

PDBx/mmCIF Ecosystem for NMR Structures

Hamid R. Eghbalnia, Kumaran Baskaran, Jonathan R. Wedell, Hongyang Yao,

Dimitri Maziuk, Michael M. Gryk, and Jeffrey C. Hoch

Department of Molecular Biology and Biophysics, UConn Health,

263 Farmington Avenue, Farmington, CT 06030, USA



The Macromolecular Crystallographic Information File (mmCIF), also known as PDBx/mmCIF¹, is a standard text file format for representation and exchange of experimentally determined three-dimensional (3D) macromolecular structure data. It is the adopted standard for all data processing and annotation by the Worldwide Protein Data Bank (wwPDB, wwpdb.org). Because mmCIF can be easily extended to include representations that characterize unique structural features such as disorder, multiple conformers, and dynamic parameters, it is extremely well-suited for aligning data from NMR studies with atomic coordinates. PDB-Dev (<https://pdb-dev.wwpdb.org>) uses the mmCIF ecosystem to represent biological assemblies derived using integrative and hybrid methods. The dictionary and format are machine-readable, assuring that the wwPDB partners can more effectively represent, biocurate, validate and distribute structural biology data.

wwPDB data deposition

Model data
PDBx/mmCIF
coordinate data and
meta data

Experimental data
NMR-STAR/NEF
Chemical shifts,
restraints, peak list and
other experimental data

wwPDB accepts data coming from structure determination studies whereas BMRB accepts all additional NMR related data including raw experimental data.

PDBx/mmCIF format

Standardization
Format, syntax and
data standards

Extensibility
Easily extendable to
support different
methods like NMR,
EM, etc..

PDBx/mmCIF

Metadata
Dictionary can be
extended to support
any type of metadata

FAIR
Findable, Accessible,
Interoperable and
Reusable

New extended PDB ID

- All the four character PDB accession codes (PDB ID) will be consumed by 2028
- New extended PDB ID (12 characters) has been introduced (example: pdb_00001abc)
- PDB entries with extended PDB IDs issued will not be compatible with legacy PDB format
- A 5-years plan will be provided to help users transition from legacy PDB format to PDBx/mmCIF format
- Resources and training materials made available (2024-2028) <https://www.wwpdb.org/documentation/new-format-for-pdb-ids>
- "beta" PDB archive organized using extended PDB IDs will be provided for user adoption (2026), which becomes main archive (2027)
- wwPDB encourages scientific journals, PDB users and depositors to transition to mmCIF format and adopt new PDB ID format as earlier as possible.

PDBx/mmCIF example

The figure below illustrates a partial mapping between the legacy PDB and PDBx/mmCIF file formats with a canonical zinc finger domain structure from PDB ID 1ZAA.

