Received: 21 December 2020 | Revised: 26 August 2021 | Accepted: 16 September 2021

DOI: 10.1002/prot.26250

RESEARCH ARTICLE

PROTEINS WILEY

Evolution of the SARS-CoV-2 proteome in three dimensions (3D) during the first 6 months of the COVID-19 pandemic

Joseph H. Lubin[1,2] | Christine Zardecki[1,3] | Elliott M. Dolan[1,2] | Changpeng Lu[1] | Zhuofan Shen[1,2] | Shuchismita Dutta[1,3,4] | John D. Westbrook[1,3,4] | Brian P. Hudson[1,3] | David S. Goodsell[1,3,4,5] | Jonathan K. Williams[2] | Maria Voigt[1,3] | Vidur Sarma[1] | Lingjun Xie[1,2] | Thejasvi Venkatachalam[1] | Steven Arnold[1] | Luz Helena Alfaro Alvarado[6] | Kevin Catalfano[7] | Aaliyah Khan[8] | Erika McCarthy[9] | Sophia Staggers[10] | Brea Tinsley[11] | Alan Trudeau[12] | Jitendra Singh[13] | Lindsey Whitmore[14] | Helen Zheng[15] | Matthew Benedek[16] | Jenna Currier[17] | Mark Dresel[2] | Ashish Duvvuru[17] | Britney Dyszel[18] | Emily Fingar[19] | Elizabeth M. Hennen[20] | Michael Kirsch[19] | Ali A. Khan[19] | Charlotte Labrie-Cleary[19] | Stephanie Laporte[21] | Evan Lenkeit[2] | Kailey Martin[18] | Marilyn Orellana[17] | Melanie Ortiz-Alvarez de la Campa[22] | Isaac Paredes[23] | Baleigh Wheeler[24] | Allison Rupert[24] | Andrew Sam[2] | Katherine See[25] | Santiago Soto Zapata[19] | Paul A. Craig[25] | Bonnie L. Hall[24] | Jennifer Jiang[1] | Julia R. Koeppe[19] | Stephen A. Mills[16] | Michael J. Pikaart[17] | Rebecca Roberts[18] | Yana Bromberg[26] | J. Steen Hoyer[27] | Siobain Duffy[27] | Jay Tischfield[28] | Francesc X. Ruiz[29] | Eddy Arnold[2,29] | Jean Baum[2] | Jesse Sandberg[30] | Grace Brannigan[30,31] | Sagar D. Khare[1,2,4] | Stephen K. Burley[1,2,3,4,32]

[1]Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA

[2]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA

[3]Research Collaboratory for Structural Bioinformatics Protein Data Bank, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA

[4]Rutgers Cancer Institute of New Jersey, Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA

[5]The Scripps Research Institute, La Jolla, California, USA

[6]Grinnell College, Grinnell, Iowa, USA

[7]University of Notre Dame, Notre Dame, Indiana, USA

[8]University of Maryland Baltimore County, Baltimore, Maryland, USA

[9]Stevens Institute of Technology, Hoboken, New Jersey, USA

[10]Frostburg State University, Frostburg, Maryland, USA

[11]Youngstown State University, Youngstown, Ohio, USA

[12]University of Central Florida, Orlando, Florida, USA

[13]New York City College of Technology, Brooklyn, New York, USA

[14]Howard University, Washington, District of Columbia, USA

[15]Watchung Hills Regional High School, Warren, New Jersey, USA

[16]Xavier University, Cincinnati, Ohio, USA

[17]Hope College, Holland, Michigan, USA

[18]Ursinus College, Collegeville, Pennsylvania, USA

[19]SUNY Oswego, Oswego, New York, USA

[20]Roger Williams University, Bristol, Rhode Island, USA

[21]Brandeis University, Waltham, Massachusetts, USA

[22]University of Puerto Rico-Rio Piedras, San Juan, Puerto Rico

[23]John Jay College, New York, New York, USA

[24]Grand View University, Des Moines, Iowa, USA

[25]Rochester Institute of Technology, Rochester, New York, USA

[26]Department of Biochemistry and Microbiology, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA

[27]Department of Ecology, Evolution and Natural Resources, School of Environmental and Biological Sciences, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA

[28]Department of Genetics, Rutgers, The State University of New Jersey, and Human Genetics Institute of New Jersey, Piscataway, New Jersey, USA

[29]Center for Advanced Biotechnology and Medicine, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA

[30]Center for Computational and Integrative Biology, Rutgers, The State University of New Jersey, Camden, New Jersey, USA

[31]Department of Physics, Rutgers, The State University of New Jersey, Camden, New Jersey, USA

[32]Research Collaboratory for Structural Bioinformatics Protein Data Bank, San Diego Supercomputer Center, University of California, San Diego, La Jolla, California, USA

Correspondence
Sagar D. Khare, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ, USA.
Email: khare@chem.rutgers.edu

Stephen K. Burley, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ, USA.
Email: stephen.burley@rcsb.org

Funding information
Division of Biological Infrastructure, Grant/Award Number: DBI-1832184; Division of Chemical, Bioengineering, Environmental, and Transport Systems, Grant/Award Number: CBET1923691; National Cancer Institute, Grant/Award Number: R01 GM133198; National Institute of Allergy and Infectious Diseases; National Institute of General Medical Sciences, Grant/Award Numbers: GM008339, GM135141, R01 GM132565; National Institutes of Health, Grant/Award Numbers: GM136431, R0137 AI027690; National Science Foundation, Grant/Award Numbers: IUSE 1709170, IUSE 1709278, IUSE 1709355, IUSE 1709805; U.S. Department of Energy, Grant/Award Number: DE-SC0019749; New Jersey Space Grant Consortium; Busch Biomedical Foundation

Abstract

Understanding the molecular evolution of the SARS-CoV-2 virus as it continues to spread in communities around the globe is important for mitigation and future pandemic preparedness. Three-dimensional structures of SARS-CoV-2 proteins and those of other coronavirusess archived in the Protein Data Bank were used to analyze viral proteome evolution during the first 6 months of the COVID-19 pandemic. Analyses of spatial locations, chemical properties, and structural and energetic impacts of the observed amino acid changes in >48 000 viral isolates revealed how each one of 29 viral proteins have undergone amino acid changes. Catalytic residues in active sites and binding residues in protein–protein interfaces showed modest, but significant, numbers of substitutions, highlighting the mutational robustness of the viral proteome. Energetics calculations showed that the impact of substitutions on the thermodynamic stability of the proteome follows a universal bi-Gaussian distribution. Detailed results are presented for potential drug discovery targets and the four structural proteins that comprise the virion, highlighting substitutions with the potential to impact protein structure, enzyme activity, and protein–protein and protein–nucleic acid interfaces. Characterizing the evolution of the virus in three dimensions provides testable insights into viral protein function and should aid in structure-based drug discovery efforts as well as the prospective identification of amino acid substitutions with potential for drug resistance.

KEYWORDS
coronavirus, COVID-19, databases, protein, evolution, molecular, pandemics, SARS-CoV-2, viral proteins

1 | INTRODUCTION

Rising numbers of COVID-19 infections and deaths worldwide show that we must prepare for the next outbreak when (it is no longer a matter of if) another coronavirus jumps the species barrier and infects humans. SARS-CoV-2, the causative agent of the COVID-19 global pandemic, is a member of the coronavirus family of RNA viruses that cause disease in mammals and birds.[1] Coronaviruses have the longest

RNA virus genomes of all known single-stranded RNA viruses. Their RNA-dependent RNA polymerases act together with RNA helicases and proofreading exonucleases to carry out efficient and relatively faithful copying of the lengthy genome.[2] Proofreading notwithstanding, coronavirus genome replication is not perfect, and coronaviruses do evolve as they passage serially from one host to the next. Today in the time of COVID-19, genome sequence-based "fingerprinting" of the virus in near real time has provided very detailed accounts of how the virus has moved around the globe since late 2019 as infected individuals, many of them asymptomatic, traveled from continent to continent.[3,4] Viral genome fingerprinting has also enabled detailed analyses of the impact of amino acid changes in particular proteins that modulate infectivity, etc. (e.g.,[5]). A systematic analysis of how these genomic changes affect the three-dimensional structures of the SARS-CoV-2 proteins, and what, if any, impact observed changes may have had on the functions of these proteins is critical for effective molecular surveillance against SARS-CoV-2 and other coronaviruses that could jump the species barrier.

Preparedness against future coronavirus pandemics requires an understanding of the conservation and mutability of viral proteins that are drug design targets. For example, a key coronaviral enzyme is nonstructural protein 5 (nsp5 or main protease or Mpro), which is highly conserved across all known coronaviruses. SARS-CoV-2 nsp5 is 95% identical in amino acid sequence to that of its SARS-CoV-1 counterpart and highly structurally similar (Figure 1). 3D structures of nsp5 are conserved among all known coronaviruses. We had every opportunity in the wake of the SARS-CoV-1 epidemic to discover and develop a drug targeting SARS-CoV-1 nsp5 (and effective against other coronavirus main proteases). Structure-guided approaches using PDB ID 1Q2W[6] and the many structures of SARS-CoV-1 nsp5 subsequently released by the PDB would almost certainly have yielded one (possibly multiple) potent and selective enzyme inhibitor(s) with good drug-like properties and an acceptable safety profile. A safe and effective drug targeting SARS-CoV-1 nsp5 would almost certainly be working today for SARS-CoV-2.[7] Looking ahead, structure-based drug design efforts in academia and industry are likely to yield several new drugs targeting nsp5 and other key viral proteins. An understanding of viral evolution will be essential for ensuring the effectiveness of these drugs, and potential combinations in treating SARS-CoV-2 infections.

Viral evolution is shaped by the interplay of mutational tolerance and selection pressures due to vaccines or drugs. For example, use of drugs targeting viral proteases of HCV and HIV has led to the emergence of drug resistance mutations (DRMs). DRMs maintain native-like substrate processing but abolish or significantly diminish drug binding, thereby escaping drug action and gaining an evolutionary advantage. DRMs arise because of the mutational tolerance of the active site residues of viral proteases: in the absence of the drug, there is little evolutionary disadvantage to DRMs, and they simply lurk in the population at low levels until selection pressure is applied. Because DRMs are selected from the pool for pre-existing diversity of viral variants, it is important to understand existing diversity of protein variants in the population, and test drug candidates against such variants to minimize the risk of DRM emergence. Identifying the list of DRMs associated with each drug would enable personalized tailoring of drug cocktails, such that probability of acquiring multiple DRMs can be minimized. DRMs identified in SARS-CoV-2 drug targets may also be present in novel coronaviruses that cross the species barrier. Effective emergency countermeasures administered to prevent future pandemics would, thus, also benefit from an understanding of the DRM landscape of existing therapeutics. Structure-based modeling of existing variants in the population is expected to aid prospective identification of DRMs.

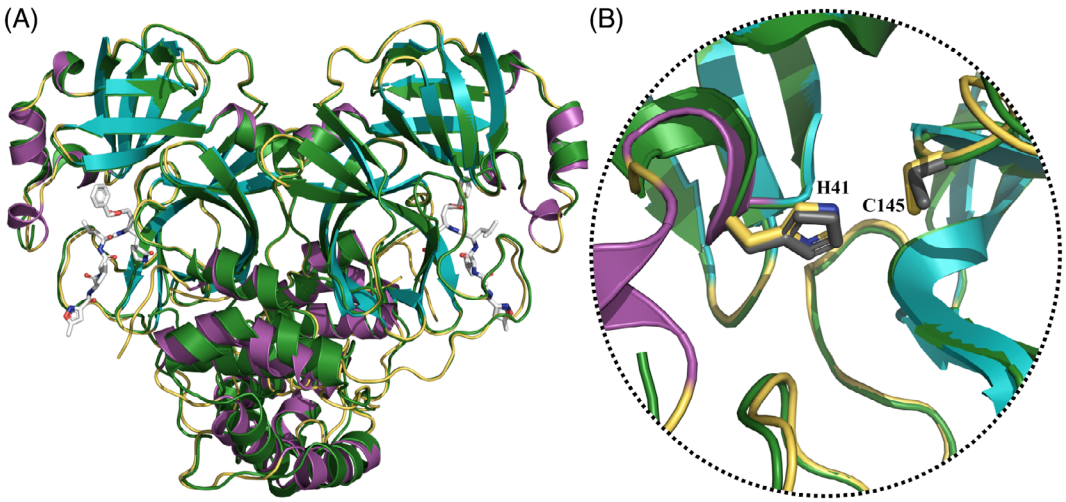

FIGURE 1 (A) Ribbon representation of the experimental structure of SARS-CoV-2 nsp5 (PDB ID 6LU7[8]), with color coding magenta (α-helices), cyan (β-sheets), and gold (loops) overlaid with SARS-CoV-1 nsp5 (PDB ID 1Q2W[6]), colored in green. Substrate analog inhibitor present in PDB ID 6LU7 is shown as an atomic stick figure with atom color coding white (carbon), red (oxygen), and blue (nitrogen). (B) The active site of both proteases, with the catalytic dyad (H41 and C145 in 6LU7) shown, with 6LU7 in gold and 1Q2W in gray

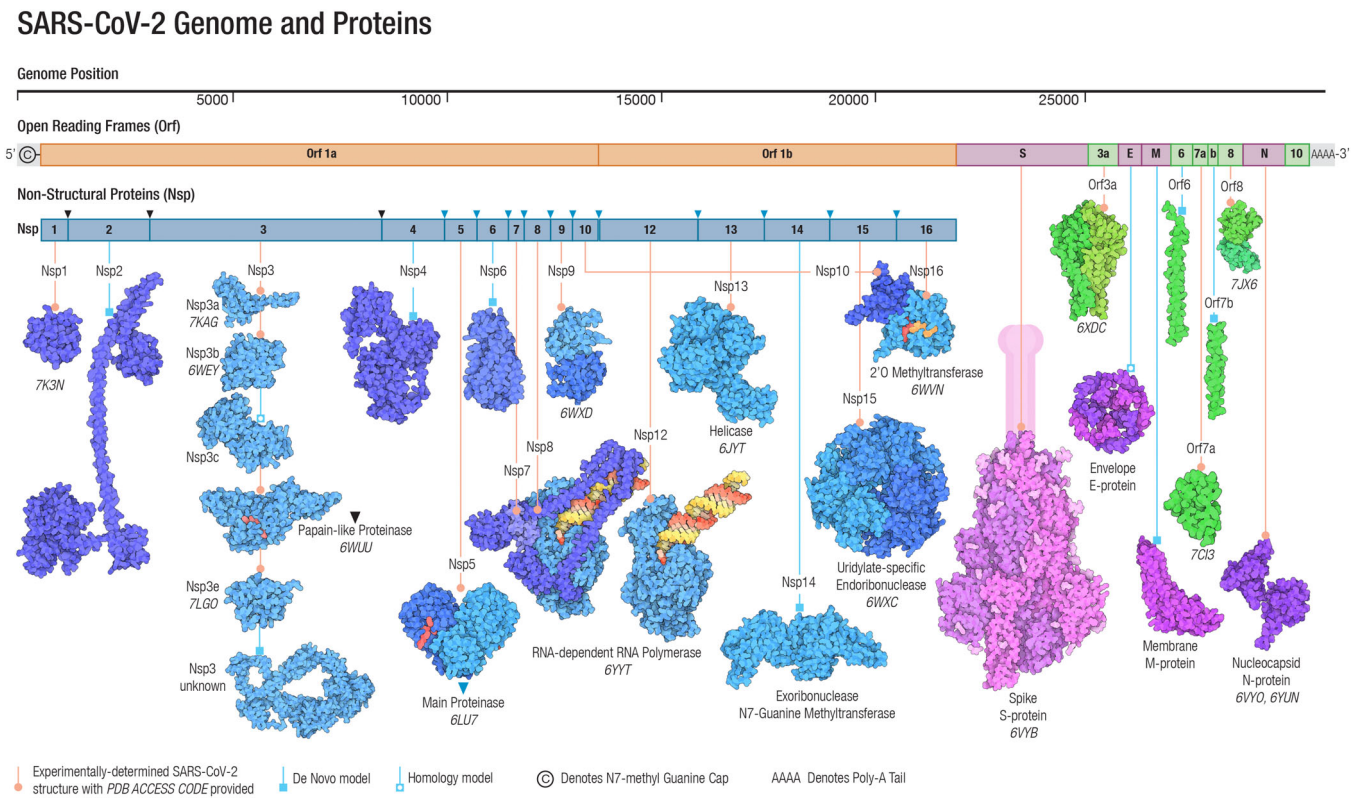

FIGURE 2 Architecture of the SARS-CoV-2 genome and proteome, including nsps derived from polyproteins or pp1a and pp1ab (shades of blue), virion structural proteins (pink/purple), and open reading frame proteins (Orfs, shades of green). Polyprotein cleavage sites are indicated by inverted triangles for Papain-like Proteinase (PLPro, black) and the Main Protease (nsp5, blue). The double-stranded RNA substrate-product complex of the RNA-dependent RNA polymerase (shown as the nsp7-nsp8$_2$-nsp12 heterotetramer and separately with only nsp12) is color coded (yellow: product strand, red: template strand). Transmembrane portions of the Spike S-protein are shown in cartoon form (pink). The source of the structural models used for analyses for all study proteins are indicated (experimentally-determined, computational homology model, or de novo predicted model)

Herein, we report a comprehensive study of how the SARS-CoV-2 proteome (Figure 2) evolved in 3D during the first 6 months of the pandemic between late 2019 and June 25th, 2020. We combined viral genome sequence data assembled by GISAID (https://www.gisaid.org), the wealth of experimental 3D structure information for SARS-CoV-2 and other coronavirus proteins available from the open-access Protein Data Bank or PDB,[9–11] and computed structural models in cases where experimentally-determined structures were not available.

2 | RESULTS AND DISCUSSION

2.1 | Sequence analyses

Viral genome sequencing and alignments of more than 48 000 individual isolates revealed protein sequence variation between December 2019 and late June 2020. We investigated the spatial locations, chemical properties, and structural and energetic impacts of the observed amino acid changes with reference to the original viral genome/proteome sequence publicly released in January 2020.

