	SKILLS CENTER STANDARD OPERATING PROCEDURE	**A BIOFIZZ** **PRODUCTON**
Navigating the Protein Data Bank 2 **Module Hours: 3**	**Effective Date: 04/05/2021** **PRQs: PDB1**	**Revision # 1.0** **A. Siclair** **Checked: M. Guzie**

BACKGROUND

One of the most important concepts to have surfaced in biology is the connection between protein structure and function. It has become increasingly notable that this relationship is imperative to bodily systems and that disruption of protein structure can lead to the establishment of numerous different diseases and phenotypes, such as Alzheimer's and prion diseases. The RCSB PDB and the worldwide PBD curate published structures of biomolecules and allow researchers to access biomolecules structure information.

The Research Collaboratory for Structural Bioinformatics (RCSB) protein data bank, established by Walter Hamilton in 1971, began its existence with a mere 7 protein structures to its name (RCSB PDB). As of 2015, over 100,000 structures have been documented in the archive

Figure 1: 1QUN, X-ray structure of the FIMC-FIMH chaperone adhesin complex from uropathogenic E. coli

in numerous different organisms, one of which is seen in Figure 1 (RCSB PDB). This program is advantageous because it allows for collaboration between researchers on an international scale, something that is invaluable when considering the big picture in research. The protein data bank information has contributed to fields including pharmacological advancement, such as the establishment of Gleevec for treatment of chronic myeloid leukemia, and will continue to aid in treatment progressions in the future at a macromolecular level (Sullivan et al., 2017).

The RCSB PDB is one of the first stops when analyzing a protein, whether focusing on structures or interactions, because the PDB file can be used within other applications such as UCSF Chimera for visualization. By developing an understanding of the fundamental tools in PDB, one can establish a solid foundation for further analysis down the line.

	SKILLS CENTER STANDARD OPERATING PROCEDURE	**A BIOFIZZ** **PRODUCTON**
Navigating the Protein Data Bank 2 **Module Hours: 3**	**Effective Date: 04/05/2021** **PRQs: PDB1**	**Revision # 1.0** **A. Siclair** **Checked: M. Guzie**

1. PURPOSE

The purpose of this procedure is to become comfortable with navigating the RCSB protein data bank (PDB) and to understand the steps needed to analyze the quality of a PDB model.

2. SCOPE

This procedure applies to qualified skills center users.

3. RESPONSIBILITY
   3.1.   It is the responsibility of the user to understand and perform the procedure described in this document.
   3.2.   It is the responsibility of the user performing the procedure to fully document any deviations from the written procedure.
   3.3.   It is the responsibility of the user to become trained on the procedure.

4. DEFINITIONS
   4.1.   Protein data bank – An international database that houses information about 3D macromolecule structures such as proteins and nucleic acids.
   4.2.   PDB identifier – The 4-digit alpha-numerical code that corresponds to a specific structure in the PDB database.
   4.3.   wwPDB validation report – A document that provides an assessment of PDB model quality using widely accepted standards and criteria. (wwPDB, 2020)
   4.4.   wwPDB x-ray validation report – A document that summarizes the quality of the x-ray structure and highlights specific concerns by considering the atomic model, the diffraction data and the fit between the atomic model and the diffraction data. (wwPDB, 2020)
   4.5.   Deposition – The act of uploading data for a particular structure and granting it with a PDB identifier for filing in the PDB archives.
   4.6.   Rfree – A measure of the fit of the model to a small subset of the experimental data, which was not used in model refinement. (wwPDB, 2020)
   4.7.   Clashscore – A value that quantifies the number of clashes per 1000 atoms, including hydrogens (wwPDB, 2020)
   4.8.   MolProbity – A structure-validation web service that provides broad-spectrum solidly based evaluation of model quality at both the global and local levels for both proteins and nucleic acids. (Chen et al., 2010)

4.9. Ramachandran outlier value – The percentage of Ramachandran outliers with respect to the total number of residues in the entry for which the outlier assessment is available. (wwPDB, 2020)

