Received: 3 May 2021 | Revised: 5 October 2021 | Accepted: 12 October 2021

DOI: 10.1002/prot.26266

RESEARCH ARTICLE

PROTEINS WILEY

Conformational variability of loops in the SARS-CoV-2 spike protein

Samuel W. K. Wong | Zongjun Liu

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada

Correspondence
Samuel W. K. Wong, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada.
Email: samuel.wong@uwaterloo.ca

Funding information
Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: RGPIN-2019-04771

Abstract

The SARS-CoV-2 spike (S) protein facilitates viral infection, and has been the focus of many structure determination efforts. Its flexible loop regions are known to be involved in protein binding and may adopt multiple conformations. This article identifies the S protein loops and studies their conformational variability based on the available Protein Data Bank structures. While most loops had essentially one stable conformation, 17 of 44 loop regions were observed to be structurally variable with multiple substantively distinct conformations based on a cluster analysis. Loop modeling methods were then applied to the S protein loop targets, and the prediction accuracies discussed in relation to the characteristics of the conformational clusters identified. Loops with multiple conformations were found to be challenging to model based on a single structural template.

KEYWORDS
conformational ensembles, COVID-19, decoy selection, loop modeling, protein structure prediction, sequence variants

1 | INTRODUCTION

The COVID-19 disease is caused by the SARS-CoV-2 strain of coronavirus and its continued spread remains a concern since the first reported infections in late 2019.[1] The SARS-CoV-2 viral genome encodes for four main structural proteins: spike, envelope, membrane, and nucleocapsid.[2] The spike (S) protein is of particular importance as it facilitates viral entry into host cells via its receptor binding domain (RBD), which recognizes human angiotensin-converting enzyme 2 (ACE2).[3] Current vaccines being administered[4] achieve efficacy against SARS-CoV-2 by enabling the human body to produce a modified version of its S protein; this in turn induces the production of neutralizing antibodies against the disease.[5]

Toward the development of such therapeutic interventions, many structure determination efforts have focused on the S protein, with the first standalone experimental structure of the full-length S protein obtained via cryo-electron microscopy in mid-February 2020.[6] Soon thereafter, the structure of the S protein RBD bound in a complex with ACE2 was also determined.[7] As of January 13, 2021, there were 203 structures deposited in the Protein Data Bank (PDB[8]) associated with the SARS-CoV-2 S protein. These include studies of the standalone S protein,[9] the S protein interacting with potential antibodies,[10,11] and the S protein interacting with various forms of ACE2.[12] Finally, with the emergence of S protein sequence variants, structures corresponding to mutations are also being studied, with D614G being a common example.[13] While individual PDB structures generally provide static snapshots of protein conformations, it is well-known that proteins exhibit dynamic movement.[14,15] The local dynamics of atoms and residues are partially depicted via crystallographic B-factors.[16] Larger motions are also possible: for the SARS-CoV-2 S protein, a well-documented example is the ability of its RBD to adopt "up" (or open) and "down" (or closed) states, where the "up" state is the conformation capable of binding to ACE2.[6] Overall then, the PDB is a rich source of data for examining the conformational variability of the S protein, given the number of times its structure has been solved experimentally.

This article focuses on the loop conformations of the S protein. Protein loops are the flexible connecting regions between regular

secondary structures, and are where protein disorder is most likely to occur.[17] This greater disordered nature of loops may be manifest in a PDB structure via missing atomic coordinates or atoms with high B-factors.[18] Accurate structure prediction for loops is both challenging and necessary, to construct useful models for downstream therapeutic applications.[19] Loops are of particular importance as they are often associated with protein function, such as providing binding recognition sites and facilitating protein–protein interactions.[20] For example, an extended loop of the SARS-CoV-2 S protein RBD interacts directly with loops of ACE2, as evidenced by the PDB structure of the RBD-ACE2 complex.[21] Dynamic structural changes can occur both in larger regions of a protein (e.g., the SARS-CoV-2 RBD), as well as in individual loops adopting conformational rearrangements to carry out protein function in accordance with their environment.[22] Thus, when a protein has been solved many times in the PDB, we may be able to observe distinct conformations among some of its loops, given their potential for disorder and structural variability. In particular for the SARS-CoV-2 S protein, the PDB also documents sequence variants arising from mutations to some of its loop regions,[23] and the possible structural impacts of mutations can also be studied more broadly via computational methods.[24–26] Mutations to the S protein are especially of concern as they can lead to more infectious variants of SARS-CoV-2.[27]

The task of structure prediction for flexible loops with multiple distinct conformations has been found to be more challenging than for rigid or inflexible ones.[28] Most loop prediction methods are designed to identify the most likely conformation, for example, with the lowest potential energy.[29–34] Such methods are typically trained on loop sets where a single conformation for each loop is taken from the PDB and assumed to represent the ground truth,[35] and thus tend to be more successful at accurately predicting inflexible loops with one "correct" solution. Accuracy is typically measured by computing the root-mean-squared deviation (RMSD) of the backbone atoms from the predicted loop conformation to the corresponding one in the PDB. In order to study loops that can adopt multiple conformations, prediction methods might instead be applied to generate an ensemble of decoys, which often involves a combination of sampling and scoring steps.[36] Then, the success of different methods could be assessed on the basis of whether their generated ensembles include decoys that are close to each of the known conformations.[28] For the SARS-CoV-2 S protein, this kind of assessment is a good test on the ability of current methods to explore a range of likely conformations, especially if further mutations were to occur in the flexible loop regions.

These considerations motivate the main contributions of this article. First, we identify the loop regions and sequence variants from the known PDB structures of the SARS-CoV-2 S protein, and use cluster analysis to classify each loop according to whether it has been observed to adopt multiple distinct conformations or a single conformation only. Second, we apply four current loop prediction methods on the identified loop regions, to generate ensembles of decoys for each one. Third, we discuss the results of these methods and the effectiveness of their application to modeling the loops of the S protein, along with the insights gained via our analyses.

2 | MATERIALS AND METHODS

2.1 | Data preparation and selection of loop targets

The 3-D structures of the SARS-CoV-2 S protein were downloaded from the PDB at the RCSB website (https://rcsb.org) on January 13, 2021, by navigating to the page in the "COVID-19 coronavirus resources" section entitled "Spike proteins and receptor binding domains." We extracted the S protein structures that are not bound to other molecules and have sequence length greater than 1000. This facilitates study of the S protein loop conformations within the context of a (mostly) full-length S protein structure, while without explicit interaction with other proteins. A total of 63 S protein PDB structures satisfied these criteria, most of which are provided as S protein trimers. We treated each chain as an individual sample and thus extracted a total of 193 S protein chains. Some realignments of the corresponding amino acid sequences were required in order to keep the residue numbers consistent across all chains; this was accomplished with the ClustalO service in Jalview.[37]

For each S protein chain, we first used DSSP[38] to determine the secondary structure classification of each residue. The eight-state DSSP classification was reduced to the traditional three types of helix (H), sheet (E), and coil (C) following the conventions in the SPIDER3[39] secondary structure prediction method: we map DSSP's "G," "H," and "I" to H; "E" and "B" to E; the remaining three states are mapped to C. Due to structural variability, the classified type (H, E, or C) for a given residue position may not always agree among the 193 S protein chains. Thus, we define a loop region for our study as follows: a segment of five or more consecutive residues where over 50% of the protein chains at each position are classified as type C. Further, if two such segments are separated by only one E or H type residue (i.e., where less than 50% of the chains are type C at that position), we treat the two combined segments (including that connecting residue) as a single loop region.

With the starting and ending positions of loops defined in this manner, we check for the presence of sequence variants in each loop region among the S protein chains. If multiple distinct residue sequences are observed for a loop region, we shall treat each unique sequence separately for further analysis. This allows us to document the possible impact of mutations on the loop conformations. Thus, we shall say that a loop instance consists of its starting and ending positions together with its unique residue sequence. We then consider the structural variability of each loop instance. To account for the potential disordered nature and structural uncertainties of loops, we extract both the atomic coordinates and B-factors from the PDB chains. Taking all chains that have no missing coordinates or B-factors within the loop residues, we compute their pairwise RMSD matrix based on the loop's backbone (N, C_α, C, and O) atoms. The RMSD calculation is applied after the backbone atoms of the loop residues for each pair are optimally superimposed using the Kabsch algorithm.[40] This is the "local RMSD"[41,42] that compares the loop region only, and so is not sensitive to orientation differences in the rest of the

structure. Based on that distance matrix, we apply hierarchical clustering with average linkage (UPGMA[43]) and a distance cutoff of 1.5 Å[28] to form initial clusters of loop conformations.

