wwPDB validation report FAQs

Table of Contents

 This FAQ last updated: 05 June 2019

1. Type of reports

1.1. What are the different types of validation reports?

Four different types of validation reports are produced at different stages of the preparation, deposition, annotation and public release of a macromolecular structure.

Validation Server Preliminary Report	This kind of report is produced by the Validation Server whenever you want to validate your structure/data, at any time prior to deposition. See also FAQ: What is a "Preliminary" Validation Report?
On Deposition Preliminary Report	This report is produced during the initial deposition process using the OneDep deposition system http://deposit.wwpdb.org/deposition/. It includes a pink diagonal watermark Not For Manuscript Review on every page. See also FAQ: What is a "Preliminary" Validation Report?
Validation Report for Manuscript Review	This report is produced once at the annotation stage after a *PDB ID* has been issued for the structure. Its title page shows the *PDB ID*, Title and the deposition date. It includes a pink diagonal watermark For Manuscript Review on every page. The confidential validation reports are sent **only** to the depositors (who may choose to submit them with their manuscripts).
Report for a publicly released PDB Entry	Validation reports are currently available for all released PDB entries determined by X-ray crystallography, NMR or EM. See wwPDB Validation Reports help top page for details on how to obtain the reports. Reports for all PDB entries will be updated annually.

For each of the 4 types of reports two different lengths of report are available:

Summary — This is the short form where, if outliers are found, only the most significant 5 outliers of each category are listed.

Full — This is the longer form where every outlier is listed.

That if there are no or fewer than 5 outliers found in every category checked the *Summary* and *Full* reports will be identical

1.2. What is a 'Preliminary' Validation Report?

Preliminary reports are produced by the Validation Server and when a structure is initially deposited using the OneDep deposition system http://deposit.wwpdb.org/deposition/.

Preliminary reports include a pink diagonal watermark Not For Manuscript Review on every page. The preliminary report is not a proof of deposition and should not be submitted to journals.

The reports are described as "preliminary" because there are currently limitations in the checks made:

- The sequence of the structure is taken from the input coordinate file and no check is made that this matches the sequence of the macromolecule from the depositor and/or external databases.
- Mogul results are only given for ligands that can be matched against a publicly-released wwPDB Chemical Component Definition (wwPDB CCD) for the chemical component id. The matching requires the ligand chemical component id (aka 3-letter code) and atom names in the uploaded file match. On deposition once the structure is annotated a wwPDB CCD will be created for the ligand and Mogul results will be shown in the Confidential Report produced at this stage (see FAQ: What are the different types of validation reports?).

1.3. Which report should be submitted to the journal for manuscript review?

The report produced at the annotation stage after the entry deposition. Its title page shows the *PDB ID*, Title and the deposition date and includes a pink diagonal watermark For Manuscript Review on every page.

1.4. The validation report only lists up to 5 outliers - what if I want to see more?

If the report you are looking at truncates the list of outliers then you are looking at a "summary" report. To see all the outliers found, download "full" report instead.

1.5. The validation report is too long with very long tables of outliers - how can I see a shorter version?

You are looking at a "full" report that lists all outliers. If you look instead at the "summary" report only the worst five outliers will be listed for each category.

2. Validation reports and the deposition process

2.1. What information will be made available about a deposited structure before the coordinates and experimental data are released?

Similar to a typical manuscript submission process (where information is confidential until a paper is published), only limited information is made publicly available prior to release of a PDB entry (see also next FAQ about timings).

The confidential validation reports are sent **only** to the depositors (who may choose to submit them with their manuscripts). They include the results of geometry checks, structure factor validation, and ligand validation. Coordinates and experimental data are not included in the validation report.

Mandatory submission of wwPDB validation reports has not had an impact on the number of submissions to IUCr journals.

2.2. What is the timing on making information about a deposited structure publicly available? What is the timing for releasing the coordinates, etc.?

The validation reports are generated as part of the current wwPDB curation pipelines, and do not affect any timings. Please see the Release of PDB Entries section of the wwPDB Processing Procedures and Policies Document for more information about timings.

2.3. How much control do authors have over the release of information?

During deposition, validation reports are only provided to the depositors, and are not provided by the wwPDB to journals or other third parties. Once a structure is released a validation report will be made available for it like other PDB entries.

2.4. I have deposited a structure with the PDB but have not received a report. Why?

The likely reason is the Confidential Validation Report is only produced after the structure has been annotated by a biocurator and a PDB id has been issued. The annotation process takes time.

