bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.006197. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

In-Silico evidence for two receptors based strategy of SARS-CoV-2

Edoardo Milanetti*†,[1,2] Mattia Miotto*,[1,2] Lorenzo Di Rienzo*,[2]
Michele Monti,[3,4] Giorgio Gosti,[2] and Giancarlo Ruocco[1,2]

[1] *Department of Physics, Sapienza University, Piazzale Aldo Moro 5, 00185, Rome, Italy*
[2] *Center for Life Nanoscience, Istituto Italiano di Tecnologia, Viale Regina Elena 291, 00161, Rome, Italy*
[3] *Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Dr. Aiguader 88, 08003 Barcelona*
[4] *RNA System Biology Lab, department of Neuroscience and Brain Technologies, Istituto Italiano di Tecnologia, Via Morego 30, 16163, Genoa, Italy*

We propose a novel numerical method able to determine efficiently and effectively the relationship of complementarity between portions of proteins surfaces. This innovative and general procedure, based on the representation of the molecular iso-electron density surface in terms of 2D Zernike polynomials, allows the rapid and quantitative assessment of the geometrical shape complementarity between interacting proteins, that was unfeasible with previous methods. We first tested the method with a large dataset of known protein complexes obtaining an overall area under the ROC curve of 0.69 in the blind recognition of binding sites and then applied it to investigate the features of the interaction between the Spike protein of SARS-Cov-2 and human cellular receptors. Our results indicate that SARS-CoV-2 uses a dual strategy: its spike protein could also interact with sialic acid receptors of the cells in the upper airways, in addition to the known interaction with Angiotensin-converting enzyme 2.

I. INTRODUCTION

At the time of writing, the new coronavirus, also known as SARS-CoV-2, which causes the severe acute respiratory syndrome [1, 2], has caused approximately the death of approximately 17000 and the infection of 390000 people. The COVID-19 outbreak represents a serious threat to public health [3], and the World Health Organization officially declared it a pandemic.

To date 7 coronavirus strains are known to infect humans and no approved therapies or vaccine against them are available [3].

In the past two decades, in addition to SARS-CoV-2, two other β-coronavirus have caused three of the most severe epidemics worldwide: SARS-CoV [4, 5] and MERS-CoV [6] which respectively cause the Severe Acute Respiratory Syndrome (SARS), and the Middle East Respiratory Syndrome (MERS).

The characteristics of the interactions between these viruses and the human cell receptors are being carefully studied in order to shed light on both diffusion speed and mortality rate differences between SARS-CoV-2 and the others, with a special regard to SARS-CoV.

Indeed, SARS-CoV spread across 26 countries in six continents and caused a total of 8,096 cases and 774 deaths (9.6%) [7], with an incubation period of 1 to 4 days [8]. On the other side, it has been demonstrated that the latency of SARS-CoV-2 varies from 3-7 days on average, up to 14 days [2]. Thus, the average latency of SARS-CoV-2 is slightly longer than that of SARS-CoV [7]. Moreover, it is estimated from epidemiological data that individuals infected with SARS-CoV-2 are contagious from the beginning of the incubation period, and that between the incubation period and the end of the infection each infected individual transmits the infection to about 3.77 other people [9].

SARS-CoV-2, similarly to SARS-CoV and MERS-CoV, attacks the lower respiratory system, thus causing viral pneumonia. However, this infection can also affect the gastrointestinal system, heart, kidney, liver, and central nervous system [2, 10, 11].

To face the emergency of this epidemic it is essential to shed light on the interaction mechanisms between the virus and the human cell receptors.

It is well characterised how SARS-CoV infection is mediated by the high-affinity interactions between the receptor-binding domain (RBD) of the spike (S) glycoprotein and the human-host Angiotensin converting enzyme 2 (ACE2) receptor [12–14]. The spike protein is located on the virus envelope and promotes the attachment to the host cell and the fusion between the viral and cellular membrane. [15, 16].

Recently, it has also been proven that several critical residues in SARS-CoV-2's RBD provide favorable interactions with human ACE2, consistent with SARS-CoV-2's capability to infect the cell [17, 18]. On the experimental side, it has also been confirmed by in vivo experiments that SARS-CoV-2's entry is mediated by lung cell Ace2 receptors [19]. More importantly the structure of the spike-Ace2 receptor complex has been recentely determined by Crio-em [20]. In conclusion, it is now understood that SARS-CoV-2 binds to the ACE2 receptor to infect the host cell using its spike protein's RBD, even if it had most likely evolved from SARS-CoV independently [21].

Since the ACE2 molecule is known to be a human entry receptor, the understanding of the molecular mechanisms of interaction between the ACE2 receptor and the spike protein of the virus can be a key factor designing new

*The authors contributed equally to the present work.
†Corresponding author: edoardo.milanetti@uniroma1.it

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.006197. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

drug compounds. With this aim, computational methods based on both sequence and structure studies of proteins represent a powerful tool [22].

Indeed, the development of effective computational methods for predicting the binding sites of proteins can improve the understanding of many molecular mechanisms [23–25]. Several methods to analyze protein interaction have used protein surface information [26–29].

A very promising strategy to study molecular interaction is to determine, using deep learning methods, chemico-physical features of the molecular surface [25]. This method allows to efficiently detect binding sites but it unfortunately has the drawbacks of any other deep learning approach: it requires the definition and training of several parameters and the creation of a sufficiently large training and test database. These method also requires the analysis of all possible orientations and relative position of the binding sites.

In this work we present a new and general method to efficiently describe the shape of molecular surfaces and apply it to study the interactions between spike proteins and the corresponding receptors involved in the SARS-CoV, MERS-CoV and SARS-CoV-2 infection mechanisms.

The method proposed here, for the first time, describes regions of molecular surface with the 2D Zernike formalism. Indeed, each local patch of a 3D protein molecule can be represented as a surface on a 2D square grid, while retaining both distance and angular information with respect to a specific reference system. Applying the Zernike expansion we compute - completely unsupervised - numerical descriptors that summarize the patch geometry and we use them to compute the structural similarity between any two patches. The Zernike descriptors are rotationally invariant, ensuring to the method an high efficiency and a low computational cost since it avoids the necessity of any preliminary structure orientation. Overall, these operation substantially reduces the search space dimensionality, thus allowing our method to be able to explore several more protein regions compered to previous methods.

