CEITEC

Central European Institute of Technology

BRNO | CZECH REPUBLIC

UNIVERSITAS MASARYKIANA BRUNENSIS

Validation of biomacromolecular structures - motivation

Radka Svobodová

EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND
INVESTING IN YOUR FUTURE

2007–13
OP Research and Development for Innovation

Validation: Why to validate?

Structural biology community found that some published structures contained serious errors

Nightmare before Christmas

Retraction

WE WISH TO RETRACT OUR RESEARCH ARTICLE "STRUCTURE OF MsbA from *E. coli*: A homolog of the multidrug resistance ATP binding cassette (ABC) transporters" and both of our Reports "Structure of the ABC transporter MsbA in complex with ADP·vanadate and lipopolysaccharide" and "X-ray structure of the EmrE multidrug transporter in complex with a substrate" (*1*–*3*).

The recently reported structure of Sav1866 (*4*) indicated that our MsbA structures (*1*, *2*, *5*) were incorrect in both the hand of the structure and the topology. Thus, our biological interpretations based on these inverted models for MsbA are invalid.

An in-house data reduction program introduced a change in sign for anomalous differences. This program, which was not part of a conventional data processing package, converted the anomalous pairs (I+ and I–) to (F– and F+), thereby introducing a sign change. As the diffraction data collected for each set of MsbA crystals and for the EmrE crystals were processed with the same program, the structures reported in (*1*–*3*, *5*, *6*) had the wrong hand.

The error in the topology of the original MsbA structure was a consequence of the low resolution of the data as well as breaks in the electron density for the connecting loop regions. Unfortunately, the use of the multicopy refinement procedure still allowed us to obtain reasonable refinement values for the wrong structures.

The Protein Data Bank (PDB) files 1JSQ, 1PF4, and 1Z2R for MsbA and 1S7B and 2F2M for EmrE have been moved to the archive of obsolete PDB entries. The MsbA and EmrE structures will be recalculated from the original data using the proper sign for the anomalous differences, and the new Cα coordinates and structure factors will be deposited.

We very sincerely regret the confusion that these papers have caused and, in particular, subsequent research efforts that were unproductive as a result of our original findings.

GEOFFREY CHANG, CHRISTOPHER B. ROTH, CHRISTOPHER L. REYES, OWEN PORNILLOS, YEN-JU CHEN, ANDY P. CHEN

Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037, USA.

References

1. G. Chang, C. B. Roth, *Science* **293**, 1793 (2001).
2. C. L. Reyes, G. Chang, *Science* **308**, 1028 (2005).
3. O. Pornillos, Y.-J. Chen, A. P. Chen, G. Chang, *Science* **310**, 1950 (2005).
4. R. J. Dawson, K. P. Locher, *Nature* **443**, 180 (2006).
5. G. Chang, *J. Mol. Biol.* **330**, 419 (2003).
6. C. Ma, G. Chang, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 2852 (2004).

Validation: Why to validate?

Garbage in, garbage out

Interesting articles:

- Matthews, B. W. (2007) Five retracted structure reports: inverted or incorrect? Protein science : a publication of the Protein Society, 16, 1013–6.
- Johnston, C. A., Kimple, A. J., Giguere, P. M., and Siderovski, D. P. (2008) RETRACTED: Structure of the Parathyroid Hormone Receptor C Terminus Bound to the G-Protein Dimer Gb1g2. Structure, 16, 1086–1094.

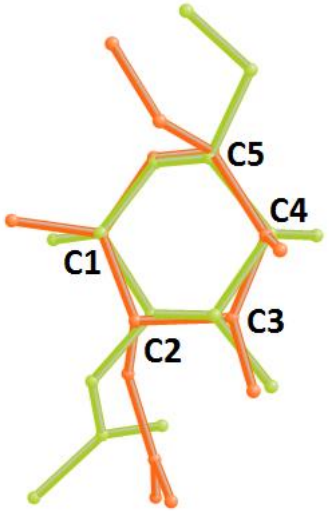
Can we still find wrong structures in PDB?

