Ligand representation and validation in the Protein Data Bank

Akira R. Kinjo, Haruki Nakamura, Genji Kurisu
PDBj, Institute for Protein Research, Osaka University

WORLDWIDE PDB PROTEIN DATA BANK
wwpdb.org

Protein Data Bank

- PDB: 1st Open Access digital resource in biology (est. in 1971 with 7 entries)
- Initially, managed jointly by data centers in US and UK
- Today, single global PDB macromolecular structure archive (>138,000 entries)

Nature New Biology **233**, page 223 (1971)

(>26,000 entries with sugars)

CRYSTALLOGRAPHY

Protein Data Bank

A repository system for protein crystallographic data will be operated jointly by the Crystallographic Data Centre, Cambridge, and the Brookhaven National Laboratory. The system will be responsible for storing atomic coordinates, structure factors and electron density maps and will make these data available on request. Distribution will be on magnetic tape in machine-readable form whenever possible. There will be no charge for the service other than handling costs. Files will be updated as new material is received. The total holding will be announced annually in the organic bibliographic volumes of the reference series "Molecular Structures and Dimensions" published for the Crystallographic Data Centre and the International Union of Crystallography by Oosthoek's, Utrecht.

The success of the proposed system will depend on the response of the protein crystallographers supplying data. These will be

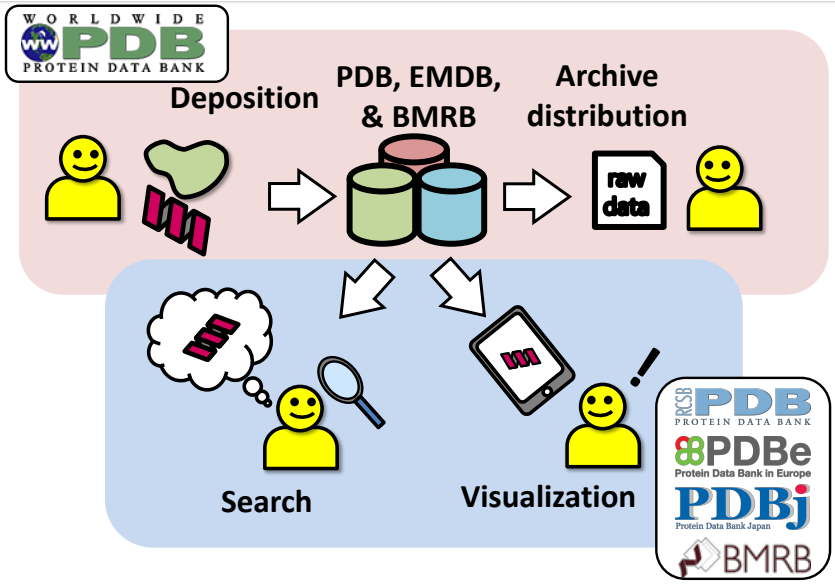
Worldwide Protein Data Bank (wwPDB)

- Ensures data are freely and globally available
- Members
  - RCSB PDB (US)*Archive Keeper
  - PDBj (Osaka University, Japan)
  - PDBe (EMBL-EBI)
  - BioMagResBank (University Wisconsin, Madison, US)

  Founding Members
- Collaborate on data processing and annotation
- Each site provides different websites that offer different services and views of the data

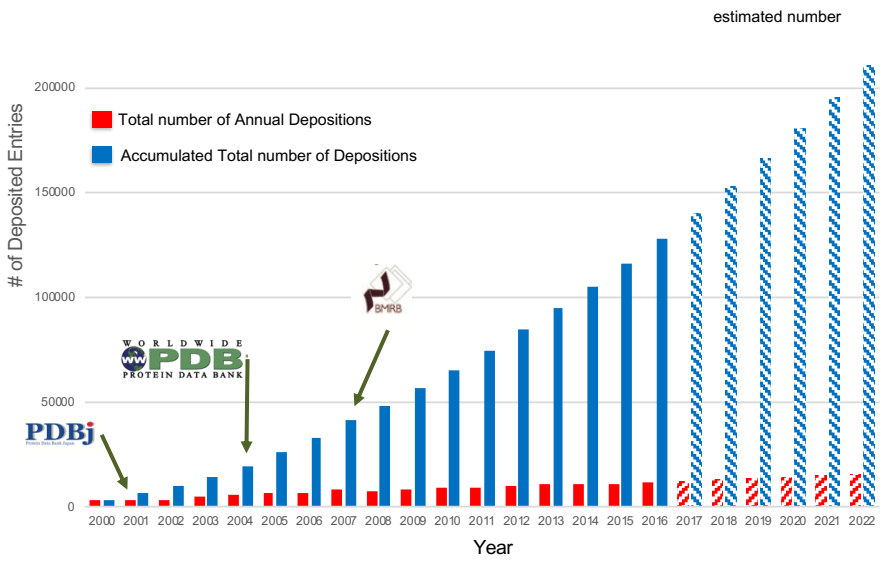
2

wwPDB collaboration



3

Growing Number of Depositions



4

PDB File formats from wwPDB

- (Legacy) PDB format
  - *NOT RECOMMENDED!*
- **mmCIF**
  - The canonical format of the wwPDB.
  - Ver. 5 released.
- **PDBML**
  - “direct translation” of mmCIF into XML.
- **PDB/RDF**
  - Translation of PDBML into RDF/XML (the standard format for the Semantic Web).