We firstly apply the method to a large dataset of protein-protein complexes (Protein-Protein Dataset), where we test the ability of our method to detect interacting regions from non-interacting regions with the use of our definition of complementarity. Our method recognize interactions in the large Protein-Protein Database with a good precision. Furthermore, thanks to the low computational cost of building an invariant description, we can blindly sample the entire surfaces of 2 interacting proteins, and retrieve their binding sites.

Then, we use our formalism to study the interaction between the spike protein and its membrane receptors, comparing SARS-Cov-2 with both SARS-COV and MERS. We demonstrate that the actual region of binding between spike protein and ACE2 human - both in SARS-CoV an SARS-CoV-2 - have an higher complementarity with respect to other randomly sampled exposed receptor region.

Furthermore, we also analize in detail the structural properties of the MERS spike protein which, like several other proteins belonging to coronaviruses family, is able to interact with sialic acids [30]. Among other coronaviruses, the bovine coronavirus (BCoV), and the two human coronaviruses OC43 and HKU1 employ sialoglycan-based receptors with 9-O-acetylated sialic acid (9-O-Ac-Sia) as key component [31]. These viruses bind to cell surface components containing N-acetyl-9-O-acetylneuraminic acid and this interaction is essential for initiation of an infection [32]. In particular, we propose here a possible alternative mechanism of SARS-CoV-2 cellular infection, through spike protein interaction with sialic acid receptors of the upper airways, similarly to what has been shown for the MERS spike protein [33]. We indeed highlights that a surface region in the N-terminal domain of SARS-CoV-2 spike is very similar to the MERS spike sialic acid binding region, and present a compatible charge. These observations suggest that these two pockets potentially share a analogous function. This second infection mechanism for SARS-CoV-2 could explain its high diffusion speed.

II. RESULTS

In the last decade the 3D Zernike formalism has been widely applied for the characterization of molecular interactions [29, 34–36]: in this work we adopted a new representation, based on the 2D Zernike polynomials, which allows the quantitative characterization of protein surface regions. As shown in Fig.1, our computational protocol associates to each molecular patch an ordered set of numbers (the expansion coefficients) that describes its shapes.

Through this compact description it is possible to both analyze the similarity between 2 different regions - suggesting for example a similar ligand for 2 binding regions - and to study the complementarity between 2 interacting surfaces.

To validate this method, we firstly collected a large structural dataset of protein-protein complexes and we characterize their binding sites through 2D Zernike. We test the method ability to recognize the higher complementarity observed between pairs of interacting surfaces compared to the lower complementarity found when the 2 surfaces are non-interacting.

We thus analyze in detail the interactions of the SARS-CoV-2 spike protein with its membrane receptors, comparing SARS-CoV-2 properties with SARS-CoV and MERS-CoV.

Binding regions unsupervised recognition

We selected a structural dataset, composed of about 4500 experimentally determined protein-protein com-

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.006197. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