It should also be noted that validation reports are only provided for new depositions of X-ray crystal, NMR or 3DEM structures and subject to the successful completion of the underlying calculations. For instance, no reports are currently created for structures determined by neutron diffraction and occasionally a software problem may preclude generation of a report for a

structure (we endeavour to fix any such problems as soon as possible, where needed in collaboration with the authors of the software).

3. The Validation Server

3.1. Can I get a validation report for my own intermediate or unpublished structure?

Validation reports can now be generated on demand by using the wwPDB Validation Server https://validate.wwpdb.org. This now works for structures produced for X-ray structures, 3D electron microscopy and NMR methods.

3.2. Where can I get more information on the wwPDB Validation Server?

Please see http://wwpdb.org/validation/validation-servers.

3.3. Is the validation package available for me to assess my structures?

The wwPDB Validation Server https://validate.wwpdb.org allows the production of reports for your structures.

The validation package software has not been made available for distribution as it includes a lot of PDB-specific codes that makes it difficult to use elsewhere. Instead a publically accessible standalone validation server is provided to enable a user to produce validation reports prior to deposition. Most of the tools used by the validation server are available separately.

4. Validation reports for existing PDB structures

4.1. Can I get a validation report for an existing structure in the PDB?

In 2014, validation reports for all existing X-ray crystal structures in the PDB archive were made publicly available through the wwPDB ftp sites. Validation reports for a particular PDB entry are also available from the entry's page at RCSB PDB, PDBe or PDBj. Validation reports for all Nuclear Magnetic Resonance (NMR) and 3D Cryo Electron Microscopy (3DEM) structures already represented in the global PDB archive were made publicly available in May 2016 (announcement). The wwPDB periodically recalculates statistical distributions as well as validation reports for all entries. All these data will be made publicly available to encourage downstream use by software developers, bioinformaticians and other PDB users. In addition, wwPDB expects to reconvene Validation Task Forces once every 5 years or so to assess the validation pipelines, reports and protocols and to recommend any changes, updates or additions.

4.2. Do all entries have a validation report?

There are a few structures determined by X-ray crystallography, 3D Cryo Electron Microscopy or NMR methods for which the validation report PDF or multi percentile slider assessments are missing. In addition, there are also cases where one of the components of the validation report has not successfully run and so is not available in the report and sliders. Periodically as part of the development process each failure is examined and improvements made to resolve as many as possible.

4.3. The percentile ranks for an entry have changed - why?

As the PDB archive continues to grow, we periodically recalculate the statistics underlying the percentile ranks. This process causes small changes in the percentile ranks of existing as new, normally “better” structures are added to the archive.

5. Validation report contents

5.1. What to do if you think a validation report for a structure gets things wrong.

Please let us know by emailing validation@mail.wwpdb.org providing as much detail as possible of the problem.

5.2. Why do REMARK 500 and the validation report differ?

Because different programs are used in preparing REMARK 500 and the validation report. In some cases the validation reports metrics are more up to date, for instance the Molprobity Ramachandran analysis is based on more data than the REMARK 500 analysis that uses the older Kleywegt and Jones (1996) study.

5.3. Why are there clashes reported between hydrogen atoms that are not present in the deposited model?

The MolProbity clashscore works by adding hydrogen atoms to the structure and then analyzing whether there are clashes. This is particularly useful in finding regions that could be improved in structures refined without explicit or riding hydrogen atoms (and less so in structures where hydrogen atoms are considered in refinement).

5.4. What to do about reported clashes?

If the structure has a poor clash score then this could indicate that:

- it is poorly built overall or
- there are regions that are poorly built or
- the refinement has not been allowed converge or
- the relative weighting of the geometry versus X-ray term in refinement has gone wrong.

The validation report listing of clashes is not that useful. The Coot program has a useful feature "Validate" "Probe clashes" that uses (and requires) the MolProbity reduce and probe programs. This allows visualization of where in the structure the clashes arise and might point out where some rebuilding could be indicated (see https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/web/docs/coot.html#Molprobity-Tools-Interface).

5.5. How can I view validation results in a molecular graphics program like Coot?

The most recent versions of the Coot program already has a facility to load the XML file produced as part of the validation process and provide a GUI that provides a more interactive way to click and be shown where outliers are found in a structure. Currently this facility is only available for released PDB entries loaded using the Coot menu option: *File, Fetch PDB using Accession Code*. We are working on a plugin to allow the procedure to work with validation XML files downloaded from the Validation Server or the OneDep deposition system.

