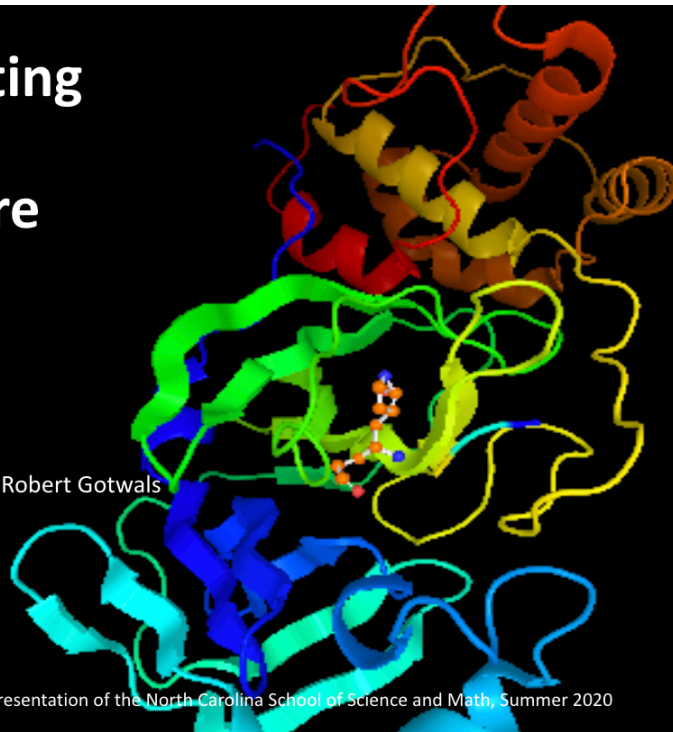


# Computing Protein Structure

By Em Ambrosius and Robert Gotwals

Computing CoVID-19: A Presentation of the North Carolina School of Science and Math, Summer 2020



COMPUTING CoVID-19: SUMMER 2020

## COMPUTING PROTEIN STRUCTURE

*Developers:*

*Em Ambrosius and Robert Gotwals*

*June 20, 2020*

COMPUTING CoVID-19. COPYRIGHT HELD BY THE NORTH CAROLINA SCHOOL OF  
SCIENCE AND MATH, MAY 1, 2020. ALL RIGHTS RESERVED.

---

## INTRODUCTION

In the first part of this course, we are going to use a variety of *bioinformatics* (computational biology) tools to look at a variety of proteins that are important in the study of SARS-CoV-2 and its associated disease, COVID-19.

There is a reason why we are asking you to study protein structure at the very beginning of this course. Without it, you can't understand journal articles, you won't understand what the various computing tools are showing you, and you certainly won't be able to understand or study the interactions between proteins, such as the ACE-2 protein receptor that serves as the connection point for the SARS-CoV-2 virus, and drugs (such as remdesivir) that also interact with the ACE-2 protein.

If you have had a basic course in biology (Honors or AP Biology), then this activity should serve as a review. If you have not studied protein structure in those courses or in some other biology program, then this lab is really important.

For this activity and many others that ask you to look at protein structures, we'll use a public domain database known as the Protein Data Bank (**PDB**).

### 1.1 BASICS OF PROTEIN STRUCTURE

Being able to look at a protein structure and describe its characteristics, primarily its structure, is critically important. When you are shown a protein structure on some online resource, including Wikipedia, typically you are seeing the *secondary* structure, with its alpha helix and beta sheet features. What does that mean?

Figure 1 shows the four levels of protein structure:

1. Primary: the most basic protein structure is the *primary* structure, the string of amino acids that make up the protein. Amino acids are also called *peptides* or *residues*, and the string of amino acids is called the *sequence*. Just about every protein computing tool has an option to show the list of amino acids, and most have the ability to highlight specific amino acids based on some specific property or characteristic.