Following, we incorporate B-factors to ensure that the clusters formed are statistically distinct. Recall that the B-factor can be expressed in terms of the mean-square amplitude of atomic oscillations u^2 around their measured positions: $B = 8\pi^2 \langle u^2 \rangle$. Using an isotropic Gaussian approximation for the corresponding coordinate uncertainties, we can determine whether the difference in backbone coordinates between a loop pair is significantly different with 95% confidence (see Appendix A for details). If none of the chains in one cluster are significantly different from any chains in another cluster, we merge them into a single cluster. Clusters composed entirely of chains with poor structure resolution (>3 Å) after this step are removed from further analysis as the atomic coordinates are unlikely to be sufficiently reliable for making detailed structural comparisons. Each remaining cluster then represents a distinct group of S protein chains which have a similar conformation for that loop instance. We consider a loop instance to have multiple distinct conformations if this analysis results in two or more such clusters of conformations; otherwise, we say that loop instance essentially adopts only a single conformation. We select a representative from each cluster by taking the chain with resolution ≤3 Å that is closest to the geometric centroid of the cluster.

Our full list of S protein loop targets for study thus consists of all the cluster representatives obtained from the above steps.

2.2 | Loop modeling methods

To study the conformational variability of the identified S protein loop targets, we make use of several loop modeling methods. We focus on methods that incorporate sampling-based techniques for loop construction, which are suitable for stochastically generating an ensemble of decoys that represent plausible conformations for a loop. We include Rosetta's next-generation KIC (NGK) algorithm,[32] the DiSGro algorithm,[33] and the PETALS algorithm,[34] which are ab initio methods that explore the conformational space with the guidance of an energy or scoring function; these do not directly make use of any structure templates of known loop conformations. We also include the Sphinx algorithm,[30] which is a hybrid method that begins with loop structure fragments obtained from sequence alignment and then completes the loop construction by ab initio sampling.

Using each of the methods, we generate an ensemble of 500 decoys for each loop target. The input (or template) structure is the loop target's representative PDB chain, prepared by removing the coordinates of the loop residues: following loop modeling conventions, we treat the backbone atoms from the starting residue's C atom to the ending residue's C_α atom as unknown. The generated decoys are compared with the loop structures from each known conformation for that loop region. The backbone RMSD is used to assess the accuracy of the decoys. Two types of RMSDs are calculated, as in Choi and Deane[41]: local RMSD (which superimposes the backbone of the loop residues, as in Section 2.1) and global RMSD, which superimposes the backbone atoms of the two residues on either side of the loop (rather than the backbone of the loop residues themselves) prior to the calculation. Global RMSD, as often reported in loop modeling studies, also considers the decoy's orientation to the rest of the structure. For loop regions with multiple conformations or mutations, decoy generation is carried out multiple times, once using each representative PDB as input; taken together, we may thus assess whether decoys generated from different PDB inputs have good coverage of the conformational space for that loop region.

The scoring function associated with each method provides a ranking of its 500 generated decoys for a loop target. Thus, it is of interest to assess how well each method's top-ranking decoys can predict the possible conformations of the loop region. We use three RMSD statistics for this purpose: (a) lowest RMSD among the 500 decoys, (b) RMSD of the top-ranked decoy, and (c) lowest RMSD among the top-five ranked decoys. The first RMSD statistic evaluates the method according to its ability to construct native-like conformations, without regard to whether its scoring function can select the best prediction. The second RMSD statistic corresponds to typical loop modeling assessment, where the top-ranked decoy is selected as the prediction. However, this approach of selecting a single prediction would be less informative if the loop region has multiple conformations. Thus, we also use the third RMSD statistic: by selecting multiple (i.e., the top five) decoys, we can examine whether these top-ranking decoys are structurally distinct and accurately represent the different known conformations.

We briefly describe how each of the loop modeling methods is run. The NGK algorithm[32] is included in the Rosetta protein modeling suite (available at https://www.rosettacommons.org/), and we used the version provided in Rosetta release 2020.50 on December 18, 2020. NGK improves on a previous kinematic closure method, which consists of local conformational sampling and Monte Carlo minimization steps performed over two (coarse and full-atom) stages. The program outputs the lowest energy loop structure found in each run, and so to obtain the desired ensemble of decoys we ran the program 500 times, following the recommended settings in the online guide (https://guybrush.ucsf.edu/benchmarks/benchmarks/loop_modeling). The DiSGro algorithm[33] uses a distance-guided sequential chain-growth method to stochastically sample loop structures. We ran the authors' program to generate 100 000 conformations for the best possible coverage of the conformational space, then used their scoring function to select the 500 decoys with the lowest energy. The PETALS algorithm[34] uses a sequence of propagation and filtering steps to explore the conformational space and locate low-energy structures. We ran the authors' program with 60 000 seeds and outputted 30 000 decoys, then used an updated scoring function to select the 500 top-ranked decoys, see Appendix B for details. The Sphinx algorithm[30] begins by searching a database for suitable fragments according to loop sequence alignments; loop decoy backbones are then constructed by sampling and ranked with a coarse-grained energy function, after which side chains are added and SOAP-Loop[44] is used to obtain the final ranking of decoys. Sphinx is hosted on the

SAbPred server,[45] for which we automated the loop target submissions and used the "general protein" option; no PDB blacklist was necessary as the fragment database had not yet been updated to contain any COVID-19 S protein structures.

3 | RESULTS AND DISCUSSION

3.1 | Loop targets of the SARS-CoV-2 S protein

Applying the procedures in Section 2.1 to the 193 standalone S protein chains, a total of 44 loop regions were identified in the SARS-CoV-2 S protein. Their starting and ending residue positions are listed in the first column of Table 1. Then, 32 of the 44 loops lie within the S1 subunit, with 13 in the N-terminal domain and 11 in the RBD; for example, loops 475–487 and 495–506 have been previously noted to form contacts with ACE2 during binding.[46] Loop sequences are shown in the second column of Table 1. There are five loop regions with sequence variants in the PDB: 380–394, 410–416, 600–608, 614–620, and 891–897. For these loop regions, the most common variant in the PDB is shown first, followed by the other variants which have their mutated residue indicated in bold. The mutation that has received the most attention thus far is D614G.[13,47,48] In total, there are 50 loop instances, that is, the combination of a loop's residue positions and unique amino acid sequence. The third column of Table 1 shows the number of PDB chains that contain a complete backbone (i.e., atomic coordinates and B-factors) for each loop instance.

The final column lists the representative PDB chains for each loop instance, obtained by the procedure for constructing clusters as described in Section 2.1. Thus, for example, there are 180 S protein chains that contain the loop at positions 329–338; clustering by pairwise RMSD identified two distinct conformations among structures with resolution ≤3 Å; 6x29A and 7kdkC were chosen to represent these clusters (which included 155 and 21 chains respectively), being the chains with resolution ≤3 Å closest to the cluster centroids. We illustrate the 329–338 loop example in the top panels of Figure 1: a histogram of all pairwise RMSDs of the loop backbone (among the 180 S protein chains that contain this loop) is shown on the left, while a close-up of the part of the S protein chain containing the loop is shown on the right. The histogram shows distinct peaks at pairwise loop RMSDs of 0.4–0.6 Å and 2.0–2.4 Å, from which clustering identified the two distinct conformations colored dark blue and turquoise. In contrast, the bottom panels of Figure 1 show another length 10 loop region (555–564) but with little structural variability: the pairwise RMSDs do not exceed around 1.5 Å and clustering identified just one main conformation (colored in red).

