Received: 7 October 2021 | Revised: 7 November 2021 | Accepted: 9 November 2021

DOI: 10.1002/pro.4238

TOOLS FOR PROTEIN SCIENCE

PDBsum extras: SARS-CoV-2 and AlphaFold models

Roman A. Laskowski | **Janet M. Thornton**

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge

Correspondence
Roman A. Laskowski, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
Email: roman@ebi.ac.uk

Abstract

The PDBsum web server provides structural analyses of the entries in the Protein Data Bank (PDB). Two recent additions are described here. The first is the detailed analysis of the SARS-CoV-2 virus protein structures in the PDB. These include the variants of concern, which are shown both on the sequences and 3D structures of the proteins. The second addition is the inclusion of the available AlphaFold models for human proteins. The pages allow a search of the protein against existing structures in the PDB via the Sequence Annotated by Structure (SAS) server, so one can easily compare the predicted model against experimentally determined structures. The server is freely accessible to all at http://www.ebi.ac.uk/pdbsum.

KEYWORDS

3D protein structure, AlphaFold, PDB, PDBsum, protein database, protein structure analysis, SARS-CoV-2, schematic diagrams

1 | INTRODUCTION

The PDBsum web server was developed at University College London (UCL) in 1995[1] and moved to the EMBL-EBI in 2001 where it now resides. It provides a largely pictorial compendium of the proteins and their complexes in the Protein Data Bank (PDB),[2] with analyses of protein secondary structure, schematic diagrams for protein–ligand, protein–DNA, and protein–protein interactions, PROCHECK analyses of structural quality, and many others.[3–7] The molecules and their interactions can be viewed in 3D using the molecular viewers RasMol,[8] Jmol,[9] PyMOL,[10] Strap,[11] and the JavaScript viewer 3Dmol.js.[12] Users can upload their own PDB files to receive a full PDBsum analysis. The two most recent additions to the server are described here: pages listing and analyzing the proteins of the SARS-CoV-2, and pages for each of the available AlphaFold models for human proteins.

2 | SARS-COV-2 PROTEINS

Since the outbreak of COVID-19 (coronavirus disease 2019) in China in December 2019,[13] and its subsequent promotion to a global pandemic on March 11, 2020, experimental scientists around the world have been solving the structures of the virus's constituent proteins. These consist of four structural proteins—spike, membrane, envelope, and nucleocapsid—16 nonstructural proteins forming its replicase/transcriptase complex, and nine putative accessory factors.[14] It has been a remarkable effort with 1,540 structures deposited in the PDB as of October 1, 2021. Figure 1 shows how these structures are distributed across the proteins. Not surprisingly most of the effort has focused on the spike protein which the virus uses to enter and infect the cells of the host organism. Also highly targeted have been NSP5 (nonstructural protein 5), which is the virus's main protease and is crucial for its replication and transcription, and