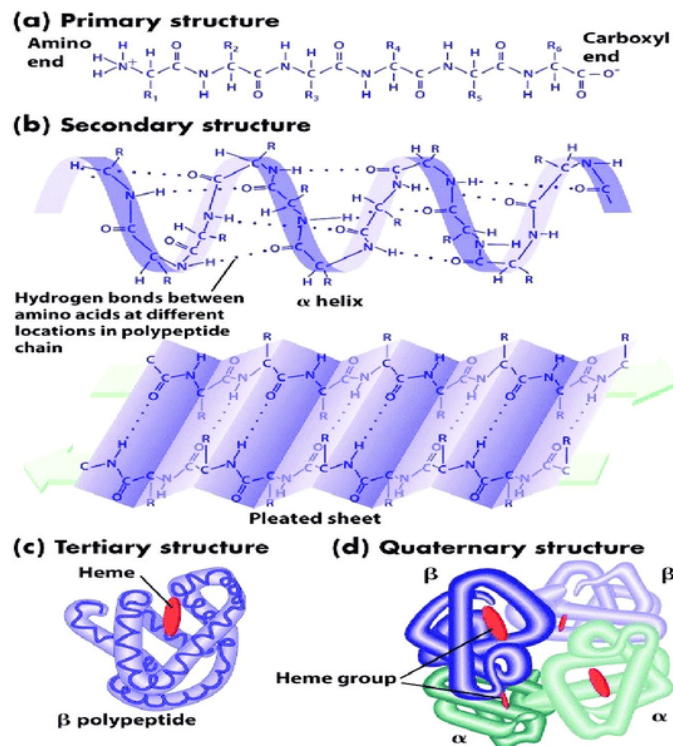


Figure 1: Four levels of protein structure

Figure 2 shows the one-letter codes for the 20 amino acids. Most of them are the first letter of the amino acid, such as "V" for valine, but some are other letters, such as "Y" for tryptophan. **Memorizing** the amino acids and their one-letter codes is strongly recommended, especially if you plan on staying in the biology business!

Note that amino acids also have a three-letter code, and you will see these as well. Again, memorizing these is a good idea!

Amino acid	Three letter symbol	One letter symbol*
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Figure 2: One letter amino acid codes

2. Secondary: Figure 3 shows the three basic types of secondary structures, alpha-helices, beta-sheets, and turns/coils. These structures are important in that they are influential on how external compounds, like drugs (also known as *ligands*) will interact, or *bind* to the protein. This protein-ligand binding determines whether or not a drug will change the structure of the protein, which in turn affects its ability to do what it does. For example, in SARS-CoV-2, we need a drug that blocks the ACE-2 protein on a host cell, thus preventing the virus from attaching, and then using that attachment to enter the host cell so it can replicate.

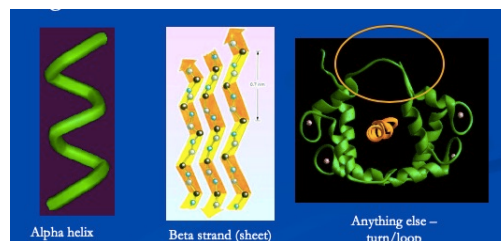


Figure 3: Three types of secondary structure

3. Tertiary: Proteins fold into three-dimensional structures, and how proteins fold is considered a *Grand Challenge* problem, one of the most scientifically and computationally challenging problems known today. Figure 4 shows how a secondary structure will fold

into this 3-D structure. How this protein folds is also fundamental in determining how the protein behaves and interacts in the body and with compounds coming into the body, such as viruses and drugs.

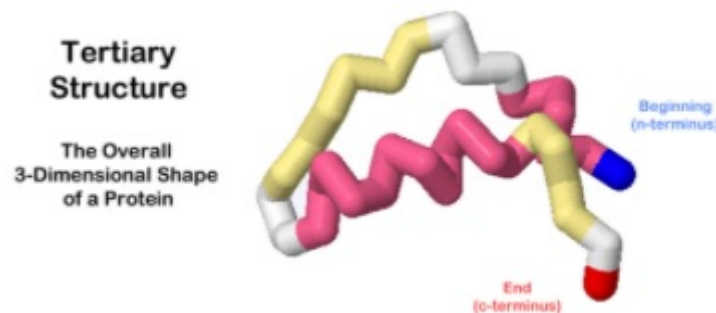


Figure 4: Tertiary structure of a protein ([1])

4. Quaternary: Most proteins consist of several *chains*. These chains, consisting of primary-secondary-tertiary components, bind together to form the final protein structure. In Figure 5, the four chains – A, B, C, and D – have been color-coded to make it easy to differentiate them.

