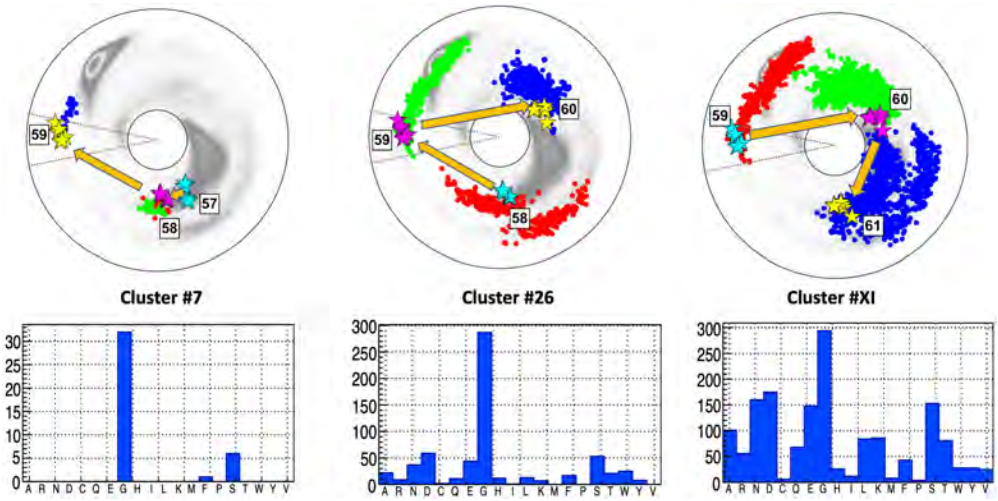

FIG. 6: (Color online) The (κ_i, τ_i) spectrum for $i = 57 - 61$; this engages sites $55 - 62$ along the Cα backbone. a) The spectrum for P:$(57, 58, 59)$, in the background of the best matching cluster #7 in [31]. b) The spectrum for A:$(58, 59, 60)$, in the background of the best matching cluster #26 in [31]. c) The spectrum for F:$(59, 60, 61)$, in the background of the best matching cluster #XI in [31]. The most common amino acid in all three clusters is G.

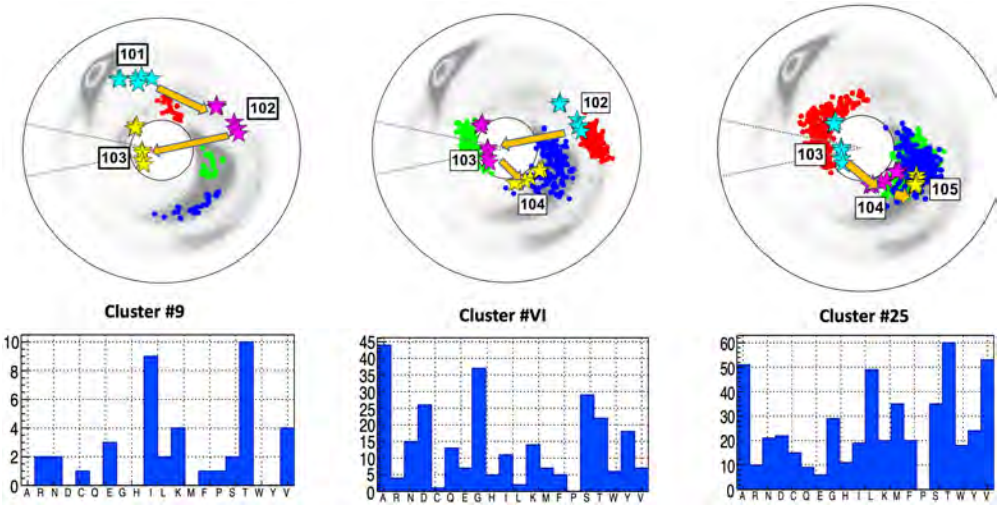

FIG. 7: (Color online) The (κ_i, τ_i) spectrum for $i = 101 - 105$; this engages sites $99 - 106$ along the Cα backbone. Panel a) The spectrum for P:$(101, 102, 103)$, in the background of the best matching cluster #7 in [31]. The most common amino acids in the cluster is T, followed by I. Panel b) The spectrum for A:$(102, 103, 104)$, in the background of the best matching cluster #26 in [31]. The most common amino acid in the cluster is A, followed by G. Panel c) The spectrum for F:$(103, 104, 105)$, in the background of the best matching cluster #XI in [31]. The most common amino acid in the cluster are T and L, followed by A and V.