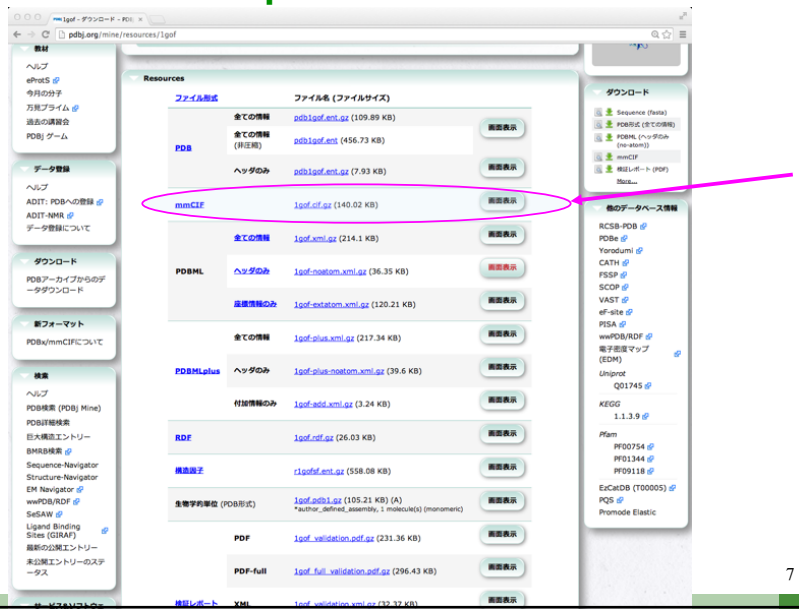
5

Other data provided from wwPDB

- Validation Report
  Translation to the RDF format is on going.
- PDB archive of Structure factors (for crystal structures)
- BMRB archive of NMR distance restraints
- EMDB archive of Cryo-TEM maps

6

mmCIF: an example



7

mmCIF basics

- Data are divided into “categories”.
  - *_category.item*
  - e.g., _entry.id → “entry” is the category name, “id” is an item of the “entry” category.
  - “_entry.id 1GOF” → The value of “id” item of “entry” category is “1GOF”.
- Two ways of presenting data.
  - key-value: if only one value exists for an item.
  - loop: if more than one item exists for an item.

8

More about mmCIF

- Context-free grammar (STAR [Self-defining Text Archive and Retrieval])
- All the categories and items are defined in the PDBx/mmCIF dictionary.
- For details, see http://mmcif.wwpdb.org/

9

A closer look at mmCIF

```
data_1GOF
#
_entry.id   1GOF
#
_audit_conform.dict_name        mmcif_pdbx.dic
_audit_conform.dict_version     5.287
_audit_conform.dict_location    http://mmcif.pdb.org/dictionaries/ascii/mmcif_pdbx.dic
#
_database_2.database_id     PDB
_database_2.database_code   1GOF
#
….
```

This tag starts a data ("datablock" is the unit of data [entry])

entry ID (PDB ID) is "1GOF".

Provenance information is also included.

10

Example of key-value pairs.

```
_cell.entry_id              1GOF
_cell.length_a              98.000
_cell.length_b              89.400
_cell.length_c              86.700
_cell.angle_alpha           90.00
_cell.angle_beta            117.80
_cell.angle_gamma           90.00
_cell.Z_PDB                 4
_cell.pdbx_unique_axis      ?
#
```

The last "#" is a convention to indicate the end of a category.

11

Example of “loop” structure

```
loop_
_entity.id
_entity.type
_entity.src_method
_entity.pdbx_description
_entity.formula_weight
_entity.pdbx_number_of_molecules
_entity.details
_entity.pdbx_mutation
_entity.pdbx_fragment
_entity.pdbx_ec
1 polymer     man 'GALACTOSE OXIDASE' 68579.250 1   ? ? ? 1.1.3.9
2 non-polymer syn 'COPPER (II) ION'   63.546    1   ? ? ? ?
3 non-polymer syn 'SODIUM ION'        22.990    1   ? ? ? ?
4 non-polymer syn 'ACETIC ACID'       60.052    2   ? ? ? ?
5 water       nat water               18.015    316 ? ? ? ?
#
```

Start of a loop

A list of items.
“One item per line” is just a convention.

- Each item is whitespace-delimited.
- In the same order as the item list (above).
- Use quotes (') for data with whitespace.

The last “#” is a convention to indicate the end of a loop.