4.10.   Ramachandran outlier – A residue whose combination of its φ and ψ torsion angles is unusual, as assessed by MolProbity. (wwPDB, 2020)

4.11.   Sidechain outliers – The percentage of residues with an unusual sidechain conformation with respect to the total number of residues for which the assessment is available. (wwPDB, 2020)

4.12.   Real-space R value (RSR) = A measure of the quality of fit between a part of an atomic model (in this case, one residue) and the data in real space. (wwPDB, 2020)

4.13.   RSR Z-score outliers – The percentage of RSRZ outliers with respect to the total number of residues for which RSRZ was computed. (wwPDB, 2020)

4.14.   Resolution – The minimum distance needed to distinguish between 2 atoms.

4.15.   Chirality – The inability of a molecule to be superimposed onto its mirror image.

4.16.   Clash overlap – The magnitude of a clash, defined as the difference between the observed interatomic distance and the sum of the van der Waals radii of the atoms involved. (wwPDB, 2020)

4.17.   Torsion angle – The angle formed by three consecutive bonds in a molecule and defined by the angle created between the two outer bonds. A protein backbone has 3 different torsion angles: phi, psi and omega. (CMBI, 2020)

4.18.   Chi torsion angles – The torsion angles that describe protein side chain conformation. (wwPDB, 2020)

4.19.   RNA suite – Used to assess torsion angles in the RNA, and consists of the torsion angles between the sugars in two RNA nucleotides and is identified by the 3' nucleotide. (wwPDB, 2020)

5. MATERIALS/EQUIPMENT

5.1.   RCSB Protein Data Bank: https://www.rcsb.org/

6. PROCEDURE

6.1. Accessing a Validation Report for an X-ray PDB Model

6.1.1. Go to the RCSB PDB: https://www.rcsb.org/
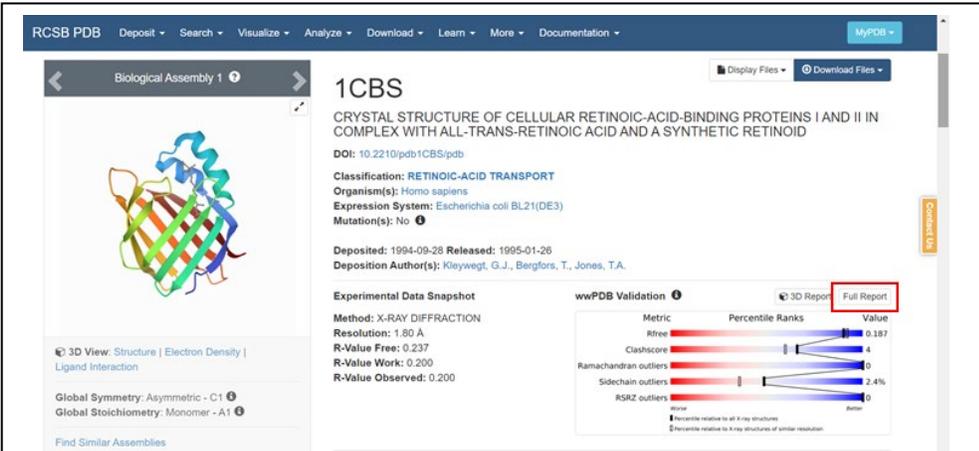
6.1.2. Search for a molecule of interest as described in PDB SOP 1.

-    PDB identifier used for this example: 1CBS

	SKILLS CENTER STANDARD OPERATING PROCEDURE	**A BIOFIZZ** **PRODUCTON**
Navigating the Protein Data Bank 2 **Module Hours: 3**	**Effective Date: 04/05/2021** **PRQs: PDB1**	**Revision # 1.0** **A. Siclair** **Checked: M. Guzie**

- Note: Each different experimental method used to collect a PDB model might have some variation in the structure of the validation report. Since 1CBS was collected via X-ray, this SOP covers this specific type of validation report. Some variances may occur with other experimental methods. See this link for more information on other experimental methods: https://www.wwpdb.org/validation/validation-reports

6.1.3. Once on the page for the desired PDB model, click the Full Report button to pull up the validation report for the molecule. (Figure 2)
- A new window will pop up with the validation report.