The initial hierarchical clustering step resulted in 137 clusters for the 50 loop instances. Based on the B-factor calculations, 17 of the 137 clusters did not have statistically distinct atomic coordinates compared to other clusters, and so merging these resulted in 120 clusters. All of the 17 clusters being merged had also failed to contain structures with sufficient resolution (≤3 Å). A further 45 of the 120 clusters contained no ≤3 Å structures, which led to two of the loop instances being omitted: 66–83 and 600–608 with the Q607E mutation. The final 75 clusters thus covered 48 loop instances; 17 of the 48 had multiple distinct conformations (ranging from 2 to 5). By choosing the centroid of each cluster as its representative conformation, a diverse set of 41 different PDB chains with ≤3 Å resolution can be seen in Table 1. It should be noted that the exact number and composition of clusters will depend on the algorithm (i.e., cutoff and criterion) chosen. Here, using a cutoff of 1.5 Å with UPGMA, the average RMSD between members of different clusters will be at least 1.5 Å. For example, if we used a cutoff of 1.5 Å with WPGMA[43] instead, 42 of the 50 loop instances maintain the same final clustering results; WPGMA would have found 82 representative conformations for the 48 loop instances. Overall, we consider the clusters in Table 1 to provide a fairly stable characterization of the structural variability present in these loops.

The final 75 clusters in Table 1 differ in their size and within-cluster variation. There were 4 singleton clusters (defined by a single chain only), and 61 clusters were defined by at least four chains and two distinct PDB codes (and often significantly more). These high chain counts per cluster enable more cluster statistics to be examined, compared to related studies, for example, Marks et al.[28] where clusters were defined by at most five chains (except in one case). Here, loop instances with multiple conformations tend have a dominant cluster that is defined by at least two-thirds of the available chains; the one exception is 841–848, which is also the most structurally variable loop with five distinct clusters. For each of the 61 well-represented clusters, we computed the average within-cluster RMSD (i.e., between all pairs of members in that cluster) as a measure of its breadth of movement, and a histogram is shown in Figure 2. The average breadth over all 61 clusters is 0.72 Å. The list of clusters grouped according to their breadth d is shown in Table 2, where 16 clusters are fairly tight with $d \le 0.5$ Å, 36 clusters have $0.5 < d \le 1.0$, and the 10 loosest clusters have $d > 1.0$ Å. It might be expected that shorter loops tend to form tighter clusters as they have a smaller conformational space; indeed, this pattern can be seen as the average loop length of clusters in these three groups are 6.5, 12.1, and 13.0 respectively. The larger clusters also tend to be tighter: the average cluster size in these three groups are 127, 108, and 49, respectively. However, we note that these are overall patterns only; for example, the cluster for the longest loop 783–816 is defined by 142 chains and has only a moderate $d = 0.81$.

It is well-known that the SARS-CoV-2 RBD as a whole can adopt an "up" or "down" conformational state.[6] Here, 7 of the 17 loop instances with multiple conformations were located within the RBD. Notably, both 475–487 and 495–506 which interact with ACE2 are among these. Thus, we examined whether this higher propensity for multiple conformation loops within the RBD might be associated with the chains having an "up" or "down" RBD state, even when the S protein chain is considered in isolation. We took PDB 6zge,[49] where it is known that chain A has a "down" RBD and chain B has an "up" RBD. Then, each of the 193 S protein chains was classified as "up" or "down" according to whether its backbone RMSD to 6zgeB or 6zgeA was smaller. Based on this criterion, the loop at 370–375 has both

TABLE 1 SARS-CoV-2 S protein loops. The first column shows the starting and ending positions of each identified loop region. The second column shows the loop sequences; if there are sequence variants in the PDB, the most common variant is listed first, and other variants have their mutated residues marked in bold. The number of PDB chains containing that loop instance is shown in the third column. The rightmost column lists the representative PDB chains for each loop instance; if a loop instance has multiple conformations, each chain listed corresponds to one distinct conformation (cluster). The number of PDB chains represented by each cluster is shown in parentheses; these may not sum up to the third column since clusters with poor structure resolution (all chains >3 Å) are omitted

Region	Sequence	#Chains	Representative conformations
14–27	QCVNLTTRTQLPPA	36	6zgeA(24), 7dddC(12)
31–46	SFTRGVYYPDKVFRSS	185	7a4nB(185)
56–60	LPFFS	185	6xr8A(185)
66–83	HAIHVSGTNGTKRFDNPV	11	none (all PDBs >3 Å resolution)
108–116	TTLDSKTQS	169	6zoxB(169)
130–140	VCEFQFCNDPF	168	6xluB(145), 7kdkC(5), 7kdlA(4)
146–168	HKNNKSWMESEFRVYSSANNCTF	38	6zgiB(27), 7dddC(9)
172–187	SQPFLMDLEGKQGNFK	52	7df3B(39), 6zp0B(12)
210–222	INLVRDLPQGFSA	154	6vxxA(152)
230–236	PIGINIT	185	6vxxA(185)
245–263	HRSYLTPGDSSSGWTAGAA	26	6zgiB(24)
280–284	NENGT	185	6x79B(185)
304–310	KSFTVEK	185	7a4nB(185)
320–324	VQPTE	185	6zoxC(181), 6xm3A(4)
329–338	FPNITNLCPF	180	6x29A(155), 7kdlB(21)
343–348	NATRFA	181	6zgeC(181)
370–375	NSASFS	182	6vxxA(139), 6zgiC(42)
380–394	YGVSPTKLNDLCFTN	170	7kdlC(164)
	YGV**C**PTKLNDLCFTN	12	6x79B(12)
410–416	IAPGQTG	179	7kdkA(178)
	IAP**C**QTG	3	6zoxB(3)
422–430	NYKLPDDFT	182	6xr8B(178), 6xm0B(2)
438–451	SNNLDSKVGGNYNY	93	6xr8A(85), 7kdlB(4)
454–472	RLFRKSNLKPFERDISTEI	96	6zgeC(95)
475–487	AGSTPCNGVEGFN	92	7dddA(87), 6xm0B(1)
495–506	YGFQPTNGVGYQ	124	6zp0A(118), 6xm0B(2), 7kdlB(3)
517–523	LLHAPAT	168	6zoxA(163), 6xm0A(2), 6xm0B(1), 6xm3A(2)
526–537	GPKKSTNLVKNK	181	7ad1B(26), 6x29B(154)
555–564	SNKKFLPFQQ	185	7kdkC(185)
578–583	DPQTLE	185	6zoxB(185)
600–608	PGTNTSNQV	170	7kdlA(169)
	PGTNTSN**E**V	12	none (all PDBs >3 Å resolution)
614–620	DVNCTEV	103	6xm4C(98)
	GVNCTEV	42	7kdkA(42)
	NVNCTEV	6	7a4nB(6)
624–641	IHADQLTPTWRVYSTGSN	26	6xm0B(18)
656–663	VNNSYECD	185	7kdkB(185)
697–710	MSLGAENSVAYSNN	185	6vxxB(185)
783–816	AQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRS	144	6zp0C(142)
825–836	KVTLADAGFIKQ	39	6xluB(2), 6xm3B(5), 6xm3C(1), 6zgiA(25)
841–848	LGDIAARD	43	6xluC(6), 6xm4B(1), 6zgeB(20), 6xm3B(6), 7dddB(6)

(Continues)

TABLE 1 (Continued)

Region	Sequence	#Chains	Representative conformations
862–866	PPLLT	185	6zoxB(185)
891–897	GAALQIP	176	7kdkB(176)
	GPALQIP	9	7a4nB(9)
908–913	GIGVTQ	185	7a4nB(185)
968–976	SNFGAISSV	188	6zp0C(185), 6xraC(3)
1033–1046	VLGQSKRVDFCGKG	188	7kdkA(188)
1106–1112	QRNFYEP	188	7kdkC(188)
1124–1132	GNCDVVIGI	188	6xm0A(185), 6xraC(3)
1135–1141	NTVYDPL	161	7kdkB(158), 6xraC(3)

Abbreviation: PDB, Protein Data Bank.