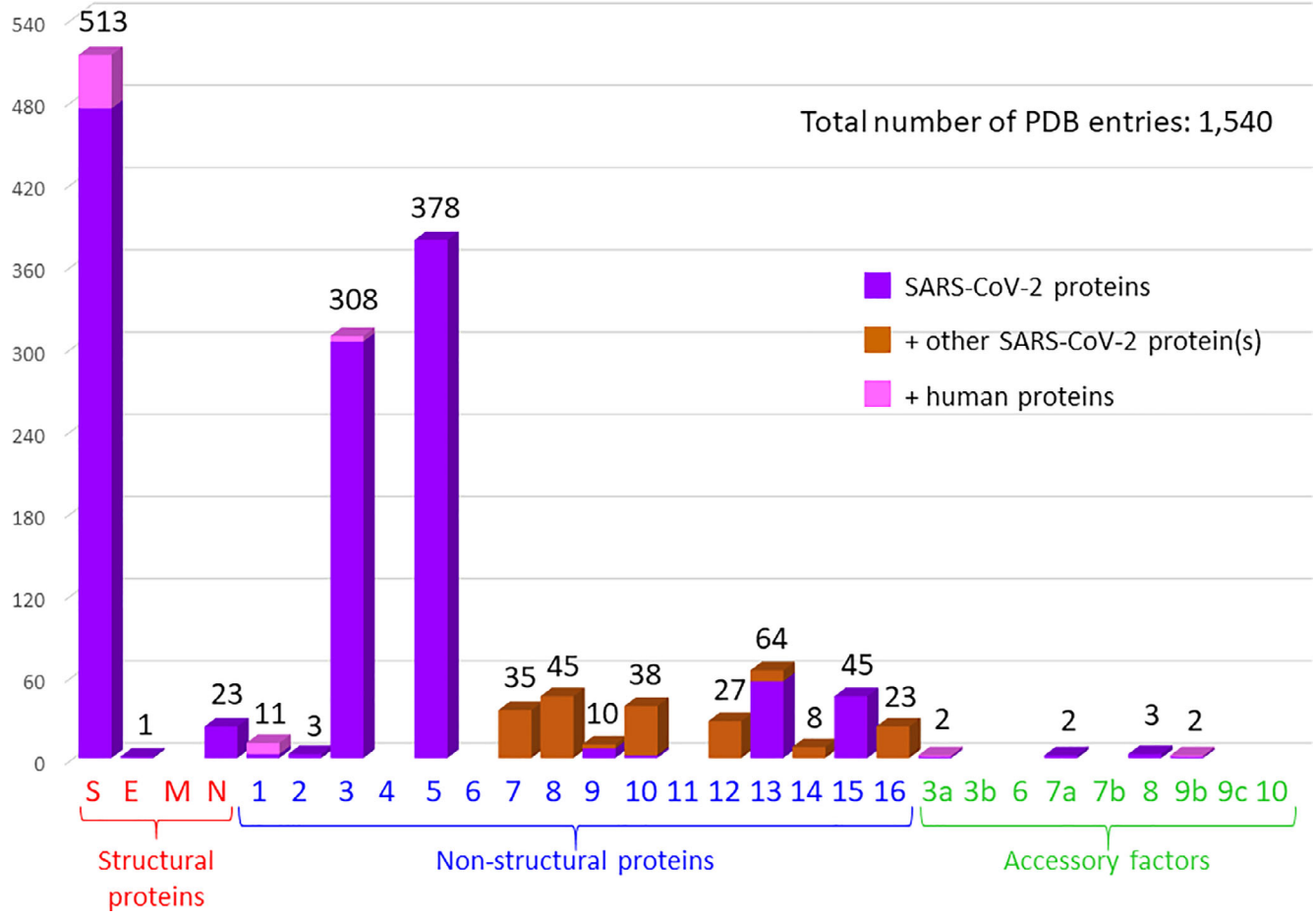

FIGURE 1 Histogram showing the numbers of PDB entries for each SARS-CoV-2 protein as of October 1, 2021. The 4 structural proteins—spike (S), envelope (E), membrane (M), and nucleocapsid (N)—are labeled on the *x*-axis in red, the 16 nonstructural proteins (Nsp1-16) are labeled in blue, and the 9 putative accessory factors in green. The bars in the plot are colored purple to represent structures which contain only that protein, brown if the structures also contain another SARS-CoV-2 protein, and magenta where the structures are complexed with a human protein. The numbers at top of the bars give the count of PDB entries containing the given protein. Note, these numbers do not sum to 1,540 because some of the PDB entries contain two or more SARS-CoV-2 proteins in complex

NSP3 which is involved with NSP4 and NSP6 in viral replication.

The main SARS-CoV-2 page in PDBsum can be accessed either via the home page, https://www.ebi.ac.uk/pdbsum, or directly via https://www.ebi.ac.uk/thornton-srv/databases/pdbsum/covid-19.html. It lists all the SARS-CoV-2 proteins and their corresponding PDB entries. Figure 2a shows the first few entries for the main protease, NSP5, together with various links to more detailed analyses. Each PDB entry in the list is represented by a purple schematic diagram of the protein's secondary structure which indicates the coverage provided by the structure—for example, in the case of the larger proteins, the PDB entry may only be of one or two domains. Also listed are any ligands or drug molecules bound to the protein. The structure can be displayed in 3D using one of four molecular viewers, either via the icons on the left, below the thumbnail image, which display the top structure in the list, or by clicking on the eye-icon next to any PDB entry in the list. The viewers are 3Dmol.js, JSmol, RasMol, and PyMOL. The first two are JavaScript-based so do not require any preinstalled software, while the last two are programs that are freely available to download and install.