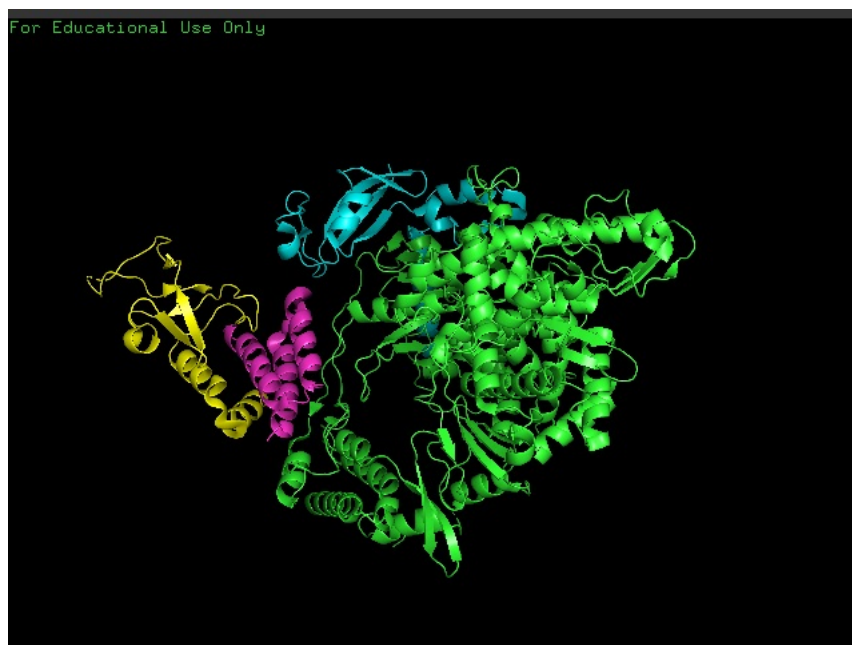


Figure 5: Quaternary structure of a protein [2]

In this next section, we will explore how protein structures are discovered, and the use of a resource known as the Protein Data Bank (PDB), an online resource for storing and sharing

protein structures by and for the scientific community.

## 1.2 THE PROTEIN DATA BANK

The Protein Data Bank (**PDB**) is the premier repository for protein structures that have been *elucidated*, or determined, by some method, such as X-ray crystallography or the newer method of Cryo-EM (electron microscopy). Once a protein has been isolated from some organism (humans, mice, bacteria, viruses, etc.) and then converted into a crystal, the coordinates (bond lengths and angles) of the atoms in the protein are determined using a crystallography method. After each molecule is uploaded it is validated by software analysis to ensure that the structure quality meets widely accepted standards and criteria. These validation protocols also ensure that the PDB entries contain the same information regardless of how the submission was processed. [3] The results are posted to the PDB, where anyone can download them and use them as needed.

The worldwide Protein Data Bank is a database for three-dimensional structural data of large biomolecules including proteins, nucleic acids, and complex assemblies. Founded in 1971 as a joint venture between Cambridge Crystallographic Data Centre and Brookhaven National Library, this database is committed to allowing public access to biomolecular data worldwide. In 2003, with the formation of the Worldwide Protein Databank (wwPDB) access to protein structural data became truly worldwide. It began with the founding members of the Protein Data Bank of Europe (PDBe), the Research Collaboratory for Structural Bioinformatics (RCSB), the Protein Data Bank of Japan (PDBj) and the Biological Magnetic Resonance Data Bank (BMRB) joined in 2006. Each individual member organization can deposit a new structural dataset, process it, and distribute it online. All submissions are reviewed, annotated, and validated before distribution.[4]

Each member organization of wwPDB has slightly different capabilities and specialties. The Biological Magnetic Resonance Data Bank collects NMR data from any experiment. This captures assigned chemical shifts and is able to calculate hydrogen exchange rates, pKa values, and relaxation parameters. RCSB PDB is the simplest (while still advanced) searching mechanism for macromolecules and ligands. They specialize in tabular reports, visualization tools, and sequence/structure comparisons. Additionally, RCSB PDB has educational resources under PDB 101. PDBj supports browsing in Japanese, Chinese, and Korean to allow for a greater amount of scientists to have access to the resources stored in the entire Protein Data Bank. PDBj is able to easily identify functionally or evolutionarily conserved motifs through the annotation of sequence and structural similarities. PDBe has multiple search and browse facilities allowing for more advanced searches into protein folding and motifs. Additionally,

PDBe has advanced visualization and validation services of NMR and EM structures.