12

Atomic coordinates in 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
```

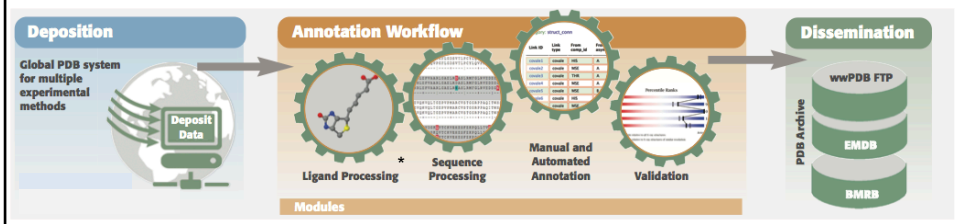
13

Main category groups

- _entity (Info about molecules)
  - entity, entity_poly, pdbx_entity_nonpoly, ...
- _atom (Info about atoms)
  - atom_site
- _struct (structural info)
  - struct, struct_conf, struct_sheet, struct_conn, pdbx_struct_assembly, ...
- _chem_comp (chemical components)
  - chem_comp
- _citation (literature info)
  - citation, citation_author, ...

14

wwPDB Common Deposition & Annotation



- Enables workload balancing and has increased productivity
- Better quality assurance of polymer sequences and ligand chemistry
- PDBx/mmCIF is now the master file format
- Validation based on recommendations from expert task forces
- Federation with other Data Resources (e.g., EMDB, SASBDB, ...)

15

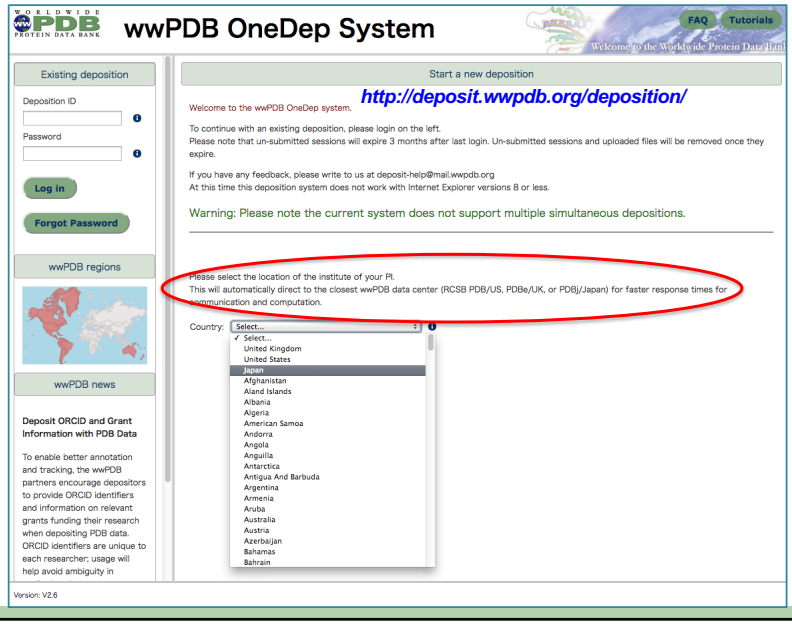
wwPDB Common Deposition & Annotation

https://wwpdb.org

OneDep

16

wwPDB Common Deposition & Annotation

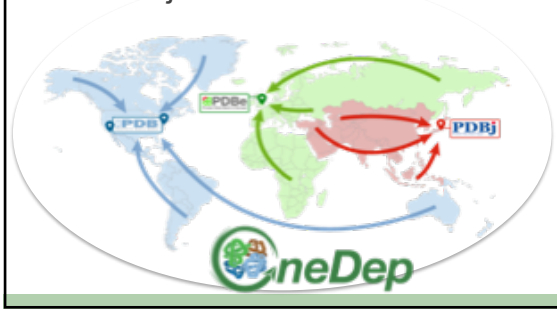


http://deposit.wwpdb.org/deposition/

17

wwPDB Common Deposition & Annotation

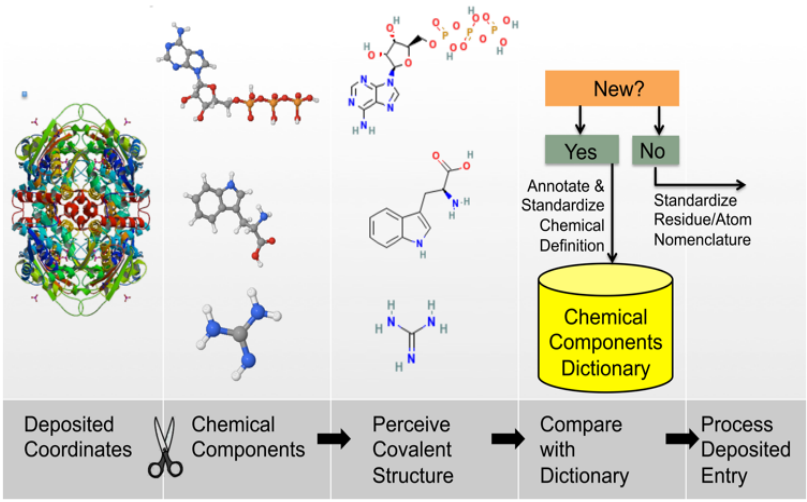
- As of 2016 region-based processing of D&A-deposited entries:
  - RCSB PDB: Americas & Oceania
  - PDBe: Europe & Africa
  - PDBj: Asia & Middle east



Year	Total Depositions	Processed By		
		RCSB PDB	PDBj	PDBe
2000	2983	2297	158	528
2001	3287	2408	383	496
2002	3565	2401	657	507
2003	4830	3135	1026	669
2004	5508	3082	1614	812
2005	6678	3563	2110	1005
2006	7282	4252	1945	1085
2007	8130	4703	2299	1128
2008	7073	4106	1994	973
2009	8300	5069	2173	1058
2010	8878	5464	2041	1373
2011	9250	5938	1816	1496
2012	9972	6408	1888	1676
2013	10566	6652	2128	1786
2014	10364	6038	1781	2545
2015	10958	4845	2100	4013
2016	11614	5326	2238	4050
2017	2577	1579	394	604
TOTAL	131815	77266	28745	25804

18

Chemical Component Deposition Pipeline



19

Chemical Component Dictionary (CCD)

- Complete descriptions of constituent small molecules in experimentally-determined 3D macromolecular structures in the PDB
- Data items include
  - Atom Nomenclature
  - Connectivity/Chirality
  - Chemical Formula, InChI/SMILES, etc.
  - Molecular Names
  - Idealized 3D Structure
  - 3D Structure Examplar from PDB Archive

20

Chemical Component Dictionary Entry

a)

```
data_HYP
#
_chem_comp.id                                   HYP
_chem_comp.name                                 4-HYDROXYPROLINE
_chem_comp.type                                 "L-PEPTIDE LINKING"
_chem_comp.pdbx_type                            ATOMP
_chem_comp.formula                              "C5 H9 N O3"
_chem_comp.mon_nstd_parent_comp_id              PRO
_chem_comp.pdbx_synonyms                        HYDROXYPROLINE
_chem_comp.pdbx_formal_charge                   0
_chem_comp.pdbx_initial_date                    1999-07-08
_chem_comp.pdbx_modified_date                   2008-04-29
_chem_comp.pdbx_ambiguous_flag                  N
_chem_comp.pdbx_release_status                  REL
_chem_comp.pdbx_replaced_by                     ?
_chem_comp.pdbx_replaces                        ?
_chem_comp.formula_weight                       131.130
_chem_comp.one_letter_code                      ?
_chem_comp.three_letter_code                    HYP
_chem_comp.pdbx_model_coordinates_details       ?
_chem_comp.pdbx_model_coordinates_missing_flag  N
_chem_comp.pdbx_ideal_coordinates_details       Corina
_chem_comp.pdbx_ideal_coordinates_missing_flag  N
_chem_comp.pdbx_model_coordinates_db_code       1OB6
_chem_comp.pdbx_processing_site                 RCSB
#
```

b)