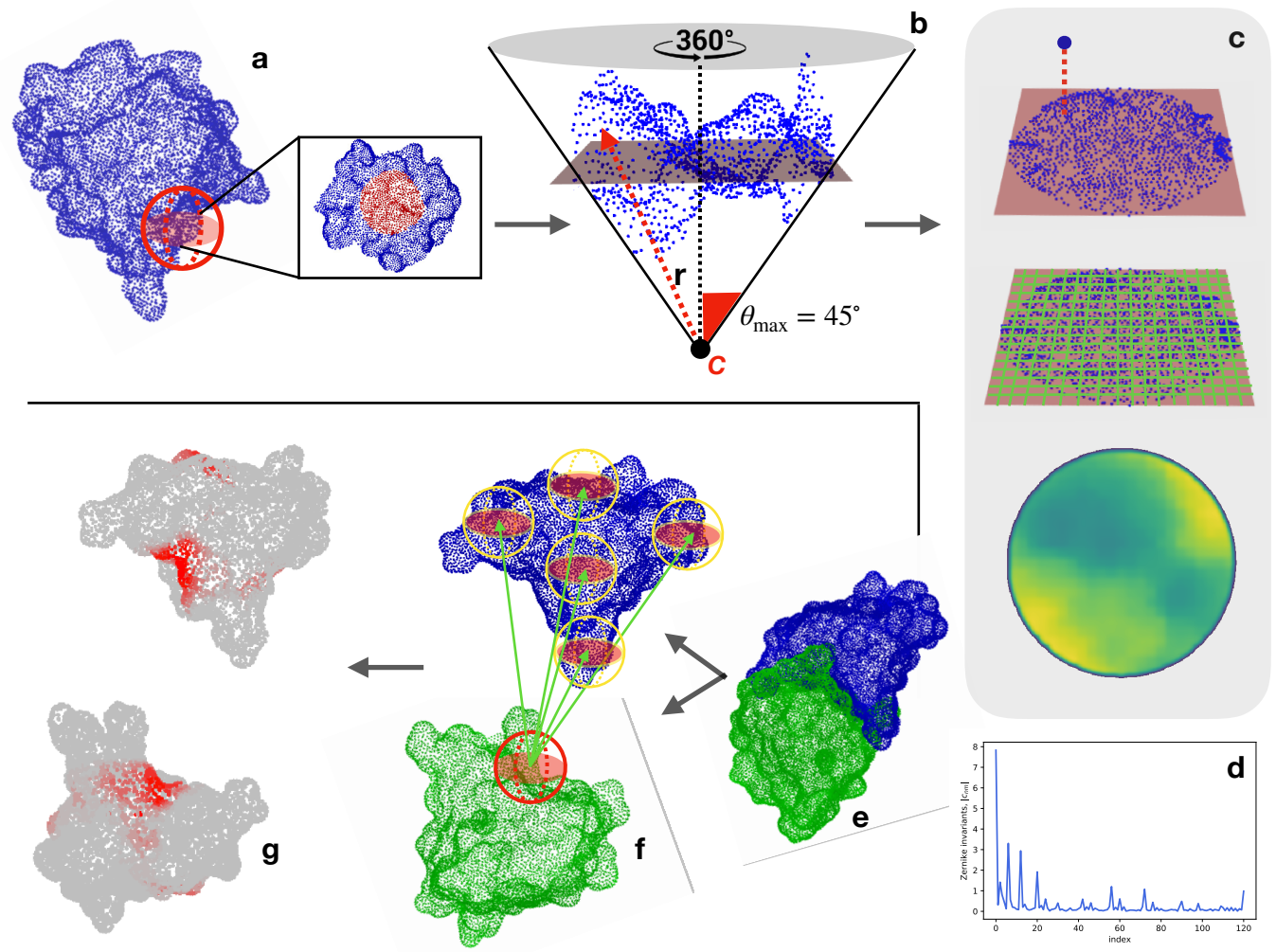

FIG. 1: **Computational protocol for the characterization of each surface region and for the blind identification of the binding sites. a)** Molecular solvent-accessible surface of a protein (in blue) and example of patch selection (red sphere). **b)** The selected patch points are fitted with a plane and reoriented in such a way that the z-axis passes through the centroid of the points and is orthogonal to the plane. A point C along the z-axis is defined, such as that the largest angle between the perpendicular axis and the secant connecting C to a surface point is equal to 45°. Finally, to each point, its distance, r with the point C is evaluated. **c)** Each point of the surface is projected on the fit plane, which is binned with a square grid. To each pixel, the average of the r values of the points inside the pixel is associated. **d)** The resulting 2D projection of the patch can be represented by a set of 2D Zernike invariant descriptors. **e-f)** Given a protein-protein complex (PDB code: 3B0F, in this example), for each surface vertex we select a patch centered on it and compute its Zernike descriptors. To blindly identify the two binding sites, each sampled patch is compared with all the patches of the molecular partner and each vertex is associated with the minimum distance between its patch and all the patches of the molecular partner is associated with. **g)** A Smoothing process of the surface point values is applied in order to highlight the signal in the regions characterized mostly by low distance values (high shape complementarity).

plexes, from a recent paper that presented a state of the art patch recognition computational method [25].

For each complex, we have selected the interacting regions and we have characterized them with the 2D Zernike invariant descriptors. Therefore, each binding site is associated with a one dimensional vector, allowing us to easily compare the shape of protein regions with the euclidean distance between their Zernike descriptors. Two regions are complementary when they are characterized by a low distance between their corresponding Zernike vectors[34].

To test the ability of the method to recognize two interacting regions, we have compared how much the distances between the Zernike descriptors (see Methods) of a pair of interacting binding sites is smaller than the distances between 2 random patches. In particular, for each protein-protein complex we define the real distance as the distance between the interacting surfaces, while the random set is formed by the comparison between the binding site of one protein and the binding sites extracted from

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.006197. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

the other complexes of the dataset.

Our *unsupervised* method has a good ability to recognize the true pair of binding regions with respect to random patches with an AUC of the ROC of 0.69 comparing patches of $R_s = 9\AA$. Note that the state of art *supervised* method described in [25] achieves an AUC of 0.68, when are considered only shape descriptors with similar patches of $9 - 12\AA$ of geodedic lenght.

Furthermore, the much lower computational time needed for the calculation of the 2D Zernike descriptors allows an extensive sampling of the surfaces of a pair of proteins in complex. Centering on each surface point a molecular patch, we generate for each protein a very high number of Zernike descriptors. Comparing all the patches of the two proteins, we label each surface point with the *binding propensity*, that is the maximum complementarity recorded between the Zernike descriptors of the patch and all the others belonging to the molecular partner surface. The real binding region is expected to be demarcated and mostly composed of elements with high complementarity. To make the binding region high complementarity more evident, we smooth the signal by attributing to each vertex of the surface the average value of the vertices closer than 6 $\AA$ to it (see Method).

As an example in Fig.1 we report the result of this method for a specific case (PDB code: 3B0F), where this procedure clearly identify the binding regions of the two proteins. For a subset of 20 protein-protein complexes - the smallest in terms of number of surface vertices -, we perform a blind search of the interacting regions. The results are very promising, since the average value of the AUC of the corresponding ROC is 0.65, with fourteen out of forty proteins having an AUC higher than 0.70.

Comparison between the complementarity of the SARS-CoV and SARS-CoV-2 spike protein with the human ACE2 receptor

The excellent and very promising results found in the large dataset encouraged us to investigate the pressing problem of the coronavirus interaction with the host cell. We first analyzed the shape complementarity between the spike proteins of SARS-CoV and SARS-CoV-2 in complex with human ACE2 receptor [12, 20]. It is interesting to notice that the contact between spike and ACE2 receptor both for SARS-CoV and SARS-CoV-2 occurs in two separate interacting regions (see Fig. 2), meaning that we need to investigate the two interacting regions separately. When comparing the two raw distances, we found that ACE2-SARS-CoV distance is smaller than the ACE2-SARS-CoV-2 one, even if for both complexes the complementarity is much higher than the one would find in other random regions of the complexes (see Figure 2). Note that for an appropriate comparison, we need to define a suitable ensemble of random patches. Indeed, the random regions are sampled from the molecular surface of the spike protein imposing that the center of the two patches have the same distance as the binding region observed in the experimental complex. Then, both the real spike binding region and the ensemble of 1000 sampled regions are compared with the receptor binding sites.

The results of this analysis are shown in Fig. 2. We show the distances distribution of the random regions and we report the distance between the real binding regions, both for the Ace2-SARS-CoV and Ace2-SARS-CoV-2 complex. As the method works in recognizing interacting patches, real binding regions show an higher complementarity (lower distance) than the randomly sampled regions.

Furthermore, this analysis shows that ACE2 receptor has a slightly higher shape complementarity with SARS-CoV than with SARS-CoV-2 spike protein. However, the results are quite comparable among them, in accordance with experimental data [3].