The chain between sites 101 and 105 is shown in Figures 7. It has one vertex near a β-point at 102. This is immediately followed by the vertex 103 that is proximal both to a

flattening point and to the inflection point. The following vertex 104 has also a very small bond angle value. The overall shape of the trajectory suggests a mutation hot-spot with inflection point perestroika that converts the site 103 from a vertex that is proximal to the flattening point into a vertex that is proximal to a β-point. It is also plausible that there has been a recent mutation with ensuing inflection point perestroika, that has converted the local topology by moving the vertex 103 from the vicinity of a β-point to the vicinity of a flattening point.

The statistical analysis shows that A, which is the smallest amino acids after G, has the highest propensity in the case of the adjacent cluster, shown in Figure 7 b). The propensity of A is also larger than that of G in the following cluster shown in Figure 7 c). Both clusters have also small RMSD value. On the other hand, in the histogram of the preceding cluster shown in Figure 7 a) the amino acid A is absent. However, RMSD value is not very small, and the Figure also shows that cluster #9 can not be good match to the spike protein segment P:(101,102,103): The distance between the observed τ value at the site 103 deviates from the τ-values in the cluster by some 150 degrees. Thus the conclusion is that the cluster #9 should be used with care, for a mutation prediction.

The Figure 8 a) shows the present 3D spike protein structure in the neighborhood of the site 103. In the Figure 8 b) the amino acid G has been replaced by A using Chimera [32]. An inspection of the interatomic distances show that A can be substituted for G without encountering steric clashes. But in the case of the other amino acids T, V and L that also have a high propensity in the cluster of Figure 7 c), a substitution using Chimera leads to steric clashes.

The conclusion os the present analysis is that the site 103G is a potential G→A mutation hot-spot.

The case of 287D

The Figures 9 show the neighborhood of the site 287D, located in the NTD domain of subunit S1. The site is both preceded and followed by a site that is proximal to a β-point. The folding index has value -1 since the chain encircles the inflection point in counterclockwise direction. But since 286 and 288 are both very close to a β-point the chain passes back-and-forth very close to the inflection point.: The links 286-287 and 287-288 have

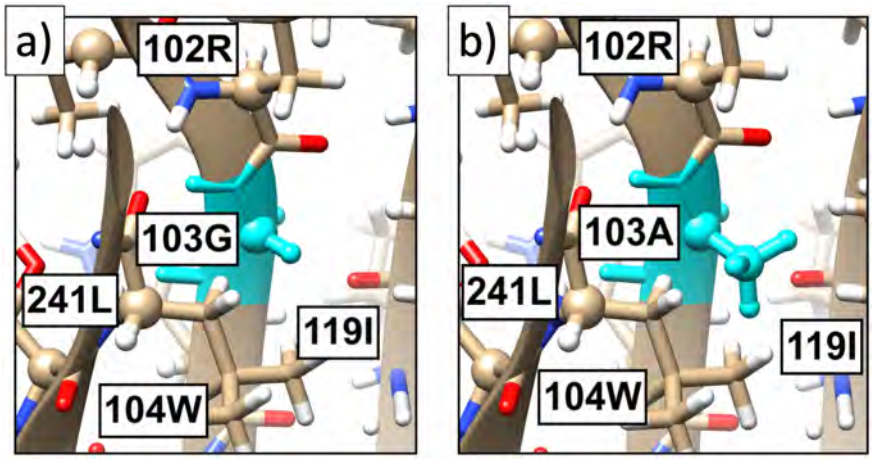

FIG. 8: (Color online) a) The present all-atom structure of the spike protein in the neighborhood of site 103. b) The substitution of A in place of 103G, as predicted by Chimera [32]. There are no apparent steric hinders for the substitution.

a very similar topology to the links 102-103 and 103-104 in Figures 7.