2.1 | Ligand analyses

Many of the proteins have ligands bound, and there are several pages devoted to these. The first is the ligand clusters page, accessed by clicking on the green and blue ligand icon below the downloadable list of structures (Figure 2a). This highlights any binding sites and how different molecules bind within them. For each protein, one PDB entry is selected as the reference. Its sequence will be identical to the relevant protein (i.e., no variants) and have the highest resolution and

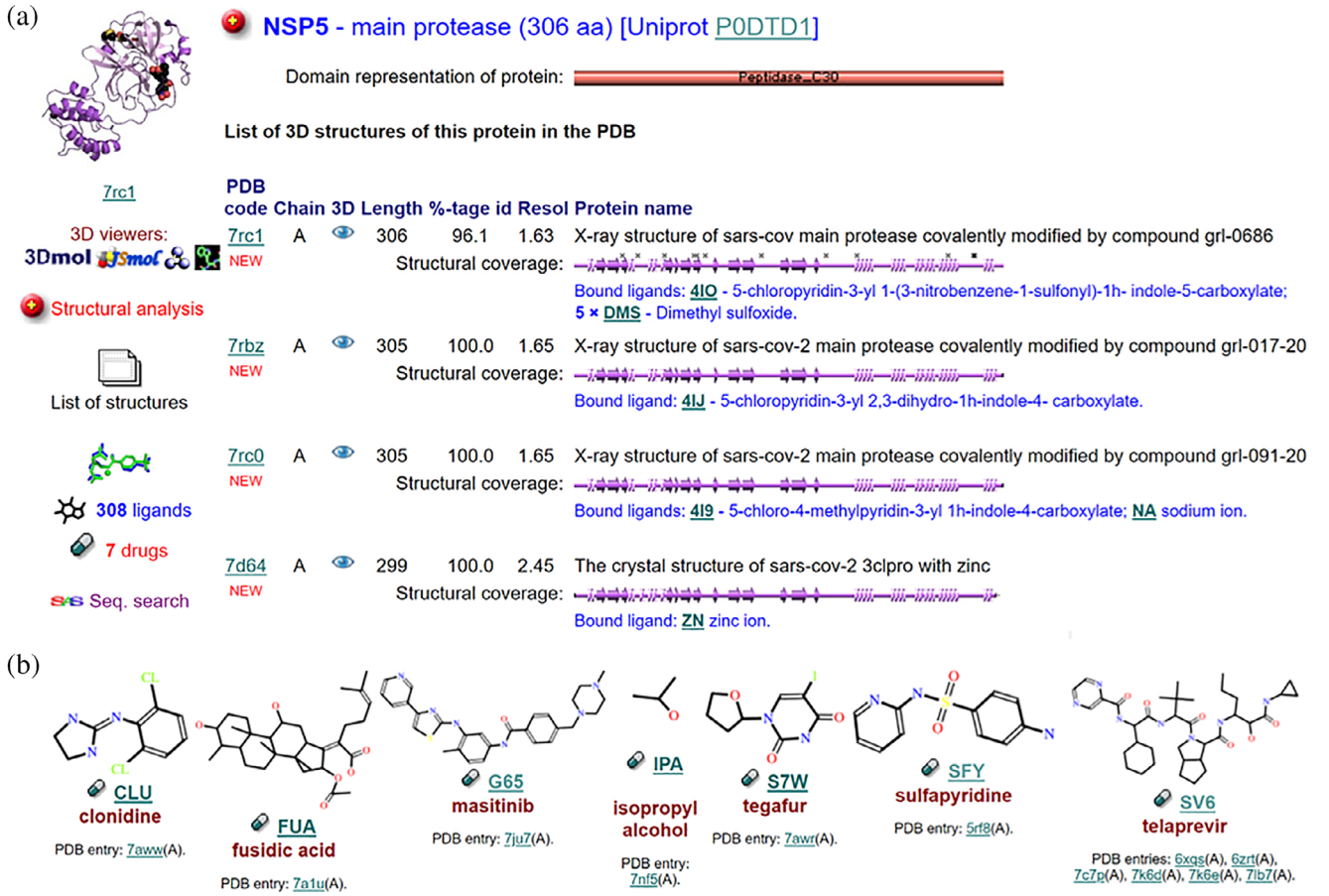

FIGURE 2 (a) An extract from the PDBsum SARS-CoV-2 page showing the virus's main protease, NSP5, and its first four PDB entries, out of 378 (as accessed on October 1, 2021). The brown cylinder at the top illustrates the domain organization which, in this case, is just a single domain. For each of the four listed PDB structures, the purple schematic diagram of the protein's secondary structure indicates the PDB entry's structural coverage of the protein—in this case all four cover the entire protein. The most recent structures (i.e., released in the current week) are marked NEW and listed first; subsequent structures are ordered by their percentage coverage of the protein and their resolution. The first structure here, 7rc1, has only 96.1% sequence identity to the viral protein, the altered amino acid positions being denoted by the small black crosses on the diagram. The thumbnail image in the top left corner is of this first entry and can be viewed in 3D using any of the molecular viewers shown by the icons below it: 3Dmol.js, JSmol, RasMol, and PyMOL. The red sphere and cross icon leads to a full structural analysis of the