Every one of the 29 SARS-CoV-2 study proteins listed in Table 1 underwent changes in amino sequence, generating an average of approximately one unique sequence variant (USV) per study protein amino acid residue (lowest: nsp10 at ~0.59 USVs/residue; highest: Orf3a at ~2.46 USVs/residue). Protein sequence differences were entirely restricted to nonsynonymous changes in one or more residues. No insertions or deletions were detected in any of the 29 study proteins. Most USVs reflect a single amino acid change in the protein sequence (~66.8%). Smaller proportions of the USVs showed accumulation of two (~25.4%), three (~6.8%), four (~0.8%), or rarely five or more (~0.2%) amino acid substitutions. Where multiple substitutions were observed in a study protein USV, visual inspection of GISAID metadata typically revealed that they accumulated serially, but no systematic effort was made to track sequence changes as a function of sample collection date or geographic location. The modest degree of amino acid sequence variation observed for each of the 29 study proteins analyzed herein is consistent with previous studies of coronavirus evolution, which underscore the importance of the 3′-to-5′ exoribonuclease activity of nsp14 (reviewed in[2]). In contrast, RNA viruses that do not possess proofreading enzymes (e.g., hepatitis C virus) exhibit significantly higher rates of amino acid substitution.[12]

TABLE 1 Summary statistics from analysis of GISAID dataset (downloaded 06/25/2020)

Study protein	Clean protein sequences analyzed	Clean protein sequences unchanged	Unique protein sequence variants	Protein length (residues)	Average unique protein sequences variants per residue	Homo-oligomeric chains	Structural model	Structure model determination method	X-ray structure resolution (Å)
nsp1	46414	45315	212	179	1.18	1	7K3N	XRD	1.65
nsp2	41579	28543	838	638	1.31	1		De Novo	NA
nsp3a[a]	37181	35364	223	206	1.08	1	7KAG	XRD	3.21
nsp3b[a]	37181	36151	181	206	0.88	1	6WEY	XRD	0.95
nsp3c[a]	37181	35665	229	332	0.69	1	H-2w2g	Homology	NA
PLPro[a]	37181	36133	225	343	0.66	1	6WUU	XRD	2.79
nsp3e[a]	37181	36114	152	172	0.88	1	7LGO	XRD	2.45
unk[a]	37181	34614	455	686	0.66	1		De Novo	NA
nsp4	45306	42803	380	500	0.76	1		De Novo	NA
nsp5	46797	43884	217	306	0.71	2	6YB7	XRD	2.16
nsp6	46691	39758	262	290	0.90	1		De Novo	NA
nsp7[b]	48670	47876	68	83	0.82	1	6YYT	CEM	2.90
nsp8[b]	48335	47635	144	198	0.73	1	6YYT	CEM	2.90
nsp9	48686	48289	82	113	0.73	2	6WXD	XRD	2.00
nsp10[c]	46850	46507	81	139	0.58	1	6WVN	XRD	2.00
nsp12[b]	44203	10266	730	932	0.78	1	6YYT	CEM	2.90
nsp13	44120	39652	466	595	0.78	2	6JYT	XRD	2.80
nsp14	31465	29600	335	527	0.64	1		De Novo	NA
nsp15	42022	40208	326	346	0.94	6	6WXC	XRD	1.85
nsp16[c]	42287	41118	206	298	0.69	1	6WVN	XRD	2.00
S-protein	33290	7743	1190	1273	0.93	3	RBD: 6M17 Close: 6VXX	CEM CEM	2.902.80
Orf3a	45932	27554	677	275	2.46	2	6XDC	CEM	2.90
E-protein	48552	48052	82	75	1.09	5	H-5x29	Homology	NA
M-protein	47326	45423	181	222	0.82	1		De Novo	NA
Orf6	48490	47935	76	61	1.25	1		De Novo	NA
Orf7a	41969	41146	181	121	1.50	1	7CI3	XRD	2.20
Orf7b	43211	42939	56	43	1.30	1		De Novo	NA
Orf8	47796	42120	195	121	1.61	2	7JX6	XRD	1.61
N-protein	45635	26486	889	419	2.12	1	N: 6YVO C: 6YUN	XRD XRD	1.701.44

Abbreviations: CEM, cryo-electron microscopy; De novo, ab initio structure prediction using Rosetta; Homology, homology modeling; XRD, X-ray diffraction.
[a]Part of nsp3.
[b]Part of RDRP.
[c]Part of methyltransferase.

2.2 | Mapping locations of observed sequence variations in 3D

Experimental structures or computed 3D structural models were assembled for all 29 study proteins and their respective USVs (see Section 4). For each study protein, we identified amino acid substitutions mapping to sites in the polypeptide chain buried in the hydrophobic core, exposed on the macromolecule surface, and present in the "boundary" layer between the core and the surface (Table 2). Not surprisingly, most of the amino acid substitutions occur on the protein surface (~53.0%) or within the boundary layer (~38.3%). Very few occur in the protein core (~8.7%). Characterization of the nature of each substitution (conserved, nonconserved) revealed that non-conservative amino acid changes were common (~64.3%), albeit less so if they occurred in the core (~54.3%) or the boundary layer (~55.0%), rather than on the protein surface (~69.1%). (N.B. A minority of USVs for some study proteins could not be modeled in 3D due to incomplete structural information.)

To further examine the types of amino acid changes in the viral proteome, we generated location-based substitution matrices from the observed USVs for each study protein and for the entire viral proteome (Figure 3). Substitutions to or from all 20 amino acids were

TABLE 2 Analysis results for 3D spatial locations and energetics of all 29 study protein USVs

	Residue counts			USVs		USV substitution count						Substitutions			Conservation		3D mapping substitutions			Energetic impact			
Study protein	Surface	Boundary	Core	Total	Modeled	1	2	3	4	5	6+	Total	Single USV	Max occurrences	Conserved	Nonconserved	Surface	Boundary	Core	Outlier	More stable	Neutral	Less stable
nsp1	79	79	21	212	126	200	11	1	0	0	0	211	200	5	67	144	79	47	8	10	5	15	96
nsp2	392	214	32	838	838	489	260	81	7	1	0	639	456	205	219	420	640	578	67	35	131	159	513
nsp3a[a]	62	45	0	223	105	202	17	4	0	0	0	219	200	5	64	155	67	40	8	6	10	26	63
nsp3b[a]	69	72	28	181	142	170	11	0	0	0	0	174	158	3	66	108	68	58	21	10	17	24	91
nsp3c[a]	103	134	27	229	174	215	12	1	1	0	0	225	210	7	83	142	91	82	14	8	35	46	85
PLPro[a]	145	141	31	225	209	215	10	0	0	0	0	221	209	3	71	150	117	83	19	11	30	48	120
nsp3e[a]	61	50	4	152	87	145	5	2	0	0	0	147	140	7	50	97	56	26	11	3	10	18	56
UNK[a]	392	222	72	455	455	395	55	4	0	0	1	444	383	7	176	268	338	151	41	28	83	115	229
nsp4	250	202	48	380	380	327	45	5	3	0	0	362	323	16	158	204	248	169	27	17	63	93	207
nsp5	202	308	102	217	217	189	26	2	0	0	0	211	189	8	80	131	106	109	32	10	24	48	135
nsp6	123	116	51	262	262	180	81	1	0	0	0	232	192	70	103	129	137	165	43	13	40	36	173
nsp7-nsp8$_2$-nsp12	514	701	139	934	840	444	427	55	4	1	3	811	655	443	328	483	379	857	107	37	80	79	644
nsp9	112	76	32	82	82	79	2	1	0	0	0	85	84	2	25	60	59	20	7	5	15	23	39
nsp10-nsp16	160	185	81	286	269	266	19	1	0	0	0	282	260	4	105	177	128	124	37	12	41	46	170
nsp13	412	610	174	466	463	363	62	34	6	1	0	417	336	40	165	252	307	221	87	24	71	143	224
nsp14	215	257	55	335	335	306	26	1	0	1	1	339	307	6	134	205	172	176	32	18	38	64	215
nsp15	546	1296	240	326	326	298	28	0	0	0	0	319	294	7	117	202	122	198	34	18	60	61	187
S-protein	969	1551	396	1190	689	327	675	171	13	3	1	922	652	805	312	610	348	824	72	35	100	138	415
Orf3a	173	178	36	677	400	303	339	33	2	0	0	428	257	198	145	283	239	316	33	18	55	69	258
E-protein	211	80	0	82	56	79	3	0	0	0	0	81	78	3	41	40	38	20	1	2	9	24	21
M-protein	114	85	24	181	181	162	16	2	0	0	1	184	168	5	74	110	125	67	15	10	22	40	109
Orf6	61	0	0	76	76	73	2	0	0	1	0	80	78	2	28	52	82	0	0	3	2	38	33
Orf7a	77	37	7	181	95	171	9	1	0	0	0	175	160	3	64	111	67	29	3	3	7	24	61
Orf7b	44	0	0	56	56	52	4	0	0	0	0	57	54	2	21	36	60	0	0	3	5	18	30
Orf8	102	84	22	195	166	147	43	5	0	0	0	172	142	31	53	119	133	53	23	12	15	29	110
N-protein	137	93	12	889	262	429	185	237	36	2	0	577	344	272	132	445	205	79	9	12	39	80	131

Note: Column label definitions (left to right): Study Protein: study protein or multiprotein-complex name. Residue counts: total number of residues in the study protein/complex that map to each layer. USVs: total—number of USVs for each study protein identified across all GISAID sequences; Modeled—number of USVs for which 3D structural models were computed. USV substitution count: number of single-substituted, double-substituted, etc. USVs. Substitutions: total—number of unique substitutions identified across all study protein USVs; single USV—number of substitutions that occur in only one USV; Max Occurrences—number of USVs in which the most frequent substitution occurred (independent of GISAID count). Conservation: conserved—number of conserved substitutions; nonconserved—number of nonconserved substitutions. 3D Mapping Substitutions: layer identifications counted across all modeled substitutions. Energetic Impact: Outlier—number of USVs with $\Delta\Delta G^{App}$ greater than two standard deviations above or below the mean value of $<\Delta\Delta G^{App}>$. More Stable—number of USVs with non-outlier $\Delta\Delta G^{App} \leq -1$; Neutral—number of USVs with $-1 < \Delta\Delta G^{App} < +1$; Less Stable—number of USVs with non-outlier $\Delta\Delta G^{App}$ between $+1$ and two standard deviations above the mean value of $<\Delta\Delta G^{App}>$.

[a]Part of nsp3.

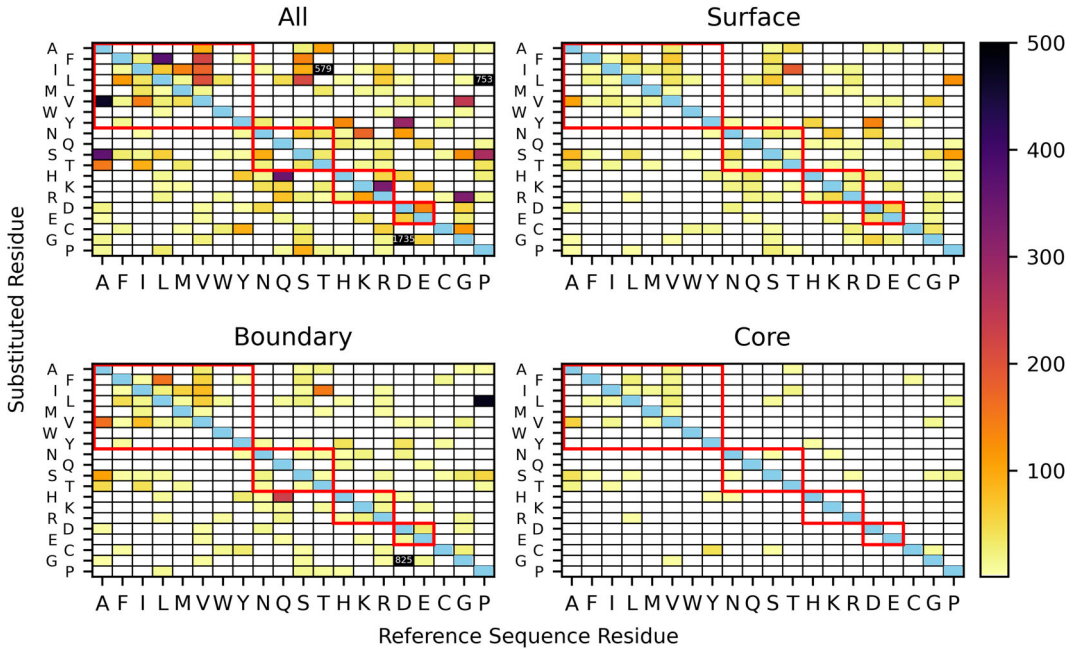

FIGURE 3 Observed counts for USV substitutions of Reference Sequence Residue (i.e., original protein reference sequence amino acid) changing to Substituted Residue for all 19 study proteins with experimentally-determined structures. (The uncertainty inherent to computationally-predicted structural models results in greater uncertainty in layer identification for those, thus only models based on experimentally-determined structures are included.) Red boxes enclose conservative substitutions for hydrophobic, uncharged polar, positively charged, and negatively charged amino acids, respectively, in order from upper left to lower right. Cysteine, glycine, and proline are excluded from these groupings. Substitutions which occurred in 500 or more USVs are also shown with a number indicating the count

observed across all 29 study proteins. Notable nonconservative changes include hydrophobic residues changing to negatively charged residues and vice versa, and glycine and proline residues changing to all types of amino acids on the surface, and to a lesser extent within the boundary layer. These trends reflect anticipated constraints imposed by protein structure on the thermodynamic stability due to amino acid substitutions. In the tightly packed environment of the hydrophobic core of a protein, fewer types of amino acid substitutions are likely to be compatible with the 3D structure, and changes that do not impair protein function are likely to be conservative. In contrast, protein boundary layers and surfaces impose far fewer constraints in terms of structural incompatibility and nonconservative substitutions.

Most of the observed nonconservative changes can be attributed to the architecture of the genetic code and single base changes in the viral RNA genome. For example, Alanine to Aspartic and Glutamic acid changes are achievable via single base changes in the second base of their respective codons. However, changes requiring double base changes (e.g., Proline to Aspartate) were also observed.

2.3 | Analyzing energetic consequences of observed sequence variations

The energetic impact of observed amino acid substitutions for each unique sequence variant of each study protein was calculated using Rosetta (Table 2, Figure 4). Most of the amino acid changes were estimated to be moderately destabilizing as judged by changes in the free energy of stabilization (apparent $\Delta\Delta G$ or $\Delta\Delta G^{App} = 0.0$ to $+15.0$ Rosetta energy units or REU; ~73.0%). A modest number were estimated to be stabilizing ($\Delta\Delta G^{App} = -0.01$ to -15.0 REU; ~22.8%). In the minority of cases, $\Delta\Delta G^{App}$ exceeded $+15.0$ REU (~4.2%). The distribution of $\Delta\Delta G^{App}$ values was used to identify outliers for each study protein (Table 2). Due to the inherent errors associated with $\Delta\Delta G^{App}$ calculations, we note that these values are best interpreted qualitatively, with numbers in the range -1 to $+1$ REU considered neutral (20.7%) in their impact on the stability of the protein. We also note that the "wild type" protein sequences are derived from the sequence of the virus first deposited in January 2020 (reference genome). The substitutions considered here, however, may have arisen in the background of other strains, and our thermodynamic analyses do not speak to the evolutionary dynamics of the virus.

Given that all modeled amino acid substitutions were detected in viruses that likely had infected human hosts when they were isolated, we assume that all modeled USVs correspond to stable, functional proteins. Most globular proteins are marginally stable, with measured free energies of stabilization $\Delta G \sim -5$ to -15 kcal/mol,[13] and tolerated amino acid substitutions are expected to have an impact within this range. Therefore, we believe that the small minority of computed large positive $\Delta\Delta G^{App}$ values represent artifacts arising from errors/approximations in our calculations (Table 2). For example, positional restraints on backbone atoms were employed when modeling USVs in Rosetta to prevent substantial departures from the reference protein

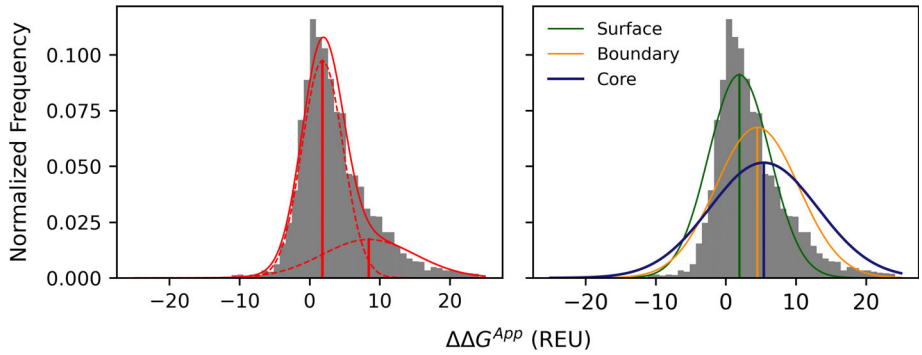

FIGURE 4 Normalized frequency histogram for $\Delta\Delta G^{App}$ calculated for all USVs aggregated across all 19 study proteins with experimentally-determined structures. (The uncertainty inherent to computationally-predicted structural models results in significant uncertainty in calculating atom-level energetics for those models, thus only models based on experimentally-determined structures were included.) Left: Overlay with fitted bi-Gaussian curve (solid red line) with fitted individual Gaussian curves (dashed red lines). The means for the individual Gaussian distributions were +1.8 REU (standard deviation or SD: 8.5) and +8.4 REU (SD: 44.2) ($R^2 = 0.92$). Right: Overlay of the same normalized frequency histogram with fitted single Gaussian curves fitted to subsets of USVs with Surface (green; mean value: +1.9 REU, SD: 19.2; $R^2 = 0.75$), Boundary (yellow; mean value: +4.5 REU, SD: 35.0; $R^2 = 0.74$), or Core (blue; mean value: +5.4 REU, SD: 59.9; $R^2 = 0.42$) substitutions. USVs with multiple substitutions were included in single Gaussian fitting when all substitutions mapped to the same region of the study protein

backbone conformation so more permissive restraints on the polypeptide chain backbone may be required to model computationally the effects of some particularly large amino acid changes. Alternatively, large positive values of $\Delta\Delta G^{App}$ may reflect shortcomings in the Rosetta energy function. Outlier cases provide a benchmark for improvements in Rosetta and other stability calculation approaches. Outliers notwithstanding, ~95% of all computationally modeled USVs yielded reasonable $\Delta\Delta G^{App}$ values. (N.B. Experimentally-determined crystal structures were not available for all study proteins, and where this was the case, computationally-predicted models were used. Computed models are considered with less confidence than experimental models. Energetic calculations, which are sensitive to even sub-Å structural perturbations, should therefore accordingly be considered with less confidence for those models.)

We next examined the distribution of energetic effects of the observed substitutions for each study protein and aggregated across all 19 study proteins with experimentally-determined structures. Several previously published experimental and theoretical studies have examined the distributions of thermodynamic stability changes due to point substitutions in individual proteins and examined the implication of these distributions for molecular evolution.[14–17] Our dataset provides an opportunity to re-examine conclusions from these studies which are, with a single exception,[18] based on limited experimental data and/or computational findings. Tokuriki et al. (2007) used FoldX-based calculations of all single substitutions in 21 different globular proteins and found that despite a diverse range of sizes and folds, the distribution of stability effects largely follows a bi-Gaussian function for each protein. They found that surface residues exhibit a narrow distribution with a modestly destabilizing mean $\Delta\Delta G^{App}$ ($<\Delta\Delta G^{App}>$), whereas core residues exhibit a wider distribution with higher positive $<\Delta\Delta G^{App}>$ values.[15] Such asymmetric distributions were also found for lattice model proteins, and were recently shown to arise from first-principle statistical mechanical considerations and a sufficiently large amino acid alphabet size.[16] Faure and Koonin (2015) obtained similar distributions across proteomes of five organisms selected from archaea, prokaryota, and eukaryota, suggesting that this distribution of energetic effects is a universal and evolutionarily conserved feature of globular protein folds.[14] In these studies, as in ours, individual $\Delta\Delta G^{App}$ values may not be accurately predictive of experimental measurements (state-of-the-art $\Delta\Delta G^{App}$ prediction methods typically have correlation coefficients ~0.7 to 0.75 compared to experimentally measured values) but the overall distributions have high information content.