	Atom identity Atom number	Residue identity chain	Residue number	The coordinates for each residue in the structure				
				structure				
				X	Y	Z		
<b>ATOM:</b> Usually protein or DNA	2	CA	GLY A 672	54.168	8.340	69.707	1.00164.94	C
	3	O	GLY A 672	52.692	8.194	69.380	1.00105.46	C
	4	O	GLY A 672	51.877	9.045	69.750	1.00106.67	O
	5	N	GLU A 673	52.359	7.101	66.691	1.00102.41	N
	6	CA	GLU A 673	50.994	6.795	66.274	1.00 89.17	C
	7	O	GLU A 673	50.624	5.325	66.585	1.00 81.77	C
	8	O	GLU A 673	51.438	4.411	66.405	1.00 81.89	O
	9	CB	GLU A 673	50.850	7.050	66.777	1.00 96.53	C
	10	CS	GLU A 673	50.252	8.399	66.439	1.00 99.19	C
	11	CD	GLU A 673	49.788	9.436	66.827	1.00115.45	C
	12	OE1	GLU A 673	49.062	7.477	66.681	1.00116.71	O
	13	OE2	GLU A 673	49.356	9.561	67.286	1.00113.58	O
	14	N	ALA A 674	49.387	5.109	69.023	1.00 67.27	N
	15	CA	ALA A 674	49.912	3.769	69.370	1.00 63.11	C
	16	C	ALA A 674	49.702	2.326	66.174	1.00 56.56	C
	17	O	ALA A 674	48.064	3.193	67.186	1.00 62.02	O
	18	CB	ALA A 674	47.616	3.866	70.189	1.00 47.04	C
	19	N	PRO A 675	49.260	1.612	66.240	1.00 55.66	N
	20	CA	PRO A 675	49.007	0.865	67.134	1.00 52.95	C
	21	C	PRO A 675	47.629	0.281	66.997	1.00 46.13	C
<b>HETATM:</b> Usually Ligand, ion, water	2517	OS	AQ4 774	25.725	0.372	53.259	1.00 75.72	C
	2518	N1	AQ4 774	24.269	0.712	53.215	1.00 63.00	N
	2519	O6	AQ4 774	23.410	-0.217	53.900	1.00 56.92	C
	2520	O7	AQ4 774	22.637	-0.309	53.572	1.00 52.23	C
	2521	O8	AQ4 774	21.501	0.476	52.546	1.00 46.19	C
<b>HETATM:</b> Usually Ligand, ion, water	2522	O9	AQ4 774	20.143	0.376	52.218	1.00 52.11	C
	2523	O1	AQ4 774	19.589	1.220	51.120	1.00 82.48	O
	2524	O10	AQ4 774	20.550	1.362	50.041	1.00 83.98	C

Figure 6: Structure of a PDB text file

PDB files are mostly text files, and therefore human-readable. Figure 7 shows an example of the structure of a PDB text file. Proteins are given a PDB ID number, a four-character code, such as 1HVR or 6LU7. Researchers typically converse about proteins by using their PDB ID code, so it's important to keep track of the PDB ID for the protein you are studying. I keep a notebook of PDB IDs that are of interest. These structures can be viewed using many third-party programs and the RCSB PDB site contains a separate viewing environment. Third-party programs include Jmol, Pymol, VMD, Rasmol, ICM-Browser, MDL Chime, UCSF Chimera, Swiss-PDB Viewer, StarBiochem, Sirius, and VisProt3DS.<sup>[4]</sup>