Figure 2: The information page for the 1CBS model.

6.2. Analyzing and Understanding a Validation Report

6.2.1. The title page gives some basic information about the model and the validation report itself.
- The type of validation report will be listed: a preliminary report is one submitted before deposition and should not be submitted to a journal, a confidential report is sent only to the depositing authors after a PDB ID has been given to the model, and a RCB public access report is available after deposition to the public. The 1CBS report is a public access report as seen on the title page. (Figure 3) (wwPDB, 2019)

- The report can either be full-length or summary: summary only describes the 5 worst outliers, while the full-length describes all outliers. The 1CBS report is full-length (Figure 3) (wwPDB, 2019)
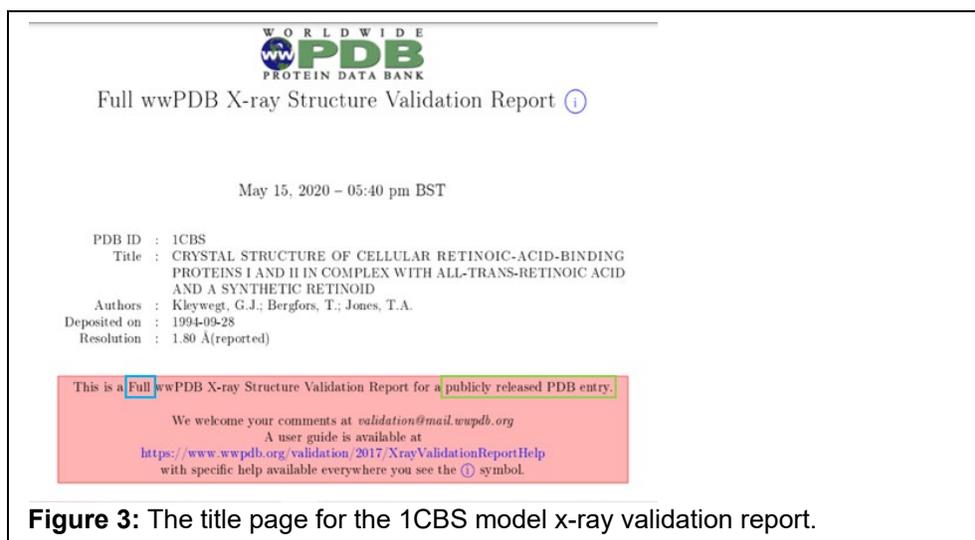

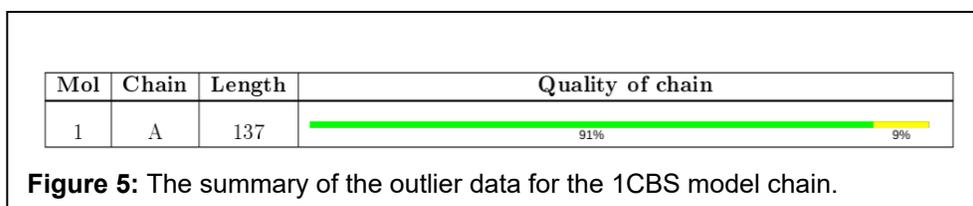**Figure 3:** The title page for the 1CBS model x-ray validation report.