In contrast with larger and more comprehensive datasets used in previous work (all substitutions at all sites in a protein), approximately one substitution per residue per study protein was sampled in the SARS-CoV-2 dataset downloaded from GISAID. To investigate whether the observed stability effects follow a similar distribution, we fit bi-Gaussian models to $\Delta\Delta G^{App}$ histograms for all USVs for all 19 study proteins with experimentally-determined structures (Figure 4). The bi-Gaussian distribution fits the calculated stability distributions better than a single Gaussian ($R^2 = 0.924$ for a bi-Gaussian and $R^2 = 0.769$ for a single Gaussian, not plotted). Individual Gaussian peaks correspond closely to the energetic impacts of surface and core substitutions, respectively (Figure 4). This trend was observed for both types of Rosetta-based stability calculations, including those in which a dampened repulsive van der Waals potential was used during the rotamer optimization step. For each calculation type, the mean destabilization calculated for the core substitution distribution is smaller than the mean value associated with the second Gaussian peak observed in the full set of substitutions, possibly due to contributions to the second peak from destabilizing boundary layer substitutions that shift the mean to higher values (and possibly to limitations of the sampling and scoring approach discussed above). Bi-Gaussian fits to $\Delta\Delta G^{App}$ distributions for each of the 29 study proteins considered individually (Supplementary Table Gaussian) show

similarly good fits for bi-Gaussian functions for globular study proteins. Robustness with respect to destabilizing effects of amino acid changes both limits and promotes viral evolution. It is, therefore, remarkable that the observed variation in the SARS-CoV-2 proteome over the first 6 months of the pandemic follows this universal trend, speaking perhaps to the relative rapidity of viral evolution due to large population sizes and imperfect replication machinery.

2.4 | Analyses of study proteins

The sections that follow provide more detailed results and discussion pertaining to USVs identified for 13 of the 29 SARS-CoV-2 study proteins, including one validated drug target [RNA-dependent RNA polymerase (RdRp, nsp7/nsp8$_2$/nsp12 heterotetramer)], five potential small-molecule drug discovery targets [papain-like proteinase (PLPro, part of nsp3), main protease (nsp5), RNA helicase (nsp13), proofreading exoribonuclease (nsp14), and methyltransferase (nsp10/nsp16 heterodimer)], plus the four structural proteins comprising the virion [spike S-protein, nucleocapsid N-protein, pentameric ion channel E-protein, and integral membrane M-protein]. Analysis results obtained for USVs of the remaining study proteins are provided in Supplementary Materials together with additional information regarding all 29 study proteins.

2.5 | Nonstructural proteins 7, 8, and 12 (nsp7/nsp8$_2$/nsp12)

The RNA-dependent RNA polymerase (RdRp) is a macromolecular machine made up of four protomers, including nsp7, two asymmetrically bound copies of nsp8, and the catalytic subunit nsp12. The resulting heterotetramer is responsible for copying the RNA genome and generating nine subgenomic RNAs.[19] nsp12 consists of three globular domains: an N-terminal nidovirus RdRp-associated nucleotidyltransferase (NiRAN), an interface domain, and a C-terminal RdRp domain. The active site of nsp12 includes residues Thr611 to Met626 (TPHLMGWDYPKCDRAM) comprising Motif A.[20] nsp12 binds to one turn of double-stranded RNA, and residues D760 and D761 bind to the 3′ end of the RNA and are essential for RNA synthesis.[21] The RNA duplex is flanked by α-helical arms formed by N-terminal segments of the two nsp8 protomers, which appear to grip the RNA and prevent its premature dissociation from the RdRp (i.e., confer processivity). Positively charged residues of nsp8 occurring within the RdRp-RNA interface include K36, K37, K39, K40, K46, R51, R57, K58, and K61. Of these, K58 interacts with the RNA duplex emerging from the active site. Any change of this residue in nsp8 yields a replication-incompetent virus.[21] Since deposition of PDB ID 6M71,[20] a plethora of RdRp structures has become available from the PDB.

Following US Food and Drug Administration (FDA) approval for remdesivir, RdRp can be reasonably regarded as being a validated drug target for treatment of SARS-CoV-2-infected individuals. Structures of SARS-CoV-2 RdRp containing incorporated remdesivir (PDB ID 7BV2[22] and PDB ID 7C2K[23]) help explain the drug's mechanism of action via delayed-chain termination[24] and provide a valuable starting point for design of second-generation RdRp inhibitors that are more potent and more selective and possibly orally bioavailable. Residues K545, R553, D623, S682, T687, N691, S759, D760, and D761 in nsp12 interact directly with remdesivir,[22] while S861 may be involved in a steric clash with the 1′-CN group of remdesivir, possibly perturbing the position of the RNA duplex.[23] Knowledge of structures of remdesivir-RdRp complexes will also provide valuable insights into potential sources of drug resistance.

The experimental structure of the RdRp-duplex RNA complex (PDB ID 6YYT[21]) was used for evolutionary analyses of nsp7, nsp8, and nsp12 (Figure 5A,B). Each protomer is considered in turn below.

2.5.1 | nsp7

Sequencing of 48 670 viral genomes identified 47 876 unchanged sequences and 68 USVs of nsp7 versus the reference sequence, with 66 single and two double substitutions (Tables 1 and 2). Most substitutions occurred in only one USV (∼91%). The most frequently observed USV for nsp7 (S25L; nonconservative, surface) was detected 562 times in the GISAID dataset.

2.5.2 | nsp8

Sequencing of 48 335 viral genomes identified 47 635 unchanged sequences and 144 USVs of nsp8 versus the reference protein sequence, with 140 single, two double, one triple, and one quintuple substitutions (Tables 1 and 2). Most substitutions occurred in only one USV (∼99%). The most frequently observed USV for nsp8 (M129I; conservative, core) was detected 124 times in the GISAID dataset. No substitutions of the essential RNA-binding residue K58 were observed. Of the remaining eight positively charged residues that face the RNA duplex, substitutions were observed for five, including K37, K40, R51, R57, and K61 (both R51L and R57L preclude salt bridge formation with RNA). Substitutions of R51 were observed in three different USVs, occurring as three distinct substitutions (R51L, R51C, R51H). Another interesting nsp8 USV is the singly observed quintuple substitution USV occurring within the N-terminal arm (A74S/S76C/A81S/V83L/S85M). This USV may be the result of a sequencing artifact, as none of the five substitutions were observed in any other USV. One other USV exhibits adjacent amino acid changes: M90S/L91F. This pair of residues occurs at the interface with nsp12 for one nsp8 protomer and near a shared interface with nsp7 and nsp12 in the other copy.

2.5.3 | nsp12

Sequencing of 44 203 viral genomes identified 10 266 unchanged sequences and 730 USVs of nsp12 versus the reference sequence,

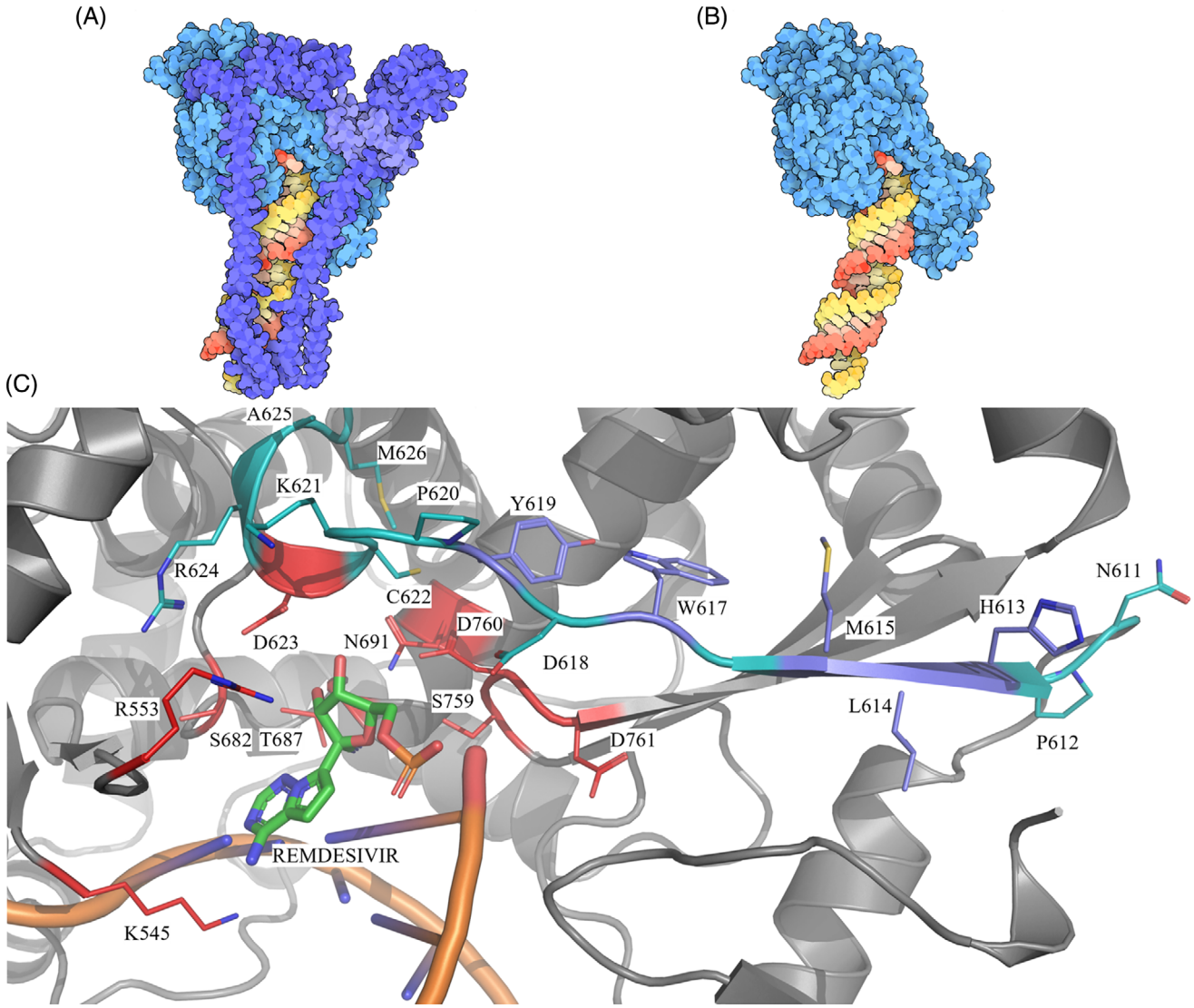

FIGURE 5 (A) Space-filling representation of the experimental structure of the nsp7/nsp8$_2$/nsp12 heterotetramer bound to double-stranded RNA (PDB ID 6YYT[21]) viewed into the enzyme active site on the anterior surface of nsp12. (B) Identical view of PDB ID 6YYT with nsp7 and nsp8 removed to reveal interactions of nsp12 with RNA. Protein color coding: nsp12-light blue; nsp8-dark blue; nsp7-blue/gray; RNA color coding: template strand-shades of red; product strand-shades of yellow. (C) Ribbon/atomic stick figure representation of the active site of nsp12 (PDB ID 7BV2[22]; mostly gray) occupied by the RNA template:product duplex (backbone shown as tubes, bases shown as sticks, colored in shades of orange) with remdesivir (shown as an atomic stick figure following enzymatic incorporation into the RNA product strand; atom color coding: C-green, N-blue, C-red, S-yellow). The active site Motif A is colored coded magenta (atom color coding for invariant residues: C-magenta, N-blue, O-dark red) and purple (atom color coding for substituted residues: C-purple, N-blue, O-dark red, S-yellow). Residues making direct or water mediated contacts with remdesivir are colored light red (atom color coding: C-light red, N-blue, O-dark red, S-yellow)

with 249 single, 424 double, 51 triple, 3 quadruple, and 3 multipoint substitutions (Tables 1 and 2). Most substitutions occurred in only one USV (~74%). More than 97% (count~32 000) of the ~44 000 GISAID dataset nsp12 sequences differing from the reference sequence carried the same P323L substitution. This substitution constitutes a distinct nsp12 clade that was first detected in the United Kingdom in January 2020 and subsequently in many other countries around the world.

Approximately 61% of the observed amino acid substitutions were nonconservative (364 nonconservative versus 228 conservative), with most of the nonconservative changes occurring in the boundary and surface portions of the 3D structure. (N.B.: Only 60 point substitutions map to the protein core.) Two of the multipoint substitutions (A97V/S520I/E522D/D523Y/A529S/L829I and T85S/I201F/V202F/V330E/I333T) were observed only once. In both cases, all substitutions were unique to that USV, suggesting that they are both the result of sequencing artifacts.

2.5.4 | nsp7/nsp8$_2$/nsp12 energetics

The vast majority of the USVs (83%) were estimated to be moderately less stable than the reference sequence ($<\Delta\Delta G^{App}> \sim +7.6$ REU). In fewer than 4% cases, the estimated change in apparent free energy of

stabilization change exceeded +19.1 REU. A minority of the USVs (~13%) were estimated to be more stable than the reference sequence (<$\Delta\Delta G^{App}$ > ~ −2.2 REU). (N.B.: Hereafter, references will be made to Tables 1 and 2 to avoid repeating the same text summarizing amino acid substitutions and energetics analyses for each of the remaining study proteins.)

2.5.5 | nsp12 active site

Of the residues in active site Motif A (Figure 5C), substitutions were observed in residues H613, L614, M615, W617, Y619, and A625 (Figure 5C). It is remarkable that all six of these residues are oriented toward the hydrophobic core of the protein, away from the active site, and should, therefore, not disrupt catalysis. No substitutions were observed for nsp12 residues that interact directly or via bridging water molecules with remdesivir (Figure 5C; K545, R553, D623, N691, D760, S759, D760).

2.5.6 | Protein–protein interfaces

The four protomers forming the RdRp heterotetramer bury significant numbers of residues within the various protein–protein interfaces. It is, therefore, difficult to be certain that a distal substitution might not have a steric influence on one or more of these interfaces. Below, we enumerate substitutions with the potential for direct effects on interfacial contacts.

Eleven substitutions involving the following six nsp7 residues could affect binding to nsp12: K7, L14, S15, S26, L40, and L41. Seven of these 11 substitutions were conservative. nsp12 substitutions at the following sites could affect binding to nsp7: T409, P412, F415, Y420, E436, A443, and D445. Y420S would break an observed hydrogen bond with D5 of nsp7. E436G/K would break an observed salt bridge with K43. Many of the nsp7 and nsp12 substitutions occurring within their contact interface were highly destabilizing, with seven giving $\Delta\Delta G^{App}$ > +10 REU.

nsp7 makes minimal contact with one copy of nsp8. Observed nsp7 substitutions at residues S25 (S25L) and S26 (S26A and S26F) involve exchange of serine for a hydrophobic residue. Both substitutions at S26 break an observed hydrogen bond with D163 of nsp8. No nsp8 D163 substitutions were identified.

The contact surface of nsp7 with the second copy of nsp8 is more extensive than with the first. nsp7 substitutions occurring within this inter-subunit interface include residues V6, T9, S15, V16, L20, L28, Q31, F49, E50, M52, S54, L56, S57, V58, L60, S61, V66, I68, and L71 (17/27 substitutions affecting all 19 nsp7 residues were conservative). S54P is a noteworthy amino acid change that inserts a Proline into the middle of an interfacial α-helix. Substitutions of the following nsp8 residues may affect binding to nsp7: residues V83, T84, S85, T89, M90, L91, M94, L95, N100, A102, I107, V115, P116, I119, L122, V131, and A150 (14 of the 21 substitutions involving these 17 sites were conservative).

Because the two nsp8 chains occur in asymmetric environments, a given substitution may alter one interface or the other, or both. Substitutions at 23 sites could affect the nsp8-nsp12 interface for one of the chains (T84, A86, L91, L95, N104, I107, V115, P116, I119, P121, L122, T123, K127, M129, V131, I132, P133, T141, A150, W154, V160, W182, and T187). Substitutions at five sites (T68, K72, R75, S76, and K79) could affect only the nsp8-nsp12 interface with the chain that wraps around nsp7. Substitutions at three sites (V83, M90, and M94) could affect both interfaces. Of these 38 substitutions across 31 sites, 19 were conservative. A P121S substitution in nsp8 could give rise to a backbone hydrogen bond with V398 of nsp12. Two Tryptophan to Cysteine substitutions (W154C and W182C) occurring in nsp8 were extreme outliers with $\Delta\Delta G^{App}$ > +30 REU, suggesting that some backbone rearrangement is necessary in response to exchange of the large Tryptophan side chains for smaller Cysteines.

In nsp12, substitutions of 25 residues could affect the interface with the first nsp8 protomer (L270, P323, T324, P328, L329, V330, V338, F340, P378, A379, M380, A382, A383, N386, V398, A399, V405, F407, W509, L514, S518, M519, S520, D523, and V675). Substitutions of 10 residues in nsp12 could affect the interface formed with the second copy of nsp8 (N414, F415, D846, I847, V848, T850, M899, M902, M906, T908). No nsp12 substitutions appear to affect contacts with both copies of nsp8. Of the 50 observed nsp12 substitutions occurring at 35 sites, 26 were conservative. The clade-defining nsp12 P323L substitution occurs at the C-terminus of an α-helix within the smaller interface between nsp12 and the first nsp8 protomer. While the structural consequences of this P→L substitution appear negligible, the computed ΔΔGApp ~8 REU. This apparent discrepancy almost certainly reflects limitations in the Rosetta energetics calculation.