```
loop_
_chem_comp_atom.comp_id
_chem_comp_atom.atom_id
_chem_comp_atom.alt_atom_id
_chem_comp_atom.type_symbol
_chem_comp_atom.charge
_chem_comp_atom.pdbx_align
_chem_comp_atom.pdbx_aromatic_flag
_chem_comp_atom.pdbx_leaving_atom_flag
_chem_comp_atom.pdbx_stereo_config
_chem_comp_atom.model_Cartn_x
_chem_comp_atom.model_Cartn_y
_chem_comp_atom.model_Cartn_z
_chem_comp_atom.pdbx_model_Cartn_x_ideal
_chem_comp_atom.pdbx_model_Cartn_y_ideal
_chem_comp_atom.pdbx_model_Cartn_z_ideal
_chem_comp_atom.pdbx_ordinal
HYP N    N   N 0 1 N N N -3.366 16.585 44.188  0.168  1.360 -0.282 1
HYP CA   CA  C 0 1 N N S -2.955 15.769 43.044 -0.384 -0.003 -0.493 2
HYP C    C   C 0 1 N N N -1.447 15.609 43.030 -1.811 -0.072 -0.013 3
HYP O    O   O 0 1 N N N -0.722 16.484 43.503 -2.233  0.764  0.750 4
HYP CB   CB  C 0 1 N N N -3.488 16.578 41.829  0.515 -0.924  0.359 5
HYP CG   CG  C 0 1 N N N -4.437 17.482 42.330  1.847 -0.159  0.505 6
HYP CD   CD  C 0 1 N N N -4.068 17.803 43.753  1.640  1.159 -0.271 7
HYP OD1  OD  O 0 1 N N N -5.693 16.815 42.294  2.917 -0.911 -0.071 8
HYP OXT  OXT O 0 1 N Y N -0.976 14.502 42.469 -2.614 -1.083 -0.433 9
HYP H    H   H 0 1 N Y N -3.980 16.047 44.765 -0.107  1.981 -1.028 10
HYP HA   HA  H 0 1 N N N -3.385 14.756 43.068 -0.325 -0.278 -1.546 11
HYP HB2  1HB H 0 1 N N N -2.567 17.141 41.398  0.066 -1.092  1.337 12
HYP HB3  2HB H 0 1 N N N -3.790 15.930 41.026  0.678 -1.873 -0.153 13
HYP HG   HG  H 0 1 N N N -4.508 18.399 41.726  2.052  0.048  1.555 14
HYP HD22 1HD H 0 0 N N N -4.956 18.005 44.370  2.018  1.065 -1.289 15
HYP HD23 2HD H 0 0 N N N -3.457 18.713 43.848  2.132  1.965  0.243 16
HYP HD1  HOD H 0 1 N N N -5.999 16.666 43.181  3.780 -0.479 -0.009 17
HYP HXT  HXT H 0 1 N N N -0.027 14.511 42.499 -3.520 -1.064 -0.098 18
#
```

Atom names

Stereochemistry & aromaticity

Model coordinates

Ideal coordinates

c)