6.2.2. The next section is the overall quality at a glance, which gives a summary of the most important indicators of model quality. (Figure 4) (wwPDB, 2020)
  - This section focuses on comparison of structures by percentile rank of quality.
  - Black vertical box = global percentile rank, calculated with respect to all X-ray structures available in the PDB archive. (Figure 4)
  - White vertical box = resolution-specific percentile rank, calculated with respect to a subset of X-ray entries in the same subset of the PDB archive, but only considering entries with comparable resolution to this entry
  - Sliders that are farther to the right, towards the blue side, indicate a higher quality structure. (Figure 4)
  - There are 5 main categories that a molecule is scored on: Rfree, Clashscore, Ramachandran outliers, Sidechain outliers and RSRZ outliers. (See definitions section for details on what these indicators describe).
  - The reported resolution of the molecule is listed too: The higher the resolution, more detail can be observed. (Figure 4)

- The table describes how many molecules were used to help calculate the metrics for the given model. (Figure 4)

Figure 4: The "Overall quality at a glance" section of the 1CBS model x-ray validation report.
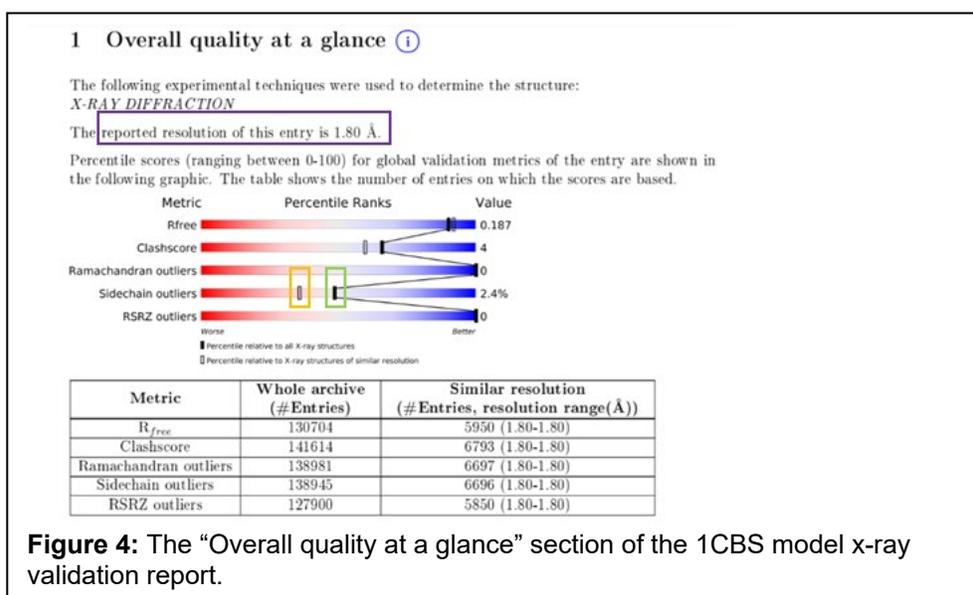
- There is also a summary of the various chains in the structure that looks at the fit of the chains compared to the electron density. There can be green, yellow, orange, and red portions in the lower bar for each chain that indicate the fraction of residues that contain outliers for a certain number of model validation criteria. (wwPDB, 2020) Green = 0, yellow = 1, orange = 2 and 3 or more = red. More green tends to indicate a better-quality model. In the case of 1CBS, there are only outliers for the 0 or 1 criteria categories. (Figure 5)
- An additional red bar above the lower bar indicates electron density outliers. (Figure 5)

Figure 5: The summary of the outlier data for the 1CBS model chain.

6.2.3. The entry composition section describes the atoms and molecules in the entry and particularly focuses on the unique molecules in the model.

6.2.4. The Residue-property plots section summarizes qualitative information about the DNA, RNA, or protein sequences of the model. This graphic uses the same color coding as seen before but applied to the annotated sequence instead. (Figure 6)

- Red dot = residue with poor fit to electron density.
- Green connector bar = consecutive stretches of residues where no outliers were detected.
- Grey = residues absent from final model.