Identification of another possible binding region of the SARS-Cov-2 spike

Although it is currently known that the spike protein of SARS-CoV-2 binds to the ACE2 receptor of host cells [18, 37], the investigation of possible other infection mechanism is a key factor in the study of this disease. Specifically, in this work [19] the authors underline the necessity to elucidate whether SARS-Cov-2 spike protein could have acquired the ability to bind with sialic acid as MERS-CoV does. In fact, it has been recently shown that besides the usual receptor (dipeptidyl-peptidase 4 receptor), MERS-CoV spike protein interacts with sialic acid molecules [33] using a well identified pocket in the N-terminal region of the protein. This makes the virus able to interact with high airways and then reach the low airway cells [38].

The recognition between spike and sialic acids occurs via a conserved groove which plays a key role for S MERS-CoV-mediated attachment to sialosides and entry into human airway epithelial cells [33].

Since the interaction of MERS-CoV spike and the sialic acids is caused mainly by hydrogen bonds and shape complementarity [39], our method is particularly suitable to find on the SARS-CoV-2 spike surface a region similar to the one involved in binding of the sialic acid in MERS-CoV spike.

Using the experimental structure of MERS-CoV spike and sialic acid molecules complexes [33], we extract its binding region and we describe it through Zernike descriptors. Then we sample the corresponding domains of both SARS-CoV and SARS-CoV-2 spike, building a molecular patch on each surface point and characterizing it with its corresponding Zernike descriptors. Each region sampled from the spike proteins of these 2 viruses are then compared with the MERS-CoV spike binding region, looking for a similar regions that can mediate an interaction with a similar ligand.

In Fig. 3, we show the results of this analysis. In partic-

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.006197. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

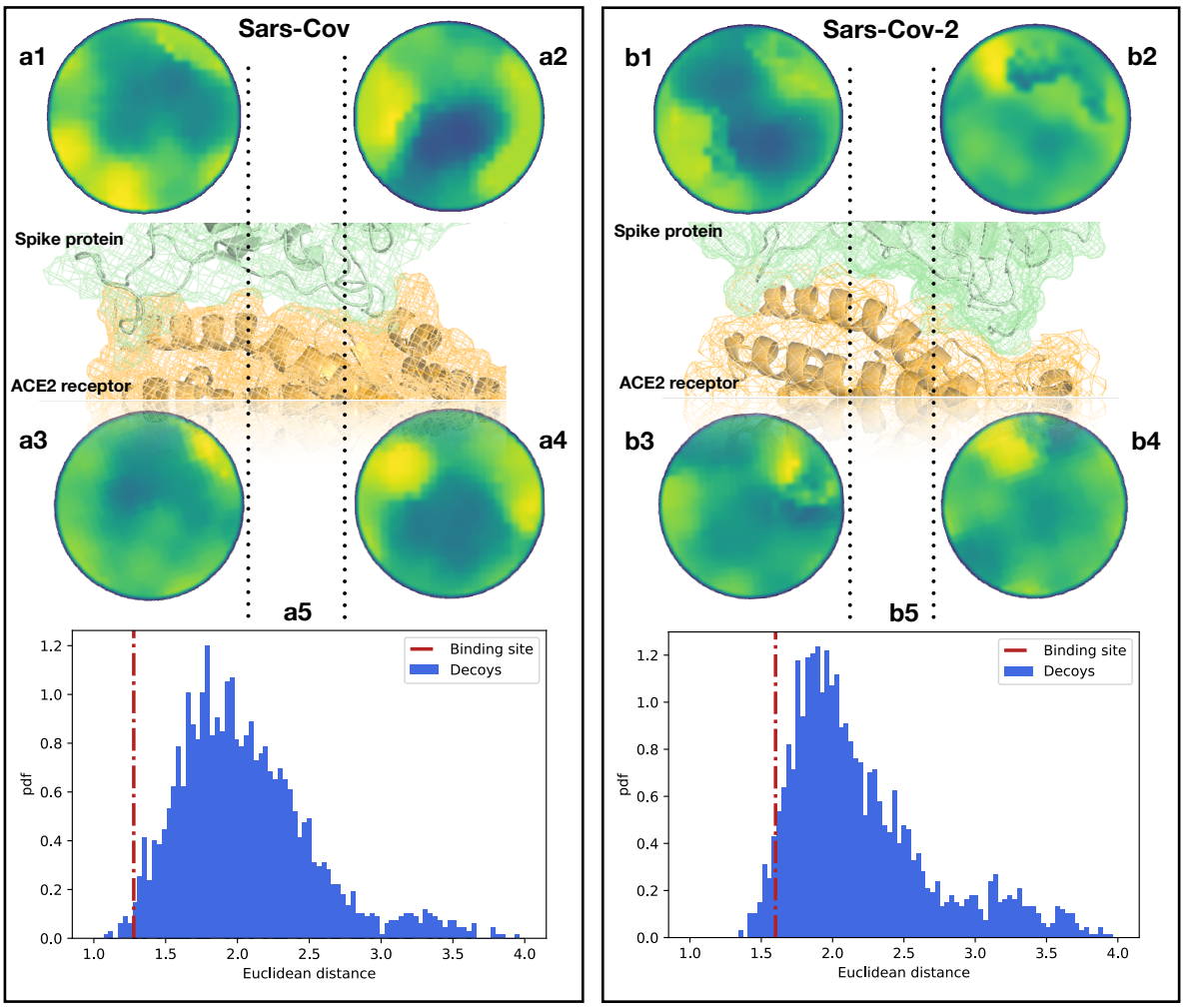

FIG. 2: **Comparison between the binding regions of SARS-CoV and SARS-CoV-2 spike protein with human ACE2. a1,2)** Patch projections in the unitary circle for the two binding regions of Sars-Cov spike protein. **a3,4)** Patch projections in the unitary circle for the binding regions of the human ACE2 receptor. **a5)** Distance distribution between ACE2 binding sites and randomly selected patches on the S spyke protein of SARS-Cov. Decoys patches are sampled taking two random regions separated by the same distance measured in the experimental structure. Red dotted line represents the distance between the real ACE2 and S spike's patches. **b)** The same as a) but for the binding site of SARS-Cov-2 and the human ACE2 receptor. The real distances are in the 1st and 5th percentiles of the distributions for SARS-CoV and SARS-CoV-2 respectively.

ular, selecting the region most similar to the MERS-CoV binding site we identified one region for both SARS-CoV and SARS-CoV-2 spike protein.