2.6 | Nonstructural protein 3 papain-like proteinase (PLPro)

The papain-like proteinase (PLPro) is a 343-residue segment occurring within the 1945 residue multidomain protein nsp3. It is one of two viral proteases responsible for processing of the polyprotein products of translation of the viral genome following infection. This enzyme cleaves the polyproteins pp1a and pp1ab at three sites (black inverted triangles in Figure 2): the nsp1/nsp2 junction and its own N- and C-termini. These three cleavage events liberate nsp1, nsp2, and nsp3. The PLPro portion of nsp3 is also implicated in cleaving post-translational modifications of ubiquitin (Ub) and ISG15 domains of host proteins as an evasion mechanism against host antiviral immune responses.[25]

PLPro is a cytoplasmic cysteine endopeptidase (EC 3.4.22.69) that catalyzes cleavage of the peptide bond C-terminal to LXGG motifs (where X is any amino acid) in the viral polyproteins. This enzyme also recognizes conserved LRGG motifs found within the C-terminal segments of Ub and ISG15 proteins. According to the MEROPS classification, PLPro belongs to the peptidase clan CA (family C16), containing

a Cys-His-Asp catalytic triad (C111–H272–D286). The first structure of SARS-CoV-2 PLPro to be made public (PDB ID 6W9C[26]) revealed a symmetric homotrimer with each enzyme monomer being highly similar to that of SARS-CoV-1 PLPro (PDB ID 2FE8[27]; r.m.s.d. ~ 0.8 Å, sequence identity~83%). Since PDB release of this initial SARS-CoV-2 PLPro structure, additional co-crystal structures of PLPro with a variety of ligands have been deposited to the PDB (list updated weekly at http://rcsb.org/covid19). In many of these structures, the enzyme is monomeric, indicating that the trimer observed in PDB ID 6W9C is almost certainly a crystal packing artifact. Comparison of the various PLPro monomer structures reveals that the enzyme does not undergo large conformational changes upon binding of inhibitors or (protein) substrates (Figure 6A). We, therefore, used the structure of an inhibited form of the enzyme (PDB ID 6WUU[28]) for evolutionary analyses of PLPro (Figure 6A).

Overall substitution trends for PLPro and energetics analysis results are summarized in Tables 1 and 2. P1640L (nonconservative, surface) and T1626I (nonconservative, surface) are the two most common USVs, observed in 48 and 47 GISAID dataset sequences, respectively. No amino acid substitutions were identified in the enzyme active site—the catalytic triad is fully preserved in all observed USVs. However, examination of apo- and inhibitor/substrate-bound structures indicates that several substitutions occur in the ISG15- and ubiquitin-binding regions of PLPro. These substitutions (e.g., F1632S, D1624G, D1625H, S1633G) mapping to the S2 and S4 α-helices of PLPro (Figure 6B) may alter the binding affinity and specificity of PLPro for interactions with host protein substrates. In cell-based assays, the interactome of SARS-CoV-2 PLPro appears to be significantly different from that of SARS-CoV-1 PLPro. SARS-CoV-2 PLPro prefers ISG15 binding to Ub whereas SARS-CoV-1 PLPro prefers Ub binding to ISG15.[29] The S2 and S4 regions are interaction hotspots in the interfaces of PLPro with ISG15 and Ub. Amino acid changes in these regions may change the protein's interactome. Finally, two observed substitutions affecting active-site

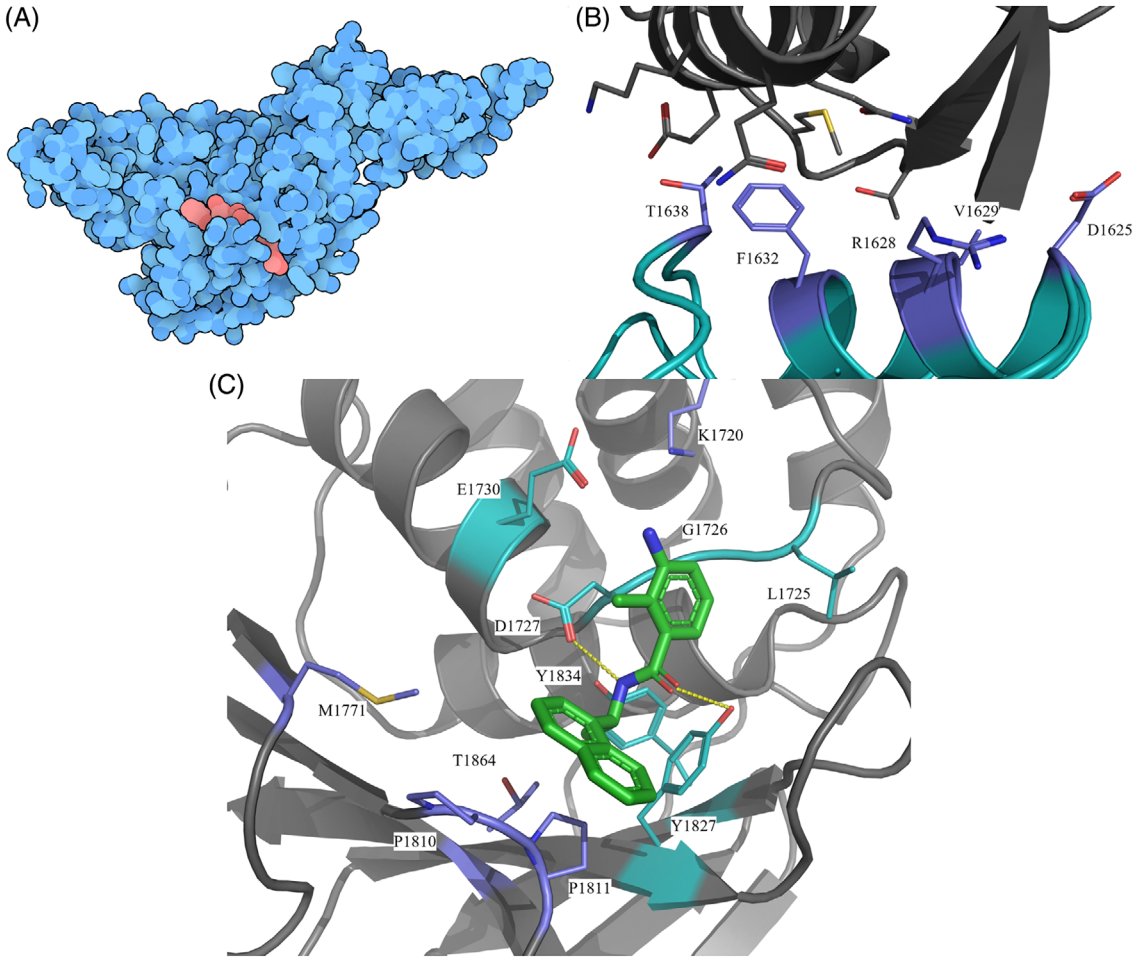

FIGURE 6 (A) Space-filling representation of the experimental structure of the PLPro monomer (blue) bound to a covalent inhibitor (Vir250; red/pink) (PDB ID 6WUU[28]). (B) Ribbon/atomic stick figure representation of the PLPro-ISG15 interface (PDB ID 6YVA[29]). Oxygen atoms are shown in red, nitrogens in blue, and sulfurs in yellow. Cartoons and carbons are gray for ISG15, purple for substituted PLPro interfacial residues, and cyan for all other PLPro residues. (C) Ribbon/atomic stick figure representation of PLPro active site (color coding as for (B)) occupied by a non-covalent inhibitor (GRL0617) shown as an atomic stick figure (atom color coding: C-green, N-blue, O-red, H-bonds-dotted yellow lines; PDB ID 7JN2[30])

proximal proline residues P1810S and P1811S may affect inhibitor binding, either by altering the backbone flexibility of the binding pocket loop or through repulsion of the hydrophobic portion of the inhibitor. They thus represent potential sites of drug resistance mutations (Figure 6C), though if they become prevalent, might also become targets for polar interactions when designing future inhibitors.

2.7 | Nonstructural protein 5 main protease (nsp5)

nsp5 is the other viral protease responsible for processing the viral polyproteins (synonyms: main protease, 3CL protease). This enzyme cleaves the longer polyprotein pp1ab at 11 sites (light blue inverted triangles in Figure 2), beginning with liberation of its own N-terminus and concluding with separation of nsp15 from nsp16 near the C-terminus of the polyprotein. nsp5 is a 306-residue cysteine endopeptidase (EC 3.4.22.69) that catalyzes cleavage of sites similar to TSAVLQ/SGFRK (where "/" denotes the cleavage site). Conserved residues Histidine 41 (H41) and Cysteine 145 (C145) constitute the catalytic dyad.[31] The first structure of SARS-CoV-2 nsp5 deposited into the PDB (PDB ID 6LU7[8]; Figure 7) revealed a symmetric homodimeric structure extremely similar to that of its SARS-CoV-1 homolog (r.m.s.d. ~ 0.8 Å, sequence identity>95% with PDB ID 1Q2W[6]). Since PDB release of this initial nsp5 structure, ~200 co-crystal structures of nsp5 with a variety of small chemical fragments and larger ligands have been deposited to the PDB (updated weekly at http://rcsb.org/covid19). Open access to this wealth of structural information spurred the launch of an international COVID-19 Moonshot effort to discover and develop drug-like inhibitors.[32] The apo nsp5 structure (PDB ID 6YB7) was used for the evolutionary analyses that follow (Figure 7A).

Overall substitution trends for nsp5 and energetics analysis results are summarized in Tables 1 and 2. G15S (nonconservative, boundary) is the most common USV, observed in 1082 sequences. The most striking change observed in the GISAID dataset involves H41, the catalytic Histidine (Figure 7B, shown in red) substitution of which is expected to eliminate catalytic activity. This substitution was detected in the H41P/L50H double substitution. It is unlikely that the loss of H41 has been compensated by the L50H substitution, given that the distance between L50 and the active site (L50:Cα-C145:Cα~16 Å versus H41:Cα-C145:Cα~7 Å) would require significant backbone rearrangement. Only one viral genome with this USV was detected in the GISAID dataset, which raises the possibility that it represents a sequencing artifact. No other observed USVs included substitutions of residue L50 to Histidine, but other amino acid changes at that site were observed within the GISAID dataset. Experimental characterization of the enzymatic activity of the H41P;L50H double substitution would resolve the issue.

Several amino acids within or adjacent to the substrate binding groove underwent substitutions (Figure 7B, shown in purple) that may affect substrate binding, including T25, M49, M165, E166, 168, 188, 189, and A191. The most dramatic alteration to the active site occurs in the triple substitution M165L;E166V;A191E. E166 lines the active site cleft, where it is thought to form a hydrogen bond with the prescissile residue of the substrate. The same residue also appears to interact with the N-terminus of the homodimeric partner. Each of these substitutions is unique to a single USV, occurring only once in the GISAID dataset. Other substitutions were observed at residues 165 and 191 in other USVs.

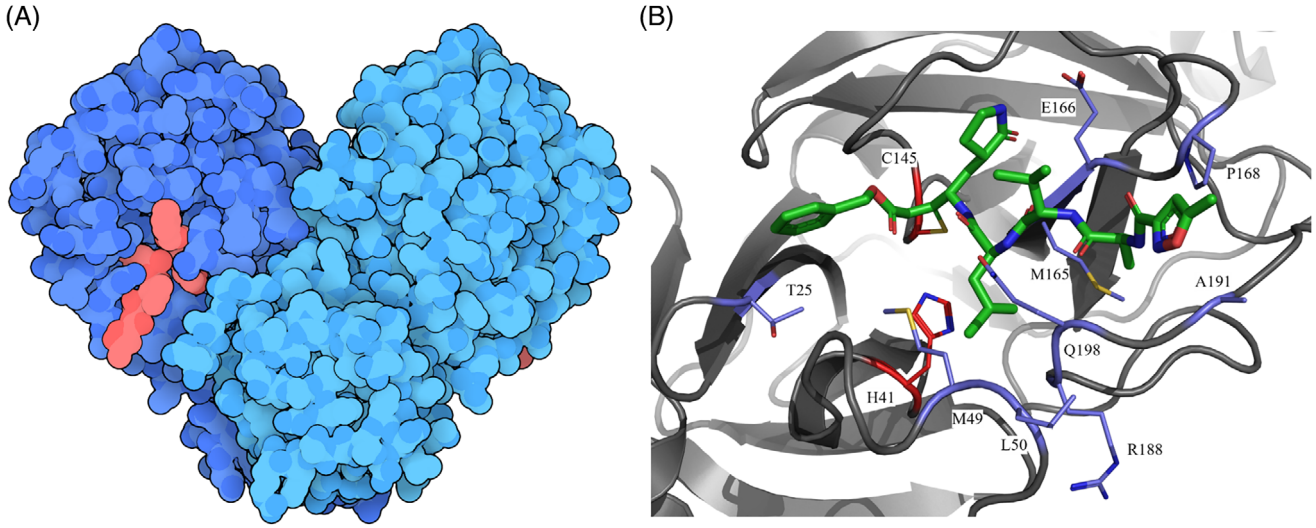

FIGURE 7 (A) Space-filling representation of the experimental structure of the nsp5 homodimer covalently bound to a substrate analogue inhibitor (PDB ID 6LU7[8]). Color Coding: nsp5 monomers-light and dark blue; substrate analogue PRD_002214 (https://www.rcsb.org/ligand/PRD_002214)-red. (B) Ribbon/atomic stick figure representation of the active site of nsp5 (gray) occupied by PRD_002214 covalently bound to C145 (atom color coding: C-green, N-blue, O-red). Catalytic residues H41 and C145 denoted with red ribbon and atomic stick figure sidechains (atom color coding: C-light red, N-blue, S-yellow). Substituted active site residues denoted with purple ribbon and atomic stick figures (atom color coding: C-purple, N-blue, O-red, S-yellow)

A number of residues occurring near the dimerization interface were also substituted, including residues M6, A7, G71, A116, S121, V125, G170, G215, M276, G278, S284, A285, Q299, G302, and T304, any one of which could affect dimerization. In several cases, Glycine residues at the interface were substituted for larger hydrogen-bonding residues, for example, G71S, G170R, G215R, and G278R. While total stability was reduced, dimer interface stability was increased in all cases except G278R (see Supplementary Table: nsp5 interfacial energies). Interestingly, all substitutions mapping to the dimer interface occurred in USVs lacking any other substitutions.

Finally, there were four cases in which substitutions to Proline (a helix breaking amino acid) occurred at positions falling within α-helical or β-strand secondary structural elements (K90P, S123P, A206P, S301P). The latter three represent the most extreme energetic outliers of all USVs, and all four were observed only once in the GISAID dataset. S123P occurs within a β-strand at the dimeric interface near the C-terminus of the homodimeric partner. The calculated destabilization resulting from these substitutions introducing Proline residues in the context of the crystal structure suggest that these variants may lead to backbone structural changes.

2.8 | Nonstructural protein 13 (nsp13)

nsp13 plays a central role in viral replication by unwinding RNA secondary structure within the 5′ untranslated region of the genome.[33] The enzyme is NTP-dependent and is also known to exhibit 5′-triphosphatase activity. nsp13 is most active in the presence of the RdRp, which suggests that the helicase is required for high-efficiency copying of the viral genome.[34] A previously published 3D electron microscopy (3DEM) structure of the nsp7-nsp8$_2$-nsp12/nsp13$_2$ heterohexamer provide a structural model for how two copies of the helicase could interoperate with RdRp during RNA synthesis (PDB ID 6XEZ[35]).

nsp13, a member of helicase superfamily 1, consists of 596 amino acid residues. It adopts a triangular pyramid-like structure consisting of five domains (Zn^{++}-binding, stalk, 1B, 1A, and 2A), with each domain directly or indirectly involved in the helicase function. There are three Zn^{++}-binding sites located within the N-terminus of the enzyme, involving conserved cysteine and histidine residues (Zn^{++}-1: C5, C8, C26, C29; Zn^{++}-2: C16, C19, H33, H39; Zn^{++}-3: C50, C55, C72, H75). NTPase activity is mediated by six conserved residues situated at the base of the 1A and 2A domains (K288, S289, D374, E375, Q404, R567). The nucleic acid binding channel is formed by domains 1B, 1A, and 2A.[36] Sequence alignment of SARS-CoV-1 nsp13 with SARS-CoV-2 nsp13 revealed near-perfect identity with a single amino acid difference (I570V). The experimental structure of SARS-CoV-1 nsp13 (PDB ID 6JYT[36]) provided the template for Rosetta computation of the SARS-CoV-2 nsp13 homology model used to analyze its evolution in 3D (Figure 8).

Overall substitution trends for nsp13 and energetics analysis results are summarized in Tables 1 and 2. The double substitution P504L;Y541C is the most common nsp13 USV, observed 1607 times in the GISAID dataset. No substitutions were observed for 11 of the

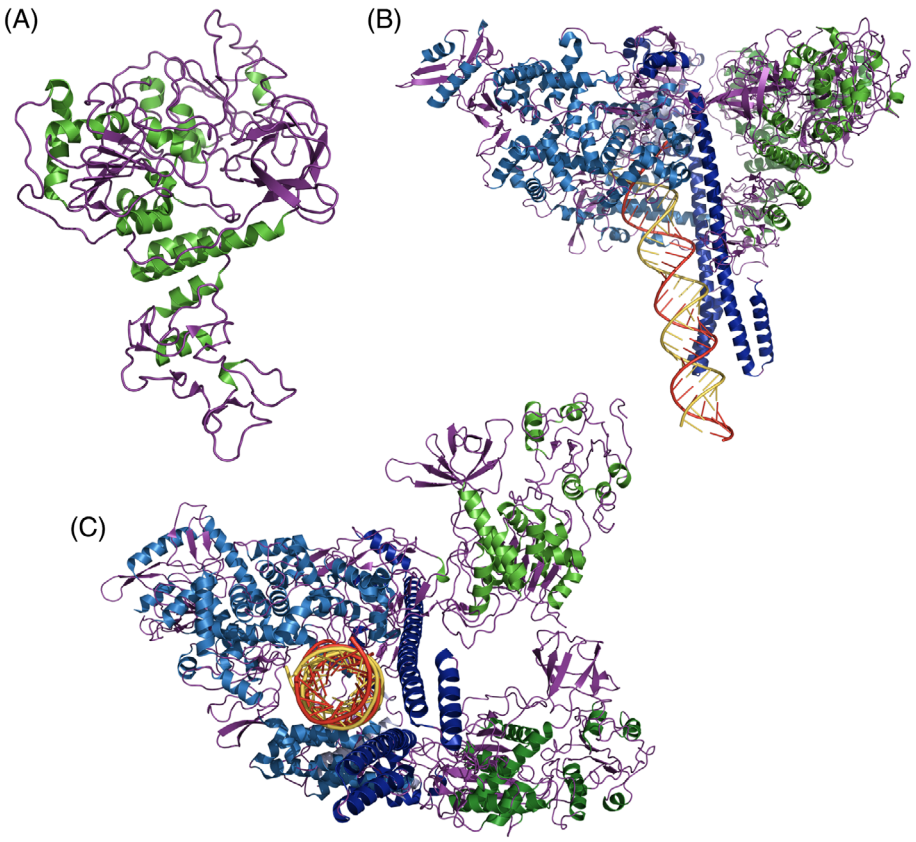

FIGURE 8 (A) Ribbon representation of the computed structural model of nsp13 (green; based on PDB ID 6JYT[36]). The RNA helicase active site is located in the upper half of the protein. (B) Ribbon representation of the experimental structure of the nsp13$_2$-nsp7/nsp8$_2$ /nsp12 heterohexamer (PDB ID 6XEZ[35]), viewed to show the RNA double helix, and (C) viewed looking down the RNA helix axis, showing the two helicase active sites presented to the RNA. (color coding for B and C: nsp13-green, otherwise same color coding as Figure 5)

12 Zn^{++}-binding residues. A single substitution was observed for Histidine 33 changing to Glutamine (H33Q), which appears unlikely to abrogate binding of Zn^{++}. Potentially important amino acid substitutions involve R337 and R339, two residues known to support helicase activity that are positioned at the entrance of the nucleic acid binding channel. Substitutions were observed in the R337L;A362V and R339L USVs. A SARS-CoV-1 R337A;R339A double substitution showed decreased helicase activity.[36] It is, therefore, likely that R337L and R339L substitutions in SARS-CoV-2 nsp13 reduced enzyme activity. Another interesting substitution involves the R567, which is important for NTP hydrolysis in SARS-CoV-1 nsp13.[36] An R567I substitution occurs in the context of the double substitution USV (V456F;R567I; GISAID dataset count = 1) and may reduce SARS-CoV-2 nsp13 helicase activity.

