RCSB Protein Data Bank Resources for Structure-facilitated Design of mRNA Vaccines for Existing and Emerging Viral Pathogens

David S. Goodsell[1,2,3] and Stephen K. Burley[1,2,4,5 *]

[1] RCSB Protein Data Bank and Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA

[2] Rutgers Cancer Institute of New Jersey, Rutgers, The State University of New Jersey, New Brunswick, New Jersey 08903, USA

[3] Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, California 92037, USA

[4] Research Collaboratory for Structural Bioinformatics Protein Data Bank, San Diego Supercomputer Center, University of California, San Diego, California 92093, USA

[5] Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA

* Correspondence: stephen.burley@rcsb.org

Keywords

mRNA vaccine, structural biology, virus structure, structure-facilitated design, surface glycoprotein, carbohydrate, SARS-CoV-2, COVID-19

Summary

Structural biologists provide direct insights into the molecular bases of human health and disease. The open-access Protein Data Bank [PDB] stores and delivers three-dimensional [3D] biostructure data that facilitate discovery and development of therapeutic agents and diagnostic tools. We are in the midst of a revolution in vaccinology. Non-infectious mRNA vaccines have been proven during the COVID-19 pandemic. This new technology underpins nimble discovery and clinical development platforms that use knowledge of 3D viral protein structures for societal benefit. The RCSB Protein Data Bank supports vaccine designers through expert biocuration and rigorous validation of 3D structures; open-access dissemination of structure information; and search, visualization, and analysis tools for structure-guided design efforts. This Resource article examines the structural biology underpinning the success of SARS-CoV-2 mRNA vaccines and enumerates some of the many protein structures in the PDB archive that could guide design of new countermeasures against existing and emerging viral pathogens.

Highlights

Atomic structures of viral surface glycoproteins inform design of mRNA vaccines
Protein Data Bank provides open access to the world archive of biomolecular structure
RCSB PDB provides essential tools for finding and analyzing biomolecular structures

Graphical Abstract

Introduction

Structural biology represents an essential tool in our quest to understand fundamental biology, biomedicine, bioenergy, and biotechnology/bioengineering in 3D at atomic resolution (Burley et al., 2018). The Protein Data Bank [PDB] was established in 1971 with just seven X-ray crystal structures of proteins as the first open-access digital data resource in biology (Protein Data Bank, 1971). Now in its 50th year of continuous operations, the PDB is the single global archive of >180,000 experimentally-determined 3D structures of proteins, DNA, and RNA. Since 2003, the global PDB archive has been managed jointly by the Worldwide Protein Data Bank partnership [wwPDB; (Berman et al., 2003; wwPDB consortium, 2019)]. Current wwPDB member organizations include the US-funded RCSB Protein Data Bank [RCSB PDB; (Berman et al., 2000; Burley et al., 2021)], the Protein Data Bank in Europe [PDBe; (Armstrong et al., 2020)], Protein Data Bank Japan [PDBj; (Kinjo et al., 2018)], the Electron Microscopy Data Bank [EMDB; (Abbott et al., 2018)], and the Biological Magnetic Resonance Bank [BMRB; (Romero et al., 2020)]. In addition to the PDB archive, wwPDB members are also jointly responsible for global management of the EMDB and BMRB archives. The wwPDB supports tens of thousands of structural biologists from all inhabited continents, who freely contribute their data to the archive and many millions of PDB data consumers [*e.g.*, researchers, educators, students, policy makers, science funders, and the curious public] living and working in every sovereign nation and territory around the world (Burley et al., 2018).

Structural biologists and the PDB are playing critical roles in efforts to improve global health and fight disease in man, animals, and agricultural crops (Burley et al., 2018; Goodsell et al., 2020). Approximately 90% of United States Food and Drug Administration [US FDA] new drug approvals between 2010 and 2016 were facilitated by open access to PDB data, much of it contributed by researchers in universities, government laboratories, and not-for-profit research institutes largely funded by taxpayer monies (Galkina Cleary et al., 2018; Westbrook and Burley, 2019). Structural biologists and the PDB have had particularly significant impacts on discovery and development of antineoplastic agents (Westbrook et al., 2020). More than 70% of new small-molecule anticancer drugs approved by US FDA in 2010-2018 were products of structure-guided drug discovery in biopharmaceutical companies [reviewed in (Burley, 2021)]. In the vast majority of cases, these for-profit drug discovery efforts were enabled by open access to PDB structures of the drug target contributed by publicly-funded researchers. Every major drug company and many smaller biotechnology companies maintain copies of the PDB archive inside their firewalls for interoperation with proprietary information. The Charter governing wwPDB operations [https://www.wwpdb.org/about/agreement] expressly forbids charging PDB, EMDB, and BMRB data depositors and data consumers, and provides access to all archival information under the most permissive Creative Commons License CC0 1.0 Universal [https://creativecommons.org/publicdomain/zero/1.0/]. Every structure housed in the PDB archive is identified with a unique code [currently four alphanumeric character codes, *e.g.*, PDB ID 6lu7, the first deposited structure of the Severe Acute Respiratory Syndrome Coronavirus 2 or SARS-CoV-2 main protease]. Minimal information regarding each PDB structure can be accessed using its dedicated wwPDB landing page DOI: 10.2210/pdb6lu7/pdb. These DOIs may be used to provide citations to individual structures [strongly recommended for citing PDB structures lacking primary literature references describing structure determinations]. Links on each landing page in turn provide access to wwPDB partner structure summary pages [*e.g.*, https://www.rcsb.org/structure/6LU7, hosted by RCSB PDB].

With the growing realization that emerging viral pathogens pose an increasing threat to global health, structural biologists have aggressively explored the basic principles of virus biology, methods for using structure-guided drug discovery to develop antiviral agents, and new ways to

apply knowledge of viral structure to create safe and effective vaccines. Since late January 2020, more than 1,300 structures of SARS-CoV-2 proteins have been deposited into the PDB [https://rcsb.org/covid19]. Of central importance to this invited Resource article for the Structure Special Issue, entitled Structural Tools for Biological Discovery, are more than 400 PDB structures of the viral surface glycoprotein. These data are informing our understanding of SARS-CoV-2 variants and enabling 3D characterization of neutralizing antibodies generated in response to infection or vaccination, or engineered for passive immunization of infected individuals.

Vaccines represent one of the great successes of medical science, providing long-term protection against multiple life-threatening infections and saving hundreds of millions of lives (Pollard and Bijker, 2021; Rappuoli et al., 2021). The tried and true approach in the fight against viral diseases has been protein-based, administering viral antigens to stimulate an immune response. Given the phenomenal success of early vaccines, many variations on this approach have been developed and deployed, including inactivated viruses [*e.g.*, poliovirus]; live cell-culture adapted viruses [*e.g.*, measles, mumps, rubella]; empty viral capsids [*e.g.*, human papillomavirus]; recombinant viral proteins [*e.g.*, hepatitis B virus surface antigen]; and, more recently, engineered nanoparticles that display viral proteins [*e.g.*, Novavax]. The protein-based approach to vaccine development, however, is a long and expensive process that must be customized to each virus. A typical vaccine may require >10 years for discovery, development, and testing before regulatory approval (Kowalzik et al., 2021). New approaches using nucleic acids are currently being developed to shorten this timeline. Rather than challenging the immune system directly with viral antigens, these vaccines deliver genetic material that encodes immunogens. Once the delivered gene is transcribed and/or translated, viral proteins are displayed on host cell surfaces and presented to the cellular immune surveillance system. Current gene-based approaches include DNA-based vaccines using engineered adenoviruses and messenger RNA [mRNA] vaccines (Pardi et al., 2018).

Recent successes of the non-infectious Pfizer-BioNTech and Moderna mRNA vaccines discovered and developed for the COVID-19 pandemic have demonstrated the promise of this new, more nimble approach to vaccine development [Figure 1] (Park et al., 2021). mRNA vaccines have many advantages; they elicit both humoral [*i.e.*, antibody] and cell-mediated immune responses, while being well-tolerated by healthy individuals with few side effects and minimal risk of anaphylaxis. They are also far less expensive and time-consuming to develop. Next generation mRNA vaccines may be rapidly deployed by simply changing the sequence of the mRNA [*e.g.*, for protection against SARS-CoV-2 variants]. For existing or newly-emerging diseases, recent experience suggests that vaccine discovery and development timelines can be substantially reduced going forward.

This invited Resource article is the product of the US-funded RCSB Protein Data Bank. The RCSB PDB delivers PDB data using two web portals: research-focused https://rcsb.org and education/outreach-focused https://pdb101.rcsb.org. Herein, we briefly introduce a short history of modern mRNA vaccines, and then present resources that are available from the RCSB.org web portal to facilitate structure-guided approaches to mRNA vaccine design against existing and emerging viral pathogens. Several case studies are presented that exemplify use of structural data in vaccine design. We also briefly describe how the structure-facilitated design of mRNA vaccines is currently informing advances in the adjacent medical field of cancer therapy.

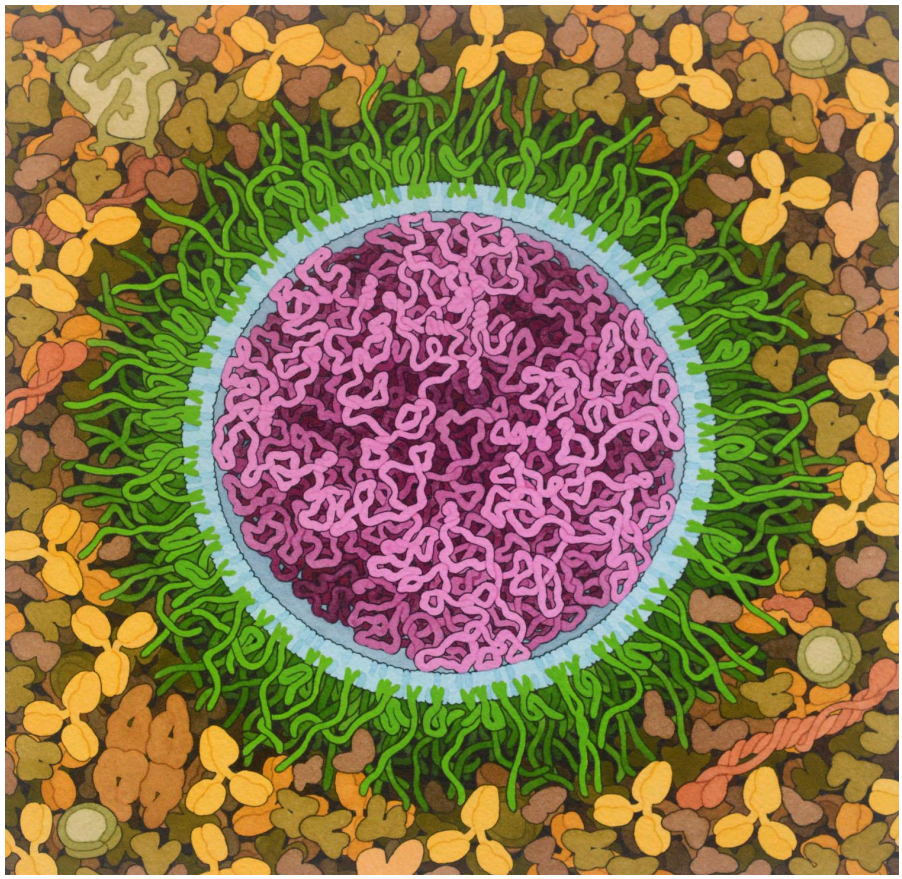

Figure 1. Idealized artistic conception of a SARS-CoV-2 mRNA vaccine. mRNA [magenta] is surrounded by a specialized lipid membrane, which typically includes PEGylated lipids [green] that protect the surface and ionizable lipids [blue] that are neutral at physiological pH, but become charged upon acidification of the endosome, thereby facilitating mRNA delivery into the cytoplasm. Surrounding the vaccine particle IgG antibodies and various human plasma proteins are depicted. Original painting [DOI: 10.2210/rcsb_pdb/goodsell-gallery-027] was created with traditional watermedia based on shapes and sizes of molecular structures taken from the PDB archive.