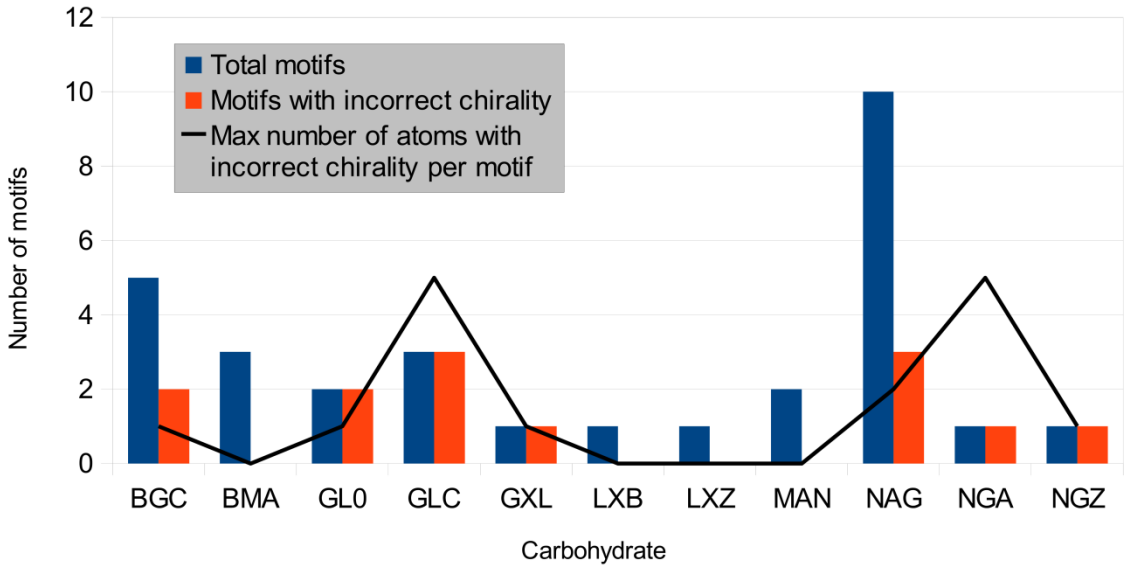
Can we still find wrong structures in PDB?

Example: Nipah G attachment glycoprotein (PDB ID 3D12, PNAS)
Contains 30 instances of 11 different carbohydrates, each with one ring and five chiral atoms.

Results:

- 13 of these ligands have incorrect chirality
- In a few cases, all chiral atoms exhibit incorrect chirality



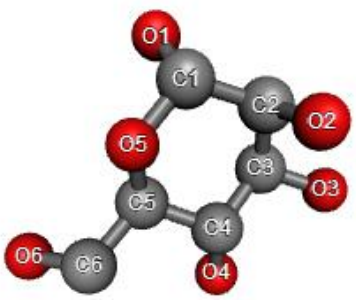


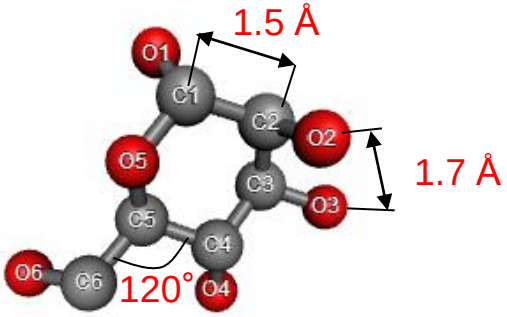
Can we still find wrong structures in PDB?

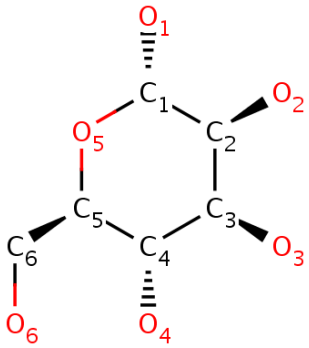
Unfortunately yes

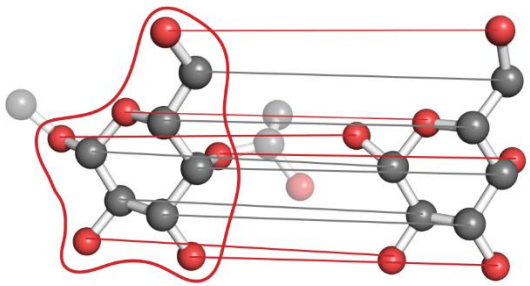
Validation approaches

What to validate?	How?	Software
Geometry (3D)	Against tabular values	**Proteins and nucl. acids:** WHAT_CHECK, PROCHECK, PROCHECK-NMR, AQUA, MolProbity, OOPS **Ligands:** ValLigURL, Mogul, Coot, PHENIX
Topology (2D)	Against a template	**Proteins and nucl. acids:** -||- **Ligands:** pdb-care, MotiveValidator, ValidatorDB







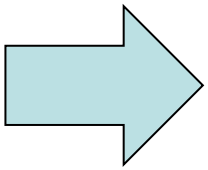


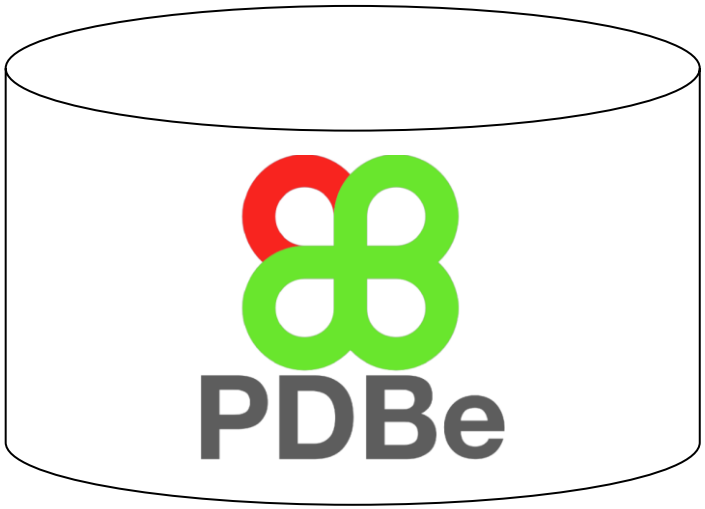
Compilation:
PDB validation reports

CEITEC

Improved data quality

- Cleaned up mmCIF data
  - Standard vocabularies
  - Experimental details, binding sites, secondary structure, antibiotic/inhibitor information, nucleic-acid parameters
- Clean mmCIF files are used in production