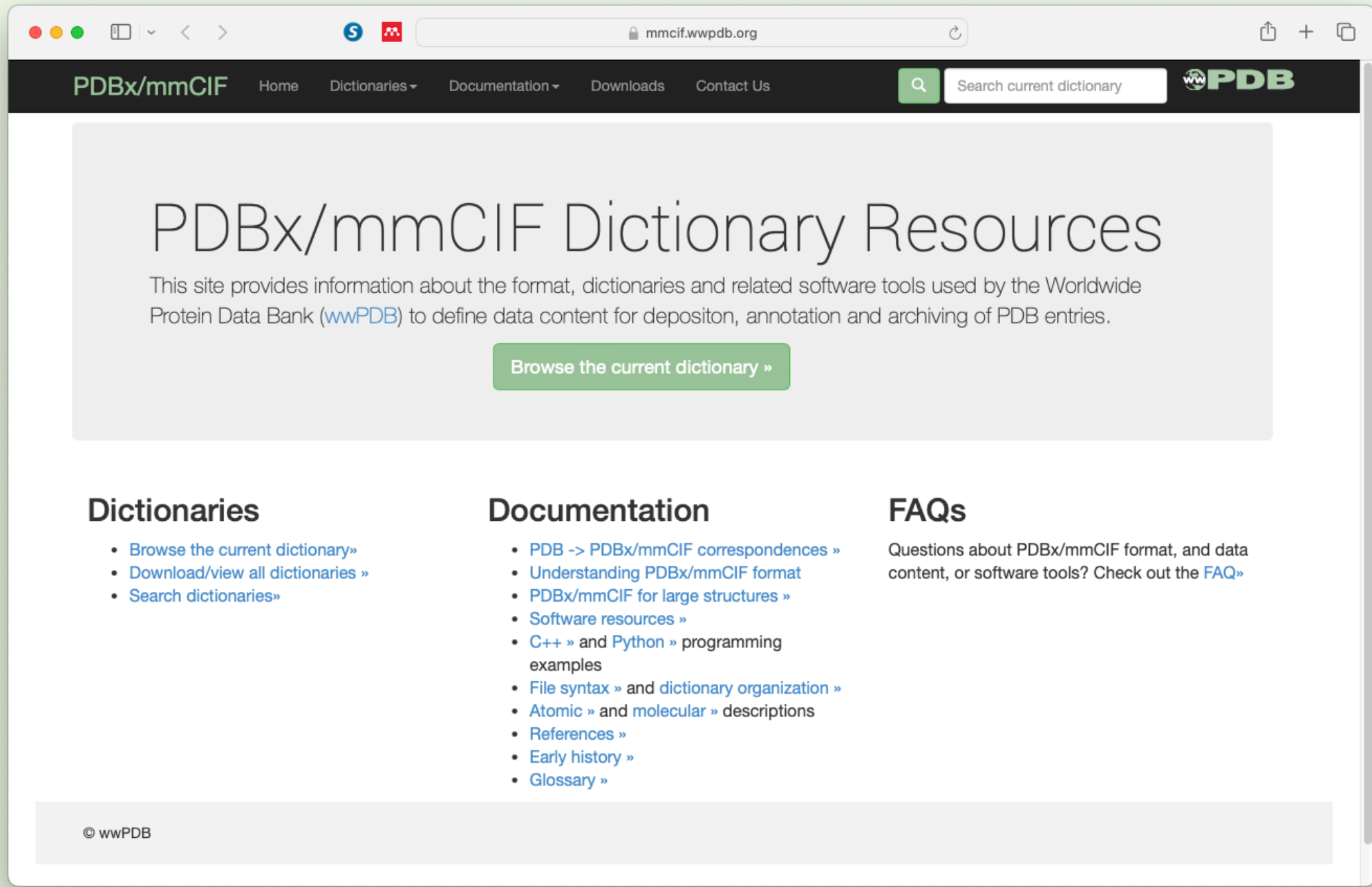
```
A
loop
_struct_conn_id
_struct_conn_type_id
_struct_conn.ptnr1_label_asym_id
_struct_conn.ptnr1_label_comp_id
_struct_conn.ptnr1_label_seq_id
_struct_conn.ptnr1_label_atom_id
_struct_conn.ptnr1_symmetry
_struct_conn.ptnr2_label_asym_id
_struct_conn.ptnr2_label_comp_id
_struct_conn.ptnr2_label_seq_id
_struct_conn.ptnr2_label_atom_id
_struct_conn.ptnr2_auth_asym_id
_struct_conn.ptnr2_auth_comp_id
_struct_conn.ptnr2_auth_seq_id
_struct_conn.ptnr2_symmetry
_struct_conn.pdbx_dist_value
_struct_conn.pdbx_value_order
metal1 metalic D ZN . ZN 1_555 C HIS 25 NE2 C ZN 201 C HIS 25 1_555 2.138
metal2 metalic D ZN . ZN 1_555 C CYS 7 SG C ZN 201 C CYS 7 1_555 2.232
metal3 metalic D ZN . ZN 1_555 C CYS 12 SG C ZN 201 C CYS 12 1_555 2.440
metal4 metalic D ZN . ZN 1_555 C HIS 29 NE2 C ZN 201 C HIS 29 1_555 1.876
#
_cell.entry_id 1ZAA
_cell.length_a 45.400
_cell.length_b 56.200
_cell.length_c 130.800
_cell.angle_alpha 90.00
_cell.angle_beta 90.00
_cell.angle_gamma 90.00
_cell._PDB 8
#
_atom_sites.entry_id 1ZAA
_atom_sites.fract_transf_matrix[1][1] 0.022026
_atom_sites.fract_transf_matrix[1][2] 0.000000
_atom_sites.fract_transf_matrix[1][3] 0.000000
#
LINK ZN ZN 201 NE2 HIS 25 1555 1555 2.14
LINK ZN ZN 201 SG CYS 7 1555 1555 2.23
LINK ZN ZN 201 SG CYS 12 1555 1555 2.44
LINK ZN ZN 201 NE2 HIS 29 1555 1555 1.88
CRYST1 45.400 56.200 130.800 90.00 90.00 90.00 C 2 2 21 8
SCALE1 0.022026 0.000000 0.000000 0.000000
SCALE2 0.000000 0.017794 0.000000 0.000000
SCALE3 0.000000 0.000000 0.007645 0.000000

B
LINK ZN ZN 201 NE2 HIS 25 1555 1555 2.14
LINK ZN ZN 201 SG CYS 7 1555 1555 2.23
LINK ZN ZN 201 SG CYS 12 1555 1555 2.44
LINK ZN ZN 201 NE2 HIS 29 1555 1555 1.88
CRYST1 45.400 56.200 130.800 90.00 90.00 90.00 C 2 2 21 8
SCALE1 0.022026 0.000000 0.000000 0.000000
SCALE2 0.000000 0.017794 0.000000 0.000000
SCALE3 0.000000 0.000000 0.007645 0.000000
```