History and initial deployment of mRNA vaccines

Prior to the COVID-19 pandemic, a number of viral pathogens had become the focus of mRNA vaccine design and development efforts [*e.g.*, respiratory syncytial virus or RSV, rabies virus, Zika virus, and human cytomegalovirus or CMV] (Pardi et al., 2018). Perfecting this promising new vaccine design technology became imperative in the face of the COVID-19 global public health emergency.

A little more than twelve months after individuals infected with SARS-CoV-2 were first identified in Wuhan in the People's Republic of China, the Pfizer-BioNTech and Moderna mRNA vaccines against SARS-CoV-2 received Emergency Use Authorization in the US and other developed countries. The initial design of the Moderna mRNA-1273 vaccine was finalized 42 days after the sequence of the SARS-CoV-2 mRNA genome was made publicly available (Hodgson, 2020). During the design phase, researchers at both Pfizer-BioNTech and Moderna had open access to a total of 18 3D structures of the extracellular portion of the SARS-CoV-1 spike protein [first PDB

ID 5x5b, publicly released 5/3/2017 (Yuan et al., 2017)]. At approximately 78% amino acid sequence identity to its SARS-CoV-2 counterpart, they would have known that the 3D structures of the two spike proteins were highly similar (Sander and Schneider, 1991). For reference, the first PDB structure of a SARS-CoV-2 spike protein [PDB ID 6vsb] was publicly released on 2/26/2020 (Wrapp et al., 2020) revealing root-mean-square deviation of about 1.5Å with PDB ID 5x5b (Yuan et al., 2017) for 917 equivalent α-carbon pairs. Since PDB ID 6vsb became publicly available, more than 400 structures of SARS-CoV-2 spike proteins have been deposited into the PDB, including those of the spike protein bound to its cellular receptor, angiotensin converting enzyme 2 [ACE2], and various neutralizing antibodies.

By mid 2021, mass vaccination programs in much of the developed world [utilizing both mRNA vaccines described above and two adenovirus-based DNA vaccines: Oxford-AstraZeneca ChAdOx1 nCoV-19 or AZD1222; Johnson & Johnson JNJ-78436735 or Ad26.COV2.S] began to turn the tide of the pandemic. Israel, for example, had administered more than 10 million doses of the Pfizer vaccine [sufficient to fully vaccinate approximately 59% of the country's estimated population of more than nine million]. Also, by mid 2021, new infections in Israel had declined from a peak of more than 10,000 per day in mid-January 2021 to less than 100, and daily fatalities in Israel had declined from a peak of more than 60 per day in late January 2021 to zero [seven-day averages]. In Israel and around the globe the race is now on to vaccinate as many medically-eligible individuals as possible before existing and emerging hyper-transmissible variants of SARS-CoV-2 cause entirely preventable deaths in geographies with low vaccination rates and put those around the world who cannot be vaccinated for medical or religious reasons at needless risk of serious illness requiring hospitalization or death.

Visualizing viral surface proteins and structure/function relationships in 3D

Structural biology is playing a central role in the design of new vaccines, by revealing the structure/function relationships of the viral targets of vaccines and by providing ways to optimize the effectiveness of the target antigens used in vaccines. mRNA viral vaccine candidates are typically designed to elicit both humoral and cellular immune responses against viral surface proteins that are readily susceptible to antibody neutralization [Figure 2]. Macromolecular crystallography [MX] and single-particle cryo-electron microscopy tools [3DEM] are being used routinely to visualize viral proteins in 3D at the atomic level. Not surprisingly, these studies have revealed structural features of the viral surface proteins that present challenges for both structural biologists and vaccine designers.

First, many of these proteins undergo significant structural transitions during the course of viral infection. Viruses typically use their surface antigens to recognize and bind to one or more cellular receptors under physiologic conditions [*i.e.*, pH 7.4]. Following endocytosis, spike protein structures can change dramatically [triggered by acidification of the local environment that is mediated by proton pumps pre-positioned in the endosomal membrane] into a fusion-competent conformation (Sollner, 2004). These phenomena make for interesting extra work by structural biologists, requiring determination of multiple structures of different conformational states and computational modeling of putative conformations that cannot be captured and studied with current experimental methods. In addition, both conformationally-dynamic loops and membrane-spanning regions frequently pose experimental challenges for structure determination, particularly with MX wherein well-ordered crystals are required. Troublesome segments of the full-length polypeptide chain are often removed or replaced with more stable, engineered sequences before they can be visualized in 3D.

Second, many viral surface proteins are glycoproteins [*i.e.*, they are decorated with carbohydrate moieties resulting from enzymatic post-translational modification]. Glycosylation plays an important functional role in shielding portions of these proteins from immune surveillance (Julien et al., 2012). Glycosylation, however, can pose difficulties in structure determination because of the static or dynamic disorder and chemical heterogeneity of glycan chains. Frequently for MX studies, wherein carbohydrates are notorious for interfering with crystallization, sites of glycosylation are mutated to eliminate post-translational modification. Single-particle 3DEM does not require crystallization, permitting structural studies of glycoproteins in their native state. For vaccine designers, however, the entire glycan may not be visible in the structure-determination experiment using either MX or 3DEM, because of dynamic disorder. As a result, atomic-level 3D structures of glycoproteins in the PDB do not always provide information regarding the full complement of covalently bound sugars.

The RCSB PDB provides a number of resources to help PDB data consumers navigate these challenges. The research-focused RCSB.org web portal maintains strong and accessible connections to over 50 biodata resources [*e.g.*, UniProt (UniProt, 2021), NCBI/RefSeq (Li et al., 2021), GlyTouCan (Tiemeyer et al., 2017), GlyCosmos (Yamada et al., 2020), GlyGen (York et al., 2020)], allowing ready access to authoritative sequence and functional information. External sequence data integration can be useful, for example, in identifying regions of the polypeptide chain that are not represented in the atomic coordinates and learning more about their functional roles. wwPDB partners recently performed a remediation of carbohydrate-containing structures across the entire PDB archive (Shao et al., 2021). Nearly 15,000 PDB structures [~10% of archival holdings at the time] were remediated, including many viral glycoproteins. These remediated structures [and those of every glycoprotein that will be deposited to the PDB in future] use standardized atom and residue nomenclature for all carbohydrates based on the 1996 IUPAC recommendations (McNaught, 1996). All glycoproteins in the PDB are now properly annotated for glycosylation sites and all glycans are uniformly represented as branched oligosaccharides, utilizing the Symbol Nomenclature for Glycans [SNFG] representation standard (Varki et al., 2015). Improved representation and annotation now support glycosylation-specific searching and analyses using the RCSB.org web portal.

PDB archival holdings for viral surface glycoproteins

Given the importance of a structural understanding of the mechanisms of viral entry and immune neutralization, the structural biology community has launched a comprehensive effort to characterize many of the viruses currently posing risks to global health. Table 1 and Figure 2 include a representative selection of the many viral glycoprotein structures currently housed in the PDB archive. Well-studied exemplars include hundreds of individual PDB structures, providing a detailed portrait of the structure and function of each glycoprotein. The RCSB.org web portal includes powerful tools for streamlining exploration of the current holdings for a particular protein. The Structure Summary Page [SSP] of a representative PDB entry may be easily found through a simple text search using the main search bar. Once a relevant text search hit is identified and selected, the user is taken to the SSP. Therein, several tools allow enumeration of related entries. Because of the diversity of viruses [many of which have variants with proteins differing slightly in amino acid sequence and structure], use of several search options may be necessary to get a comprehensive view of current archival holdings for a particular spike protein. The simplest search option supports finding all structures corresponding to a given UniProt ID. For viruses that encode multiple proteins within one or more polyproteins, however, searching on UniProt ID may return PDB structure hits other than the desired spike protein. A more consistently reliable approach allows the PDB data consumer to search for all structures with a desired sequence identity to the representative entry [available sequence identity options: 100%, 95%, 90%, 80%, 70%, 60%,

40%, 30%]. One-click “Sequence” searching at 100% identity will not return structures of variants. When searching for variants of a particular glycoprotein, we recommend performing “Sequence” searching at 95% identity, which should eliminate false positives corresponding to related viruses [*e.g.*, SARS-CoV-2 and SARS-CoV spike proteins are ~78% identical]. Occasionally, PDB data depositors have studied chimeric structures composed of sequence segments from related viruses. “Sequence” searching at lower sequence identity can help reveal such cases. [N.B.: Care must be taken to exclude false positive “Structure” search results that encompass only a small fraction of the protein sequence, such as affinity purification tags.] When studying the structures of surface glycoproteins from related viruses [*e.g.*, from SARS-CoV-2 and other coronaviruses], one click “Structure” similarity searching from an RCSB.org SSP using our Zernike Polynomial-based system (Guzenko et al., 2020) can be very effective.