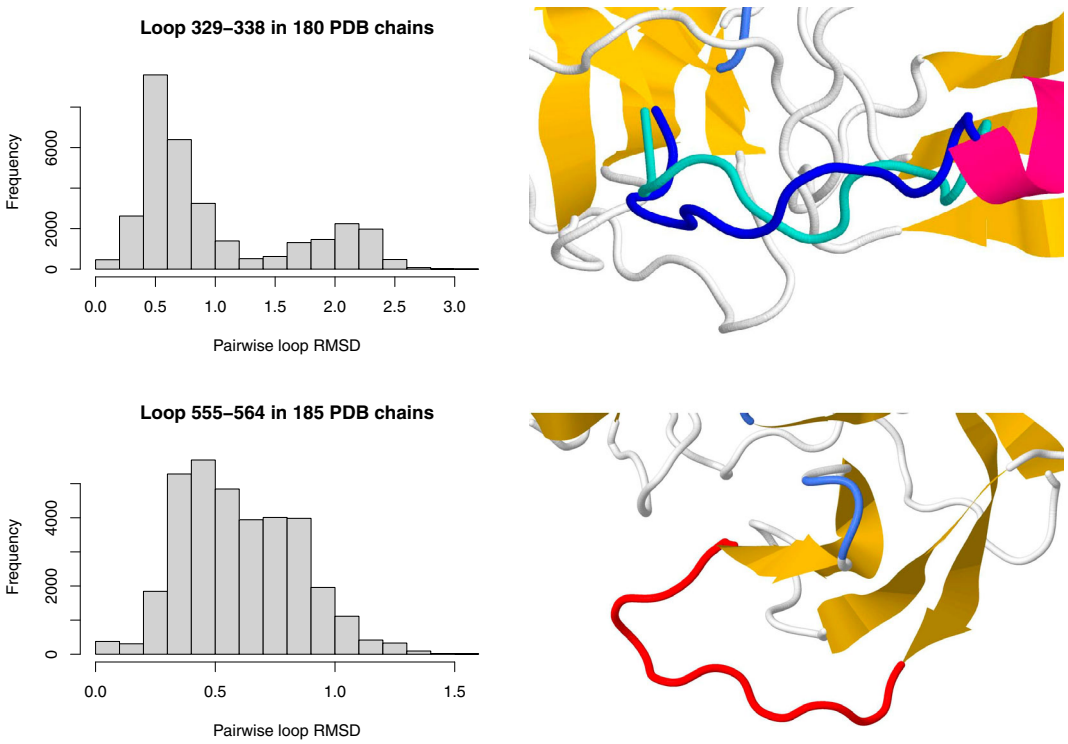

FIGURE 1 Two examples of SARS-CoV-2 S protein loops of length 10: 329–338 (top panels) and 555–564 (bottom panels). The histograms (left panels) show the pairwise root-mean-squared deviations (RMSDs) of the loop backbone among all S protein chains containing that loop: it can be seen that 329–338 exhibits higher structural variability than 555–564, due to the presence of two distinct clusters. The right panels display close-ups of the representative loop conformations: 329–338 has two distinct conformations, colored in dark blue and turquoise; 555–564 has essentially one conformation, colored in red

distinct conformations coming from "down" RBD chains, while four other loops with two conformations (329–338, 422–430, 438–451, 475–487) indeed have one conformation associated with the "up" state and the other associated with the "down" state. Of the two remaining loops, 495–506 has one conformation from a "down" RBD and two from an "up" RBD, while 517–523 has two conformations from each. Overall then, five RBD loop regions have structures that do not vary significantly with the RBD state (370–375 and the four single conformation loops in the RBD), while the other six do potentially vary.

Five loop regions had sequence variants present in the PDB, each consisting of a single point mutation. All of these loop instances had only a single conformation. Taking the representative chain for each sequence variant listed in Table 1, we computed the local loop backbone RMSD between the representatives and the results are shown in Table 3. For example, for the loop region 380–394, the sequence variants are S and C at position 383, represented by 7kdlC and 6x79B respectively; these structures have backbone RMSD 0.54 Å computed on the loop residues. For the loop 600–608, there were no high-resolution PDB structures containing the

Q607E mutation. Overall, these sequence variants do not have large impacts on the loop conformations with observed backbone differences all <1 Å, such that the conformational space of these loop regions (including variants) could be represented by a single cluster.

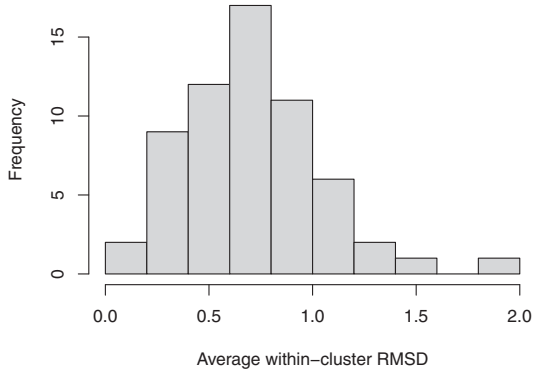

FIGURE 2 The amount of within-cluster variation for the 61 clusters defined by at least four chains and two distinct Protein Data Bank (PDB) codes. The breadth of movement observed within a cluster is measured by its average within-cluster root-mean-squared deviation (RMSD); 36 of the clusters have an average between 0.5 and 1 Å

Three of the loop targets were omitted from consideration for loop modeling, as all of their PDB chains were missing a residue immediately next to the loop: 14–27 (both conformations missing residue 13), 614–620 with the D614G and D614N mutations (both missing residue 621). Thus, the loop modeling methods were applied to a total of 71 targets.

3.2 | Loop modeling results

The four methods described in Section 2.2 were applied to model the conformations of the 71 loop targets identified in Section 3.1. Of these, 66 targets could be run successfully using all four methods. NGK and PETALS completed decoy generation for all 71 targets, while DiSGro completed 68 targets and Sphinx completed 66 targets. We focus the discussion on the results of the 66 loop targets for which all the methods could successfully generate decoys; the 5 remaining cases are discussed briefly at the end.

First, we assess the ability of methods to predict a correct loop structure. We define this loop prediction accuracy by calculating the

TABLE 2 Clusters grouped according to their breadth of movement *d* as defined by their average within-cluster RMSDs. Each cluster is listed based on its representative conformation (Table 1) together with its starting and ending residues. The average loop length and size of clusters in the three groups are shown in the rightmost columns

Breadth (*d*)	Clusters	Avg. length	Avg. size
$d \le 0.5$ Å	6xr8A_56_60, 6x79B_280_284, 7a4nB_304_310, 6zoxC_320_324, 6xm3A_320_324, 7kdkA_614_620, 7a4nB_614_620, 7kdkB_656_663, 7dddB_841_848, 6zoxB_862_866, 7kdkB_891_897, 7a4nB_891_897, 7a4nB_908_913, 6zp0C_968_976, 7kdkC_1106_1112	6.5	127
$0.5 < d \le 1.0$ Å	6zgeA_14_27, 7a4nB_31_46, 6zoxB_108_116, 6xluB_130_140, 7kdlA_130_140, 7dddC_146_168, 7df3B_172_187, 6zp0B_172_187, 6vxxA_230_236, 6x29A_329_338, 7kdlB_329_338, 6zgeC_343_348, 6vxxA_370_375, 6zgiC_370_375, 7kdlC_380_394, 6x79B_380_394, 7kdkA_410_416, 6xr8B_422_430, 6xr8A_438_451, 6zgeC_454_472, 6zp0A_495_506, 7ad1B_526_537, 6x29B_526_537, 7kdkC_555_564, 6zoxB_578_583, 7kdlA_600_608, 6xm4C_614_620, 6xm0B_624_641, 6vxxB_697_710, 6zp0C_783_816, 6xm3B_825_836, 6zgiA_825_836, 6zgeB_841_848, 7kdkA_1033_1046, 6xm0A_1124_1132, 7kdkB_1135_1141	12.1	108
$d > 1.0$ Å	7dddC_14_27, 7kdkC_130_140, 6zgiB_146_168, 6vxxA_210_222, 6zgiB_245_263, 7kdlB_438_451, 7dddA_475_487, 6zoxA_517_523, 6xluC_841_848, 6xm3B_841_848	13.0	49

Abbreviation: RMSD, root-mean-squared deviation.

TABLE 3 Backbone RMSDs between the PDB chains representing the different sequence variants, in loop regions where mutations are present. Local RMSDs are computed on the loop residues. The residues that differ between the sequence variants are highlighted in bold

Region	Sequence 1	Sequence 2	RMSD
380–394	YGV**S**PTKLNDLCFTN	YGV**C**PTKLNDLCFTN	0.54
410–416	IAP**G**QTG	IAP**C**QTG	0.40
614–620	**D**VNCTEV	**G**VNCTEV	0.67
614–620	**D**VNCTEV	**N**VNCTEV	0.62
614–620	**G**VNCTEV	**N**VNCTEV	0.51
891–897	G**A**ALQIP	G**P**ALQIP	0.23

Abbreviations: PDB, Protein Data Bank; RMSD, root-mean-squared deviation.