Interestingly the best region of SARS-Cov-2 spike exhibits an higher similarity than the pocket selected by the SARS-CoV spike. We moreover calculate the electrostatic potential of the involved surfaces with eF-surf webserver [40]. As shown in Fig. 3, in cartoon representation, the region found in the molecular surface of SARS-CoV-2 spike is very similar to the MERS spike region interacting with sialic acid, both in terms of shape and electrostatic similarity. Differently, the region identified in SARS-CoV spike exhibit an electrostatic clearly dissimilar from the sialic binding site in MERS-CoV spike, making very unlikely the interaction with sialic acid in that region.

In addition, in Fig.4, we propose a multiple sequence alignment - with software Clustal Omega[41] - between the three spike proteins.

Importantly the proposed SARS-CoV-2 binding site, besides being structurally in a surface region bordering the corresponding MERS pocket, is composed by a set of consecutive residues (sequence number 73-76) constituting an insertion in respect to SARS-CoV spike sequence. Thus, this insertion in the A-domain of the spike protein could confer to SARS-CoV-2 the capability of infecting human cells in a dual strategy, making possible the high diffusion speed of this virus.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.006197. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

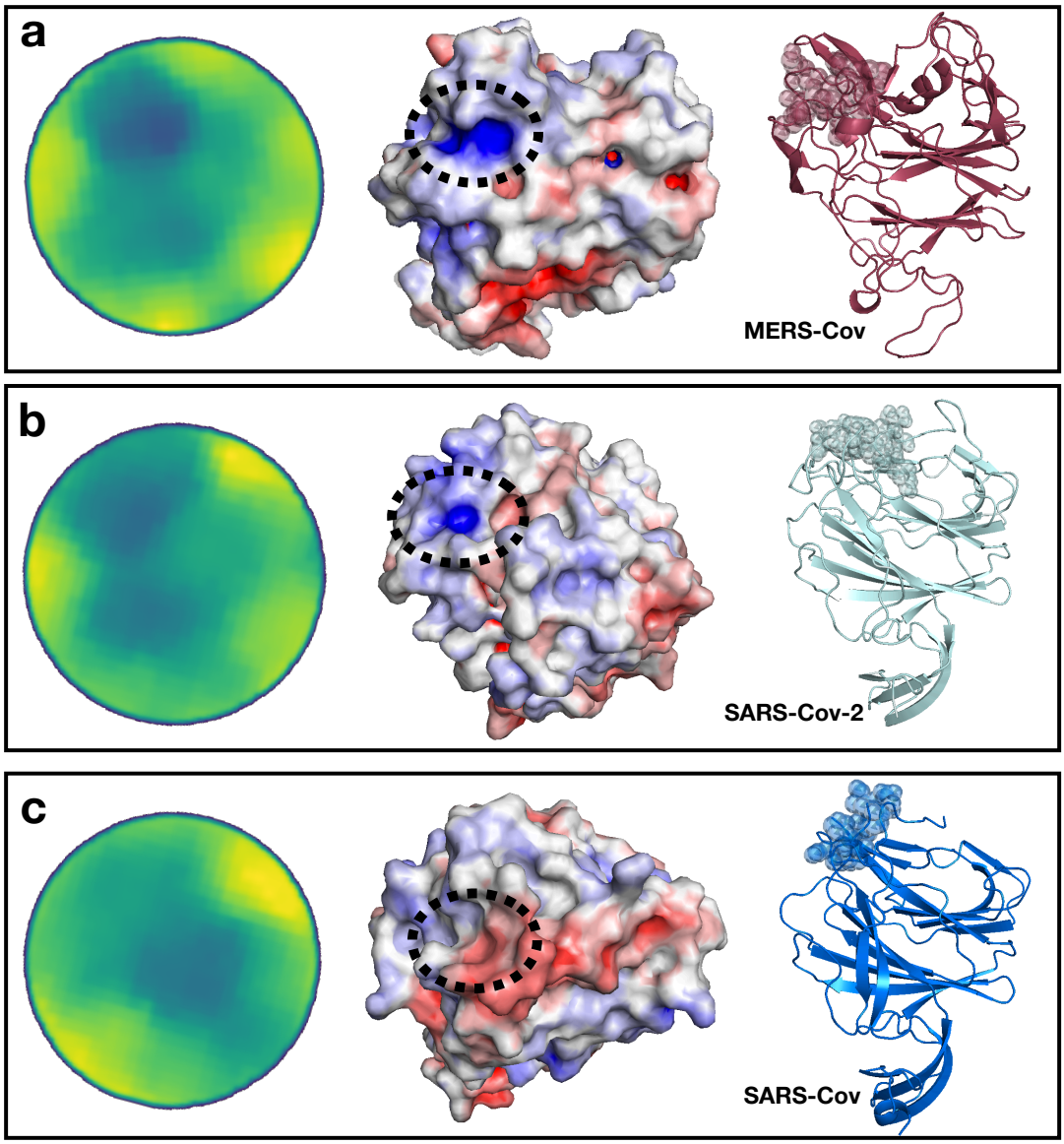

FIG. 3: **Identification of a SARS-CoV-2 spike region very similar to the sialic acid binding site on MERS-CoV spike.** **a)** From left to right, projected region of the real interaction between MERS-CoV and sialic acid, electrostatic potential surface in the same region and cartoon representation of the MERS-CoV spike protein with highlighted the binding site. **b)** Putative sialic acid binding region on SARS-CoV-2 as predicted by our Zernike-based method. From left to right, projected region of putative interaction between SARS-CoV and sialic acid, electrostatic potential surface and cartoon representation of the SARS-CoV spike protein with highlighted the binding site. **c)** Same as b) but for SARS-CoV spike protein.