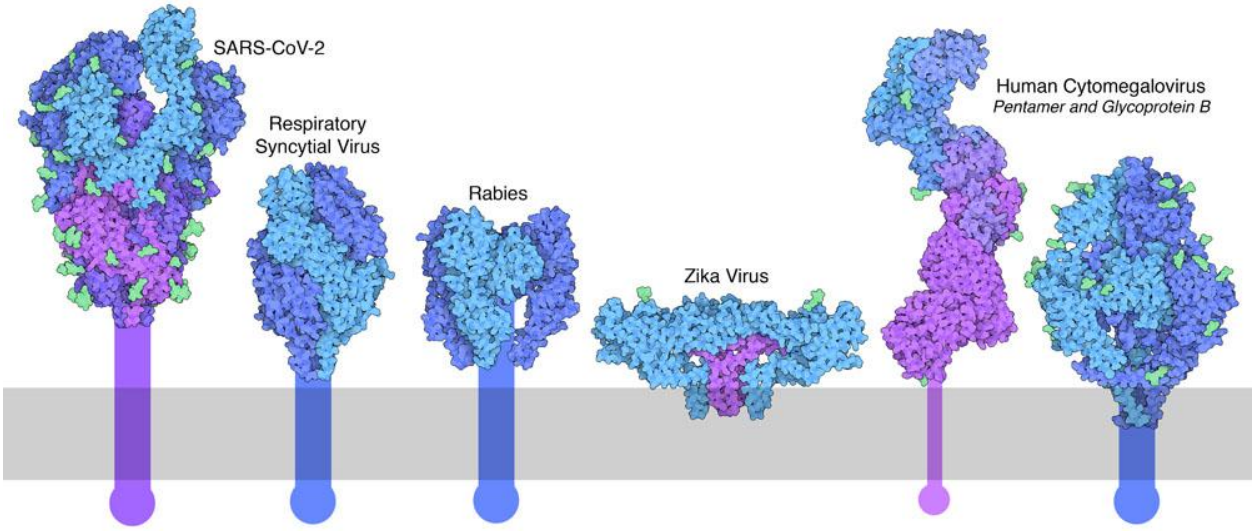

Figure 2. Selected viral surface glycoproteins currently being targeted with mRNA vaccines. Structures are available in the PDB archive for ectodomains, with proteins shown in shades of blue and purple and surface glycans [when included in the atomic coordinates] in green. The extent of the lipid bilayer is indicated in grey, and transmembrane portions are shown schematically. SARS-CoV-2 spike [PDB ID 6vyb (Walls et al., 2020)]; RSV fusion glycoprotein [PDB ID 4jhw (McLellan et al., 2013b)]; rabies virus glycoprotein [PDB ID 6lgx (Yang et al., 2020)] with trimeric assembly based on PDB ID 5i2s (Roche et al., 2007); Zika virus E [blue] and M [purple] proteins [PDB ID 5ire (Sirohi et al., 2016)]; cytomegalovirus (CMV) pentamer with subunits depicted in shades of blue and purple [PDB ID 5vob (Chandramouli et al., 2017)]; and CMV glycoprotein B [PDB ID 7kdp (Liu et al., 2021b)]. Figure created with Illustrate software [ccsb.scripps.edu/illustrate].

Table 1. Selected viral pathogen glycoprotein structures in the PDB.

Virus Name	Protein Name	PDB ID	UniProt Hits[1]	80% Hits[2]	Literature References
Coronaviruses					
SARS-CoV	Spike glycoprotein	6crz	49	43	(Kirchdoerfer et al., 2018)
SARS-CoV-2	Spike glycoprotein	6vsb	444	438	(Wrapp et al., 2020)
MERS-CoV	Spike glycoprotein	5w9n	38	32	(Pallesen et al., 2017)
Paramyxoviruses					
Respiratory syncytial virus A	Fusion glycoprotein F0	4jhw	63	82	(McLellan et al., 2013b)
Measles virus	Hemagglutinin	2zb5	6	8	(Hashiguchi et al., 2007)
Mumps virus	Hemagglutinin-neuraminidase	5b2d	4	4	(Kubota et al., 2016)
Rhabdoviruses					
Rabies virus	Rabies glycoprotein	6lgx	2	3	(Yang et al., 2020)
Vesicular stomatitis virus	Glycoprotein G	5i2s	3	5	(Roche et al., 2007)
Orthomyxoviruses					
Influenza virus	Hemagglutinin	1ruz	17	81 432[3]	(Gamblin et al., 2004)
	Neuraminidase	1nn2	11	35 218[3]	(Varghese and Colman, 1991)
Herpesviruses					
Herpes simplex virus	Envelope glycoprotein B	3nw8	12	12	(Stampfer et al., 2010)
Human cytomegalovirus	Envelope glycoprotein B	7kdp	6	6	(Liu et al., 2021b)
Flaviviruses					
Zika virus	Envelope protein E	5ire	48	42	(Sirohi et al., 2016)
Dengue virus	Envelope protein E	1tg8	2	45	(Zhang et al., 2004)
West Nile virus	Envelope protein E	2i69	1	13	(Kanai et al., 2006)
Tick-borne encephalitis virus	Envelope protein E	1svb	5	14	(Rey et al., 1995)
Hepatitis C virus	Envelope glycoprotein E2	4mwf	nd[4]	23[4]	(Kong et al., 2013)
Retroviruses					
HIV-1	Envelope glycoprotein	4nco	93	121	(Julien et al., 2013)

[1]Number of PDB structures with identical UniProt IDs *versus* the representative PDB ID, evaluated July 15, 2021.

[2]Number of PDB structures obtained using an 80% sequence identity search *versus* the representative PDB ID, then filtered using the "Scientific Name of Source Organism" Refinement on the Structure Summary Page. [N.B.: The non-intuitive cases where 80% sequence identity number is smaller than the UniProt ID search number is due in large part by the presence of PDB structures with small peptide fragments of the proteins, which are not recognized by the sequence similarity search, or inclusion of other domains in cases where the UniProt ID corresponds to a polyprotein.]

[3]Because influenza virus proteins are so diverse, a fuller representation of PDB holdings was evaluated using the “Structure” similarity search *versus* the representative PDB ID.
This page is intentionally blank.
[4]A 50% sequence identity search *versus* the representative PDB ID was used to capture multiple subtypes of HCV glycoprotein E2; the UniProt entry includes the entire genome polyprotein, so results for the UniProt search are not included here.

RCSB PDB and open access data

To make structure-enabled mRNA vaccine design possible, and indeed all structure-enabled science, standardized structure archiving, rigorous validation, expert biocuration, and facile data delivery are essential. Ample evidence of the central role played by the PDB archive has been published in peer reviewed scientific journals (Burley et al., 2018; Feng et al., 2020; Goodsell et al., 2020; Markosian et al., 2018), going well beyond the fields of structure-guided drug discovery (Burley, 2021; Westbrook and Burley, 2019; Westbrook et al., 2020) and protein structure prediction (Burley and Berman, 2021). The RCSB PDB and its wwPDB partners are dedicated to timely archiving of new results, continuing the 50 year PDB tradition of supporting scientific discovery and technical innovation based on experimental data freely contributed by structural biologists. Making good on this commitment involves complementary activities, including timely validation/biocuration and archiving of newly deposited information; open access to 3D structure data with no limitations on usage; provision of effective tools for searching and downloading archival data; and enabling web-based visualization and analysis of PDB structures.

The RCSB PDB response to the COVID-19 global pandemic highlights the PDB's long-standing adherence to the FAIR principles of Findability, Accessibility, Interoperability, and Reusability (Wilkinson et al., 2016). It is no exaggeration to state that the PDB was "walking the walk" decades before people began "talking the talk" about concepts such as FAIR and FACT [Fairness, Accuracy, Confidentiality, and Transparency] (van der Aalst et al., 2017). As illustrated in Figure 3, the first SARS-CoV-2 protein structure was deposited to the PDB within months of the initial outbreak. The shared commitment of the scientific community, including structural biologists, the PDB, and most scientific publishers was to make pandemic-related research results immediately accessible. This unprecedented level of cooperation, and our ability to build on abundant and freely-available structure data from previous coronavirus outbreaks, is supporting rapid discovery and development of multiple vaccines, neutralizing antibodies, and small-molecule drugs targeting SARS-COV-2.

In particular, we are enjoying the fruits of a "resolution revolution" in 3DEM (Kuhlbrandt, 2014), which is providing structural results for challenging biological systems at a pace far exceeding the capabilities of more established structure-determination techniques [*e.g.*, MX]. Improved sample preparation and cryo-preservation techniques, cryogenically cooled electron microscopes, direct electron detectors, and advances in software for data processing and structure determination are together providing new opportunities and posing new challenges to the scientific community, the wwPDB, and RCSB PDB. Of immediate concern is the need for new methods for assessing and validating 3DEM structural results and methods for interpretable display and exploration of large macromolecular assemblies. The wealth of structural information now available for the SARS-CoV-2 spike protein and its interactions with cell surface receptors and antibodies, described in more detail below, is testament to the power of 3DEM.

Recent admission of EMDB to the wwPDB partnership formalized a long-standing arrangement, wherein the OneDep global system for PDB structure deposition, validation, and biocuration served the needs of the 3DEM community. OneDep is the one-stop shop for deposition of atomic coordinates and supporting experimental data for structures determined using MX, 3DEM, and nuclear magnetic resonance [NMR] spectroscopy. Atomic coordinates for structures determined by MX, 3DEM, or NMR are stored in the PDB archive. Supporting experimental data are stored in the core archives jointly managed by the wwPDB; MX data are stored in the PDB archive, 3DEM maps are stored in the EMDB archive, and NMR data are stored in the BMRB archive. All three wwPDB core archives interoperate with one another.

The PDB is one of the most highly curated biological data archives. The PDB and individual structures therein are trusted by data depositors and data consumers alike. For this reason, and others discussed below, PDB usage is among the highest of any data repository in biology (Read et al., 2015).

Free availability of rigorously validated and expertly biocurated 3D structures from the PDB archive has enabled progress in the fields of structural biology and structural bioinformatics [reviewed in (Burley and Berman, 2021)] in myriad ways, including development of new structure determination methods, structure-guided drug discovery, predicting the impact of point mutations in proteins, comparative or homology protein structure modeling, protein-ligand pose prediction and scoring, prediction of protein-protein interactions, molecular dynamics simulations, and *de novo* protein structure prediction.

Open access to 3D biostructures supports established research efforts broadly in fundamental biology, biomedicine, bioenergy, and biotechnology/bioengineering, and studies of newly emerging topics. Because the vast majority of PDB data consumers are not structural biologists [and are indeed unlikely to ever contribute a structure to the archive], our RCSB.org web portal has been engineered (i) to enable searching for relevant structure(s), (ii) to provide integrated complementary information from trusted external data sources, and (iii) to support facile 3D visualization of structures:

(i) Multi-dimensional approaches to structure searching allow RCSB.org web portal users to pinpoint the data they need for their particular research or teaching needs. Simple text searching available at the top of every RCSB.org web page often suffices. A powerful, intuitive interface is also available for performing specialized searches across hundreds of structural features and descriptors with the option of using Boolean Logic to further narrow the results set. Once an initial set of structures is returned by the search system, the user can further narrow the results. A variety of options are available to manage examination of results set, such as easy "Refinement" checkboxes, which are particularly useful in cases with large numbers of structures, such as viral structures or structures related to immunoglobulin structure and binding. Browsing and detailed examination of individual structures is supported by dedicated SSPs, as described above for viral surface glycoproteins.