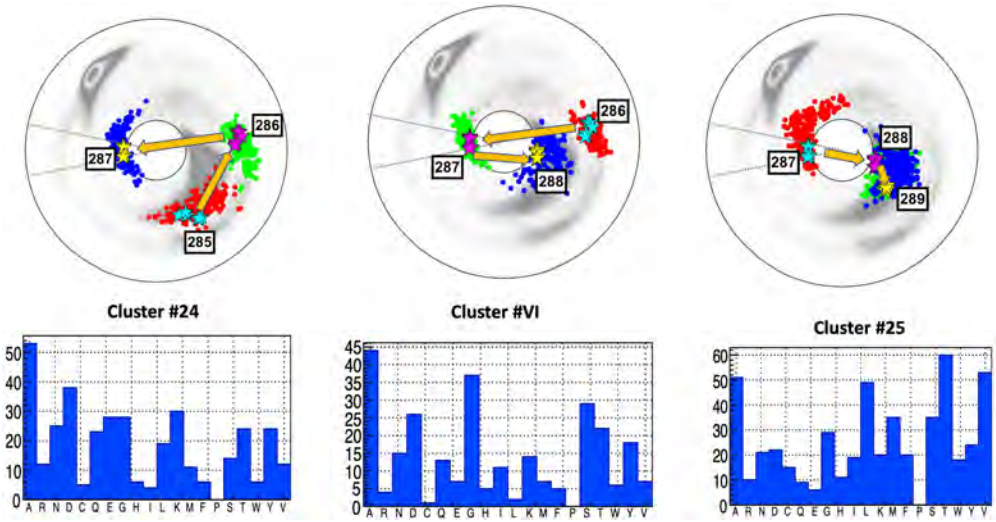

FIG. 9: (Color online) The (κ_i, τ_i) spectrum for $i = 285 - 289$; this engages sites $283 - 290$ along the Cα backbone. Panel a) The spectrum for P:$(285, 286, 287)$, in the background of the best matching cluster #24 in [31]. The most common amino acids in the cluster is A, followed by D. Panel b) The spectrum for A:$(286, 287, 288)$, in the background of the best matching cluster #VI in [31]. The most common amino acid in the cluster is A, followed by G. Panel c) The spectrum for F:$(287, 288, 289)$, in the background of the best matching cluster #25 in [31]. The most common amino acid in the cluster are T and V, followed by A and L.

The RMSD values (16) are all very small so that the three clusters identified in the Figures 9 are an excellent match. The statistical analysis in all three clusters show that A has a very high probability at the site 287; both T and V are also likely substitutions in

the following cluster, and G has also propensity in the adjacent cluster. But D that is now located at the site, is not very prominent in any of the clusters. Thus the site 287 is a very good candidate for a D→A mutation hot-spot.

The case of 316S

The site 316S is located in the junction between the NTD and the RBD domains. A mutation hot-spot in this junction is of interest as it can have a large effect to the transition between the closed and open states, the way how the RBD becomes exposed to ACE2. The site 316S is the only potential hot-spot mutation site in this junction that has been identified in the present study.

