How can PDB submission be made more automatic

Sameer Velankar
Protein Data Bank in Europe
Sustain freely accessible, interoperating Core Archives of structure data and metadata for biological macromolecules as an enduring public good to promote basic and applied research and education across the sciences.

- Manage the wwPDB Core Archives as a public good according to the FAIR Principles.
- Provide expert deposition, validation, biocuration, and remediation services at no charge to Data Depositors worldwide.
- Ensure universal open access to public domain structural biology data with no limitations on usage.
- Develop and promote community-endorsed data standards for archiving and exchange of global structural biology data.
PDBx/mmCIF Working Group

- PDBx/mmCIF is the archival data standard for the repository
- wwPDB together with the PDBx/mmCIF Working Group of community experts and methods developers oversee the evolution of the standard
- Working Group ensures that the standard is well supported by key community software tools.
- PDB hosts community workshops and maintains mmcif.wwpdb.org serving PDBx/mmCIF data dictionaries, schema and software tools.
Protein Data Bank Deposition

- Open Access to single global archive
  - Weekly release

- Regional sharing of data deposition efforts
  - RCSB PDB (US)
  - PDB Japan (PDBj)
  - PDB in Europe (PDBe)

- Partner-hosted websites offer complementary services and views of data

Protein Data Bank: the single global archive for 3D macromolecular structure data
Protein Data Bank Deposition

- Deposition
  - deposit.wwpdb.org

- Validation – validate your structures before deposition
  - validate.wwpdb.org

- Validation API
  - https://www.wwpdb.org/validation/onedep-validation-web-service-interface
Growth/Complexity

- 13,377 depositions in 2019
- Rapid growth in 3DEM
- More structures per Depositor
- Increasing data complexity
  - Increase in large complexes resolved by EM

2019 Processing Sites

- RCSB PDB: 41%
- PDBj: 25%
- PDBe: 34%

2019 Depositor Locations

- North America: 32%
- Europe: 25%
- South America: 32%
- Asia: <1%
Protein Data Bank Deposition

- How does deposition work now?

- Manual curation review within OneDep
## What data is captured?

<table>
<thead>
<tr>
<th>Deposition type</th>
<th>Database accession codes issued</th>
<th>Mandatory file uploads</th>
<th>Optional file uploads</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-ray and neutron crystallography</td>
<td>PDB</td>
<td>Atomic coordinates Structure factor data Unmerged intensity data</td>
<td>Ligand definition file or image Auxiliary files</td>
</tr>
<tr>
<td>Solution and solid-state NMR</td>
<td>PDB BMRB</td>
<td>Atomic coordinates Assigned chemical shifts Restraints used in refinement Auxiliary sequence file from AMBER</td>
<td>Spectral peak lists Ligand definition file or image Auxiliary files</td>
</tr>
<tr>
<td>Electron crystallography</td>
<td>PDB EMDB</td>
<td>Atomic coordinates Structure factor data or Mass density map volume</td>
<td>Ligand definition file or image Auxiliary files</td>
</tr>
<tr>
<td>3DEM (map and model)</td>
<td>PDB EMDB</td>
<td>Atomic coordinates Mass density map volume Entry image for public display (EMDB)</td>
<td>Any number of additional maps Any number of masks Two half maps Fourier shell correlation (FSC) curve Ligand definition file or image</td>
</tr>
<tr>
<td>3DEM (map only)</td>
<td>EMDB</td>
<td>Mass density map volume Entry image for public display</td>
<td>As above</td>
</tr>
</tbody>
</table>
What data is captured?

▪ Administrative information (e.g., author release instructions).
▪ Description of each distinct macromolecule present in the sample.
▪ Description of the experimental setup (e.g., sample preparation and data collection).
▪ Description of experimental data, refinement (e.g., crystallographic refinement statistics), and software used.
▪ Description and matching of ligands and modified polymer residues to the PDB Chemical Component Dictionary (CCD).
▪ Information on the quaternary structure and, whenever possible, experimental support for the biologically relevant assembly.
Deposition of raw data

Starting a deposition session

Can I deposit raw data to the PDB?