Validation reports

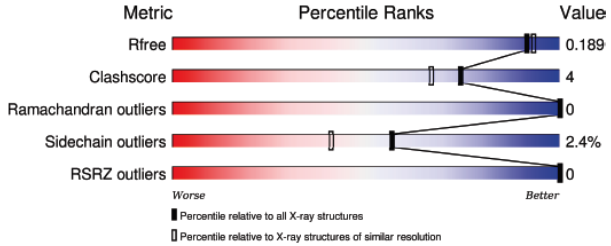
- Summary
  - Quality vs. all PDB
  - Quality vs. entries at similar resolution
  - Overview of residue-based quality for every polymer
  - Table of ligands that may need your attention

1 Overall quality at a glance

The following experimental techniques were used to determine the structure:
X-RAY DIFFRACTION

The reported resolution of this entry is 1.80 Å.

Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.



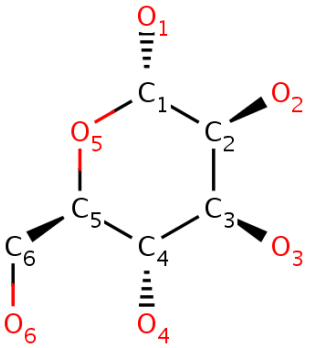
Metric	Whole archive (#Entries)	Similar resolution (#Entries, resolution range(Å))
R_{free}	91344	4533 (1.80-1.80)
Clashscore	102246	5383 (1.80-1.80)
Ramachandran outliers	100387	5320 (1.80-1.80)
Sidechain outliers	100360	5319 (1.80-1.80)
RSRZ outliers	91569	4547 (1.80-1.80)

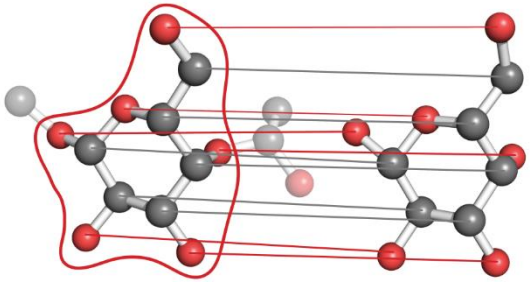
The table below summarises the geometric issues observed across the polymeric chains and their fit to the electron density. The red, orange, yellow and green segments on the lower bar indicate the fraction of residues that contain outliers for >=3, 2, 1 and 0 types of geometric quality criteria. A grey segment represents the fraction of residues that are not modelled. The numeric value for each fraction is indicated below the corresponding segment, with a dot representing fractions <=5% The upper red bar (where present) indicates the fraction of residues that have poor fit to the electron density. The numeric value is given above the bar.

Mol	Chain	Length	Quality of chain
1	A	137	91% 9%

Mol	Type	Chain	Res	Chirality	Geometry	Clashes	Electron density
1	PAQ	A	471	X	-	-	-
1	PAQ	B	471	X	-	-	-
1	PAQ	C	471	X	-	-	-
1	PAQ	D	471	X	-	-	-
6	CU	A	1744	-	-	-	X
6	CU	B	1748	-	-	-	X

Topology validation
(= validation of annotation)

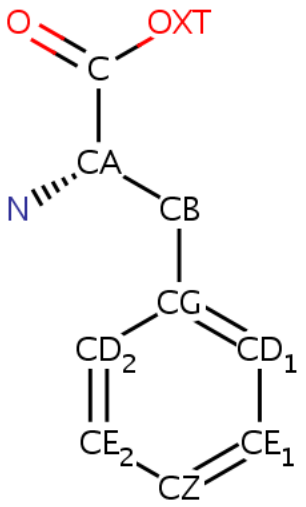




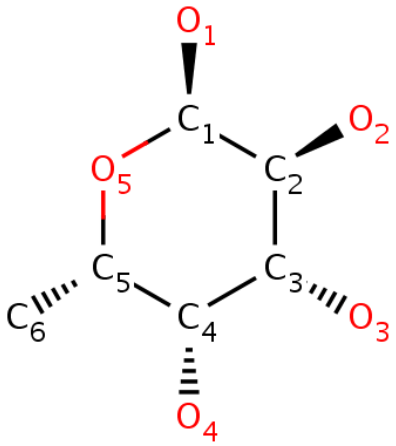
Topology validation – basic terms

Residue

- Any component of a biomacromolecule
- Examples: amino acids, nucleotides, saccharides, ions, …
- In PDB, residue is annotated via “Residue ID” – a unique 3-letter code