```
loop_
_chem_comp_bond.comp_id
_chem_comp_bond.atom_id_1
_chem_comp_bond.atom_id_2
_chem_comp_bond.value_order
_chem_comp_bond.pdbx_aromatic_flag
_chem_comp_bond.pdbx_stereo_config
_chem_comp_bond.pdbx_ordinal
HYP N   CA   SING N N 1
HYP N   CD   SING N N 2
HYP N   H    SING N N 3
HYP CA  C    SING N N 4
HYP CA  CB   SING N N 5
HYP CA  HA   SING N N 6
HYP C   O    DOUB N N 7
HYP C   OXT  SING N N 8
HYP CB  CG   SING N N 9
HYP CB  HB2  SING N N 10
HYP CB  HB3  SING N N 11
```

Connected atoms

Bond type

Stereochemistry & aromaticity

d)

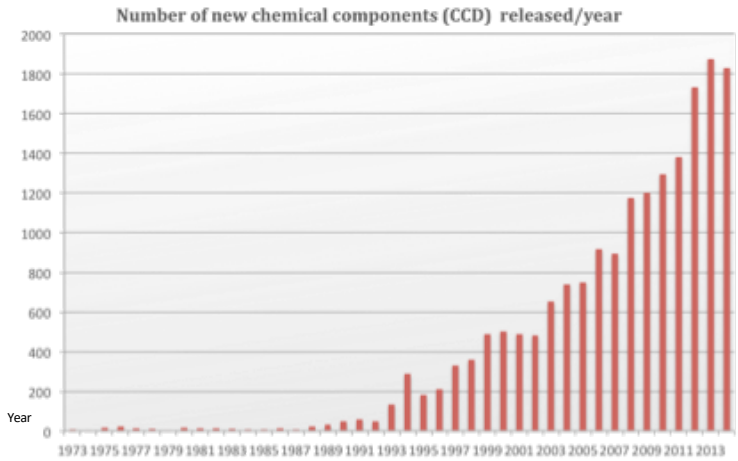
```
#
loop_
_pdbx_chem_comp_descriptor.comp_id
_pdbx_chem_comp_descriptor.type
_pdbx_chem_comp_descriptor.program
_pdbx_chem_comp_descriptor.program_version
_pdbx_chem_comp_descriptor.descriptor
HYP SMILES           ACDLabs              12.01 O=C(O)C1NCC(O)C1
HYP SMILES_CANONICAL CACTVS               3.385 O[C@H]1CN[C@@H](C1)C(O)=O
HYP SMILES           CACTVS               3.385 O[C@H]1CN[CH](C1)C(O)=O
HYP SMILES_CANONICAL "OpenEye OEToolkits" 1.7.5 C1[C@H](CN[C@@H]1C(=O)O)O
HYP SMILES           "OpenEye OEToolkits" 1.7.5 C1C(CNC1C(=O)O)O
HYP InChI            InChI                1.03  "InChI=1S/C5H9NO3/c7-3-1-4(5(8)9)6-2-3/h3-4,6-7H,1 ..."
HYP InChIKey         InChI                1.03  PMMYEEVYMWASQN-DMTCNVIQSA-N
#
loop_
_pdbx_chem_comp_identifier.comp_id
_pdbx_chem_comp_identifier.type
```

SMILES

InChI

21

Growth of Chemical Components in PDB

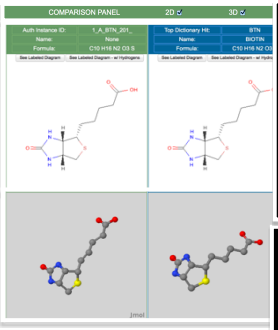


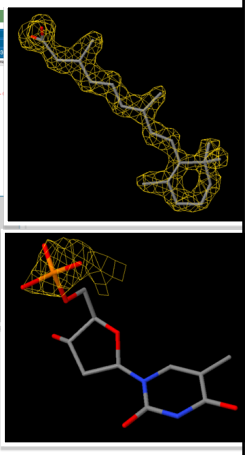
Growth trends new Chemical Components in the PDB

22

Improved Ligand Annotation

- Batch search against Chemical Component Dictionary with automated CCD ID assignment
- Captures and displays author-provided chemical information
- Comparison panel
  - 2D and 3D views of ligand for review
  - ID assignment
- Display of local ligand electron density fit





Deposited instance from coordinates (left) and the closest match in the dictionary (right)

Local ligand density display (1.5 sigma omit map)
Top: REA in entry 1CBS with LLDF=1.31 (**RSR=0.10, CC=0.95**)
Bottom: TMP in entry 3HW4 with LLDF=6.77 (**RSR=0.41, CC=0.70**)

23

wwPDB/CCDC/D3R Ligand Validation Workshop

Meeting Objectives: To bring together co-crystal structure determination experts from Academe and Industry with Crystallography and Computational Chemistry Software Developers to discuss, develop, and recommend:

- Best practices PDB archive deposition/validation of co-crystal structures
- Editorial/Refereeing/Publication standards for co-crystal structures
- Improvements in ligand representation across the PDB Archive

24

Workshop White Paper

- White Paper describing recommendations re deposition/validation and editorial/refereeing/publication standards is published in *Structure* 24, 502-508 (2016)

CellPress

Structure
Meeting Report

Outcome of the First wwPDB/CCDC/D3R Ligand Validation Workshop

Paul D. Adams,[1] Kathleen Aertgeerts,[2] Cary Bauer,[3] Jeffrey A. Bell,[4] Helen M. Berman,[5,6] Talapady N. Bhat,[7] Jeff M. Blaney,[8] Evan Bolton,[9] Gerard Bricogne,[10] David Brown,[11,12] Stephen K. Burley,[5,6,13,*] David A. Case,[6] Kirk L. Clark,[14] Tom Darden,[15] Paul Emsley,[16] Victoria A. Feher,[17,+] Zukang Feng,[5,6] Colin R. Groom,[18,*] Seth F. Harris,[8] Jorg Hendle,[19] Thomas Holder,[4] Andrzej Joachimiak,[20] Gerard J. Kleywegt,[21]