protein (see Figure 3). The icons below it lead to a downloadable listing of the structures, a page that displays the ligand clusters (i.e., ligands from all the protein's PDB entries superposed to reveal common binding patterns), a page listing all the bound ligands, and a list (shown in (b)) of the bound drug molecules and the PDB entries in which they are found. The final icon is to SAS, which searches the protein's sequence against all structures in the PDB

best *R*-factor. All other structures of the protein are then superposed onto this reference, bringing any bound ligands with them. The net result is a superposition of the binding sites and the molecules bound to them, which appear as clusters of ligands on the surface of the reference structure. The separate ligand clusters are color-coded and the ligands in each cluster are listed with links to the PDB entries in which they are found. JSmol and RasMol views can show either individual ligand clusters or the reference structure with all clusters superposed.

An alternative way of obtaining all the ligands that bind to a given protein is via the black molecule icon just below the ligand clusters one. This simply lists all the ligands found in the PDB entries of the protein, with any drug molecules marked. A separate listing of just the drug molecules, if any, can be seen by clicking on the green pill icon in Figure 2a, which gives the listing shown in Figure 2b.

2.2 | Structural analysis and variants

The red sphere icon, with yellow cross, shown in Figure 2a leads to a more detailed analysis of the protein's structure and its variants, including any variants of concern. The main part of this analysis is shown in

