Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Computers in Biology and Medicine 128 (2021) 104124

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

Techniques assisting peptide vaccine and peptidomimetic design. Sidechain exposure in the SARS-CoV-2 spike glycoprotein

B. Robson

Ingine Inc. Cleveland Ohio USA and the Dirac Foundation, Oxfordshire, UK

ARTICLE INFO

Keywords:
Spike glycoprotein
Exposure
Accessibility
Glycosylation
Conformation
Disorder
Coronavirus
SARS-CoV-2
COVID-19

ABSTRACT

The aim of the present study is to discuss the design of peptide vaccines and peptidomimetics against SARS-COV-2, to develop and apply a method of protein structure analysis that is particularly appropriate to applying and discussing such design, and also to use that method to summarize some important features of the SARS-COV-2 spike protein sequence. A tool for assessing sidechain exposure in the SARS-CoV-2 spike glycoprotein is described. It extends to assessing accessibility of sidechains by considering several different three-dimensional structure determinations of SARS-CoV-2 and SARS-CoV-1 spike protein. The method is designed to be insensitive to a distance limit for counting neighboring atoms and the results are in good agreement with the physical chemical properties and exposure trends of the 20 naturally occurring sidechains. The spike protein sequence is analyzed with comment regarding exposable character. It includes studies of complexes with antibody elements and ACE2. These indicate changes in exposure at sites remote to those at which the antibody binds. They are of interest concerning design of synthetic peptide vaccines, and for peptidomimetics as a basis of drug discovery. The method was also developed in order to provide linear (one-dimensional) information that can be used along with other bioinformatics data of this kind in data mining and machine learning, potentially as genomic data regarding protein polymorphisms to be combined with more traditional clinical data.

1. Introduction

1.1. Background

Today, the word "coronavirus" needs no introduction. However, coronaviruses have only been known since 1966 [1,2] and until 2019 they have only been considered a serious threat to humanity because of the SARS outbreak in 2002 [3]. Only recently (January 2020), the isolate obtained from patients associated with the location of the Wuhan Seafood market [4] provided the first widely available genome for study of SARS-CoV-2. This is now the preferred name for the causative agent of the current COVID-19 pandemic (with SARS-COV-1 indicating the earlier SARS virus). The final confirmed genomic sequence MN908947.3 entered on the GenBank database on January 23, 2020 was the basis of a rapid response by present author (BR) using bioinformatics analysis for design of synthetic vaccines and peptidomimetic therapeutics, and also using knowledge-gathering and processing tools [5–9]. With little or no direct experimental data for SARS-CoV-2 in January, SARS-CoV-1 was used as a reference model, because of a high degree of homology [5,6], particularly of the spike protein. This was even though the prevalent view at the time (perhaps encouraged by authorities) was that the new epidemic was not SARS [6].

At the time of the present study and preparation of the paper, relatively little time had elapsed to enable detailed general analysis of the proteins produced by the virus genome compared with other proteins. Focus by researchers had naturally jumped directly onto aspects that lead to specific proposals for vaccines and therapeutic agents. The situation is now rapidly changing (see discussion in Conclusions Section 6), although prevention and cure is still the main objective. But from the outset, there were extensive applications of traditional sequence-based bioinformatics (e.g. Refs [5–7]). Use of that depends largely on a direct mapping of one-dimensional information from coronavirus genomes rather than use of extensive three dimensional structural, energetic, and dynamic analysis, but it still yields "low lying fruit" as valuable information for the design of peptide synthetic vaccines and peptidomimetic therapeutics. In Refs [5–7,9] the strategy for this design was primarily one of finding subsequences of very roughly 12 amino acid residues in an order that is essentially conserved across many coronaviruses. The specific motivations for this approach are described in Section 1.4, but one consequence was that a sequence motif KRSFIEDLLFNKV was proposed as an important target [5–7]. The

E-mail address: barryrobson@ingine.com.

https://doi.org/10.1016/j.compbiomed.2020.104124
Received 16 September 2020; Received in revised form 6 November 2020; Accepted 11 November 2020
Available online 21 November 2020

strategy used above is not the only one that can be applied readily at the one-dimensional, sequence level. In Ref. [8] the tactic was somewhat different, by exploring a possible neglected non-covalent sialic acid glycan binding function of the SARS-CoV-2 spike protein. However, this is essentially a similar one-dimensional task of looking for subsequences characteristic of a suspected function except that the exact order of the amino acid residues involved is somewhat less important.

1.2. Purpose of the present paper

The aim of the present study is to discuss the design of peptidomimetics against SARS-COV-2, to develop and apply a method of protein structure analysis that is particularly appropriate to applying and discussing such design, and also to use that method to summarize some important features of the SARS-COV-2 spike protein sequence. It also reflects a larger effort by the present author concerned with converting complex, three-dimensional information about proteins to a one-dimensional description. Linear representations as annotations of the amino acid residue sequence are well suited for data mining and machine learning generally. Such information in combination with the clinical data in future clinical decision support systems may well be of benefit in disease diagnosis and selection of best therapy based on genomics. The above is a longer-term vision. More immediately, issues concerning the usefulness or otherwise of the motif KRSFIEDLLFNKV were a motivation for the present study. While this motif remains very promising and potentially important for several reasons discussed below in Sections 1.3 to 1.6, collaborators in the biopharmaceutical industry were initially concerned regarding the extent and duration of exposure (Section 1.5). However, the present paper is not confined to that motif, and covers the entire SARS-COV-2 spike protein. The above motif is used in this Introduction primarily to exemplify the issues that need to be addressed more generally in considering molecular defenses against SARS-CoV-2.

Given the fact that the focus is on the virus, the effort of developing a new method for assessing exposure of sidechains to solvent would not seem a priority. There is an abundance of methods for assessing atom, sidechain, or residue exposure. Early in the present project, these were explored (e.g. see Results Sections 4.1-4.3), but finally a novel algorithm was developed to meet specific requirements. A difficulty to overcome was that different methods of assessing exposure of residues to the solvent show great variation in results. That applies to assessing degree of exposure of specific residues in proteins, but it is reflected in the variations in average degree of exposure for each of the 20 naturally occurring amino acid residues, which show considerable diversity (see Results Section 4.3). Consequently, a motivation for developing the approach described in Theory Section 2 was that (a) it would make the method less sensitive to a critical interatomic distance parameter (here called R_{max}), while (b) giving plausible agreement with other measures for the average properties of the 20 naturally occurring amino acid residues that make sense (without artificially forcing that result). There were also several simpler practical considerations. These were that the approach (c) be easily adjustable to describe potential access to solvent molecules over a continuous range of sizes (although deduced indirectly from protein atom coordinates), (d) handle missing sections of protein chain in structure determinations, (e) provide a flexible way to select molecules in the system that were considered in any study as being intrinsic to the system of interest as opposed to incidental (e.g. peculiar to the experimental setup for the three dimensional structure determination), and (f) include covalently bound glycans and potential sites for glycosylation, as well as (g) deliver the results in a convenient one dimensional format, as mentioned earlier above. The main purpose for these requirements is to help develop methods to automate the design of peptidomimetics. It is an important aim, as follows.

1.3. Design of peptidomimetics

A peptidomimetic agent as the term is used here is a compound that is not simply a copy of a subsequence of interest but rather a modification of the subsequence, usually containing D-amino acids as in the case of a retroinverso construct [6,7], or unnatural amino acids or other chemical groups. Despite that, researchers still usually mean that it has a recognizable relation to the original subsequence, at least on deeper examination. Within the term "peptidomimetic" one might also include peptides or peptide-like compounds that are less obviously related, but which are experimentally generated in the laboratory in some way by starting from the synthetic peptide conforming more closely to the original subsequence. All these typically depend primarily on identification or prediction of a section of amino acid residues from the protein of interest to serve as the plausible starting point. Methods of converting these into a synthetic peptide with the required activity were described and reviewed in Ref. [6], including both general recipes and worked examples for the KRSFIEDLLFNKV motif, both as a potential for a peptidomimetic and a synthetic peptide vaccine. The intention of the present paper is not to repeat that fairly straightforward exercise for every potentially interesting subsequence in the spike protein. Rather, it is to provide a kind of "directory" of information to which a researcher may easily refer, while also armed with the above recipes as one of several tools in his or her toolbox. For example, the numeric characters in the string 79849888777 aligned beneath the characters of the subsequence GSTPCNGVEGF are exposure scores that, along with other one-character notation, indicate the following. It is region in the ACE2 (angiotensin converting enzyme type 2) binding domain which is well exposed adjacent to a more tightly binding region adhering to the ACE2 receptor binding site and a region of antibody binding, although it is disordered in two of the three spike protein monomers in many structure determinations without ACE2 or antibody. This is discussed in relation to Block 13 in Section 4.6. Note that blocks of information of this kind, for consecutive sections of spike protein sequence each 60 amino acid residues long, are numbered for convenience. A further summary string under the above sequence of exposure scores gives no evidence of shielding of this subsequence by glycosylation, nether from within this region of the sequence itself nor from nearby in space in the three-dimensional structure.

Discussion on use of subsequence as a starting point for a peptidomimetic is given in Section 5.1 and particularly in Section 5.4. Although as stated above peptidomimetics are probably best considered as a tool in drug discovery, i.e. as a useful first step amongst others in development of convenient "in a pill" drugs that are even less obviously related to the subsequence, there is at the outset always the possibility that an original (early-predicted) peptidomimetic might work and serve directly as a preventative and curative agent, e.g. in aerosols. The ease of synthesis of peptides comprising L and/or D amino acids by modern techniques means that such candidates can be early and readily eliminated, without great cost.

1.4. Importance of using conserved subsequences

A high degree of conservation of a subsequence is particularly important for design of a synthetic peptide vaccine, a peptidomimetic, and even therapeutic drugs in general, for two reasons. First, the implied lack of variation across evolutionary time, including in emergence of new virus strains, suggests that a subsequence is a motif that has a function important to replication and/or survival of the virus. Identification of such functions is of course of huge interest, but whatever that function might be found to be, it would seem desirable to see what happens if researchers can block it [9]. Second, under the selective pressure of vaccines and drugs, RNA virus evolution can accelerate rapidly so as to render those weapons useless in perhaps just a few months, and intuitively this is much less likely to occur if it already resisted change for long periods of natural evolutionary time. While SARS-CoV-2 accepts mutations at a slower rather than many RNA viruses, all RNA virus evolve relatively quickly, and it is estimated that there may be 10^{26} SARS-CoV-2 RNA molecules in the world,

representing a very large parallel computer to competing against human efforts at defense [9].

As noted by our colleagues at HitGen, in Chengdu City, Sichuan Province, China, spike glycoprotein S2 subdomain sequences are in general quite well conserved at least among β-coronaviruses, which suggests that the fusion mechanisms could be similar. S2, typically considered as spanning residues 686–1213 of the spike protein, could be an attractive target due to its high degree of sequence conservation amongst divergent human coronaviruses. For example, EK1, a pan-coronavirus fusion inhibitor, targeted the HR1 domain (894–966) of S2 protein. It could inhibit infection by many human coronaviruses. EK1C4A, one of the lipopeptides derived from EK1, has recently been proved to be a potent fusion inhibitor against SARS-CoV-2 S protein-mediated membrane fusion and pseudovirus infection [11]. The motif KRSFIEDLLFNKV, spike glycoprotein residues 814–826 in the S2' spike glycoprotein subunit, was favored in Refs [5–7] and can be used as an example, although descriptions of the principles and complexities involved are more generally applicable. The motif is particularly well conserved across the coronaviruses [5,6] and even detectable as a trace in other nidoviruses [7]. In contrast, in the receptor binding domain (RBD), which directly binds to ACE2, the sequence of S1 domain between SARS-CoV and SARS-CoV-2, especially the receptor binding motif (RBM) (50% identity), is much less conserved (64% identity) compared to that of S2 protein (90% identity), although the homologies for the RBD domain overall with related coronaviruses is much higher (74% identity). A neutralizing antibody, CR3022, targets an epitope in the RBD region, distal from the RBM site, that enables cross-reactive binding between SARS-CoV-2 and SARS-CoV and which is considered as "highly conserved". The present author, however, considered this also as rather too variable across coronaviruses for a long-term vaccine solution (or therapeutic based on it) [9].

1.5. Importance of using accessibility

A further consideration for the design of peptidomimetics against the action any protein is that the subsequence of interest is either (a) adequately exposed or (b) readily exposable by binding interactions between an antibody or therapeutic agent, or (c) at least sufficiently exposed as some stage of the life cycle of protein, rather than buried within the protein structure. This is not such an important consideration if the strategy is to block competitively a protein site Y to which a protein X, such as that of an invading pathogen, is already known (or readily shown) experimentally to bind. For example, in the case of SARS-CoV-2, the most plausible target as the host protease responsible for the final activation cleavage under normal circumstances, initially by analogy with SARS-CoV-1, is the transmembrane serine protease TMPRSS2, of which the most likely cleavage point corresponds to the arginine R in the above-mentioned motif KRSFIEDLLFNKV [5,6]. It been subsequently demonstrated experimentally that SARS-CoV-2 entry also depends on binding to angiotensin converting enzyme II (ACE2) followed by a cleavage by TMPRSS2 at the above S2' site [10]. A TMPRSS2 inhibitor approved for clinical use has been shown to block SARS-CoV-2 entry and could constitute a treatment option [10]. However, accessibility is the unavoidably important consideration for any strategy that directly attacks the protein X (e.g. Ref. [11]). That also includes attack by the antibodies raised by vaccinating the patient, or by injecting the patient with preformed antibodies, for passive immunization. Again considering the example of KRSFIEDLLFNKV, it was considered as the sequence of an exposed or exposable spike glycoprotein site because the N-terminal end of the same motif *vitro* in SARS-CoV-1 was known to be subject to cleavage not only by TMPRSS2 as above, but also by a variety of proteases *in vitro* [5].

Our colleagues at HitGen nonetheless had concerns to the extent of accessibility of the motif as a target for a direct attack by antibodies or designed agents because research on the cleavage process of S protein of MERS-CoV showed that S1/S2 cleavage comes first and consequentially S2′ site is exposed for cleavage, indicating that S2′ site is shielded to some extent [12]. As they noted, the S2′ site locates at the stem of bundle-of-flower like trimeric S proteins [13]. The down "CTD1" of S1 protein locates immediately above the S2 subunit and have direct interaction with helix linker 2. Opening of CTD1, especially by binding the receptor, would remove the steric restraints on helix linker 2, triggering the release of the S1 subunits and probably simultaneously allowing the extension of pre-fusion S2 helixes to form the post-fusion S2 long helix bundle [14]. Here the cleavage at the S1/S2 site associated with a PIGAG motif [6] appears to be required to expose the KRSFIEDLLFNKV site, cleavage of which is perhaps the most essential step, but a second step nonetheless. It makes evolutionary sense that this sequence of events may protect the KRSFIEDLLFNKV site from antibody attack until that brief period in which it is needed. That is, the S2' site in the long-lasting *pre-fusion* state may be less exposable than originally thought, so that the steric hindrance could prevent the binding of at least a conventional antibody. However, the situation is complex and partially contested by immunological data as discussed immediately below, and as analyzed in the present paper.

1.6. Complexity of the notion of accessibility

A degree and character of steric hindrance sufficient to prevent proteolytic cleavage at a site does not necessarily mean that the adjacent features are buried within the protein structure. Fig. 1 shows the above motif as comprising the N-terminus of an α-helix that is partly exposed at the protein surface. Accessibility is also a matter of degree in time as well as in space, even for the prefusion form. Binding to the motif depends on relative strengths of two kinds of binding in the manner of a "tug of war": there is a free energy of interaction between part of the spike protein that might cover the site and the rest of the spike protein, and the free energy of interaction of the site and the antibody or synthetic ligand. There is also the matter of the frequency and duration of any fluctuations that expose the site, i.e. ultimately a matter of energy barriers. Indeed, the access of the motif to a variety of proteases *in vitro* naturally suggests that the binding strength of peptide to protein interactions can be sufficient. It is often stated as around −12 kcal/mol for overall free energy of binding. Although there is a well-known difficulty in computing entropy contributions, it can be estimated the total change in intramolecular (bond rotational) entropy of a peptide ligand as potential therapeutic is roughly $T\Delta S = 1.5$/mole per residue at 300 K [6], i.e. approximately 20 kcal/mol for an analogue of a 13 residue motif, and the computed enthalpy contribution of protease-substrate interactions

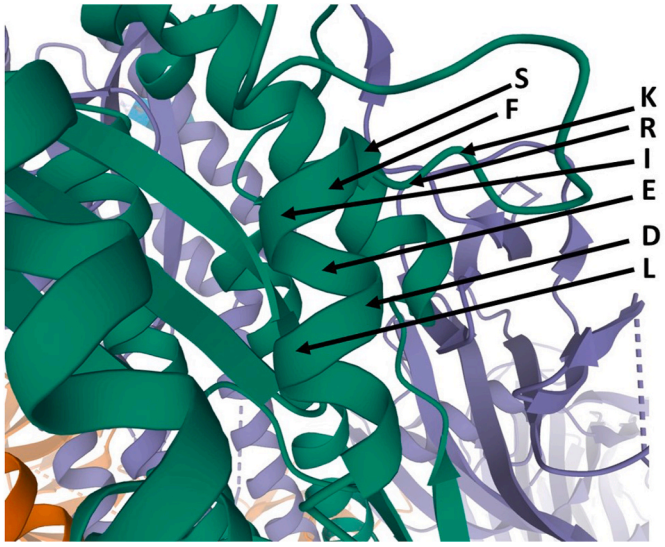

Fig. 1. 6VXX SARS-C0V-2 Spike protein Closed State Chain B (dotted lines indicate disordered part of chain).

can be as strong as −30 to −40 kcal/mol [14].