(ii) On every dedicated SSP, extensive interoperation with trusted external data sources facilitates exploration of multiple sources of information for the protein of interest. Currently more than 40 external data resources are integrated with PDB data. A robust pipeline connects individual PDB structures to corresponding UniProt (UniProt, 2021) and NCBI/RefSeq (Li et al., 2021) accession codes with mapping at the amino acid level between the 3D structure and the protein sequence in both cases. Sequence and function information in UniProt is available graphically in the RCSB Protein Explorer, and links embedded in the SSP allow direct access to UniProt pages and other PDB entries with the same UniProt ID. Reciprocally, UniProt provides a simple browser and viewing capability for PDB holdings, and direct links to wwPDB member web sites. Links are also provided to CATH (Sillitoe et al., 2021), SCOP (Andreeva et al., 2019), DrugBank (Wishart et al., 2018), PubChem (Kim et al., 2021), BindingDB (Gilson et al., 2016), and Pharos (Nguyen et al., 2017), all of which are relevant to the problem of discovery and development of viral pathogen countermeasures.

(iii) Turnkey molecular visualization is accessible on the RCSB.org web portal as a stand-alone feature and from within every dedicated SSP, using the web-native 3D molecular graphics display system Mol* (Sehnal et al., 2021). The principal advantage of Mol* *versus* other currently available web-native molecular graphic tools stems from deep integration of protein sequence information

with the 3D atomic coordinates made possible by its reliance on the PDBx/mmCIF data standard that underpins the PDB archive. This unique feature of Mol* enables navigation of PDB structures and communication with one-dimensional [1D] protein sequence features integrated from external data resources. The Mol* software library provides a technology stack for state-of-the-art data delivery, web-native molecular graphics, and analysis tools for interrogating 3D macromolecular structure data. Mol* is the collaboratively developed successor of the RCSB PDB NGL Viewer (Rose et al., 2018) and Protein Data Bank in Europe LiteMol (Sehnal et al., 2017). It works entirely within the user's web browser, avoiding the need to license, download, install, or maintain external software.

Mol* supports routine display of atoms and interatomic bonds, metal ions, and bound ligands in a variety of commonly used biomolecular representational styles and rendering of molecular surfaces for depiction of protein-protein interfaces and small-molecule binding sites. It also supports graphical display and analyses [structural interrogation] of intra- and inter-molecular contacts, and more generally intermolecular contacts between any number of polymer chains and ligands in macromolecular assemblies of any size [*e.g.*, larger than an entire ribosome]. Mol* provides a user-friendly and rapid way to visualize structures on-the-fly, streamlining exploration of many structures during the initial search phase of a study, and providing advanced options for detailed study and analysis of the most salient structures.

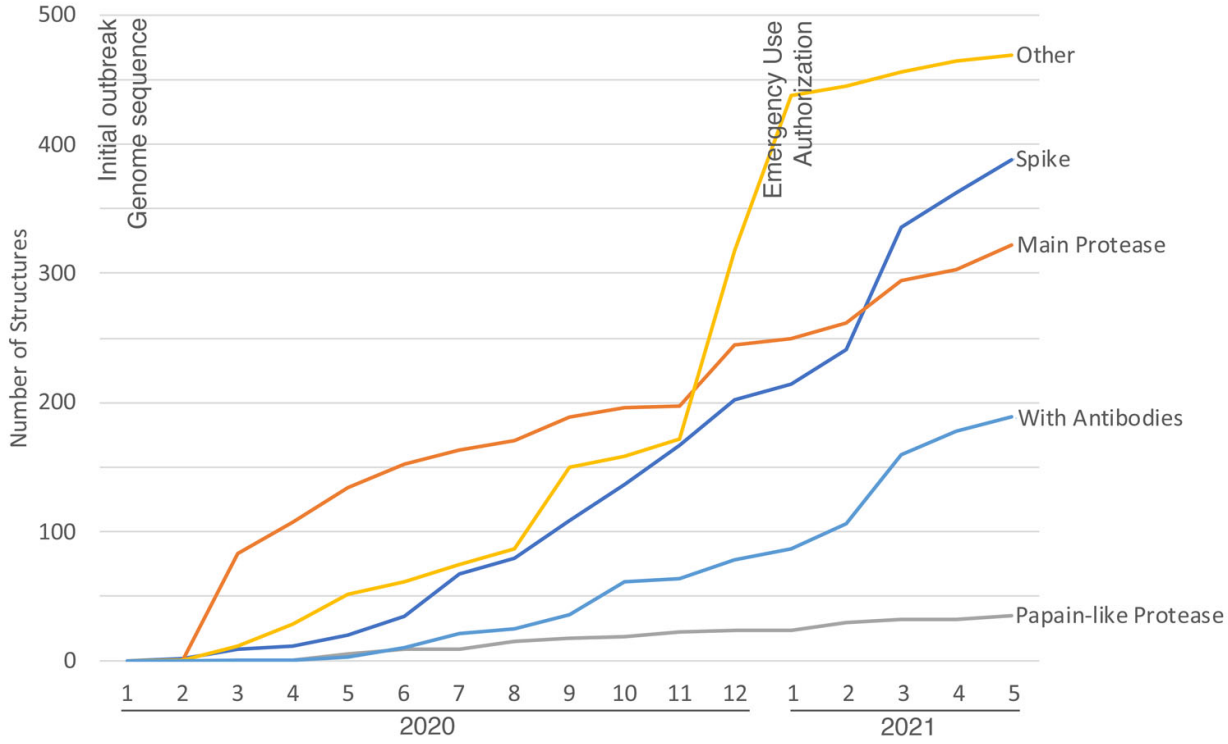

Figure 3. PDB archival holdings related to SARS-CoV-2 proteins accumulated during the COVID-19 global pandemic.

Vaccine Design Case Studies: Prefusion state stabilization of viral glycoproteins

The power of structure-facilitated vaccine design is being demonstrated in an ongoing effort to discover and develop a vaccine against respiratory syncytial virus [RSV], which is the most common cause of bronchiolitis and pneumonia in children under one year of age in the US.

Infants, young children, and older adults with chronic medical conditions are at risk of severe disease from RSV infection. Annually in the US, RSV is responsible on average for approximately 58,000 hospitalizations with 100-500 deaths among children younger than 5 years old and 177,000 hospitalizations with 14,000 deaths among adults aged 65 years or older. RSV, a member of the orthopneumovirus family, is an enveloped virus carrying a negative-sense single-stranded RNA genome that encodes 11 proteins [~15,000 nucleotides in length]. Although there is currently no vaccine against RSV, passive immunization with monoclonal antibodies [MAbs] is available to prevent RSV infection and hospitalization in infants at highest risk. F and G glycoproteins are the two major surface proteins that control viral attachment and the initial stages of infection. They are the primary targets for neutralizing antibodies during natural infection (Battles and McLellan, 2019). The G protein is produced as either a membrane-bound form that mediates viral attachment, or a secreted form involved in immune evasion. It is largely disordered, but structures of a central region have revealed details of its interactions with antibodies The most successful anti-RSV MAb palivizumab [Astra Zeneca, brand name Synagis] is directed against the A antigenic site of the surface fusion [F] glycoprotein of RSV. The fusion glycoprotein is also the target in a structure-facilitated design effort for vaccine development.

Analysis of antibody binding revealed that the most effective antigenic sites are present on the metastable pre-fusion state of the F glycoprotein, suggesting that stabilization of this conformation would lead to a more effective vaccine (Gilman et al., 2016; Magro et al., 2012). Based on structures of the glycoprotein ectodomain in pre- and post-fusion conformations, a series of stabilizing amino acid changes were predicted and tested [Figure 4]. These substitutions included addition of a disulfide bridge between a pair of amino acids that are 4.4 Å apart in the prefusion form but 124.2 Å apart in the post fusion conformation, plus several changes intended to fill small cavities and a T4-phage fibritin trimerization domain [foldon] to stabilize the C terminus (McLellan et al., 2013a). The resulting prefusion-stabilized form of the protein showed much improved RSV-neutralizing activity. This landmark study has led to subsequent work demonstrating a proof of concept for effectiveness of this approach in as subunit-based vaccine in humans (Crank et al., 2019) and use in an mRNA vaccine tested in rodent models (Espeseth et al., 2020).

Building on this work, the idea of stabilizing prefusion conformational states of surface glycoproteins has been applied to coronaviruses. Analysis of RSV and HIV surface glycoproteins identified a structurally-critical region in their interiors, linking the heptad repeat [HR1] to the central α-helices. Adding two prolines to this loop was found to stabilize the prefusion form of Middle East Respiratory Syndrome Coronavirus [MERS-CoV] glycoprotein and other coronaviruses (Pallesen et al., 2017). Similar stabilizing prolines are included in both the Moderna and BioNTech/Pfizer mRNA vaccines against SARS-CoV-2, and combined with inactivation of the furin cleavage site in the virus-vectored vaccine from Janssen/Johnson&Johnson and a subunit vaccine from Novavax [reviewed in (Dai and Gao, 2021)]. We expect that a similar approach will be applied to all manner of viral targets, as the body of structural information grows. For example, the recent structure of hepatitis C glycoprotein E2 in complex with its cellular receptor [PDB ID 7mwx, (Kumar et al., 2021)], when compared with a decade of previous structural studies of the glycoprotein alone and with neutralizing antibodies, reveals conformational changes upon acidification in preparation for membrane fusion. These structural insights will potentially allow targeted design of prefusion-stabilized glycoprotein in future vaccine development programs.

