Deposition and validation of macromolecular structures at wwPDB

Sameer Velankar, sameer@ebi.ac.uk

PDBe

EMBL-EBI

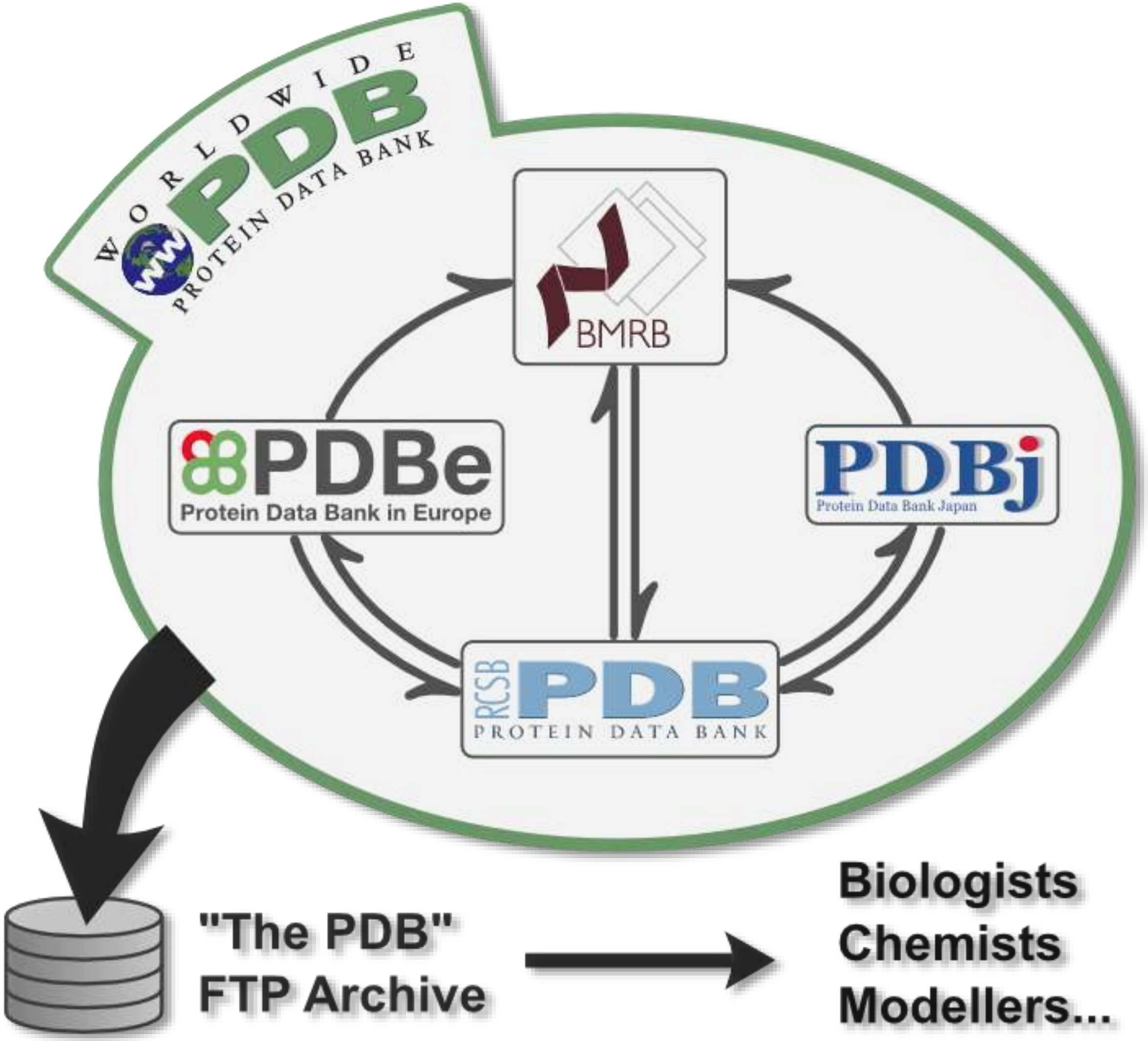
PDBe at a glance

Protein Data Bank in Europe

Bringing Structure to Biology

- Mission: ***Bringing Structure to Biology***
- Founding partner of Worldwide Protein Data Bank (wwPDB)
- Founder of Electron Microscopy Data Bank (EMDB)
- Major activities:
  - Deposition and annotation site for structural data on biomacromolecules (X-ray, NMR, EM)
  - Integrated resource to serve structural data and information
  - Liaise with structural biology community
- Guided by advisory bodies
  - PDBe, wwPDB, EMDataBank

wwPDB


WORLDWIDE PDB PROTEIN DATA BANK
BMRB
PDBe Protein Data Bank in Europe
PDBj Protein Data Bank Japan
RCSB PDB PROTEIN DATA BANK
"The PDB" FTP Archive
Biologists
Chemists
Modellers...

wwPDB partnership

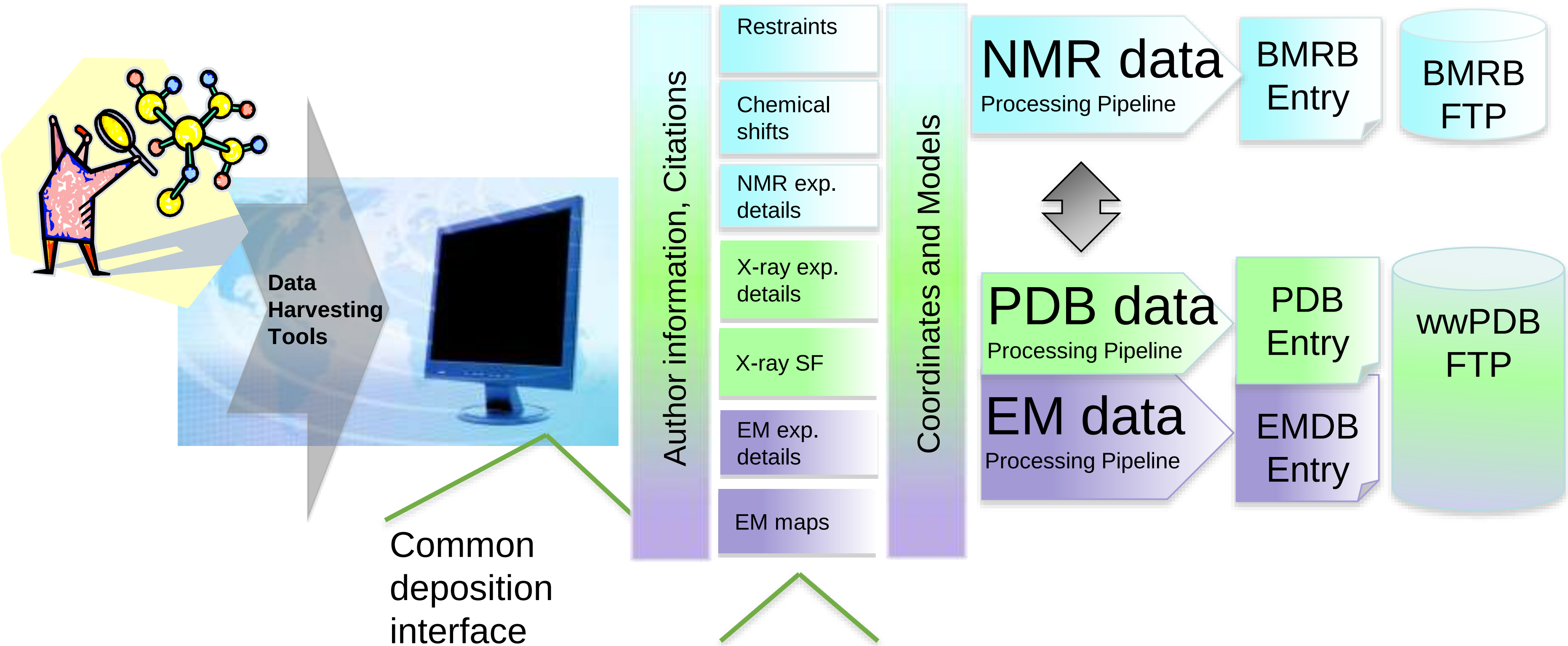
- Collaborate on “*data in*”
  - Policy issues
  - Weekly releases
  - Validation standards
  - Format specifications
  - Chemical component database
  - Deposition and annotation procedures
  - Archive quality and remediation
  - Journal interactions
  - Community interactions
- Friendly competition on “*data out*”
  - Serving PDB data with added-value
  - PDB-based services
  - Other services, resources and activities