6. Validation report contents: ligand geometry

6.1. How does PDB validate the geometry of ligand molecules?

The wwPDB validation report uses the CCDC Mogul program to validate the geometry of ligands against the Cambridge Structural Database (CSD) of small molecule organic structures. For each bond length, bond angle, torsion angle or ring in the ligands, Mogul identifies CSD structures that contain that feature and builds a distribution for the observed values. Engh & Huber showed how the CSD provides a good source for geometrical information for natural amino acids. Mogul facilitates a similar approach to be taken for ligands. For each bond, bond angle and torsion within a ligand Mogul identifies CSD small structures with a similar chemical environment and finds the distribution for the measure.

6.2. Why are there so many geometry outliers for my ligand?

There are a number of possible reasons. In some cases a poor method has been used to generate geometrical restraints for a ligand (see next FAQ). An erroneous fit to electron density or refinement with problematic X-ray weights can also lead to outliers. Outliers can also be caused by the Validation Pipeline Software providing an incorrect description of the chemistry of the molecule to the Mogul tool. This sometimes happens in preliminary validation reports but is rare for final validation reports where the PDB chemical components definition for the ligand is used as a basis for the chemical description.

6.3. What ligand geometry standards do PDB use and why are they different from REFMAC/PHENIX/BUSTER?

The wwPDB validation report uses Mogul for ligand geometry reports (see above). Good ways to derive ligand restraints including information from small molecule structures are:

- The ACEDRG program from CCP4 that uses an alternative small molecule structure open database COD.
- Phenix Elbow (particularly if you have an installed licensed Mogul program).
- Global Phasing Grade program (provided you have an installed licensed Mogul program).
- The Global Phasing Grade Web Server http://grade.globalphasing.org/. This uses Mogul but does not require installation/licenses (but it should be used only for non-confidential ligands).
- The CCP4 Pyrogen program (provided you have an installed licensed Mogul program).

The use of any of these programs should ensure that in the (new) validation report ligand geometry outliers should not arise from restraint issues.

6.4. Why are there inconsistencies in the ligand report between preliminary report generated at anonymous validation/deposition server and final report sent by PDB staff post annotation?

E.g., Outliers for ligands are not picked up at anonymous validation or deposition servers, but were picked up during PDB processing.

In order for Mogul to produce reliable results it is essential for the program to be provided with the correct bond configuration of the ligand. For the final report the PDB chemical components dictionary is used in which bond orders are defined.

Currently the preliminary validation report uses the CCDC program Gold_utils to find the connectivity and assign connectivity and bond orders from the user provided coordinates alone. This procedure normally works well but can go wrong particularly when provided with distorted ligand geometry. We are currently working on a number of improvements:

- To work with crystallographic software developers to include ligand chemical and restraint information in the mmCIF coordinate file used for deposition and validation. This will mean that reliable information will be available.
- To provide user feedback as to the chemistry provided to Mogul so that assignment problems can be quickly diagnosed. This would ideally be provided as 2D chemical diagrams in the report.
- If users provide a ligand with explicit hydrogen atoms the validator pipeline should use these in setting ligand bond orders. This should provide a potential work around solution that should work practically all the time.

6.5. Why are Mogul statistics not made available for me to analyze?

The XML file produced with the validation report includes additional Mogul information on bond and angle outliers. It would be necessary to obtain CCDC permission for the release of the full Mogul output for a ligand (because of data mining issues).

6.6. What is a Ligand Of Interest (LOI)?

A Ligand Of Interest (LOI) is a subject of the author's research. The validation report uses LOI information as selected by authors during deposition.

The ligands of Interest are defined in mmCIF in the category pdbx_entity_instance_feature

6.7. Why don't some of the ligands present in an entry have 2D geometrical quality and/or electron density fit images?

2D graphical depiction of geometrical quality analysis and/or electron density fit are provided for all instances of the ligands that have been designated as ligand of interest (LOI) by the depositor, regardless of the validation assessment, and any ligands with molecular weight greater than 250 Daltons that have outliers will be shown.

7. Validation report contents: X-ray specific

7.1. How does PDB calculate RSRZ?

RSRZ are calculated by the EDS (Electron-Density Server) component of the validation pipeline which is a re-implementation of the software used by the Uppsala EDS server (Kleywegt et al., 2004). The process is done by:

- Using the REFMAC program to calculate electron density maps based on the uploaded model and structure factors.
- The fit between the model and the 2Fo-Fc electron density map is found by calculating the real-space R-value (RSR). RSR is 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 (Jones et al., 1991). RSR is calculated using USF MAPMAN software tools (for a description see Tickle (2012)).
- The RSR Z-score (RSRZ) is a normalisation of RSR specific to a residue type and a resolution bin (Kleywegt et al., 2004). This means that RSRZ provides a comparison to the typical fit of a particular residue type for PDB structures at that resolution. RSRZ is calculated only for standard amino acids and nucleotides in protein, DNA and RNA chains.