III. DISCUSSION

A blind prediction of the interaction regions between molecules is still an open challenge, despite the great steps that have been made. However, the need for fast and reliable theoretical and computational tools, capable to guide and speed-up experiments, becomes especially important when we face emergencies like the present one. Emergency caused by the novel SARS-CoV-2 human infection, which is spreading with an impressive rate, such that the World Health Organization officially declared it a pandemic. Indeed, the last few months have seen extensive studies about the virus-host interactions focusing in particular on the various stages of the cell internalization mechanism. Recent works found that, in analogy with the case of SARS-CoV, SARS-CoV-2 uses too its spike protein to bind to ACE2 receptors, mostly associated in the lower respiratory ways. Further experimental investigations highlighted a comparable receptor binding affinity between the novel coronavirus and the older SARS-CoV, even if the binding regions display a certain degree of variability [21]. The modest difference in

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.006197. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

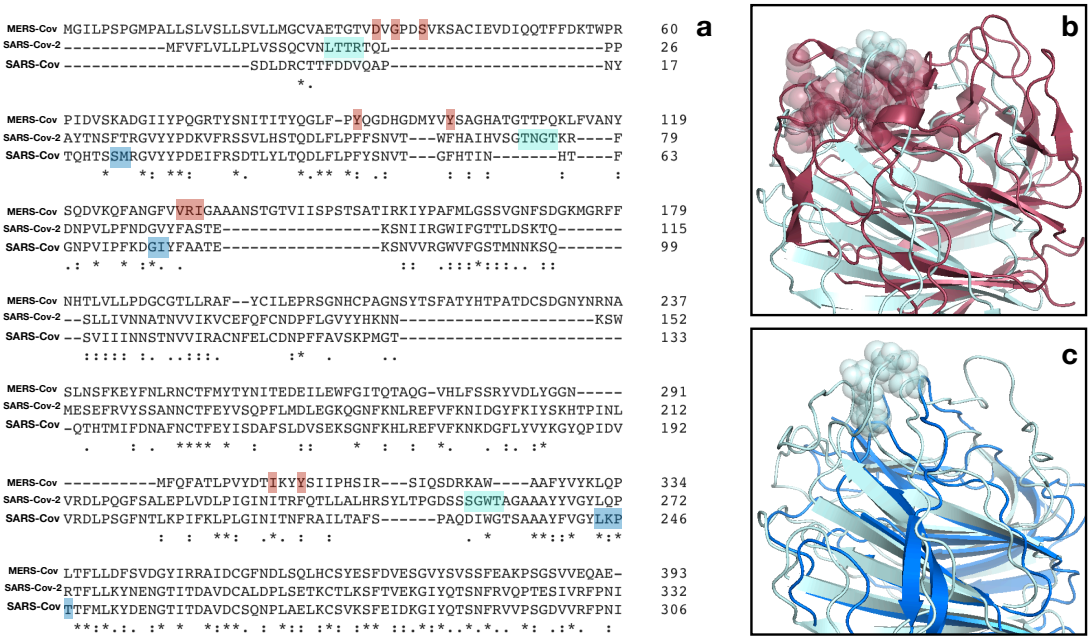

FIG. 4: **Sequence and structure comparison of the N-terminal region of MERS-CoV, SARS-CoV-2 and SARS-CoV. a)** A multiple sequence alignment between the MERS-CoV, the SARS-CoV and the SARS-CoV-2 spike protein sequence. **b)** Structural comparison between MERS-CoV and SARS-Cov-2 A-domain. The sialic acid binsing site for MERS-Cov spike and the proposed binding site on SARS-Cov-2 spike are highlighted. **c)** Structural comparison between SARS-CoV and SARS-Cov-2 A-domain. The proposed binding site on SARS-Cov-2 has no correspondence in SARS-CoV structure.

binding affinity seems insufficient to explain the higher human-human transmission rate in respect to SARS-CoV and the overall sequence variability suggest that SARS-CoV-2 may have optimized in other directions, such as in acquiring the ability of binding to other receptors [19].

In this work, we present a new fast computational method that compactly summarizes the morphological properties of a surface region of a protein. Testing the unsupervised method on a large dataset of protein-protein interactions, we proved its ability to correctly recognize the high shape complementarity occurring between interacting surfaces. Analizing the newly available experimental structures of SARS-CoV-2 Spike protein in complex with human ACE2, we found that the binding region presents indeed a comparable (slightly slower) shape complementarity with the analogous complex of SARS-CoV. Such a minimal difference enforces the hypothesis that the apparent higher fitness of SARS-CoV-2 should lie elsewhere. In particular, looking at other members of the large coronavirus family, one finds that many members developed the ability to bind to two distinct receptors, with one binding site in the C-terminal domain of the S-protein (like ACE2 for SARS-CoV and SARS-CoV-2) and the other situated in the N-terminal region, usually able to bind to sugar-like molecules. In particular, MERS-CoV has been found able to bind to sialic acid receptors both in camel, human and bat cells. Applying our method to the sialic acid binding region, which has been recently determined experimentally in MERS-CoV, we have found an exceptionally similar region in the corresponding region of SARS-CoV-2 spike. We speculate that this region, similar in structure to the MERS-CoV one and absent in SARS-CoV (see Figure 4), could be able to mediate a low-affinity but high-avidity interaction with sialic acid. In conclusion, we propose that this dual cell entry mechanism can explain the high diffusion speed this virus exhibits and we strongly encourage a more accurate investigation of this observation.

IV. METHODS

A. Datasets

Protein-Protein Dataset

A dataset of protein-protein complexes experimentally solved in x-ray crystallography is taken from [25].

We only selected pair interactions regarding chains with more than 50 residues. The Protein-Protein dataset is therefore composed by 4634 complexes.

Experimental protein structures

- Complex between SARS-CoV spike protein and human ACE receptor: PDB code 6ACJ.
- Complex between SARS-CoV-2 spike protein and human ACE receptor: PDB code 6M17.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.006197. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

- Complex between MERS spike protein and sialic acid: PDB code 6Q07.
- Unbound SARS-CoV spike protein: PDB code 6CRV.
- Unbound SARS-CoV-2 spike protein: modeled by I-TASSER server[42].