(Author list continued on next page)

25

Improved Validation

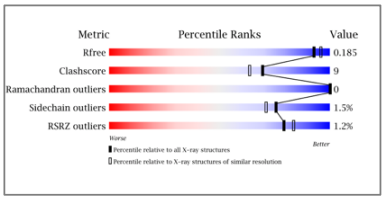
- wwPDB Validation Task Forces X-ray, NMR, SAS
- wwPDB/EMDataBank VTF for EM
- Recommendations about validating new and existing structures
  - Implemented in software pipeline
  - Produces summary report (PDF) and XML file with detailed statistics
- Validation at different stages
  - While determining/depositing the structure
  - After annotation (official; should be sent to journals)
  - Upon release (publicly available; updated annually)

26

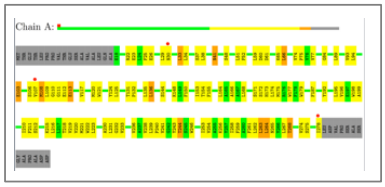
X-ray Validation Report

- 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



27

Validation Report is requested for peer review

EDITORIAL

nature structural & molecular biology

Nature Struct. Mol. Biology, 23 (10), 871, 2016

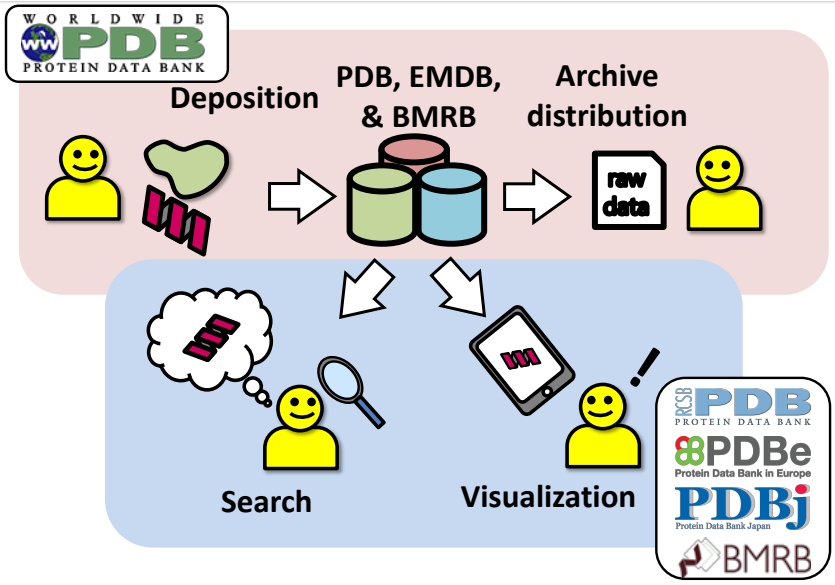
Where are the data?

Here, we announce two policy changes across Nature journals: data-availability statements in all published papers and official Worldwide Protein Data Bank (wwPDB) validation reports for peer review.

We are now taking a further step and are requesting official wwPDB validation reports for peer review. These reports are made available by the wwPDB after data deposition (http://www.wwpdb.org/validation/validation-reports). Other Nature journals will soon follow suit.

28

wwPDB collaboration



29

Worldwide Protein Data Bank (wwPDB)

30

Whi

31

DOI Landing Page Layout (Planned)

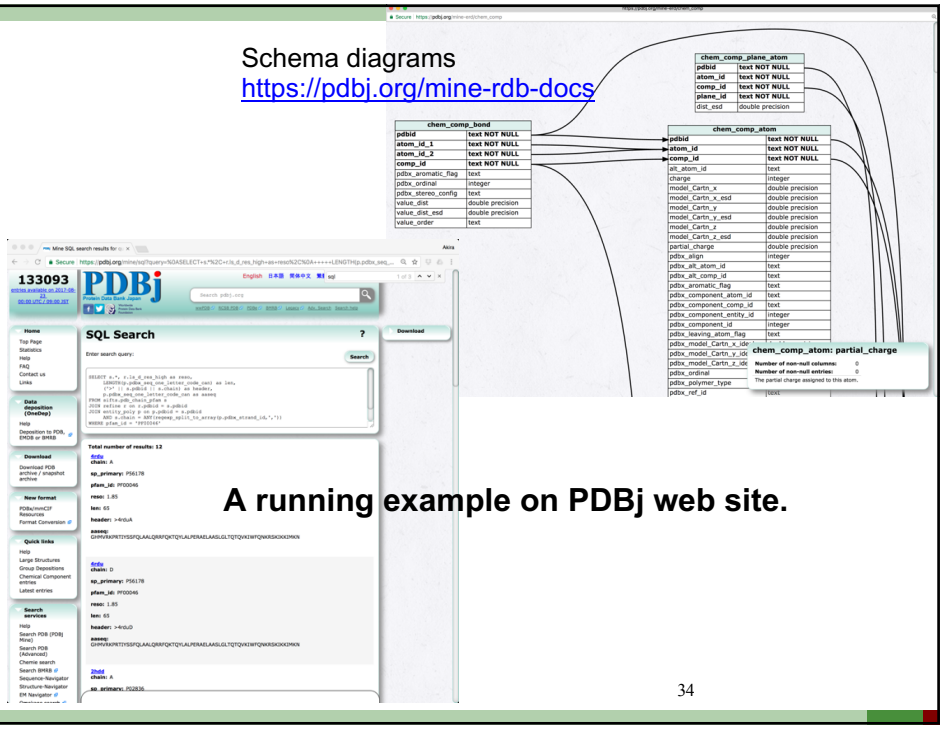
32

PDBj Mine2 RDB (https://pdbj.org)

- Relational database working behind the PDBj.
- Docs: https://pdbj.org/mine-rdb-docs
  - Complete database schema with diagrams.
- Web SQL interface: https://pdbj.org/mine
- REST API: https://pdbj.org/rest/mine2_sql
- SQL dump: ftp://ftp.pdbj.org/mine2/
  - Requires **PostgreSQL** >= 9.3
  - See https://pdbj.org/help/mine2-rdb-local-install
- Many examples: https://pdbj.org/help/mine2-sql

33

Schema diagrams
https://pdbj.org/mine-rdb-docs



A running example on PDBj web site.

34

Integration with SIFTS

- "Structure Integration with Function, Taxonomy and Sequence" developed by PDBe & UniProt.
  - https://www.ebi.ac.uk/pdbe/docs/sifts/
- Integrates *UniProt*, *NCBI Taxonomy*, *Gene Ontology*, *Pfam*, *EC code*, *PubMed*, *SCOP*, *CATH* with PDB.
- The "Quick access" data of SIFTS are integrated into the PDBj Mine RDB.
  - c.f., https://www.ebi.ac.uk/pdbe/docs/sifts/quick.html

35

New integration

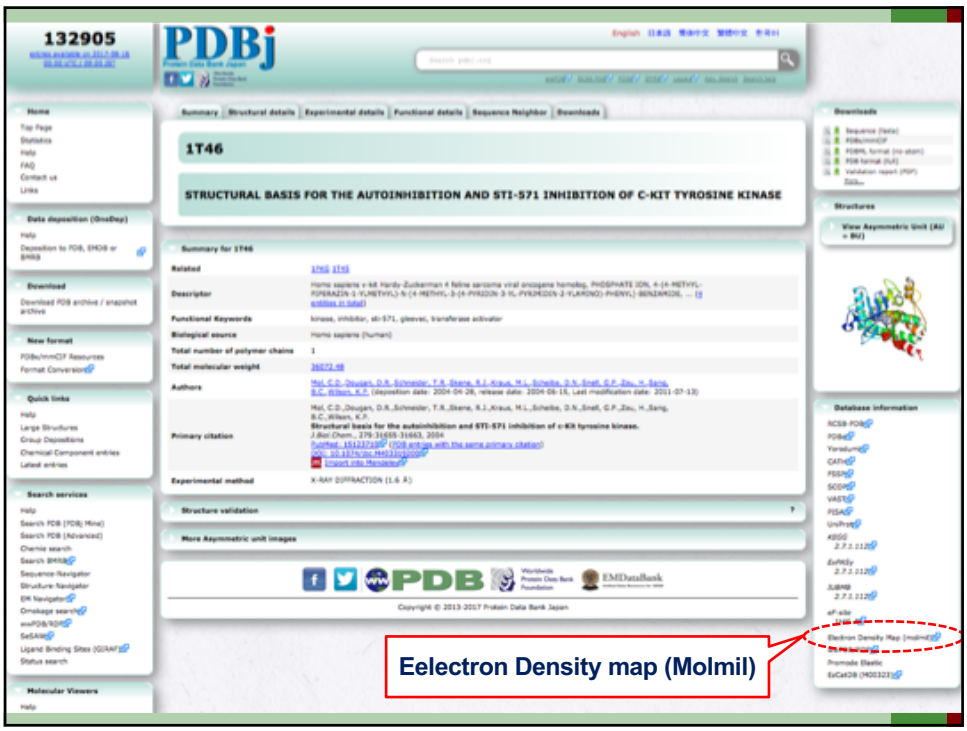
- Chemical Component Dictionary (cc)
  - PDB's (3-letter) chemical components
  - Includes InChi keys, SMILES, etc.
- Chemical Component Model Data (ccmodel)
  - Xref to *Cambridge Structure Database* (CSD)
- BIRD (prd)
  - “Biologically Interesting Molecule Reference Dictionary”
  - Peptide-like antibiotic and inhibitor molecules.

36

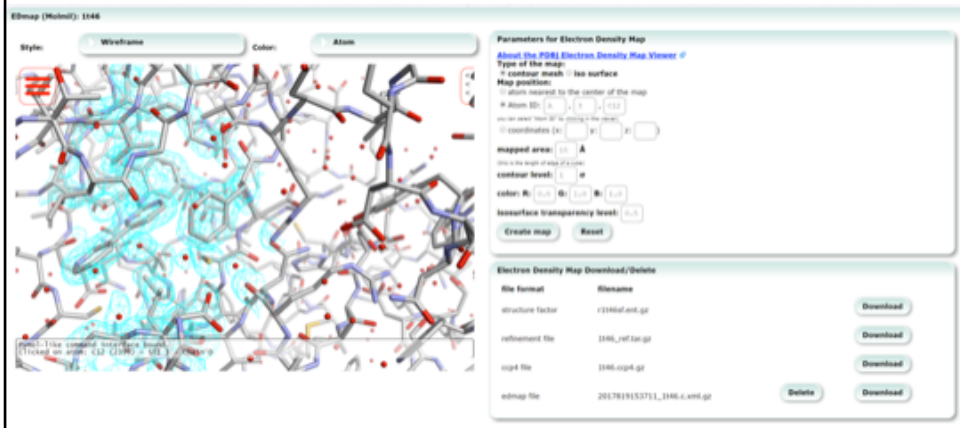
Example1: Gleevec using PDBj-Mine

37

Explore 1T46: Gleevec using PDBj-Mine (cont.)



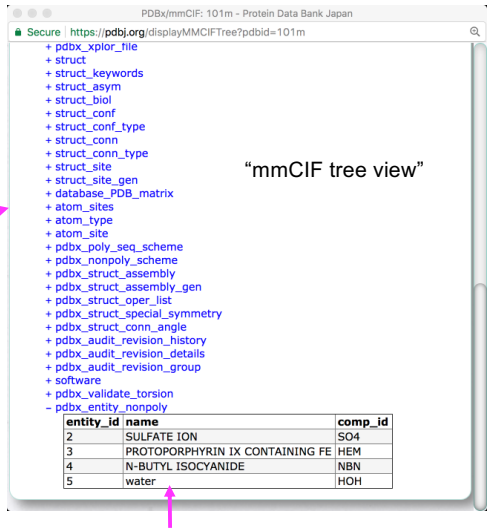
Explore 1T46: Gleevec using PDBj-Mine (cont.)



40

Example 2. PDB entries containing "HEM"