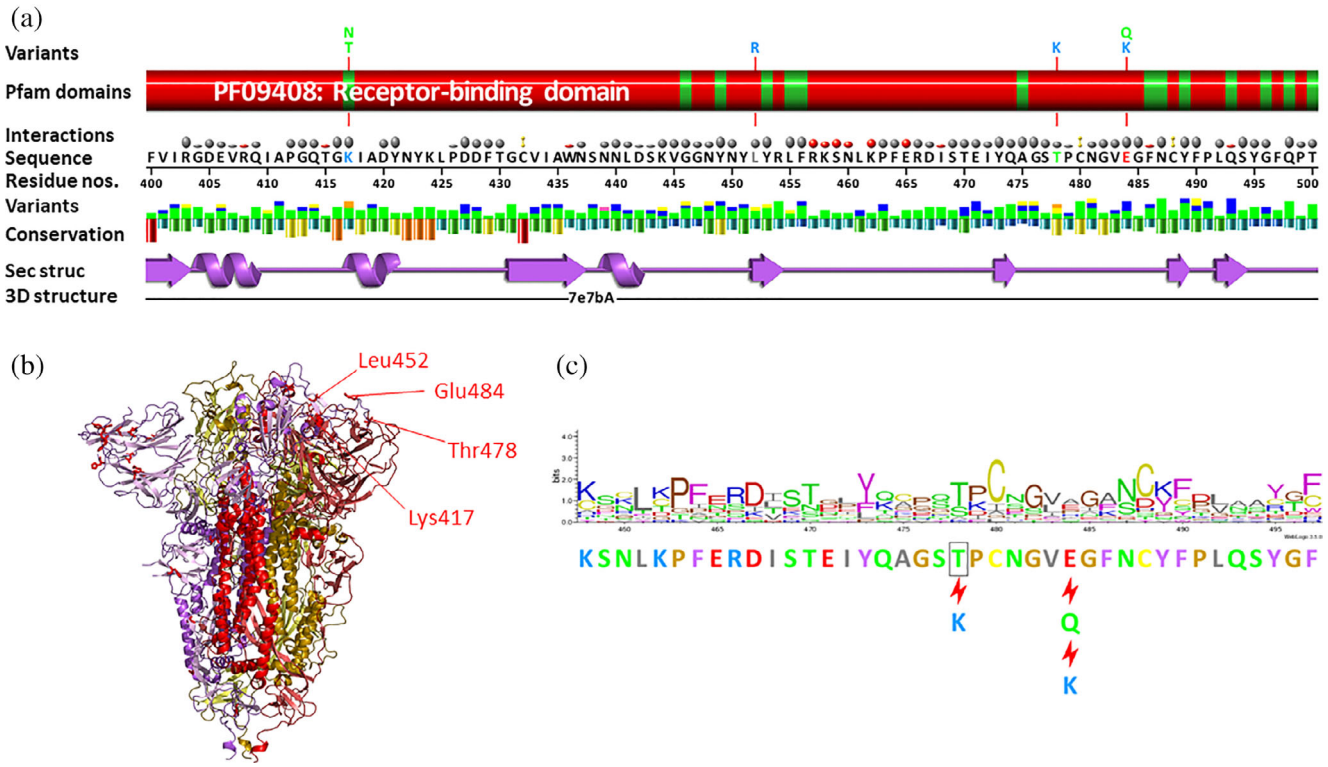

FIGURE 3 (a) A schematic diagram of part of the SARS-CoV-2 spike protein (residues 400–500) with various sequence and structural annotations. The red cylinder at the top is part of the receptor-binding domain which binds to the host's ACE-2 receptor as the first step in its entry into the host cell. The green bands identify the residues making up the receptor binding site. The variants of concern—Lys417Thr, Lys417Asn, Leu452Arg, Thr478Lys, Glu484Gln, and Gly484Lys—are shown above the domain diagram and are identified by their single-letter amino acid code, colored by residue type (blue, positive; red, negative; green, neutral). Above the protein's sequence is a line of colored blobs of different sizes corresponding to the types and numbers of intermolecular interactions observed in the 3D structures. The larger the blob, the more structures the interaction occurs in: red for interactions with ligand, and gray for protein–protein interactions. The "best" PDB entry (PDB code 7e7b, chain A), in terms of closest sequence identity and best structural quality, is shown schematically in the purple "wiring diagram", where beta strands are represented by arrows, and alpha helices by coils. Above it is a graph of residue conservation, computed from a sequence alignment obtained from a BLAST search against the UniProt database. The bars are colored from red for highly conserved to purple for highly variable. The histogram above the conservation plot shows all the variants observed to date, as obtained from the CoV-GLUE database. The colors indicate how many sequences the variant has been observed in: green, fewer than 10 sequences, to red, more than 100,000 sequences. Clicking on the various annotations gives a pop-up window with further information. (b) The structure of the spike protein trimer (PDB entry 7e7b) with the residues having variants of concern in (a) labeled in red. The top part of the structure binds to the ACE 2 receptor, while the lower part is embedded in the virus's outer membrane. (c) Sequence logo showing the alternative residues found at each position, obtained from sequence alignments of related proteins in UniProt. The plot is centered on Thr478 and obtained by clicking on this residue's conservation bar in (a). The amino acid codes are colored by residue property: blue = positive (H,K,R), red = negative (D,E), green = neutral (S,T,N,Q), gray = aliphatic (A,V,L,I,M), purple = aromatic (F,Y,W), brown = Pro&Gly (P,G), yellow = cysteine (C). The variants of concern are indicated by the red lightning bolts

Figure 3a and is based on the schematic diagrams of VarSite[15] which focuses on disease-associated variants in human proteins. Figure 3a shows part of the virus's spike protein (residues 400–500). This segment forms part of the protein's receptor binding domain (red cylinder), which binds to the host's ACE-2 receptors and initiates the virus's entry into the cell and infection of it. The green stripes indicate residues that form the binding site. Above the domain are shown six variants of concern. Three of these belong to the Delta variant—Leu452Arg, Thr478Lys, and Glu484Gln. Although none of these variants are in the binding site, they are nevertheless close to it.

The other schematics in Figure 3a provide additional information. At the bottom is a purple diagram of the protein's secondary structure. Clicking on it displays the protein structure in the 3Dmol.js viewer with all the variants of concern labeled. This annotated structure can also then be viewed in either RasMol or PyMOL. Figure 3b shows a PyMOL image of the spike protein (which is a trimer), with just the variants of Figure 3a marked on its first chain.