PHE, phenylalanine



FUC, *alpha-L-fucose*

CEITEC

Topology validation – basic terms

Types of residues:

Standard residues:

- amino acids
- nucleotides

Non-standard residues:

- modified amino acids and nucleotides

Ligands:

- Chemical compounds which form a complex with a biomacromolecule (e.g., sugar, drug, heme).
- Also ions are often referred as ligands

Topology validation – basic terms

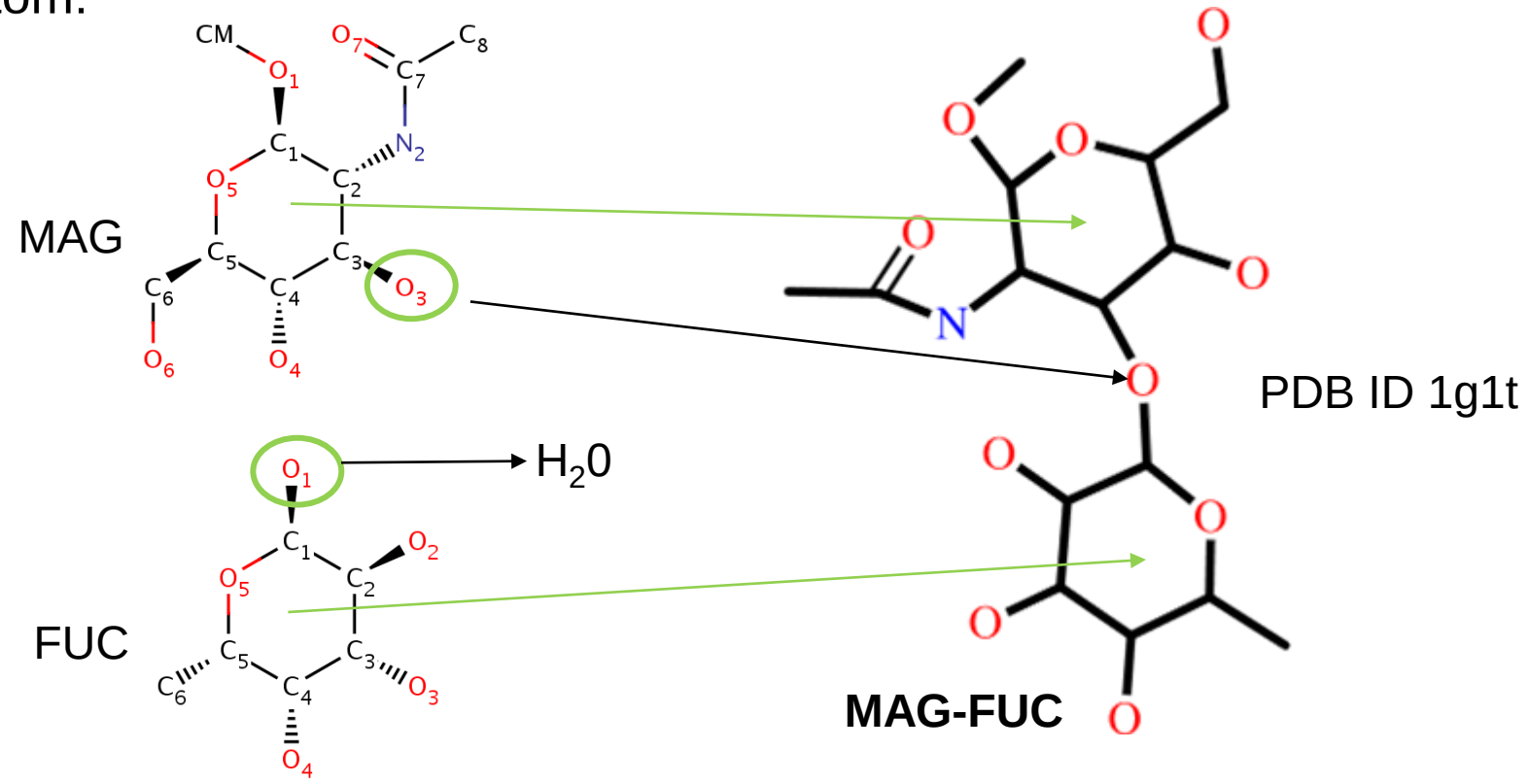
Principles of topology validation:

- Subjects of topology validation are residues
- Validated residue is compared with a model residue, which has the same Residue ID
- The model residues are taken from a reference database
- Differences between the model residue and the validated residue are reported

Topology validation – approach

Complication with “shared atoms”:

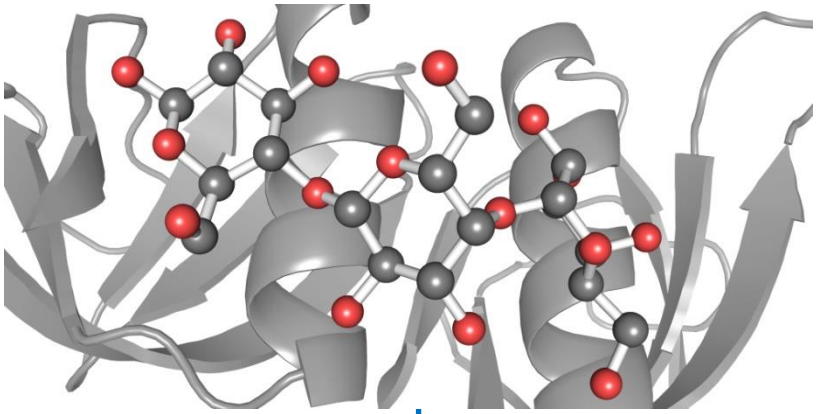
- When some residues bind together, one of them can lose an atom:



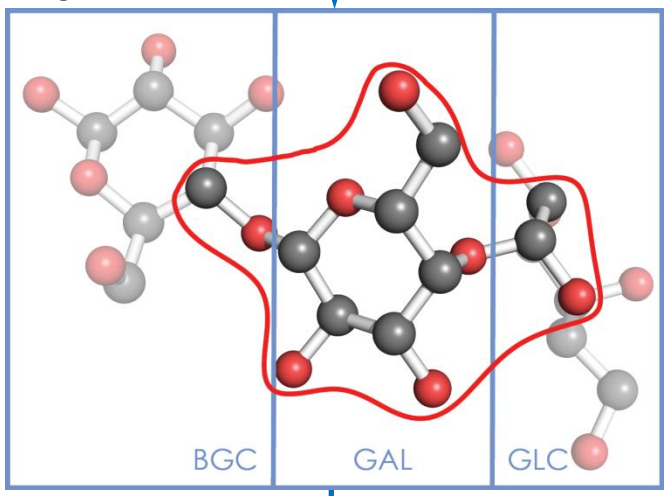
- **Solution:** When we validate a residue, we must include also its close surrounding

Topology validation - approach

Input PDB entry

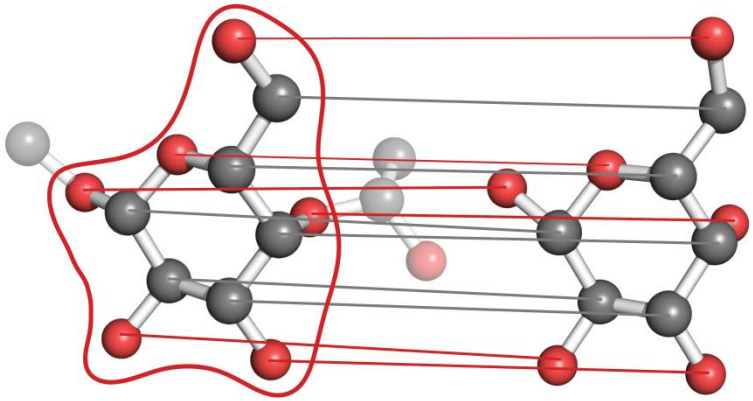


Selection of **validated residue** and its close surrounding



Input motif = validated residue + surrounding

Mapping of **input motif** to the **model residue** (via subgraph matching)



Validated motif mapped to the **Model Residue**

CEITEC

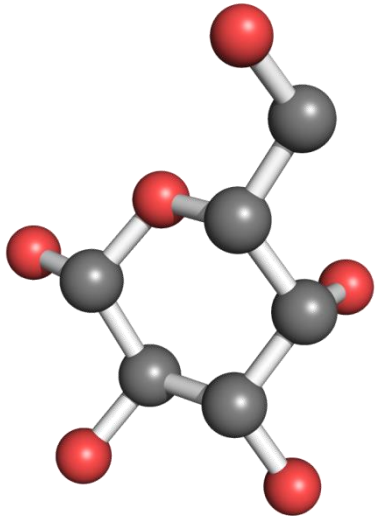
Topology validation – types of validation analyses

- **Completeness analyses**
  - Missing atoms
  - Missing rings
- **Chirality analyses**
  - Chirality on C atom, metal atom, high bond order atom, planarity
- **Advanced analyses**
  - Substitution
  - Different atom naming
  - Foreign atoms

Topology validation – types of validation analyses

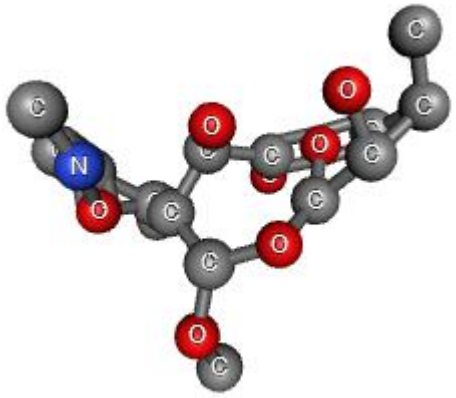
Prerequisite: Can we map the validated and the model residue?

