

# PDBx/mmCIFフォーマットによるPDB登録とデータ検証ポリシーについて

栗栖源嗣

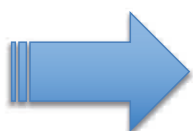
大阪大学蛋白質研究所



wwpdb.org

1

## After Solving Structure, what we should do?



<https://wwpdb.org>

2

# wwPDB Common Deposition & Annotation

<https://wwpdb.org>

The screenshot shows the wwPDB website homepage. The navigation bar at the top includes links for Validation, Deposition, Data Dictionaries, Documentation, Task Forces, Statistics, and About. The main content area is divided into three columns. The left column, 'wwPDB Members', lists various member organizations like BMRB, PDBe, PDBj, and RCSB PDB. The middle column, 'wwPDB Resources', provides links to Data Dictionaries, Annotation, and Information for Journals. The right column, 'News & Announcements', features recent updates and news items. A red circle highlights the 'Deposit Structure' button in the top right navigation area.

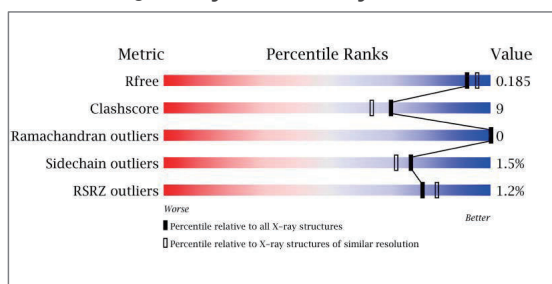


3

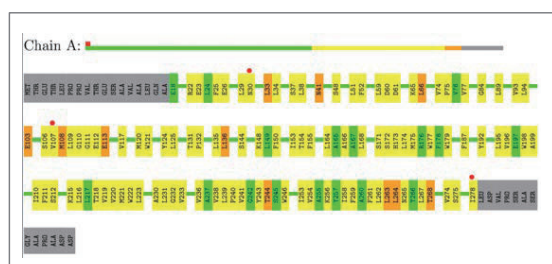
## X-ray Validation Report

- Model Quality
  - Bond lengths and angles (outlier info, RMS-Z)
  - Chirality, planarity
  - Close contacts (including worst clashes, MolProbity clash score)
  - Torsion angles (Ramachandran statistics, protein rotamers)
  - Ligand geometry (Mogul analysis)
- Residue Plots
  - Residues with model-quality outliers (0, 1, 2, >2)
  - Residues with RSR-Z > 5 are highlighted
  - Residues not observed

### Overall Quality Summary



### Residue Plots



4

## Validation software utilized for generation of wwPDB validation report (2018)


**Table 3. Component Software Packages Included in the 2017 Version of the Validation Pipeline**

Software Package	Which Section and Metric of the Report the Package Is Used for
MolProbity	model geometry: bond lengths and bond angles of standard protein residues and nucleotides, too-close contacts, Ramachandran outliers, rotamer outliers, RNA suiteness
MAXIT	model geometry: symmetry-related too-close contacts, stereochemistry issues, identification of <i>cis</i> -peptides
Mogul <small>Update (2018, CSD archive)</small>	model geometry: bond-length and bond-angle outliers in small molecules
Xtriage (Phenix) <small>Update (Phenix 1.13)</small>	crystallographic data and refinement statistics: signal-to-noise, twinning
DCC	crystallographic data and refinement statistics: $R$ , $R_{\text{free}}$ fit to crystallographic data: $R_{\text{free}}$
EDS <small>Update (Recmac 7.0v44)</small>	fit to crystallographic data: real-space $R$ outliers
Cyrange	NMR ensemble composition: identification of well-defined protein cores
RCI	NMR chemical shifts: prediction of protein backbone order parameter from chemical shifts
PANAV	NMR chemical shifts: suggested referencing corrections in chemical shift assignments

Percentile statistics reflecting the state of the archive on December 31st 2017.

Structure 25, 1916–1927, 2017<sup>5</sup>

## Stand-alone wwPDB validation server


**wwPDB Validation Service**
[FAQ](#)

Existing validation

Validation ID

Password

Log in

Forgot Password

Deposition server

Deposit your data to PDB, BMRB and EMDB at [deposit.wwpdb.org](https://deposit.wwpdb.org)

wwPDB news and announcements

**Compliance with GDPR legislation**  
wwPDB has revised its [privacy policy](#) in line with the requirements of the EU's GDPR legislation.