Depositors are encouraged to deposit raw data of their structure and then provide the corresponding DOI in the “Related Entries” user interface during structure deposition in wwPDB OneDep.

Depositors are strongly encouraged to deposit their raw data in a curated archive.

- X-ray diffraction
  - ProteinDiffraction.org - https://proteindiffraction.org
  - SBGrid - https://sbgrid.org
  - CIXID - http://www.cixid.org
- Electron microscopy
  - EMBL-EBI - https://www.ebi.ac.uk/pdbe/emdb/empiar/deposition
- NMR e.g., FIDs, in particular those relating to NOESY type spectra
  - please contact BMRB - bmrbhelp@bmrb.wisc.edu

Related experimental data sets

Please provide the DOI for a related raw data set (e.g. diffraction image data with a DOI, not related PDB entries or related citations).

Type of experimental data

<table>
<thead>
<tr>
<th>DOI for related data set</th>
<th>DOI for metadata</th>
</tr>
</thead>
</table>

Your deposition has been submitted

Thank you for your deposition at the wwPDB.

The PDB code for this deposition is 6Y47.

If you have additional raw X-ray diffraction data, please deposit at one of the raw image archives and inform us of the DOI.

proteinDiffraction.org
SBGrid
CIXID
Group deposition
https://deposit-group.rcsb.rutgers.edu/groupdeposit/

- Provided support for D3R Blind Challenges
- Early Adopters: Roche, Merck Serono, U. Marburg, U. Essex
- Requires a complete PDBx/mmCIF file with all mandatory data
- Each group needs to implement the necessary infrastructure to generate complete PDBx/mmCIF file
Mandatory submission of PDBx/mmCIF – June ’19

- Increase in the amount of meta data collected in the entries during deposition
- Work with Software community to develop automated process for deposition
- Improved consistency in biocuration between related entries
- Increased biocuration efficiency
- Curating an increasing number of entries with the same resources
Deposition of associated experimental data

- Experience of API’s within OneDep
- Interaction with SASBDB during deposition
- API services to interact with federated deposition systems
SFX/XFEL are Revolutionizing MX

- **SFX**: Serial Femtosecond X-ray crystallography
- **XFEL**: X-ray Free Electron Laser
- Time-resolved studies
  - Conformational changes in enzyme active sites
  - Response to photoexcitation
  - Membrane protein dynamics
- Rapid advances in data collection and sample handling technologies
- Linking of multiple depositions (“investigations”)

**Image:**
- **BioXFEL**
- **PAUL SCHerrer INSTITUT**
- **SwissFEL**
- **LCLS**
- **European XFEL**
- **SACLs**
- **SwissFEL**
- **Paul Scherrer Institut**
- **European XFEL**
- **SwissFEL**
Automated deposition

- Improved fidelity and completeness of 3D structure data deposited into the PDB
- Streamlining the wwPDB data deposition, validation, and biocuration system.
- Provide easy mechanism for deposition of multiple structures (SF/XFEL, Fragment screening)
- Improved efficiency for depositors and biocuration
- Consistent biocuration
How can PDB deposition be made more automatic

APIs to expose wwPDB deposition/annotation pipeline
New three-year project (PDBe and RCSB PDB)

- Funded by BBSRC and NSF
- Start date – July 2020 (RCSB PDB); Jan 2021 (PDBe)

- Gather requirements on Deposition API requirements
  - Automated deposition
  - Multi-structure deposition/annotation
    - “investigation” based deposition
    - “investigation” based annotation – Annotators to provide requirements
New three-year project (PDBe and RCSB PDB)

- Data Content Extension:
  - “Investigation” level semantics
  - Integrate validation data in mmCIF

- New Deposition Data Preparation APIs:
  - pre-deposition registration service for chemical reference data
    - Better validation of ligands before and during deposition
Questions and discussion