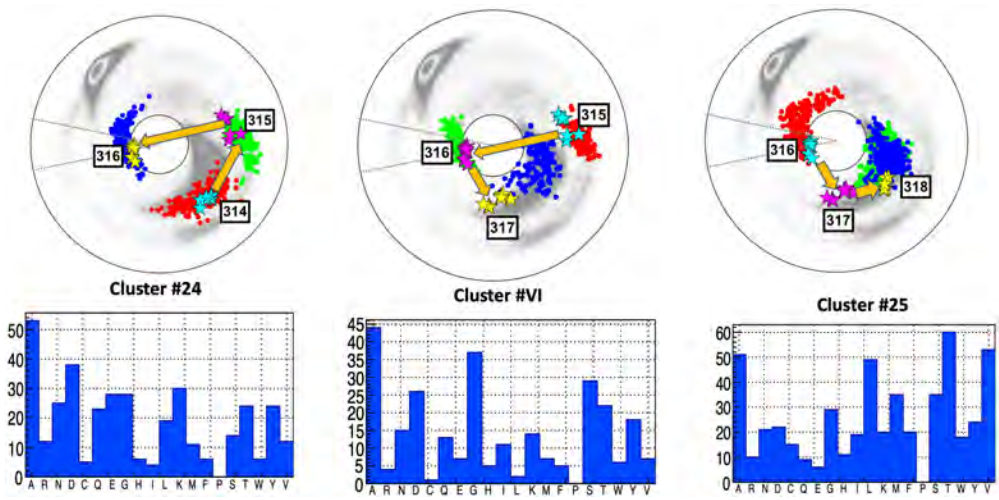

FIG. 10: (Color online) The (κ_i, τ_i) spectrum for $i = 314 - 318$; this engages sites $312 - 319$ along the Cα backbone. Panel a) The spectrum for P:$(314, 315, 316)$, in the background of the best matching cluster #24 in [31]. The most common amino acids in the cluster is A, followed by D. Panel b) The spectrum for A:$(315, 316, 317)$, in the background of the best matching cluster #VI in [31]. The most common amino acid in the cluster is A, followed by G. Panel c) The spectrum for F:$(316, 317, 318)$, in the background of the best matching cluster #25 in [31]. The most common amino acid in the cluster are T and V, followed by A and L.

The RMSD values (16) are quite small so that the three clusters identified in the Figures 10 are a good match. The chain between sites 315-317 has a very similar topology to the chains 102-104 in Figures 7 and 286-288 in Figures 8. The clusters are also the same as in the case of 287D and therefore it is concluded that the site 316 is a good candidate for a S→A mutation hot-spot.

The case of 464F

This site is located in the RBD domain of the subunit S1, and a mutation hot-spot can affect the binding to ACE2. Unfortunately, there are missing residues in the PDB data, only data for chain A in open state (PDB structure 6VYB) is available. The chain shown in Figures 11 is very similar to that in the case of 59F shown in Figures 6. The RMSD values are very low so that the three clusters are an excellent match. The statistical analysis shows that the site is a very good hot-spot candidate for a F→G mutation.

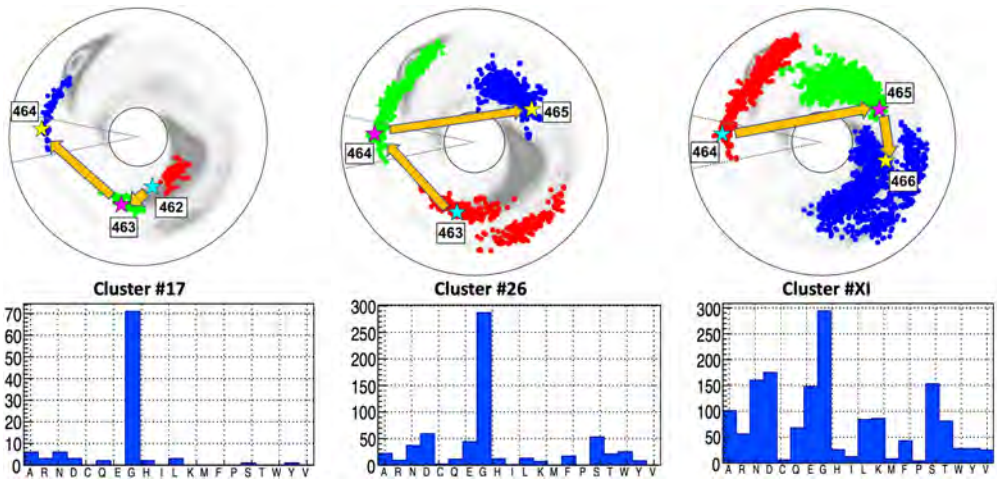

FIG. 11: (Color online) The (κ_i, τ_i) spectrum for $i = 462 - 466$; this engages sites $460 - 467$ along the Cα backbone. Panel a) The spectrum for P:$(462, 463, 464)$, in the background of the best matching cluster #17 in [31]. Panel b) The spectrum for A:$(463, 464, 465)$, in the background of the best matching cluster #26 in [31]. Panel c) The spectrum for F:$(464, 465, 466)$, in the background of the best matching cluster #XI in [31]. The most common amino acid in all three clusters is G.