WORLDWIDE wwPDB PROTEIN DATA BANK

Two major recent wwPDB projects

- Development of a new joint wwPDB Deposition and Annotation (D&A) system
  - Will handle X-ray, NMR, EM, …
  - Will be used at all wwPDB sites
  - Replaces Adit, AutoDep, EMdep, parts of Adit-NMR
- Validation using community-recommended methods will be integral part of new D&A
  - 2008: X-ray Validation Task Force (VTF)
  - 2009: NMR VTF
  - 2010: EM VTF
  - wwPDB validation pipelines being developed (at PDBe) for **X-ray**, **NMR** and **EM** based on VTF guidelines

wwPDB D&A tool



Data Harvesting Tools

Author information, Citations

Restraints

Chemical shifts

NMR exp. details

X-ray exp. details

X-ray SF

EM exp. details

EM maps

Coordinates and Models

NMR data Processing Pipeline → BMRB Entry → BMRB FTP

PDB data Processing Pipeline → PDB Entry

EM data Processing Pipeline → EMDB Entry

wwPDB FTP

Common deposition interface
- Accession ID
- Validation report

Integrated data capture

Changes for depositors

- Deposit to wwPDB - one place
  - wwPDB to define rules where deposition and annotation takes place (previous deposition, geography, load balancing, …)
  - Same deposition system at all sites
  - Will issue PDB, EMDB and BMRB identifiers

- Access to deposition interface via login
  - All communication will be through this interface
  - Access to all previous depositions
  - No more e-mailing of replacement coordinate sets

Changes for depositors

- More functionality on deposition side
  - Validation of model and data before submission
  - Added value (*e.g.*, quaternary structure) before submission
  - Additional checks for data integrity – e.g. Taxonomy and sequence information

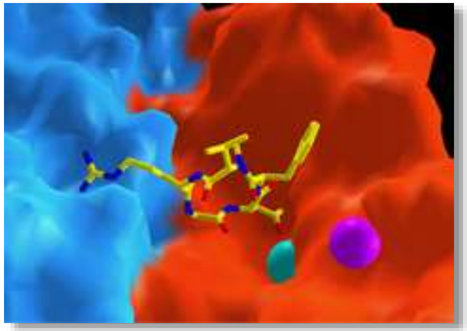
Validation by wwPDB - advantages

- Improves the quality and consistency of the PDB archive
- Supports editors and referees
- Helps users assess if an entry is suitable
- Helps users compare related entries
- Enables identification of outliers when mining the PDB
- Stimulates adoption of better protocols by the community

Errors in Structures

- **Completely wrong**
  - Wrong trace, incorrect fold of protein
  - Register errors, where trace of protein is not in keeping with sequence order.
- **Partial errors**
  - Incorrectly built loops.
  - Wrong residues built into the structure (i.e., Proline instead of Aspartic acid).
- **Bad quality**
  - Bad geometry and stereochemistry.
  - Incorrect positioning of ligands etc due to lack of experimental evidence.
- **FRAUD !!**

Much can go wrong with ligands…

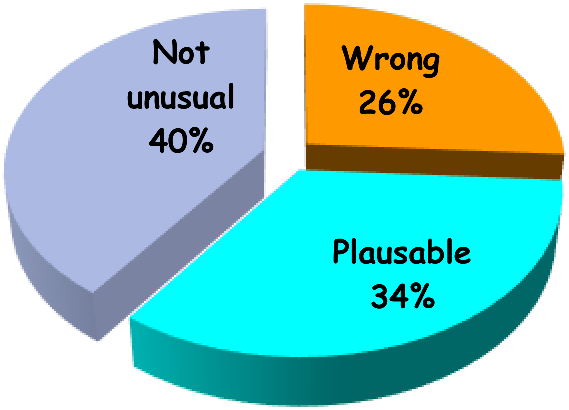


- Nowadays often problems with ligands
  - Building the wrong ligand
  - Building expected ligands without credible density
  - Mistaking an unexpected ligand for waters
  - Fitting a ligand incorrectly in ambiguous density
  - Errors in ligand conformation
  - Errors in ligand geometry or stereochemistry
- Protein crystallography is not a good method for determining the structure of small molecules
  - But for ligand design it is crucial that the ligand and the binding-site residues are modelled reliably

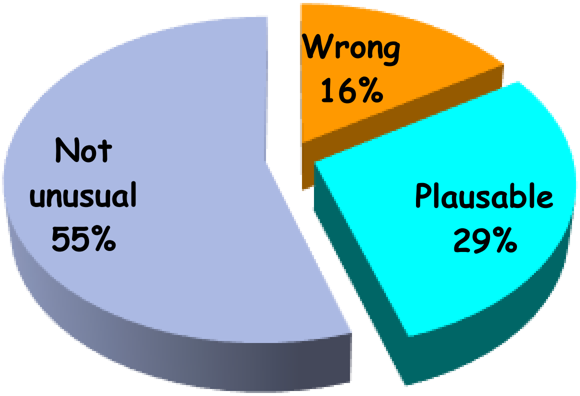
Kleywegt, *Acta Cryst.* **D63**, 94 (2007)

Validation of PDB ligand structures by CCDC

- 16% of PDB entries deposited in 2006 had ligand geometries that were almost certainly significantly in error *(in-house analysis using Relibase+/Mogul)*
- The good news - for structures before 2000 the figure was 26%