Immunological evidence of accessibility is particularly important but even here the picture is slightly fuzzy. A recent Google initiative at https://www.iedb.org/has a large amount of information that can be queried regarding the SARS-CoV-1 virus of the earlier SARS outbreak. A query for KRSFIEDLLFNKV or overlap with most of it with a BLASTp score of at least 90% suggested that a majority of at least 45 studies confirmed B epitope (antibody inducing) activity binding in SARS human or animal patient sera and a majority MHC/T-epitope activity. However, it became apparent on examining the source papers that several of them also studied several epitopes and not all uses of the above motif were covered properly or were necessarily successful. At the time of writing it appears that 84–85% supported B-epitope activity for the motif, which requires accessibility. In a few cases neutralization was obtained. A broader range of 67–90% appeared to support T-epitope capability, but while T-epitope is important for a vaccine, it does not necessarily imply accessibility in the native conformation of the spike complex. Some important examples are as follows. See for example Ref. [15]. One study found that in some animals only one determinant (Leu 803 to Ala 828) was able to induce the antisera with the binding ability to the native S protein and the neutralizing activity to the SARS-CoV-1 pseudovirus [16]. The significantly higher levels of IgG antibodies specific to three (S791 = PLKPTKRSFIEDLLF], M207 and N161) of 42 peptides investigated were detectable in the post-infection sera from 23 (51%), 27 (60%) and 19 (42%) of 45 patients, respectively. The fluorescence intensity (FI) of the anti-SARS-CoV-1 spike protein at positions 791–805 (termed anti-S791) was highest among the peptides tested [16]. An example of a case in which it appeared less successful was Ref. [17]. In any event, it appears that the region of spike polypeptide chain associated with this subsequence KRSFIEDLLFNKV has to be exposable, and exposed at some stage, for infection to occur (e.g. Refs [18,19]). Some recent work on peptides designed from other regions of SARS-COV-2 are discussion in Discussion Section 5.4.

1.7. Related work

The computational approach to sidechain exposure as implemented in the present study differs significantly from previous methods, though it would not be considered revolutionary because there are many diverse methods for measuring exposure at the surface of a protein and for assessing the physical significance of that [20]. See, for example, Refs [21–23] for review. Although developed and tested using many other proteins, the present approach was refined and calibrated to be particularly suited to studies of the SARS-CoV-2 spike protein. It also reflects interests in designing peptide synthetic vaccines and peptidomimetcs more generally, and in providing a format appropriate to such purposes as well as for efficient data mining. The core feature of the method used here is based on distances between atoms, which is not itself unique (e.g. Ref. [23]). Indeed, the process of counting atoms in specified volume or shell of space so as to derive a *radial distribution function* is common in the molecular sciences in general [24]. However, there is a simple way to make results regarding definition of neighboring atoms far less sensitive to distance criteria and it brings results into good alignment with physicochemical and other exposure properties of sidechains. It is adapted from the radial distribution function as follows.

2. Theory

Theoretical and experimental aspects of amino acid and protein structure relevant to the present study have been reviewed in Ref. [20]. Specific theoretical and experimental aspects of residue exposure in proteins are reviewed in Refs. [21,22]. As discussed in Section 1.2, the present study is part of a larger effort concerned with expressing complex, three-dimensional information about proteins into a linear description suited for data mining and automated inference from it. This was in order to form the basis of near future studies in which such data about patient proteins is combined with the clinical data in order to benefit a patient, and more immediately as a basis for the design of peptidomimetics. To do this in a way suitable for data mining and other analytic methods means that a description of some important structural of functional aspect is described as a linear sequence, analogous to the way in which protein secondary structure is usually described (i.e. as a sequence of residues in an α-helical state H, a β-pleated sheet or extended state E, and a coil or loop state C). Various methods for identifying exposed surface sites have been available for years (e.g. see Refs. [20–23]). In the present case the purpose is to describe the exposure or accessibility of the sidechains along the sequence of a specified chain (subunit). The present method bases both in terms of interactions within that specified chain (subunit) and with all other entities present such as covalently bound glycans, ligands such as antibodies host cell receptors, and so forth. Interactions with solvent molecules that may be peculiar to the experimental setup for structure determination are ignored. However, any solvent ions would normally be included on the rationale that they may form a more persistent natural complex with a protein (if this is not the case, they may be readily excluded).

The basic raw score S_{raw} for each sidechain in a protein which is described below relates to, but is not identical to, the notion of a radial distribution function for the distributions of, say, molecules in a liquid, in chemical physics [24]. It is closer to the integral of it over distance, up to a specified distance, i.e. it addresses volumes of increasing radius from the coordinates of the sidechain atom of interest. In the present study, however, the average number of atoms within a set maximum distance (in this study mostly 6.5 Å), of sidechain atoms is expressed as the *average* per *sidechain atom* for each sidechain and, from this raw score, the final score for each sidechain is determined on a scale set by maximum and minimum raw scores in the system as a whole. Recall that the scores reported are for residues in a prespecified protein molecule for which an amino acid sequence is meaningful, say a subunit such as chain B of the SARS-CoV-2 spike glycoprotein, but the other protein molecules present, say chain A and C of the above, provide interactions which are counted, as well as those within B itself. Interactions within a residue are not counted, and interactions of less than 2 Å are considered as covalently bonded to the sidechain, and so are also not counted. Sidechain atoms are considered to be the Cγ (CG) atom and all atoms in the rest of the sidechain beyond, except for including Cα (CA) in glycine GLY and Cβ (CB) in alanine ALA. Otherwise, backbone atoms, i.e. amide N, carbonyl carbon C, carbonyl oxygen O, and Cα and Cβ are ignored. No hydrogen atoms H are considered. The raw score per sidechain may therefore be more formally represented as follows.

$$S_{raw} = N(sc)^{-1} \Sigma_{i=1,2,3 \ldots N(sc)} \Sigma_{j=1,2,3 \ldots N(sys)} F(R_{min}, R_{max}) \quad (1)$$

$F(R_{min}, R_{max}) = 1$ for any atom in the system other than the residue itself at a distance equal to greater than $R_{min} = 2$ Å and less than or equal to R_{max}, and $F(R_{min}, R_{max}) = 0$ otherwise. N(sc) is the number of sidechain atoms for the amino acid residue under consideration, and N(sys) the number in the molecular system in consideration, meaning in the protein, covalently linked molecules such as glycans, and ligands and/or antibodies of interest, but excluding solvent. The index i takes the count over the N(sc) sidechain atoms and j takes it over the other N(sys) atoms in the system considered.

A special so-called "glycoscore" is of particular interest as, in the present kind of case, it indicates a degree of by the SARS-CoV-2 virus's protective coat of sialic acid glycans. This score is represented by S_{raw} when any atom in the sidechain interacts with a glycan molecule covalent linked to the residue with an atom in the range R_{min} to R_{max}. This is essentially the same as saying that S_{raw} is greater than zero and involves an interaction with a glycan. The notion is that access to a residue is partly sterically restricted by the vicinity of the glycan chain of a glycosylated residue, and note that this restriction can be in space, not confined to interactions with glycans of neighboring glycosylated residues in the sequence. It is primarily of importance when the glycoscore

itself exceeds 0, in which case it is indicated by the charter '%' in a linear description of the residue's exposure. In practice that is done for the sequence of states representing the smoothed exposure score (see below).

The exposure score S_{exp} (sometimes referred to below as an accessibility score), is defined as

$$S_{exp} = \text{int}(11*(S_{max} - S_{raw}) / (S_{max} - S_{min})) \quad (2)$$

This gives for S_{exp} a convenient score of 0–9, allowing a one-character representation 0,1,2,3,4,5,6,7,8,9 to represent the exposure state of a residue regarding its sidechain. Occasionally a score of 10 is seen, an extreme state of protrusion for a sidechain that is represented by the character X. Note that in the present study S_{max} is not fixed as a constant but is the maximum S_{raw} in the spike protein complex being considered, and S_{min} is correspondingly the minimum value. For all but the smallest protein structures, however, S_{min} and S_{max} do not change significantly from systems to system, 6.4 and 10.5 respectively being typical values, for similar reasons to the consistency of a sidechain S_{raw} value under different values of R_{max} as discussed below.

3. Methods

Protein structure coordinates were obtained from the protein Data Bank (PDB) via https://www.rcsb.org/. Many proteins have been studied in preparation for the project but 6ZP5, 6ZP7, 6ACC, 6ACD, 7BYR, 6XC3, 6LXT and 6M0J, along with 6VW1, are of major interest below. These relate to the SARS-CoV-2 spike glycoprotein except for 6ACC and 6ACD which are SARS-CoV-1 spike glycoprotein structures. Note that some structures are significantly modified by genetically engineering and represent only the relevant domains of the spike glycoprotein in complex with another molecule, here the spike protein receptor binding domain (RBD) with the ACE2 domain or antibody domains of interest as discussed below. ACE2 refers to the angiotensin converting enzyme type 2 that represents the initial SARS-CoV-2 receptor. Fab relates to the antibody heads used in the structure determination study. Initially the B chain was of primary focus because it is directly involved in most such interactions of interest, but the overall study includes the three chains A, B, and C of the spike protein (a trimer). The following represent the main reports of analyses, "smoothed" relating to the treatment of scores described shortly below (and usually represented by the abbreviation "sm." in displays).

6ZP5 spike closed, sidechain exposure.
6ZP5 spike closed, exposure smoothed + spike glycan interactions.
6ZP7 spike open, sidechain exposure.
6ZP7 spike open, exposure smoothed + spike glycan interactions.
6ACC trypsin treated down conformation of SARS-CoV-1 spike, binding ACE2.
6ACD trypsin treated up conformation of SARS-CoV-1 spike, binding ACE2.
7BYR RBD binding BD23 Fab antibody, sidechain exposure
7BYR RBD binding BD23 Fab antibody, smoothed + glycan interactions
6XC3 RBD binding antibodies CC12.1 and CR3022
6XC3 RBD binding antibodies CC12.1 and CR3022 smoothed
6LXT activated spike protein smoothed.
6M0J complex with ACE2 domain
6M0J complex with ACE2 domain smoothed/

For brevity, not all are included in every case if their exposure scores are typical and of less direct interest. In reports, the specific chain by investigated is indicated by, for example 6ZP5-B, which means the B chain subunit, except in engineered structures that encompass the features of just one subunit of the trimer.

The method is currently coded in Perl but it is easy for e.g. a Python programmer to code, given that he or she has some familiarity with analyzing protein structures and follows Eqns. (1) and (2). There is an outer loop over all atoms in the system storing consecutively in (Perl) arrays @atomName, @x, @y, @z @number, @residue @chain etc. including only those atoms (e.g. Chain B) that the user wishes to interact with the rest of the system (e.g. glycosylation, chains A and C, ligands etc.). This administrative and essentially straightforward detail controlled by regular expression matches and conditional actions (IF tests) occupies much of the program. The program should loop over the list of atom coordinates etc. in the usual way as follows. Note that there is a special separate count in regard to glycans and glycosylation.

```
for ($j=1;$j<=$atom;$j++)
        {
        next if $number[$i] == $number[$j] && $resdue[$i] eq $residue[$j];
        $distanceSqd = ($x[$i]-$x[$j])**2+($y[$i]-$y[$j])**2+($z[$i]-$z[$j])**2;
        if ($distanceSqd >= $minRangeSqd && $distanceSqd <= $maxRangeSqd)



                {
                $scoreRes[$number[$i]]++ ;
                $glycoScore[$number[$i]]++ if $residue[$j] =~ m/$glycotypes/;
                }
    }
```

Displays and printouts resemble the formats used in alignments, secondary structure prediction, and so forth, the linear sequence typically showing 50 or 60 residues per line [21]. 60 is a good choice in the present study as the SARS-CoV-2 sequence can be broken up into descriptions that are in reasonable accord with subdomains and regions of interest. In the present study there are minimally three linear strings of characters displayed per line, for example as follows, although exposure analysis relating to the same section of sequence also usually appear for several structure determinations with, frequently, different ligands.

```
HTPINLVRDLPQGFSALEPLVDLPIGInItRFQTLLALHRSYLTPGDSSSGWTAGAAAYY  266
75859799635873430537653343453421123358X~~~~~~~~~~~~~~~~X7742 7BYR-C BD23 Fab
566777777665544554444444443%%%333455689~~~~~~~~~~~~~~~~98765 7BYR-C BD23 Fab sm.
```

Note the following.

(1) The amino acid residue sequence is expressed in standard one-letter code. The number to the right is the count of residues up to the last residue on the line based primarily on the residue numbering of the Wuhan seafood market pneumonia virus isolate GenBank entry MN908947.3 aligned with 6ZP5, but some structures including the reference structure 6ZP5 miss the first part of the sequence expressed by the gene (27 residues in the 6ZP5 case), and it was convenient that these were also omitted here to enable a universal 60 residue per line layout which reasonably captures regions of interest as described above. What this means in practice is that the number system used here starts at residue 27, and misses the signal peptide MFVFLVLLPLVSSQCV 1–16 (IUPAC one letter amino acid code) and a short section NLtTRTQLPP 17–26 sometimes referred to as the signal peptide hinge.

(2) Beneath each single residue character there is a character representing exposure score i.e. a single character 0–9, or X where appropriate. In addition, a character '∼' may often be seen, indicating that the residue is probably significantly exposed, but more precisely that it is disordered in the experimental structure, i.e. it cannot be "seen" because exact coordinates are not available.

(3) To facilitate synthetic peptide vaccine and peptidomimetic design, a second sequence of residue-by-residue characters representing a smoothed exposure score is also displayed in alignment with the above. This describes the average exposure score for the residues from i-m to i+5 central residue at locus i expressed as belonging to residue i. The score is still expressed as in (2) as a single character 0–9, or X where appropriate. One of

these characters may, however, be replaced by the character '%' indicating a degree of shielding by glycosylation (See Theory Section 2). Note that not all the experimental structures used are glycosylated. Such cases will be apparent from the display.

Some other characters serve specific helpful functions. The following is a summary and explanation of characters seen in displays.

0–**9** Numeric characters 0 through to 9 express the exposure score of Eqn. (2).
X Highly exposed (exposure score 10)
~ Disordered loop
. Residue not seen at chain ends in structure determination, sometimes because of shortened sequence but often disordered loop.
% residue obscured by glcyoslation
n s Putative glycosylated asparagine (N[^P]S[^P] rule)
n t Putatative glycosylated asparagine (N[^P]T[^P] rule)
$ putative noncovalent sialic acid binding region
@ antibody binding residues
# ACE2 receptor binding residues

Sidechains and portions of protein backbone adopting more than one conformation would be coded the same way as disordered loop, using the character '~'. Such cases of e.g. two or more conformers of one sidechain were not seen explicitly in the experimental three-dimensional structures used as data, except that runs of consecutive residues were considered by the authors of the PDB publication as conformationally disordered. As is usually the case, they involve missing atoms in the PDB entry that the program developed here automatically detected. In principle, the algorithm can certainly be used to average over interactions involving two or more conformers in experimental structures because one version of the algorithm being developed averages over selected intervals in the histories of motion in molecular dynamics simulations. In contrast, in the present paper, variations in structure *between different experimental structure determinations* (i.e. different PDB entries) are explicitly represented by aligning the sequences and their associated characters describing the status of the residues. The meaning of any further minor aspects will be self-evident in the examples given in Section 4 below.

Note that likely N-glycoslylation sites are reported, with the regular expression consensus N[^P][S,T][^P]. O-glycosylation is less predictable and, in particular, any serine S and threonine T is suspect. However, total protection by glycosylation usually only extends by a surprisingly few residues, and it is the accumulative mass provides the shield. At the same time, the same distribution of glycosylation may not be the same per human patient per organ, so it is best report the basic case.

4. Results

4.1. Choice of optimum value of maximum range R_{max}

The distribution of scores depends on the value of R_{max} and some choice will be the most appropriate for the method. Table 1 shows the percentage distribution describing the above for a broad range of values of R_{max} using the SARS-CoV-2 spike glycoprotein structure 6ZP5. This is of some theoretical interest regarding the behavior and nature of the measurement and as relating the radial distribution of atoms with respect to protein sidechains. A radial distribution function in which one seeks peaks with successive shells of surrounding atoms or molecules [24]. In this present study this is influenced by the nature of the measure used (Eqns. (1) and (2)) which counts volumes from the center of each sidechain atom considered, not shells, and the dimensions of the spike which is approximately 200 Å in length. While the results are conceptually closer to the integral of a radial distribution function [24], the form departs from that because of the nature of Eqn. (1), because S_{min} and S_{max}, as well as S_{raw}, will vary with R_{max}.

The distribution of scores down each column reported for the same structure is in general, for any choice of R_{max}, a skewed normal curve exemplified as follows, emphasizing that higher scores (above 3–5 are relatively rare. The range of interest *circa* 5.5–7.5 Å for the maximum separation distance seems reasonable on physical grounds. It represents the peak of the first shell of surrounding atoms characteristic of an organic molecule, ignoring hydrogen atoms. Including hydrogen atoms, a small aromatic molecule such as toluene has approximate average diameter of *circa* 6 Å and can separate two methyl groups of approximately 2 Å radius each, making 5.5–7.5 Å reasonable for considering a 10 Å ideal "hole" if one considers for typical protein surface flexibility allowing a "squeeze between" or displacement effect. In practice, it is the optimal R_{max} for interactions within the protein that may provide shielding., as follows.