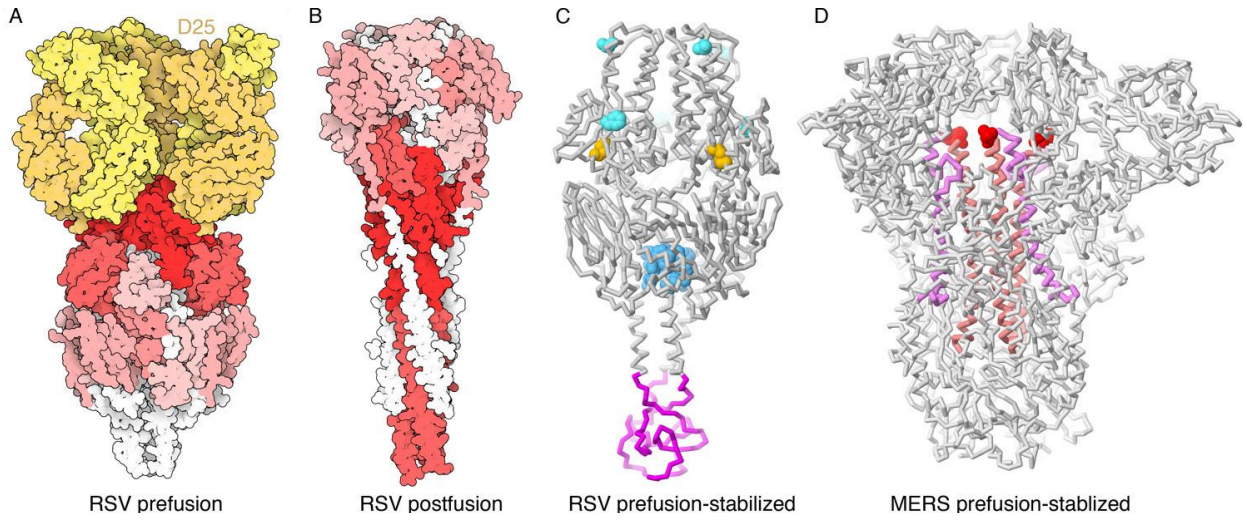

Figure 4: Stabilization of pre-fusion conformations of viral surface glycoproteins. (A) Space-filling representation of the pre-fusion conformation of the RSV glycoprotein F ectodomain [PDB ID 4jhw (McLellan et al., 2013b)], with antibody epitopes colored by neutralization sensitivity [dark red - highest; light red-intermediate, pink-lowest, white-antibody inaccessible], and a neutralizing antibody D25 Fab shown in yellow. (B) Space-filling representation of the post-fusion conformation of the RSV glycoprotein F ectodomain [PDB ID 3rrr (McLellan et al., 2011)], showing that the most neutralization sensitive epitopes are inaccessible to antibodies in the altered conformation. (C) Polypeptide chain backbone representation of the RSV F glycoprotein stabilized in the prefusion conformation [PDB ID 4mmv (McLellan et al., 2013a)] with an engineered disulfide bridge [yellow], several sites of mutation to fill pockets (turquoise and blue), and a foldon to stabilize the homotrimer [magenta]. (D) Polypeptide chain backbone representation of the MERS-CoV virus spike protein homotrimer stabilized in a prefusion conformation [PDB ID 5w9j (Pallesen et al., 2017)] with two prolines [red spheres] linking the heptad repeat [HR1, magenta] and the central helix [pink]. (A) and (B) created with Illustrate software; (C) and (D) created with Mol* (Sehnal *et al.*, 2021) at the RCSB PDB website.

Looking Ahead: The challenge of SARS-CoV-2 variants

One of the great promises of mRNA vaccines is that they will provide a timely and effective way of responding to newly emerging variants of SARS-CoV-2. Every time the virus replicates in a human [or animal] host, there is a non-zero likelihood that one or more of the viral protein sequences [and consequently 3D structures] will change. Coronaviruses have the longest RNA virus genomes of all known single-stranded RNA viruses [~30,000 nucleotides]. SARS-CoV-2 RNA-dependent RNA polymerase [multi-subunit enzymes composed of non-structural proteins or nsps 7, 8, and 12] acts in concert with an RNA helicase [nsp13] and a proofreading exonuclease [nsp14] to carry out efficient and relatively faithful copying of the lengthy genome (Denison et al., 2011). Proofreading notwithstanding, coronavirus genome replication is not perfect, and coronaviruses evolve as they *passage* serially from one host to the next (Harvey et al., 2021). A recently published study of SARS-CoV-2 protein evolution in 3D during the first six months of the pandemic examined amino acid changes in >48,000 viral isolates and documented how each one of the 29 viral proteins underwent amino acid changes (Lubin et al., 2020).

Of greatest concern at the time of writing is the Delta variant of SARS-CoV-2 [B.1.617.2, (https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/)] with spike glycoprotein

substitutions: T19R, V70F, T95I, G142D, E156-, F157-, R158G, A222V, W258L, K417N, L452R, T478K, D614G, P681R, D950N (https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html). These substitutions include a critical change in the receptor binding domain [position 452] that is thought to confer a stronger binding to the ACE2. An additional change at position 681 affects the rate of cleavage of the spike protein precursor (Cherian et al., 2021). Figure 5 illustrates the locations of some amino acid changes in the Delta variant spike protein. Initial reports suggest that both the Pfizer-BioNTech and Moderna mRNA vaccines offer high levels of protection against serious illness requiring hospitalization and death for doubly vaccinated individuals [but not for those who are singly vaccinated or previously infected with another SARS-CoV-2 variant] (Callaway, 2021). Experts around the world concur that continued infections in geographies with low rates of complete vaccination represent potential breeding grounds for new variants that could threaten a second global pandemic. Both Pfizer and Moderna have reported progress in developing booster mRNA vaccines, which could be re-designed to include Delta and/or possibly newer post-Delta variants. Structural biologists are providing critical information archived in the PDB that contribute to our understanding of new SARS-CoV-2 spike protein variants and how substitutions at various positions within the homotrimer impact function, infectivity, and virus neutralization by antibodies of either natural or man-made origin.

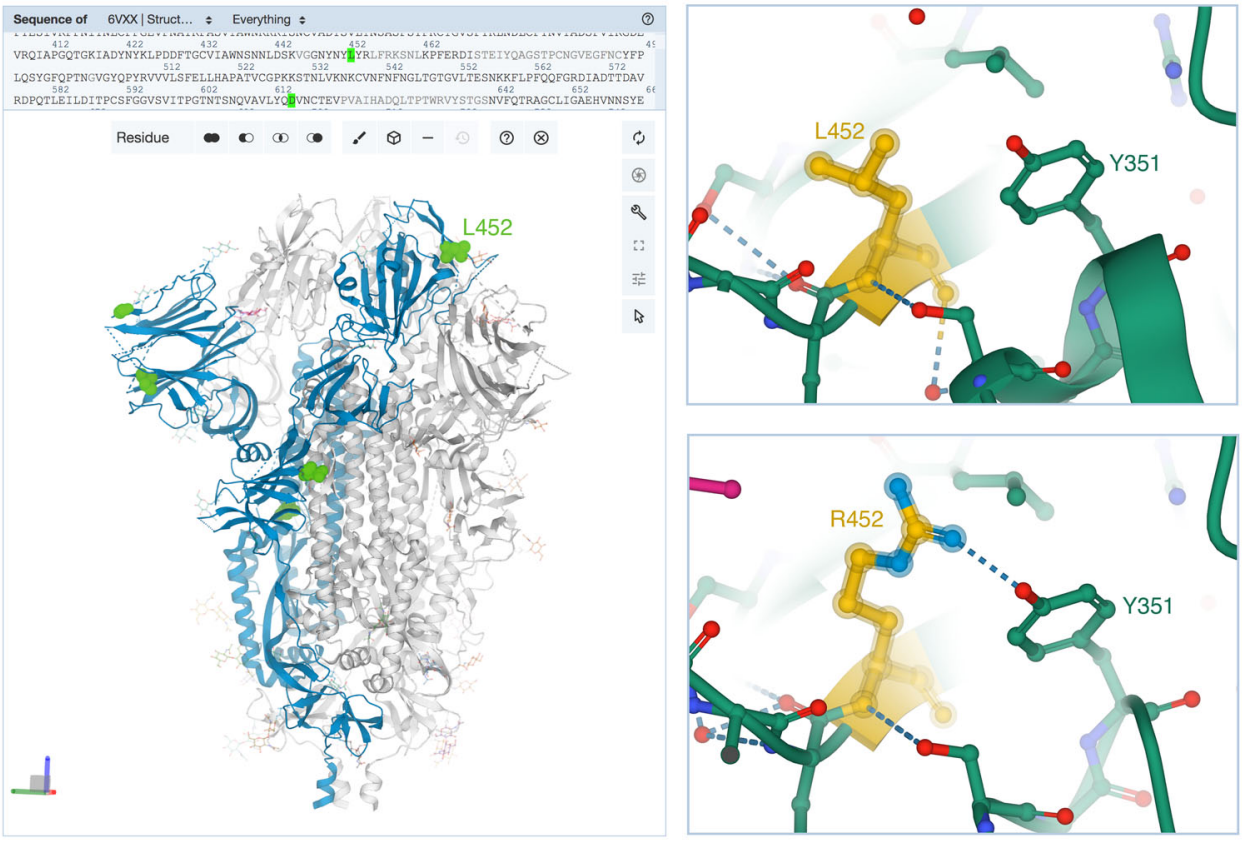

Figure 5. SARS-CoV-2 spike protein Delta variant structure. (Left) Ribbon representation of the SARS-CoV-2 spike protein homotrimer [one protomer in blue] with attached carbohydrates [pink atomic stick figures] created using the RCSB.org web portal with Mol*. Selected sites of amino acid substitutions in the Delta variant [*versus* the original viral isolate] are highlighted in green from the upper sequence selection panel [PDB ID 6vxx (Walls et al., 2020)]; (Upper right) Close-up view of L452 occurring in the receptor-binding domain was generated in Mol* by clicking on the position, triggering automatic Mol* display of the location showing interactions with neighboring amino acids [PDB ID 7ora (Liu et al., 2021a)]. (Lower right) Consequences of the L452R substitution showing a new hydrogen bond (dashed line) with Y351 [PDB ID 7orb (Liu et al., 2021a)]. Atom color coding: C-green or yellow; N-blue; O-red. Figure created with Mol* (Sehnal *et al.*, 2021) at the RCSB PDB website.

The RCSB.org web portal 3D Protein Feature View (PFV) is equipped with tools that assist users in relating changes in 3D structure due to sequence/structure variation with changes in function. Based on data integrated from UniProt, the PFV includes mappings of variant locations with assorted protein sequence features, including domains, proteolytic processing sites, glycosylation sites, *etc*. The 1D graphical view is also linked to a Mol* structure viewer, allowing protein sequence features to be visualized in the context of atomic level 3D structure.

Figure 6 exemplifies uses of the 3D PFV, illustrating the structure of the SARS-CoV-2 spike protein D614G variant implicated in increased infectivity. Analysis of >33,000 viral genomes sequenced before late June 2020, revealed that ~74% of the viral isolates possessed the D614G amino acid substitution in their spike proteins (Lubin et al., 2020). Structural characterization of an engineered D614G spike protein by 3DEM revealed a significantly increased populations of

conformations in which the ACE2 receptor binding domains occur in the open state, presumed to be due to loss of an interaction with T859 in a neighboring subunit [PDB ID 6xs6 (Yurkovetskiy et al., 2020)]. Alteration of the equilibrium between closed and open states is thought to explain increased infectivity. Figure 6 Right shows that G614 is ~8Å from the sidechain of T859 of another copy of the protein within the trimer. In the structure of the spike protein from the original viral isolate, D614 makes a hydrogen bond with T859, stabilizing the structure in a closed state that is unable to interact with ACE2 [data not shown].