RMSD to the closest loop structure among all chains containing that loop instance. Thus, for this task, a good prediction can be close to *any* cluster member among *any* of the loop's known conformations (clusters), which accounts for the possible within-cluster variation (Figure 2) and treats loop structures in all the chains as an equi-energetic ensemble. Loop targets representing regions with multiple conformations can score well by this definition as long as a method can predict any one of the known conformations. For example, there are three targets for the loop 130–140 corresponding to its three conformations, represented by 6xluB, 7kdkC, and 7kdlA; decoys generated using 6xluB as input are compared to loop structures in all 154 chains of the three clusters combined, and likewise for 7kdkC and 7kdlA. We categorized the targets according to whether they belong to loop instances with multiple conformations or not; these categories are denoted as "Multiple conf." and "Single conf." in Table 4, containing 40 and 26 loop targets, respectively. Table 4 displays the three RMSD statistics described in the Section 2—lowest RMSD among the 500 decoys, RMSD of the top-ranked decoy, and lowest RMSD among the top-five ranked decoys—using both local and global RMSD calculations and averaged over the loop targets for each method. On average, all four methods can generate decoys at <1 Å local RMSD and <1.5 Å global RMSD from a correct structure. However, it remains difficult to correctly rank the generated decoys, with the RMSDs of the top-ranked decoy often substantially higher than the best decoy available. When each method is allowed to choose five decoys, then it is more likely that at least one of the five is close to a correct structure; for example, NGK's average accuracy improves from 2.31 to 1.60 Å (global RMSD). Further, the difficulty of the loop prediction task tends to vary by target category: for all four methods, the average top decoy RMSD for loops with multiple conformations are higher than for single conformation loops, whether considering local or global RMSDs.

To visualize these results, the global RMSD of the top decoy is plotted against loop length for each method in Figure 3. It is clear that the prediction difficulty and the variance of prediction RMSDs tend to increase with loop length, with methods consistently achieving <2 Å RMSD accuracy only for the shortest loops (≤6 residues). This is sensible since the size of the conformational space increases with loop length, with long loops (>12 residues) often posing a challenge for methods to sample adequately.[50] The plots also indicate that hardest targets for a given loop length tend to be those from multiple conformations, especially for the two most accurate methods (NGK and PETALS). The average lengths of loop targets in the "Single conf." and "Multiple conf." categories are similar (9.7 vs. 10.0 residues). The detailed results for each target individually are given in Table S1 of the Supporting Information.

If one is allowed to select the best prediction among *all* targets for a loop instance, then results for loops with multiple conformations improve dramatically (e.g., taking the lowest RMSD of all decoys generated from 6xluB, 7kdkC, and 7kdlA together as the result for the loop 130–140); the average global RMSD for the top decoy in multiple conformation loops decreases to just 1.05 Å for NGK and 1.74 Å for PETALS. However, this is generally not a realistic scenario in practice, as often just a single template would be available for constructing predictions. In this sense, our findings on the difficulty of predicting multiple conformation loops are less categorical compared to Marks et al.[28] for the targets in this S protein dataset. For these S protein targets, multiple conformation loops are more difficult to predict when a single template is used, but not when we can choose the best prediction among all available templates; for the dataset considered by Marks et al.,[28] the difficulty still remained when choosing the best prediction among all available templates, albeit accounting for less possible within-cluster variation as their clusters had much less representation in the PDB.

In addition to loop length, we also examine whether the cluster characteristics, namely their size (as measured by the number of chains) and breadth (as measured by the average within-cluster RMSD in Figure 2), are associated with prediction difficulty. For each

TABLE 4 RMSD metrics for assessing the loop prediction accuracy of the four methods. The loop backbone RMSDs shown are averaged over single conformation targets ($n = 26$), multiple conformation targets ($n = 40$), and all targets ($n = 66$). The columns "Min.," "Top," and "Top-5" refer, respectively, to the lowest RMSD among the 500 decoys, RMSD of the top-ranked decoy, and lowest RMSD among the top-five ranked decoys. Prediction accuracy is defined as the RMSD to the closest loop structure among all chains containing that loop instance

		Local RMSD			Global RMSD		
Method	Target category	Min.	Top	Top-5	Min.	Top	Top-5
DiSGro	Single conf.	0.76	1.81	1.28	0.97	2.66	1.73
	Multiple conf.	0.96	1.95	1.56	1.47	3.60	2.95
	All	0.88	1.90	1.45	1.27	3.23	2.47
NGK	Single conf.	0.42	1.06	0.85	0.58	1.93	1.62
	Multiple conf.	0.66	1.42	1.08	1.07	2.55	1.59
	All	0.56	1.28	0.99	0.87	2.31	1.60
PETALS	Single conf.	0.68	1.24	0.98	0.98	2.06	1.51
	Multiple conf.	0.85	1.58	1.33	1.42	3.00	2.32
	All	0.78	1.44	1.19	1.25	2.63	2.00
Sphinx	Single conf.	0.64	1.49	1.15	1.11	2.75	2.09
	Multiple conf.	0.74	1.77	1.31	1.34	3.53	2.46
	All	0.70	1.66	1.25	1.25	3.22	2.31

Abbreviations: PDB, Protein Data Bank; RMSD, root-mean-squared deviation.

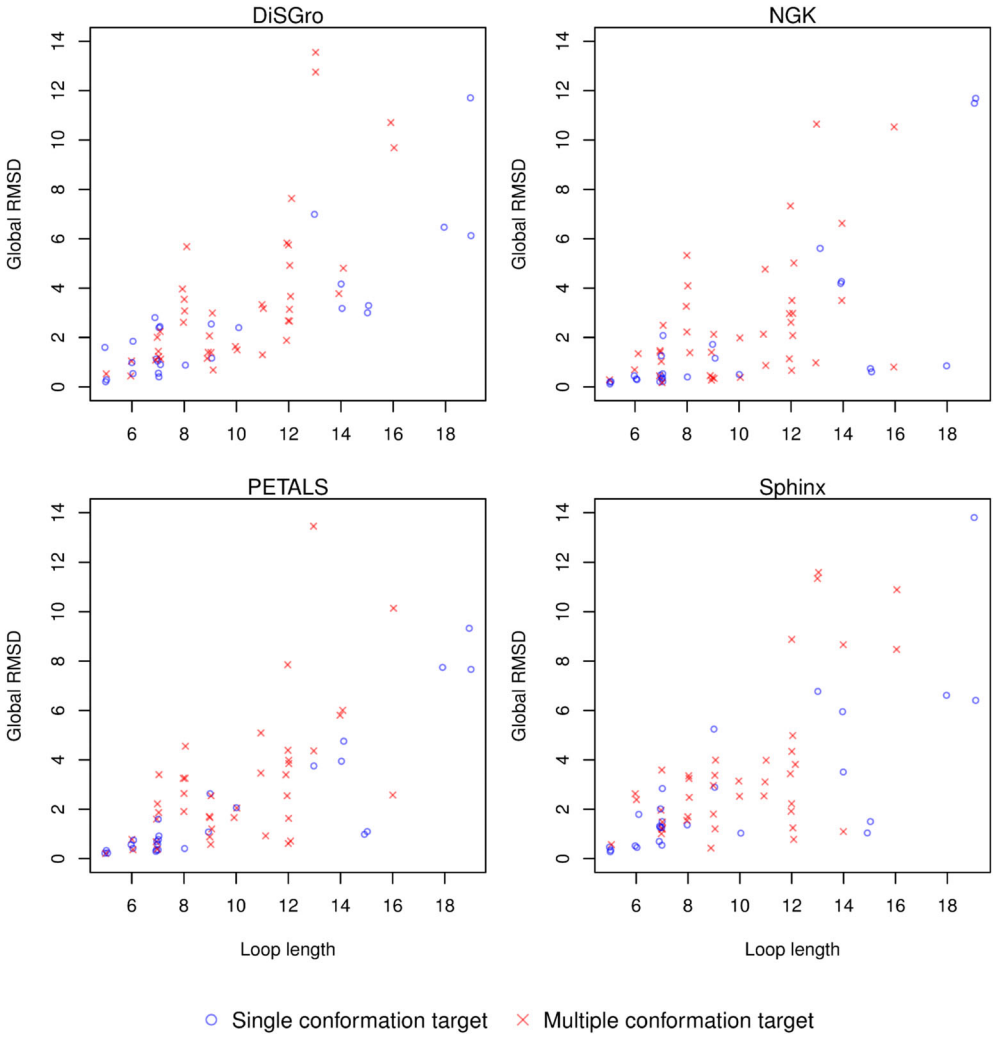

FIGURE 3 Loop prediction accuracy for each of the four methods, visualized by plotting the global root-mean-squared deviation (RMSD) of the top decoy versus loop length. Prediction difficulty increases with loop length, with methods consistently achieving <2 Å RMSD only for the shortest loops (≤6 residues). The hardest targets for a given loop length tend to be those from multiple conformations, especially for the two most accurate methods (NGK and PETALS). Slight jitter is added along the x-axis to the points for readability