4.2. Choice of more specific value of R_{max} based on antibody binding

The final choice of 6.5 Å was also influenced by calibrating R_{max} to highlight effects of exposure change on binding antibody (e.g. see Table 2) and is consistent with the findings reported below in Sections 4.3 and 4.4. However, this approach was less persuasive than expected, for the following reasons. The idea of studying antibody binding is that one expects the number of exposed sidechains to fall at the interface between the spike and an antibody and one wishes to choose a value for R_{max} that emphasizes this. The trend should be that lower scores of exposure fall, and higher scores rise. Overall studies do suggest that, and the best evidence for this is seen in the scores in the range 2–6 for R_{max} in the range 5.5–7.5 Å and particularly 6.5–7.5 Å, and clearest when directly expressed in terms of the counts of the different scores rather than percentages for the different number of scores, as is done in Table 2. However, the overall picture is complicated, and the expected behavior is less marked than has been shown in relation to many other experimental structures.

Examination shows that this is for an interesting reason. In the case of the SARS-CoV-2 spike, a major effect of antibody interactions is to interact with a disordered loop and impose a more ordered structure, as shown later below. This means that the initial disordered loop is not available in the experimental structure interpretations for calculation of the sidechain exposures. This is not a problem conceptually because it is

Table 1
Distribution *by Percentage* of accessibility scores for sidechains in SARS-CoV-2 spike glycoprotein PDB 6ZP5.

R_{max} =	3	4	5	6	6.5	7	8	9	10	15	20	25	50	75	100
0	2	1	2	2	5	4	2	2	5	2	4	4	2	9	**41**
1	0	2	3	7	13	14	11	13	13	7	6	6	4	10	20
2	1	6	7	15	**16**	15	15	14	14	13	12	7	9	13	13
3	2	13	13	**18**	15	**17**	**18**	**16**	15	**20**	**17**	10	12	**15**	11
4	5	17	**21**	17	15	14	15	**16**	**16**	17	14	13	14	14	6
5	18	**24**	18	14	13	13	14	13	11	14	14	15	15	13	5
6	**27**	19	15	12	11	11	11	10	10	11	13	**17**	**16**	10	2
7	17	12	10	8	7	6	8	9	8	9	9	14	13	7	2
8	19	5	7	4	4	4	5	5	5	6	6	8	10	6	1
9	9	1	3	2	2	2	2	2	2	2	4	4	6	4	0
10 (X)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2
Distribution of accessibility scores for sidechains. In SARS-CoV-2 spike glycoprotein.

Number of scores of this rank	maximum separation of atoms = 5.5 Å (between centers) in spike (6ZP5)	maximum separation of atoms = 5.5 Å (between centers) in spike + BD23 antibody 7BYR)	maximum separation of atoms = 6.5 Å (between centers) in spike (6ZP5)	maximum separation of atoms = 6.5 Å (between centers) in spike + BD23 antibody (7BYR)	maximum separation of atoms = 7.5 Å (between centers) in spike (6ZP5)	maximum separation of atoms = 7.5 Å (between centers) in spike + BD23 antibody (7BYR)
0	20	13	45	56	23	39
1	43	21	123	117	120	122
2	110	85	152	173	125	184
3	158	164	146	173	178	174
4	181	219	137	141	132	121
5	145	172	121	123	136	138
6	126	163	106	111	99	118
7	86	93	62	58	73	58
8	49	55	36	35	44	32
9	25	13	15	12	13	12
10 (X)	0	0	0	0	0	0

reasonable to assume that there is a high degree of exposure of sidechains in such a loop, especially for practical purposes of considering antibody binding (since the residues can readily be moved in space), and so the exposure is in effect decreased. Indeed, for most purposes an implied score of 10 may not be unreasonable for residues in any loop that is too disordered to appear in a structural determination, even if it is a dynamic random coil [20]. The fact that the precise exposure values cannot, however, be seen for the disordered region, the choice of 6.5 was also further validated as in Section 4.3 and 4.4 and, importantly, by examining the more detailed effects of score changes in regions that were ordered even without binding of large structures, as will be seen in displays of sequence details later below.

Although no score of 10 was obtained in these structures, it does appear in some structures for SARS-CoV-2 spike proteins as shown later below.

4.3. Comparison with other methods

Before developing the above algorithm, several methods were tried for calculating exposure of sidechains [20–22], particularly two that may be considered as corresponding to two extremes of a continuum of methods. At one end is the "rolling ball" approach. This is exemplified by the method of Lee and Richards [23], which is almost certainly the traditional and most widely used algorithm for determination of solvent accessible surface of a protein. There are many implementations, e.g. http://legacy.ccp4.ac.uk/html/areaimol.html. Here, solvent accessible surface is defined as the locus of the center of a probe sphere (representing a solvent molecule) as it rolls over the van der Waals surface of the protein. At the other end are geometric approaches based only on the coordinates of the protein atoms. Arguably, the example which the most purely geometric is an early method of the present author in which a vector is calculated from the centroid C of the protein to the centroid I of each ith sidechain. Vectors are then calculated from I to the centroids of each jth sidechain, and the average cosine between vector C–I and I-J over all J is calculated. See Fig. 13.3 of Ref [20]. Recall that the algorithm in the present paper was developed subsequently to facilitate certain special requirements (see Section 2.1). Tests on *circa* 30 arbitrarily selected protein structures per method gave similar results to the approach used in the present paper after conversion of the results to the same desired representation. Note that exposed residues in the trimer and those contacting between the trimer can be listed in response to a PDB entry code such as 6ZP5 at the EBI site https://www.ebi.ac.uk/msd-srv/prot_int/cgi-bin/piserver. However, there were important differences between the methods. As opposed to methods that address the outer surface, the present method includes as solvent-accessible the linings of cavities which could accommodate a smaller potentially therapeutic molecule. In principle, access to these could take place during folding, or because of fluctuations or conformational changes in the life of the spike. Distinct and sizable cavities are rare in globular proteins (most are well packed), but cavity-like volumes do exist in and around the regions of interface between the three monomers. There are *circa* 900 residues regarded as accessible in the spikes of the viruses studied in the present paper, and some 140 interact between trimers. A third to a half of these interactions are in the following which are the longest interfacing sections in entry 6ZP5 B chain in which most residues are in contact between monomers.

(a) GVSPTKLNDLCFTNVY (277-292),
(b) KKFLPFQQFGRDIALT (408-423)
(c) YTMSLGAENSVAYSNNSIAIP (517-536)
(d) QALNTLVKQLSSNFGA (754-769)

There are 13 residues out of 16 in (a) are in contact between monomers with low exposure scores of 0–5 in the method used in the paper (with R_{max} = 6.5 Å), of which 4 resides are tight contact with scores or 0,1,2 and one with a marginal score of 3. These numbers are typical of the similar long sections for monomer interactions in all the spike protein structures examined. Importantly, however, the rest of the residues interacting between the monomers are distributed among numerous much shorter sections of between 1 and 5 residues in length. Many of these shorter regions can be considered as separated and immediately adjacent to accessible residues by Lee-Richards, vector, and the present method, i.e. still accessible to a small therapeutic molecule between the monomers.

See Section 4.6 for examples of details regarding the changes that occur in the exposures of sidechain in the spike protein when binding various molecules and groups. It is pleasing for validation of the method that these are almost entirely in directions that might be expected: the finding is that the exposure score associated with many sidechains at the binding interface decreases substantially. However, it is the differences from other methods that is of greater interest. A key feature of the present method is that it indicates the extent that sidechains are exposed, not the exposure of the backbone. Knowing the subsequences in which all or most sidechains are exposed at the protein surface is important for design of peptidomimetics for two reasons.

(i) It is the sidechains that are responsible for molecular recognition, and sidechain by sidechain detail can be important. Despite the above comment on agreement with expectation, not all exposed sidechains make a tight interaction with another protein within what may be considered a binding region. The designer may have a freer hand in making advantageous changes to these. Relevant for small molecule therapeutic design is that, of the residues that could be considered as the binding region of 6XC3 CC12.1 +

Table 3
Comparison of exposure measures by other methods.

	Standard view	Phys 1	Phys 2	Phys 3	Exposed $R_{max} = 6.5$ Å	Exposed $R_{max} = 10.0$ Å	Exposed $R_{max} = 4.5$ Å	Not in contact with hydro-phobic core	(Moel-bert et al.	Semanta et al. [22] A	(Semanta et al. [22]) B
LYS	+/−	+1.8	−1.5	−3.0	5.9	5.6	7.3	+2.0	0.61	9.8	85
GLU	+/−	+1.6	−0.74	−2.5	5.4	5.5	6.0	+1.9	0.59	10.6	64
ARG	+/−	+1.0	−2.5	−3.0	5.1	4.6	6.5	+0.9	0.54	16.3	77
ASP	+/−	+2.4	−0.1	−2.5	5.1	5.3	5.1	+2.1	0.61	11.0	48
ASN	P	+0.4	−0.8	0	5.1	5.7	5.4	−0.3	0.57	12.0	48
HIS	+/− P	+1.4	−0.4	0	4.5	4.5	5.8	−0.05	0.42	15.2	47
GLN	P	+0.2	−0.8	0	4.4	4.3	5.8	−0.3	0.57	12.8	60
PRO	H	−0.3	+0.12	+1.4	4.2	4.4	4.9	−1.8	0.50	10.6	43
THR	P	+0.1	0	+0.4	4.0	4.4	4.3	−0.5	0.48	11.3	36
SER	P	+0.3	−0.2	−0.3	3.8	4.5	4.1	−0.2	0.57	9.3	28
TYR	P,H	+0.2	0	+2.3	3.6	3.2	5.1	−1.2	0.32	20.6	40
TRP	H	−0.2	+0.8	+3.4	3.2	3.0	4.5	−1.4	0.28	24.8	38
GLY	–	+1.1	+0.5	0	3.2	3.5	4.2	0	0.59	–	–
LEU	H	−0.7	+1.5	+1.8	3.2	3.0	5.6	−1.2	0.32	14.8	23
VAL	H	−0.5	+1.8	+1.5	2.7	3.1	4.2	−1.0	0.31	12.8	19
ILE	H	−0.8	+1.4	+1.8	2.7	2.6	4.5	−1.25	0.27	15.3	21
PHE	H	−0.6	+1.2	+2.5	2.7	2.3	5.0	−1.3	0.29	20.0	25
ALA	H	+0.3	+0.6	+0.5	2.6	3.3	3.8	−0.6	0.40	7.2	18
MET	H	−0.4	+0.6	+1.3	2.3	1.6	5.5	−1.0	0.37	16.5	29
CYS	H	+0.2	+0.3	+1.0	2.1	2.7	3.5	−0.5	0.27	13.4	10

CR3022, some 17 residues have an exposure score of 4 or higher in the present method within the subsequence regard as closely involved in antibody binding. More directly relevant still is that some 41 have an exposure score of 4 or higher in the present method in the subsequence considered as involved in the ACE2 virus receptor binding (PDB entry 6MOJ), and 27 of these have scores of 6 or higher. Binding in these cases is evidently cemented primarily by interacting residues with exposure scores of 0, 1, or 2, there being 24 such residues in the case of ACE2 binding. Other methods that are widely available and well known do not give these details appropriate to peptidomimetic design.

(ii) Retro-inverso peptidomimetics have the defect that backbone amide and carbonyl groups are interchanged [6]. If the interactions of the spike protein with a human protein target primarily involve sidechains, this problem for the *retro*-inverso peptide as an antagonist is minimized and the *retro*-inverso approach arguably could be a "near perfect" method of making an analogue composed of D-amino acid residues (subject, of course, to the results of synthesis and testing). See Ref. [6] and Discussion Section 5 in the present paper.

Much software is less well suited to the present purpose. The vector approach [20] clearly emphasizes loops of which stick out well from the surface, and like the "rolling ball" approach as normally implemented does not make the required distinction between sidechain and backbone. This is also the case for what is probably the most widely used program for looking at surfaces and interactions, routinely available as a viewer at the Protein Data Bank site https://www.rcsb.org/3d-view/6ZP5/1. Selecting the viewer NGL (WebGL), and then selecting Style and then Surface, produces a surface view of the spike protein. Using such methods to (a) understand sidechain interactions and (b) detect differences over several structures and related but different proteins are laborious tasks for present purposes. This is primarily because they provide a perspective that is somewhere between a two and three dimensional one that does not lend itself to peptidomimetic design which is sequence based. One can of course compute energies of interaction and consider the contributions of specific sidechains using the techniques reviewed in Ref. [7]. They include some, developed by the present author in collaboration with others, that are particularly appropriate. Nonetheless, such approaches are algorithms for drug design; they are much less applicable when the ligand has yet to be designed and they are not directly appropriate to the first step of recognizing parts of proteins that may serve as targets.

The difference between any Lee-Richards kind of approach and that of the present paper is noticeable in the shorter sections that form interfaces between the monomers. For example, GLTGT (396–400) that the Lee-Richard approach considers as not exposed has intermediate exposure by the present method with scores 35546 respectively, which is moderately exposed in the present approach. Typically, they have an average exposure score of 3 by the present method but are often bordered by regions that are buried by the current score but more exposed by Lee-Richards. For example, SFGG (442–445) has an average exposure score of 3 followed by a run of scores of 2, indicating less exposure. In several cases it is noticeable that the origin of the differences is because the present method not only focuses on the sidechain rather than the residue as a whole, but also treats the sidechain more precisely as the outer part of the sidechain which carries molecular recognition, from the Cγ carbon outward ward, with the exception of glycine and alanine that use Cα and Cβ atoms respectively. In sequence (b) KKFLPFQQFGRDIALT above, KKF has exposure scores of 561 by the present method, indicating that the lysine residues K are moderately exposed at the ends but that the aromatic ring of phenylalanine F is more buried. The glycine G and the arginine R have exposure scores in the present approach of 0, 4 respectively, which emphasizes that the glycine is well buried but the distal end of the arginine is moderately exposed.

Some implementations of available methods do focus on sidechains, or can be more easily adapted or repurposed to do so, and they give a variety of different results with different purposes and merits. One way to discuss and

summarize these is to compare the average scores for the twenty naturally occurring different sidechains. It is important to keep in mind that exposures are calculated for each amino acid *in situ* not on notions of hydrophilic and hydrophobic character so that, for example, tyrosine would have a different value at different occurrences of it in the protein. However, it would be expected that the core values per amino acid reside sidechain types based on any algorithm should (a) essentially follow the findings of other methods for average sidechain exposure, and (b) that this should at least approximately follow the polar/nonpolar hydrophilic/hydrophobic character of residue sidechains [22], and this is so in the present case. Table 3 Column 6 shows the ranking of average scores per amino acid type at an R_{max} of 6.5 Å for the B chain of the SARS-CoV-2 spike protein including glycosylation (PDB entry 6ZP5), including interaction with chains A and C and all covalently bound glycans. The other columns are compared with this ranking. Column 7 and 8 show the scores at a large separation of $R_{max} =$ 10 Å, and R_{max} of 4.5 Å which is a close contact between sidechains. Although these measures were a relevant factor in selecting an R_{max} of 6.5 Å as optimal choice, the results are not particularly sensitive to changes in R_{max}. They are in reasonable accord with Column 2 that indicates the standard view of the sidechain as likely to be highly exposed because it is charged (+/-), or possibly exposed because it is polar (P) and can form hydrogen bond to solvent, or hydrophobic (H), i.e. non-polar. Alanine A and glycine G are noticeably in some disagreement between the standard view and the present method which seeks to focus on the molecular recognition aspect. Columns 3–5 headed Phys 1, Phys 2, and Phys 3 reflecting discussions regard physicochemical properties in Ref. [20], but are effectively respectively equivalent to the values in the Wikipedia entry for hydrophobicity scales, values originally due to Tanford, and also due to Levitt. See Ref. [20] for discussions, and notably those associated with Fig. 8.5 in that text. Column 9 headed "not in contact with hydrophobic core" is a somewhat different approach in which a hydrophobic core of residues is identified and then the remaining residues are considered as likely to be in the region close to the surface (Table 12.1B of Ref [20]). See also Refs [21, 22] for similar recent reviews and comparisons by other authors. Column 10 relates to the study by Moelbert et al. [21] using a Lee-Richards approach with a probes sphere of 1.4 Å. Columns [11] relates to a study by Semanta et al. [22]. using two methods. Method A uses the notion of the number of atoms in contact with a sidechain atom, which is probably the approach most similar to the present algorithm, except that they focused on close atomic contacts analogous of up to 4.5 Å between sidechain atoms. It is for comparison that Column 8 compares results for the present author's current method at 4.5 Å. Semanta et al. also compared the Lee-Richards approach (last column).

Having compared other approaches, a major consideration motivated the adoption of the current algorithm. The approach is in good accord with polar and non-polar properties of amino acid residues and not least expectations based on chemical structure, while at the same time being remarkably insensitive to choice of value for R_{max} (due to the nature of Eqns (1) and (2) However, this insensitivity naturally has its limits. As R_{max} is dramatically increased, the above converge to the range of 4–6 Å and there is even an inversion of the size of score measures of polar and nonpolar residues. For example, at 50 Å, LYS takes the value 5.2 comparable to its value above, but TRP rises to 6.8 and SER and THR take values of 3.8 and 4.0 respectively, while most nonpolar residues take a value of around 5 Å. It is noteworthy, however, that for all physically meaningful ranges, the score for tryptophan is somewhat higher than expected on the basic of size and hydrophobic character. However, it is commonly associated with surface loops in many proteins [21], and it appears particularly high in SARS-CoV-2 because tryptophan residues are not in general abundant and many are likely to be involved in (non-covalent) binding host cell sialic acid glycans [8], in a manner similar to influenza hemagglutinin.