Correct residue



Degenerated structure

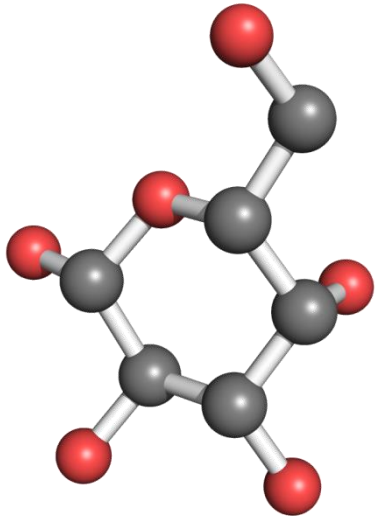
1IVG_17_7716 (MAN)



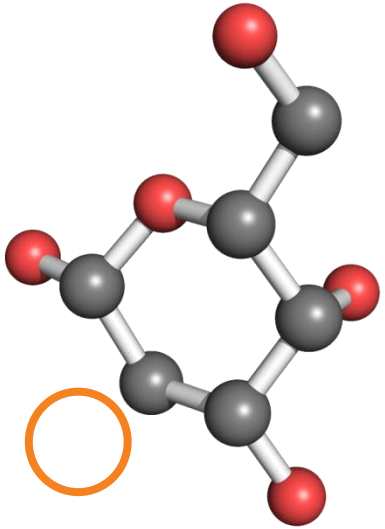
Topology validation – types of validation analyses

ERRORS – INCOMPLETE STRUCTURE

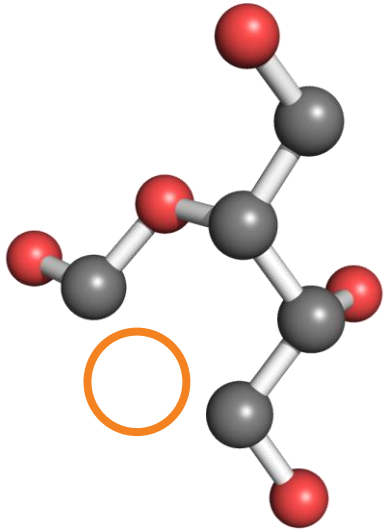
Correct residue



Missing atom



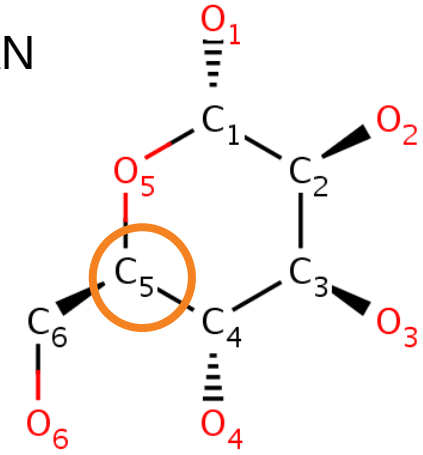
Missing ring

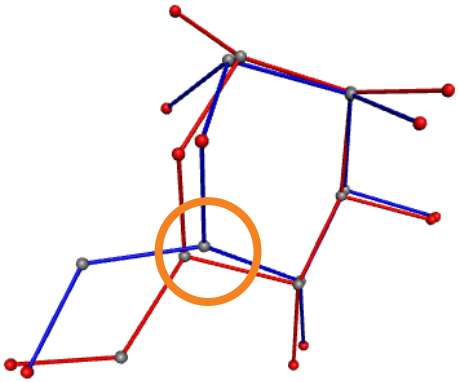


Topology validation – types of validation analyses
ERRORS – CHIRALITY

Wrong chirality on C atoms

MAN

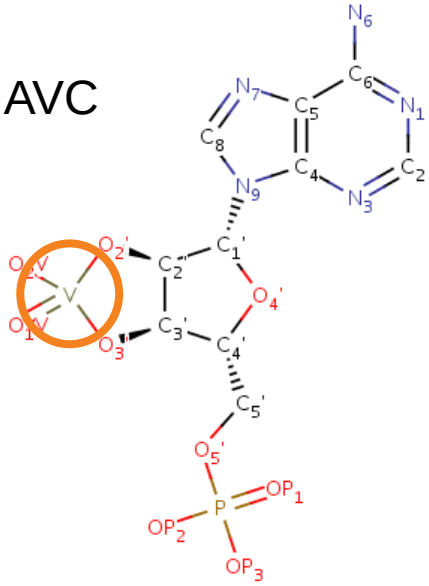


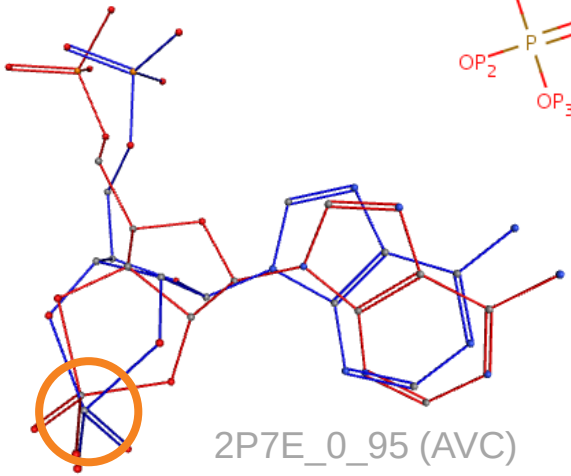


1E4M_16_4280 (MAN)