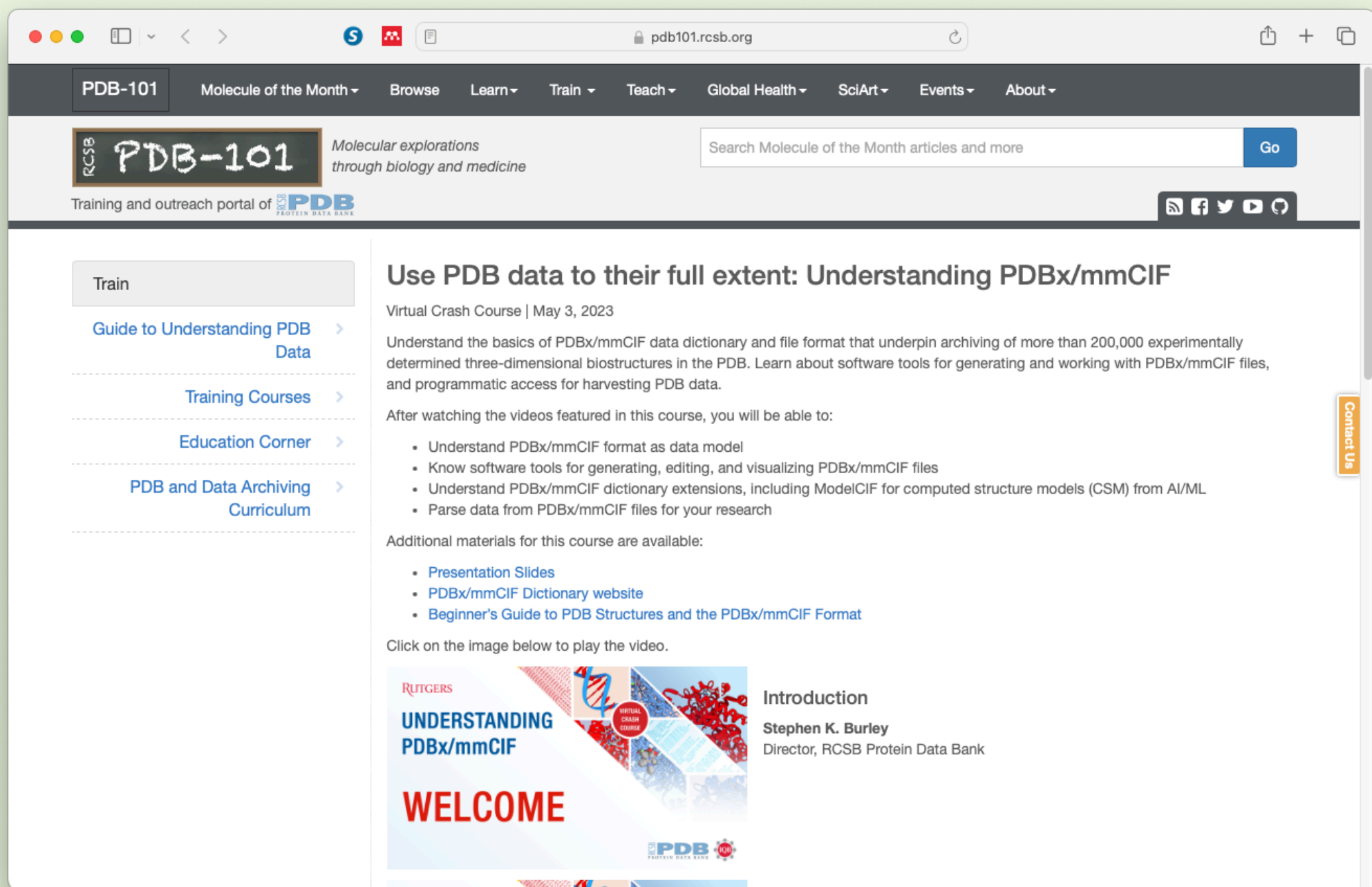
(A) Partial PDBx/mmCIF file for PDB ID 1ZAA. N.B.: Every data value has a key and multiple rows of data may be described in a table. The yellow highlighting describes the category and attributes. For the _struct_conn category, green depicts the residue numbers and cyan the component type. (B) Equivalent metadata records in legacy PDB format. Similar color coding depicts the mapping between category keys and record names as in (A), with LINK records highlighting the residue number and cyan the chemical component type. Inset figure, one of the zinc finger domains in 1ZAA depicting the side-chains that interact with the bound zinc ion codified in (A) and (B).