Start a new validation

Welcome to the wwPDB validation system.  
This server runs the performs the same validation as you would observe during the deposition process. This service is designed to help you check your model and experimental files prior to start of deposition.  
To continue with an existing validation, please login on the left.  
To start a new validation, please complete the form below. Upon completion, you will be emailed login information specific to your new validation.

Your e-mail address

Password (optional, or we will provide one)  
This is a shared "group password" (6 to 16 alphanumeric characters)

Country

Experimental method  
☐ X-Ray Diffraction  
☐ Electron Microscopy  
☐ Solution NMR  
☐ Neutron Diffraction  
☐ Electron Crystallography  
☐ Solid-state NMR  
☐ Fiber Diffraction

Please copy this code : 56819

Privacy policy  
☐ Tick to indicate that you have read and accepted the wwPDB policy on personal data privacy, including what data wwPDB collects, how the data is stored and shared. [www.wwpdb.org/about/privacy](https://www.wwpdb.org/about/privacy)



# PDB Core Archive Update I

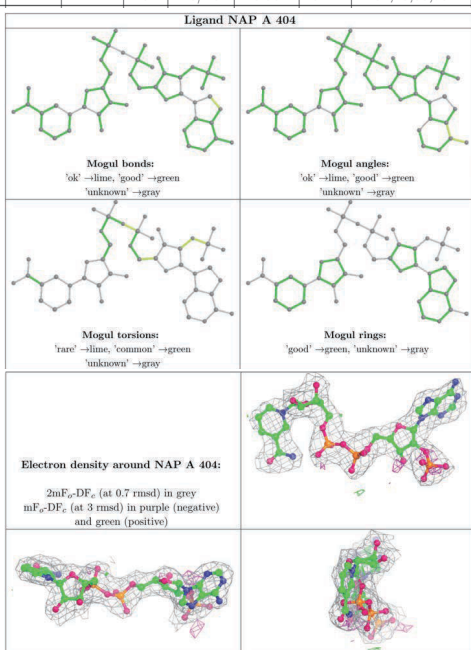
## Improved Ligand Validation

- Adapted software from Global Phasing Ltd. under a formal collaboration agreement with all wwPDB partners
- Benefits:
  - Provides geometrical quality in 2D depiction
  - Provides electron density fit for X-ray in 2D depiction
- Now mandatory at deposition: identification of Ligand(s) Of Interest (LOI, author's research focus)
  - 2D depictions provided in wwPDB Validation Report for all LOIs

7

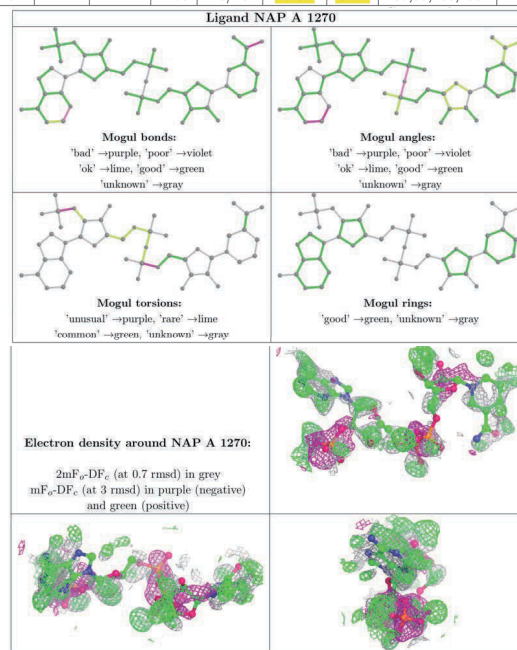
## Ligand Validation- NADP Examples

Mol	Type	Chain	Res	Atoms	RSCC	RSR	B-factors( $\text{\AA}^2$ )	Q<0.9
-----	------	-------	-----	-------	------	-----	-----------------------------	-------



PDB entry 5zix (Better data quality)

Mol	Type	Chain	Res	Atoms	RSCC	RSR	B-factors( $\text{\AA}^2$ )	Q<0.9
-----	------	-------	-----	-------	------	-----	-----------------------------	-------



PDB entry 1zk4 (Worse data quality)

8

# PDB File formats from wwPDB

- (Legacy) PDB format
  - *NOT RECOMMENDED!*
- PDBx/mmCIF
  - The canonical format of the wwPDB.
  - Ver. 5.299 released.
- PDBML
  - “direct translation” of mmCIF into XML.
- PDB/RDF
  - Translation of PDBML into RDF/XML (the standard format for the Semantic Web).