Pre 2000



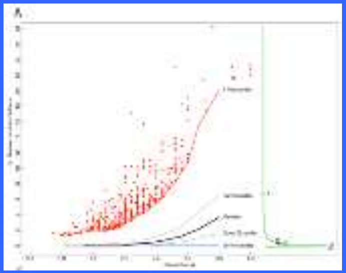
2006

See also: Liebeschuetz *et al.*, *JCAMD* **26**, 169 (2012)

wwPDB X-ray Validation Pipeline

External reference files (*e.g.*, Engh & Huber)

Distributions



Validation information

Validation pipeline 1.0

MolProbity

Xtriage

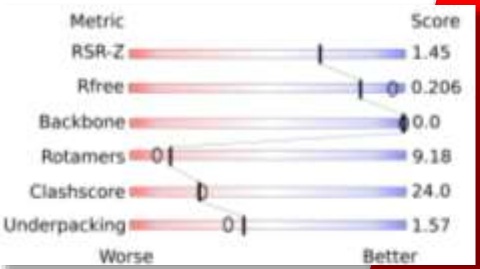
EDS

Mogul

Percentiles

PDF maker

Deposited data (coordinates & reflections)



PDF report for depositor & referees - Statistics and plots for the entry, per chain, per residue, and list of unusual features

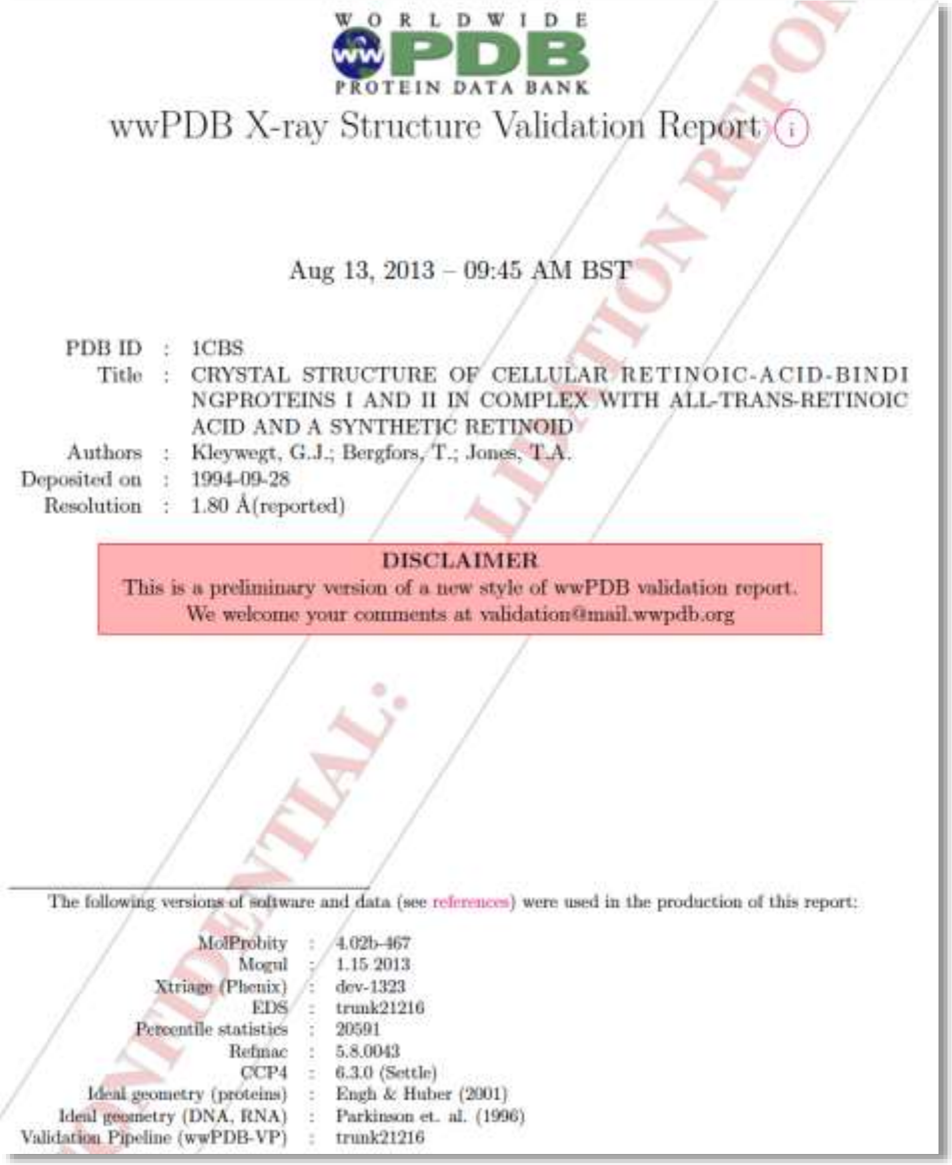
Read *et al.*, *Structure* **19**, 1395 (2011)
Gore *et al.*, *Acta Cryst.* **D68**, 478 (2012)

wwPDB X-ray validation pipeline

- Version 1.0 essentially ready for production use
  - X-ray validation pipeline – reports are sent out to depositors from 1 August
  - stand-alone server from September
- After version 1.0:
  - WhatCheck (e.g., DACA, unusual backbone)?
  - pdbcare (carbohydrates)?
  - LabelIt (spacegroup errors)?
  - DDQ (e.g., uninterpreted density)?
  - Ligand summary? Better real-space fit criterion for non-polymers? Figures of ligand plus binding site plus density?

Validation reports

- Front cover
    - Deposition info
    - Software info



WORLDWIDE PDB PROTEIN DATA BANK

wwPDB X-ray Structure Validation Report

Aug 13, 2013 – 09:45 AM BST

PDB ID	:	1CBS
Title	:	CRYSTAL STRUCTURE OF CELLULAR RETINOIC-ACID-BINDINGPROTEINS I AND II IN COMPLEX WITH ALL-TRANS-RETINOIC ACID AND A SYNTHETIC RETINOID
Authors	:	Kleywegt, G.J.; Bergfors, T.; Jones, T.A.
Deposited on	:	1994-09-28
Resolution	:	1.80 Å(reported)

DISCLAIMER
This is a preliminary version of a new style of wwPDB validation report.
We welcome your comments at validation@mail.wwpdb.org