Wrong chirality on metal atom

AVC



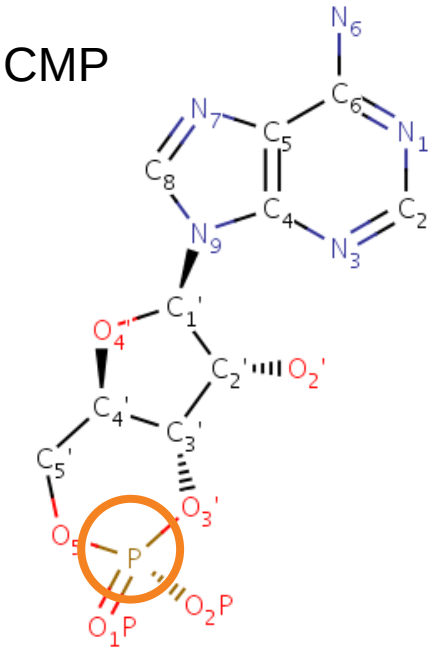


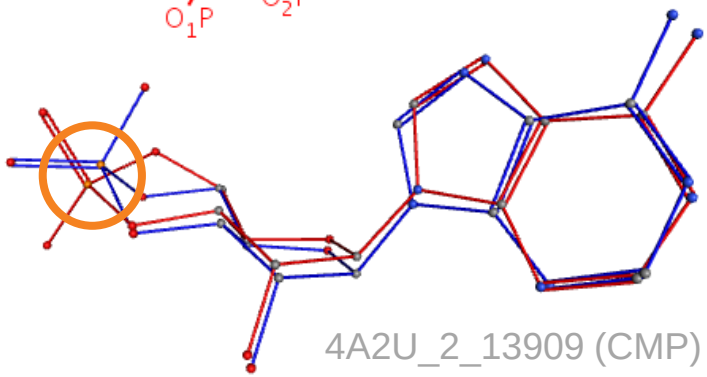
2P7E_0_95 (AVC)

Topology validation – types of validation analyses
ERRORS – CHIRALITY II

Wrong chirality on atom having high order bonds

CMP

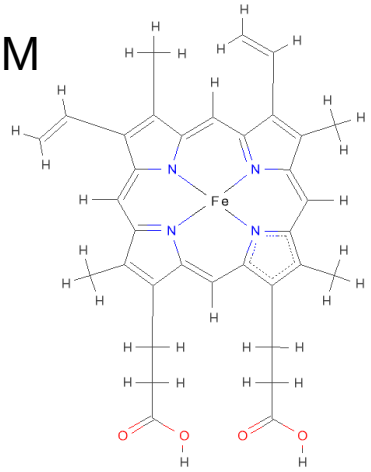


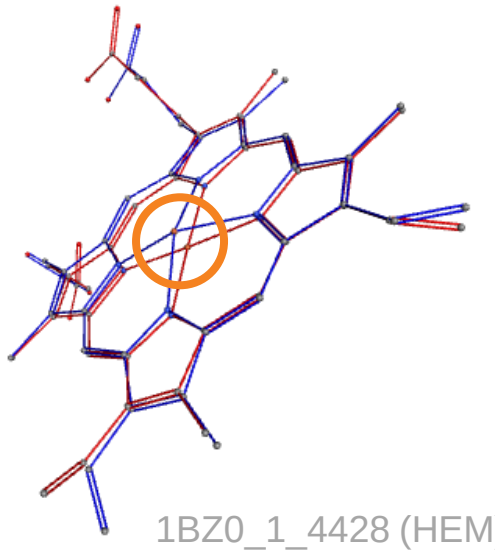


4A2U_2_13909 (CMP)

Wrong chirality (planar)

HEM





1BZ0_1_4428 (HEM)

Correct chirality (Tolerant)

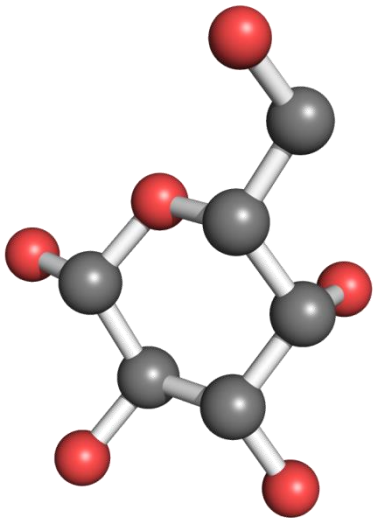
=

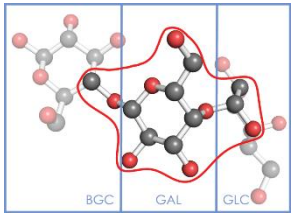
Correct residues + residues having only these issues:

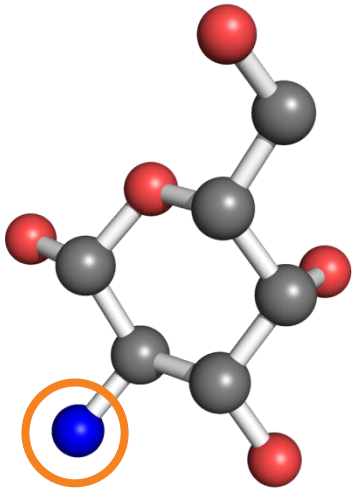
- **Wrong chirality (planar)**
- **Wrong chirality on atom having high order bonds**

Topology validation – types of validation analyses
WARNINGS

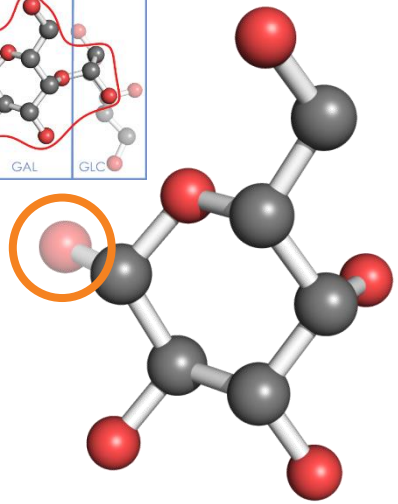
Correct residue





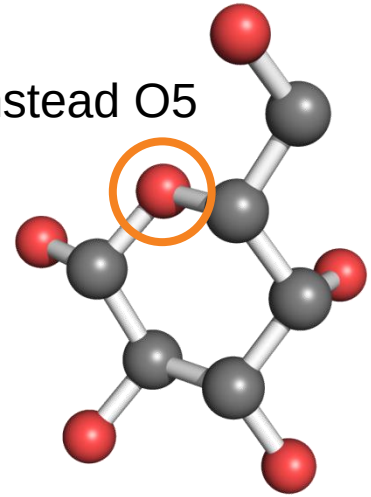


Substitution



Foreign atom
(= atom from neighboring residue)

O1 instead O5



Different atom name

ADVANCED WARNINGS:

Alternate locations

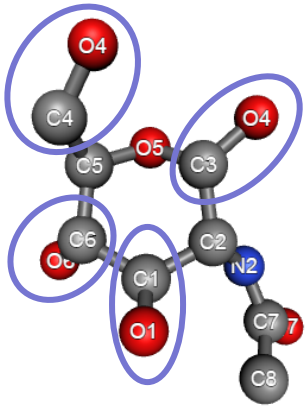
Zero RMSD with model

CEITEC

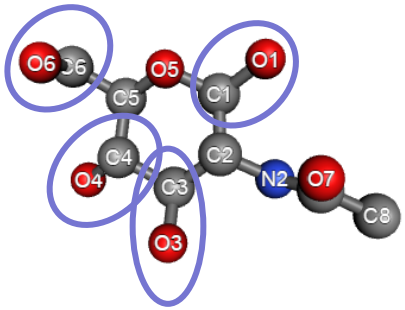
Different atom names - example

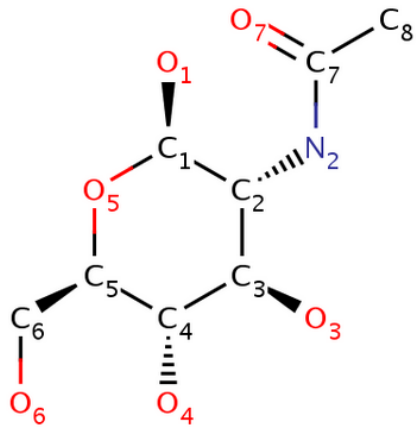
4ACQ_37_39697 (NAG)

Validated Motif



Model





Different Atom Names 7

Model	Motif	Type
C1	C3 C 39699	N N'
C3	C1 C 39697	N N'
C4	C6 C 39702	N N'
C6	C4 C 39700	N N'
O3	O1 O 39706	N N'
O4	O6 O 39709	N E'
O6	O4 O 39707	N E'

Topology validation
Software tools

PDB care (Lütteke et al., 2006):

- Tool focused on carbohydrates validation
- First application, which implements topology validation
- Performs basic validation analyses (missing atoms, missing rings, wrong chirality)

MotiveValidator (http://ncbr.muni.cz/MotiveValidator) :

- Tool, which allows validation of all residues
- Performs basic validation analyses + reports basic warnings (substitutions, foreign atoms, different naming)

ValidatorDB (http://ncbr.muni.cz/ValidatorDB):

- Database, containing validation results for all* ligands and non-standard residues in PDB (weekly updated)
- Performs basic validation analyses + advanced validation analyses (report degenerated residue, distinguish type of chirality error)
- Reports basic warnings + next warnings (Alternate locations, Zero RMSD with model)

* Except amino acids, nucleotides, and small residues (<7 heavy atoms)

Thank you for your attention