9

## Atomic coordinates in PDBx/mmCIF

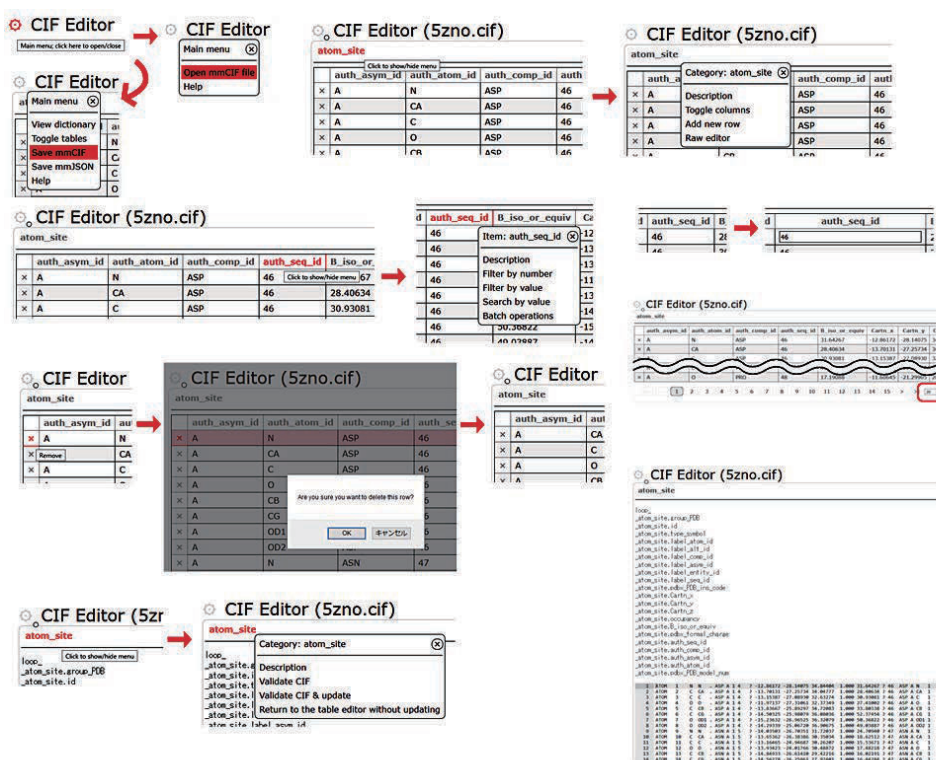
```

loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.Cartn_x_esd
_atom_site.Cartn_y_esd
_atom_site.Cartn_z_esd
_atom_site.occupancy_esd
_atom_site.B_iso_or_equiv_esd
_atom_site.pdbx_formal_charge
_atom_site.auth_seq_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
ATOM 1 N N . ALA A 1 1 ? 38.840 0.236 1.012 1.00 34.65 ? ? ? ? ? 1 ALA A N 1
ATOM 2 C CA . ALA A 1 1 ? 38.356 -0.999 0.357 1.00 42.26 ? ? ? ? ? 1 ALA A CA 1
ATOM 3 C C . ALA A 1 1 ? 37.098 -1.547 1.056 1.00 41.25 ? ? ? ? ? 1 ALA A C 1
ATOM 4 O O . ALA A 1 1 ? 36.619 -0.946 2.028 1.00 29.44 ? ? ? ? ? 1 ALA A O 1
ATOM 5 C CB . ALA A 1 1 ? 39.398 -2.114 0.379 1.00 40.70 ? ? ? ? ? 1 ALA A CB 1
ATOM 6 N N . SER A 1 2 ? 36.610 -2.666 0.495 1.00 32.67 ? ? ? ? ? 2 SER A N 1
ATOM 7 C CA . SER A 1 2 ? 35.411 -3.244 1.202 1.00 34.90 ? ? ? ? ? 2 SER A CA 1
ATOM 8 C C . SER A 1 2 ? 35.683 -4.740 1.081 1.00 38.30 ? ? ? ? ? 2 SER A C 1
ATOM 9 O O . SER A 1 2 ? 36.827 -5.147 0.747 1.00 28.59 ? ? ? ? ? 2 SER A O 1
ATOM 10 C CB . SER A 1 2 ? 34.063 -2.660 0.823 1.00 24.49 ? ? ? ? ? 2 SER A CB 1
ATOM 11 O OG . SER A 1 2 ? 33.031 -3.308 1.686 1.00 20.37 ? ? ? ? ? 2 SER A OG 1