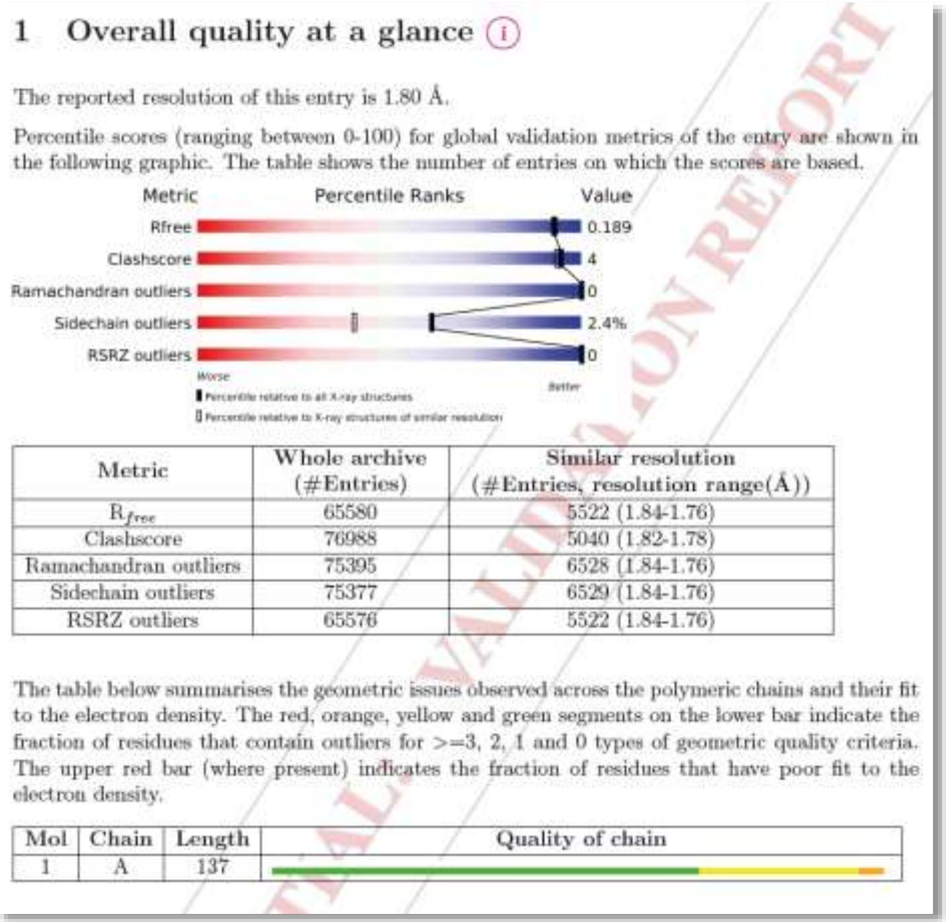
The following versions of software and data (see references) were used in the production of this report:

MolProbity	:	4.02b-467
Mogul	:	1.15 2013
Xtriage (Phenix)	:	dev-1323
EDS	:	trunk21216
Percentile statistics	:	20591
Refmac	:	5.8.0043
CCP4	:	6.3.0 (Settle)
Ideal geometry (proteins)	:	Engh & Huber (2001)
Ideal geometry (DNA, RNA)	:	Parkinson et. al. (1996)
Validation Pipeline (wwPDB-VP)	:	trunk21216

Validation reports

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

http://wwpdb.org/validation.html



1 Overall quality at a glance ⓘ

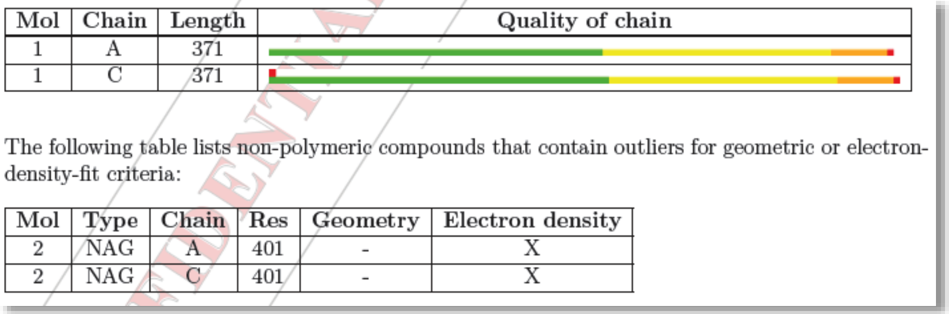
The reported resolution of this entry is 1.80 Å.

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}	65580	5522 (1.84-1.76)
Clashscore	76988	5040 (1.82-1.78)
Ramachandran outliers	75395	6528 (1.84-1.76)
Sidechain outliers	75377	6529 (1.84-1.76)
RSRZ outliers	65576	5522 (1.84-1.76)

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. The upper red bar (where present) indicates the fraction of residues that have poor fit to the electron density.

Mol	Chain	Length	Quality of chain
1	A	137	



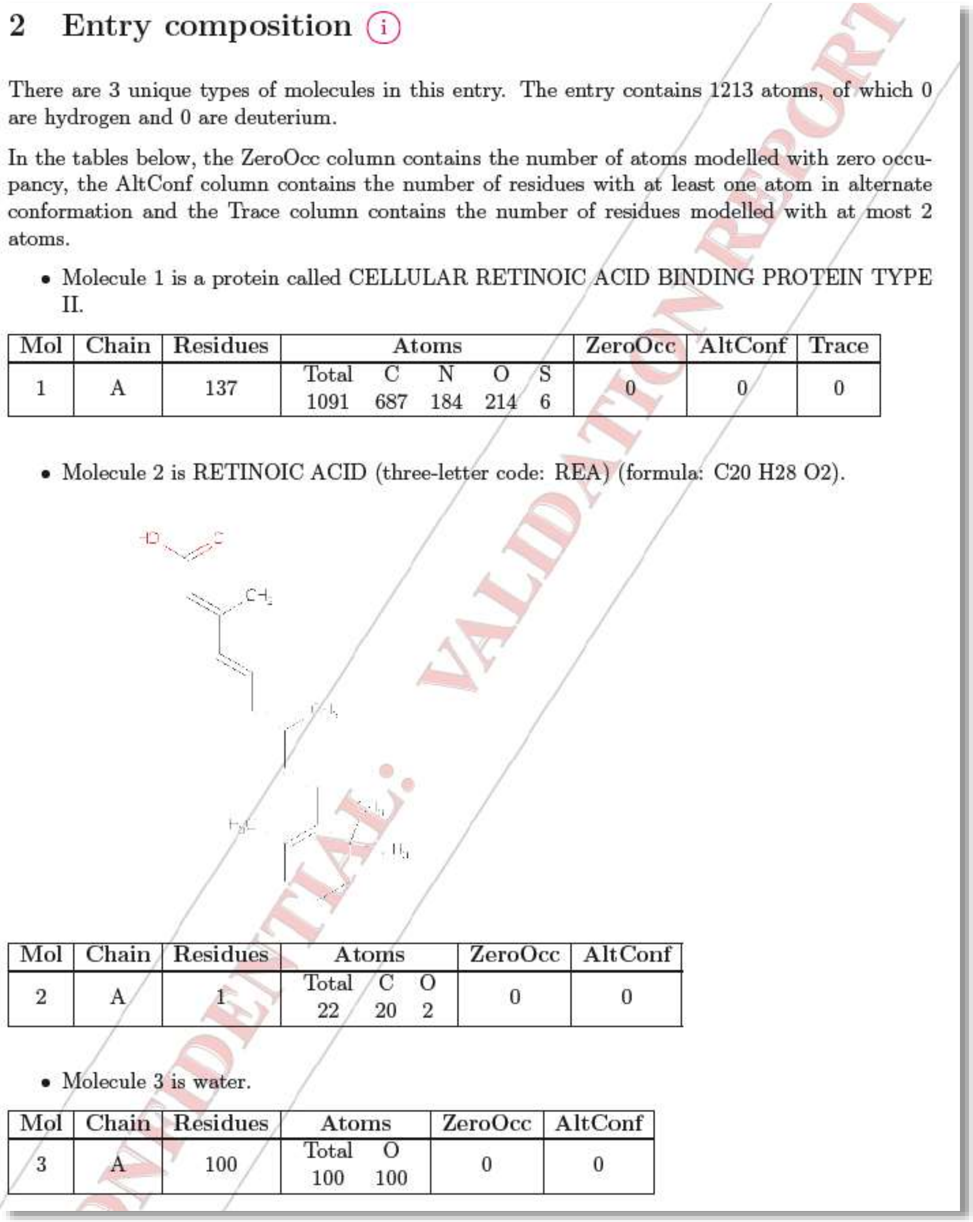
Mol	Chain	Length	Quality of chain
1	A	371	
1	C	371	