PDBx/mmCIF resource

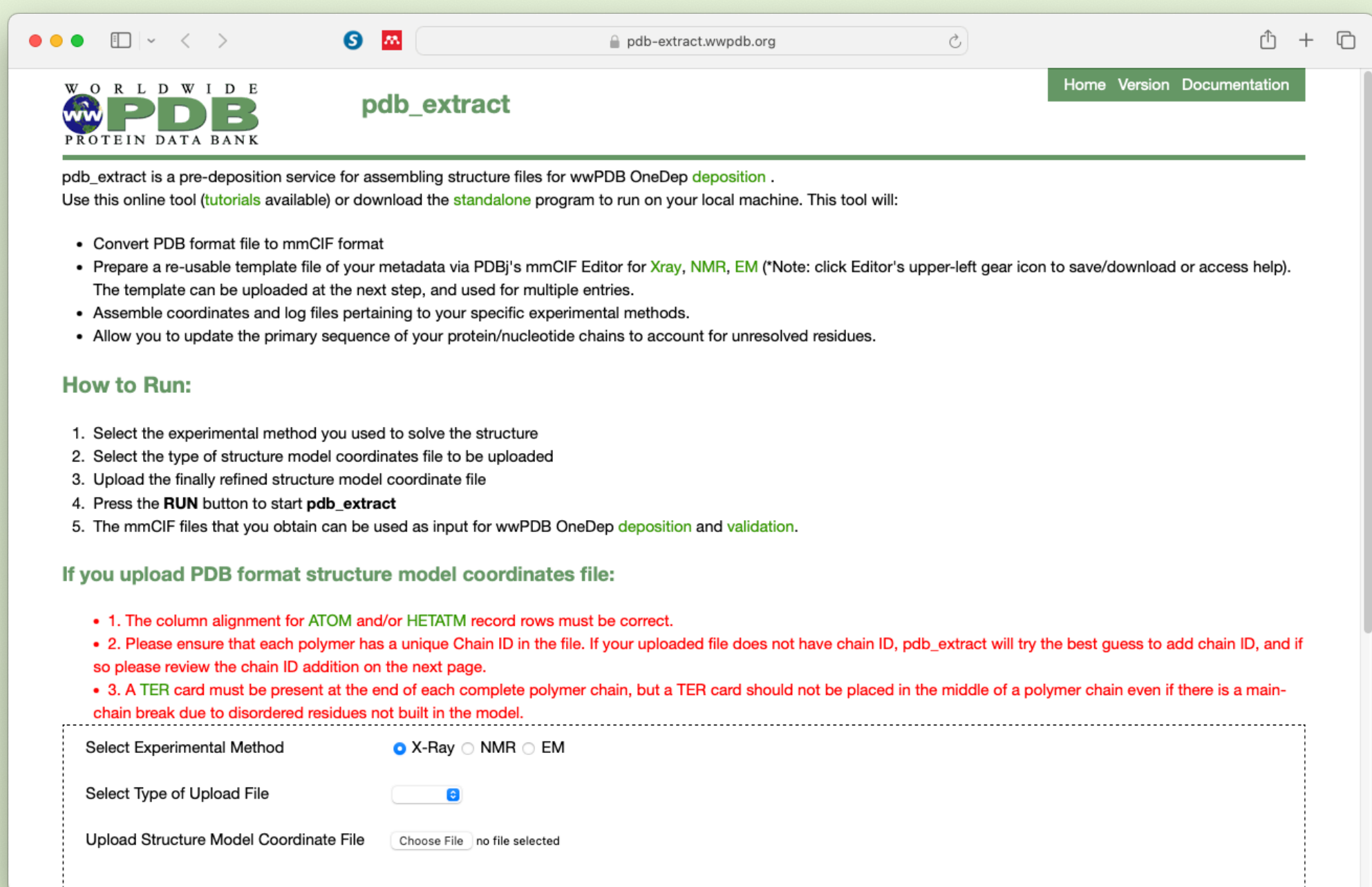
Dictionary resource <https://mmcif.wwpdb.org/>