4.4. Descriptive classification of exposures

At $R_{max} = 6.5$ Å the average score per residue sidechain, i.e. over all residue sidechains, is 3.8. A score analogous to that for a sidechain but corresponding to the centroid of the spike protein was 1.4. Taking account of this as well as all of the considerations discussed above (Sections 4.1-4.3), it continues to appear reasonable to use a maximum separation R_{max} of 6.5 and also to use the following descriptions or classifications based on the exposure score.

0-2 buried, in the protein interior.
3-5 partially exposed, typical of a reasonably flat protein surface.
6-8 well exposed, "elevated above" the surrounding sidechains.
9-X highly exposed, protruding well above the surrounding sidechains.

A score of 4 or more may be considered worthy of examination as reasonably, or possibly readily, exposed.

4.5. Exposure analysis of the S1 head region

The following illustrates the basic features of the analysis and its format as introduced more generally in Methods Section 3. This layout is conserved in the subsequent Sections, so that one may "walk though" an account of the SARS-CoV-2 spike glycoprotein sequence. For cross reference and to avoid any ambiguity arising in different layouts in publication, the output for sections of sequence and sections of sequence of different spike proteins are called BLOCKs and are numbered. For brevity in this paper, not every different structure is compared in a block. The numbering 18–331, 27–332 below refers to two common descriptions of what is commonly considered as the S1 N-terminal domain (NTD). See BLOCK 1.

```
BLOCK 1.
S1 N-TERMINAL DOMAIN (NTD) 18-331, 27-332
(...
AYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSnVtWFHAIHVSGTnGtKRFDNPVLPF  86
.856153332134763243555664854022265624447~~~~~~~~~~~~~~687351 6ZP5-A closed
.544444343434444455555555544444443345678 9~~~~~~~~~~~~~~~988765 6ZP5-A closed sm.
763515333212374323354565364512236451334 9~~~~~~~~~~~~~~~87351 6ZP5-B closed
5%%444333333333344444444444444433%%%%554~~~~~~~~~~~~~~~~55555 6ZP5-B closed sm.
884614434322467324374466373412336472445 8~~~~~~~~~~~~~~~88261 6ZP5-C closed sm.
555554434444444444544454444444444456789~~~~~~~~~~~~~~~~98765 6ZP5-C closed sm.
.856153332234763353655664854022265624447~~~~~~~~~~~~~~~687351 6ZP7-A open
.544444344444444555555555444444433456789~~~~~~~~~~~~~~~988765 6ZP7-A open sm.
763515333212374323354565364512236451334 9~~~~~~~~~~~~~~~87351 6ZP7-B open
5%%444333333333344444444444444433%%%%554~~~~~~~~~~~~~~~~55555 6ZP7-B open sm.
884614334222368141364466273402336472444 8~~~~~~~~~~~~~~~88261 6ZP7-C open
555544443334444445544444444444433345678 9~~~~~~~~~~~~~~~98765 6ZP7-C open sm.
682605323213356324254454564113353723557~~~~~~~~~~~~6X766451 7BYR-A BD23 Fab
876654333333333344444454444444433345678 8~~~~~~~~~~~~~99876554 7BYR-A BD23 Fab sm.
773626343232486313253463465412426262354 7~~~~~~~~~~~~~~986362 7BYR-B BD23 Fab
987654434444434444444444444444433345678 8~~~~~~~~~~~~~~988765 7BYR-B BD23 Fab sm.
784514333233564235364465476412337462454 6~~~~~~~~~~~~~~987746 7BYR-C BD23 Fab
987654434333334444444555544444444456788~~~~~~~~~~~~~~X98775 7BYR-C BD23 Fab sm.
                        $$$$$$$$$$$$$ putative noncovalent host glycan binding
```

Note the appearance of shielding by spike protein glycans, indicated by '%'. The above block of sequence of interest was predicted as a potential site for binding host sialic acid glycans [8] as indicated by the characters '$' that applies to the N-terminal side of the tryptophan W which is a signature of such a potential binding. Note that the asparagine n or threonine t appears glycosylated in several related structures examined and may impede host sialic acid binding to the protein side-chains, but it was it was not considered an impediment, There can be direct interactions between virus and host sialic acid glycans. The above disordered loop indicated by ~~~~~~~~~~~~~~~~~ is similar in all chains A, B, and C of SARS-CoV-2 structures examined here, but it is more ordered in 6ACC SARS-CoV-1. Compare the following block which is essentially similar in all the structures examined. It is not involved in any of the interactions (ACE2, FAB etc.) studied here, but it is of sufficient size to be of future interest. Associated with the tryptophan W there is some weaker potential for host sialic acid glycan binding by the criteria of Ref [8], but it is below the 100 threshold.

It is interesting to align and compare with the above the PDB entry 6ACC (SARS-CoV-1 trypsin-cleaved and low pH-treated SARS-CoV-1 spike glycoprotein and ACE2 complex, ACE2-free conformation with three RBD in down conformation). The FASTA sequence file aligned with the SARS-COV-2 sequences of 6ZP5 etc. Indicates a deletion in the place of the disordered loop, but the region and its surrounds are not reported as disordered in the PDB entry. See BLOCK 2.

experimental structure data seems to suggest that there are deletions in the sequence that are really disordered structures invisible to the structure determination, and certainly a new computer program analyzing the PDF file could make that mistake, so it is useful to note the following. In addition to extensive missing sections at the N- and C-termini, the following appear to be the regions not available in the SARS-COV-1 spike protein 6ACC structure but present in the FASTA description of the sequence, confirmed as non-deletions by searches with BLASTp via https://blast.ncbi.nlm.nih.gov/Blast.cgi.

PAQD
HTVSLLRSTSQKS (verified by BLASTp, 100% matches over many SARS strains)
DAGFMKQYGECLGDINARDL (verified by BLASTp, 100% matches over many SARS strains)
YDPLQPELDSF (verified by BLASTp, 100% matches over many SARS strains)

The situation is similar, i.e. it is not a deletion, for most closely related SARS-COV-1 sequences such as that of the spike protein of SARS coronavirus from Frankfurt, GebnBank AAP33697.1, and SARS-like bat and civet strains, e.g. GenBank AGZ48828.1 and AAU04664.1 respectively. None of the above appear to be significantly related to consensus sequences (i.e. potential motifs with particular functions) when submitted to e.g. Prosite (https://prosite.expasy.org/cgi-bin/prosite/prosite_search_full.pl).

The following blocks of sequence are of less interest in the current analysis, but it is noteworthy that a disordered loop is partly more ordered in the complex with Fab despite the apparent absence of direct spike-Fab interactions. See BLOCK 3 and 4.

```
BLOCK 2.
AYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSnVtWFHAIHVSGTnGtKRFDNPVLPF 6ZP5 (SARS-COV-2)
tQHTSSMRGVYYPDEIFRSDTLYLTQDLFLPFYSnVtGFHTInHt-------FGNPVIPF 6ACC-B (SARS-COV-1)
5646246233234644233735654955124382824045356935-------43467451
6655444444443344444455555445454433444444444455-------55555544
```

```
BLOCK 3.
NDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNnAtNVVIKVCEFQFCNDPFLGVYYH  146
6820210134878637133102538895622322646644333l3~~~~~~~~~~~~~~~
433333334444555444334%4555%555544%4%%%%%%5555~~~~~~~~~~~~~~~
6821220254788743120103424463512221326521123112164598656~~~~~ 7BYR-B BD23 Fab
443334444444544433322333333333333333%%%%3%2%223445567899~~~~~ 7BYR-B BD23 Fab sm.

BLOCK 4.
KNNKSWMESEFRVYSSANnCtFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSK  206
~~~~~~~~~~X76~~~976555646~~~~~~~~~~~~~~~6151200226944222124 6ZP5-B closed
~~~~~~~~~~677~~~%%%6655%4~~~~~~~~~~~~~~~4433333333333344334 6ZP5-B closed sm.
~~~~~~~~~~X76~~~976555656~~~~~~~~~~~~~~~6151201326944122124 6ZP7-B open
~~~~~~~~~~667~~~%%%6665%4~~~~~~~~~~~~~~~4433333333333344334 6ZP7-B open sm.
~~~~~~~~~X843545566466454765485~~~~~~~~~~7732220124685423201l 7BYR-A BD23 Fab
~~~~~~~~~987766555%55555%67789~~~~~~~~~~8765434333433344444 7BYR-A BD23 Fab sm.
~~~~~~~~~~~~~~~~~~8466877486~~~~~~~~~~~~7423231234785533227l 7BYR-B BD23 Fab
~~~~~~~~~~~~~~~~~~7666666666~~~~~~~~~~~~5544344444444455444 7BYR-B BD23 Fab sm.
~~~~~~~~~~~~6646546465976566~~~~~~~~~~~~7631131114673425223l 7BYR-C BD23 Fab
~~~~~~~~~~~~877%6%%66666678899~~~~~~~~~~8765433333334444444l 7BYR-C BD23 Fab sm.
```

Such deletions appear as a common theme in the descriptions later below, i.e. as cases in which a section of sequence conceptually inserted in an ordered region of SARS-COV-1 is represented by a disordered loop in SARS-COV-2. There appear to be occasions in the past in which

The above disordered loops ~~~~~~~~ are ordered in SARS-CoV-1 6ACC, with only a small deletion in the vicinity as shown in the following. They are similar in all chains A, B, and C. See BLOCK 5.

```
BLOCK 5.
KNNKSWMESEFRVYSSA--NnCtFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSK 6ZP5  (SARS-COV-2)
MNNKsQSVIIINNstNVVIRACNFELCDNPFFAVSKPMGTQTHTMIFDNAFNCt--FEYISD 6AAC-B (SARS-COV-1)
3558141223335673333611227448355320488938488745659778 56--577749
4334433333434444443333444444444555555666666666766666 66--666666
```

The following blocks are also of slightly less interest in the present study but note the '$' characters indicating suspected binding to host cell glycan binding, notably sialic acid glycan [8] (possibly heparin may bind in this region). See BLOCK 6.

```
BLOCK 6.
HTPINLVRDLPQGFSALEPLVDLPIGInItRFQTLLALHRSYLTPGDSSSGWTAGAAAYY  266
737397774348635104265432363614615246 53~~~~~~~~~~~~~~~~~~~~~1 6ZP5-B closed
455566666654444443333344444%%%44433333~~~~~~~~~~~~~~~~~~~~~3 6ZP5-B closed sm.
737297764348525104255432364614515246 43~~~~~~~~~~~~~~~~~~~~~1 6ZP5-B open
45555666565444444333334444%%%44433333~~~~~~~~~~~~~~~~~~~~~3 6ZP5-B open sm.
748597586358733104366534365634624325 3799~~~~~~~~~~~~~~~XX7332 7BYR-A BD23 Fab
55566666665444444444445455 4%%%4444566789~~~~~~~~~~~~~~987655 7BYR-A BD23 Fab sm.
658~~~~~~997453042564454449244242388~~~~~~~~~~~~~~~~~~~~~434 7BYR-B BD23 Fab
678~~~~~~98766554444455554444566788~~~~~~~~~~~~~~~~~~~~~~765 7BYR-B BD23 Fab sm.
75859799635873430537653343453421123358X~~~~~~~~~~~~~~~~X7742 7BYR-C BD23 Fab
566777777665544554444444443%%%333455689~~~~~~~~~~~~~~~~98765 7BYR-C BD23 Fab sm.
                                                  $$$$$$$$$$
```

Also worthy of note is that in the corresponding region of the SARS-COV-1 spike glycoprotein structure 6ACC there is again a case in which the disordered loop region is instead a deletion, and the region is not disordered. See BLOCK 7.

```
BLOCK 7.
HTPINLVRDLPQGFSALEPLVDLPIGInItRFQTLLALHRSYLTPGDSSSGWTAGAAAYY 6ZP5 (SARS-COV-2)
YQPIDVVRDLPSGFNTLKPIFKLPLGInItNFRAIL----------TAFSIWGTSAAAYF 6AAC-B (SARS-COV-1)
65747599746857441736764445463441423 4----------2137X333494233
45666666666665555554445555444443333 4----------56788876543434
```

In SARS-COV-2 cases above the disordered loop ~~~~~~~ is again essentially similar in all chains A,B, but C, but this consistency does not hold throughout the sequence in this section for all SARS-COV-2 structures examined: in the following block there appears a disorganized loop in 7 BYR chain A which is not present in the other structures. See BLOCK 8.

```
BLOCK 8.
VGYLQPRTFLLKYNEnGtITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESI 326
124166551124137414145221255673231314474362862345263365846761 6ZP5-B closed
34333333343333%%%%%333334443333434434444455544444444555555444 6ZP5-B closed sm.
11406555111413741414512125567323131447436286234526426584676 1 6ZP7-B open
33333333333333%%%%%333333433333434333344455544444444555554444 6ZP7-B open sm.
135166350124~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 7BYR-A BD23 Fab
443333333333~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ 7BYR-A BD23 Fab sm.
023156240023134524145221244564131304564372763446255254747751 7BYR-B BD23 Fab
44332223223322%%%333333334333333333334444555455454445555444444 7BYR-B BD23 Fab sm.
124167541124236621241675411242366252562333566642423255654728 7BYR-C BD23 Fab
4443333334433%3%444433333344 33%3%4%444444444444444444444555555 7BYR-C BD23 Fab sm.
```

In this case, there is no deletion in the corresponding region of the SARS-COV-1 6ACC structure, which does not seem surprising as the disordered loop appears peculiar to the 7BYR A chain. See BLOCK 9.

```
BLOCK 9.
VGYLQPRTFLLKYNEnGtITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESI 6ZP5 (SARS-COV-2)
VGYLKPTTFMLKYDENGTITDAVDCSQNPLAELKCSVKSFEIDKGIYQTSNFRVVPSGDV 6ACC-B (SARS-COV-1)
22528834112522872524633114443414241447458287346646412652657 2
44444444443434444444333333333333444555556555554445555444445
```

4.6. Exposure analysis of the S1 receptor binding domain (RBD) region

This region is of interest in exhibiting antibody binding in 6XC3. A loop is not disordered prior to binding, and further to comments made in Section 4.2, the effect on sidechain exposure of the binding of antibody elements is less dramatic than might be expected, but the scores in bold certainly indicate a diminution. Recall that 0–2 was considered buried in Section 4.4 on the basis of considerations in Sections 4.1-4.3. Not all sidechains would, of course, be exposed prior to binding. See BLOCK 10.

Here there is no such evidence of disorder in ATSTGN corresponding to KGVGN in SARS-COV-2 structures, a subsequence that finishes at the beginning of the next the next block. In the next block, the fall in exposure scores is less pronounced, and a major factor is the binding to disordered loops and producing order in them. Not all glycosylation groups show up in all molecular viewers but the present software finds them in the PDF file. For example, the glycosylating NAG A1322 shielding lysine K 347 in NL**K**PFERDIST in 6 ZP-B does not show in all

```
BLOCK 10.
S1 RECEPTOR BINDING DOMAIN (RBD) 333-526,331-527
      (...
VRFPNItNLCPFGEVFnAtRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTK  386
32135488837456436388444142154753482225446357576355361752136 5 6ZP5-B closed
4444%%555556%%55%%5444444444444444444444444555%5%%554444444443 6ZP5-B closed sm.
32135488737446336388434132154753482225446357576355362752136 5 6ZP7-B open
4444%%555555%%5%%4444444434444444444444555%5%%555444444443 6ZP7-B open sm.
53~~~~~~648534336698463175557775364636447458588546452374536 5 7BYR-A BD23 Fab
78~~~~~~987%%%5%%555555555555555555555555%%66666%555544544455 7BYR-A BD23 Fab sm.
31146776937446335386242033062673481225256356477256362452145 4 7BYR-B BD23 Fab
4444%555655%%%5%%4444333%3334444444444444%%55%55%554444443343 7BYR-B BD23 Fab sm.
42349887736545446588442235255743593335466546688456563563367 5 7BYR-C BD23 Fab
55555666666%%%5%%554544444445555555555555%66666%6665555555555 7BYR-C BD23 Fab sm.
......X98254473364981220321646723825264464665873553735766468 6M0J RBD+ACE2
......99876%%%5%%5444344333344445555455555%%55%5655555555655566 6M0J RBD+ACE2
.......X836459336497133043175782393526455445698254230323334 6 6XC3 RBD+CC12.1+CR3022
.......X987%%%5%%5444444444445555555554554%%55%%%44443333334 44 6XC3 sm.
binding loop                                @@@@@@@@@@@@@@@@@@@@ 6XC3 CC12.1+CR3022
```

Similar comments apply to the next important block regarding both antibody and ACE2 binding. Note that ACE2 binding is to and ordered loop that is disordered in most structures. See BLOCK 11.

```
BLOCK 11.
LNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVG  446
22522311031021438664363345567472146224237753112114847766~~~~ 6ZP5-B closed
33332222122233334444554545544444443444443333343344%456667~~~~ 6ZP5-B closed sm.
22522312031021438664363335567472146123237753212114847766~~~~ 6ZP7-B open
33332222122233334444554545544444443333443333343344%456667~~~~ 6ZP7-B open sm.
23675313141121212747434576766323824557777321111250267 65~~  7BYR-A BD23 Fab
2333456789999 9X9877777776665433222334444445555555444433~~  7BYR-A BD23 Fab sm.
~~~~226321040200100130623613335575610221232168152121023137 72 7BYR-B BD23 Fab
~~~~33332222111111112223333444433333332222333333332223333444 7BYR-B BD23 Sm.
25743313142121143266574545678765237232346772133325416866 4~~~ 7BYR-C BD23 Fab
44444333222222333444555666655555544444444444443344456789~~~ 7BYR-C BD23 Fab sm.
23766413140011014062383445686520320132289262122133027621687 6M0J RBD+ACE2
55544333321112222333344555544443332333433443333233334444554 6M0J RBD+ACE2 sm.
                                                  ######### ACE2 binding loops
35766323140101011052482333655242000123228925112113302872 17X~ 6XC3 RBD+CC12.1+CR3022
54444433221111112223334444433332221223333333333333456778~ 6XC3 sm.
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@ binding loop   6XC3 CC12.1+CR3022
binding
```