method, we consider a target to be successfully predicted if the top decoy has a global RMSD of <2 Å, and to be a failure otherwise. Based on this criterion, DiSGro, NGK, PETALS, and Sphinx had 25 (48%), 35 (67%), 31 (60%), and 25 (48%) successes, respectively, out of the 52 loop targets representing conformational clusters defined by at least four chains and two distinct PDB codes. We use the Welch *t* test to provide a simple assessment of whether the mean of each variable is significantly different between successes and failures, and the results are shown in Table 5 for the four methods. The sign of the *t*-statistic indicates whether successes (positive *t*-statistic) or failures (negative *t*-statistic) are associated with larger values of that variable; for example, the *t*-statistics for loop length are all negative, so successes are associated with shorter loop lengths as expected from Figure 3. Each of the three variables is significantly associated with prediction success ($p < .01$ for all tests, except cluster size for the Sphinx method with $p = .011$). Targets with longer loop lengths, smaller cluster sizes, and larger cluster breadths tend to be more difficult to predict successfully, regardless of which loop modeling method is used.

Next, we focus on the loop instances with multiple distinct conformations, to assess how well the decoys generated from a specific PDB input can represent *all* the known conformations for that loop instance. Taking the loop 130–140, for example: the decoys generated using 6xluB are compared to the loop structures in the clusters represented by 6xluB, 7kdkC, 7kdlA, and the RMSD to the closest structure in each cluster is recorded; the average of the RMSDs to these three clusters then provides an overall result for 6xluB; the same is done using the decoys from 7kdkC and 7dklA. The results are summarized in Table 6 using the same RMSD metrics, averaged over the targets in the multiple conformation categories. This task is noticeably more challenging than the prior prediction task, as evidenced by RMSDs in Table 6 which are all larger than the corresponding values in the "Multiple conf." rows of Table 4 for all four methods. While the top decoy RMSDs are expected to increase relative to Table 4, a substantial increase still occurs when taking the entire decoy set ("Min." column, e.g., 1.07–2.18 Å global RMSD for NGK) and when allowing methods to choose the top five decoys ("Top-5" column, e.g., 1.59–2.85 Å global RMSD for NGK), whether considering local or global RMSD. This suggests that building the loop using the atomic environment of a single structural template may preclude the methods from being able to locate and predict all the possible loop conformations; Marks et al.[28] observed a similar

TABLE 5 Comparing prediction successes and failures of the four methods, according to loop length, cluster size, and cluster breadth. Prediction success is defined as a global RMSD of <2 Å for the top decoy. The Welch *t*-statistics (with degrees of freedom in brackets) and *p*-values for each variable are shown. Positive *t*-statistics indicate that successes have a larger mean than failures. The tests are based on the loop targets representing conformational clusters defined by at least four chains and two distinct PDB codes

	Welch *t*-test results			
Variables	**DiSGro**	**NGK**	**PETALS**	**Sphinx**
Loop length	$t(41.1) = -6.32$	$t(30.3) = -3.53$	$t(36.5) = -5.20$	$t(49.4) = -3.23$
	$p < .001$	$p = .0015$	$p < .001$	$p = .0022$
Cluster size	$t(49.2) = 4.18$	$t(31.1) = 3.91$	$t(41.9) = 3.10$	$t(5.0) = 2.63$
	$p < .001$	$p < .001$	$p = .0034$	$p = .011$
Cluster breadth	$t(47.7) = -4.62$	$t(23.1) = -3.52$	$t(35.0) = -3.08$	$t(48.2) = -2.94$
	$p < .001$	$p = .0018$	$p = .0040$	$p = .0050$

Abbreviations: PDB, Protein Data Bank; RMSD, root-mean-squared deviation.

TABLE 6 RMSD metrics for the loop instances with multiple conformations. The loop backbone RMSDs shown are averaged over the targets in the multiple conformation category, where decoys generated from each target are compared to all known conformations for that loop instance and RMSDs are calculated to the closest structure in each cluster. The columns "Min.," "Top," and "Top-5" refer, respectively, to the lowest RMSD among the 500 decoys, RMSD of the top-ranked decoy, and lowest RMSD among the top-five ranked decoys

	Local RMSD			Global RMSD		
Method	**Min.**	**Top**	**Top-5**	**Min.**	**Top**	**Top-5**
DiSGro	1.36	2.40	2.00	2.50	4.76	4.05
NGK	1.19	2.01	1.65	2.18	3.84	2.85
PETALS	1.28	2.11	1.86	2.56	4.26	3.60
Sphinx	1.14	2.24	1.80	2.28	4.70	3.65

Abbreviation: RMSD, root-mean-squared deviation.

phenomenon in their dataset. The detailed results for each loop target individually are given in Table S2 of the Supporting Information.

The multiple conformation loop instances in the RBD were not more difficult to predict. Methods located known conformations from their loop targets at a comparable level of accuracy versus those outside the RBD; for example, average global RMSDs for assessing the representation of all the conformations in the top five decoys were 2.55 versus 3.07 Å for NGK and 4.21 versus 3.94 for DiSGro. The average length of these loop targets in the RBD is 9.9 residues, and similar to the average length (10.0) among all multiple conformation targets. The loop regions with sequence variants in the PDB had little structural variability (Table 3) and were not expected to pose additional challenges for the loop modeling methods. Detailed results for each sequence variant confirm this, and are provided in Table S3 of the Supporting Information.

Five loop targets were omitted from the above analyses due to challenges encountered when running the methods. The two very long loops in the set, namely 146–168 and 783–816, were particularly difficult, with DiSGro and Sphinx unable to generate decoys possibly due to their lengths. The 146–168 loop has two conformations, both of which could be predicted moderately well by PETALS (top decoy global RMSDs: 2.18 for 6zgiB conformation, 2.39 for 7dddC conformation) and NGK (top decoy global RMSDs: 2.80 for 6zgiB conformation, 2.45 for 7dddC conformation). The length 34 loop (783–816) is very challenging, and no method could give useful results (top decoy global RMSDs: 26.8 for NGK, 12.0 for PETALS). The Sphinx webserver was also unable to generate decoys for 31–46 and 320–324 (6xm0A conformation) possibly due to a lack of suitable templates. Further, some of Sphinx's jobs were unable to complete the full SOAP-Loop ranking steps; thus, we used the 500 SOAP-Loop ranked decoys if they were available, and otherwise selected its top 500 decoys from the coarse-grained ranking stage for our analysis. Detailed results for these five targets are provided in Table S4 of the Supporting Information.

4 | CONCLUSION

In this article, we studied the conformations of loops in the SARS-CoV-2 S protein. We extracted all SARS-CoV-2 S protein loop regions, examined their sequence and structural variability based on the available structures in the PDB, and applied loop modeling methods to assess how well the loop conformations could be predicted. Then, 44 loop regions were identified, and as the structure of the S protein has been experimentally solved many times, 17 loop instances were observed to have substantive structural variability and be able to adopt multiple distinct conformations according to a cluster analysis. The clusters gave insights into the amount of structural uncertainty present in these loops, and there were quantifiable differences in their sizes and breadths.