7.2. Why does the validation report list RSRZ outliers when my examination of the refined electron density map shows that these residues have a good fit to density?

Currently the validation report assesses the fit to electron density using a map calculated using the REFMAC program as part of the EDS procedure (see previous FAQ). If the validation report lists residues as RSRZ outliers that in your examination of electron density maps have a good fit then this most likely indicates that the REFMAC did not calculate maps correctly. This happens occasionally and please accept our apologies for this. The procedure currently is known to have shown problems with:

- Lowish resolution structures that have been refined using Phenix where hydrogen atoms are involved.
- Low resolution anisotropic structures.
- Structures refined with Phenix twin option.

We are working on alternative procedures to improve the reliability of the fit to map validation.

7.3. Why is the EDS server not made available for me to re-calculate RSR?

The wwPDB Validation Service https://validate.wwpdb.org is a standalone server where the Validation Pipeline software can be run for any model or structures the user wants to examine. The same software is used as that used on deposition and this can be used to find RSR and RSRZ.

7.4. Why are calculated R factors different from what the author reported, especially for BUSTER?

The RCSB utility DCC is used to re-calculate R factors for an entry using the uploaded coordinates and structure factors. DCC uses either REFMAC, phenix-refine or CNS but currently does not support BUSTER. BUSTER R factors are not directly comparable to the other programs as it uses a different approach to the X-ray maximum likelihood calculation.

7.5. How do I convert calculated electron density map coefficients to MTZ?

The map coefficient cif files can be converted to MTZ using either

- CCP4 Convert the cif files to MTZ files using cif2mtz and then mtzutils to merge the two MTZ files together

The small shell script below will do this and the resulting MTZ file can be opened in Coot with the "Auto Open MTZ" option.

```
#!/usr/bin/env sh


FILE1=$1

FILE2=$2
```

```
OUTPUT_FILE=$3

if [ -z "$FILE1" -o -z "$FILE2" -o -z "$OUTPUT_FILE" ]
then
    echo "convert_to_mtz.sh IN_FILE1.cif IN_FILE2.cif OUT_FILE.mtz"
    exit 1
fi

FILE1_MTZ=$FILE1.mtz
FILE2_MTZ=$FILE2.mtz

echo "converting ${FILE1} to MTZ"

cif2mtz hklin $FILE1 hklout $FILE1_MTZ <<eof
END
eof

echo "converting ${FILE2} to MTZ"

cif2mtz hklin $FILE2 hklout $FILE2_MTZ <<eof
END
eof

echo "merging ${FILE1_MTZ} and ${FILE2_MTZ} to ${OUTPUT_FILE}"

mtzutils hklin1 $FILE1_MTZ hklin2 $FILE2_MTZ hklout $OUTPUT_FILE <<eof
EXCLUDE 2 FOM
```

```
END

eof
```

- Phenix cif_as_mtz. The MTZ files are then merged using CCP4's mtzutils.

The small shell script below will do this and the resulting MTZ file can be opened in Coot with the "Auto Open MTZ" option.

```
#!/usr/bin/env sh


FILE1=$1

FILE2=$2

OUTPUT_FILE=$3


if [ -z "$FILE1" -o -z "$FILE2" -o -z "$OUTPUT_FILE" ]

then

    echo "convert_to_mtz.sh IN_FILE1.cif IN_FILE2.cif OUT_FILE.mtz"

    exit 1

fi


FILE1_MTZ=$FILE1.mtz

FILE2_MTZ=$FILE2.mtz


echo "converting ${FILE1} to MTZ"


phenix.cif_as_mtz ${FILE1} --output_file_name=${FILE1_MTZ}


echo "converting ${FILE2} to MTZ"
```

```
phenix.cif_as_mtz ${FILE2} --output_file_name=${FILE2_MTZ}


echo "merging ${FILE1_MTZ} and ${FILE2_MTZ} to ${OUTPUT_FILE}"


mtzutils hklin1 $FILE1_MTZ hklin2 $FILE2_MTZ hklout $OUTPUT_FILE <<eof

EXCLUDE 2 FOM

END

eof
```

7.6. How do I use calculated electron density map coefficients in Coot?

Using the cif files is a two step process

1. convert the two cif files to an MTZ file (7.5)
2. in Coot use the "Auto open MTZ" option to open the output MTZ file(s)