The following table lists non-polymeric compounds that contain outliers for geometric or electron-density-fit criteria:

Mol	Type	Chain	Res	Geometry	Electron density
2	NAG	A	401	-	X
2	NAG	C	401	-	X

Validation reports

- Entry contents
  - Inventory



2 Entry composition ⓘ

There are 3 unique types of molecules in this entry. The entry contains 1213 atoms, of which 0 are hydrogen and 0 are deuterium.

In the tables below, the ZeroOcc column contains the number of atoms modelled with zero occupancy, the AltConf column contains the number of residues with at least one atom in alternate conformation and the Trace column contains the number of residues modelled with at most 2 atoms.

- Molecule 1 is a protein called CELLULAR RETINOIC ACID BINDING PROTEIN TYPE II.

Mol	Chain	Residues	Atoms					ZeroOcc	AltConf	Trace
			Total	C	N	O	S			
1	A	137	1091	687	184	214	6	0	0	0

- Molecule 2 is RETINOIC ACID (three-letter code: REA) (formula: $C_{20}H_{28}O_2$).

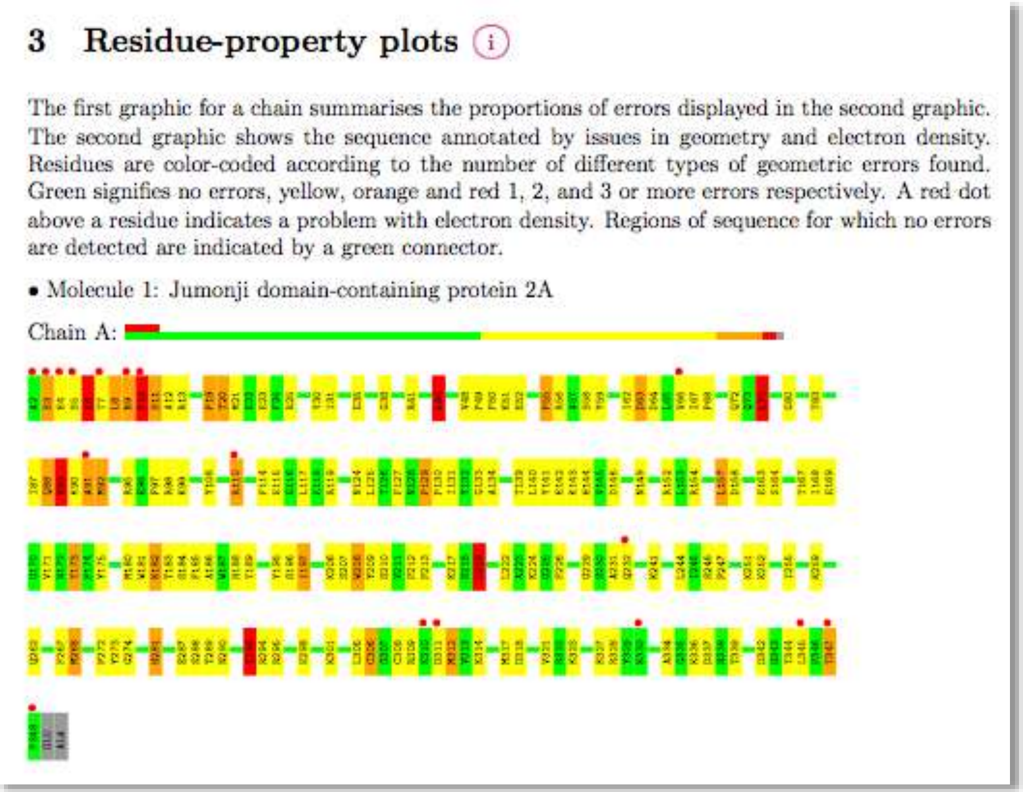
Mol	Chain	Residues	Atoms			ZeroOcc	AltConf
			Total	C	O		
2	A	1	22	20	2	0	0

- Molecule 3 is water.

Mol	Chain	Residues	Atoms		ZeroOcc	AltConf
			Total	O		
3	A	100	100	100	0	0

Validation reports

- Residue quality
  - One plot per polymer
  - Coloured by number of ***types*** of geometric outliers
  - Grey if not modelled
  - Red dots: poor density (RSR-Z > 2, as in EDS)



3 Residue-property plots ⓘ

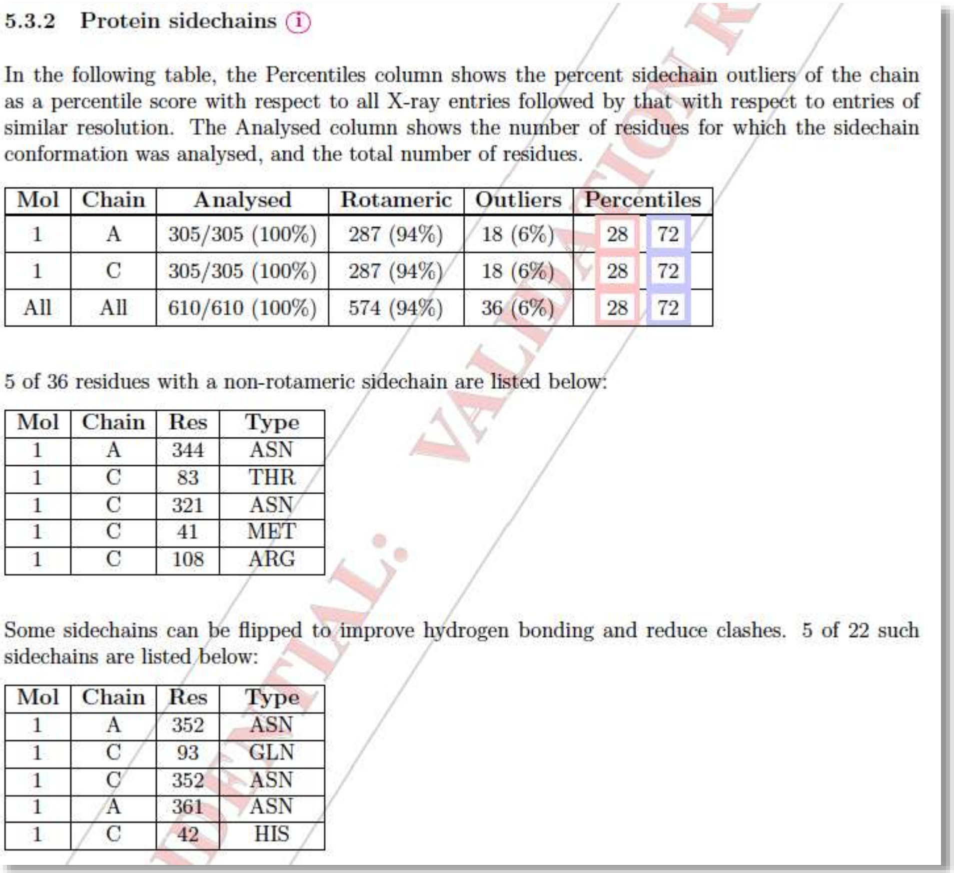
The first graphic for a chain summarises the proportions of errors displayed in the second graphic. The second graphic shows the sequence annotated by issues in geometry and electron density. Residues are color-coded according to the number of different types of geometric errors found. Green signifies no errors, yellow, orange and red 1, 2, and 3 or more errors respectively. A red dot above a residue indicates a problem with electron density. Regions of sequence for which no errors are detected are indicated by a green connector.