Above the sequence in Figure 3a is a line of colored blobs representing contacts to other molecules: red for ligands and gray for other proteins. The size of the blobs indicates the proportion of PDB entries that exhibit the interaction. Two histograms below the sequence, one pointing up and the other down, give information on natural variants and on residue conservation, respectively. The former come from the CoV-GLUE website (http://cov-glue-viz.cvr.gla.ac.uk)[16] which maintains a database of mutations, insertions, and deletions observed in the virus's protein sequences as deposited in GISAID (Global Initiative on Sharing All Influenza Data).[17] The colors of the bars indicate how many sequences the variant has been observed in: green corresponds to fewer than 10 sequences, then blue, yellow, orange, pink, and red which signifies the variant has been observed in more than 100,000 sequences. The lower histogram gives the residue conservation which is computed from a sequence alignment obtained from a BLAST[18] search of the sequence against UniProt.[19] The conservation score is computed using the ScoreCons[20] method. Clicking on any of the bars gives a sequence logo,[21] like that shown in Figure 3c, which displays the alternative residues found at each position in the BLAST output.

3 | AlphaFold MODELS

At the end of 2020, DeepMind, a London-based AI company now part of Google's parent company, Alphabet Inc., announced that their AlphaFold 2 system[22] had significantly outperformed all other methods in the biennial Critical Assessment of protein Structure Prediction (CASP).[23] The models it produced were of a quality approaching that of experimental determination. In mid-2021, they released their source code and made public almost 350,000 protein models from various species, including human.[24] This resulted in much excitement in the structural biology community. A website providing access to individual models, and to downloads of all models by species, was set up by a collaboration between DeepMind and EMBL-EBI.[25]

One of the benefits of having these models is that the hypothetical 3D structures of many proteins for which no structural information is available—not even from very distant homologs—can now form the basis of functional studies.[26]

The available AlphaFold models for all human proteins have been added to PDBsum—23,391 models in all, corresponding to 20,504 proteins. There are more models than proteins because very long proteins, longer than 2,500 residues, are represented by several models, each 1,400 residues long and progressively shifted by 200 residues along the sequence. See, for example, PDBsum entry A005.

Accommodating the new models into PDBsum required generating a pseudo PDB identifier of four character for each, starting at A001 and incrementing via numbers and lower- and upper-case letters to A65h. Each model is accessible via its UniProt accession, as well as this pseudo PDB code.

Figure 4 shows the PDBsum page for the AlphaFold model of the human enzyme ATP-dependent 6-phosphofructokinase, muscle type (UniProt accession P08237, pseudo PDB code A13M). All the standard PDBsum analyses are given apart from those that are not relevant—that is, those involving interactions with other molecules as all the models are of a single protein chain with no bound ligands, DNA/RNA, or other protein chains.

The most obvious difference between the page shown in Figure 4 and a typical PDBsum page is in the coloring of the protein's 3D structure and secondary structure schematic diagrams. Instead of being colored by chain, the protein is colored using the same scheme as in the AlphaFold website, namely by the confidence of the prediction. This is quantified by the pLDDT score (predicted local distance difference test). Regions with very high confidence (pLDDT > 90) are colored dark blue, confidently predicted regions (90 > pLDDT > 70) are light blue, regions of low confidence (70 > pLDDT > 50) are yellow, and very low confidence regions (pLDDT < 50) are orange. The orange regions can, essentially, be ignored. They are often seen as elongated strings floating aimlessly away from the rest of the protein and are unlikely to have any basis in reality. A number of the very poor models resemble balls of this orange spaghetti and do not look like viable proteins at all (e.g., A26x, which is the model for human semenogelin-2, UniProt accession Q02383).

One particular benefit of having the models in PDBsum is that one can easily compare them with existing, experimentally determined structures in the PDB. Clicking the SAS[27] icon will run SAS (Sequence annotated by Structure) to search the protein's sequence against all the sequences in the PDB, including the human AlphaFold models. From the resultant alignment one can tell if the AlphaFold model is a novel one, or likely based on one or more existing PDB structures. Furthermore, one can superpose the AlphaFold model on one or more of the matched PDB entries and view in either RasMol or JSmol to see how well the model matches the experimental structures.