2.9 | Nonstructural protein 14 proofreading exoribonuclease (nsp14)

nsp14 is a 527-residue protein that acts as both a proofreading exoribonuclease and a methyltransferase to synthesize the N7-methyl-guanine cap 5′ for the mRNA-like genome.[37,38] It is encoded as part of polyprotein pp1ab and is excised by nsp5. Following excision, it is thought to form a 1:1 complex with nonstructural protein 10 (nsp10) to proofread newly formed RNAs synthesized by the RdRp heterotetramer.[39] (N.B.: nsp10 also forms a heterocomplex with nsp16, for which there is an experimental structure available from the PDB [see nsp10/nsp16 section below]). At the time of writing there were no publicly available structures of SARS-CoV-2 nsp14. A computed homology model was used to analyze the evolution of nsp14, based on the structure of SARS-CoV-1 nsp14 (PDB ID 5C8S[40]), with which it shares ~95% sequence identity (Figure 9). Superposition of the methyltransferase catalytic centers of SARS-CoV-2 nsp14 and SARS-CoV-1 nsp14 revealed 100% conservation of active site residues, including both the cap binding residues (N306, C309, R310, W385, N386, N422, and F426) and the S-adenosyl methionine (SAM) binding residues (D352, Q354, F367, Y368, and W385). The active site of the exoribonuclease proofreading domain of nsp14 contains a D-E-D-D-H motif (D90, E92, D243, D273, H268), which is identical to the corresponding motif found in SARS-CoV-1 Nps14.[40]

Overall substitution trends for nsp14 and energetics analysis results are summarized in Tables 1 and 2. A320V (conservative, core) was the most common substitution, occurring in six USVs with a total GISAID dataset count of 327. A320V also occurred in four double substitution USVs (A320V/D496N, A320V/K349N, A320V/P355S, A320V/A323S). F233L (conservative, core) was the second most common substitution, occurring in 4 USVs, and observed in 273 independently sequenced genomes. It occurred in both a single substitution USV (F233L) and in three double substitution USVs (F233L/A360V, A23S/F233L, F233L/S461P). Two USVs (sequenced in same geographic location) had surprisingly large numbers of amino acid changes and very large $\Delta\Delta G^{App}$ values. The first had five substitutions (T193K/D352E/D358E/Y361K/E364Q), none of which were observed in single substitution USVs. The other had 14 substitutions (Y64F/N67Y/Y69F/P70L/N71Y/M72L/I74F/E77V/I80F/R81S/H82L/V83F/W86C/I87F) with only P70L being observed in another USV as a single substitution. Given the large number of substitutions, extremely unfavorable apparent stabilization energy changes ($\Delta\Delta G^{App}$ ~ 20 REU and ~56REU, respectively), and the fact that they were detected only once, we believe that both USVs are the result of sequencing artifacts. No substitutions were observed within the active site of the exoribonuclease proofreading domain. The methyltransferase domain displayed a high level of conservation with only three of 12 active site residues substituted. Two guanine cap binding residues (N306 and F426) were found substituted, with N306S (conservative, surface) observed as a single amino acid change and F426L observed once in the double-substitution USV F426L;S448Y. One SAM binding residue was substituted: Q354H (nonconservative, boundary) was observed in five independently sequenced viral genomes.

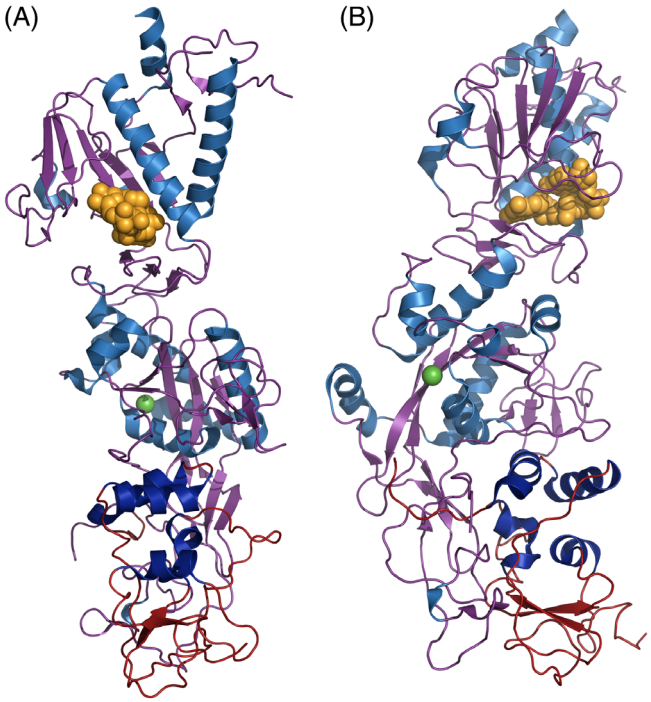

FIGURE 9 (A) Ribbon representation of the computed structural model of the nsp10/nsp14 heterodimer bound to GpppA and S-adenosyl homocysteine (based on PDB ID 5C8S[40]). (B) Rotated 90° about the vertical. Color coding: nsp14-light blue (α-helices) and purple (β-sheets and loops); nsp10-dark blue (α-helices) and red (β-sheets and loops); GpppA-yellow/orange; Exoribonuclease active site Mg^{++} cation: green

While we did not generate structural models of the nsp14/nsp10 heterodimer, the structure of SARS-CoV-1 nsp14/nsp10 heterodimer (PDB ID 5C8S[40]) allowed us to predict which SARS-CoV-2 amino acid changes may affect nsp10/nsp14 heterodimer formation. Sixteen nsp14 sites of substitution (T5, P24, H26, L27, K47, M62, N67, Y69, V101, N129, T131, K196, V199, I201, P203, and F217, giving a total of 21 distinct substitutions) and eight nsp10 sites of substitution (T12,

A18, A20, Y30, A32, I81, K93, and K95, giving a total of 13 distinct substitutions) were mapped to the putative nsp14/nsp10 interface, of which 18 were conservative and 16 were nonconservative. The most prevalent substitutions were T12 (surface, T12I and T12N), A32 (surface, A32S and A32V), H26Y (surface), and P203 (surface, P203L and P203S).

2.10 | Nonstructural proteins 10 and 16 methyltransferase (nsp10/nsp16)

Nonstructural proteins nsp10 and nsp16 are both found within pp1ab, from which they are excised by nsp5. Together, nsp10 and nsp16 form a stable heterodimer that functions as a methyltransferase, acting on the 2′ OH of the ribose of the first nucleotide of the viral genome (i.e., 5′(m7Gp)(ppAm)[pN]$_n$, where Am denotes 2′-O-ribose methyl-adenosine). This process renders the viral cap structure indistinguishable from that of eukaryotic cap-1, thereby disguising the viral genome so that it resembles cellular RNAs typically found in multicellular organisms and protecting the viral genome from cellular 5′ exonucleases. Enzyme activity of nsp16 depends on SAM as a cofactor, which donates the methyl group from the methionine group for transfer to the ribose of the capped viral RNA.[41] (N.B.: Capping of the viral RNA is carried out by the N7-guanine methyltransferase domain of nsp14[40]). The structure of the SARS-CoV-2 nsp10/nsp16 heterodimer (PDB ID 6WVN[42]) revealed a heterodimer extremely similar to that of its SARS-CoV-1 homolog (sequence Identities ~93% (for nsp10) and ~98% (for nsp16); r.m.s.d. ~ 1.1 Å for PDB ID 6WVN[42] versus PDB ID 2XYQ[43]).

The SAM binding site includes residues N43, G71, G73, G81, D99 (3 interactions), D114, C115, D130, and M131.[44] The N7-methyl-GpppA binding site consists of residues K24, C25, L27, Y30 (2 interactions), K46, Y132, K137 (2 interactions), K170, T172, E173, H174, S201 (2 interactions), and S202 (4 interactions). Efficient catalytic activity of nsp16 depends on heterodimerization with nsp10, which possesses two zinc-binding motifs (PDB ID 6ZCT[45]). The two Zn^{++}-binding sites of nsp10 are composed of residues C74, C77, H83, and C90, and C117, C120, C128, and C130, respectively.

Polar interactions within the nsp10/nsp16 interface include nsp10:L45-nsp16:Q87; nsp10:G94-nsp16:R86; nsp10:K93-nsp16:S105; nsp10:K43-nsp16:K138; nsp10:Y96-nsp16:A83; and nsp10:A71/G94-nsp16:D106. There is also a salt bridge between H80 and D102 in the SARS-CoV-1 nsp10/nsp16 heterodimer.[41] At the time of analysis, there was one PDB structure of SARS-CoV-2 nsp10 alone (PDB ID 6ZCT[45]). A dozen co-crystal structures of the SARS-CoV-2 nsp10/nsp16 heterodimer are available from the PDB, together with nearly 20 structures of nsp10/nsp16 from SARS-CoV-1 and MERS CoV. In the case of SARS-CoV-1, nsp10 also forms a heterodimer with nsp14 (e.g., PDB ID 5C8S[40]). Evolutionary analyses of the nsp10/nsp16 heterodimer that follow were carried out using PDB ID 6WVN[42] (Figure 10).

Overall substitution trends for nsp10 and nsp16 and energetics analysis results are summarized in Tables 1 and 2. Several observed substitutions are noteworthy. Two USVs involving SAM binding residues in nsp16 include D99N (nonconservative; core) and D114G (nonconservative; surface), both of which may alter binding affinity to the SAM moiety due to loss of the negative charge upon substitution. Indeed, modeling indicates reduced stability ($\Delta\Delta G^{App}$ ~ 7REU in the case of D114G). M131I (conservative; boundary) may also affect SAM binding. By perturbing SAM binding, these substitutions may influence the ability of the enzyme to methylate the first ribose of the viral cap, although these predictions await experimental testing. USVs involving 7-methyl-GpppA binding residues in nsp16 include K24N (nonconservative; surface), D75Y (nonconservative; surface), and S202F (nonconservative; boundary). All these substitutions had destabilizing effects, with $\Delta\Delta G^{App}$ > 7 REU for S202F. D75Y appears to form a new hydrogen bond with the 7-methyl-GpppA, which would slightly shift its position in the binding pocket (Figure 10). Only one nsp10 USV affected the Zn^{++}-binding residue C130 (C130S;D131H), which would be unlikely to abrogate cation binding.

A number of sites near the protein–protein interface were also substituted, any one of which may affect heterodimer stability, including nsp10 residues K43, T47, T58, F68, and K93; and nsp16 residues P37, G39, M41, V44, T48, G77, V78, P80, R86, T91, D108, T110, M247, and P251. Nine of the interfacial substitutions were conservative and mildly destabilizing, although nsp16 M247I had a more pronounced effect with $\Delta\Delta G^{App}$ > 10 REU. Of the 16 nonconservative interfacial substitutions

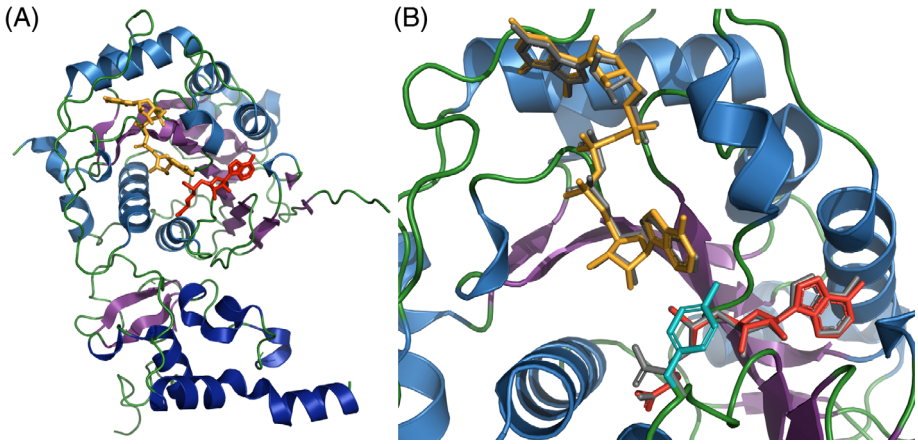

FIGURE 10 Ribbon and stick figure representation of the experimental structure of the nsp10 (dark blue)/nsp16 (light blue) heterodimer bound to N7-methyl-GpppA and SAM (PDB ID 6WVN[42]). Color coding: β-sheets—purple; loops—green; nsp16 α-helices—light blue; nsp10 α-helices—dark blue; N7-methyl-GpppA—yellow; SAM—red. Left: full complex. Right: active site, showing D75Y, with the WT residue and both ligands in gray, and the substituted residue in cyan

V78G was most common, appearing in 42 GISAID sequences and three USVs, in two cases occurring concurrently with amino acid changes for P80 (boundary) (P80A and P80L), suggesting that greater flexibility in this region of the protein may be tolerated. Four substitutions were identified that could introduce new hydrogen bonds spanning the heterodimer interface (P37S, G39S, M41T, and G77R), although each of these substitutions appears mildly destabilizing as judged by the results of $\Delta\Delta G^{App}$ calculations with Rosetta.

2.11 | Structural spike surface glycoprotein (S-protein)

The SARS-CoV-2 spike protein (S-protein) is a membrane-anchored homotrimeric class I fusion protein, that is 1273 residues in length and contains 22 N-linked glycosylation sites[46] per monomer (Figure 11A). The S-protein supports viral entry via host cell attachment and virion-host membrane fusion. Attachment to a host cell is mediated through the interaction of the S-protein receptor-binding domain (RBD, located in domain S1) with the angiotensin-converting enzyme 2 (ACE2) receptor (Figure 11B). Fusion of the virion to the host cell membrane occurs after cleavage of the S-protein between the S1 and S2 domains, with an additional cleavage (S2′) occurring near the fusion peptide (FP) domain, which is responsible for anchoring to the host cell membrane.

The first experimental structures of the S-protein deposited to the PDB include the pre-fusion state of the S-protein in two conformations—one with all three RBDs in a closed conformation (PDB ID 6VXX[47]) and one with RBD protruding upwards (PDB ID 6VSB[48]). A subsequently deposited PDB structure (PDB ID 6X2B[49])

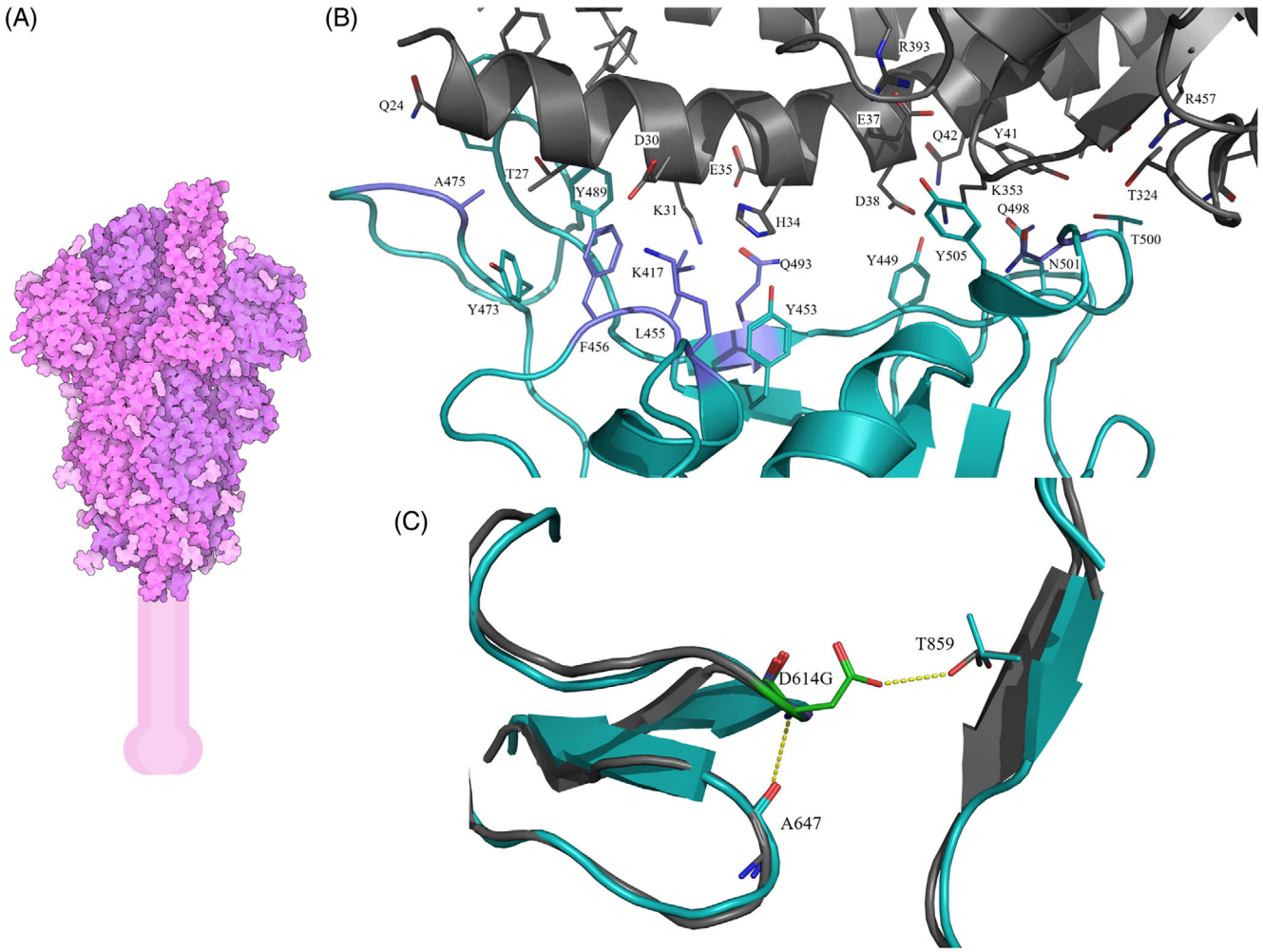

FIGURE 11 (A) Space-filling representation of the experimental structure of the S-protein homotrimer with one RBD protruding upwards (PDB ID 6VSB[48]); color coding: RBD up monomer-dark pink, RBD down monomers purple, N-linked carbohydrates-light pink). Membrane spanning portions are depicted in cartoon form. (B) Ribbon/atomic stick figure representation of the RBD interacting with ACE2 (PDB ID 6LZG[53]). RBD ribbon color: cyan or purple (substituted residues), atom color coding: C-cyan or purple, N-blue, O-red). ACE2 ribbon color: gray; atom color coding: C-gray, N-blue, O-red. (C) Ribbon/atomic stick figure representation of the D614 reference sequence structure (PDB ID 6VSB[48]; D614 ribbon color: cyan; atom color coding: C-cyan, N-blue, O-red) overlayed on the D614G substitution structure (PDB ID 6XS6[55]; D614G ribbon color-gray; atom color coding: C-gray, N-blue, O-red). H-bonds denoted with dotted yellow lines

revealed two upwards protruding RBDs; however, only a single RBD is necessary for ACE2 binding. It is not yet known if protrusion of the RBD from the S-protein trimer is necessary for binding to ACE2 or, as a recent meta-analysis of cryo-EM data suggests[50] that interconversion of the RBD between closed and open states represents an intrinsic property of the S-protein. Structures of the S-protein RBD were determined by X-ray crystallography early in the pandemic, both bound to full-length ACE2 receptor (PDB ID 6M17[51]) and bound to relevant ACE2 binding domains (PDB ID 6M0J[52]; PDB ID 6LZG[53]).