Loops' frequent association with protein function, together with their more disordered nature compared to regular secondary structures, means that their accurate modeling is an important problem in structural biology. Specifically for the S protein, loop regions we identified include 475–487 and 495–506, which correspond to key loops known to be involved in binding with ACE2. These are referred to as

"Loop 3" and "Loop 4" in Williams et al.,[51] where molecular dynamics simulations revealed "Loop 3" to be highly flexible in the unbound state, including the possibility of a conformation that inhibits ACE2 binding. Interestingly, our results also showed that 475–487 was one of the most difficult loops to predict, with all four methods struggling with the 6xm0B template (global RMSD of top decoy >10 Å, Table S2). Exploring the conformational variability of "Loop 3" thus provides a fuller range of structural states that the development of therapeutics might target before the S protein binds to ACE2.[51] More generally, high-quality loop models are a crucial part of protein structures used in the computational drug discovery process.[19]

We found that the structurally flexible loops with multiple conformations in the S protein tended to be more challenging for loop modeling methods to predict a correct structure, compared to relatively inflexible loops with a single conformation. Prediction accuracies were strongly associated with loop length, due to the larger conformational space of longer loops. Further, it was very challenging for methods to predict all known conformations from a single structural template. Our results thus highlight limitations of current loop prediction methods, most of which were designed to predict a single "correct" conformation. These echo some of the findings in Marks et al.,[28] but with some important distinctions. First, we were able to more fully consider cluster size and breadth in the analysis, thanks to the large number of S protein chains in the PDB. Second, we did not construct a curated set of high and low flexibility loops specifically, but rather considered all S protein loops which cover a wider range of loop structural variability. In effect, a much larger proportion of loops (17 of 44 in our study) may be considered highly flexible, if other structures were to be solved this many times. Third, the multiple conformation targets in our dataset were easier to predict than those of Marks et al.[28] when allowing the best decoys across all structural templates to be chosen. Overall, this work provides insight into the abilities of current loop prediction methods for a key protein associated with the ongoing COVID-19 disease, and identifies the loops where structural flexibility could play a role as the SARS-CoV-2 virus continues to evolve. Future study in loop modeling protocols might better incorporate multiple conformation loops in their training data and improve prediction accuracies for longer loops.

Finally, we note one limitation of this study, namely our focus on loops rather than more global protein structure. In this sense, more global structural variability across S protein chains may have hindered the ability of methods to locate all the distinct loop conformations from a single input structure, since the rest of the protein chain is held fixed. Additionally, we found the observable changes to loop structures from known sequence variants in the PDB to be small. There could be more global structural changes due to mutation not detected by the current analysis, for example, the D614G mutation.[13] Nonetheless, loops deserve careful study in their own right, due to their functional importance. Further study could focus on larger-scale variability in the S protein structure, leveraging the rich source of experimental data available in the PDB to better understand COVID-19.

ACKNOWLEDGMENT

This work was partially supported by Discovery Grant RGPIN-2019-04771 from the Natural Sciences and Engineering Research Council of Canada.

CONFLICT OF INTEREST

Both the authors declare no potential conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/prot.26266.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the RCSB Protein Data Bank.

ORCID

Samuel W. K. Wong https://orcid.org/0000-0002-7325-7267

REFERENCES

1. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med*. 2020;382:727-733.
2. Jiang S, Hillyer C, Du L. Neutralizing antibodies against SARS-CoV-2 and other human coronaviruses. *Trends Immunol*. 2020;41(5):355-359.
3. Shang J, Ye G, Shi K, et al. Structural basis of receptor recognition by SARS-CoV-2. *Nature*. 2020;581(7807):221-224.
4. Polack FP, Thomas SJ, Kitchin N, et al. Safety and efficacy of the bnt162b2 mRNA COVID-19 vaccine. *N Engl J Med*. 2020;383(27):2603-2615.
5. Sewell HF, Agius RM, Kendrick D, Stewart M. COVID-19 vaccines: delivering protective immunity. *BMJ*. 2020;371:m4838.
6. Wrapp D, Wang N, Corbett KS, et al. Cryo-EM structure of the 2019-ncov spike in the prefusion conformation. *Science*. 2020;367(6483):1260-1263.
7. Lan J, Ge J, Yu J, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*. 2020;581(7807):215-220.
8. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235-242.
9. Cai Y, Zhang J, Xiao T, et al. Distinct conformational states of SARS-CoV-2 spike protein. *Science*. 2020;369(6511):1586-1592.
10. Schoof M, Faust B, Saunders RA, et al. An ultrapotent synthetic nanobody neutralizes SARS-CoV-2 by stabilizing inactive spike. *Science*. 2020;370(6523):1473-1479.
11. Shi R, Shan C, Duan X, et al. A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. *Nature*. 2020;584(7819):120-124.
12. Guo L, Bi W, Wang X, et al. Engineered trimeric ACE2 binds viral spike protein and locks it in "three-up" conformation to potently inhibit SARS-CoV-2 infection. *Cell Res*. 2021;31(1):98-100.
13. Yurkovetskiy L, Wang X, Pascal KE, et al. Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell*. 2020;183(3):739-751.
14. Henzler-Wildman K, Kern D. Dynamic personalities of proteins. *Nature*. 2007;450(7172):964-972.
15. Mittermaier A, Kay LE. New tools provide new insights in NMR studies of protein dynamics. *Science*. 2006;312(5771):224-228.

16. Schneider B, Gelly J-C, de Brevern AG, Černỳ J. Local dynamics of proteins and DNA evaluated from crystallographic B factors. *Acta Crystallogr D Biol Crystallogr*. 2014;70(9):2413-2419.
17. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. *Structure*. 2003;11(11):1453-1459.
18. Shehu A, Clementi C, Kavraki LE. Modeling protein conformational ensembles: from missing loops to equilibrium fluctuations. *Proteins: Struct Funct Bioinform*. 2006;65(1):164-179.
19. Muhammed MT, Aki-Yalcin E. Homology modeling in drug discovery: overview, current applications, and future perspectives. *Chem Biol Drug Des*. 2019;93(1):12-20.
20. Espadaler J, Querol E, Aviles FX, Oliva B. Identification of function-associated loop motifs and application to protein function prediction. *Bioinformatics*. 2006;22(18):2237-2243.
21. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science*. 2020;367(6485):1444-1448.
22. Papaleo E, Saladino G, Lambrughi M, Lindorff-Larsen K, Gervasio FL, Nussinov R. The role of protein loops and linkers in conformational dynamics and allostery. *Chem Rev*. 2016;116(11):6391-6423.
23. Zhang J, Cai Y, Xiao T, et al. Structural impact on SARS-CoV-2 spike protein by D614G substitution. *Science*. 2021;372(6541):525-530.
24. Chen J, Wang R, Wang M, Wei G-W. Mutations strengthened SARS-CoV-2 infectivity. *J Mol Biol*. 2020;432(19):5212-5226.
25. Sedova M, Jaroszewski L, Alisoltani A, Godzik A. Coronavirus3d: 3d structural visualization of COVID-19 genomic divergence. *Bioinformatics*. 2020;36(15):4360-4362.
26. Wong SW. Assessing the impacts of mutations to the structure of COVID-19 spike protein via sequential Monte Carlo. *J Data Sci*. 2020;18(3):511-525.
27. Li Q, Wu J, Nie J, et al. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell*. 2020;182(5):1284-1294.
28. Marks C, Shi J, Deane CM. Predicting loop conformational ensembles. *Bioinformatics*. 2018;34(6):949-956.
29. Liang S, Zhang C, Zhou Y. Leap: highly accurate prediction of protein loop conformations by integrating coarse-grained sampling and optimized energy scores with all-atom refinement of backbone and side chains. *J Comput Chem*. 2014;35(4):335-341.
30. Marks C, Nowak J, Klostermann S, et al. Sphinx: merging knowledge-based and ab initio approaches to improve protein loop prediction. *Bioinformatics*. 2017;33(9):1346-1353.
31. Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B. Loop modeling: sampling, filtering, and scoring. *Proteins: Struct Funct Bioinform*. 2008;70(3):834-843.
32. Stein A, Kortemme T. Improvements to robotics-inspired conformational sampling in Rosetta. *PloS One*. 2013;8(5):e63090.
33. Tang K, Zhang J, Liang J. Fast protein loop sampling and structure prediction using distance-guided sequential chain-growth Monte Carlo method. *PLoS Comput Biol*. 2014;10:e1003539.
34. Wong SW, Liu JS, Kou S. Fast de novo discovery of low-energy protein loop conformations. *Proteins: Struct Funct Bioinform*. 2017;85(8):1402-1412.
35. Fiser A, Do RKG, Šali A. Modeling of loops in protein structures. *Protein Sci*. 2000;9(9):1753-1773.
36. Barozet A, Bianciotto M, Vaisset M, Simeon T, Minoux H, Cortés J. Protein loops with multiple meta-stable conformations: a challenge for sampling and scoring methods. *Proteins: Struct Funct Bioinform*. 2021;89(2):218-231.
37. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*. 2009;25(9):1189-1191.
38. Kabsch W, Sander C. Dictionary of protein secondary structure—pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983;22(12):2577-2637.
39. Heffernan R, Yang Y, Paliwal K, Zhou Y. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics*. 2017;33(18):2842-2849.
40. Kabsch W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr A*. 1976;32(5):922-923.
41. Choi Y, Deane CM. FREAD revisited: accurate loop structure prediction using a database search algorithm. *Proteins: Struct Funct Bioinform*. 2010;78(6):1431-1440.
42. Karami Y, Rey J, Postic G, Murail S, Tufféry P, De Vries SJ. Dareus-loop: a web server to model multiple loops in homology models. *Nucleic Acids Res*. 2019;47(W1):W423-W428.
43. Sokal RR. A statistical method for evaluating systematic relationships. *Univ Kansas Sci Bull*. 1958;38:1409-1438.
44. Dong GQ, Fan H, Schneidman-Duhovny D, Webb B, Sali A. Optimized atomic statistical potentials: assessment of protein interfaces and loops. *Bioinformatics*. 2013;29(24):3158-3166.
45. Dunbar J, Krawczyk K, Leem J, et al. Sabpred: a structure-based antibody prediction server. *Nucleic Acids Res*. 2016;44(W1):W474-W478.
46. Ali A, Vijayan R. Dynamics of the ACE2–SARS-CoV-2/SARS-CoV spike protein interface reveal unique mechanisms. *Sci Rep*. 2020;10:14214.
47. Grubaugh ND, Hanage WP, Rasmussen AL. Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear. *Cell*. 2020;182(4):794-795.
48. Zhang L, Jackson CB, Mou H, et al. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat Commun*. 2020;11:6013.
49. Wrobel AG, Benton DJ, Xu P, et al. SARS-CoV-2 and bat ratg13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. *Nat Struct Mol Biol*. 2020;27(8):763-767.
50. Li J, Abel R, Zhu K, Cao Y, Zhao S, Friesner RA. The VSGB 2.0 model: a next generation energy model for high resolution protein structure modeling. *Proteins: Struct Funct Bioinform*. 2011;79(10):2794-2812.
51. Williams JK, Wang B, Sam A, Hoop CL, Case DA, Baum J. Molecular dynamics analysis of a flexible loop at the binding interface of the SARS-CoV-2 spike protein receptor-binding domain. *Proteins: Struct Funct Bioinform*. 2021. doi:10.1002/prot.26208
52. Wang G, Dunbrack RL Jr. Pisces: a protein sequence culling server. *Bioinformatics*. 2003;19(12):1589-1591.
53. Miranda LJV. PySwarms, a research-toolkit for particle swarm optimization in Python. *J Open Source Softw*. 2018;3(21):433.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Wong SWK, Liu Z. Conformational variability of loops in the SARS-CoV-2 spike protein. *Proteins*. 2022;90(3):691-703. doi:10.1002/prot.26266