```
SELECT pdbid
FROM pdbx_entity_nonpoly
WHERE comp_id = 'HEM'
```



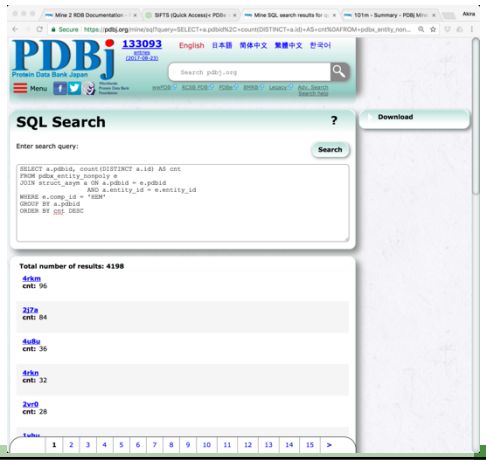
"mmCIF tree view"

"pdbx_entity_nonpoly" category

41

PDB entries containing “HEM” sorted by the number of HEM's in asymmetric unit

```
SELECT a.pdbid, count(DISTINCT a.id) AS cnt
FROM pdbx_entity_nonpoly e
JOIN struct_asym a ON a.pdbid = e.pdbid
                  AND a.entity_id = e.entity_id
WHERE e.comp_id = 'HEM'
GROUP BY a.pdbid
ORDER BY cnt DESC
```



Example 3. BIRD: Biologically Interesting Molecule Reference Dictionary

- See https://www.wwpdb.org/data/bird
- Antibiotics, inhibitors, etc.

In 1KQE:

```
_pdbx_molecule_features.prd_id     PRD_000154
_pdbx_molecule_features.name       'MINI-GRAMICIDIN A DIMER'
_pdbx_molecule_features.type       Polypeptide
_pdbx_molecule_features.class      Antibiotic
_pdbx_molecule_features.details
;THE N-TERMINI OF THE TWO IDENTICAL PEPTIDES, EACH
 A TRUNCATED GRAMICIDIN A WERE LINKED BY A SUCCINIC
 ACID IN A HEAD-TO-HEAD MANNER.
;
#
```

43

Combining with BIRD

Find PDB entries containing antibiotics
of molecular weight less than 1000 Da.