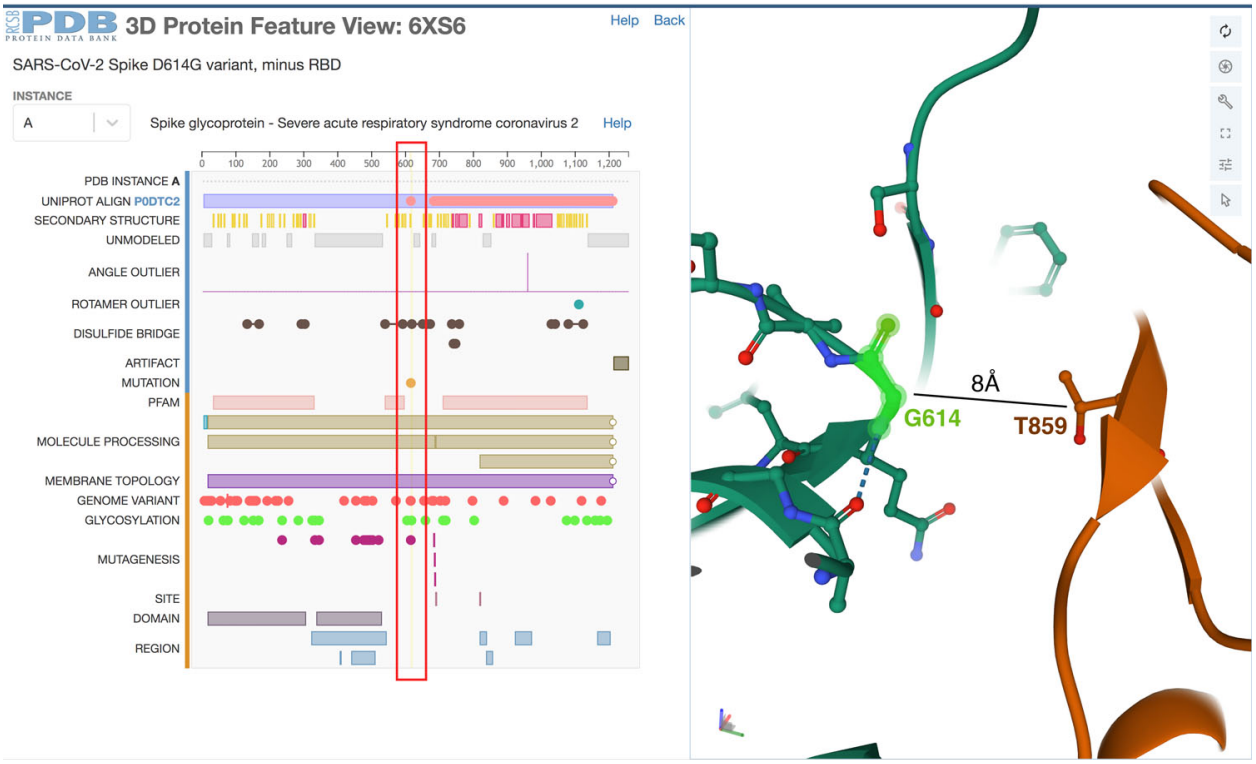

Figure 6. Screenshot of the RCSB PDB 3D Protein Feature View for the D614G substituted form of the SARS-CoV-2 spike protein PDB ID 6xs6 (Yurkovetskiy et al., 2020). (Left) The location of the substitution is selected using the sequence-based display in the Protein Feature View (red box). (Right) The amino acid G614 is highlighted with a green halo using Mol* structure viewer, revealing loss of a stabilizing interaction with T859 in a neighboring protomer within the homotrimer.

Design of future anti-viral mRNA vaccines and potential anti-cancer therapies

The wealth of 3D structure information available for the SARS-CoV-2 spike protein and its potential to influence design of second generation COVID-19 mRNA vaccines suggests that there is much to be optimistic about for design of additional mRNA vaccines against other viral pathogens. mRNA vaccines are non-infectious, do not become integrated into the host genome, stimulate both humoral and cell-based immunity, are well tolerated by healthy individuals with few side effects, can be rapidly designed, and are less expensive to develop and manufacture. They have enormous potential as vehicles for addressing viral diseases more broadly. A number of candidate mRNA vaccines are currently being evaluated in clinical trials for Zika virus, metapneumovirus, parainfluenzavirus, cytomegalovirus, rabies, and others (Wadhwa et al.,

2020), while contributing to a multi-pronged effort to combat the COVID-19 pandemic (Park et al., 2021).

mRNA vaccines are also making inroads as experimental agents for treating human cancers. The past decade has witnessed a revolution in cancer immunotherapy with approval of monoclonal antibodies that recognize cell-surface antigens [first CTLA-4, followed by PD-1 and then PD-L1] that are responsible for downregulation of T-cell responses to cancers expressing neoantigens specific to the tumor. In a subset of antibody-treated individuals [with the exact proportion depending on the type of cancer], interdiction of the T-cell immune checkpoint prevents the malignant cell from "persuading" the T-cell that it should not be killed. President Jimmy Carter was on the verge of death due to late-stage melanoma when he received pembrolizumab [Merck, brand name Keytruda]. At the time of writing, Carter was in long-term remission and may well have been cured. Current challenges with immune checkpoint therapies include their relatively low response rates and ascertaining why some individuals respond to antibody treatment while others do not. mRNA vaccines represent a potentially important means of increasing the likelihood of long-term remissions for some cancers.

Both Moderna and BioNTech have been active in this arena with their respective lipid nanoparticle formulations for mRNA delivery. Other cancer vaccine approaches include naked synthetic mRNA, an individual's own dendritic cells that have been manipulated and expanded *ex vivo*, protamine formulations, and self-amplifying mRNA [SAM] vaccines [reviewed in (Miao et al., 2021)]. A SAM vaccine vehicle carries viral replication machinery capable of self-amplifying over 1-2 months and inducing more potent and persistent immune responses. SAM platforms are expected to support significant antigen production following vaccination with very low doses [*versus* those used for non-replicating mRNA vaccines]. As for design of antiviral mRNA vaccines, the challenge will be to deliver a genetic payload that encodes the right antigen. Indeed, selection of tumor-associated or tumor-specific antigens [TAAs or TSAs] preferentially expressed in malignant cells appears to be the problem. BioNTech's BNT111 vaccine encoding four TAAs identified in melanoma cells has yielded T-cell responses in early stage clinical trials. Alphavax has also reported encouraging results for their AVX701 SAM, which encodes carcinoembryonic antigen or CEA. While certainly promising, mRNAs encoding one or more TAAs may not prove to be broadly effective. Individuals with highly variable TAAs present in their polyclonal tumors due to errors in DNA replication may not respond to vaccines that deliver mRNAs encoding wild-type proteins. There is also the problem of autoimmune attack of normal tissues that may encode TAAs.

An alternative approach to antigen selection focuses on neoantigens, which derive from random somatic mutations in malignant cells. The success of the immune checkpoint antibodies is thought to result from T-cell recognition of neoantigens as non-self-proteins. The challenge of this approach is identification of oligopeptide fragments of mutant proteins that have high immunogenicity. Moderna and Merck, a leader in antibody immunotherapy, are collaborating on development of mRNA-5671, which encodes four well-characterized somatic mutations of KRAS affecting amino acid glycine 12 [see PDB ID 4lr2 for the 3D structure of KRAS G12C (Ostrem et al., 2013)], for use as a monotherapy and in combination with pembrolizumab in participants with KRAS mutant advanced or metastatic non-small cell lung cancer, colorectal cancer, or pancreatic cancer. BioNTech is currently collaborating with Genentech [a member of the Roche group] on individualized neoantigen specific therapy using its mRNA delivery platform, which aims to customize vaccines according to the repertoire of immunogenic neoantigens detected in a particular tumor.

In all of these exciting developments, the PDB archive has played an essential role by providing open access to the global corpus of structural knowledge and the RCSB PDB has provided the resources to find and utilize this information. This has streamlined the structural understanding of the basic mechanisms of carcinogenesis, facilitated the structure-based design of antineoplastic drugs, and provided detailed structural understanding of antibodies and their neutralization of cancer antigens. Indeed, a recent analysis revealed that open access to 3D structural information facilitated the discovery and development of the majority of recent antineoplastic agents, including 25 biologics (Westbrook *et al.*, 2020).

While no one expects that every new effort involving an mRNA vaccine will succeed, it appears likely that at least some of them will work and improve the range of options available to intervene medically and improve the human condition. Strategic investments by Pfizer/BioNTech and in parallel by Moderna [with considerable financial assistance from the US government and philanthropic contributors] in the face of the global pandemic transformed the landscape for mRNA vaccine design and development. The public now knows much more about the science and technology of mRNA vaccines and how they can generate live-saving results in a short time. Absent COVID-19, it would have taken much longer for the technology to mature and become broadly available. There is much more to come and the RCSB PDB team looks forward to ensuring that the Protein Data Bank contributes to the good of humankind for another 50 years in this arena and more broadly across the biological and biomedical sciences.

Acknowledgements

First and foremost, the authors thank the tens of thousands of structural biologists who deposited structures to the PDB since 1971 and the many millions around the world who consume PDB data. We also thank Christine Zardecki for assistance with manuscript preparation. Finally, the authors gratefully acknowledge contributions to the success of the PDB archive made by all members of RCSB PDB [past and present] and our PDBe, PDBj, EMDB, and BMRB wwPDB partners.

Funding

RCSB PDB is jointly funded by the National Science Foundation [DBI-1832184, PI: S.K. Burley], the US Department of Energy [DE-SC0019749, PI: S.K. Burley], and the National Cancer Institute, National Institute of Allergy and Infectious Diseases, and National Institute of General Medical Sciences of the National Institutes of Health under grant R01GM133198 [PI: S.K. Burley]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Declaration of Interests

The authors declare no conflicts of interest with the contents of this article.

Author Contributions

DSG and SKB contributed to the conceptualization, investigation, visualization and writing of this article. SKB is responsible for supervision, project administration, and securing funding for the RCSB PDB.

References

Abbott, S., Iudin, A., Korir, P.K., Somasundharam, S., and Patwardhan, A. (2018). EMDB Web Resources. Curr Protoc Bioinformatics *61*, 5.10.11-15.10.12. 10.1002/cpbi.48.

Andreeva, A., Kulesha, E., Gough, J., and Murzin, A.G. (2019). The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Research *48*, D376-D382. 10.1093/nar/gkz1064.

Armstrong, D.R., Berrisford, J.M., Conroy, M.J., Gutmanas, A., Anyango, S., Choudhary, P., Clark, A.R., Dana, J.M., Deshpande, M., Dunlop, R., et al. (2020). PDBe: improved findability of macromolecular structure data in the PDB. Nucleic Acids Res *48*, D335-D343. 10.1093/nar/gkz990.

Battles, M.B., and McLellan, J.S. (2019). Respiratory syncytial virus entry and how to block it. Nat Rev Microbiol *17*, 233-245. 10.1038/s41579-019-0149-x.

Berman, H.M., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. Nature Structure Biology *10*, 980. 10.1038/nsb1203-980.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-242. 10.1093/nar/28.1.235.

Burley, S.K. (2021). Impact of structural biologists and the Protein Data Bank on small-molecule drug discovery and development. J Biol Chem, 100559. 10.1016/j.jbc.2021.100559.

Burley, S.K., and Berman, H.M. (2021). Open-access data: A cornerstone for artificial intelligence approaches to protein structure prediction. Structure *29*, 515-520. 10.1016/j.str.2021.04.010.

Burley, S.K., Berman, H.M., Christie, C., Duarte, J.M., Feng, Z., Westbrook, J., Young, J., and Zardecki, C. (2018). RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. Protein Sci *27*, 316-330. 10.1002/pro.3331.