The case of 1080A

According to the histogram in Figure 4 the fusion core between the HR1 and HR2 domains has a large number of sites that are proximal to a flattening point. However, the detailed investigation does not reveal any excellent candidate for a mutation hot-spot. An example is the site 1080A, the analysis results are summarized in Figures 12.

The RMSD value for preceding cluster P:(1078,1079,1080) is not very good, but the value is excellent both for the adjacent A:(1079,1080,1081) and following F:(1080,1081,1082) clusters. Moreover, the chain shows that the site 1080 is an excellent candidate for a bi-flattening perestroika. But the statistical analysis reveals that the amino acid A, presently

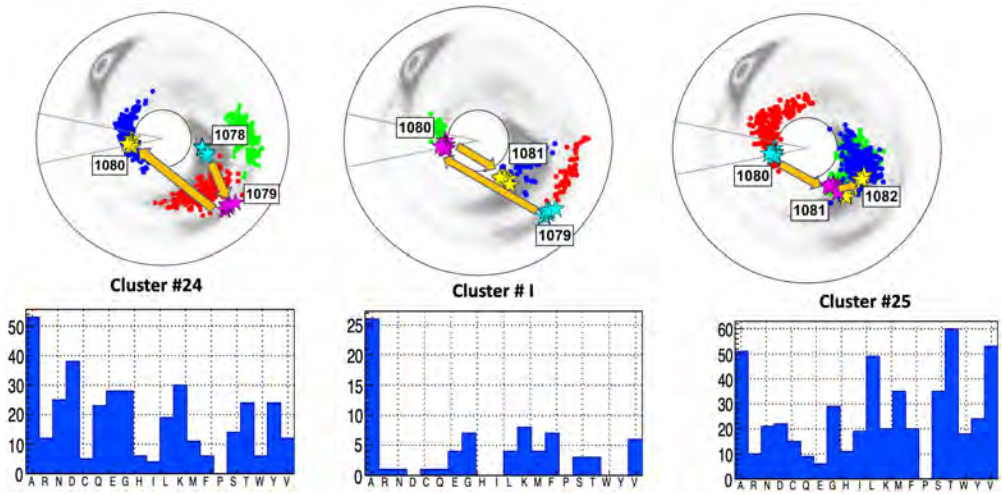

FIG. 12: (Color online) The (κ_i, τ_i) spectrum for $i = 1078 - 1082$; this engages sites $1076 - 1083$ along the Cα backbone. Panel a) The spectrum for P:$(1078, 1079.1080)$, in the background of the best matching cluster #24 in [31]. The most common amino acids in the cluster is A. Panel b) The spectrum for A:$(1079, 1080, 1081)$, in the background of the best matching cluster #I in [31]. The most common amino acid in the cluster is A, followed by G. Panel c) The spectrum for F:$(1080, 1081, 1082)$, in the background of the best matching cluster #25 in [31]. The most common amino acid in the cluster are T and V, followed by A and L.

at this site, is also the most probable one and with very high probability in the case of the adjacent cluster. Thus, it appears that a recent mutation may have taken place at this site with the amino acid A as the substitution. The statistical analysis of the third cluster shows a relatively high propensity for T, V and L but all three are subject to steric hindrances and they also have a very low propensity in the adjacent cluster. The conclusion is that even though the vertex is proximal to a bi-flattening perestroika, it does not qualify as a mutation hot-spot. It is more likely that due to proximity of the bi-flattening perestroika, the site has exceptional flexibility with an important role in the fusion process. The residue 1080A can be a good target for the development of fusion inhibitors.