- Molecule 1: Jumonji domain-containing protein 2A

Chain A:

Validation reports

- Model quality
  - Bond lengths and angles
  - Torsion angles (Ramachandran, rotamers)
  - Clashes
  - Separately for standard residues, non-standard residues, ligands, carbohydrates
  - Generally: information about distribution, outlier stats, percentile scores, list of up to 5 (worst) outliers



5.3.2 Protein sidechains ⓘ

In the following table, the Percentiles column shows the percent sidechain outliers of the chain as a percentile score with respect to all X-ray entries followed by that with respect to entries of similar resolution. The Analysed column shows the number of residues for which the sidechain conformation was analysed, and the total number of residues.

Mol	Chain	Analysed	Rotameric	Outliers	Percentiles	
1	A	305/305 (100%)	287 (94%)	18 (6%)	28	72
1	C	305/305 (100%)	287 (94%)	18 (6%)	28	72
All	All	610/610 (100%)	574 (94%)	36 (6%)	28	72

5 of 36 residues with a non-rotameric sidechain are listed below:

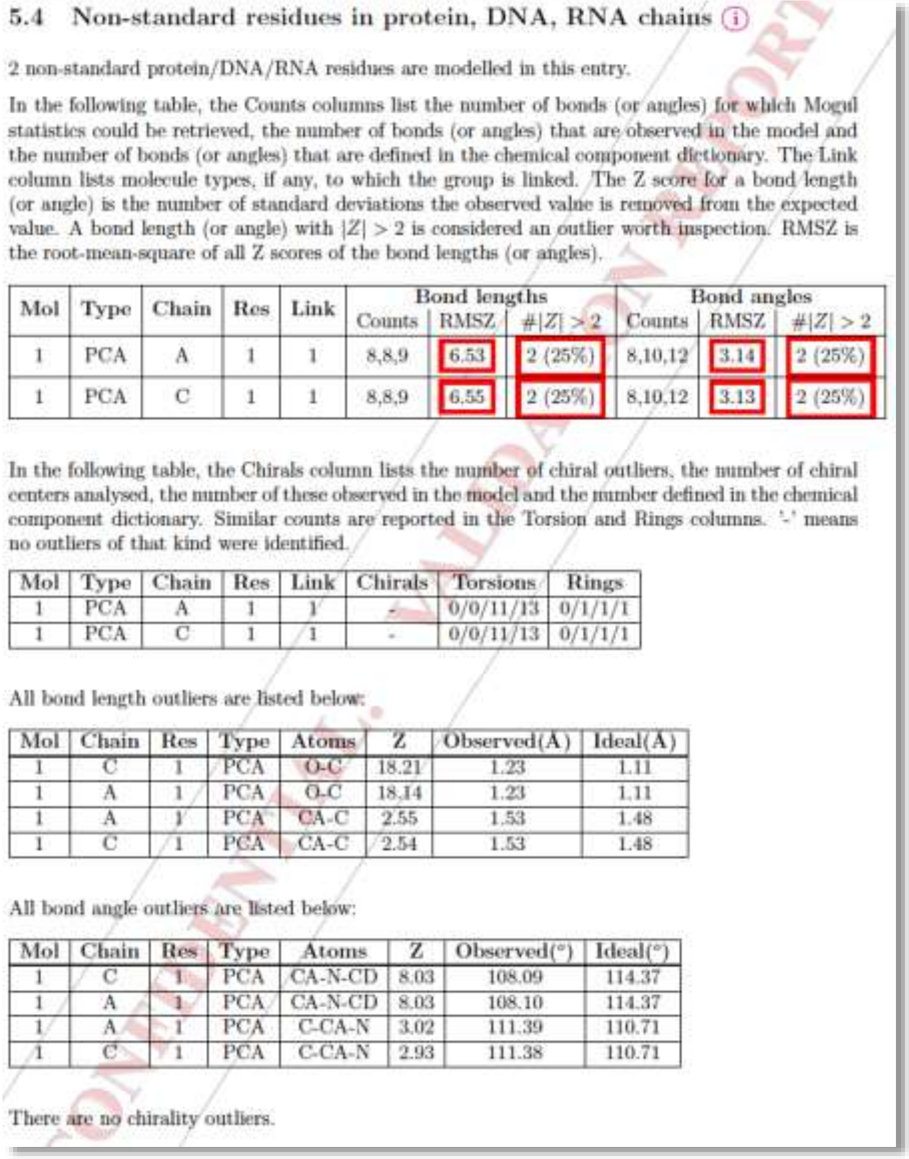
Mol	Chain	Res	Type
1	A	344	ASN
1	C	83	THR
1	C	321	ASN
1	C	41	MET
1	C	108	ARG

Some sidechains can be flipped to improve hydrogen bonding and reduce clashes. 5 of 22 such sidechains are listed below:

Mol	Chain	Res	Type
1	A	352	ASN
1	C	93	GLN
1	C	352	ASN
1	A	361	ASN
1	C	42	HIS

Validation reports

- Geometry validation of ligands and non-standard entities
  - Mogul (CCDC)
- wwPDB will get CSD coordinates for new and existing compounds



5.4 Non-standard residues in protein, DNA, RNA chains ⓘ

2 non-standard protein/DNA/RNA residues are modelled in this entry.