	Atom identity Atom number	chain	Residue identity Residue number	The coordinates for each residue in the structure					
				X	Y	Z			
<b>ATOM:</b>  Usually protein or DNA	ATOM	2	CA	GLY A 672	54.168	8.340	69.707	1.00104.94	C
	ATOM	3	O	GLY A 672	52.692	8.194	69.380	1.00105.46	C
	ATOM	4	O	GLY A 672	51.877	9.045	69.750	1.00106.67	C
	ATOM	5	N	GLU A 673	52.359	7.191	66.691	1.00102.41	N
	ATOM	6	CA	GLU A 673	50.994	6.785	66.274	1.00 69.17	C
	ATOM	7	O	GLU A 673	50.624	5.325	66.585	1.00 81.77	C
	ATOM	8	O	GLU A 673	51.438	4.411	66.405	1.00 81.88	O
	ATOM	9	CB	GLU A 673	50.850	7.050	66.777	1.00 96.53	C
	ATOM	10	CG	GLU A 673	50.252	8.399	66.438	1.00 99.19	C
	ATOM	11	CD	GLU A 673	48.786	9.436	66.827	1.00115.45	C
	ATOM	12	OE1	GLU A 673	48.062	7.477	66.681	1.00116.71	O
	ATOM	13	OE2	GLU A 673	49.356	9.561	67.286	1.00113.58	O
	ATOM	14	N	ALA A 674	49.387	5.109	69.023	1.00 67.27	N
	ATOM	15	CA	ALA A 674	48.912	3.769	69.370	1.00 63.11	C
	ATOM	16	C	ALA A 674	48.702	2.926	66.174	1.00 50.54	C
	ATOM	17	O	ALA A 674	48.064	3.193	67.186	1.00 62.02	O
	ATOM	18	CB	ALA A 674	47.616	3.866	70.189	1.00 47.04	C
	ATOM	19	N	PRO A 675	49.260	1.612	66.240	1.00 55.66	N
	ATOM	20	CA	PRO A 675	49.007	0.665	67.134	1.00 52.95	C
	ATOM	21	C	PRO A 675	47.629	0.261	66.997	1.00 46.19	C
<b>HETATM:</b>  Usually Ligand, ion, water	HEM 2517	O5	AQ4 774	25.725	0.972	53.250	1.00 75.72	C	
	HEM 2518	H1	AQ4 774	24.209	0.712	53.215	1.00 63.33	C	
	HEM 2519	O6	AQ4 774	23.410	-0.217	53.900	1.00 56.32	C	
	HEM 2520	O7	AQ4 774	22.037	-0.309	53.572	1.00 52.23	C	
	HEM 2521	O8	AQ4 774	21.501	0.476	52.546	1.00 48.18	C	
	HEM 2522	O9	AQ4 774	20.143	0.376	52.218	1.00 52.11	C	
	HEM 2523	O1	AQ4 774	19.589	1.220	51.120	1.00 82.48	C	
	HEM 2524	O10	AQ4 774	20.530	1.362	50.043	1.00 83.98	C	

Figure 7: Structure of a PDB text file

Figure 8 shows a list of all of the PDB files that are relevant to the coronavirus pandemic.