Compare the corresponding region in the SARS-COV-1 structure 6ACC. See BLOCK 12.

```
BLOCK 12.
LNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVG 6ZP5-B (SARS-COV-2)
LNDLCFSNVYADSFVVKGDDVRQIAPGQTGVIADYNYKLPDDFMGCVLAWNTRNIDATST 6ACC-B (SARS-COV-1)
35642315221122124164494445696686235334338993222212327875147 9
44444332222222223344455556655555544445554444443333444455566 6
```

viewers but shows up well as in contact with the lysine in the space-filling mode of NGL (WebGL) viewer. See BLOCK 13.

```
BLOCK 13.
GNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQ  506
~~X95565~~~~~63854653788~~~~~~~~~~~~~~~~~~~85555~~~~~~~~~~~~ 6ZP5-B closed
~~676666~~~~~65%56%66666~~~~~~~~~~~~~~~~~~65554~~~~~~~~~~~~ 6ZP5-B closed sm.
~~X95465~~~~~63854537888~~~~~~~~~~~~~~~~~~~85555~~~~~~~~~~~~ 6ZP7-B open
~~666666~~~~~65%55%66666~~~~~~~~~~~~~~~~~~~55544~~~~~~~~~~~~ 6ZP7-B open sm.
~985553~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~X856672479~~8573 7BYR-A BD23 Fab
~332222~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~2223334445555555 7BYR-A BD23 Fab sm.
203723314337571643135456624334888385723661321112221146258463 7BYR-B BD23 Fab
34433334435%%%%%44%4%%4%4555555655665544332225%12222334444443 7BYR-B BD23 Fab sm.
~~595653~~~~~838654847999~~~~~~~~~~~~~~~9X9866476779~~~~9893 7BYR-C BD23 Fab
~~999999998777%8899XX9988~~~~~~~~~~~~~~~~XX9887778899~~~~9876 7BYR-C BD23 Fab sm.
@@@@@@@@@@                          @@@@@@@@@@@@@@@@          BD23-B binding loops
225514111248761854555767924524878487867632251133120145027442 6M0J RBD+ACE2
444333344344455655656555556565666776655444332222223333333332 6M0J RBD+ACE2 sm.
################################################################## ACE2 binding loops
~58252123477 41X64665767932512798498887773236125514145813842 3 6XC3 RBD+CC12.1+CR3022
~766544434445556556666555555556667776666554443334444434332 6XC3 sm.
      @@@@@                  @@@@@@@@@@@@@@@@@@@@@@      CC12.1+CR3022 binding loops
```

One may of course test the method by obtaining structures of proteins with bound molecules and comparing them with results for structures in their absence. Block 13 naturally provides several interesting examples of this. For subsequence NCYFPLQSYGFQ, where the binding domain of ACE2 binds (#) the receptor binding domain RBD of 6M0J, the residues have high exposure scores or are disordered (indicated by ~) in the structures without ACE2. These scores are of at least 4 and usually much higher at 6 to 10 (X). However, this falls to 322511331201 in 6M0J RBD + ACE2 indicating the significant fall in exposure in intimate binding between spike protein and ACE2 residues. Similarly, where Fab binds (indicated by @), the residues have high exposure scores or are disordered (~) in the structures without Fab of at least 4 and usually 5–10, but that this falls in 7BYR, B chain which to which BD23 Fab is bound. For example, CYFPLQSYGFQ is ~~~~X856672479 in chain A of 7BYR and 9X9866476779 in chain C of 7BYR falls to 211122211 in chain B of 7BYR to which Fab binding occurs. In the case of binding CR3022 in 6XC3, the same region has exposure 32361255141 with some residues that are exposed with scores 5 and 6, but where the low scores are found on inspection to agree with intimate binding. One is of course interested in potential sites throughout the protein for which a receptor or protease, antibody or ligand might bind, but for which an experimentally determined three-dimensional structures were not yet available at the time of this study. In that regard note that RLFRKS in Block 13, a disordered loop in most cases, is a possible heparin binding site. Similarly, there are future novel synthetic ligands, such as peptidomimetics composed of D-amino acid discussed later below, that might have a preventative or therapeutic action, so all exposed sites are potentially of interest. Preliminary proofs of the method have been done by comparing results for structures in which "incidental" solute molecules that are present in the experimental structure, e.g. to aid crystallization, but of which coordinates are usually *excluded* from the analysis, but are retained for comparison, notably dimethyl sulfoxide (DMS 1332) in 6ZP7 that obscures VSPTKL 382–387 in chain B. The idea also applies, albeit "in reverse", to molecules that are of biological relevance. By "in reverse" is meant that their coordinates are normally *included* in the present analysis, but results can then compared with case in which they are excluded from the analysis. In all the above cases, it is again found that scores in the absence of the obscuring molecule or group are of at least 4 and usually much higher at 6 to 10 (X), but fall to less than 4 when the ligand or group is included. In the present study, molecules or groups that are best considered as normally being present most notably include glycosylating groups, typically indicated as ligands or groups NAG, NAM, MAN, or BMA. Decreased exposure is especially clear because here one sees the character '%' indicating at least partial shielding of a sidechain in three dimensional space by a glycosylation group which may, or may, not, be covalently bonded to residues nearby in the sequence. An interesting example relates to the shielding of lysine K 347 in NLKPFERDIST and other sidechains in 6 ZP-B in Block 13 as indicated by the annotation character %. This subsequence has no features that would suggest a glycosylation covalently linked to it, except perhaps O-glycosylation at the serine S and/or threonine T at the C-terminal end, but it was initially surprising and of concern because the N-acetyl glucosamine NAG A1322 (the sugar molecule providing that shielding) does not show in all viewers including Mol* (Javascript) viewer which is the default choice at the https://www.rcsb.org/3d-view/site. However, atoms of NAG A1322 show up well as obstructing access to a cleft containing many of atoms of the lysine K 347 of the B chain space in the space-filling mode of NGL (WebGL) viewer. The nitrogen atom NZ which is well exposed if the glycosylation is removed is particularly close to NAG atoms. Somewhat similar comments apply to the glutamate E 350 in the same subsequence, which lies quite close to the lysine headgroup. The lysine and glutamate charged head groups (positive and negative respectively) would normally expect to be exposed because of the heavy free energy cost of removing water, and appear accessible as a recognition site in the cleft in the absence of this shielding, at least for smaller ligands. Any access would certainly seem likely be obstructed by the glycolsylation.

It is interesting that the above important disordered loops ~~~~~~~~ that are disordered in 6ZP5 chains B and C are ordered in 6 ZP5 A chain, i.e. the SARS-CoV-2 spike in prefusion state (1-up closed conformation), as discussed in some detail as a worked example with a molecular graphic in Discussion Section 5.1. It is also interesting that they are ordered in 6ACC SARS-CoV-1 (where they are similar in all chains A, B, and C), as follows. See BLOCK 14.

```
BLOCK 14.
GNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQ 506
9968243125497719545657679345469784978779626622762727484684 73 6ZP5-A (SARS-COV-2)
8777655554555566666666566666666677777766665554445555555443 6ZP5-A (SARS-COV-2)
GNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCTPPA-LNCYWPLNDYGFYTTTGIGYQ 6ACC-B (SARS-COV-1)
448755335446782753244567774369689494869948725652426695694830
555555555555555555555556667676677877766666655555555556555544
```

The next block continues to have some residue interactions with the above antibodies. See BLOCK 15.

```
BLOCK 15.
Spike domain 1 C-terminal residual domain 527-588, 528-588
                    (...
PYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG  566
~~~2112234575473301046955855676304221335546021372956157214 40 6ZP5-B closed
~~~4333334444444444444445566555444333333333434444444%44433444 6ZP5-B closed sm.
~~~2112134575483301046955855675304221335445021372946157214 50 6ZP7-B open
~~~3333334444444444444455666555444333333333334444444%44444444 6ZP5-B open sm.
1341122232697645203358966855477315252575556122473966157316 50 7BYR-A BD23 Fab
4332223344444445555555566665555444444444444444454455%55544444 7BYR-A BD23 Fab sm.
1220112123454453301155755865677303132234345121373868147201 30 7BYR-B BD23 Fab
3322122223333333333344455665555443333333333344444%%%44433333 7BYR-B BD23 Fab sm.
022111223477X7964155..................................... 6M0J RBD+ACE2
22222234455656677888..................................... 6M0J RBD+ACE2 sm.
032111313658X8X641525X................................... 6XC3 RBD+CC12.1+CR3022
2222223445666666677889................................... 6XC3 sm.
          @@@@@   @@@@
```

4.7. Exposure analysis of the SD2 domain region

```
BLOCK 16.
Spike domain 2 (SD2) 589-626, plus S1/S2 Boundary Cleavage site
                     (...
RDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTnTsNQVAVLYQDVnCtEVSNVFQT  626
44734455023212666212155364221111022364477323143635365 9~~~~~ 6ZP5-B closed
3444343333333%%%33444433333222222233%%%4444444444%%%%~~~~~~ 6ZP5-B closed sm.
45623445123112666212155365221111021364376323143635365 9~~~~~ 6ZP7-B open
3444343333333%%%33444333333222222233%%%4444444444%%%%6~~~~~ 6ZP7-B open sm.
46745343112139675314255364222212121486588333254536377X977443 7BYR-A BD23 Fab
444444433444444444444433333322223344445555545445%%%899887%5 7BYR-A BD23 Fab sm.
43624445112113566223147464221112122386586333243535376 9~~~~~ 7BYR-B BD23 Fab
3333333333333%%%3444443333332222334444445544434444%%%89~~~~~ 7BYR-B BD23 Fab sm.

BLOCK 17.                                          ...)
DQLTPTWRVYSTGSNVFQTRAGCLIGAEHVnNsYECDIPIGAGICASYQTQTNSPGSASS  686
~~~~~~~~~~~~9965542942024437536746846263223301336 7~~~~~~~~~~ 6ZP5-B closed
~~~~~~~~~~~~666666%4444444444%5%%5555544433334 4%44~~~~~~~~~~ 6ZP5-B closed sm.
~~~~~~~~~~~~8965542942024437536745846263333301336 7~~~~~~~~~~ 6ZP7-B open
~~~~~~~~~~~~66666%4443444444%5%%5555544433333444 4~~~~~~~~~~ 6ZP7-B open sm.
~~~~~~~~~~~~~~~~~~~~941124537647756736363333402~~~~~~~~~~~~~~ 7BYR-A BD23 Fab
~~~~~~~~~~~~~~~~~~~~%444444445%%%55555444345667~~~~~~~~~~~~~~ 7BYR-A BD23 Fab sm.
~~~~~~~~~~~X766663383102332664673673726322302~~~~~~~~~~~~~~~ 7BYR-B BD23 Fab
~~~~~~~~~~~998765%4%333333344%%555544433445679~~~~~~~~~~~~~~ 7BYR-B BD23 Fab sm.
         S1/S2 Cleavage site RARR replaced by GSAS:  RRARˆ
```

This of interest as regards the S1/S2 cleavage site. See BLOCK 16 and 17.

Loop PRRARS in the native SARS-COV-2 structure is a possible heparin binding site. Note that the S1/S2 Cleavage site RARR in the engineered and determined structure is replaced by GSAS. The above important disordered loops ~~~~~~~~ are more ordered in 6ACC SARS-CoV-1. They are similar in all chains A,B, and C. The trypsin-treated structures in SARS-CoV-1 are not disordered in the C-chain. See BLOCK 18.

4.8. Exposure analysis of the S2 neck and S2' cleavage region

This region is of interest as containing the KRSFIEDLLFNKV discussed in Introduction and used as an example. Relevant sections are in bold. While indeed it fairly buried in many structures, it is exposed and immediately proceeded by a disordered loop in the case of BD23 Fab antibody head binding, *suggesting that a combined attack by antibodies may be effective*. See BLOCK 19, 20 and 21.

```
BLOCK 19
S2 (KNECK) REGION
Cleavage domain 678-845
VASQSIIAYTMSLGAENSVAYSnNsIAIPTnFtISVTTEILPVSMTKTSVDCTMYICGDS  746
~~964121052444664635244622122222423245542313353333435311247 4 6ZP5-B closed
~~4%44433333444444%4%4%%%33%32%%%333333333333333333333344434 6ZP5-B closed sm.
~~964111052355664635244612122222423245532312353333435311137 4 6ZP7-B closed
~~4444433334445455%5%%%%%2%22%%%333333333333333333333334434 6ZP7-B closed sm.
57895321153454664526245622112333423355433333353333457432146 4 7BYR-A BD23 Fa
98765544444444454445%4%%%3%332%%%333333344333344444344445444 7BYR-A BD23 Fab sm.
56895231053345665436255512111233423245432223363333446322246 3 7BYR-B BD23 Fab
87654444444454555%%%%%2322%%%333533333333333344443334444444 7BYR-B BD23 Fab sm.

BLOCK 20.
TECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFnxfsFsQIL  806
771364265344346424732543254137126612343642285488652462424423 6ZP5-B closed
445544444444444444444444433444333433333444455555%%%5544%%%%45 6ZP5-B closed sm.
771364265344346424732542254137126612343642185488652462424323 6ZP7-B closed
445544444444444444444444433343333343333434455555%%%%44%%%%44 6ZP7-B closed sm.
861362264345346434633543264147136513443552275488552572434433 7BYR-A BD23 Fab
444444444444444444444444433443433444444444555555%5554%%%%55 7BYR-A BD23 Fab sm.
762353355244345333632543153136025423342761274376642462435533 7BYR-B BD23 Fab
444444434444444444433333333333333333434444445544%4%4%%%%%%45 7BYR-B BD23 Fab sm.
782473376435446423632543264146136513443752275486652572424523 7BYR-C BD23 Fab
555555544544444444444444444444444444444444555555%5%5544%%%%55 7BYR-C BD23 Fab sm.

BLOCK 21.
S2' cleavage and KRSFIEDLLFNKV Highly Conserved Motif Region
                                        ...)
PDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLT  866
557783573161122226327 8~~~~~~~~~~~~~~~~~~~~~~~~~96641335456424 6ZP5-B closed
455555544433333344555 5~~~~~~~~~~~~~~~~~~~~~~~~~55555544444444 6ZP5-B closed sm.
557783573061022226327 8~~~~~~~~~~~~~~~~~~~~~~~~~96651345456424 6ZP5-B open
455555544333333334555 55~~~~~~~~~~~~~~~~~~~~~~~~55555544444544 6ZP5-B open sm.
5599~~~842621323474489~~~~~~~~~~~~~~~~~~~~~~~~95752335456434 7BYR-A BD23 Fab
6788~~~776654334456789~~~~~~~~~~~~~~~~~~~~~~~~98766544445554 7BYR-A BD23 Fab sm.
5487~~~640521222564477~~~~~~~~~~~~~~~~~~~~~~~~X6552435456434 7BYR-B BD23 Fab
6677~~~765544334456789~~~~~~~~~~~~~~~~~~~~~~~~98766544445544 7BYR-B BD23 Fab sm.
5688~~~740621323474489~~~~~~~~~~~~~~~~~~~~~~~~88663246556344 7BYR-C BD23 Fab
6788~~~7666543344577 89~~~~~~~~~~~~~~~~~~~~~~~~98876555455554 7BYR-C BD23 Fab sm
        ^ S2' Cleavage site
```

```
BLOCK 18.
PDB-6ACC Trypsin-cleaved spike protein C-chain in SARS-CoV-1
DQLTPAWRIYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASVAYTMSLGADSS
~~~~~~~~~596964349674203242784985683736443330138432063466 68
~~~~~~~~~666666655444444445556666666554433345678876654455 66
```

There is a degree of protection by surrounding atoms except for the K, F, N and TLA in KRSFIEDLLFNKVTLA. The above important disordered loops ~~~~~~~~ are present in 6ACC SARS-CoV-1. Loop and α-helix N-terminus SKPSKR is a possible heparin binding site. Again, all chains A, B, and C there are essentially the same. In view of the proposed importance in the present author's previous publications as to actual exposure or exposability, the trypsin-treated SARS-CoV-1 which has bound ACE2 is of particular interest. See BLOCK 22.

```
BLOCK 22.
PDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLT 6ZP (SARS-COV-2)
PDB-6ACC Trypsin-cleaved spike protein A-chain in SARS-CoV-1
PDPLKPTKRSFIEDLLFNKVTLA~~~~~~~~~~~~~~~~~~~~ICAQKFNGLTVLPPLLT
468X98785163133333754899~~~~~~~~~~~~~~~~~~~~X5969542356567435
6666666655444444456789X~~~~~~~~~~~~~~~~~~~~X9877655555455655
PDB-6ACC Trypsin-cleaved spike protein B-chain in SARS-CoV-1
PDPLKPTKRSFIEDLLFNKVTLA~~~~~~~~~~~~~~~~~~~~ICAQKFNGLTVLPPLLT
468X98784163133333754899~~~~~~~~~~~~~~~~~~~~X5969542356567435
6666666655444444456789X~~~~~~~~~~~~~~~~~~~~X9877655555455655
PDB-6ACC Trypsin-cleaved spike protein C-chain in SARS-CoV-1
PDPLKPTKRSFIEDLLFNKVTLA~~~~~~~~~~~~~~~~~~~~ICAQKFNGLTVLPPLLT
468X98784162133337754899~~~~~~~~~~~~~~~~~~~~X5969542356567435
6666666655444445556789X~~~~~~~~~~~~~~~~~~~~X9877655555455655
```

See Section 5.1 for further discussion on this concerning the exposure of the motif and the effect of antibody binding at a remote site on that. Also, the story for SARS-COV-1 is essentially the same as for the SARS-COV-2 structures around the first part KRSFIEDLLF of the conserved motif KRSFIEDLLFNKV, but the part NKVTLA is more highly exposed. It is also shortly followed by a disordered loop like that in SARS-COV-2. At a more detailed level, there are significant differences in intramolecular interactions in SARS-COV-1 and SARS-COV-2 as discussed below. Put together with the above observations, the particular residues involved in recognition that would explain high conservation are KR.F… … NKVTLA, but it remains to understand why the whole of this subsequence is well conserved. On the whole KRSFIEDLLFNKV in the conserved motif appears exposable but the rather buried nature of IEDLLF in the motif is interesting because the negatively charged glutamate E and a negatively charged aspartate D are typically hydrated in proteins, i.e. exposed to surrounding water hydrogen atoms and often involved with positive metal ions. In a few cases in proteins they can form internal salt bridges particularly with positively charged sidechains lysine K and arginine R and sometimes histidine H. This appears to be the case with the aspartate residue D. Even so, the energetically favorable coulombic charge-charge interaction is typically outweighed by an unfavorable desolvation of interacting charges, such that charged sidechains of glutamate E, aspartate D, lysine K, arginine R, and often histidine H, typically prefer solvent exposure. Recall from Section 4.3 the following top six most exposed sidechains even when using exposure scores taken only from the spike protein: K 5.9, E 5.4, R 5.1, D 5.1. Histidine H, often just partially charged, is at sixth position at 4.5, but superseded by uncharged asparagine N at 5.1.