APPENDIX A

B-factor analysis

Let (x_{11}, y_{11}, z_{11}), ..., (x_{1N}, y_{1N}, z_{1N}) and (x_{21}, y_{21}, z_{21}), ..., (x_{2N}, y_{2N}, z_{2N}) denote the measured backbone coordinates for the pair being compared, with corresponding B-factors denoted by B_{11}, ...B_{1N} and B_{21}, ... B_{2N}.

Since the B-factor is defined as $B = 8\pi^2 \langle u^2 \rangle$, a Gaussian approximation gives the variance in each measured x, y, and z coordinate as $B/(3 \cdot 8\pi^2)$. For the ith atom, the coordinate difference between the pair is a random vector (H_{xi}, H_{yi}, H_{zi}) with a multivariate Gaussian distribution with mean vector $(x_{1i} - x_{2i}, y_{1i} - y_{2i}, z_{1i} - z_{2i})$ and a diagonal covariance matrix with the value $\sigma_i^2 = (B_{1i} + B_{2i})/(3 \cdot 8\pi^2)$ along its diagonal.

By the properties of the multivariate Gaussian,

$$\frac{(H_{xi} - (x_{1i} - x_{2i}))^2 + (H_{yi} - (y_{1i} - y_{2i}))^2 + (H_{zi} - (z_{1i} - z_{2i}))^2}{\sigma_i^2}$$

has a chi-squared distribution with 3 degrees of freedom, denoted χ_3^2. Similarly, considering all the atoms together, a χ_{3N}^2 random variable is defined by

$$\sum_{i=1}^{N} \frac{(H_{xi} - (x_{1i} - x_{2i}))^2 + (H_{yi} - (y_{1i} - y_{2i}))^2 + (H_{zi} - (z_{1i} - z_{2i}))^2}{\sigma_i^2}.$$

The pair of loop backbones are not different if it is plausible that $(H_{xi}, H_{yi}, H_{zi}) = (0, 0, 0)$ for all N atoms, that is, all the coordinate differences are zero. This corresponds to computing the statistic

$$T = \sum_{i=1}^{N} \frac{(x_{1i} - x_{2i})^2 + (y_{1i} - y_{2i})^2 + (z_{1i} - z_{2i})^2}{\sigma_i^2}$$

and comparing T to the quantiles of the chi-squared distribution with $3N$ degrees of freedom. Taking a significance level of $\alpha = 0.05$, let c denote the 0.95 quantile of the χ_{3N}^2 distribution. Then, the pair is considered significantly different if $T > c$.

APPENDIX B

Updated scoring function for PETALS algorithm

In this work, we also tested a strategy for improving the energy function accuracy of the PETALS algorithm, in its ability to rank generated loop decoys. The set of structures used for training is the same as that described in Wong et al.,[34] namely, the CulledPDB list by PISCES[52] on March 14, 2015 with maximum 20% sequence identity, resolution 2.0 Å, and R-factor cutoff 0.25, thus ensuring no SARS-CoV-2 S protein structures were present. Loop regions were extracted via DSSP, from which we compiled 10 786 loops with lengths ranging from 5 to 10 residues.

The PETALS algorithm was first used to generate 200 decoys for each loop, and for each decoy, we computed: RMSD to the native conformation, 210 distance-based energy terms corresponding to each pair of atom types defined in DiSGro's energy function,[33] and a backbone torsion term.[34] We then define $\hat{y}_{ij}$ as the predicted energy of the ith loop's jth decoy according to

$$\hat{y}_{ij} = T_{ij} + \sum_{k=1}^{210} \beta_k E_{ijk},$$

where β_k's are coefficients associated with each energy term E_{ijk} to be trained, and T_{ij} is the torsion term. Then, define the square-error loss function

$$\sum_{i=1}^{N} \sum_{j=1}^{200} w_{ij} \left(f(\hat{y}_{ij}) - f(\mathrm{RMSD}_{ij}) \right)^2, \tag{B1}$$

where RMSD_{ij} is the RMSD to native and w_{ij} is the weight associated with the ith loop's jth decoy, N is the number of training loops, and f is a mapping function associated with the rank of that decoy. The decoys with the lowest RMSDs are the ones that best resemble the true conformation; thus the goal is to train the β_k's to minimize this loss function so that the rankings of the predicted energies and the rankings of the RMSD values match as closely as possible.

We chose $f(\cdot)$ to be a function that maps values into quantile bins. Specifically, we ranked the 200 predicted energies $\{\hat{y}_{ij}\}_{j=1}^{200}$ from smallest to largest, then assigning $f = 1$ to the best 10%, $f = 2$ to the next 10%, until $f = 10$ for the last 10%. We ranked the 200 RMSD values $\{\mathrm{RMSD}_{ij}\}_{j=1}^{200}$ and assigned values of f the same way. Positive weights w_{ij} were assigned to the top five quantile bins, with higher weights for the better ranked predicted energies: 1.0 for the best 10%, 0.9 for the next 10%, until 0.6 for 5th quantile bin, and zero for the rest. We used 80% of the loops as training data and 20% as validation data. As gradient information was unavailable due to the discrete nature of the model, the PySwarms[53] implementation of Particle Swarm Optimization was used to minimize the square error loss function in Equation (B1).