Figure 6: The residue-property plot for 1CBS, showing minimal outliers.
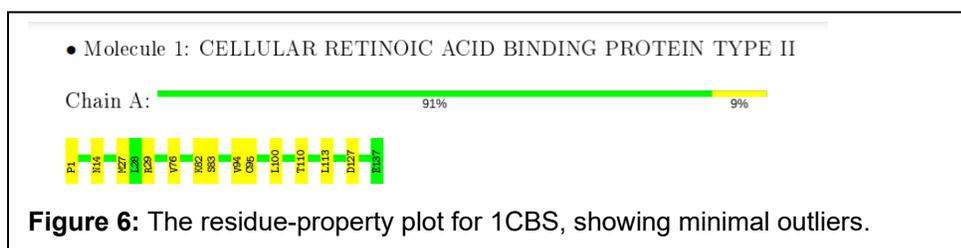
6.2.5. The next section, data and refinement statistics, gives insight into the data itself as well as the source of the information.

- For detailed information on all of the included metrics, access section 4 of the wwPDB x-ray validation guide at: https://www.wwpdb.org/validation/2017/XrayValidationReportHelp

6.2.6. The model quality section houses some of the most important and detailed information for the validation and quality determination. First, standard geometry is assessed. This includes bond lengths, bond angles, chirality, and planarity.

- For low resolution models, bond lengths and angles should output low RMSZ values closer to 0 while high resolution models should output higher values closer to 1. (wwPDB, 2020)
- When assessing planarity, the validation programs used look at side chain, peptide and main chain deviations for potential outliers. (wwPDB, 2020)
- When looking at chirality, the chiral volume is calculated and then the sign is evaluated to look for "deviations". If the sign of the calculated volume is wrong, then the handedness of the chirality is wrong and is counted as a deviation. (wwPDB, 2020)

6.2.7. The model quality section also looks at too-close contacts, meaning that it analyzes steric clashes between non-bonded atoms involving van der Waals. (wwPDB, 2020)

- There is evaluation of clashes within the same asymmetric unit and within different asymmetric units.
- MolProbity calculates the clashscore (see definitions).
- Symmetry-related clashes are those that occur between different asymmetric units.
- The first table gives basic insight into the various components of the clash. (Figure 7)

Mol		Chain	Non-H	H(model)	H(added)	Clashes	Symm-Clashes
1		A	1091	0	1106	7	0
2		A	22	0	27	2	0
3		A	100	0	0	2	0
All		All	1213	0	1133	9	0

Mol	The molecule identifier
Chain	The instance identifier
Non-H	The number of non-hydrogen atoms modelled.
H(model)	The number of hydrogen atoms modelled.
H(added)	The number of hydrogen atoms added by MolProbity.
Clashes	The number of clashes in which the atoms in this instance of the molecule are involved.
Symm-clashes	The number of symmetry-related clashes in which the atoms in this instance of the molecule are involved.

Figure 7: The summary of the clashes in an example molecule (right) and a key explaining what each component describes (left).

- If asymmetric clashes are present, further details will be given in a second table. (Figure 8)

Atom-1	Atom-2	Interatomic distance (Å)	Clash overlap (Å)
1:A:82:LYS:HG2	1:A:100:LEU:HD21	1.56	0.87
2:A:200:REA:H8	2:A:200:REA:H181	1.81	0.61
1:A:82:LYS:HG2	1:A:100:LEU:CD2	2.32	0.59
2:A:200:REA:C8	2:A:200:REA:H181	2.41	0.51
1:A:1:PRO:HB2	1:A:113:LEU:HD12	1.97	0.47
1:A:76:VAL:HG23	3:A:310:HOH:O	2.16	0.46
1:A:29:ARG:NH2	3:A:387:HOH:O	2.48	0.45
1:A:83:SER:HB3	1:A:95:CYS:SG	2.57	0.44
1:A:94:VAL:HG22	1:A:110:THR:HG22	2.03	0.41

There are no symmetry-related clashes.

Atom-1	The molecule identifier, instance identifier, residue number, residue name and atom name for the first atom. where applicable, the chain identifier is prefixed with model number and an alternative conformation identifier is shown as a suffix to the atom name.
Atom-2	Identifies the second atom in the clash.
Interatomic	The distance between **Atom-1** and **Atom-2** in Å.

Figure 8: Further details about the clashes in the 1CBS model (top) and a key explaining what each component describes (bottom).