In SARS-COV-*2* 6ZP5 the picture is somewhat more variable between chains A, B, and C than it is in SARS-COV-1 6ACC, as follows.

(a) In SARS-COV-1 6ACC the residues IEDLLF of the motif are sunk into the structure, although the carboxyl groups glutamate E and aspartate D carboxyl end groups appear accessible to solvent. Nonetheless, there are internal intramolecular interactions with sidechain end groups. The glutamate E forms hydrogen bonds to serine S in the KRSFIEDLLFNKV motif and the threonine T in subsequence AAYTAA in an α-helix on the C-terminal side of the above disordered loop in the same chain (monomer), and the aspartate D forms a salt bridge by two hydrogen bonds to the positively charged lysine K of the KRSFIEDLLFNKVTLA motif in the same chain. These interactions are similar in chains A, B, and C although in chain C the aspartate D appears to be less intimately involved with the lysine K and seems more accessible to solvent.

(b) In SARS-COV-*2* 6ZP5 the residues IEDLFF of the motif are again sunk into the structure, and again the carboxyl groups glutamate E and aspartate D appear accessible to solvent. Nonetheless, there are again intramolecular interactions. In chain B the glutamate E forms two hydrogen bonds to the sidechain and backbone amide of serine S in the KRSFIEDLLFNKV motif, but not as intimately (as in SARC-COV-1) with threonine T in subsequence AAYTAA in the α-helix on the C-terminal side of the above disordered loop in the same chain. In chain A it also forms a hydrogen bond to the backbone amide of serine S 1055 (1028 in the PDB entry). In contrast to SARS-COV-1 6ACC, the aspartate D seems more exposed to possible interactions with water, and lacks the intimate salt bridge with lysine in B though there is a hydrogen bond to the arginine R in the motif in chain A, and into the preceding lysine K in KPSKRSFIEDLL … in chain C.

4.9. *Remaining spike glycoprotein sequence*

This section is of interest regarding the changes of exposure that occur in forming the activated spike. It is evident that there are radical changes in exposure, and important increases in exposure. While of considerable interest, the brief duration of this state may make use of synthetic peptide vaccine, peptidomimetics, and derived traditional drugs less likely. However, at the time of writing, this needs to be proven experimentally and certain antibodies and ligands may have utility in trapping this form prior to cell entry, or simply virus replication. See BLOCK 23, 24, 25, 26 and 27.

```
BLOCK 23.
CR 847-911, HR1 912-984, CH-BH 985-1071, SD3 1072...
DEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQ  926
6632162131211011101521334633220132020311441514221024449310526ZP5-B closed
4433333222211111112223333333%22221122222233322223333333%33%%6ZP5-B closed sm.
6632162031211001100521334633120122020311441513221013449310516ZP7-B open
4433333222211111111223333332%%2211111222222322222233333%33%%6ZP7-B open sm.
6633263131212112101521333733221233121522232624322134547321627BYR-A BD23 Fab
4444433322221122222233333333%33222222233333333333333333%44%%7BYR-A BD23 Fab sm.
7622263221211001100421434633220133120411421514221023448310627BYR-B BD23 Fab
4443333222211111111223333333%32222222222222222223333%33%33%%7BYR-B BD23 Fab sm.
.....630561063 6XLT-A activated
.....544333322 6XLT-A activated sm.
.....053106 6XLT-B activated
.....333322 6XLT-B activated sm.
.....721631163 6XLT-C activated
.....444333322 6XLT-C activated sm.

BLOCK 24.
FNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDP  986
2242237356535678854571462256347255312612434131533554043314626ZP5-B closed
2%%33444%5565556555554444444444444443333333333333333344344 6ZP5-B closed sm.
2241237346425677854471462256347255312622435251533425404431566ZP7-B close
2%%33444%4555555555554444444444444443333333333333333344344 6ZP7-B closed sm.
3352338446445667864572563355346246323624435242543635414442467BYR-A BD23 Fab
33444%44%5566666666555544444444444443443344344443444444444447BYR-A BD23 Fab sm.
2252338336443677875462443355346245323514535133554636513573577BYR-B BD23 Fab
33%4444445555566555555444444444443343433334444444445445454457BYR-B BD23 Fab sm.
0230023226232203011511611261071124516------144234414614451576XLT-A activated
2222222222222222222222222333322233333------333333333344445676XLT-A activated sm.
1230033326123220301261261127106212410------244225323613461696XLT-B activated
2222222222222222222222233333322233333------333333333344445676XLT-B activated sm.
0230032236223120401261261126107223 5------512442354247144616 9XLT-C activated
2222222222222222222222333333333333------343333333344444456676XLT-C activated sm.

BLOCK 25.
PEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKG  1046
6605641361042224013331352435136254305511311110222228444312726ZP5-B spike closed
4444433333322222223333333334433343333322222112223333334444333 6ZP5-B spike closed sm.
6605541460042225014331452434136154305501211110222229444312826ZP7-B spike open
4444433333322222223333333334433343333322221112223333334444332 6ZP5-B spike open sm.
6505542351042334233235243414624430561142212023238445423736BYR-A BD23 Fab
4444433333333333333333333344333433333322222223333344444443376BYR-A BD23 Fab sm.
5606432461142224023221352434146154205511421110122237445323636BYR-B BD23 Fab
5444433333333222222233333333333333333222221222223333444433376BYR-B BD23 Fab sm.
4568467267372462706820550157456846726737246.....6XLT-A activated sm.
8998876555554554444444333567899887655555545.....6XLT-A activated sm.
46X9493582673723628158205604611540561316368.....6XLT-B activated
89X9887665555545544444443333333333334456779.....6XLT-B activated sm.
56X9856826738356281683165056017406813 39X8X.....6XLT-C activated
89X9987765555555554544444433443444556789 9X.....6XLT-C activated sm.

BLOCK 26.
YHLMSFPQSAPHGVVFLHVTYVPAQEKnFtTAPAICHDGKAHFPREGVFVSnGtHWFVTQ  1103
5002021111352120201012236675441231015866020158221015553210326ZP5-B spike closed
2222122222222222112233%%%%%333333%33334443322333%%%%3%433 6ZP5-B spike closed sm.
4002021111352120201013236674440230015866020158221015553210326ZP7-B spike open
2222111222222211111223 3%%%%%333333%33334443222233 3%%%%3%433 6ZP7-B spike open sm.
5012032222462221201023336565351241114866021157431012666431036BYR-A BD23 Fab
3333222222222222222222333%%%%4433334%33334443323334 4%%%%3%433 6BYR-A BD23 Fab sm.
5012021232462110201012247564331241114866020257331013666431036BYR-B BD23 Fab
2222222222222222211223 33%%%4333333%33334443323334 3%%%%3%433 6BYR-B BD23 Fab sm.

BLOCK 27.
RNFYEPQIITTDNTFVSGNCDVVIGIVnNtVYDPLQPELDSFKEELDKYFKnHtSPDVDL  1143
5534456722473035367487276445572654574758.....6ZP5-B spike closed
344%444%4444444454555555%55%%55555567889.....6ZP5-B spike closed sm.
5434456722373025367486266445472654674768.....6ZP7-B spike open
33%%444%4444444454455555%55%%55555667889.....6ZP7-B spike open sm.
4345455845784238367687584827838.....7BYR-A BD23 Fab
4444555555555556656666666666666.....7BYR-A BD23 Fab sm.
4445456845774237377586485737838.....7BYR-B BD23 Fab
4444555555555556555666666666666.....7BYR-B BD23 Fab sm.
```

The remaining sequence is as follows. It is missing in many of the experimentally determined three dimensional structures.

```
GDISGInAsVVNIQKEIDRLNEVAKNLnEs
LIDLQELGKYEQYIKWPWYIWLGFIAGLIA
IVMVTIMLCCMTSCCSCLKGCCSCGSCCKF
DEDDSEPVLGVKLHYT
```

There is a possible host sialic glycan binding site $ above at (H/I) WKWPWYIWL [9].

5. Discussion

5.1. Principles of use and some general results

Several tactics are available for the design of peptidomimetics, encouraged by the fact that since the 1990s there has been considerable progress in synthesizing peptides that are purer, larger peptides (e.g. Refs [25–29]). With the rise of COV-19, a great deal of diverse work has also been done that relates to the immunogenicity of the spike protein

and subsequences including glycosylated subsequences in it (e.g. Refs. [30–32]), including some novel methods of assessing accessibility as a basis of design (e.g. Ref. [33]). There has also been a growing understanding of the biological properties of peptidomimetics that are partly or wholly composed of D-amino acid residues [34], and how to select those likely to have the required activity [35]. As noted by vanPatten et al., peptidomimetics based on D-amino acids have a very desirable property of resistance to a patient's proteases, but other options include design of peptidomimetics with reduced and functionalized amide bonds, peptoids, urea peptidomimetics, peptide sulfonamides, oligocarbamates, azapeptides, β-peptides, and N-modified peptides [36].

The focus in this paper is on the virus responsible for COVID-19, and somewhat less so on the algorithm used. As discussed in Introduction Section 1.7, and in Theory and Results, there are many methods of assessing the surface and surface exposure. There are indeed many standard methods. For example, entering https://www.rcsb.org/3d-view/6ZP5/1, selecting the viewer NGL (WebGL), and then selecting Style and then Surface, produces a surface view of the spike protein. Such views, as well as space filling views (atoms as spheres with van der Waals's radii), were useful in confirming the validity of the present software program in cases where obscuring of a sidechain by a glycan bound to a residue remote in the sequence seemed perhaps surprising. An example of lysine K 347 and glutamate E 350 in NLKPFERDIST was discussed in Section 4.6 in relation to Block 13. There are nonetheless several reasons for developing the new software in the present study. Those of a more theoretical and fundamental nature were discussed above, e.g. in Results Section 4.3. Some matters of practical convenience are as follows.

(1) At the time of this study and of writing, the present author was not aware of any layout such as that of the above "blocks", and in particular for the SARS-COV-2 spike protein, that would facilitate peptidomimetic design and discussion of it.
(2) A standard format and single character annotation was preferred as a one-dimensional representation as described in Methods Section 3, to facilitate incorporation in a larger automated approach including analysis by data mining and machine learning including parameterizing of inference nests and neural nets (including "Deep Learning"), and flexible user intervention. It is understood that the description of shielding by glycosylation (by the % character) is not exhaustive, and many human tissues can produce coronaviruses with different glycosylation patterns. However, by comparing many aligned sequences of structures (perhaps combined with prediction of glycosylation), the present method raises "red flags" as to the need to provide sophisticated glycan technology.
(3) It was felt that many current available methods such as surface-reporting algorithms did not exactly capture the notion and degree of exposure as would facilitate design of peptidomimetics. For example, for *retro*-inverso peptidomimetics, sidechains are correctly placed relative to each other in three-dimensional space but backbone amide and carbonyl groups are interchanged, so a run of high exposure measures indicate that all the sidechains "stick out" at the surface, and form the basis of interaction, is important.
(4) It is the functional ends of residue sidechains that provide molecular recognition (in contrast, backbone hydrogen bonding groups are universal to polypeptide chains).
(5) A range of 6.5 Å seems more suited as to what is in the neighborhood of sidechain atoms responsible for recognition, and methods and results by other authors either did not use this distance, or took a non-distance approach. As noted in this paper, 6.5 Å does include opportunity for inclusion of a small organic molecule or side-group of a larger binding molecule between sidechains. However, as shown in Tables 1–3 it is an optimal value at which the notion of neighboring sidechains and glycosylating groups might the defined, even simply as the point at which results vary less with R_{max} because the derivative of the exposure score with respect to R_{max} is zero at the peak value in that general range of close separation distance. It was felt that a slight shift in parameters in most current methods made the results for, and notion of, exposure change excessively, whereas because of Eqn. (2) this is much less so.

The primary and ultimate purpose of the present approach is automation of design of peptide peptidomimetics, especially with use of D-amino acids in mind; however, the general idea for using the current output in an unautomated way for a variety of purposes is simple. A researcher interested in a list of potential candidate subsequences to use the basis of synthetic vaccines or peptidomimetics can use the above output to look for runs of residues with scores indicating high exposure such as in 6ZP5 chain A aligned beneath the characters of the subsequence GSTPCNGVEGF, i.e. 79849888777, and 66677777766 beneath that as the summary sequence in which measures are smoothed over neighboring residues. The numeric characters in the string along with other one-character notation, also combined with alignments with similar experimentally determined structures and other related proteins, provide several kinds of information. The above specific example is perhaps a more obvious choice because it is involved in the ACE2 binding region that has been well studied, but it is insightful to see how the present output is consistent with a known case in which comparison can be made with prior knowledge expectations (see Fig. 2).

Alignments in the blocks can show variation in glycosylation. GSTPCNGVEGFNCYF and neighboring residues is not a glycosylated region in 6ZP5, nor obscured by glycosylation. Not all the experimental structures are glycosylated and in actual infection glycosylation may vary with tissue, so that any asparagine N, serine S or threonine T are particularly suspect as glycosylation sites. Gutamine Q which is not present, and tyrosine Y which is present, can also occasionally be found glycosylated in proteins. Being in a disordered state in chains B and C

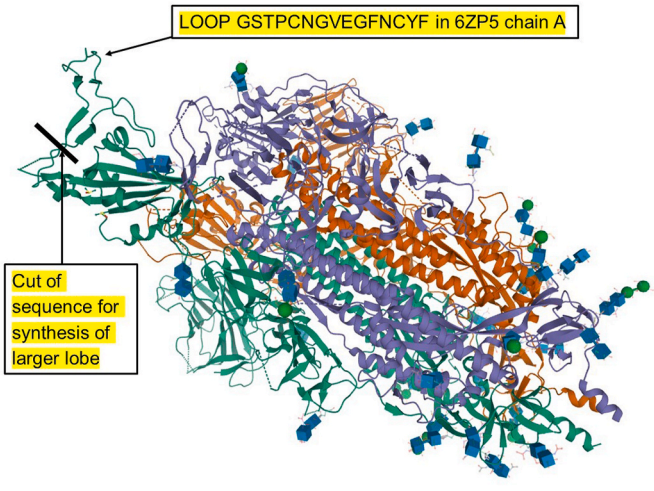

Fig. 2. Location of loop GSTPCNGVEGFNCYF in the Spike Glycoprotein.

and other structures, it is immediately likely to be a loop or part of one, as shown for the A chain in Fig. 2. It is an ordered loop in many structures binding ACE2 or antibodies, although it is disordered in many structures in the absence of binding. That binding of such ligands stabilizes the conformation of nearby loops is a notable feature in the present results. Inspection of blocks 13 and 14 above shows that the larger version of this sequence GSTPCNGVEGFNCYF is also well exposed overall with the exception of the second cystine C, and that loop is close by a disulfide bond between the two cystine residues C. Use of this subsequence and extended sections of it are discussed as a possible peptidomimetic in Section 5.4. The cut shown in Fig. 2 relates to the largest subsequence closing the above that is of particular interest in the present paper as discussed later below.