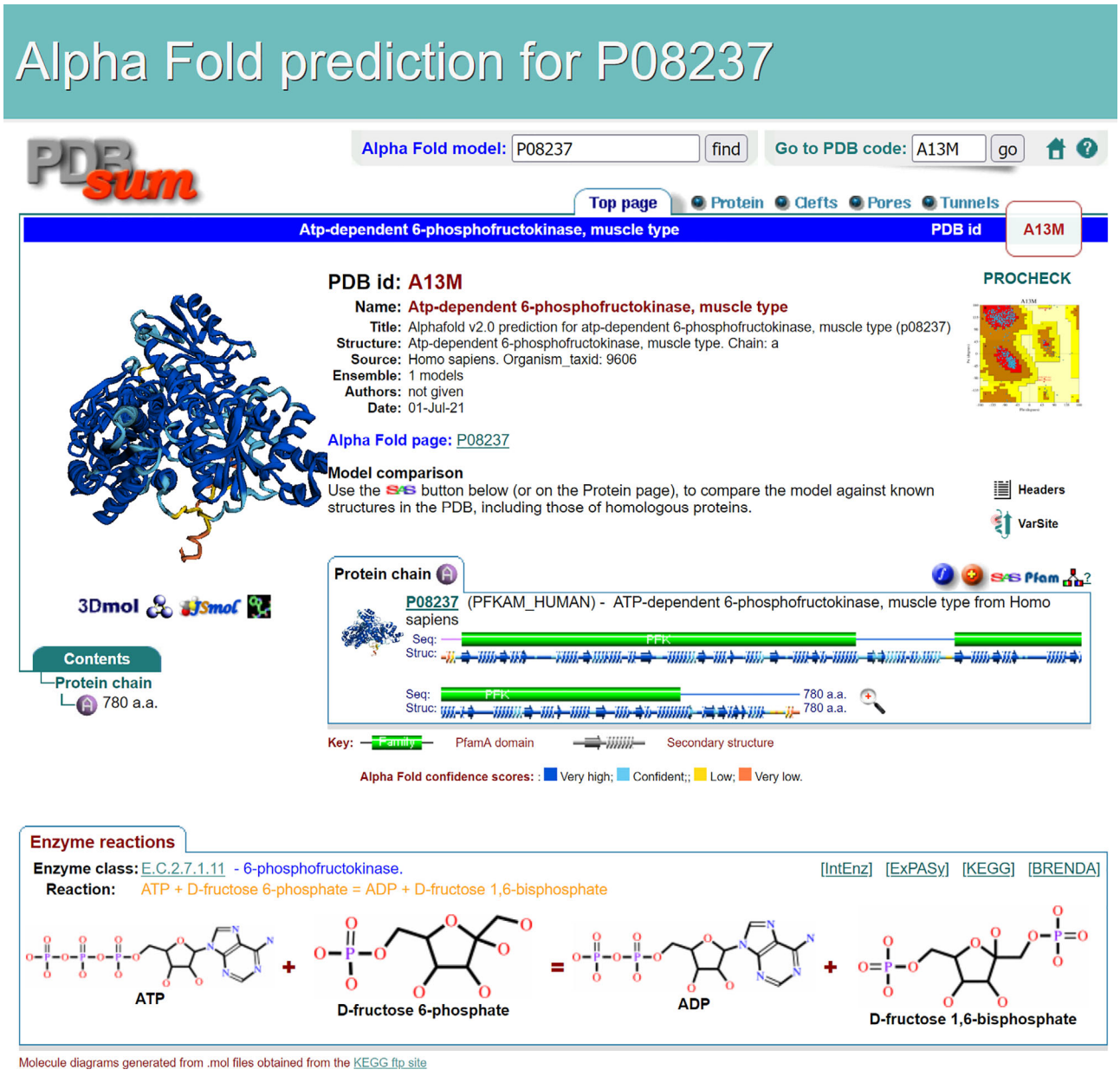

FIGURE 4 PDBsum page for the AlphaFold model of the human enzyme ATP-dependent 6-phosphofructokinase, muscle type (UniProt accession P08237). The mail difference here from a standard PDBsum page is that the thumbnail image and schematic diagrams of the protein's secondary structure are colored according to the pLDDT score: dark blue for regions predicted with very high confidence (pLDDT > 90), light blue for confidently predicted regions (90 > pLDDT > 70), yellow for regions of low confidence (70 > pLDDT >50), and orange for the regions of very low confidence (pLDDT < 50)

This will be clearer when the human AlphaFold models are added to our VarSite[15] database which is closely linked to PDBsum. VarSite maps disease-associated variants in human genes to the corresponding protein sequences and structures (https://www.ebi.ac.uk/thornton-srv/databases/VarSite). A link on each PDBsum AlphaFold page already links to the associated VarSite page. This currently shows the structural coverage of each protein by existing PDB entries, but, by the time this paper is published, will also show the AlphaFold model coverage and hence which parts of the protein are only available in the latter model.