In the following table, the Counts columns list the number of bonds (or angles) for which Mogul statistics could be retrieved, the number of bonds (or angles) that are observed in the model and the number of bonds (or angles) that are defined in the chemical component dictionary. The Link column lists molecule types, if any, to which the group is linked. The Z score for a bond length (or angle) is the number of standard deviations the observed value is removed from the expected value. A bond length (or angle) with $|Z| > 2$ is considered an outlier worth inspection. RMSZ is the root-mean-square of all Z scores of the bond lengths (or angles).

Mol	Type	Chain	Res	Link	Bond lengths			Bond angles		
					Counts	RMSZ	#$\|Z\| > 2$	Counts	RMSZ	#$\|Z\| > 2$
1	PCA	A	1	1	8,8,9	6.53	2 (25%)	8,10,12	3.14	2 (25%)
1	PCA	C	1	1	8,8,9	6.55	2 (25%)	8,10,12	3.13	2 (25%)

In the following table, the Chirals column lists the number of chiral outliers, the number of chiral centers analysed, the number of these observed in the model and the number defined in the chemical component dictionary. Similar counts are reported in the Torsion and Rings columns. '-' means no outliers of that kind were identified.

Mol	Type	Chain	Res	Link	Chirals	Torsions	Rings
1	PCA	A	1	1	-	0/0/11/13	0/1/1/1
1	PCA	C	1	1	-	0/0/11/13	0/1/1/1

All bond length outliers are listed below:

Mol	Chain	Res	Type	Atoms	Z	Observed(Å)	Ideal(Å)
1	C	1	PCA	O-C	18.21	1.23	1.11
1	A	1	PCA	O-C	18.14	1.23	1.11
1	A	1	PCA	CA-C	2.55	1.53	1.48
1	C	1	PCA	CA-C	2.54	1.53	1.48

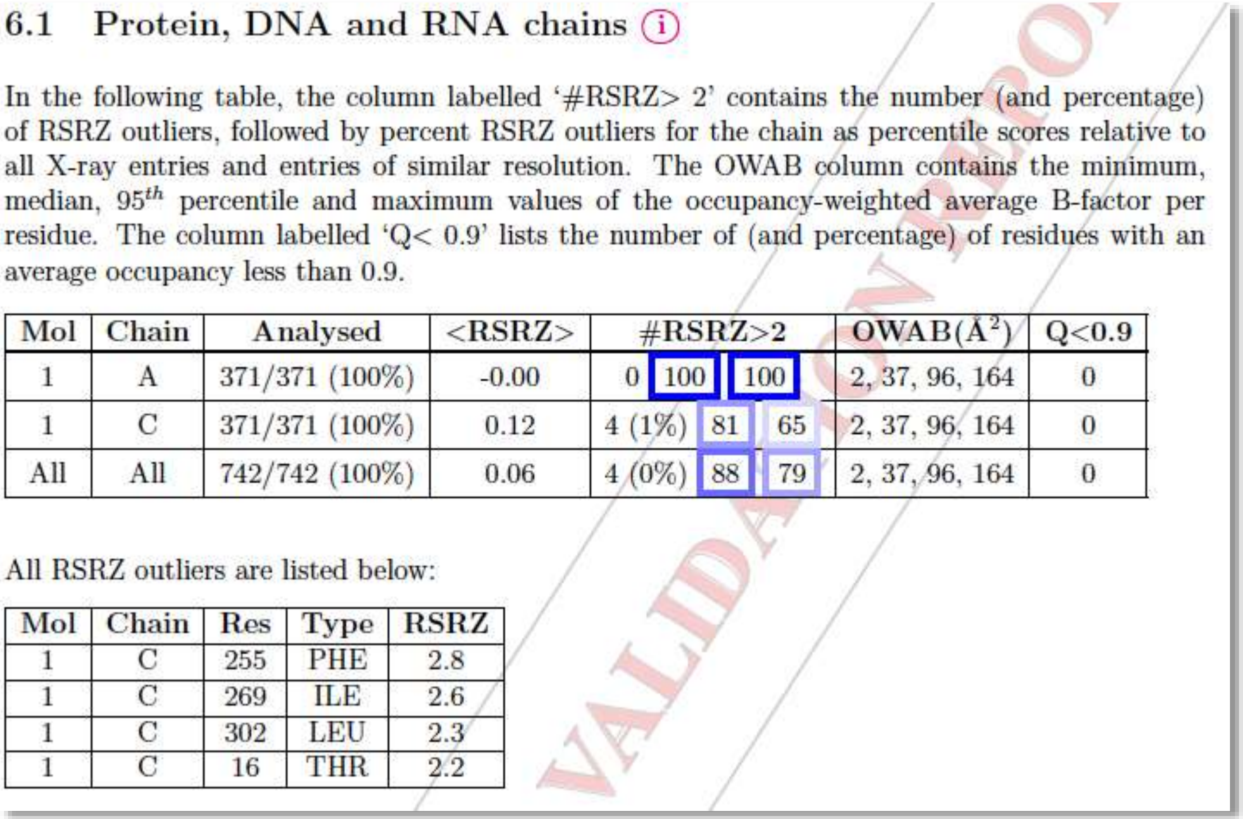
All bond angle outliers are listed below:

Mol	Chain	Res	Type	Atoms	Z	Observed(°)	Ideal(°)
1	C	1	PCA	CA-N-CD	8.03	108.09	114.37
1	A	1	PCA	CA-N-CD	8.03	108.10	114.37
1	A	1	PCA	C-CA-N	3.02	111.39	110.71
1	C	1	PCA	C-CA-N	2.93	111.38	110.71

There are no chirality outliers.

Validation reports

- Model/data fit proteins, DNA, RNA
  - RSR and RSR-Z (EDS)



6.1 Protein, DNA and RNA chains (i)

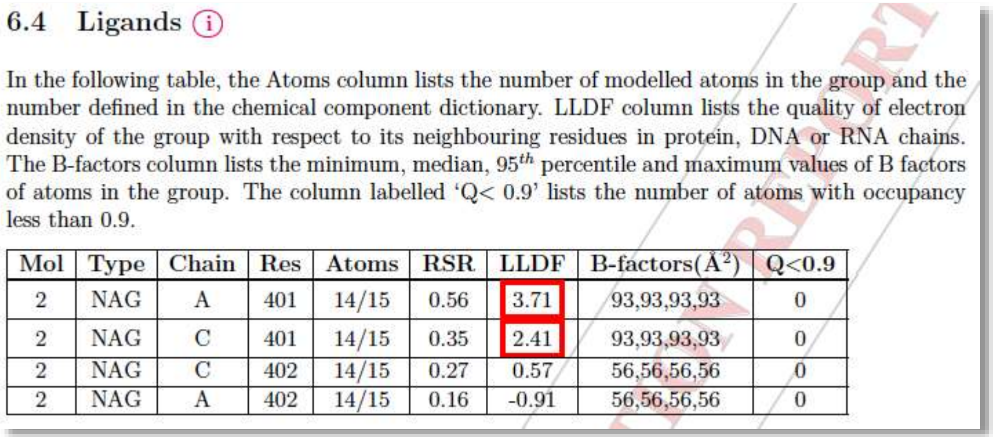
In the following table, the column labelled '#RSRZ> 2' contains the number (and percentage) of RSRZ outliers, followed by percent RSRZ outliers for the chain as percentile scores relative to all X-ray entries and entries of similar resolution. The OWAB column contains the minimum, median, 95th percentile and maximum values of the occupancy-weighted average B-factor per residue. The column labelled 'Q< 0.9' lists the number of (and percentage) of residues with an average occupancy less than 0.9.