Similarly, a researcher noting an exposed sequence 488837456 may have some confidence that these provide enough recognition surface (provided by the sidechains), to be at least worthy of consideration, but there is further information. It refers to the central part of VRFPNItNLCPFGEVFnAtRF with exposure scores 321354888374 564363884. The sequence starts and ends with N-glycosylation and that is realized in some of the structure determinations. However, the high degree of sidechain exposure above remains in structural determinations in which that glycosylation is realized. It obscures six sidechains as shown by the summary string 4444%%555556%%55%%544. Recall that the sequence of scores in which glycosylation annotation % is added is a locally averaged exposure score: 488837456 remains the exposure scores for the central section of residues. Similarly, a researcher interested in a section of sequence because of some prior knowledge, e.g. SLLIVNnAtNVVIKVCEF, may rightly be warned away that 122213265211231121 indicates low exposure in part because that site is well shielded by glycosylated at least in some structures, as indicated by 333333%%%%3%2%2234, unless perhaps that laboratory is well versed in glycosylation technology.

There are several occasions when binding at one point by a ligand such as an antibody or ACE2 has effects along the sequence. For example, sequence (a) below that starts in block 13 is normally associated with a loop ~~~ as shown in (b), but on interacting with antibody at the residues indicated by @ in (c), the antibody not only greatly decreases exposure for FPLQ but also stabilizes the loop and actually enhances access to small organic ligands at HAPATV.

```
(a) FPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKK
(b) 85555~~~~~~~~~~~~~~~~2112234575473301046 9
(c) @@@@@@                      @@@@@    @@@@
(d) 3612551414581384230321113136 58X8X641525X
```

See Section 5.3 below concerning the interesting effect of antibody binding on remote sites.

5.2. Further comments on use

The typical procedure for design of synthetic peptide vaccines and peptidomimetic agents should be obvious, when considered alongside the worked example in Ref. [6], noting that the algorithm described here is a tool to be used amongst several. Notably, it is to be used alongside standard bioinformatics procedures such as BLASTp and Clustal Omega [6]. Recall that a subsequence is a section of sequence that is potentially a special biological signal or "motif", but that to serve as the basis of vaccine it is not required that it has a biological function. Caution is required in that Ref. [6] described some rules-of-thumb for synthetic vaccines, fairly well known in the field, that change slightly the sequence for best effect as an isolated peptide used in a vaccine. For example, the exact point of starting and ending the peptide can depend on choice of linker to a carrier protein, or at the point where the resulting N-terminal + NH_3 - or C-terminal -COO^- of the peptide mimics a correspondingly charged sidechain in the original subsequence. As discussed below, there can be even more drastic changes for a peptidomimetic agent, such as the "retroinverso" approach, i.e. using only D-amino acids and reversing the sequence [6]. Some chemical and peptidomimetic companies are making the author's proposal available commercially for research purposes, but they do not include these modifications that are described in Ref. [6].

Of usual interest are those subsequences in which the exposure score, and particularly the smoothed exposure score is reasonably exposed i.e. 4 or more (characters 4–9, X), and even disordered loop (character '~'), make a section of sequence at least worthy of consideration. Of particular interest, nonetheless, should be those residues that are especially well exposed (characters 5–9,X, ~). Again, they should not be shielded by the virus glycan coat (shielded is indicated by character '%'), unless the laboratory has sufficiently sophisticated technology for managing the preparation and attachment of glycans [6]. Again, recall that a subsequence is a section of sequence that is potentially a special biological signal or "motif", but that to serve as the basis of vaccine it is not required that it has a biological function. They should also ideally be well conserved across coronaviruses, if the virus is not to evolve quickly against a vaccine or drug directed at that subsequence, hence the reference to use of BLASTp and Clustal Omega above.

5.3. The example of KRSFIEDLLFNKV

The subsequence KRSFIEDLLFNKV [5–7] remains a good example. Based on data from immunological lab research discussed in Introduction Section 1.6, the site appears spontaneously exposable and so the peptide forms the basis of a vaccine that is even neutralizing for SARS-CoV-1. As discussed in Introduction, the section KPS**KRSFIEDLLFNKV**TLADA is at least partly obscured (835**731611222263**278~~) until ACE2 binding and S1/S2 cleavage has occurred. But as shown in the present paper, it can even be extensively exposed (~~~**766654334457**89~~) when antibodies bind, even at a remote location to the cleavage site. The two leucine amino acid residues, LL, remain fairly obscured (~~~766654**33**4457789~~~) however, but that some sidechains that are exposed overall do so is usually the case in subsequences (often *circa* half are). There is simply a need for sufficient sidechains pointing outward for the recognition. Consequently, at very least a combination of a general vaccine and a synthetic construct based on KRSFIEDLLFNKV could work. But also, since the FAb binds well away from the S2' cleavage point, the site may be exposable to a direct synthetic vaccine alone, and serviceable as a basis for a vaccine, as the experimental immunology suggests for SARS-CoV-1 (instruction Section 1.6). This motif is still worth investigating, not least because it is highly conserved.

Aspects of the above may be illustrated in terms of actual use of the output format. One may deduce that the binding of antibody to chain C 7BYR-C BD23 Fab opens up KPS as a loop and increases the exposure of RSFIEDLLFN in (a) below, i.e. compared with exposure scores in (b) that become those of (c) on binding antibody. Here (b) is structure 6ZP5 chain B and (c) is 7BYR chain C with bound BD23 Fab. Note that (d) as 6ZP5 chain C and (e) as PDB-6ACC chain C for the Trypsin-cleaved structure for SARS-COV-1 spike with ACE were already somewhat more exposed without the bound Fab.

```
(a) PDPSKPSKRSFIEDLLFNKVTLA
(b) 55778357316111222263278~
(c) 6788~~~766654334457789~
(d) 67885858416224334744 8X9
(e) 468X98784163133337548 99
(f)          ^ S2' Cleavage
```

Note also that in the ACE2 bound trypsin-treated SARS-COV-1 spike the opening effect of ACE2 binding is less dramatic than remote Fab binding. In all these cases, however, it is reasonable to say that this site does have significant exposure of sidechains for design of small molecules. Importantly, the opening effect of antibodies at remote sites could form the basis of a synthetic vaccine that attacks a remote site as well as the above motif.

5.4. D-peptide mimetics and an example based on GSTPCNGVEGF

The following are examples only, but they illustrate how far one can go with peptidomimetics when defined more broadly as starting with an amino acid residue subsequence of the protein target. The ability to synthesize larger peptides and proteins by ultrastructural chemistry (i.e. not by ribosomes) facilitates experimental studies for more complex peptides, and importantly the use of peptides made partly or solely out of D-amino acids (e.g. Refs [25,26]). It opens the possibility of less biodegradable peptidomimetics that can persist several days in the circulation after injection, before degradation by other endogenous, and promises an era of more refined bionanotechnology for peptidomimetics [27,28]. Because of their low biodegradability, the compounds developed are more like those of traditional in-a-pill drugs even though they can require injection or use of aerosol to avoid non-enzymatic acid hydrolysis in the stomach. Immune response by the T cell system to a larger D-peptide as a D-protein is not possible because of resistance to proteolysis that is required for presentation of fragments and immune memory. Synthesizing structures of D-amino acids is no more difficult than for L, although D-amino acids are more expensive. The challenge being addressed since the 1990s has been the size of the structures sometimes required, which (whether composed of L or D amino acids) need higher speed synthesis to avoid side reactions and chemical ligation techniques to join sections of polypeptide chain. One approach is to generate (in the "wet" laboratory) combinatorial chemistry libraries of D-peptides and select them against, for example the SARS-COV-2 receptor ACE2. Unfortunately, the astronomic numbers of combinations makes success less likely than a targeted sequence-based approach, although of course any such targeted approach as discussed below can also be a powerful starting point for combinatorial chemistry.

The sequence-based computational method described in this paper is well suited to design of peptides made largely of D-amino acids. As well as detecting surface regions, use of D-amino acids presents a problem that must be addressed concerning how to choose the required sequence for the peptide being synthesized. With inclusion of D-amino acids, the required peptide can differ considerably from the subsequence or motif in the target protein on which it is based. The sequence will likely need to be changed to mimic the interaction surface. Typically, one wants to make an antagonist. For example, to address SARS-COV-2 infection one may wish to inhibit binding of the spike protein to its receptor ACE2 or to host sialic acid glycoproteins, or to proteases for the S1/S2 or S2' cleavage sites, all by mimicking the structure of the binding site on the target spike protein. The method discussed and worked though in some detail in Ref. [6] for the KRSFIEDLLFNKV motif was the *retro-inverso* strategy. Here the selected sequence from the target protein is written backwards and then synthesized out of D-amino acids. While the sidechains are then in the correct position when the peptide is induced into a similar conformation, the effect is as if the backbone amide N–H groups and carbonyl C═O groups have been interchanged, compared with the original biological structure in the target protein. As noted in Ref. [6], this will likely not matter if the main interactions with the peptide involve the sidechains rather than the backbone, and this was an important motivation for the current approach that focuses on the accessible sidechains rather than the backbone.

A potential peptidomimetic antagonist binding ACE2 is suggested by the section of sequence containing PLQSYGFQ, say VEGFNCYFPLQSYGFQPTN, in block 13 in Results Section 4.6. PLQSYGFQ is partly disordered with scores of mostly 5 or higher in structures where ACE2 is not bound, but exposure drops to scores of 0–3, e.g. 11331201, indicating a tight binding of ACE2 in structures with ACE2. An example way of proceeding would be an attempt to explore use of a *retro*-inverso peptidomimetic containing the sequence written backward, i.e. it would contain PLQSYGFQ but now backward as QFGYSQLP, which is then synthesized of D-amino acids.

However, the above region has already been well studied in relation to ACE2 as a pharmacophore and, as a good example, there is also a case for the region which overlaps with the above on the N-terminal side. A peptidomimetic could still sterically interfere with the spike-ACE2 interaction. That is the subsequence GSTPCNGVEGFNCYF in the ACE2 binding region. It was the loop shown in Fig. 2. To check as to whether there are reasonable choices regard to ends of the synthetic peptide sequence with the charged N and C-termini, which are of course involved in peptide bonds and not charged in the original spike protein structure, it is useful to consider the larger sequence containing the above subsequence, e.g. EIYQAGSTPCNGVEGFNCYF. One reason for that choice is that all the sidechains are seen to be well exposed with scores 93454697849787796266 and beneath that 66666666777777666655 for smoothed scores in the summary score (in the 6ZP5 A chain). As a comparison, it is useful to start by considering use of the subsequence as a peptide for a synthetic vaccine (in which case we do not consider the sequence backward), the cystine C provides a convenient linker to a carrier protein. In that case, however, it is desirable to replace the second cystine C by its natural analogue serine S, to avoid unwanted disulfide bonding between peptides. In using that particular strategy, there are for that subsequence no particularly obvious ways to shorten the sequence in such a way that charged sidechains in the sequence are mimicked by charged N- and C-termini of the synthetic peptide. One may truncate at the N-terminus of the subsequence as written, though it leaves a charged amino group NH_3+ and so may be best N-acetylated (CH_3(C═O)-), and truncate at the C-terminus as written, though it leaves a charged COO- that may be best N-methyl amidated (-(NH)CH_3). These make an interesting comparison with the *retro*-inverso peptidomimetic, where the first step is to write the sequence backward: FYCNFGEVGNCPTSGAQYIE: in this case omitting the C-terminal glutamic acid E will allow the carboxy terminus to emulate that sidechain. The natural disulfide bridge (but now in mirror image) should be retained because it provides the native conformational constraints that can improve biological activity or confer thermostability, e.g. by reduction followed by oxidation at reduced concentration. Consequently, a plausible proposal is as follows.

N-Ac-dextro-[FYCNFGEVGNCPTSGAQYI]

Note that the three glycine residues G which are achiral due to absence of sidechain slightly reduces slightly the cost of production that D-amino acids present.

Also of potential interest for a larger peptidomimetic is the longer section of sequence enclosing the above as follows, for which in most structures most smoothed exposures scores are 5 or more (except for one at 4), and which essentially comprises the lobe (though not the larger

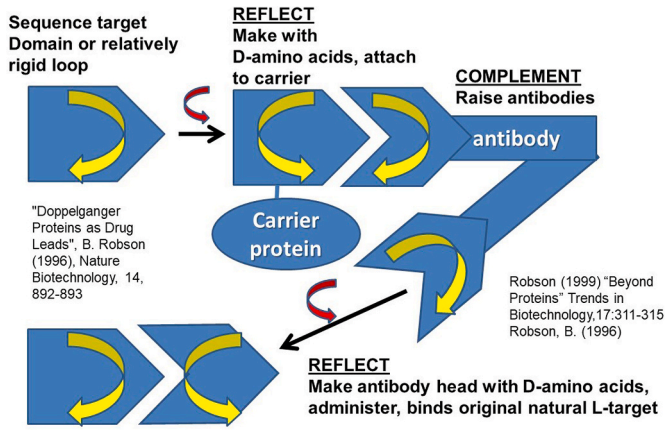

Fig. 3. The reflect-complement-reflect method.

compact domain) between the antiparallel strands of β-peated sheets that form the neck of the loop in Fig. 2. These have the sequence LYRL and PLQS but they do not exhibit a clear alternation of residue-by-residue exposure scores that is typical of well defined pleated sheet structures in which one face is more exposed than the other. As shown by the cut point in Fig. 2, this lobe is an integral conformational unit (but not a compact domain) which comprises 46 residues: NYLYRLFRKS NLKPFERDIS TEIYQAGSTP CNGVEGFNCY FPLQSY. The segments LYRL and PLQS come together in space to form the short antiparallel β-pleated sheet that forms the neck of this structure by the cut in Fig. 2. There is the issue that (see Block 13) NLKPFERDIST appears to partly shielded by glycosylation in some structures, which would suggest the need for some fairly sophisticate glycosylation technology. Putting that aside, tackling structures of this size as a peptidomimetic, and even the whole compact ACE2 binding domain or more, is today possible. This is valuable because a stably folded domain (essentially a "miniprotein") is more likely to retain the required recognition surface of van der Waals', electrostatic, and hydrophobic interactions. However, even the D-peptidomimetics based on the short subsequence GSTPCNGVEGFN-CYF might well be sufficiently stable due to the stabilizing effect of the disulfide bridge, and more synthetic stabilizing links might be introduced.

Unfortunately, *retro*-inverso peptides have had rather limited success as peptidomimetics, presumably largely because (when backbone hydrogen bonding hydrogen groups are involved in intermolecular interactions with the target), the backbone amide hydrogen bond donors and carbonyl hydrogen bond receptors are interchanged. As a less well known and less well studied example for developing larger peptides made of D-amino acids, Fig. 3 illustrates a promising approach developed by the present author which has yet to be fully explored, but may help overcome this problem by addressing peptides with a more compact fold form [27–29]. The approach involves synthesizing the target subsequence as a peptide but made of D-amino acids, raining antibodies against that by linking to a carrier protein, and sequencing and then synthesizing the resulting antibody heads using D-amino acids. The target and source of subsequence of interest could be the spike protein itself, or a receptor such as ACE2 or host protease that activates the spike. In contrast to the above discussion of *retro*-inverso approach discussed above, it is of interest to take this slightly more unusual choice of attacking the spike protein. This does have the advantage of being potentially refinable so that a peptidomimetic is likely to interfere with host receptor or enzyme function. Whichever is chosen, however, the approach in Fig. 3 is said to be "promising" above because all component steps have been shown feasible. Antibody heads are a particularly difficult chemical synthesis, but the challenges have been overcome by the present author and colleagues [25]. Larger proteins than Fab antibody heads, notably superoxide dismutase, can be synthesized using D-amino acids and they fold in mirror image [26]. The resulting D-protein made as described in Ref. [26] had the same activity as the natural L protein, because the substrate is achiral. When the substrate or ligand is not chiral, D-proteins will function in mirror image, once folded [27,28]. The required step of raising the required antibodies against the D-peptide or D-protein is possible by presenting it as a hapten covalently combined with a suitable carrier protein, which is of course done in, for example, a sheep or cow, not the human patient. While it is well known that D-peptides are far less immunogenic than many considered in such studies, it is also being seen that peptides containing D-amino acids can be immunogenic and raise antibodies in judicious laboratory conditions [34]. It is also the case that a reflect-complement-reflect approach based on phage display has been productive in peptidomimetic design [36]. In principle, use of the animal immune system for the reflect-complement-reflect method is likely to be the most powerful approach: there are processes of refinement and maturation of molecular recognition by the immune system that show considerable sophistication, even though not perhaps yet fully understood. It is therefore pleasing that antibody and immune response have been obtained to SARS-COV-1 spike protein peptides (See Section 61 and e.g. Refs. [16,17,37–40]) and recently SARS-COV-2 spike protein peptides [36–40]. Importantly, D-amino acid *retro*-inverso peptides of somewhat similar size and amino acid content have long been shown to capable of raising antibodies when attached to a suitable carrier protein (e.g. Ref. [41]).