Overall substitution trends for the S-protein and energetics analysis results are summarized in Tables 1 and 2. The most commonly observed amino acid change from the reference sequence was D614G, a nonconservative substitution occurring in the SD2 boundary region of the S1 domain (Figure 11C). This substitution appears 21 014 times as a single point substitution and 3523 times in double or multipoint substitution contexts, accounting for ∼68% (805/1190) of all USVs and ∼74% (24 537/33290) of all sequenced genomes downloaded from GISAID. While this substitution is estimated to be slightly destabilizing versus the reference sequence (∼+0.6 REU), it seems to have emerged early in the pandemic and G614 is now the dominant form of the S-protein worldwide.[5] The question of if and why G614 is preferred versus D614 continues to be debated. It has been hypothesized that this substitution confers increased infectivity, possibly by reducing the pre-emptive shedding of the S1 domain and increasing the total amount of S-protein incorporated into virions.[54] A recent cryo-EM-based structural characterization of an engineered D614G S-protein revealed a significantly increased population of conformations in which RBDs are in the open state (PDB ID 6XS6[55]). Interestingly, the measured binding affinity of the G614 spike for ACE2 was slightly lower compared to the D614 variant. The increased population of open conformations in G614 was correlated with loss of inter-protomer contacts in the trimeric spike between D614 from the S1 domain and T859 from the S2 domain. This contact was postulated to be a "latch" that favors the closed state (Figure 11C).

Definitive elucidation of the effects of D614G and other substitutions on S-protein stability would require measuring impacts on the stability of all states (pre-fusion, post-fusion, open, closed). Moreover, amino acid changes may impact the structure and stability of complexes with binding partners (ACE2 and other possible co-receptors) and proteases responsible for S-protein cleavage. In this work, we limited our analysis of substitutions to two S-protein PDB structures available in June 2020: a pre-fusion all-closed RBD conformation (PDB ID 6VXX[47]), and the RBD-ACE2 complex (PDB ID 6M17[51]). Our methodology could be extended to other structures that continue to be determined at a fast clip, including antibody-bound or inhibitor-bound structures.

2.11.1 | Receptor binding domain substitutions

The most prevalent RBD substitution is the T478I (count = 57), which is in a portion of a loop that contacts ACE2, though residue 478 does not appear to be in direct contact itself. Interestingly, most substitutions directly interfacing with ACE2 were primarily neutral or destabilizing, with none improving binding affinity by more than −1 REU.

2.11.2 | Cleavage-site substitutions

It was recognized early in the pandemic that the S-protein possesses a potential furin cleavage site (residue 681-PRRAR/SV-residue687). Furin cleavage is thought to represent another mechanism for transition into a fusion-compatible state,[56] thereby contributing to virulence. However, the virus was still found to be infectious upon deletion of the furin cleavage site, indicating that it may not be required for viral entry[47] but may affect replication kinetics.[56] In that context, it is remarkable that several substitutions are observed within the putative furin cleavage site (P681L/S/H, R682Q/W, R683P/Q, A684T/S/V, S686G). Others have reported that amino acid changes occurred in the furin cleavage site.[57] Furin cleavage requires a polybasic motif, but the enzyme is not very stringent, suggesting that these altered sites may still be proteolytically cleaved.[58]

Prior to virus entry, the S-protein undergoes a second cleavage at the S2′ site (residue 811-KPSKR/SFI-residue 818), which exposes the fusion peptide. This component in the S2 domain fusion machinery attaches to the host cell membrane to initiate membrane fusion. The identity of the enzyme(s) responsible for the cleavage at this site is not known, although given the cleavage site sequence it is thought that it is a furin-like enzyme.[59,60] We identified several substitutions within the S2′ cleavage domain, including P812L/S/T, S813I/G, F817L, I818S/V. Further experimental study of these substitutions and the replication properties of these altered viruses may provide insight into the role played by furin cleavage in SARS-CoV-2 infection and virulence.

2.11.3 | Fusion machinery substitutions

Following cleavage at the S2′ site, the S-protein fuses the viral membrane with the host cell endosomal membrane. S2′ cleavage exposes the fusion peptide (loosely defined as residues 816–855), which then inserts into the host cell membrane. SARS-CoV-1 and SARS-CoV-2 fusion peptide sequences are very similar (∼93% sequence homology).[61] Our analyses, however, identified many USVs in which amino acid changes in this segment occurred during the pandemic (i.e., L821I, L822F, K825R, V826L, T827I, L828P, A829T, D830G/A, A831V/S/T, G832C/S, F833S, I834T). The active conformation and mode of insertion of the SARS-CoV-2 fusion peptide have not been experimentally characterized, making the impact of these substitutions impossible to assess. It may be significant that many of the observed amino acid changes in the fusion peptide are conservative.

A partial structure of the post-fusion state of the S-protein was determined early in the pandemic (PDB ID 6LXT[62]). During the final stages of membrane fusion, the HR1 and HR2 domains of class I fusion proteins assemble into a 6-helix bundle.[61] HR2 sequences of

SARS-CoV-1 and SARS-CoV-2 are identical. Differences in HR1 sequences between the two viruses suggest that SARS-CoV-2 HR2 makes stronger interactions with HR1.[62] Several substitutions occur on the solvent accessible surface of the HR1 domain (e.g., D936Y, S943P, S939F) and do not seem to participate in stabilizing interactions with HR2. It is, therefore, unclear how these nonconservative amino acid changes might affect the packing or stability of the post-fusion S-protein. Other residues in HR2 undergoing substitutions during the pandemic (e.g., K1073N, V1176F) or in the transmembrane or cytoplasmic tail domains (e.g., G1219C, P1263L) are not present in the post-fusion structure of the 6-helix bundle. Future experimental work to determine the conformation of the FP, HR1, HR2, and TM domains along the entire membrane fusion pathway should help to elucidate substitutions affecting these segments of the S-protein.

2.11.4 | N-terminal domain substitutions

The N-terminal domain (NTD) of the S-protein includes the first ~300 residues. Thus far, the function of the NTD has not been experimentally characterized. It is the target of neutralizing antibodies obtained from convalescent serum of individuals previously infected with SARS-CoV-2,[63] and the site of many substitutions identified in this work. Interestingly, the S-protein NTD of MERS-CoV utilizes sugar-binding receptors as a secondary means of interaction with host cells. Awasthi and co-workers have proposed that the SARS-CoV-2 S-protein NTD may do the same. Their computational modeling results suggest that that the NTD β4-β5 (69-HVSGTNGTKRF-79) and β14-β15 (243-ALHRSYLTPGDSSSGWTAGA-262) loop regions form a sialoside-binding pocket that would support engagement of host cell sialic acid moieties.[64] Our analyses documented that virtually all of the residues in these loops underwent amino acid changes during the pandemic (β4-β5: H69Y, V70F, S71F/Y, G72R/E/W, T73I, N74K, G75R/V/D, T76I, K77M/N, R78M/K, F79I; β14-β15: A243S/V, H245Y/R, R246I/S/K, S247R/N/I, Y248S, L249S/F, T250N, P251S/H/L, G252S, D253G/Y, S254F, S255F/P, S256P, G257S/R, W258L, A260S/V, G261V/S/D/R, A262S/T). Unfortunately, these loop regions are largely absent from the 3DEM structures used in our analysis (PDB ID 6VXX[47]; PDB ID 6VSB[48]), presumably because they are largely unstructured. Notwithstanding the paucity of 3D structural information, many of these substitutions would likely disrupt stabilizing electrostatic interactions between NTD and sialic acid derivatives postulated by Awasthi and coworkers.[64] Experimental work will be required to evaluate SARS-CoV-2 NTD interactions with sialic acid and how amino acid changes in the NTD affects binding to host cells.

2.12 | Structural nucleocapsid protein (N-protein)

The nucleocapsid N-protein (422 residues in length) forms a ribonucleoprotein (RNP) complex with viral RNA to protect and stabilize it within the viral envelope. The N-terminal domain (NTD) is responsible for nucleotide binding, while the C-terminal domain (CTD) is responsible for dimerization.[65] They are connected by a serine/arginine-rich (SR) linker region that is thought to be intrinsically disordered based on amino acid composition. Experimental structures for the N- and C-terminal domains of the SARS-CoV-2 N-protein (PDB ID 6VYO[66]; PDB ID 6YUN[67]) were used for the evolutionary analysis (Figure 12). Residues for which 3D structural information were not available include 1–48, 174–247, and 365–422.

Overall substitution trends for the N-protein and energetics analysis results are summarized in Tables 1 and 2. The most frequently observed USV (R203K/G204R) observed 11 425 times affects two residues within the SR linker region for which there is no 3D structural information. R203K (conservative, atomic coordinates are not present in either PDB structure) is the most common substitution, observed 13 130 times, and occurring in 272 USVs. The R203K/G204R double substitution also appears in most of the triple point substitutions (228/237 triples, 35/36 quadruples, 1/2 quintuples). Another interesting USV includes the 5-point substitution, R36Q/R203K/G204R/T135I/K373N). The NTD contains several basic residues (Arginine and Lysine) that are located in the finger subdomain and appear likely to interact with the RNA. Several substitutions in these finger-domain residues were observed in various USVs (e.g., R92S, R93L, R88L). If and how these may affect RNA-binding remains to be investigated.

FIGURE 12 Ribbon representation of the experimental structures of N-protein domains (PDB IDs 6VYO[66] and 6YUN[67]). [N. B. The relative orientations of the N-terminal (upper: residues 49–173) and C-terminal (lower: residues 248–364) domains was chosen arbitrarily. No structural information is currently available for residues 1–48, 174–247, and 365–422]

2.13 | Structural protein ion channel envelope protein (E-protein)

The integral membrane E-protein is the smallest of the SARS-CoV-2 structural proteins (75 residues). It plays important roles in virus-like particle production and maturation. Coronavirus E-proteins are co-translationally inserted into the endoplasmic reticulum (ER) and transported to Golgi complexes.[68] Although it is abundantly expressed within the cell, only a modest number of copies are incorporated into the viral envelope (estimated number/virion~20 for SARS-CoV-1,[69]). Instead, most of the protein participates in virion assembly and budding together with the SARS-CoV-2 integral membrane M-protein (also a virion structural protein). Additional functions of the E-protein are thought to include preventing M-protein aggregation and inducing membrane curvature.[70] Recombinant coronaviruses lacking E-proteins display weakened maturation, reduced viral titers, or yield incompetent progeny, highlighting its role in maintaining virion integrity.

The E-protein consists of a shorter hydrophilic N-terminal segment, a longer hydrophobic transmembrane domain (TMD), and a hydrophilic C-terminal domain. An amphipathic α-helix within the TMD oligomerizes into an homopentameric arrangement perpendicular to the plane of the lipid bilayer forming an ion-conducting viroporin.[70] Residues lining the pore include N15, L19, A22, F26, T30, I33, and L37. The NMR structure of the SARS-CoV-1 E-protein (PDB ID 5X29[71]) served as the template for generating the computed structural model of the SARS-CoV-2 E-protein that was used for analyzing its evolution in 3D (Figure 13). The N-terminal seven residues and the C-terminal ten residues were omitted from the homology model, because they were not reported in the SARS-CoV-1 NMR structure.

Overall substitution trends for the E-protein and energetics analysis results are summarized in Tables 1 and 2. S68F (nonconservative, structural location unknown) is the most common USV, observed 107 times in the GISAID dataset. The most intriguing changes in the protein are L37R and L37H USVs, located near the entrance to the pore (Figure 13). The changes of Leucine to Arginine or Histidine are notable because the canonical transmembrane domain lacks charged residues. The SARS-CoV-1 E-protein is preferentially selective for cations, although it can transport anions.[72] Substitution of L37 to a positively charged residue may affect ion passage selectivity and/or its ability to transport ions. L30H was recorded twice in GISAID, confirming that these variants are viable. The quantification of the consequences of substitutions to L30 on viral viability may be a subject for experimental investigation as this may affect the transport of ions facilitated by the E-protein.

SARS-CoV-1 E-protein is N-linked glycosylated at N66.[73] At the time of writing, there were no published reports pertaining to SARS-CoV-2 E-protein glycosylation. The corresponding residue in SARS-CoV-2 E-protein is N66, which underwent substitution to Histidine in a single USV (N66H) that would abrogate glycosylation. Observed

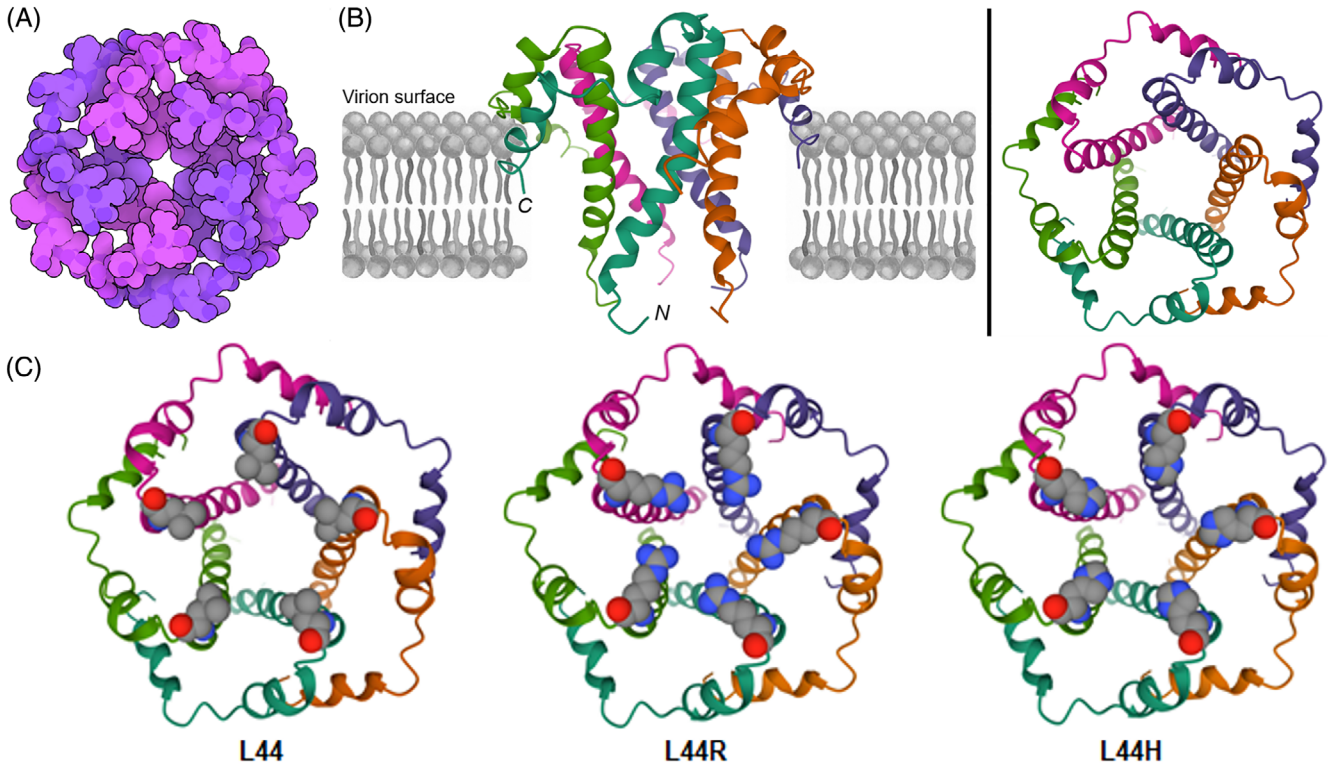

FIGURE 13 (A) Space-filling representation of the computed structural model of the E-protein with individual protomers shown with shades of pink and purple. (B) Ribbon representation with each protomer shown using a different color viewed parallel to the membrane (left, membrane shown, N- and C-termini labeled) and down the five-fold axis from the virion surface (right). (C) Pore-lining substitutions L37R and L37H compared to L37 in the reference sequence (residue 37 is shown in a color-coded space-filling representation; C-gray; O-red; N-blue)

amino acid substitutions involving loss or gain of other potential sites of N-linked and O-linked glycosylation include A41S, C43S, N48S, S50G, P54S, S55F, S68C, S68F, and S68Y.

2.14 | Structural integral membrane protein (M-protein)

The integral membrane M-protein (222 residues in length) is the most abundant structural protein in the SARS-CoV-1 virion.[74] It is co-translationally inserted into the ER and transported to Golgi complexes,[73] where it is responsible for directing virus assembly and budding via interactions with E-, N-, and S-proteins. The SARS-CoV-2 M-protein is predicted to consist of a small glycosylated amino-terminal ectodomain, a triple-membrane spanning domain, and a carboxyl-terminal endodomain that extends 6–8 nm into the viral particle. The C-terminal portion of coronaviral M-proteins bind to the N-protein within the cell membrane of the ER or Golgi complex, stabilizing the nucleocapsid and the core of the virion. M-proteins also interacts with the E-protein to trigger budding, and with the S-protein for incorporation into virions.[75] Following assembly, virions are transported to the cell surface and released via exocytosis. The M-protein is believed to exist as a dimer in the cell membrane and may adopt two conformations that allow it to bend the membrane and interact with N-protein/RNA RNP.[76] Sequence alignment of SARS-CoV-2 M-protein to its SARS-CoV-1 homolog revealed high sequence identity (~90%). The M-protein structural model used for analyzing evolution in 3D was computed by the David Baker Laboratory during a CASP competition (CASP-C1906 Stage 2, Figure 14).