Mol	Chain	Analysed	<RSRZ>	#RSRZ>2			OWAB(Å^2)	Q<0.9
1	A	371/371 (100%)	-0.00	0	100	100	2, 37, 96, 164	0
1	C	371/371 (100%)	0.12	4 (1%)	81	65	2, 37, 96, 164	0
All	All	742/742 (100%)	0.06	4 (0%)	88	79	2, 37, 96, 164	0

All RSRZ outliers are listed below:

Mol	Chain	Res	Type	RSRZ
1	C	255	PHE	2.8
1	C	269	ILE	2.6
1	C	302	LEU	2.3
1	C	16	THR	2.2

Validation reports



6.4 Ligands ⓘ

In the following table, the Atoms column lists the number of modelled atoms in the group and the number defined in the chemical component dictionary. LLDF column lists the quality of electron density of the group with respect to its neighbouring residues in protein, DNA or RNA chains. The B-factors column lists the minimum, median, 95th percentile and maximum values of B factors of atoms in the group. The column labelled 'Q< 0.9' lists the number of atoms with occupancy less than 0.9.

Mol	Type	Chain	Res	Atoms	RSR	LLDF	B-factors($\AA^2$)	Q<0.9
2	NAG	A	401	14/15	0.56	3.71	93,93,93,93	0
2	NAG	C	401	14/15	0.35	2.41	93,93,93,93	0
2	NAG	C	402	14/15	0.27	0.57	56,56,56,56	0
2	NAG	A	402	14/15	0.16	-0.91	56,56,56,56	0

- Model/data fit ligands etc.
  - RSR as usual
  - Can't usually compute RSR-Z due to few/no occurrences in PDB
  - New: "***LLDF***" – *Local Ligand Density Fit* = Z-score of ligand RSR relative to nearby polymeric residues (incl symmetry)

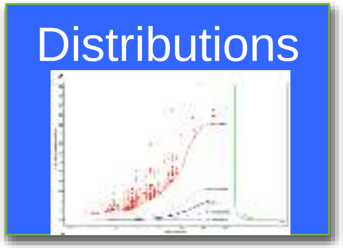
LLDF = (RSR(ligand) - <RSR(site)>) / σ(RSR(site))

wwPDB validation plans

- X-ray
  - Started sending out reports from 1 August
  - Stand-alone server in September
  - Integrated with new D&A (both for depositors and annotators) when it goes into production next year
- NMR
  - VTF recommendations will develop over time
  - Start with geometric checks, shifts and ensemble analysis
- EM
  - Not many accepted validation standards yet
  - Start with geometric checks and “sanity checks”

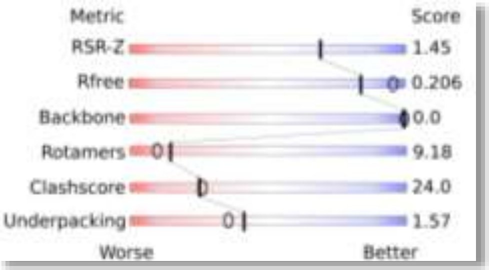
Validation information is not static

- Validation methods develop
  - Better criterion for Ligand RSR-Z
  - Methods to handle hybrid experimental data
- Every year we will update "distributions"
  - Example RSR-Z for Trp
    - Trp between 2.4 and 2.6Å:
      - 2012: 58321 observations, <> = 0.1419, σ = 0.0537
      - 2008: 26794 observations, <> = 0.1602, σ = 0.0660
      - RSR=0.25 ➔ RSR-Z=2.0 (2008: 1.4)
- Percentile scores will change


Distributions

RSR-Z (RSR, aa, d) =(RSR - <RSR(aa,d)>) / σ(RSR(aa,d))

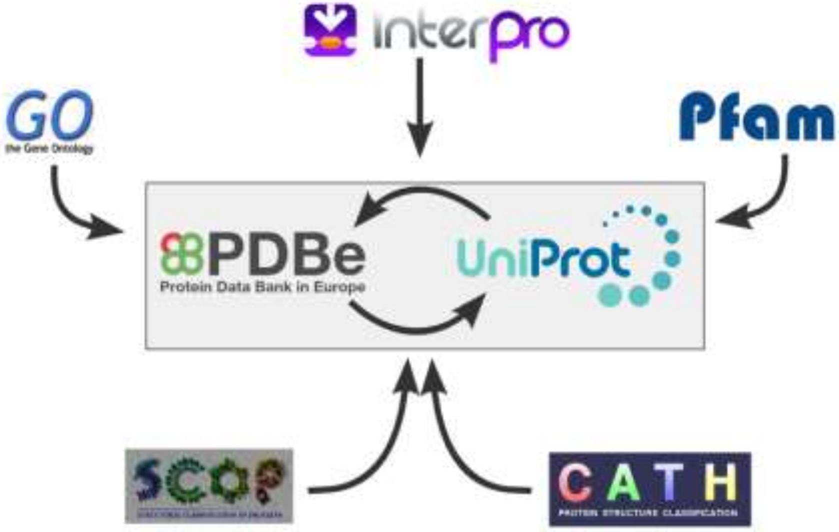
aa = residue type
d = resolution (in shells of 0.2Å)
Calculated using >57,000 EDS entries



Similar challenge faced with cross –reference information

- Cross reference information is not static
  - Species information
  - Sequence cross-reference
  - Related entries in other structure DB – e.g. EMDB, CATH, SCOP
  - Function information – the understanding may evolve over time
  - Biological assembly information – New experimental information may result in better understanding of quaternary assembly

SIFTS (pdbe.org/sifts)



- Structure Integration with Function, Taxonomy and Sequence
- Collaboration between UniProt and PDBe
- Data available for every polypeptide, not only the ones with UniProt mappings.
- Both GO and InterPro annotations based on PDB sequence
- Resources (InterPro, GO, UniProt, CATH, SCOP, Pfam, IntEnz) updated weekly.
- Better handling of UniProt accessions that have become obsolete.
- Ability to identify when a given accession appeared the first time.

Velankar et al., Nucleic Acids Research 41, D483 (2013)

What else might change?

- The data representation evolves over time
  - e.g. Representation of inhibitors and antibiotics
- Remediation of archive to improve consistency and data integrity

Acknowledgments

- PDBe staff
- All our wwPDB and EMDataBank colleagues
- All our other collaborators - VTF members and software developers who have contributed to the validation pipeline
- All our funders

PDBe staff and funders

wellcome trust

EMBL-EBI

BBSRC
bioscience for the future

NATIONAL INSTITUTES OF HEALTH

The Worldwide D&A Project Team