DISCUSSION

The spike protein is a prominent target for the development of vaccines and therapeutic drugs against the SARS-CoV-2 virus. But for a durable antiviral one needs to know, how to identify those amino acids that are prone for a mutation and how to predict the biological consequences of any potential mutation. The methodology that is presented here aims to pinpoint those hot-spot residues where a mutation can be expected to have sub-

stantial conformational consequences. Since conformation is pivotal for a protein's function, the knowledge of hot-spots is important for better understanding of how the spike protein operates and how it can be incapacitated. For this the local topology of spike protein's Cα backbone is scrutinized; the local backbone topology can only change when a structural bifurcation occurs at a critical site. In the case of a protein backbone a pre-requisite for such a bifurcation is the presence of a flattening point. Therefore, the flattening points along the spike protein backbone are first localized and classified. The geometry in the neighborhood of a flattening point is then investigated to deduce its potential for a mutation hot-spot. The detailed methodology is developed using the notorious D614G mutation site of the spike protein as an example; the site is proximal to a flattening point and the present approach correctly predicts its D$\rightarrow$G mutation. It is found that several topologically very similar potential hot-spot sites are located in particular in the N-terminal domain and in the fusion core between the two heptapeptide repeat sequences. A number of these are analyzed in detail, as examples. Those in the N-terminal domain are found to be good candidates for mutation hot-spots, while the potential hot-spot sites in the junction between the heptapeptide repeat sequences appear to be more stable. In particular, the residue 1080A that has been analyzed as an example seems to be stable against mutation. But it seems to be prone to a change in local topology, due to its proximity to a bi-flattening point. This can make the residue 1080A into a good target for the development of structure based fusion inhibitors.

CONCLUSIONS

The topological techniques that are employed here have already proven themselves to be most powerful in several Physics problems. Since conformation is pivotal for a protein's function, topology should be a most effective tool also in protein research. The present methodology combines the investigation of local Cα backbone topology with a statistical analysis of PDB structures. This brings a downside, as the present structural data on the spike protein has been measured with quite a low resolution; the PDB structures that are used in the present study have been determined using electron microscopy with a resolution no better than around 2.8 - 3.5 Å and there are even some missing residues in available structures. The low resolution causes also some uncertainty in the correct identification of sites that are proximal to a flattening point, for a conclusive investigation better resolution

spike protein structures are needed. At the same time the statistical analysis that is used here for the identification of the pertinent PDB cluster is also preliminary, as it utilizes a somewhat limited set of structural clusters that are measured with a very high resolution. Therefore, besides a need for improvement in the structural data on spike protein, a further development and refinement of the statistical methodology is also needed to ensure a higher reliability of the present methodology.

AUTHOR CONTRIBUTIONS

AJN proposed the research; XP performed the calculations; AJN and XP analyzed the data and wrote the manuscript.

ACKNOWLEDGEMENTS

XP is supported by Beijing Institute of Technology Research Fund Program for Young Scholars. AJN is supported by the Carl Trygger Foundation Grant CTS 18:276, by the Swedish Research Council under Contract No. 2018-04411, and by COST Action CA17139. AJN is also partially supported by Grant No. 0657-2020-0015 of the Ministry of Science and Higher Education of Russia.

---

[1] Zhu, N., D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, et al., 2020. A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine* .

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

[3] Andersen, K. G., A. Rambaut, W. I. Lipkin, E. C. Holmes, and R. F. Garry, 2020. The proximal origin of SARS-CoV-2. *Nature medicine* 26:450–452.

[4] Benvenuto, D., M. Giovanetti, A. Ciccozzi, S. Spoto, S. Angeletti, and M. Ciccozzi, 2020. The 2019-new coronavirus epidemic: evidence for virus evolution. *Journal of medical virology* 92:455–459.

[5] Paraskevis, D., E. G. Kostaki, G. Magiorkinis, G. Panayiotakopoulos, G. Sourvinos, and S. Tsiodras, 2020. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. *Infection, Genetics and Evolution* 79:104212.

[6] Riou, J., and C. L. Althaus, 2020. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Eurosurveillance* 25:2000058.

[7] Zhao, S., Q. Lin, J. Ran, S. S. Musa, G. Yang, W. Wang, Y. Lou, D. Gao, L. Yang, D. He, et al., 2020. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International journal of infectious diseases* 92:214–217.

[8] Chan, J. F.-W., S. Yuan, K.-H. Kok, K. K.-W. To, H. Chu, J. Yang, F. Xing, J. Liu, C. C.-Y. Yip, R. W.-S. Poon, et al., 2020. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet* 395:514–523.