Burley, S.K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G., Christie, C.H., Dalenberg, K., Costanzo, L.D., Duarte, J.M., et al. (2021). 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 Research *49*, D437-D451. 10.1093/nar/gkaa1038.

Callaway, E. (2021). Delta coronavirus variant: scientists brace for impact. Nature *595*, 17-18. 10.1038/d41586-021-01696-3.

Chandramouli, S., Malito, E., Nguyen, T., Luisi, K., Donnarumma, D., Xing, Y., Norais, N., Yu, D., and Carfi, A. (2017). Structural basis for potent antibody-mediated neutralization of human cytomegalovirus. Sci Immunol *2*. 10.1126/sciimmunol.aan1457.

Cherian, S., Potdar, V., Jadhav, S., Yadav, P., Gupta, N., Das, M., Rakshit, P., Singh, S., Abraham, P., Panda, S., and team, N. (2021). Convergent evolution of SARS-CoV-2 spike mutations, L452R, E484Q and P681R, in the second wave of COVID-19 in Maharashtra, India. bioRxiv, 2021.2004.2022.440932. 10.1101/2021.04.22.440932.

Crank, M.C., Ruckwardt, T.J., Chen, M., Morabito, K.M., Phung, E., Costner, P.J., Holman, L.A., Hickman, S.P., Berkowitz, N.M., Gordon, I.J., et al. (2019). A proof of concept for structure-based vaccine design targeting RSV in humans. Science *365*, 505-509. 10.1126/science.aav9033.

Dai, L., and Gao, G.F. (2021). Viral targets for vaccines against COVID-19. Nat Rev Immunol *21*, 73-82. 10.1038/s41577-020-00480-0.

Denison, M.R., Graham, R.L., Donaldson, E.F., Eckerle, L.D., and Baric, R.S. (2011). Coronaviruses. RNA Biology *8*, 270-279. 10.4161/rna.8.2.15013

Espeseth, A.S., Cejas, P.J., Citron, M.P., Wang, D., DiStefano, D.J., Callahan, C., Donnell, G.O., Galli, J.D., Swoyer, R., Touch, S., et al. (2020). Modified mRNA/lipid nanoparticle-based

vaccines expressing respiratory syncytial virus F protein variants are immunogenic and protective in rodent models of RSV infection. NPJ Vaccines *5*, 16. 10.1038/s41541-020-0163-z.
Feng, Z., Verdiguel, N., Di Costanzo, L., Goodsell, D.S., Westbrook, J.D., Burley, S.K., and Zardecki, C. (2020). Impact of the Protein Data Bank Across Scientific Disciplines. Data Science Journal *19*, 1-14. 10.5334/dsj-2020-025.
Galkina Cleary, E., Beierlein, J.M., Khanuja, N.S., McNamee, L.M., and Ledley, F.D. (2018). Contribution of NIH funding to new drug approvals 2010-2016. Proc Natl Acad Sci U S A *115*, 2329-2334. 10.1073/pnas.1715368115
1715368115 [pii].
Gamblin, S.J., Haire, L.F., Russell, R.J., Stevens, D.J., Xiao, B., Ha, Y., Vasisht, N., Steinhauer, D.A., Daniels, R.S., Elliot, A., et al. (2004). The structure and receptor binding properties of the 1918 influenza hemagglutinin. Science *303*, 1838-1842.
Gilman, M.S., Castellanos, C.A., Chen, M., Ngwuta, J.O., Goodwin, E., Moin, S.M., Mas, V., Melero, J.A., Wright, P.F., Graham, B.S., et al. (2016). Rapid profiling of RSV antibody repertoires from the memory B cells of naturally infected adult donors. Sci Immunol *1*. 10.1126/sciimmunol.aaj1879.
Gilson, M.K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., and Chong, J. (2016). BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res *44*, D1045-1053. 10.1093/nar/gkv1072.
Goodsell, D.S., Zardecki, C., Di Costanzo, L., Duarte, J.M., Hudson, B.P., Persikova, I., Segura, J., Shao, C., Voigt, M., Westbrook, J.D., et al. (2020). RCSB Protein Data Bank: Enabling biomedical research and drug discovery. Protein Sci *29*, 52-65. 10.1002/pro.3730.
Guzenko, D., Burley, S.K., and Duarte, J.M. (2020). Real time structural search of the Protein Data Bank. PLoS Comput Biol *16*, e1007970. 10.1371/journal.pcbi.1007970.
Harvey, W.T., Carabelli, A.M., Jackson, B., Gupta, R.K., Thomson, E.C., Harrison, E.M., Ludden, C., Reeve, R., Rambaut, A., Consortium, C.-G.U., et al. (2021). SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol *19*, 409-424. 10.1038/s41579-021-00573-0.
Hashiguchi, T., Kajikawa, M., Maita, N., Takeda, M., Kuroki, K., Sasaki, K., Kohda, D., Yanagi, Y., and Maenaka, K. (2007). Crystal structure of measles virus hemagglutinin provides insight into effective vaccines. Proc Natl Acad Sci U S A *104*, 19535-19540. 10.1073/pnas.0707830104.
Hodgson, J. (2020). The pandemic pipeline. Nat Biotechnol *38*, 523-532. 10.1038/d41587-020-00005-z.
Julien, J.P., Cupo, A., Sok, D., Stanfield, R.L., Lyumkis, D., Deller, M.C., Klasse, P.J., Burton, D.R., Sanders, R.W., Moore, J.P., et al. (2013). Crystal structure of a soluble cleaved HIV-1 envelope trimer. Science *342*, 1477-1483. 10.1126/science.1245625.
Julien, J.P., Lee, P.S., and Wilson, I.A. (2012). Structural insights into key sites of vulnerability on HIV-1 Env and influenza HA. Immunol Rev *250*, 180-198. 10.1111/imr.12005.
Kanai, R., Kar, K., Anthony, K., Gould, L.H., Ledizet, M., Fikrig, E., Marasco, W.A., Koski, R.A., and Modis, Y. (2006). Crystal structure of west nile virus envelope glycoprotein reveals viral surface epitopes. J Virol *80*, 11000-11008. 10.1128/JVI.01735-06.
Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., et al. (2021). PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res *49*, D1388-D1395. 10.1093/nar/gkaa971.
Kinjo, A.R., Bekker, G.J., Wako, H., Endo, S., Tsuchiya, Y., Sato, H., Nishi, H., Kinoshita, K., Suzuki, H., Kawabata, T., et al. (2018). New tools and functions in data-out activities at Protein Data Bank Japan (PDBj). Protein Science *27*, 95-102. 10.1002/pro.3273.
Kirchdoerfer, R.N., Wang, N., Pallesen, J., Wrapp, D., Turner, H.L., Cottrell, C.A., Corbett, K.S., Graham, B.S., McLellan, J.S., and Ward, A.B. (2018). Stabilized coronavirus spikes are

resistant to conformational changes induced by receptor recognition or proteolysis. Sci Rep *8*, 15701. 10.1038/s41598-018-34171-7.
Kong, L., Giang, E., Nieusma, T., Kadam, R.U., Cogburn, K.E., Hua, Y., Dai, X., Stanfield, R.L., Burton, D.R., Ward, A.B., et al. (2013). Hepatitis C virus E2 envelope glycoprotein core structure. Science *342*, 1090-1094. 10.1126/science.1243876.
Kowalzik, F., Schreiner, D., Jensen, C., Teschner, D., Gehring, S., and Zepp, F. (2021). mRNA-Based Vaccines. Vaccines (Basel) *9*. 10.3390/vaccines9040390.
Kubota, M., Takeuchi, K., Watanabe, S., Ohno, S., Matsuoka, R., Kohda, D., Nakakita, S.I., Hiramatsu, H., Suzuki, Y., Nakayama, T., et al. (2016). Trisaccharide containing alpha2,3-linked sialic acid is a receptor for mumps virus. Proc Natl Acad Sci U S A *113*, 11579-11584. 10.1073/pnas.1608383113.
Kuhlbrandt, W. (2014). Biochemistry. The resolution revolution. Science *343*, 1443-1444. 10.1126/science.1251652.
Kumar, A., Hossain, R.A., Yost, S.A., Bu, W., Wang, Y., Dearborn, A.D., Grakoui, A., Cohen, J.I., and Marcotrigiano, J. (2021). Structural insights into hepatitis C virus receptor binding and entry. Nature. 10.1038/s41586-021-03913-5.
Li, W., O'Neill, K.R., Haft, D.H., DiCuccio, M., Chetvernin, V., Badretdin, A., Coulouris, G., Chitsaz, F., Derbyshire, M.K., Durkin, A.S., et al. (2021). RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. Nucleic Acids Res *49*, D1020-D1028. 10.1093/nar/gkaa1105.
Liu, C., Ginn, H.M., Dejnirattisai, W., Supasa, P., Wang, B., Tuekprakhon, A., Nutalai, R., Zhou, D., Mentzer, A.J., Zhao, Y., et al. (2021a). Reduced neutralization of SARS-CoV-2 B.1.617 by vaccine and convalescent serum. Cell. 10.1016/j.cell.2021.06.020.
Liu, Y., Heim, K.P., Che, Y., Chi, X., Qiu, X., Han, S., Dormitzer, P.R., and Yang, X. (2021b). Prefusion structure of human cytomegalovirus glycoprotein B and structural basis for membrane fusion. Sci Adv *7*. 10.1126/sciadv.abf3178.
Lubin, J.H., Zardecki, C., Dolan, E.M., Lu, C., Shen, Z., Dutta, S., Westbrook, J.D., Hudson, B.P., Goodsell, D.S., Williams, J.K., et al. (2020). Evolution of the SARS-CoV-2 proteome in three dimensions (3D) during the first six months of the COVID-19 pandemic. bioRxiv. 10.1101/2020.12.01.406637.
Magro, M., Mas, V., Chappell, K., Vazquez, M., Cano, O., Luque, D., Terron, M.C., Melero, J.A., and Palomo, C. (2012). Neutralizing antibodies against the preactive form of respiratory syncytial virus fusion protein offer unique possibilities for clinical intervention. Proc Natl Acad Sci U S A *109*, 3089-3094. 10.1073/pnas.1115941109.
Markosian, C., Di Costanzo, L., Sekharan, M., Shao, C., Burley, S.K., and Zardecki, C. (2018). Analysis of impact metrics for the Protein Data Bank. Sci Data *5*, 180212. 10.1038/sdata.2018.212.
McLellan, J.S., Chen, M., Joyce, M.G., Sastry, M., Stewart-Jones, G.B., Yang, Y., Zhang, B., Chen, L., Srivatsan, S., Zheng, A., et al. (2013a). Structure-based design of a fusion glycoprotein vaccine for respiratory syncytial virus. Science *342*, 592-598. 10.1126/science.1243283.
McLellan, J.S., Chen, M., Leung, S., Graepel, K.W., Du, X., Yang, Y., Zhou, T., Baxa, U., Yasuda, E., Beaumont, T., et al. (2013b). Structure of RSV fusion glycoprotein trimer bound to a prefusion-specific neutralizing antibody. Science *340*, 1113-1117. 10.1126/science.1234914.
McLellan, J.S., Yang, Y., Graham, B.S., and Kwong, P.D. (2011). Structure of respiratory syncytial virus fusion glycoprotein in the postfusion conformation reveals preservation of neutralizing epitopes. J Virol *85*, 7788-7796. 10.1128/JVI.00555-11.
McNaught, A.D. (1996). International Union of Pure and Applied Chemistry and International Union of Biochemistry and Molecular Biology - Joint Commission on Biochemical Nomenclature - Nomenclature of carbohydrates - Recommendations 1996. Pure and Applied Chemistry *68*, 1919-2008. 10.1351/pac199668101919.