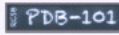



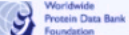
		Macromolecular Structures Enabling Breakthroughs in Research and Education	Enter search term(s) <a href="#">Advanced Search</a>   <a href="#">Browse Annotations</a>
<div>      </div>			
Search	History	Browse Annotations	MyPDB Help
<p><b>QUERY:</b> PDB ID(s) IN (6X2G, 6XB0, 6XB1, 6XB2, 6XA4, 6XAA, 6XA9, 6XCH, 6XBG, 6XBH, 6XBI, 6XDC, 6XDH, 6Z2E, 6Z4U, 6M5I, 7BQ7, 7C8R, 7C8T, 6X6P, 7BYR, 5RHB, 5RHC, 5RHD, 5RHE, 5RHF, 6X4I, 6YZ5, 6YZ7, 6Z2M, 6Z43, 6M1V, 7BWJ, 7BZF, 7C2K, 6WPS, 6WPT, 6X29, 6X2A, 6X2B, 6X2C, 6WZO, 6WZQ, 6X1B, 7BW4, 7C2I, 7C2J, 7C01, 6WZU, 5RGT, 5RGU, 5RGV, 5RGW, 5RGX, 5RGY, 5RGZ, 5RH0, 5RH1, 5RH2, 5RH3, 5RH4, 5RH5, 5RH6, 5RH7, 5RH8, 5RH9, 5RHA, 6Y2G, 6Y2F, 6Y2E, 6W02, 6W01, 6Y84, 6W41, 6W4H, 6VSB, 6W4B, 6W61, 6W63, 6W75, 6WV1, 6W6Y, 6VXS, 6VWW, 6VYO, 6VYB, 6VXX, 6YB7, 5R84, 5R83, 5R7Y, 5R80, 5R82, 5R81, 5R7Z, 5REA, 5REC, 5REB, 5REE, 5RED, 5REG, 5REF, 5RE9, 5RE8, 5RE5, 5RE4, 5RE7, 5RE6, 5RFB, 5RFA, 5RFD, 5RFC, 5RFF, 5RFE, 5RFH, 5RFG, 5REY, 5REX, 5RF9, 5REZ, 5RF2, 5REP, 5RF1, 5RES, 5RF4, 5RER, 5RF3, 5REU, 5RF6, 5RET, 5RF5, 5REW, 5RF8, 5REV, 5RF7, 5REI, 5REH, 5REK, 5REJ, 5REM, 5REL, 5REO, 5RF0, 5REN, 5RFZ, 5RFY, 5RFR, 5RFQ, 5RFT, 5RFS, 5RFV, 5RFU, 5RFX, 5RFW, 5RFJ, 5RFI, 5RFL, 5RFK, 5RFN, 5RFM, 5RFP, 5RFO, 5RG0, 6M03, 6M17, 6M0J, 6M3M, 6LU7, 6LVN, 6LXT, 6LZG, 6W9C, 5R8T, 6M71, 6W9Q, 6Y13, 7BTF, 6WEN, 6WCF, 5RG1, 5RG2, 5RG3, 5RGG, 5RGH, 5RGI, 5RGJ, 5RGK, 5RGL, 5RGM, 5RGN, 5RGO, 5RGP, 5RGQ, 5RGR, 5RGS, 6M2N, 6M2Q, 6YLA, 6WIQ, 6WJI, 6WJT, 7BQY, 7BV2, 7BV1, 6LZE, 6M0K, 7BUY, 6W37, 6WEY, 6WKP, 6WKQ, 6WLC, 6YHU, 6YMO, 6YNO, 6YOR, 6WKS, 6WNP, 6WOJ, 6WQF, 6WQ3, 6WQD, 6WRH, 6YT8, 6YWK, 6YWL, 6YWM, 6WRZ, 6WTC, 6WVN, 6YVA, 6YYT, 6YZ1, 7BRO, 7BRP, 7BRR, 7BZ5, 6WTJ, 6WTK, 6WTM, 6WTT, 6YVF, 6YZ6, 6WUJ, 6WX4, 6WXC, 6WXD, 6YUN, 7C22)</p>			

Figure 8: Coronavirus PDB list, as of June 2020

Figure 9 shows a sample entry for the 7BV1 structure, an important SARS-CoV-2 protein structure file. Notice the date on this entry (April 22, 2020). This protein was isolated from the SARS-CoV-2 virus, using cryo-EM at a resolution of 2.8 angstroms. This entry contains three proteins: nsp12, nsp7, and nsp8, proteins that we will look at in this course. It also has one *ligand*, or small-atom molecule. Ligands are typically drugs, but they can also be co-



factors, atoms or molecules that support the functioning of the protein. In this case, there is a zinc (Zn) atom that serves as a co-factor but is listed as a ligand.