```
SELECT mf.pdbid, rm.name
FROM pdbj.pdbx_molecule_features mf
JOIN prd.pdbx_reference_molecule rm
       ON rm.prd_id = mf.prd_id
WHERE  rm.class = 'Antibiotic'
AND    rm.formula_weight < 1000.0
```

The “prd” schema (for some historical reasons...)

https://pdbj.org/mine-rdb-docs?schema=prd

44

List BIRD entries or their types according to popularity

```
SELECT prd_id, name, COUNT(pdbid)
FROM  pdbx_molecule_features
GROUP BY prd_id,name
ORDER BY COUNT DESC
```

Total number of results: 835

prd_id: PRD_000020
name: D-Phe-Pro-Arg-CH2Cl
count: 51

prd_id: PRD_000238
name: Ac-Asp-Glu-Val-Asp-CMK
count: 46

prd_id: PRD_000142
name: Cyclosporin A
count: 30

prd_id: PRD_000398
name: N-((2S)-2-{(N-acetyl-L-threonyl-L-isoleucyl)amino}hexyl)-L-norleucyl-L-glutaminyl-N~5~-{amino(iminio)methyl}-L-ornithinamide
count: 29

prd_id: PRD_001243
name: CARFILZOMIB, bound form
count: 28

prd_id: PRD_000454
name: Saquinavir
count: 27

prd_id: PRD_000557
name: Pepstatin
count: 24

```
SELECT type, COUNT(pdbid)
FROM pdbx_molecule_features
GROUP BY type
ORDER BY COUNT DESC
```

Total number of results: 17

type: Peptide-like
count: 1032

type: Oligopeptide
count: 146

type: Cyclic peptide
count: 130

type: Polypeptide
count: 117

type: Glycopeptide
count: 38

type: Cyclic depsipeptide
count: 23

type: Thiopeptide
count: 18

type: Peptaibol
count: 15

type: Non-polymer
count: 8

type: Cyclic lipopeptide
count: 3

type: Lipopeptide
count: 3

45

Example 4. Combining with CC model

Find PDB entries containing a compound corresponding to a Cambridge Structure Database (CSD) entry.

```
SELECT p.pdbid, p.id, p.name, r.db_code
FROM pdbj.chem_comp p
JOIN ccmodel.pdbx_chem_comp_model m
       ON m.comp_id = p.id
JOIN ccmodel.pdbx_chem_comp_model_reference r
       ON r.model_id = m.model_id
WHERE r.db_name = 'CSD' AND r.db_code = 'YARXEW'
```

The “ccmodel” schema.

https://pdbj.org/mine-rdb-docs?schema=ccmodel

46

Combining with CC

Find PDB entries containing monomers with the given InChIKey.

The “pdbj” schema is the default and can be omitted.

```
SELECT p.pdbid, p.id
FROM pdbj.chem_comp p
JOIN cc.pdbx_chem_comp_descriptor cc
       ON cc.comp_id = p.id
WHERE cc.type = 'InChIKey'
AND   cc.descriptor = 'ZKHQWZAMYRWXGA-KQYNXXCUSA-N'
```

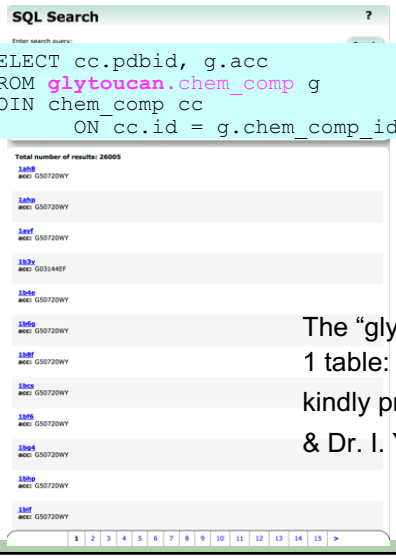
Chemical Component Dictionary entries are under the “cc” schema.

For complete information:
https://pdbj.org/mine-rdb-docs?schema=cc

47

Integration with GlyTouCan

Listing PDB ID's that contain GlyTouCan entries.



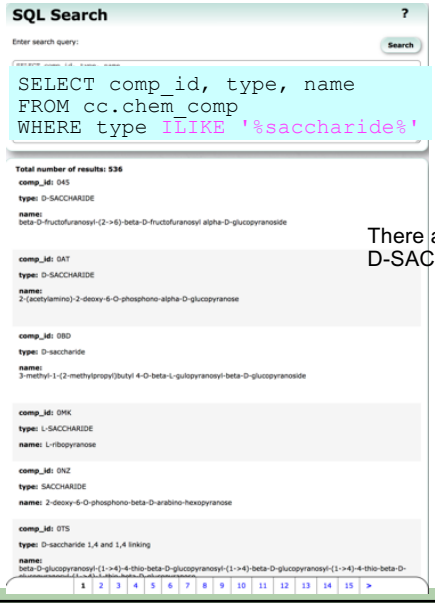

```
SELECT cc.pdbid, g.acc
FROM glytoucan.chem_comp g
JOIN chem_comp cc
        ON cc.id = g.chem_comp_id
```

The “glytoucan” schema contains only 1 table: chem_comp

kindly provided by the GlyTouCan team & Dr. I. Yamada.

48

Another CC example (sugars!)




```
SELECT comp_id, type, name
FROM cc.chem_comp
WHERE type ILIKE '%saccharide%'
```

Find all “**saccharides**” from the Chemical Component Dictionary.

There are too many varieties in annotation: D-SACCHARIDE, D-saccharide, etc. :(

49

If you want to do complicated queries,
we may be able to help!

Feel free to ask any questions at:
https://pdbj.org/contact?tab=PDBjmaster

50

Acknowledgements

51