- If symmetric clashes are present, further details will be given in a third table. There are no symmetric clashes in 1CBS. (Figure 8)

6.2.8. The model quality section also looks at torsion angles if protein molecules make up a part or all of the model.

- The protein backbone section is displayed in a table format, with columns written as explained in Figure 9. If Ramachandran outliers exist, then 5 random outliers are chosen to be displayed in the summary report. (wwPDB, 2020)

Mol	The molecule identifier
Chain	The instance identifier
Analysed	The first number here is the number of residues in the chain for which MolProbity output is available. The second number is the total number of residues in the chain. Phi and psi angles cannot be analysed for terminal residues, non-standard residues or for residues with incompletely modelled main chain.
Favoured, Allowed, Outliers	The number (and percentage) of residues in the favoured, allowed and outlier regions respectively, of the residue-specific phi-psi plots.
Percentiles	The percentile score based on the percentage of Ramachandran outliers in the chain. These are given relative to the whole archive (first value) and relative to structures of a similar resolution (second value). The colours around the percentile values correspond to the slider positions in the *Overall quality* section of the report, **as described above**

Figure 9: Details about the protein backbone table in the torsion angles section.

- When looking at the protein side chains, there will be differing torsion angle values depending on the residue type. MolProbity compares the side chain of the new model to the side chains in existing, high quality models to determine if the side chain is an outlier or not. (wwPDB, 2020) The side chain section is displayed in a table format, with columns written as explained in Figure 10.

Mol	The molecule identifier
Chain	The instance identifier
Analysed	The first number here is the number of residues in the chain which were analysed by MolProbity. The second number is the total number of residues in the chain. Chi torsion angles cannot be analysed for non-standard residues or for residues with incompletely modelled sidechains.
Rotameric, Outliers	The number (and percentage) of residues with favoured, and unusual chi torsion angles respectively.
Percentiles	The absolute and relative percentile scores based on the percentage of sidechain outliers in the chain. These are given relative to the whole archive (first value) and relative to structures of a similar resolution (second value). The colours around the percentile values correspond to the slider positions in the *Overall quality* section of the report, **as described above**

Figure 10: Details about the protein side chain table in the torsion angles section.
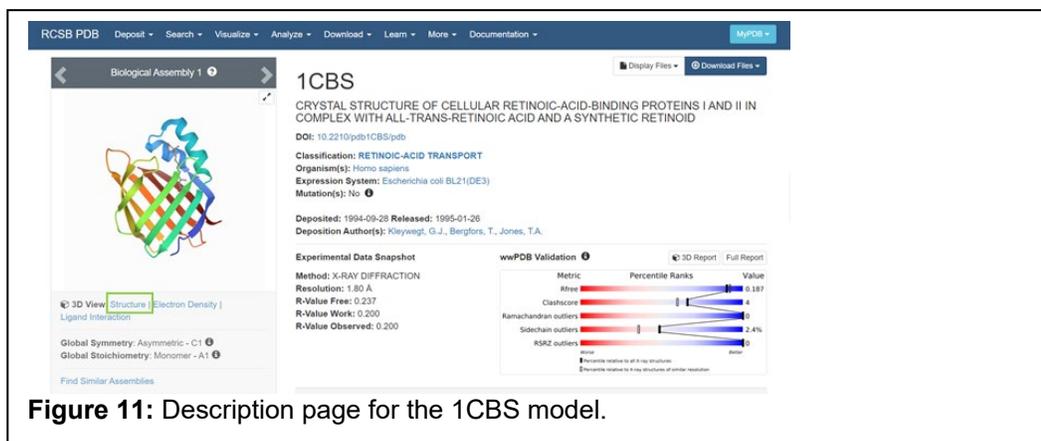
- The asp, glu and his residues are occasionally flipped in the molecules for hydrogen bonding purposes. If this happens in the model, it is expressed in a table format.
- The torsion angle section also assesses the RNA by looking at suites (see definitions) and the backbone and reports the results in a table format. The suiteness parameter is the overall outcome determined by MolProbity. (wwPDB, 2020) The rest of the columns serve similar functions to the other torsion angle sections.
- This section also analyzes other miscellaneous things like non-standard residues, carbohydrates, ligands and other polymers using multiple different programs. (wwPDB, 2020) The evaluated criteria include bond lengths, bond angles, torsions and isolated rings. (wwPDB, 2020)
- The polymer linkage issues section identifies any chain breaks in the model.