[9] Phan, L. T., T. V. Nguyen, Q. C. Luong, T. V. Nguyen, H. T. Nguyen, H. Q. Le, T. T. Nguyen, T. M. Cao, and Q. D. Pham, 2020. Importation and human-to-human transmission of a novel coronavirus in Vietnam. *New England Journal of Medicine* 382:872–874.

[10] Kissler, S. M., C. Tedijanto, E. Goldstein, Y. H. Grad, and M. Lipsitch, 2020. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* 368:860–868.

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

[12] Huang, Y., C. Yang, X.-f. Xu, W. Xu, and S.-w. Liu, 2020. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacologica Sinica* 1–9.

[13] Walls, A. C., Y.-J. Park, M. A. Tortorici, A. Wall, A. T. McGuire, and D. Veesler, 2020. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* .

[14] Dai, W., B. Zhang, X.-M. Jiang, H. Su, J. Li, Y. Zhao, X. Xie, Z. Jin, J. Peng, F. Liu, et al., 2020. Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease. *Science* 368:1331–1335.

[15] Gordon, D. E., J. Hiatt, M. Bouhaddou, V. V. Rezelj, S. Ulferts, H. Braberg, A. S. Jureka, K. Obernier, J. Z. Guo, J. Batra, et al., 2020. Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science* .

[16] Li, J.-Y., C.-H. Liao, Q. Wang, Y.-J. Tan, R. Luo, Y. Qiu, and X.-Y. Ge, 2020. The ORF6, ORF8 and nucleocapsid proteins of SARS-CoV-2 inhibit type I interferon signaling pathway. *Virus research* 286:198074.

[17] Gordon, D. E., G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, K. M. White, M. J. OMeara, V. V. Rezelj, J. Z. Guo, D. L. Swaney, et al., 2020. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 1–13.

[18] Hoffmann, M., H. Kleine-Weber, and S. Pöhlmann, 2020. A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Molecular Cell* .

[19] Hoffmann, M., H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche, et al., 2020. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* .

[20] Xia, S., M. Liu, C. Wang, W. Xu, Q. Lan, S. Feng, F. Qi, L. Bao, L. Du, S. Liu, et al., 2020. Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. *Cell research* 30:343–355.

[21] Korber, B., W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhattacharya, B. Foley, et al., 2020. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* 182:812–827.

[22] Arnold, V., 1996. On the Number of Flattening Point on Space Curves. *AMS Trans., Ser* 171:11.

[23] Aicardi, F., 2000. Self-linking of spatial curves without inflections and its applications. *Functional Analysis and Its Applications* 34:79–85.

[24] Uribe-Vargas, R., 2004. On singularities," perestroikas" and differential geometry of space curves. *ENSEIGNEMENT MATHEMATIQUE* 50:69–102.

[25] Yurkovetskiy, L., X. Wang, K. E. Pascal, C. Tomkins-Tinch, T. P. Nyalile, Y. Wang, A. Baum, W. E. Diehl, A. Dauphin, C. Carbone, et al., 2020. Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell* .

[26] Spivak, M. D., 1970. A comprehensive introduction to differential geometry, volume 5. Publish or perish.

[27] Hu, S., M. Lundgren, and A. J. Niemi, 2011. Discrete Frenet frame, inflection point solitons, and curve visualization with applications to folded proteins. *Physical Review E* 83:061908.

[28] Lundgren, M., A. J. Niemi, and F. Sha, 2012. Protein loops, solitons, and side-chain visualization with applications to the left-handed helix region. *Physical Review E* 85:061909.

[29] Lundgren, M., A. Krokhotin, and A. J. Niemi, 2013. Topology and structural self-organization in folded proteins. *Physical Review E* 88:042709.

[30] Hinsen, K., S. Hu, G. R. Kneller, and A. J. Niemi, 2013. A comparison of reduced coordinate sets for describing protein structure. *The Journal of Chemical Physics* 139:09B649_1.

[31] Peng, X., J. He, and A. J. Niemi, 2015. Clustering and percolation in protein loop structures. *BMC structural biology* 15:22.

[32] Shapovalov, M. V., and R. L. Dunbrack Jr, 2011. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* 19:844–858.