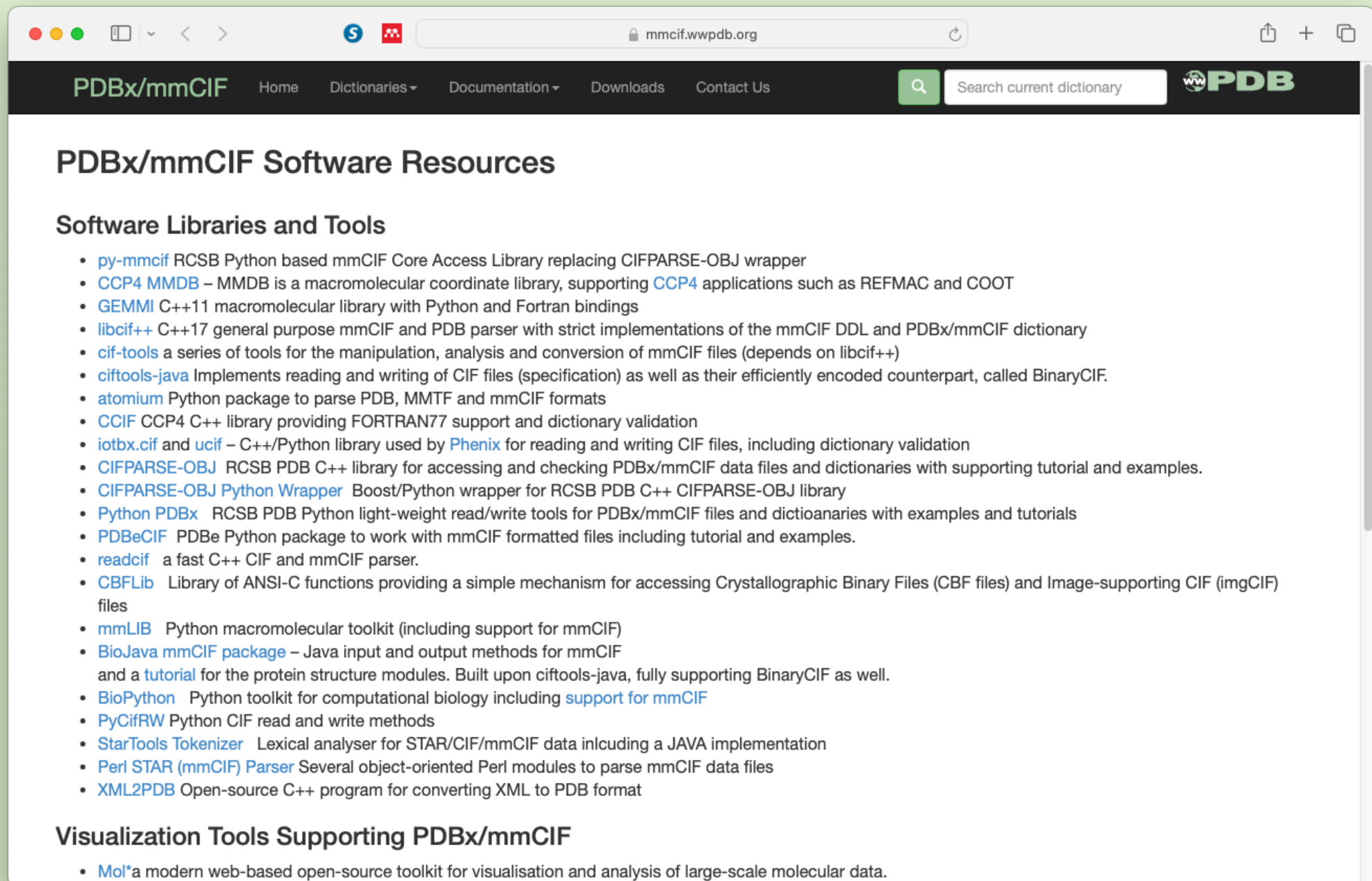
Virtual crash course <https://pdb101.rcsb.org/train/training-events/mmcif>