Miao, L., Zhang, Y., and Huang, L. (2021). mRNA vaccine for cancer immunotherapy. Mol Cancer *20*, 41. 10.1186/s12943-021-01335-5.
Nguyen, D.T., Mathias, S., Bologa, C., Brunak, S., Fernandez, N., Gaulton, A., Hersey, A., Holmes, J., Jensen, L.J., Karlsson, A., et al. (2017). Pharos: Collating protein information to shed light on the druggable genome. Nucleic Acids Res *45*, D995-D1002. 10.1093/nar/gkw1072.
Ostrem, J.M., Peters, U., Sos, M.L., Wells, J.A., and Shokat, K.M. (2013). K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. Nature *503*, 548-551. 10.1038/nature12796.
Pallesen, J., Wang, N., Corbett, K.S., Wrapp, D., Kirchdoerfer, R.N., Turner, H.L., Cottrell, C.A., Becker, M.M., Wang, L., Shi, W., et al. (2017). Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. Proc Natl Acad Sci U S A *114*, E7348-E7357. 10.1073/pnas.1707304114.
Pardi, N., Hogan, M.J., Porter, F.W., and Weissman, D. (2018). mRNA vaccines - a new era in vaccinology. Nat Rev Drug Discov *17*, 261-279. 10.1038/nrd.2017.243.
Park, J.W., Lagniton, P.N.P., Liu, Y., and Xu, R.H. (2021). mRNA vaccines for COVID-19: what, why and how. Int J Biol Sci *17*, 1446-1460. 10.7150/ijbs.59233.
Pollard, A.J., and Bijker, E.M. (2021). A guide to vaccinology: from basic principles to new developments. Nat Rev Immunol *21*, 83-100. 10.1038/s41577-020-00479-7.
Protein Data Bank (1971). Crystallography: Protein Data Bank. Nature (London), New Biol. *233*, 223-223. 10.1038/newbio233223b0.
Rappuoli, R., De Gregorio, E., Del Giudice, G., Phogat, S., Pecetta, S., Pizza, M., and Hanon, E. (2021). Vaccinology in the post-COVID-19 era. Proc Natl Acad Sci U S A *118*. 10.1073/pnas.2020368118.
Read, K.B., Sheehan, J.R., Huerta, M.F., Knecht, L.S., Mork, J.G., Humphreys, B.L., and N.I.H. Big Data Annotator Group (2015). Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study. PLoS One *10*, e0132735. 10.1371/journal.pone.0132735.
Rey, F.A., Heinz, F.X., Mandl, C., Kunz, C., and Harrison, S.C. (1995). The envelope glycoprotein from tick-borne encephalitis virus at 2 A resolution. Nature *375*, 291-298. 10.1038/375291a0.
Roche, S., Rey, F.A., Gaudin, Y., and Bressanelli, S. (2007). Structure of the prefusion form of the vesicular stomatitis virus glycoprotein G. Science *315*, 843-848. 10.1126/science.1135710.
Romero, P.R., Kobayashi, N., Wedell, J.R., Baskaran, K., Iwata, T., Yokochi, M., Maziuk, D., Yao, H., Fujiwara, T., Kurusu, G., et al. (2020). BioMagResBank (BMRB) as a Resource for Structural Biology. Methods Mol Biol *2112*, 187-218. 10.1007/978-1-0716-0270-6_14.
Rose, A.S., Bradley, A.R., Valasatava, Y., Duarte, J.M., Prlić, A., and Rose, P.W. (2018). NGL viewer: web-based molecular graphics for large complexes. Bioinformatics *34*, 3755–3758.
Sander, C., and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins : Structure, Function, and Genetics *9*, 56-68. 10.1002/prot.340090107.
Sehnal, D., Bittrich, S., Deshpande, M., Svobodova, R., Berka, K., Bazgier, V., Velankar, S., Burley, S.K., Koca, J., and Rose, A.S. (2021). Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Res *49*, W431–W437. 10.1093/nar/gkab314.
Sehnal, D., Deshpande, M., Varekova, R.S., Mir, S., Berka, K., Midlik, A., Pravda, L., Velankar, S., and Koca, J. (2017). LiteMol suite: interactive web-based visualization of large-scale macromolecular structure data. Nat. Methods *14*, 1121-1122. 10.1038/nmeth.4499.
Shao, C., Feng, Z., Westbrook, J.D., Peisach, E., Berrisford, J., Ikegawa, Y., Kurisu, G., Velankar, S., Burley, S.K., and Young, J.Y. (2021). Modernized Uniform Representation of Carbohydrate Molecules in the Protein Data Bank. Glycobiology. 10.1093/glycob/cwab039.

Sillitoe, I., Bordin, N., Dawson, N., Waman, V.P., Ashford, P., Scholes, H.M., Pang, C.S.M., Woodridge, L., Rauer, C., Sen, N., et al. (2021). CATH: increased structural coverage of functional space. Nucleic Acids Res *49*, D266-D273. 10.1093/nar/gkaa1079.
Sirohi, D., Chen, Z., Sun, L., Klose, T., Pierson, T.C., Rossmann, M.G., and Kuhn, R.J. (2016). The 3.8 A resolution cryo-EM structure of Zika virus. Science *352*, 467-470. 10.1126/science.aaf5316.
Sollner, T.H. (2004). Intracellular and viral membrane fusion: a uniting mechanism. Curr Opin Cell Biol *16*, 429-435. 10.1016/j.ceb.2004.06.015.
Stampfer, S.D., Lou, H., Cohen, G.H., Eisenberg, R.J., and Heldwein, E.E. (2010). Structural basis of local, pH-dependent conformational changes in glycoprotein B from herpes simplex virus type 1. J Virol *84*, 12924-12933. 10.1128/JVI.01750-10.
Tiemeyer, M., Aoki, K., Paulson, J., Cummings, R.D., York, W.S., Karlsson, N.G., Lisacek, F., Packer, N.H., Campbell, M.P., Aoki, N.P., et al. (2017). GlyTouCan: an accessible glycan structure repository. Glycobiology *27*, 915-919. 10.1093/glycob/cwx066.
UniProt, C. (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res *49*, D480-D489. 10.1093/nar/gkaa1100.
van der Aalst, W.M.P., Bichler, M., and Heinzl, A. (2017). Responsible Data Science. Business & Information Systems Engineering *59*, 311-313. 10.1007/s12599-017-0487-z.
Varghese, J.N., and Colman, P.M. (1991). Three-dimensional structure of the neuraminidase of influenza virus A/Tokyo/3/67 at 2.2 A resolution. J Mol Biol *221*, 473-486. 10.1016/0022-2836(91)80068-6.
Varki, A., Cummings, R.D., Aebi, M., Packer, N.H., Seeberger, P.H., Esko, J.D., Stanley, P., Hart, G., Darvill, A., Kinoshita, T., et al. (2015). Symbol Nomenclature for Graphical Representations of Glycans. Glycobiology *25*, 1323-1324. 10.1093/glycob/cwv091.
Wadhwa, A., Aljabbari, A., Lokras, A., Foged, C., and Thakur, A. (2020). Opportunities and Challenges in the Delivery of mRNA-based Vaccines. Pharmaceutics *12*. 10.3390/pharmaceutics12020102.
Walls, A.C., Park, Y.J., Tortorici, M.A., Wall, A., McGuire, A.T., and Veesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell *181*, 281-292 e286. 10.1016/j.cell.2020.02.058.
Westbrook, J.D., and Burley, S.K. (2019). How Structural Biologists and the Protein Data Bank Contributed to Recent FDA New Drug Approvals. Structure *27*, 211-217. 10.1016/j.str.2018.11.007.
Westbrook, J.D., Soskind, R., Hudson, B.P., and Burley, S.K. (2020). Impact of Protein Data Bank on Anti-neoplastic Approvals. Drug Discov Today *25*, 837-850 10.1016/j.drudis.2020.02.002.
Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data *3*, 1-9. 10.1038/sdata.2016.18.
Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res *46*, D1074-D1082. 10.1093/nar/gkx1037.
Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.L., Abiona, O., Graham, B.S., and McLellan, J.S. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science *367*, 1260-1263. 10.1126/science.abb2507.
wwPDB consortium (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res *47*, D520-D528. 10.1093/nar/gky949.
Yamada, I., Shiota, M., Shinmachi, D., Ono, T., Tsuchiya, S., Hosoda, M., Fujita, A., Aoki, N.P., Watanabe, Y., Fujita, N., et al. (2020). The GlyCosmos Portal: a unified and comprehensive web resource for the glycosciences. Nat Methods *17*, 649-650. 10.1038/s41592-020-0879-8.

Yang, F., Lin, S., Ye, F., Yang, J., Qi, J., Chen, Z., Lin, X., Wang, J., Yue, D., Cheng, Y., et al. (2020). Structural Analysis of Rabies Virus Glycoprotein Reveals pH-Dependent Conformational Changes and Interactions with a Neutralizing Antibody. Cell Host Microbe *27*, 441-453 e447. 10.1016/j.chom.2019.12.012.
York, W.S., Mazumder, R., Ranzinger, R., Edwards, N., Kahsay, R., Aoki-Kinoshita, K.F., Campbell, M.P., Cummings, R.D., Feizi, T., Martin, M., et al. (2020). GlyGen: Computational and Informatics Resources for Glycoscience. Glycobiology *30*, 72-73. 10.1093/glycob/cwz080.
Yuan, Y., Cao, D., Zhang, Y., Ma, J., Qi, J., Wang, Q., Lu, G., Wu, Y., Yan, J., Shi, Y., et al. (2017). Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. Nature communications *8*, 15092. 10.1038/ncomms15092.
Yurkovetskiy, L., Wang, X., Pascal, K.E., Tomkins-Tinch, C., Nyalile, T.P., Wang, Y., Baum, A., Diehl, W.E., Dauphin, A., Carbone, C., et al. (2020). Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. Cell *183*, 739-751 e738. 10.1016/j.cell.2020.09.032.
Zhang, Y., Zhang, W., Ogata, S., Clements, D., Strauss, J.H., Baker, T.S., Kuhn, R.J., and Rossmann, M.G. (2004). Conformational changes of the flavivirus E glycoprotein. Structure *12*, 1607-1618. 10.1016/j.str.2004.06.019.