5.5. *Other approaches with peptidomimetics*

As noted earlier above, other approaches include use of functionalized amide bonds, peptoids, urea peptidomimetics, peptide sulfonamides, oligocarbamates, azapeptides, β-peptides, and N-modified peptides [36]. Those are essentially the options in the chemistry part of the peptidomimetic strategy, but there also possible variations in overall strategy [42]. There is a comprehensive review on these lines by Vagner et al. have [43]. They take the fairly general view that peptidomimetics are compounds whose essential elements (pharmacophore) mimic a natural peptide or protein in 3D space and which retain the ability to interact with the biological target and produce the same biological effect. However, taken out of context that would be a little too general, as the same might be said of a traditional, small organic in-pill drug. Like the present author, the authors of Ref [42], and almost certainly most researchers in the pharmaceutical industry, they see the main role of peptidomimetics not as an end in itself but as a first step for in-a-pill drug discovery. They note that the design process begins by (a) exploring structure-activity relationships to define a minimal active sequence or major pharmacophore elements and identifying the key residues that are responsible for the biological effect, then (b) the researcher applies structural constraints to probe the three dimensional arrangements of these features, the peptide complexity is reduced, and the basic pharmacophore model is defined by its critical structural features in three dimensional space [36]. The present paper essentially relates to step (a) above, although it is notable that the reflect-complement-reflect method is in effect using an antibody head

synthesized from D-amino acids and could have a neutralizing effect, perhaps initially by inducing conformational changes in the spike protein, without being considered to interact at a specific pharmacophore. As to the step (b) above in which the peptidomimetic is refined, this is less necessary in the case of a *retro*-inverso peptide. That is because the kind of output presented in the present paper can be used to focus on regions in which the sidechains are exposed and so avoid the defect of the method that backbone amide and carbonyl groups are interchanged. Whereas there is always a role for further refinement by computational methods, there comes a point at which experiment by synthesis and testing (which are not usually arduous) is the best approach. With the backbone defect avoided for *retro*-inverso peptides, refinement is more likely to consist of overcoming the problem that the peptide has a higher conformational entropy than the same segment in the source protein, so reducing the binding free energy. This is a main reason for the choice of the loop in Fig. 2, as it is conformationally constrained by a disulfide bond both in the source protein and the synthetic peptide. A very recent review of peptidomimetics by Mabonga and Kappo [44] puts focus on modeling the peptidomimetic at the target binding site. This again relates largely to refinement stage (b) above. It also potentially opens up the argument that if extensive peptide modeling is the focus, then there is also the option of automatically "evolving" a peptide to fit at the required binding site [45], and in some sense this is possible by evolving from nothing at all, except to add progressively residues to the peptide which "best fit" the binding site. In that strategy, a bioinformatics approach starting from a part of the protein sequence is, in principle, less essential. Nonetheless, in the past laboratories of the present author and past collaborators, most of the successful approaches to diagnostic, vaccine and peptide design at least started from a natural sequence and a bioinformatics approach [6] even when computational chemistry was subsequently applied extensively and usefully.

6. Conclusion

Despite the comments in Section 1.1 regarding the relative lack of detailed general analysis of the proteins produced by the SARS-COV-2 genome in the understandable rush for vaccines and antiviral agents, there have been many other studies of the accessible surface including the glycosylation of the SARS-C-V-2 glycoprotein (e.g. Refs. [30–32]) in order to facilitate development of such weapons against the virus. As one may expect, SARS-CoV-2 has also stimulated some unusual approaches of this general kind [33]. In hindsight, it may emerge as one of the most intensely studied aspects of this kind, over such a short duration of time, for any specific protein up to 2020. Indeed, since the present work and first preparation of the present paper, there have been an explosive number of publications that relate to use of synthetic peptides as weapons against the COVID-19 virus. It is not the intent to give a comprehensive review, but Refs. [36–40] provide excellent introductions and reviews, as do many of the websites of many commercial enterprises selling peptides. Nonetheless, at the time of writing there do not seem to be detailed studies of peptidomimetics composed entirely of D-amino acid residues. More work with glycosylated peptides would be useful. Typically, the glycans covalently bound to the spike protein are described as a viral "shield" [32], implying a barrier to attacking the virus and hence not an "Achilles heel", although they can certainly be involved in antigenicity (and hence a consideration in the design of a vaccine) and in the binding to host proteins; it is simply that the technology for taking them into account synthetically is a little more difficult [7]. It is therefore appropriate that, as in the present study, the influence of these glycans on accessibility is reported separately.

As emphasized at several points in this paper, the computational method described was partly developed because other approaches to exposure were less convenient for design of at least of the kind of peptidomimetics based on D-amino acid residues, and for the author's approach and way of working. There was a particular incentive for "not reinventing the wheel" whenever tools already exist. That is because, as discussed in many preceding papers including those on SARS-COV-2 studies [6,9], the author's normal approach is to develop and apply automated inference technology that interfaces with available tools and data that already exist on the Internet. That is especially true for pre-existing bioinformatics and protein structure tools [46]. However, when suitable tools do not meet the criteria, new software is developed. The present approach conveniently gives required details in one step. While no claim is made that the information cannot be obtained by other means, the present method reveals fine detail conveniently located in one final output of a single approach. The one exception to that is alignment of sequences. Alignment is an important part of the approach to confirm insertions and deletions between different related proteins, highlight parts of the structure that are absent or cannot be seen in the experimental structure, and also to indicate conserved sites that are likely important to the virus and less likely to evolve to escape from vaccine, diagnostic, and pharmaceuticals [5–7]. Preexisting tools such as Clustal Omega were used as described in the preceding publications [5–9]. Probably the feature differing most from previous computational structure analysis methods is the calibration of a first shell of atoms around each sidechain, then expressed on an averaged per sidechain atom basis, and (importantly) which is then made less sensitive by normalizing the final measure against the minimum and maximum scores encountered (Eqn. (2)). However, there is a huge history of studies on exposure and accessibility than spans half a century and which have been reported in many languages, so some specific papers may have been missed. In any event, the algorithm including its displays (printouts) for the SARS-CoV-2 spike glycoprotein presented here should be helpful in designing synthetic peptide vaccines and peptidomimetics against the spike protein [5–7], even though the experimental three dimensional structures are often glycosylated or less extensively glycosylated than can occur in some animal tissues. The output should be an important consideration when seen alongside information about the extent to which subsequences are conserved [5–7]. Here they have here been shown against regions of antibody/Fab and ACE2 binding, and they may similarly be usefully displayed against linear output for conservation of the subsequences, secondary structure, and B-epitope and T-epitope character. The value of this in data mining and machine learning, and in AI-based approaches generally, should be evident.

Finally, it will be interesting to compare the present results with high grade molecular dynamics simulations of the spike protein and complexes with other molecules, and these are commencing, with attention focused by the findings of the present paper. Ultimately, nonetheless, these are not experiment but remain simulations with known limitations, i.e. the limits of pairwise potential functions, coping properly with electrostatic fields inducing polarization, and difficulty in bringing entropy to a realistic convergence. In contrast, the present paper is rooted in the beginnings of the exploding amount of experimentally derived three-dimensional knowledge concerning SARS-COV-2 spike protein structure.

Declaration of competing interest

Papers in this series are provided to the community to promote the more general applications of the thinking of Professor Paul A. M. Dirac in human and animal medicine in accordance with the charter of The Dirac Foundation, to emphasize the advantages and simplicity of the basic form of the Hyperbolic Dirac Net, to encourage its use, and to propose at least some of the principles of the associated Q-UEL, a universal exchange language for medicine, as a basis for a standard for interoperability. Those using Q-UEL knowledge algorithms and/or novel bioinformatics algorithms in regard to the study of SARS-COV-2, and synthetic vaccines and preventative or therapeutic compounds recommended by using them, are presented to the community in a similar spirit. These mathematical and engineering principles are used, amongst many others in an integrated way, in the algorithms and internal architectural features of the BioIngine.com, a distributed system developed by Ingine Inc. Cleveland, Ohio, for the mining of, and inference from, Very Big Data for commercial purposes.

References

[1] D.A. Tyrrell, M.L. Bynoe, Cultivation of viruses from a high proportion of patients with colds, Lancet 1 (1966) 76–77.

[2] D. Hamre, J.J. Procknow, A new virus isolated from the human respiratory tract, Proc Soc Exp Biol Med 121 (1966) 190–193.

[3] P.S. Masters, The molecular biology of coronaviruses, Adv. Virus Res. 66 (2006) 193–292.

[4] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Meng, J. Wang, Y. Lin, J. Yuan, Z. Xie, J. Ma, W.J. Liu, D. Wang, W. Xu, E. C. Holmes, G.F. Gao, G. Wu, W. Chen, W. Shi, W. Tan, Genomic characterization and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. www.thelancet.com. Published online January 29, 2020, https://doi.org/10.1016/S0140-6736(20)30251-30258, 2020.

[5] B. Robson, Preliminary Bioinformatics Studies on the Design of Synthetic Vaccines and Preventative Peptidomimetic Antagonists against the Wuhan Seafood Market Coronavirus. Possible Importance of the KRSFIEDLLFNKV Motif, Circulated and Published in January on ResearchGate, 2020, https://doi.org/10.13140/RG.2.2.18275.09761.

[6] B. Robson, Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus, Comput. Biol. Med. (2020). February, 103670.

[7] B. Robson, COVID-19 coronavirus spike protein analysis for synthetic vaccines, a peptidomimetic antagonist, and therapeutic drugs, and analysis of a proposed Achilles' heel conserved region to minimize probability of escape mutations and drug resistance, Comput. Biol. Med. 121 (2020) 103749. June.

[8] B. Robson, Bioinformatics studies on a function of the SARS-CoV-2 spike glycoprotein as the binding of host sialic acid glycans, Comput. Biol. Med. 122 (2020) 103849. July.

[9] B. Robson, The use of knowledge management tools in viroinformatics. Example study of a highly conserved sequence motif in Nsp 3 of SARS-CoV-2 as a therapeutic target, Comput. Biol. Med. 125 (2020) 103963. August.

[10] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche, M.A. Müller, C. Drosten, S. Pöhlmann, SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor, Cell 181 (2) (2020) 271–280, 16 April.

[11] S. Xia, M. Liu, C. Wang, et al., 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 Res. 30 (4) (2020) 343–355, https://doi.org/10.1038/s41422-020-0305-x.

[12] J.K. Millet, G.R. Whittaker, Host cell entry of Middle East respiratory syndrome coronavirus after two-step, furin-mediated activation of the spike protein, Proceedings of the National Acadamy of Sciences USA 111 (42) (2014) 15214–15219.

[13] Jaimes, A., André, M., Millet, J. K., and Whittaker, J. R. , Structural modeling of 2019-novel coronavirus (nCoV) spike protein reveals a proteolytically-sensitive activation loop as a distinguishing feature compared to SARS-CoV and related SARS-like coronaviruses, BIOREX, https://www.biorxiv.org/content/10.1101/2020.02.10.942185v1PLOS Pathogens | August 13, 2018)..

[14] J. Trylska, V. Tozzini, C.A. Chang, A.J. McCammon, HIV-1 protease substrate binding and product release pathways explored with coarse-grained molecular dynamics 92 (12) (15 June 2007) 4179–4187.

[15] H. Zhang, G. Wang, J. Li, N. Nie, X. Shi, G. Lian, W. Wang, X. Yin, Y. Zhao, X. Qu, M. Ding, H. Deng, Identification of an antigenic determinant on the S2 domain of the severe acute respiratory syndrome coronavirus spike glycoprotein capable of inducing neutralizing antibodies, J. Virol. 78 (13) (2004 Jul) 6938–6945.

[16] S. Shichijo, N. Keicho, H.T. Long, T. Quy, Phi, L.D. Ha, V.V. Ban, S. Itoyama, C.-J. Hu, N. Komatsu, T. Kirikae, F. Kirikae, S. Shirasawa, M. Kaji, T. Fukuda, M. Sata, T. Kuratsuji, K. Itoh, T. Sasazuki, Assessment of synthetic peptides of severe acute respiratory syndrome coronavirus recognized by long-lasting immunity, Tissue Antigens 64 (5) (2004 Nov) 600–607.

[17] Y. He, Y. Zhou, H. Wu, B. Luo, J. Chen, W. Li, S. Jiang, Identification of immunodominant sites on the spike protein of severe acute respiratory syndrome (SARS) coronavirus: implication for developing SARS diagnostics and vaccines, J. Immunot. 173 (6) (2004 Sep 15) 4050–4057.

[18] M. Hoffmann, H. Kleine-Weber, S. Pöhlman, A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells, Mol. Cell 78 (4) (2020) 7709–7784.

[19] Bonnin,A., Danneels,A., Dubuisson, J., Goffard, A. and Sandrine Belouzard,S., HCoV-229E spike protein fusion activation by trypsin-like serine proteases is mediated by proteolytic processing in the S2 region.

[20] B. Robson, J. Garnier, Introduction to Proteins and Protein Engineering, Elsevier Press, 1988.

[21] S. Moelbert, E. Emberly, C. Tang, Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins, Protein Sci. 3 (3) (2004) 752–762.

[22] U. Uttamkumar Samanta, R.P. Bahadur, P. Chakrabarti, Quantifying the accessible surface area of protein residues in their local environment, Protein Eng. Des. Sel. 15 (8) (2002) 659–667.

[23] B. Lee, F.M. Richards, The interpretation of protein structures: estimation of static accessibility, J. Mol. Biol. 55 (1971) 379–400.

[24] L.M. Surhone, M.T. Timpledon, S.F. Marseken, Radial Distribution Function, VDM Publishing, 2010.

[25] L.E. Canne, G.M. Figliozzi, B. Robson, M.A. Siani, R.J. Simon, N-Alkoxy amid backbone protection in BOC chemistry : improved synthesis of a difficult sequence, Protein Sci. 5 (Suppl.1) (1996), 72.

[26] M.A. Siani, L.E. Canne, R.J. Simon, G.M. Figliozzi, B. Robson, D.A. Thompson, R. J. Simon, Chemical synthesis and activity of D-superoxide dismutase, Protein Sci. 5 (Suppl.1) (1996) 72.

[27] B. Robson, Beyond Proteins Trends in Biotechnology 17 (1990) 311–315.

[28] B. Robson, Doppelganger proteins as drug leads, Nat. Biotechnol. 14 (1996) 892–893.

[29] B. Robson, Pseudoproteins: non-protein protein-like machines, The Sixth Foresight Conference on Molecular Nanotechnology (1998). https://foresight.org/Conferences/MNT6/Abstracts/Robson/index.html. (Accessed 16 September 2020).

[30] D. Zhou, X. Tian, R. Qi, C. Peng, W. Zhang, Identification of 22 N-glycosites on spike glycoprotein of SARS-CoV-2 and accessible surface glycopeptide motifs: implications for vaccination and antibody therapeutics, Glycobiology, cwaa052, June 10 (2020).

[31] A.C. Walls, Y.-J. Park, A. Tortorici, A. Wall, T.A. McGuire, D. Veesler, Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein, Cell 181 (2) (2020).

[32] N. Vankadari, J.A. Wilce, Emerging COVID-19 coronavirus: glycan shield and structure prediction of spike glycoprotein and its interaction with human CD26, Emerg. Microb. Infect. 9 (1) (2020).

[33] R.C. Penner, Conserved high free energy sites in human coronavirus spike glycoprotein backbones, J. Comput. Biol. (2020). Published Online 13 May.

[34] M.H.V. Van Regenmortel, S. Muller, D-peptides as immunogens and diagnostic reagents, Curr. Opin. Biotechnol. 9 (4) (1998) 377–382.

[35] K. Wiesehan, D. Willbold, Mirror-image phage display: aiming at the mirror, Chembiochem 4 (9) (2003) 811–815.

[36] S. VanPatten, M. He, A. Altiti, K.F. Cheng, M.H. Ghanem, Y. Al-Abed, Evidence supporting the use of peptides and peptidomimetics as potential SARS-CoV-2 (COVID-19) therapeutics, Future Med. Chem. 12 (18) (2020) 1647–1656.

[37] Z. Cormier, Lab-Made 'Miniproteins' Could Block the Coronavirus from Infecting Cells. Synthetic Peptides that Mimic Human Antibodies for COVID-19 Could Be Cheaper and Easier to Produce, Sceintific America, 2020. October, https://www.scientificamerican.com/article/lab-made-miniproteins-could-block-the-coronavirus-from-infecting-cells/.

[38] B.K. Maiti, Potential role of peptide-based antiviral therapy against SARS-CoV-2 infection, ACS Pharmacology & Translational Science 3 (4) (2020) 783–785.

[39] R. Ling, Y. Dai, B. Huang, W. Huang, J. Yu, X. Lu, Y. Jiang, In silico design of antiviral peptides targeting the spike protein of SARS-CoV-2, Peptides 130 (2020) 170328, 020, https://www.sciencedirect.com/science/article/pii/S0196978120300772.

[40] V. Khavinson, N. Linkova, A. Dyatlov, B. Kuznik, R. Umnov, Peptides: Prospects for Use in the Treatment of COVID-19, Molecules, vols. 25–04389, MDPI, 2020.

[41] G. Guichard, N. Benkirane, G. Zeder-Lutz, M.H. van Regenmortel, J.P. Briand, S. Muller, Antigenic mimicry of natural L-peptides with retro-inverso-peptidomimetics, Procedings of the National Acadamy of Sciences 91 (21) (1994) 9765–9769.
[42] A. Trabocch, A. Guarna, Peptidomimetics in Organic and Medicinal Chemistry: the Art of Transforming Peptides into Drugs, John Wiley and Sons, 2014.
[43] J. Vagner, H. Qu, V.J. Hruby, Peptidomimetics, a synthetic tool of drug discovery, Curr. Opin. Chem. Biol. 12 (3) (2008) 292–296 (2008).
[44] L. Mabonga, A.P. Kappo, Peptidomimetics: a synthetic tool for inhibiting protein–protein interactions in cancer, Int. J. Pept. Res. Therapeut. 26 (2020) 225–241.
[45] D. Frenkel, D.E. Clark, J. Li, C.W. Murray, B. Waszkowycz, B. Robson, D. R. Westhead, PRO_LIGAND: an approach to de novo molecular design. 4. Application to the design of peptides, J. Comput. Aided Mol. Des. 9 (1995) 213–225.
[46] B. Robson, Extension of the Quantum Universal Exchange Language to precision medicine and drug lead discovery. Preliminary example studies using the mitochondrial genome, Comput. Biol. Med. 117 (2020), 103621.