```

10

<https://pdbj.org/cif-editor>



11

## Deposition Notes

- Model coordinates
  - TER records
  - Coordinates for Chimera protein
  - ALA models
- Sample sequence
  - Complete polymer sequence
  - Non-poly residues
  - UNK
  - Source organism
- Deposition efficiency
- Communication, contact authors & ORCiD
- Release

12

# Chimera protein (polymer1-linker-polymer2)

N ---- LEU LYS ALA <GLY SER GLY> LYS ILE GLU GLU GLY LYS LEU VAL ILE -----C

ATOM	95	N	LYS	A	19	23.835	-13.001	26.015
ATOM	96	CA	LYS	A	19	24.415	-12.460	27.237
ATOM	97	C	LYS	A	19	24.521	-10.943	27.107
ATOM	98	O	LYS	A	19	24.200	-10.203	28.041
ATOM	104	N	ALAA	20		24.954	-10.493	25.930
ATOM	105	CA	ALAA	20		25.008	-9.069	25.602
ATOM	106	C	ALAA	20		23.707	-8.629	24.943
ATOM	107	O	ALAA	20		23.686	-7.686	24.149

ATOM	123	N	LYS	A	24	23.102	-4.528	18.734
ATOM	124	CA	LYS	A	24	24.010	-3.416	18.990
ATOM	125	C	LYS	A	24	23.870	-2.274	17.983
ATOM	126	O	LYS	A	24	24.740	-1.408	17.900
ATOM	132	N	ILE	A	25	22.794	-2.279	17.205
ATOM	133	CA	ILE	A	25	22.483	-1.122	16.372
ATOM	134	C	ILE	A	25	21.929	-0.030	17.288
ATOM	135	O	ILE	A	25	21.002	-0.276	18.049

ATOM	109	N	GLY	A	100	22.624		
ATOM	110	CA	GLY	A	100	21.342		
ATOM	111	C	GLY	A	100	20.794		
ATOM	112	O	GLY	A	100	20.917		
ATOM	113	N	SER	A	101	20.197		
ATOM	114	CA	SER	A	101	20.100		
ATOM	115	C	SER	A	101	20.155		
ATOM	116	O	SER	A	101	19.169		
ATOM	119	N	GLY	A	102	21.304		
ATOM	120	CA	GLY	A	102	21.478		
ATOM	121	C	GLY	A	102	22.240		
ATOM	122	O	GLY	A	102	22.047		

13

## ALA model

When an amino acid residue is disordered due to low density

Side chain atoms cannot be assigned, and the residue is often modeled as:

- ALA model
- GLY model
- SER model

Etc.

**However, the residue name in the coordinates should not be changed to MATCH with the sequence even without the atoms in the side chain.**

# Sample Sequence

Fill in a complete polymer sequence used for experiment

- Please include

HIS- or other Expression tags, Linker, Residues missing from the coordinates due to disorder

Coordinates: . . . . . LVVVTNNLR . . . RIPGIRIED . . . ITLMELILEH . . . . .

Sequence: **GSHMALVVVTNNLR****EFERIPGIRIED****GSGITLMELILEHHHHH**

- Please don't include

Residues cleaved from the macromolecules prior to or during the experiment

15

# DO NOT include ligands

Sample sequence is a list of the consecutive chemical components covalently linked in a linear fashion to form a polymer.

Please don't include:

- Metal ions, Chemical components or groups covalently linked to side-chains (in peptides)
- Floating Metal ions, Chemical components or groups

incorrect sequence: NLREFERIPG**(NAG)**IRIEDYTYITLMELILEHHH**(NAG)(ZN)**

(Should be) NLREFERIPGIRIEDYTYITLMELILEHHH

16



## Use "UNK" or "N" ONLY in TWO special cases

UNK : unknown amino acid

N or DN : unknown nucleotide

1. You don't know the sequence
2. You know the sequence but don't know how the coordinates align with the sequence

(Example)