Overall substitution trends for the M-protein and energetics analysis results are summarized in Tables 1 and 2. T175M (nonconservative, surface) is the most common USV, observed 746 times in the GISAID data set (~39% of the observed variant M-proteins). An N5S substitution affects the sole N-linked glycosylation site in the small ectodomain. Given that M-protein glycosylation is not essential for maintaining virion morphology or growth kinetics,[77] it is unclear if M-protein function is affected by the N5S substitution.

3 | IMPLICATIONS FOR THE ONGOING PANDEMIC AND DISCOVERY AND DEVELOPMENT OF EFFECTIVE COUNTERMEASURES

Our analyses of SARS-CoV-2 genome sequences archived by GISAID documented that every one of the 29 study proteins underwent amino acid changes versus the original reference sequence during the first 6 months of the pandemic. Most of these substitutions occurred infrequently. Approximately two thirds of the substitutions were non-conservative, and most appear to have arisen from single or double nucleotide changes in the RNA genome. Computational 3D structure modeling of the USVs demonstrated that substitutions primarily occurred in the boundary layers and the surfaces of the viral proteins. Most of the amino acid changes appear to be moderately destabilizing, as judged by the results of energetics ($\Delta\Delta G^{App}$) calculations. Given that most of the viral genomes archived by GISAID were obtained from samples provided by infected individuals, we believe that the viruses and hence the viral proteins were functional and capable of causing disease in humans. Where multiple substitutions were detected in a USV, we believe that most were the product of cumulative changes. At least one of the observed amino acid changes in multisubstitution USVs was almost always detected as a single

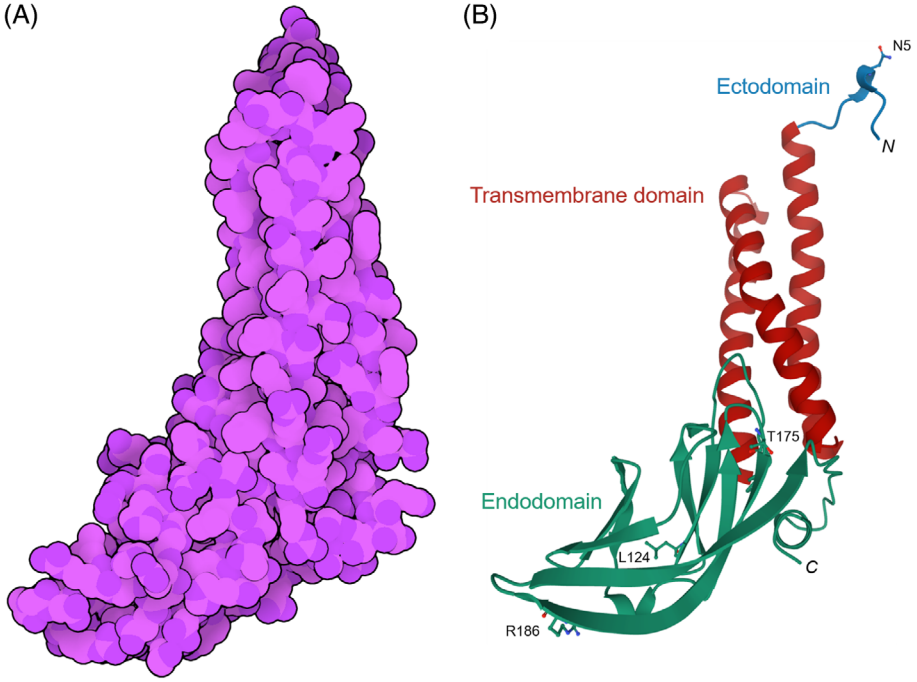

FIGURE 14 (A) Space-filling representation of the computed structural model of the M-protein protomer. The glycosylated N-terminus is located at the apex of the structure. (B) Ribbon/atomic stick figure representation (color coding: ectodomain-blue, transmembrane α-helices-red, endodomain-green). N- and C-termini are labeled, together with residues N5, L124, T175, and R186 (shown in ball and stick representation; atom color coding: C-green, O-red, N-blue)

substitution in another USV derived from a sample collected earlier in the pandemic. There is every reason to believe that the pool of viruses circulating in humans and other mammals (e.g., *Mustela lutreola* or European mink) around the world today will continue to diverge from the reference sequence. We have made 3D structure models of 7462 USVs and our analysis results freely available under the most permissive Creative Commons CC0 1.0 Universal license to facilitate the work of research groups using experimental and computational tools to characterize SARS-CoV-2 protein function and study the structural and functional consequences of the myriad substitutions observed during the first half of 2020.

Some, almost certainly not all, of the 29 viral proteins analyzed herein represent promising targets for discovery and development of small-molecule anti-viral agents. At the time of writing, one small-molecule drug (remdesivir targeting the RdRp) has received full approval from the US FDA. This compound was originally discovered during the search for an Ebola virus therapeutic. Although it failed to demonstrate efficacy in clinical trials for Ebola victims, the safety profile encouraged the sponsor company (Gilead Sciences Inc.) to successfully repurposed the drug for treatment of SARS-CoV-2 infected individuals. Open access to PDB structures of remdesivir bound to the RdRp sets the stage for structure-guided discovery of second generation nucleoside analogs with superior potency and/or selectivity, more desirable drug-like properties, or better Absorption-Distribution-Metabolism-Excretion profiles (e.g., improved oral bioavailability to avoid intravenous administration).[78] Open access to our computed 3D structural models of 840 RdRp USVs will provide useful information that may enable drug hunting teams to anticipate potential sources of drug resistance during selection for candidates slated for in vitro pre-clinical development studies.

Open access to PDB structures of other essential SARS-CoV-2 enzymes (and those of their closely related SARS-CoV-1 homologs) have already facilitated initiation of structure-guided drug discovery campaigns for PLPro, nsp5, nsp13, nsp14, and nsp10/nsp16. As for RdRp, free availability of computed 3D structural models of nearly 1500 USVs may provide useful information pertaining to potential causes of drug resistance. Knowledge of sequence (and 3D structure) variation during the pandemic could also be used to prioritize these potential drug targets using quantitative assessments of active site conservation. The best drug discovery targets could be those proteins observed to undergo the fewest amino acid changes in their active (or drug-binding) site during the first 6 months of the pandemic. It is also possible that inhibitors making contacts with residues that are not engaged by substrates will be more susceptible to the emergence of drug resistance.

The S-protein is the target of both monoclonal antibodies (for passive immunization) and vaccines. At the time of writing, several monoclonal antibodies had already received Emergency Use Authorization (EUA) from the US FDA (e.g., bamlanivimab; sponsor company Eli Lilly and Co.). The Pfizer/BioNTech mRNA vaccine had received full FDA approval and the Moderna mRNA vaccine was granted under EUA. Open access to a host of PDB structures of the S-protein in various conformational states and in complexes with host cell proteins and Fab fragments of monoclonal antibodies will facilitate the work of research teams focused on discovery and development of second-generation monoclonal antibodies and vaccines. Free availability of 689 3D structural models of S-protein USVs may provide insights into potential efficacy failures due to amino acid changes in the S-protein that interfere with viral antigen recognition by antibodies (monoclonal or humoral) or T-cells while preserving ACE2 receptor binding.

4 | MATERIALS AND METHODS

4.1 | Project history

This work was initiated by research interns (undergraduates and one high school student) hosted virtually during the summer of 2020 by the Rutgers University Institute for Quantitative Biomedicine (IQB), the Rutgers University RISE Program, and the US-funded RCSB Protein Data Bank headquartered at Rutgers.[79–81] Prior to the online five-week research program, participating students and mentors received 1 week of online training in 3D molecular visualization and computational bioinformatics in the IQB "Summer of the Coronaverse" Online Boot Camp.[82] The methods used in the research study were developed, evaluated, and refined during the online Boot Camp. Supervision of the research phase was provided by IQB graduate students, postdoctoral fellows, and RCSB Protein Data Bank scientific staff, all of whom served as mentors in the Boot Camp. The research interns worked collaboratively in teams, carrying out multiple sequence alignments, constructing phylogenetic trees, computing 3D structural models of viral proteins, visualizing 3D structures, and analyzing the structural, functional, and energetic consequences of SARS-CoV-2 protein amino acid substitutions identified during the first 6 months of the pandemic. All computed 3D structural models and results of the sequence/energetics analyses are described in the main body of this paper and accompanying Supplementary Materials. The computed 3D structural models and energetics results are made freely available under Creative Commons license CC0 1.0 Universal for researchers wishing to perform further computational and experimental studies (see https://doi.org/10.5281/zenodo.5521766).

4.2 | SARS-CoV-2 genome

The SARS-CoV-2 genome resembles a single-stranded cellular messenger RNA, ~29.9 kb in length with a 7-methyl-G 5′ cap, a 3′ poly-A tail, and more than 10 open reading frames or Orfs (Figure 1). Viral proteins are expressed in two ways. Translation of two long polyproteins occurs initially, yielding the machinery required to copy the viral genome. Subsequent expression of multiple sub-genomic mRNAs produces the four structural proteins present in virions and other proteins designated as Orf3a, Orf6, Orf7a, Orf7b, Orf8, Orf9b, Orf14, and possibly the hypothetical protein Orf10. The nonstructural proteins (nsps) are expressed within the shorter polyprotein 1a (pp1a, encompassing nsp1-nsp11) and the longer polyprotein 1ab (pp1ab,

encompassing nsp1-nsp16). Both pp1a and pp1ab require two virally-encoded proteases for processing into individual nsp protomers (Figure 2). nsp3 includes a papain-like protease (PLPro) domain, which is responsible for polypeptide chain cleavage at three sites within the N-terminal portions of both polyproteins (dark blue inverted triangles in Figure 1). Ten additional polypeptide chain cleavages are carried out by nsp5 (light blue inverted triangles in Figure 2), also known as the main protease or the 3C-like protease. The structural proteins present in mature virions include the S-protein (surface spike glycoprotein, responsible for viral entry), the N-protein (nucleocapsid protein), the E-protein (a pentameric ion channel), and the M-protein (a second integral membrane protein found in the viral lipid bilayer).

4.3 | SARS-CoV-2 study protein sequences

Pre-aligned protein sequences were downloaded in FASTA format from the GISAID website (gisaid.org)[83,84] on June 25th, 2020. Sequence alignments for each of the SARS-CoV-2 proteins (hereafter study proteins) were constructed by removing non-human sequences from the alignment; removing truncated sequences; removing incompletely determined sequences (i.e., those with one or more "X" in lieu of an amino acid one-letter code); and eliminating duplicates. Study protein sequences made public by researchers in the People's Republic of China on January 10th 2020 (GenBank accession code MN908947.3)[85] were defined as the "reference sequence" for each individual study protein and all unique sequence variant (USV) or amino acid substituted forms of individual study proteins were compared with their respective reference sequence. We have assumed that none of observed USVs yielded study proteins that either failed to fold or lost necessary biochemical functionality for other reasons, because it is likely given the timing of specimen collection that all the viral RNAs were isolated from infected individuals and are, therefore, presumed to have been infectious. For sequence identity calculations, GenBank accession code AY278741.1 was used as the source of SARS-CoV-1 protein reference sequences.

4.4 | Experimentally-determined structures of study proteins from the PDB archive

Atomic coordinates for the experimental structures of 19 study proteins were downloaded from the PDB archive via the RCSB PDB website (RCSB.org), including nsp1 (PDB ID 7K3N[86]), nsp3a (PDB ID 7KAG[87]), nsp3b (PDB ID 6WEY[88]), Papain-like Proteinase (PLPro; nsp3d; PDB ID 6WUU[28]), nsp3e (PDB ID 7LGO[89]), nsp5 (PDB ID 6YB7[90]), nsp7 (part of the RDRP; PDB ID 6YYT[21]), nsp8 (part of the RDRP; PDB ID 6YYT[21]), nsp9 (PDB ID 6WXD[91]), nsp10 (part of the methyltransferase; PDB ID 6WVN[42]), nsp12 (part of the RDRP; PDB ID 6YYT[21]), nsp13 (PDB ID 6JYT[36]), nsp15 (PDB ID 6WXC[92]), nsp16 (part of the methyltransferase; PDB ID 6WVN[42]), S-protein (PDB ID 6VXX[47]; PDB ID 6M17[51]), Orf3a (PDB ID 6XDC[93]), Orf7a (PDB ID 7CI3[94]), Orf8 (PDB ID 7JX6[95]), and the N-protein (PDB ID 6VYO[66]; PDB ID 6YUN[67]).

4.5 | Computed structural models of study proteins

Swiss-Model[96] was the source of the computed structural model nsp3c (part of nsp3) using 75% sequence identical template SARS-CoV-1 nsp3c (part of nsp3) (PDB ID 2W2G[97]).

The computed structural model for nsp14 (https://robetta.bakerlab.org/results.php?id=15671) was downloaded from the Robetta-based predictions from the website for Seattle Structural Genomics Center for Infectious Disease (https://www.ssgcid.org/cttdb/molecularmodel_list/?organism__icontains=COVID-19).

The computed structural model for the SARS-CoV-2 E-protein were generated using the solution state NMR structure of the SARS-CoV-1 E-protein embedded in lyso-myristoyl phosphatidylglycerol micelles (PDB ID 5X29, model 1[71]) as a template, and substituting differing residues using the MUTATE feature of VMD.[98] The structural model was then subjected to 10 000 steps of energy minimization in vacuum using NAMD 2.13[99] and the CHARMM 36 force field.[100]

Computed structural models for the seven remaining study proteins were obtained from the Rosetta-based Baker group predictions (TS131) CASP website (https://predictioncenter.org/caspcommons/targetlist.cgi; Model 1 was chosen), including nsp2, UNK (part of nsp3), nsp4, nsp6, the M-protein, Orf6, and Orf7b.

nsp11, Orf9b, Orf14, and hypothetical protein Orf10 were excluded from consideration owing to lack of sequence and/or 3D structure data.

4.6 | Molecular visualization and graphics

The RCSB Protein Data Bank web-native molecular graphics tool (Mol*[101]) was used for visual inspection and comparison of reference and amino-acid-substitute study proteins. Space-filling representation figures were generated using Illustrate.[102] Ribbon/atomic stick figure representation figures were generated using Mol* and PyMOL.[103]

4.7 | Rosetta-based analyses of substitution location(s), conservation, and energetics

PyRosetta[104] was used to analyze each study protein and its observed USVs. All residue pairs with C_α-C_α distance <5.5 Å were considered neighbors, and residue pairs with C_α-C_α distance <11 Å were also considered neighbors if their C_α-C_β vectors were at an angle <75°. Residue layer identifications were performed on reference (rather than substituted) study protein structures, based on side chain neighbors within a cone centered on the C_α-C_β vector, which is independent of side chain conformation. The Layer Determination Factor (LDF) is defined as $\mathrm{LDF} = ((\cos[\theta] + 0.5)/1.5)^2/(1 + \exp[d - 9])$, where θ is the

angle between the C_α–C_β vector of a given residue and that of a neighbor, and *d* is the C_α–C_α distance between residue and neighbor. LDF is summed over nearby neighbors and if its value is <2, the residue is considered surface. If it is >5.2, the residue is considered core. Otherwise, it is considered boundary.

Amino acid substitution conservation was determined by whether a residue change stayed within a residue type group as follows: hydrophobic (A, F, I, L, M, V, W, Y), negatively charged (D, E), positively charged (H, K, R), and uncharged hydrophilic (N, Q, S, T) and any substitution to a residue outside the native residue's group was considered nonconservative. Changes to or from Glycine, Proline, or Cysteine were considered nonconservative. Amino acid substitutions in study proteins were identified by alignment with the reference sequence.

Experimental structures and computed structural models of study proteins were prepared for computational analyses using the Rosetta FastRelax protocol, employing atom positional restraints to limit significant changes to backbone geometry. Homo-oligomeric proteins were modeled using the symmetric protein modeling framework in Rosetta.[105] Integral membrane proteins were modeled using Rosetta membrane protein modeling framework.[106]

Structural models for study protein USVs were computed by replacing the reference side chain atomic coordinates in the starting model with those of the substituted amino acid(s) and performing three rounds of Monte Carlo optimization of rotamers for all side chains falling within an 8 Å radius of the substitution(s), followed by gradient-based energy minimization of the entire structure, with atom positional restraints to limit significant changes to backbone geometry. Computed structural model optimizations were performed with three different combinations of scoring functions based on previous work,[107] including "hard-hard," indicating that both side chain optimization and structure minimization were performed with default van der Waals repulsion term in the Rosetta scorefunction, "soft-soft" indicating that for both steps, a different scorefunction was used that has dampened van der Waals repulsion (in this case, the backbone was entirely prevented from moving during minimization), and "soft-hard" indicating that the soft-repulsive score function was used for side chain rotamer optimization, while the hard-repulsive scorefunction was used for energy minimization. The scorefunctions used were REF2015[108] and REF2015_soft[107] for soluble proteins, and franklin2019[109] for integral membrane proteins (with a dampened van der Waals repulsion weight in the case of soft repulsion).

Energetic consequences of amino acid substitutions were determined by performing identical side chain optimization and energy minimization on both wild type and substituted models thrice and subtracting the total energy of the lowest-scoring wild-type model from that of the lowest-scoring substituted model (dividing by the number of symmetric chains where applicable). The "soft-hard" protocol emerged as the preferred method because it generated the lowest number of outliers. Only USVs in which a unique set of substitutions occurred at residue positions that were present in the available study protein structures were included in the energy analyses (7462 USVs).

ACKNOWLEDGMENTS

We thank all the many structural biologists who have deposited coronavirus protein structures to the PDB archive since 2002. We also thank Drs. David Baker and Ivan Anischenko for providing computed structural models generated by the David Baker Laboratory, Ms. Virginia Jiang, and Dr. Scott Banta for help with Rosetta calculations involving integral membrane proteins, and Dr. Andrew Brooks for advice regarding sequencing artifacts. We gratefully acknowledge contributions from all members of the Research Collaboratory for Structural Bioinformatics PDB and our Worldwide Protein Data Bank partners. RCSB PDB is jointly funded by the National Science Foundation (DBI-1832184), the US Department of Energy (DE-SC0019749), and the National Cancer Institute, National Institute of Allergy and Infectious Diseases, and National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH) under grant R01 GM133198. The Khare laboratory has been funded by NIH NIGMS (R01 GM132565) and NSF (CBET1923691) and a Rutgers University Center for COVID-19 Response and Pandemic Preparedness award. The Baum laboratory is funded by NIH Grant GM136431 and a Rutgers University Center for COVID-19 Response and Pandemic Preparedness award. J.H.L. was funded by NIH NIGMS T32 Training Grants GM008339 and GM135141. E.A. acknowledges support by a National Institutes of Health MERIT Award (R0137 AI027690) and a Rutgers Center for COVID-19 Response and Pandemic Preparedness Award. J.S. and G.B. are funded by the Busch Biomedical Foundation. The BASIL Consortium is funded by NSF IUSE grants 1709170, 1709355, 1709805, and 1709278. We gratefully acknowledge support for L.H.A.A., A.K., E.M., S.S., B.T., A.T., L.W., and M.O.-A. by the Rutgers University RISE (Research Intensive Summer Experience) Program, an NSF REU for A.T., S.S., L.W., and a New Jersey Space Grant Consortium Award for L.H.A.A. and B.T.