B. Computation of molecular surfaces

We use DMS [43] to compute the solvent accessible surface for all proteins structure, given their x-ray structure in PDB format [44], using a density of 5 points per $\AA^2$ and a water probe radius of 1.4 $\AA$. The unit normals vector, for eac point of the surface was calculated using the flag $-n$.

C. Patch selection and space reduction

Given a molecular surface described as a set of point in a three dimensional Cartesian space, and a region of interest on this surface, we define a surface patch, Σ as the intersection of the region of interest and the surface. In principle, the region of interest can have an arbitrary shape, in this work we chose to use a spherical region having radius R_s and one point of the surface as center (see Figure 1a). Once the patch is selected, we fit a plane that passes though the points in Σ, and we reorient the patch in such a way to have the z-axis perpendicular to the plane and going through the center of the plane. Then given a point C on the z-axis we define the angle θ as the largest angle between the perpendicular axis and a secant connecting C to any point of the surface Σ. C is then set in order that $\theta = 45°$. r is the distance between C and a surface point. We then build a square grid, and we associate each pixel with the mean r of the points inside it. This 2D function can be expanded in the basis of the Zernike polynomials: the norm of the coefficients of this expansion constitute the Zernike invariant descriptors, which are invariant under rotation in the space. In the next section we provide a brief summary of the main features of the Zernike basis. There are several good reviews, like [45] that offer more detailed discussions. A schematic representation of the procedure is shown in Figure 1b-d.

D. 2D Zernike polynomials and invariants

Given a function $f(r, \phi)$ (polar coordinates) defined inside the region $r < 1$ (unitary circle), it is possible to represent the function in the Zernike basis as

$$f(r, \phi) = \sum_{n=0}^{\infty} \sum_{m=0}^{m=n} c_{nm} Z_{nm} \tag{1}$$

with

$$c_{nm} = \frac{(n+1)}{\pi} \langle Z_{nm} | f \rangle = \\ = \frac{(n+1)}{\pi} \int_0^1 drr \int_0^{2\pi} d\phi Z^*_{nm}(r, \phi) f(r, \phi). \tag{2}$$

being the expansion coefficients. The Zernike polynomials are complex functions, composed by a radial and an angular part,

$$Z_{nm} = R_{nm}(r) e^{im\phi}. \tag{3}$$

where the radial part for a certain couple of indexes, n and m, is given by

$$R_{nm}(r) = \sum_{k=0}^{\frac{n-m}{2}} \frac{(-1)^k (n-k)!}{k! \left(\frac{n+k}{2} - k\right)! \left(\frac{n-k}{2} - k\right)!} r^{n-2k} \tag{4}$$

In general, for each couple of polynomials, one finds that

$$\langle Z_{nm} | Z_{n'm'} \rangle = \frac{\pi}{(n+1)} \delta_{nn'} \delta_{mm'} \tag{5}$$

which ensures that the polynomials can form a basis and knowing the set of complex coefficients, $\{c_{nm}\}$ allows for a univocal reconstruction of the original image (with a resolution that depends on the order of expansion, $N = max(n)$). We found that $N = 20$, which corresponds to a number of coefficient of 121, allows for a good visual reconstruction of the original image.

By taking the modulus of each coefficient ($z_{nm} = |c_{nm}|$), a set of descriptors can be obtained which have the remarkable feature of being invariant for rotations around the origin of the unitary circle.

The shape similarity between two patches can then be assessed by comparing the Zernike invariants of their associated 2D projections. In particular, the similarity between patch i and j is measured as the Euclidean distance between the invariant vectors, i.e.

$$d_{ij} = \sqrt{\sum_{k=1}^{M=121} (z_i^k - z_j^k)^2} \tag{6}$$

E. Evaluation of similarity and complementarity

When comparing patches, the relative orientation of the patches before the projection in the unitary circle must be evaluated. Intuitively, if we search for similar regions we must compare patches that have the same orientation once projected in the 2D plane, i.e. the solvent exposed part of the surface must be oriented in the same

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.006197. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

direction for both patches, for example as the positive z-axis. If instead we want to assess the complementarity between two patches, we must orient the patches contrariwise, i.e. one patch with the solvent-exposed part toward the positive z-axis ('up') and the other toward the negative z-axis ('down').

F. Blind search of binding sites

The velocity of the procedure that from a patch in the 3D surface produces the set of invariant descriptors, allows for a vast screening of couples of surfaces in order to look for both similar and also complementary regions. In order to identify the binding region given a couple of proteins, we can associate to each point of one surface a vector of Zernike invariants associate to the 'up' patch having that point as center and another set of invariants to each point of the other surface (with 'down' orientation). Then for each point i of say, protein 1, we can compute the Euclidean distance with all the points of the other surface associated to protein 2 and associate to point i the minimum found distance and viceversa for protein 2 (see Figure 1e-f). A smoothing process of the surface point values is applied in order to highlight the signal in the regions characterized mostly by low distance values (high shape complementarity).

Acknowledgment

The authors would like to thank Prof. Gian Gaetano Tartaglia for the very helpful discussions.