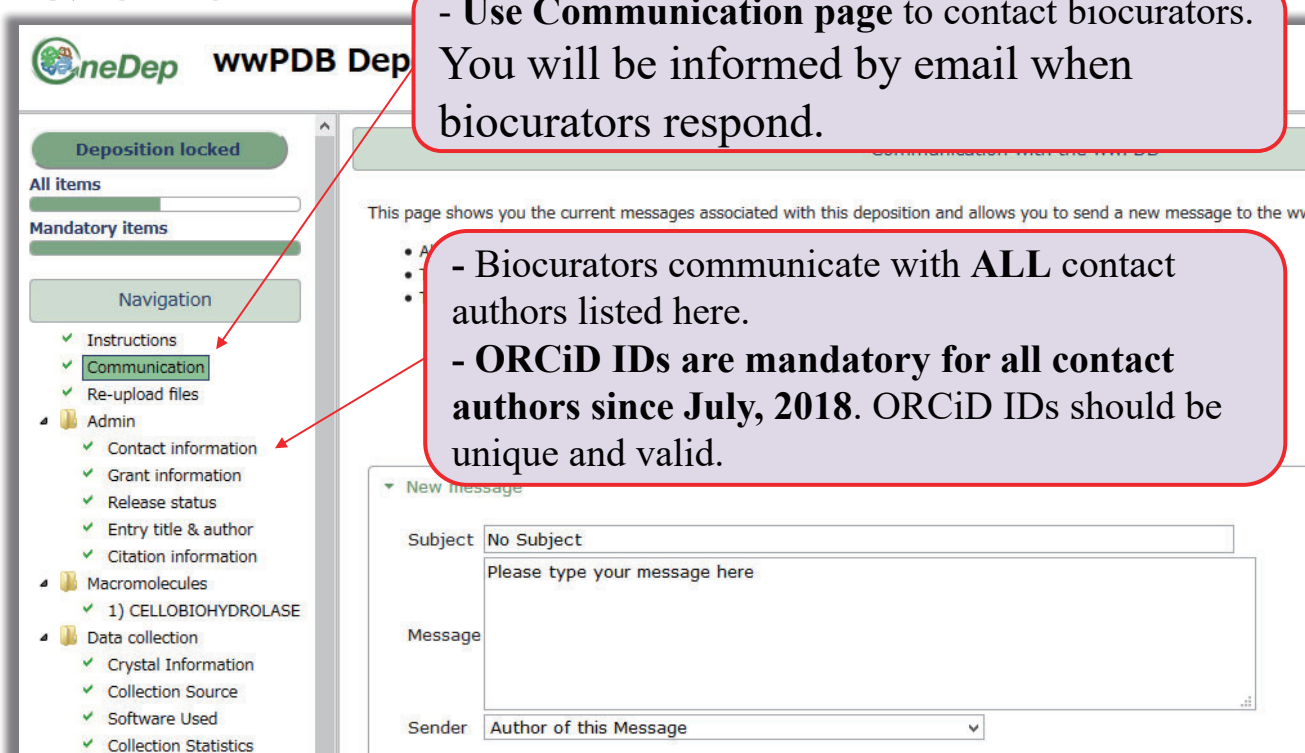
You know the sequence, but are not sure if the first residue seen in the density is really the first residue of the sequence.

(Note)

Please use "UNK" or "N" in the coordinates only when you use "UNK" or "N" in sequence.

17

## Communication, Contact authors & ORCiD



- Use **Communication page** to contact biocurators. You will be informed by email when biocurators respond.

- Biocurators communicate with **ALL** contact authors listed here.

- **ORCiD IDs are mandatory for all contact authors since July, 2018.** ORCiD IDs should be unique and valid.

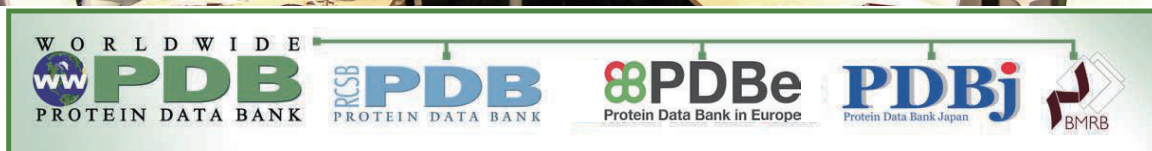
18

# Release

- Do not ask for replacement of coordinates just before the release
- When you refer to PDB in your paper, please refer as **“the coordinates are deposited to the wwPDB”**.
  - Your entries are processed by PDBj, RCSB or PDBe
- Release instructions (REL, HPUB, HOLD) cannot be changed by depositors
  - Ask biocurators after submission

19

# Acknowledgements



20