PDBjとwwPDBの活動方針について
Recent Activities of PDBj and wwPDB

栗栖源嗣（Genji Kurisu）
大阪大学蛋白質研究所，日本蛋白質構造データバンク
（PDBj, Institute for Protein Research, Osaka University）

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)

CRYSTALLOGRAPHY

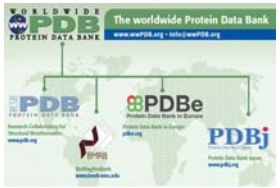
Protein Data Bank

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 accepted either "raw" or refined, in machine-readable form or as manuscripts. Laboratories intending to join the scheme should communicate with Mrs Olga Kennard or Dr

Protein Data Bank Japan
http://pdbj.org/

Since 2001, PDBj has been managed at Institute for Protein Research, Osaka University as a member of the wwPDB, to curate and process the deposited data for an open and single archive.



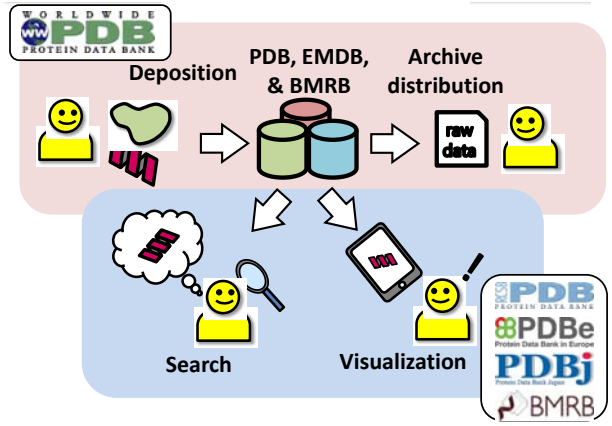
Worldwide Protein Data Bank (wwPDB)

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

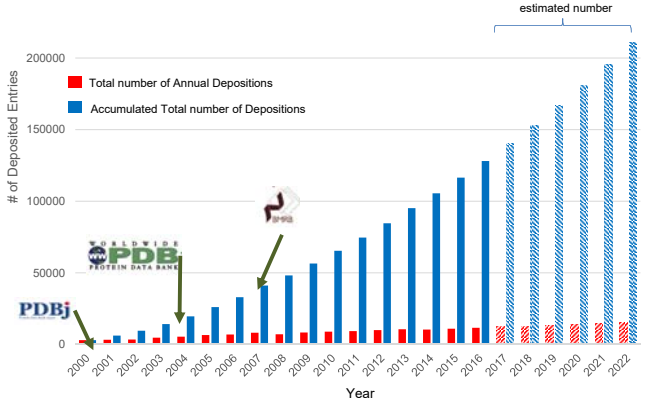
annotation

- Each site provides different websites that offer different services and views of the data

wwPDB collaboration



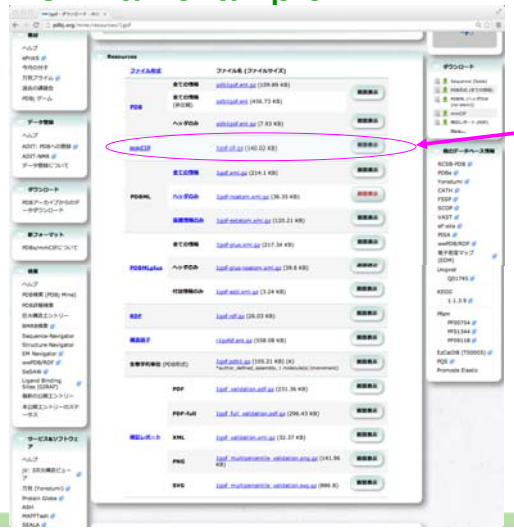
Growing Number of Depositions



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).