[1] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, et al., The Lancet **395**, 497 (2020).
[2] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, et al., New England Journal of Medicine (2020).
[3] A. C. Walls, Y.-J. Park, M. A. Tortorici, A. Wall, A. T. McGuire, and D. Veesler, Cell (2020).
[4] C. Drosten, S. Günther, W. Preiser, S. Van Der Werf, H.-R. Brodt, S. Becker, H. Rabenau, M. Panning, L. Kolesnikova, R. A. Fouchier, et al., New England journal of medicine **348**, 1967 (2003).
[5] T. G. Ksiazek, D. Erdman, C. S. Goldsmith, S. R. Zaki, T. Peret, S. Emery, S. Tong, C. Urbani, J. A. Comer, W. Lim, et al., New England journal of medicine **348**, 1953 (2003).
[6] A. M. Zaki, S. Van Boheemen, T. M. Bestebroer, A. D. Osterhaus, and R. A. Fouchier, New England Journal of Medicine **367**, 1814 (2012).
[7] J. Xu, S. Zhao, T. Teng, A. E. Abdalla, W. Zhu, L. Xie, Y. Wang, and X. Guo, Viruses **12**, 244 (2020).
[8] J. Lessler, N. G. Reich, R. Brookmeyer, T. M. Perl, K. E. Nelson, and D. A. Cummings, The Lancet infectious diseases **9**, 291 (2009).
[9] Y. Yang, Q. Lu, M. Liu, Y. Wang, A. Zhang, N. Jalali, N. Dean, I. Longini, M. E. Halloran, B. Xu, et al., medRxiv (2020).
[10] E. Prompetchara, C. Ketloy, and T. Palaga, Asian Pacific J. allergy Immunol **10** (2020).
[11] S. Su, G. Wong, W. Shi, J. Liu, A. C. Lai, J. Zhou, W. Liu, Y. Bi, and G. F. Gao, Trends in microbiology **24**, 490 (2016).
[12] F. Li, W. Li, M. Farzan, and S. C. Harrison, Science **309**, 1864 (2005).
[13] F. Li, Journal of virology **82**, 6984 (2008).
[14] W. Li, C. Zhang, J. Sui, J. H. Kuhn, M. J. Moore, S. Luo, S.-K. Wong, I.-C. Huang, K. Xu, N. Vasilieva, et al., The EMBO journal **24**, 1634 (2005).
[15] R. L. Graham and R. S. Baric, Journal of virology **84**, 3134 (2010).
[16] L. Kuo, G.-J. Godeke, M. J. Raamsman, P. S. Masters, and P. J. Rottier, Journal of virology **74**, 1393 (2000).
[17] L. Du, Y. He, Y. Zhou, S. Liu, B.-J. Zheng, and S. Jiang, Nature Reviews Microbiology **7**, 226 (2009).
[18] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche, et al., Cell (2020).
[19] P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang, et al., Nature pp. 1–4 (2020).
[20] R. Yan, Y. Zhang, Y. Li, L. Xia, Y. Guo, and Q. Zhou, Science (2020).
[21] K. Andersen, A. Rambaut, W. Lipkin, E. Holmes, and R. Garry, ARTIC Network **17** (2020).
[22] C. Wu, Y. Liu, Y. Yang, P. Zhang, W. Zhong, Y. Wang, Q. Wang, Y. Xu, M. Li, X. Li, et al., Acta Pharmaceutica Sinica B (2020).
[23] B. R. Donald, *Algorithms in structural molecular biology* (MIT Press, 2011).
[24] T. Kortemme, L. A. Joachimiak, A. N. Bullock, A. D. Schuler, B. L. Stoddard, and D. Baker, Nature structural & molecular biology **11**, 371 (2004).
[25] P. Gainza, F. Sverrisson, F. Monti, E. Rodolà, D. Boscaini, M. Bronstein, and B. Correia, Nature Methods **17**, 184 (2020).
[26] A. Shulman-Peleg, R. Nussinov, and H. J. Wolfson, Journal of molecular biology **339**, 607 (2004).
[27] D. Duhovny, R. Nussinov, and H. J. Wolfson, in *International workshop on algorithms in bioinformatics* (Springer, 2002), pp. 185–200.
[28] K. A. Sharp and B. Honig, Annual review of biophysics and biophysical chemistry **19**, 301 (1990).
[29] S. Daberdaku and C. Ferrari, Bioinformatics **35**, 1870 (2019).
[30] M. A. Tortorici, A. C. Walls, Y. Lang, C. Wang, Z. Li, D. Koerhuis, G.-J. Boons, B.-J. Bosch, F. A. Rey, R. J. de Groot, et al., Nature structural & molecular biology **26**, 481 (2019).
[31] R. J. Hulswit, Y. Lang, M. J. Bakkers, W. Li, Z. Li, A. Schouten, B. Ophorst, F. J. van Kuppeveld, G.-J. Boons, B.-J. Bosch, et al., Proceedings of the National Academy of Sciences **116**, 2681 (2019).
[32] C. Schwegmann-Weßels and G. Herrler, Glycoconjugate journal **23**, 51 (2006).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.24.006197. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC 4.0 International license.

[33] Y.-J. Park, A. C. Walls, Z. Wang, M. M. Sauer, W. Li, M. A. Tortorici, B.-J. Bosch, F. DiMaio, and D. Veesler, Nature Structural & Molecular Biology **26**, 1151 (2019).

[34] V. Venkatraman, Y. D. Yang, L. Sael, and D. Kihara, BMC bioinformatics **10**, 407 (2009).

[35] D. Kihara, L. Sael, R. Chikhi, and J. Esquivel-Rodriguez, Current Protein and Peptide Science **12**, 520 (2011).

[36] L. Di Rienzo, E. Milanetti, R. Lepore, P. P. Olimpieri, and A. Tramontano, Scientific reports **7**, 1 (2017).

[37] Y. Wan, J. Shang, R. Graham, R. S. Baric, and F. Li, Journal of virology (2020).

[38] W. Li, R. J. G. Hulswit, I. Widjaja, V. S. Raj, R. McBride, W. Peng, W. Widagdo, M. A. Tortorici, B. van Dieren, Y. Lang, et al., Proceedings of the National Academy of Sciences **114**, E8508 (2017).

[39] M. A. Tortorici, A. C. Walls, Y. Lang, C. Wang, Z. Li, D. Koerhuis, G.-J. Boons, B.-J. Bosch, F. A. Rey, R. J. de Groot, et al., Nature Structural & Molecular Biology **26**, 481 (2019).

[40] K. Kinoshita and H. Nakamura, Bioinformatics **20**, 1329 (2004).

[41] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, et al., Molecular systems biology **7** (2011).

[42] J. Yang and Y. Zhang, Nucleic acids research **43**, W174 (2015).

[43] F. M. Richards, Annual review of biophysics and bioengineering **6**, 151 (1977).

[44] H. M. Berman, P. E. Bourne, J. Westbrook, and C. Zardecki, in *Protein Structure* (CRC Press, 2003), pp. 394–410.

[45] V. Lakshminarayanan and A. Fleck, Journal of Modern Optics **58**, 545 (2011).