CONFLICT OF INTEREST

The authors declare no competing financial interests.

AUTHOR CONTRIBUTIONS

Sagar D. Khare and Stephen K. Burley developed the research plan. All authors helped to assemble the data, develop and execute analysis strategies, prepare figures and tables, and write the manuscript.

DATA AVAILABILITY STATEMENT

The computed 3D structural models and energetics results are made freely available under Creative Commons license CC0 for researchers wishing to perform further computational and experimental studies (see https://iqb.rutgers.edu/covid-19_proteome_evolution and https://doi.org/10.5281/zenodo.5521766).

ORCID

Christine Zardecki https://orcid.org/0000-0002-4149-1745
Jonathan K. Williams https://orcid.org/0000-0002-7272-6885
Sagar D. Khare https://orcid.org/0000-0002-2255-0543

REFERENCES

1. Chen Y, Liu Q, Guo D. Emerging coronaviruses: genome structure, replication, and pathogenesis. *J Med Virol*. 2020;92(4):418-423.
2. Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. Coronaviruses. *RNA Biol*. 2011;8(2):270-279.
3. Wang R, Hozumi Y, Yin C, Wei GW. Decoding SARS-CoV-2 transmission and evolution and ramifications for COVID-19 diagnosis, vaccine, and medicine. *J Chem Inf Model*. 2020;60:5853-5865.
4. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 2018;34(23):4121-4123.
5. Korber B, Fischer WM, Gnanakaran S, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell*. 2020;182(4):812-827. e819.
6. Pollack A. Company says it mapped part of SARS virus. *The New York Times*. 2003;2003:C2.
7. Burley SK. How to help the free market fight coronavirus. *Nature*. 2020;580(7802):167.
8. Jin Z, Du X, Xu Y, et al. Structure of M(pro) from SARS-CoV-2 and discovery of its inhibitors. *Nature*. 2020;582(7811):289-293.
9. Protein Data Bank. Crystallography: protein data bank. *Nature (London), New Biol*. 1971;233(42):223-223.
10. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235-242.
11. wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res*. 2019;47(D1): D520-D528.
12. Simmonds P, Bukh J, Combet C, et al. Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. *Hepatology*. 2005;42(4):962-973.
13. Privalov PL, Gill SJ. Stability of protein structure and hydrophobic interaction. *Adv Protein Chem*. 1988;39:191-234.
14. Faure G, Koonin EV. Universal distribution of mutational effects on protein stability, uncoupling of protein robustness from sequence evolution and distinct evolutionary modes of prokaryotic and eukaryotic proteins. *Phys Biol*. 2015;12(3):035001.
15. Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS. The stability effects of protein mutations appear to be universally distributed. *J Mol Biol*. 2007;369(5):1318-1332.
16. Razban RM, Shakhnovich EI. Effects of single mutations on protein stability are Gaussian distributed. *Biophys J*. 2020;118(12):2872-2878.
17. Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. Thermodynamic prediction of protein neutrality. *Proc Natl Acad Sci U S A*. 2005;102(3):606-611.
18. Nisthal A, Wang CY, Ary ML, Mayo SL. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc Natl Acad Sci U S A*. 2019;116(33):16367-16377.
19. Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The architecture of SARS-CoV-2 transcriptome. *Cell*. 2020;181(4):914-921. e910.
20. Gao Y, Yan L, Huang Y, et al. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science*. 2020;368(6492):779-782.
21. Hillen HS, Kokic G, Farnung L, Dienemann C, Tegunov D, Cramer P. Structure of replicating SARS-CoV-2 polymerase. *Nature*. 2020; 584(7819):154-156.
22. Yin W, Mao C, Luan X, et al. Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science*. 2020;368(6498):1499-1504.
23. Wang Q, Wu J, Wang H, et al. Structural basis for RNA replication by the SARS-CoV-2 polymerase. *Cell*. 2020;182(2):417-428. e413.
24. Gordon DE, Jang GM, Bouhaddou M, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*. 2020; 583(7816):459-468.
25. Clementz MA, Chen Z, Banach BS, et al. Deubiquitinating and interferon antagonism activities of coronavirus papain-like proteases. *J Virol*. 2010;84(9):4619-4629.
26. Osipiuk J, Azizi S-A, Dvorkin S, et al. Structure of papain-like protease from SARS-CoV-2 and its complexes with non-covalent inhibitors. *Nat Commun*. 2021;12:(1). http://dx.doi.org/10.1038/s41467-021-21060-3
27. Ratia K, Saikatendu KS, Santarsiero BD, et al. Severe acute respiratory syndrome coronavirus papain-like protease: structure of a viral deubiquitinating enzyme. *Proc Natl Acad Sci U S A*. 2006;103(15): 5717-5722.
28. Rut W, Lv Z, Zmudzinski M, et al. Activity profiling and crystal structures of inhibitor-bound SARS-CoV-2 papain-like protease: a framework for anti-COVID-19 drug design. *Sci Adv*. 2020;6(42): eabd4596.
29. Shin D, Mukherjee R, Grewe D, et al. Papain-like protease regulates SARS-CoV-2 viral spread and innate immunity. *Nature*. 2020;587: 657-662.
30. Osipiuk J, Tesar C, Endres M, et al. The crystal structure of papain-like protease of SARS CoV-2 in complex with PLP_Snyder441 inhibitor. 2020. https://doi.org/10.2210/pdb7JN2/pdb.
31. Huang C, Wei P, Fan K, Liu Y, Lai L. 3C-like proteinase from SARS coronavirus catalyzes substrate hydrolysis by a general base mechanism. *Biochemistry*. 2004;43(15):4568-4574.
32. Chodera J, Lee AA, London N, von Delft F. Crowdsourcing drug discovery for pandemics. *Nat Chem*. 2020;12(7):581.
33. Miao Z, Tidu A, Eriani G, Martin F. Secondary structure of the SARS-CoV-2 5′-UTR. *RNA Biol*. 2020;18:1-10.
34. Adedeji AO, Marchand B, Te Velthuis AJ, et al. Mechanism of nucleic acid unwinding by SARS-CoV helicase. *PLoS One*. 2012;7(5):e36521.
35. Chen J, Malone B, Llewellyn E, et al. Structural basis for helicase-polymerase coupling in the SARS-CoV-2 replication-transcription complex. *Cell*. 2020;182(6):1560-1573. e1513.
36. Jia Z, Yan L, Ren Z, et al. Delicate structural coordination of the severe acute respiratory syndrome coronavirus Nsp13 upon ATP hydrolysis. *Nucleic Acids Res*. 2019;47(12):6538-6550.
37. Khailany RA, Safdar M, Ozaslan M. Genomic characterization of a novel SARS-CoV-2. *Gene Rep*. 2020;19:100682.
38. Shannon A, Le NT, Selisko B, et al. Remdesivir and SARS-CoV-2: structural requirements at both nsp12 RdRp and nsp14 exonuclease active-sites. *Antiviral Res*. 2020;178:104793.
39. Eckerle LD, Becker MM, Halpin RA, et al. Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. *PLoS Pathog*. 2010;6(5):e1000896.
40. Ma Y, Wu L, Shaw N, et al. Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. *Proc Natl Acad Sci U S A*. 2015;112(30):9436-9441.
41. Chen Y, Su C, Ke M, et al. Biochemical and structural insights into the mechanisms of SARS coronavirus RNA ribose 2′-O-methylation by nsp16/nsp10 protein complex. *PLoS Pathog*. 2011;7(10): e1002294.
42. Rosas-Lemus M, Minasov G, Shuvalova L, et al. High-resolution structures of the SARS-CoV-2 2′-O-methyltransferase reveal strategies for structure-based inhibitor design. *Sci Signal*. 2020;13(651): eabe1202.
43. Decroly E, Debarnot C, Ferron F, et al. Crystal structure and functional analysis of the SARS-coronavirus RNA cap 2′-O-methyltransferase nsp10/nsp16 complex. *PLoS Pathog*. 2011;7(5):e1002059.
44. Rosas-Lemus M, Minasov G, Shuvalova L, et al. The crystal structure of nsp10-nsp16 heterodimer from SARS-CoV-2 in complex with S-adenosylmethionine. *bioRxiv*. 2020; 2020.2004.2017.047498. https://doi.org/10.1101/2020.04.17.047498
45. Rogstam A, Nyblom M, Christensen S, et al. Crystal structure of non-structural protein 10 from severe acute respiratory syndrome coronavirus-2. *Int J Mol Sci*. 2020;21(19):7375.
46. Watanabe Y, Allen JD, Wrapp D, McLellan JS, Crispin M. Site-specific glycan analysis of the SARS-CoV-2 spike. *Science*. 2020; 369(6501):330-333.

47. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*. 2020;181(2):281-292. e286.
48. 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.
49. Henderson R, Edwards RJ, Mansouri K, et al. Controlling the SARS-CoV-2 spike glycoprotein conformation. *Nat Struct Mol Biol*. 2020; 27(10):925-933.
50. Melero R, Sorzano COS, Foster B, et al. Continuous flexibility analysis of SARS-CoV-2 spike prefusion structures. *IUCrJ*. 2020;7(Pt 6): 1059–1069.
51. 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.
52. 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.
53. Wang Q, Zhang Y, Wu L, et al. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell*. 2020;181(4):894-904. e899.
54. 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(1):6013.
55. 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. e738.
56. Johnson BA, Xie X, Kalveram B, et al. Furin cleavage site is key to SARS-CoV-2 pathogenesis. *bioRxiv*. 2020. https://doi.org/10.1101/2020.1108.1126.268854
57. Xing Y, Li X, Gao X, Dong Q. Natural polymorphisms are present in the Furin cleavage site of the SARS-CoV-2 spike glycoprotein. *Front Genet*. 2020;11:783.
58. Shiryaev SA, Chernov AV, Golubkov VS, et al. High-resolution analysis and functional mapping of cleavage sites and substrate proteins of furin in the human proteome. *PLoS One*. 2013;8(1):e54290.
59. Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, Decroly E. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res*. 2020;176:104742.
60. Hoffmann M, Kleine-Weber H, Schroeder S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell*. 2020;181(2):271-280. e278.
61. Tang T, Bidon M, Jaimes JA, Whittaker GR, Daniel S. Coronavirus membrane fusion mechanism offers a potential target for antiviral development. *Antiviral Res*. 2020;178:104792.
62. Xia S, Liu M, Wang C, 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*. 2020;30(4):343-355.
63. Chi X, Yan R, Zhang J, et al. A neutralizing human antibody binds to the N-terminal domain of the spike protein of SARS-CoV-2. *Science*. 2020;369(6504):650-655.
64. Awasthi M, Gulati S, Sarkar DP, et al. The Sialoside-binding pocket of SARS-CoV-2 spike glycoprotein structurally resembles MERS-CoV. *Viruses*. 2020;12(9):909.
65. Kang S, Yang M, Hong Z, et al. Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *Acta Pharm Sin B*. 2020;10(7):1228-1238.
66. Chang C, Michalska K, Jedrzejczak R, et al. Crystal structure of RNA binding domain of nucleocapsid phosphoprotein from SARS coronavirus 2. 2020. https://doi.org/10.2210/pdb6VYO/pdb.
67. Zinzula L, Basquin J, Bohn S, et al. High-resolution structure and biophysical characterization of the nucleocapsid phosphoprotein dimerization domain from the Covid-19 severe acute respiratory syndrome coronavirus 2. *Biochem Biophys Res Commun*. 2021;538:54-62.
68. Ruch TR, Machamer CE. The coronavirus E protein: assembly and beyond. *Viruses*. 2012;4(3):363-382.
69. DeDiego ML, Alvarez E, Almazan F, et al. A severe acute respiratory syndrome coronavirus that lacks the E gene is attenuated in vitro and in vivo. *J Virol*. 2007;81(4):1701-1713.
70. Schoeman D, Fielding BC. Coronavirus envelope protein: current knowledge. *Virol J*. 2019;16(1):69.
71. Surya W, Li Y, Torres J. Structural model of the SARS coronavirus E channel in LMPG micelles. *Biochim Biophys Acta Biomembr*. 2018; 1860(6):1309-1317.
72. Verdia-Baguena C, Nieto-Torres JL, Alcaraz A, et al. Coronavirus E protein forms ion channels with functionally and structurally-involved membrane lipids. *Virology*. 2012;432(2):485-494.
73. Chen SC, Lo SY, Ma HC, Li HC. Expression and membrane integration of SARS-CoV E protein and its interaction with M protein. *Virus Genes*. 2009;38(3):365-371.
74. Siu YL, Teoh KT, Lo J, et al. The M, E, and N structural proteins of the severe acute respiratory syndrome coronavirus are required for efficient assembly, trafficking, and release of virus-like particles. *J Virol*. 2008;82(22):11318-11330.
75. Tseng YT, Chang CH, Wang SM, Huang KJ, Wang CT. Identifying SARS-CoV membrane protein amino acid residues linked to virus-like particle assembly. *PLoS One*. 2013;8(5):e64013.
76. Neuman BW, Kiss G, Kunding AH, et al. A structural analysis of M protein in coronavirus assembly and morphology. *J Struct Biol*. 2011; 174(1):11-22.
77. Voss D, Pfefferle S, Drosten C, et al. Studies on membrane topology, N-glycosylation and functionality of SARS-CoV membrane protein. *Virol J*. 2009;6:79.
78. Westbrook JD, Soskind R, Hudson BP, Burley SK. Impact of protein data Bank on anti-neoplastic approvals. *Drug Discov Today*. 2020;25: 837-850.
79. Burley SK, Bhikadiya C, Bi C, et al. RCSB protein data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering, and energy sciences. *Nucleic Acid Res*. 2021;49:D437-D451.
80. Goodsell DS, Zardecki C, Di Costanzo L, et al. RCSB protein data bank: enabling biomedical research and drug discovery. *Protein Sci*. 2020;29:52-65.
81. Burley SK, Berman HM, Christie C, et al. RCSB protein data bank: sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Sci*. 2018; 27(1):316-330.
82. Burley SK, Bromberg Y, Craig P, et al. Virtual boot camp: COVID-19 evolution and structural biology. *Biochem Mol Biol Educ*. 2020;48: 511-513.
83. Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data – from vision to reality. *Euro Surveill*. 2017;22(13). https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494
84. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall*. 2017;1(1):33-46.
85. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):265-269.
86. Semper C, Watanabe N, Savchenko A. Structural characterization of nonstructural protein 1 from SARS-CoV-2. *iScience*. 2021;24(1):101903.
87. Stogios PJ, Skarina T, Chang C, et al. Crystal structure of the ubiquitin-like domain 1 (Ubl1) of Nsp3 from SARS-CoV-2. 2020. https://doi.org/10.2210/pdb7KAG/pdb.
88. Frick DN, Virdi RS, Vuksanovic N, Dahal N, Silvaggi NR. Molecular basis for ADP-ribose binding to the Mac1 domain of SARS-CoV-2 nsp3. *Biochemistry*. 2020;59(28):2608-2615.
89. Stogios PJ, Skarina T, Di Lio R, et al. Crystal structure of the nucleic acid binding domain (NAB) of Nsp3 from SARS-CoV-2. 2021. https://doi.org/10.2210/pdb7LGO/pdb.

90. Owen CD, Lukacik P & Strain-Damerell CM et al. SARS-CoV-2 main protease with unliganded active site (2019-nCoV, coronavirus disease 2019, COVID-19). 2020. https://doi.org/10.2210/pdb6YB7/pdb.
91. Littler DR, Gully BS, Colson RN, Rossjohn J. Crystal structure of the SARS-CoV-2 non-structural protein 9, Nsp9. *iScience*. 2020;23(7):101258.
92. Kim Y, Wower J, Maltseva N, et al. Tipiracil binds to uridine site and inhibits Nsp15 endoribonuclease NendoU from SARS-CoV-2. *Commun Biol*. 2021;4(1):193.
93. Kern DM, Sorum B, Mali SS, et al. Cryo-EM structure of SARS-CoV-2 ORF3a in lipid nanodiscs. *Nat Struct Mol Biol*. 2021;28(7):573-582. http://dx.doi.org/10.1038/s41594-021-00619-0
94. Zhou Z, Huang C, Zhou Z, et al. Structural insight reveals SARS-CoV-2 ORF7a as an immunomodulating factor for human CD14(+) monocytes. *iScience*. 2021;24(3):102187.
95. Hall PD, Nelson CA, Fremont DH, Center for Structural Genomics of Infectious Diseases (CSGID). Crystal structure of the SARS-CoV-2 ORF8 protein. 2020. https://doi.org/10.2210/pdb7JX6/pdb.
96. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. 2018;46(W1):W296-W303.
97. Tan J, Vonrhein C, Smart OS, et al. The SARS-unique domain (SUD) of SARS coronavirus contains two macrodomains that bind G-quadruplexes. *PLoS Pathog*. 2009;5(5):e1000428.
98. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph*. 1996;14(1):33-38.
99. Phillips JC, Hardy DJ, Maia JDC, et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J Chem Phys*. 2020;153(4):044130.
100. MacKerell AD Jr, Bashford D, Bellott M, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*. 1998;102(18):3586-3616.
101. Sehnal D, Bittrich S, Deshpande M, et al. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res*. 2021;49:W431-W437.
102. Goodsell DS, Autin L, Olson AJ. Illustrate: software for biomolecular illustration. *Structure*. 2019;27:1716-1720.
103. *The PyMOL molecular graphics system* . [computer program]. 2002.
104. Chaudhury S, Lyskov S, Gray JJ. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*. 2010;26(5):689-691.
105. DiMaio F, Leaver-Fay A, Bradley P, Baker D, Andre I. Modeling symmetric macromolecular structures in Rosetta3. *PLoS One*. 2011;6(6):e20450.
106. Alford RF, Koehler Leman J, Weitzner BD, et al. An integrated framework advancing membrane protein modeling and design. *PLoS Comput Biol*. 2015;11(9):e1004398.
107. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins: Struct Funct Genet*. 2011;79(3):830-838.
108. Alford RF, Leaver-Fay A, Jeliazkov JR, et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J Chem Theory Comput*. 2017;13(6):3031-3048.
109. Alford RF, Fleming PJ, Fleming KG, Gray JJ. Protein structure prediction and design in a biologically realistic implicit membrane. *Biophys J*. 2020;118(8):2042-2055.

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: Lubin JH, Zardecki C, Dolan EM, et al. Evolution of the SARS-CoV-2 proteome in three dimensions (3D) during the first 6 months of the COVID-19 pandemic. *Proteins*. 2022;90(5):1054-1080. doi:10.1002/prot.26250