mmCIF: an example



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
```

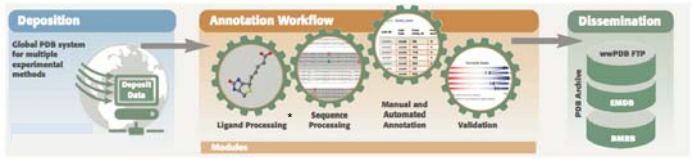

File Versioning:

Current Issues:

- Loss of connection between PDB ID and Publication under current wwPDB Obsolete/Supersede Policy

- Enable revisions to entries updated by the Depositor of Record (e.g., Version 1-0 → 1-1; 1-0 → 2-0)
  - wwPDB will NOT assign a new PDB ID going forward (for Depositor of Record revision only)
- Introduce new PDB ID code format
  - With PDB prefix and extension of 4 characters

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, …)

wwPDB Common Deposition & Annotation

https://wwpdb.org

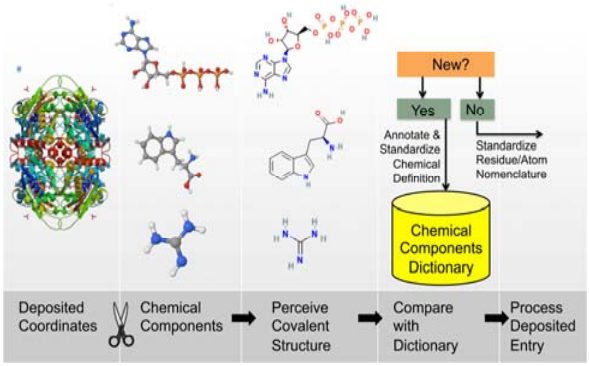
OneDep

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	2097	158	528
2001	3297	2408	393	496
2002	3965	2401	697	907
2003	4830	3136	1026	669
2004	5658	3082	1614	812
2005	6678	3563	2110	1005
2006	7282	4252	1946	1086
2007	8130	4703	2290	1129
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	10666	6652	2129	1786
2014	10364	6038	1781	2545
2015	10958	4845	2100	4013
2016	11614	5326	2238	4050
2017	2577	1579	394	604
TOTAL	131815	77266	26745	25804

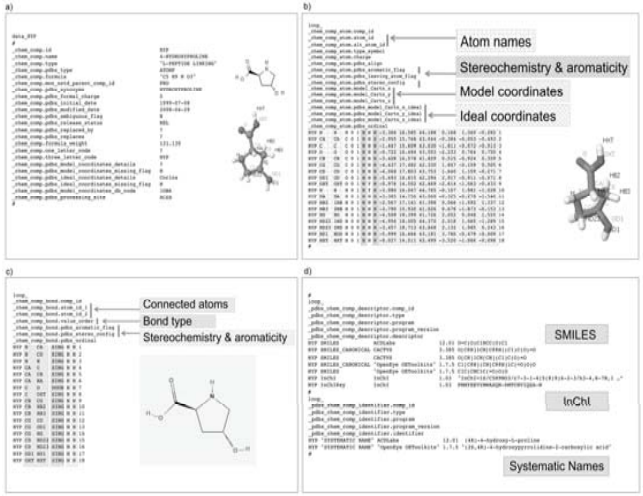
Chemical Component Deposition Pipeline



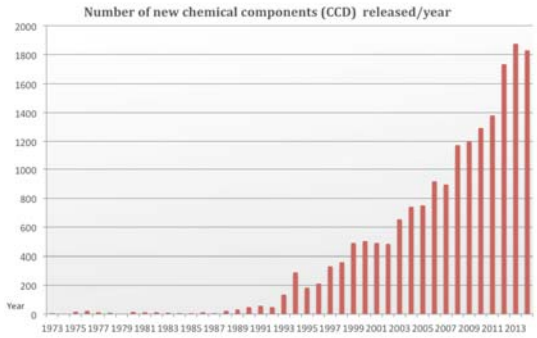
Chemical Component Dictionary (CCD)

- 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

Chemical Component Dictionary Entry



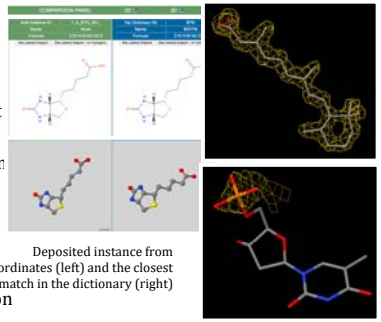
Growth of Chemical Components in PDB



Growth trends new Chemical Components in the PDB

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**)

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

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]

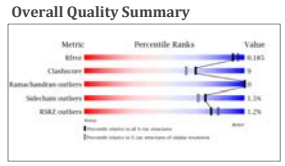
(Author list continued on next page)