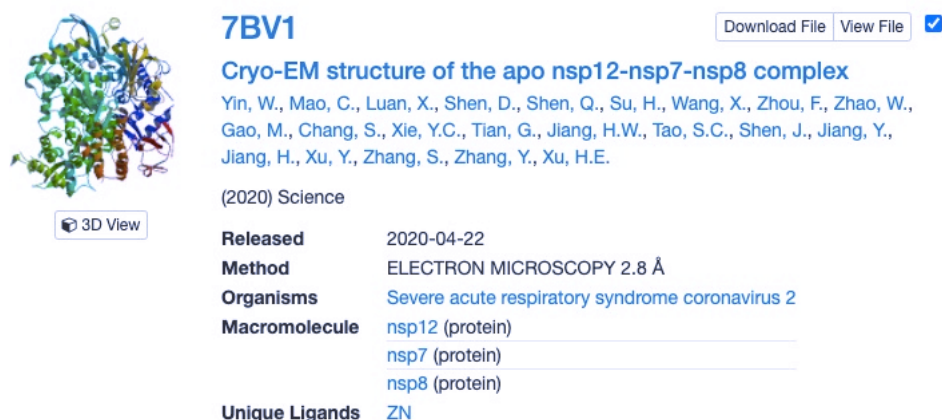


Figure 9: PDB ID 7BV1, an important SARS-CoV-2 structure

2

## STUDENT ACTIVITY

**NOTE! The majority of the steps for the activity will be demonstrated in the webinar. These instructions are meant only as short reminders of the available resources related to PDB and expectations concerning the scavenger hunt of SARS-CoV-2 proteins in the Protein Data Bank.**

For this lab, you are asked to answer some questions based on the lab reading (this document) and a PDB "scavenger hunt" (Section 2.2) for the PDB IDs 6W41 and 6VXX.

### 2.1 PROTEIN DATA BANK RESOURCES

The following resources will aid in finding the necessary information about the Protein Data Bank. Review each of the resources below to learn more about the worldwide Protein Data Bank and the partner organizations. [3]

1. Introduction to PDB Data

2. [Worldwide PDB](#) [3]
3. [PDB 101 \(Protein Data Bank educational branch\)](#)
4. [Wikipedia overview of the Protein Data Bank](#) [4]
5. [RCSB PDB](#)
6. [Biological Magnetic Resonance Data Bank](#)
7. [PDB of Europe](#)
8. [PDB of Japan](#)
9. [User guide to the wwPDB EM validation reports](#)

## 2.2 SCAVENGER HUNT OF SARS-CoV-2 PROTEINS

This Scavenger Hunt will look at the properties of the SARS-CoV-2 proteins [6W41](#) [5] and [6VXX](#) [6]. A Google sheet is available [here](#) for you to make a copy of and keep track of what you are finding to aid in completing the quiz.

### 2.2.1 6W41 QUESTIONS [5]

1. What is the title of this protein file?
2. What organisms did it come from?
3. What date was this structure deposited?
4. Who authored the structure deposition?
5. What journal was the paper the accompanies this structure published in?
6. How many atoms are contained within this structure?
7. How many unique protein chains are contained within this structure?
8. What is the sequence length of the longest unique protein chain within this structure?
9. What is the resolution of this structure in angstroms?
10. What ion is present in this structure as a ligand?

### 2.2.2 6VXX QUESTIONS [6]

1. What organism was this protein expressed in?
2. What method was used to obtain this data?
3. What was the electron dose used in the experiment that obtained this structure? (answer in electrons per square angstrom)
4. How many alpha helices are seen in the secondary structure of one unique protein chains?
5. What percent of the secondary structure is covered in beta sheets?
6. How many total poly-peptide chains are in the structure?
7. What type of amino acid is the 753rd in the chain sequence? (answer in a one letter abbreviation of the amino acid)
8. In the **validation report**, what percentage of chain A has quality classified as green?
9. In the **validation report**, how many sulfur atoms are in chain B of the protein?
10. How many residues are in chain B of the protein?

---

## REFERENCES

- [1] Tertiary structure.
- [2] Introduction to pymol, 2009.
- [3] Helen Berman, Kim Henrick, and Haruki Nakamura. Announcing the worldwide Protein Data Bank. *Nature Structural & Molecular Biology*, 10(12):980, 2003.
- [4] Wikipedia Contributors. Protein data bank, May 2020.
- [5] M. Yuan, N. C. Wu, X. Y. Zhu, and I. A. Wilson. 6w41: Crystal structure of sars-cov-2 receptor binding domain in complex with human antibody cr3022, Mar 2020.
- [6] A. C. Walls, Y. J. Park, M. A. Tortorici, A. Wall, Seattle Structural Genomics Center for Infectious Disease (SSGCID), A. T. McGuire, and D. Veasley. 6vxx: Structure of the sars-cov-2 spike glycoprotein (closed state), Mar 2020.