6.2.9. The "Fit of model and data" section compares the molecules in the entry model to the deposited experimental data.

- The protein, DNA and RNA section focuses on comparing the residues to the experimental electron density to look for outliers.
- The section also looks at the same other miscellaneous categories as seen in the end of the torsion angles section.

6.3. Comparing the Model to an Electron Density Map

6.3.1. Open the RCSB PDB page for the model of interest (EX: 1CBS)

6.3.2. Click on the Structure option for one of the given biological assemblies. (Figure 11)

	SKILLS CENTER **STANDARD OPERATING PROCEDURE**	**A BIOFIZZ** **PRODUCTON**
Navigating the Protein Data Bank 2 **Module Hours: 3**	**Effective Date: 04/05/2021** **PRQs: PDB1**	**Revision # 1.0** **A. Siclair** **Checked: M. Guzie**

Figure 11: Description page for the 1CBS model.

6.3.3. Once the new window opens, ensure that the Mol Javascript viewer is the selected option, noted at the bottom right corner of the page. (Figure 12)

6.3.4. Click on the Density option and press Enable. (Figure 12)

Figure 12: Enabling the Density function in the Mol Javascript viewer for the 1CBS model.
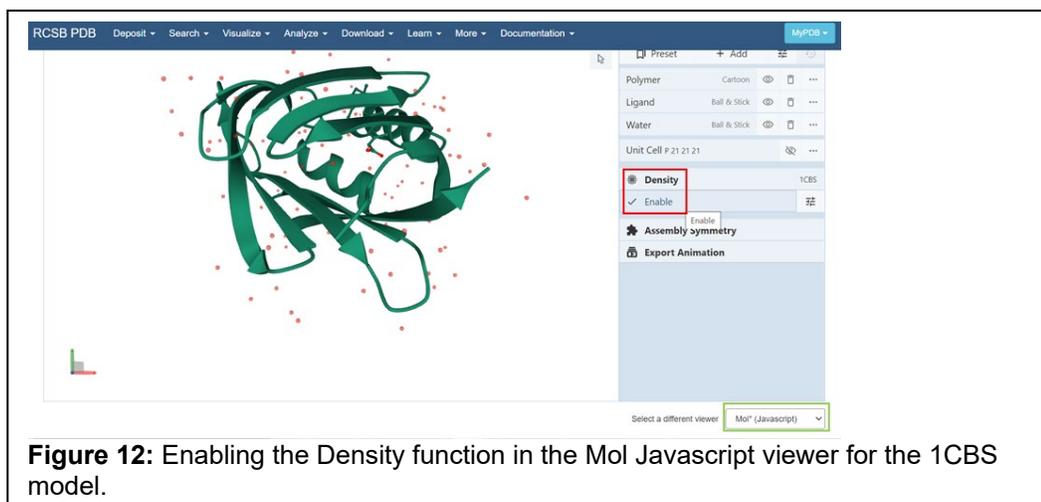
6.3.5. Clicking on a specific residue will pull up the density for the region within a 5-angstrom radius. (PDBe, 2017)

6.3.6. 3 different density map options will pop up after enabling the density function.

- Blue = density that should surround the atoms
- Red = negative density, indicating extra atoms in this location
- Green = positive density, indicating that atoms are missing in this location

- Less red and green density tends to mean that the model is a better fit for the electron density map.