ACKNOWLEDGMENT

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare that there is no potential conflict of interest.

AUTHOR CONTRIBUTIONS

Roman A. Laskowski: Software (lead). **Janet Thornton:** Funding acquisition (lead); supervision (lead).

DATA AVAILABILITY STATEMENT

The PDBsum server is freely available at http://www.ebi.ac.uk/pdbsum.

ORCID

Roman A. Laskowski https://orcid.org/0000-0001-5528-0087
Janet M. Thornton https://orcid.org/0000-0003-0824-4096

REFERENCES

1. Laskowski RA, Hutchinson EG, Michie AD, Wallace AC, Jones ML, Thornton JM. PDBsum: A web-based database of summaries and analyses of all PDB structures. Trends Biochem Sci. 1997;22:488–490.
2. wwPDB consortium. Protein data bank: The single global archive for 3d macromolecular structure data. Nucleic Acids Res. 2019;47:D520–D528.
3. Laskowski RA. PDBsum: Summaries and analyses of PDB structures. Nucleic Acids Res. 2001;29:221–222.
4. Laskowski RA, Chistyakov VV, Thornton JM. PDBsum more: New summaries and analyses of the known 3d structures of proteins and nucleic acids. Nucleic Acids Res. 2005;33:D266–D268.
5. Laskowski RA. PDBsum new things. Nucleic Acids Res. 2009;37:D355–D359.
6. de Beer TA, Berka K, Thornton JM, Laskowski RA. PDBsum additions. Nucleic Acids Res. 2014;42:D292–D296.
7. Laskowski RA, Jablonska J, Pravda L, Varekova RS, Thornton JM. PDBsum: Structural summaries of PDB entries. Protein Sci. 2018;27:129–134.
8. Sayle RA, Milner-White EJ. RasMol: Biomolecular graphics for all. Trends Biochem Sci. 1995;20:374–376.
9. Herraez A. Biomolecules in the computer: Jmol to the rescue. Biochem Mol Biol Educ. 2006;34:255–261.
10. DeLano WL. The PyMOL molecular graphics system. Palo Alto, CA: DeLano Scientific, 2002.
11. Gille C, Frommel C. Strap: Editor for structural alignments of proteins. Bioinformatics. 2001;17:377–378.
12. Rego N, Koes D. 3Dmol.Js: Molecular visualization with webgl. Bioinformatics. 2015;31:1322–1324.
13. Listings of WHO's response to Covid-19. 2020. https://www.who.int/news/item/29-06-2020-covidtimeline.
14. Gordon DE, Jang GM, Bouhaddou M, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature. 2020;583:459–468.
15. Laskowski RA, Stephenson JD, Sillitoe I, Orengo CA, Thornton JM. VarSite: Disease variants and protein structure. Protein Sci. 2020;29:111–119.
16. Singer J, Gifford R, Cotten M, Robertson D. CoV-GLUE: A web application for tracking SARS-CoV-2 genomic variation. Preprints, 2020060225 (doi: 10.20944/preprints202006.0225.v1). 2020.
17. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data—From vision to reality. Euro Surveill. 2017;2017(22):30494.
18. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and Psi-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–3402.
19. The UniProt Consortium. UniProt: The universal protein knowledgebase in 2021. Nucleic Acids Res. 2020;49:D480–D489.
20. Valdar WS. Scoring residue conservation. Proteins. 2002;48:227–241.
21. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator. Genome Res. 2004;14:1188–1190.
22. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–589.
23. CASP14. 2021. https://predictioncenter.org/casp14/.
24. Tunyasuvunakool K, Adler J, Wu Z, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021;596:590–596.
25. Alphafold protein structure database. 2021. https://www.alphafold.ebi.ac.uk.
26. Akdel M, Pires DEV, Porta Pardo E, et al. A structural biology community assessment of AlphaFold 2 applications. bioRxiv. 2021, doi: 10.1101/2021.09.26.461876
27. Milburn D, Laskowski RA, Thornton JM. Sequences annotated by structure: A tool to facilitate the use of structural information in sequence analysis. Protein Eng. 1998;1998(11):855–859.

How to cite this article: Laskowski RA, Thornton JM. PDBsum extras: SARS-CoV-2 and AlphaFold models. Protein Science. 2022;31:283–9. https://doi.org/10.1002/pro.4238