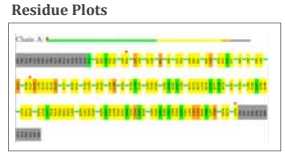
Improved Validation

- 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)

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





Validation Report is requested for peer review

EDITORIAL

nature structural & molecular biology

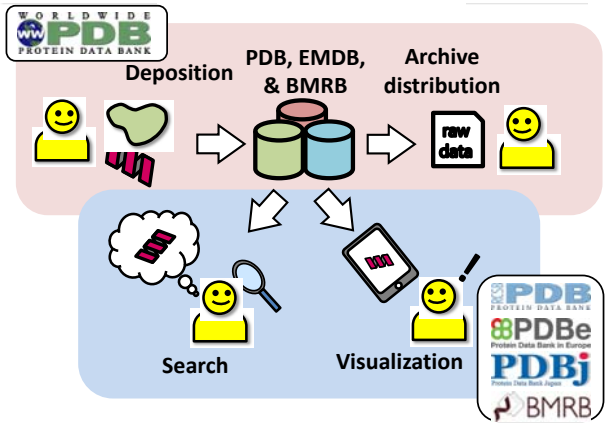
Nature Struct. Mol. Biology, 23 (10), 871, 2016

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.

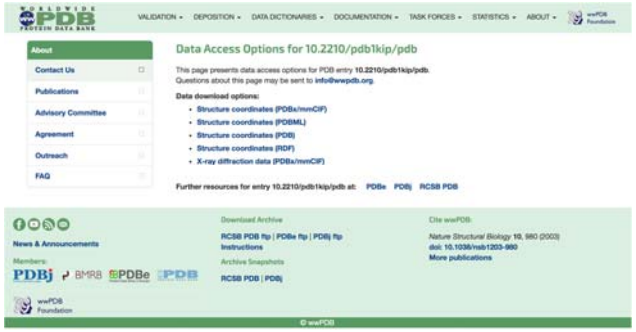
We are now taking a further step and are requesting official wwPDB validation reports for peer review. These reports are made available by the wwPDB after data deposition (http://www.wwpdb.org/validation/validation-reports). Other Nature journals will soon follow suit.

wwPDB collaboration



Worldwide Protein Data Bank (wwPDB)

DOI Landing Page Layout (Planned)



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

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

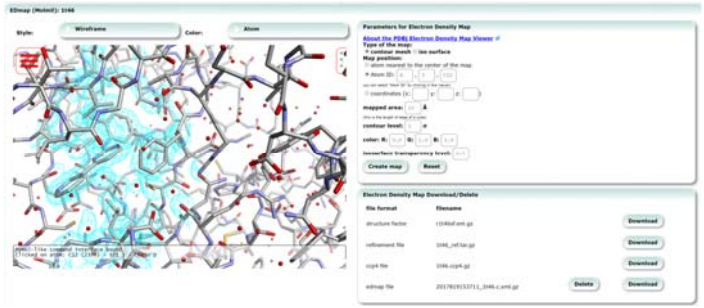
New integration

- 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

Example1: Gleevec using PDBj-Mine

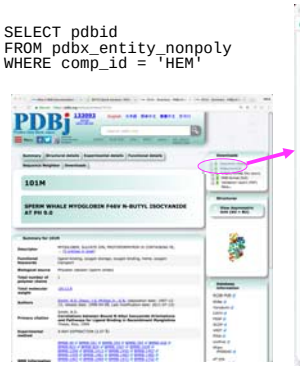
Explore 1T46: Gleevec using PDBj-Mine (cont.)

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



Example 2. PDB entries containing “HEM”

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



“mmCIF tree view”

“pdbx_entity_nonpoly” category

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.
;
#
```

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

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
```

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

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

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

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

Acknowledgements