6.3.7. For each type of density map, it is possible to change the opacity and/or enable a wire frame outline. (Figure 13)

- Clicking the eye icon next to a specific density map will cause that map to be ommitted from the viewer screen. (Figure 13)
- Changing the ISO value, utilizing the bar next to the eye icon, changes the amount of density available for analysis in the given region. (Figure 13)

Figure 13: When a residue is first clicked, a semi-transparent outline of the density maps around that region will appear (left). The opacity and presence of a wire frame for each of the 3 types of density map can be altered (middle), as seen by the alteration of the green electron density map for the 1CBS model. The green wireframe indicates that there should be atoms present in this region that are missing in this model. The control panel for the density maps (right).

7. REFERENCES

Chen, V.B., Arendall III, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S. & Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D, D66,* 12-21. http://scripts.iucr.org/cgi-bin/paper?S0907444909042073

CMBI. (2020, Feb 1). *Torsion Angles.* CMBI. https://www3.cmbi.umcn.nl/wiki/index.php/Torsion_angle

Kleywegt, G.J., Bergfors, T. & Jones, T.A. (2020) Full wwPDB X-ray Structure Validation Report. *wwPDB.* https://www.wwpdb.org/validation/samples/ 1cbs/1cbs_validation.pdf

[Protein Data Bank in Europe – PDBe]. (2017, Aug 14). *Mini Tutorial: Viewing electron*

	SKILLS CENTER STANDARD OPERATING PROCEDURE	**A BIOFIZZ** **PRODUCTON**
Navigating the Protein Data Bank 2 **Module Hours: 3**	**Effective Date: 04/05/2021** **PRQs: PDB1**	**Revision # 1.0** **A. Siclair** **Checked: M. Guzie**

density in Litemol. [Video].Youtube. https://www.youtube.com/watch ?v=HF6xVLY3lec

RCSB Protein Data Bank. (1995). *1CBS: Crystal structure of cellular retinoic-acid binding proteins I and II in complex with all-trans-retinoic acid and a synthetic retinoid.* RCSB PDB. https://www.rcsb.org/structure/1CBS

wwPDB Foundation. (2020) *User guide to the wwPDB x-ray validation reports.* wwPDB. https://www.wwpdb.org/validation/2017/XrayValidationReportHelp

wwPDB Foundation. (2019) *wwPDB validation report FAQs.* wwPDB. https://www.wwpdb.org/validation/2017/FAQs

wwPDB Foundation. (n.d.) *wwPDB Validation Reports.* wwPDB. https://www.wwpdb.org/valida.tion/validation-reports

	SKILLS CENTER STANDARD OPERATING PROCEDURE	A BIOFIZZ PRODUCTON
Navigating the Protein Data Bank 2 **Module Hours: 3**	**Effective Date: 04/05/2021** **PRQs: PDB1**	**Revision # 1.0** **A. Siclair** **Checked: M. Guzie**

8. MODULE MASTERY TASK

This task will test your ability to analyze the quality of a PDB model in the RCSB Protein Data Bank.

1. Choose a protein using the basic search, narrow the results to a specific organism of your choosing. You are encouraged to find a model that was determined via X-ray since this SOP focuses on this specific validation report category. Document the PDB code and organism of the protein you chose, then open the validation report.
2. Is the chosen validation report a summary or full-length?
3. Describe the percentile rank graphic of the chosen model. How do the global percent rank boxes compare to the resolution-specific percentile rank boxes? Based solely on this graphic, does your model seem to be of higher quality? Why or why not?
4. What is the reported resolution of your model? Is this a high or low resolution model?
5. Pick a chain from your model and look at the chain analysis section in "Overall quality at a glance". Which chain did you choose? (Only necessary to answer if more than one chain is present)
6. What percentage of the chosen chain has outliers for 0 model validation criteria? For 2 model validation criteria?
7. Are there any electron density outliers in the chosen chain?
8. Pull up the electron density maps for your model and choose a residue. At first glance, which color does there appear to be the most of at your chosen residue?
9. Change the iso values for each of the different electron density maps. What does increasing the ISO value do to each of the different maps?
10. Omit 2 of the electron density map subtypes from the viewer so that you can focus in on one type. Implement the wire frame function and increase the opacity level. Document your chosen ISO value. Submit the resulting image.