Create your mmCIF <https://pdb-extract.wwpdb.org/>



Software resource <https://mmcif.wwpdb.org/docs/software-resources.html>



NMR-STAR data model

BMRB uses NMR-STAR² data model to represent NMR data derived from various NMR experiments. NMR-STAR is a tag-value format similar to PDBx/mmCIF driven by NMR-STAR data dictionary. The PDBx/mmCIF and the NMR-STAR data dictionaries share common tags used to represent structure models. The NMR-STAR dictionary is focused more towards modeling NMR experimental data and PDBx/mmCIF dictionary includes data models relevant to other experimental techniques like X-Ray crystallography and Cryo-EM. The only difference between NMR-STAR and PDBx/mmCIF is that NMR-STAR uses save frames to organize and group data items.

NMR-STAR is the archival format for NMR data at the world wide Protein Data Bank

Why PDBx/mmCIF?

- **Flexibility:** mmCIF format allows for more flexibility in representing complex structural data
- **Richer Metadata:** mmCIF provides a more extensive set of data fields and metadata descriptors
- **Improved Representation of Biomolecular Assemblies:** mmCIF supports more sophisticated descriptions of macromolecular assemblies, including quaternary structures and complexes
- **Support for Large Structures:** mmCIF is better suited for representing large and complex structures, such as those determined by NMR spectroscopy
- **Future-Proofing:** mmCIF is a more modern and extensible format compared to the PDB format, which has limitations in representing certain types of structural data. By adopting mmCIF, researchers can future-proof their data and ensure compatibility with evolving standards and technologies

Overall, the mmCIF format offers a more comprehensive and flexible framework for representing NMR structures, allowing for better integration, analysis, and exchange of structural data within the scientific community.

Join us!

BMRB is leading the effort to extend the PDBx/mmCIF data model to faithfully represent the structure models derived from NMR studies. In association with wwPDB partners, BMRB is in the process of forming a PDBx/mmCIF working group for the NMR community. If you are interested to contribute to the data modeling efforts and become a NMR PDBx/mmCIF working group member, feel free to contact BMRB at help@bmrbo.io

1. Westbrook, et al., (2022) PDBx/mmCIF Ecosystem: Foundational Semantic Tools for Structural Biology JMB 434: 167599 doi: 10.1016/j.jmb.2022.167599
2. Ulrich, et al. NMR-STAR: comprehensive ontology for representing, archiving and exchanging data from nuclear magnetic resonance spectroscopic experiments. J Biolomol NMR 73, 5–9 (2019). doi: 10.1007/s10858-018-0220-3

PDF version of the poster



BMRB is supported by NIGMS through grant R24GM150793