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/mmCIF[1], 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.

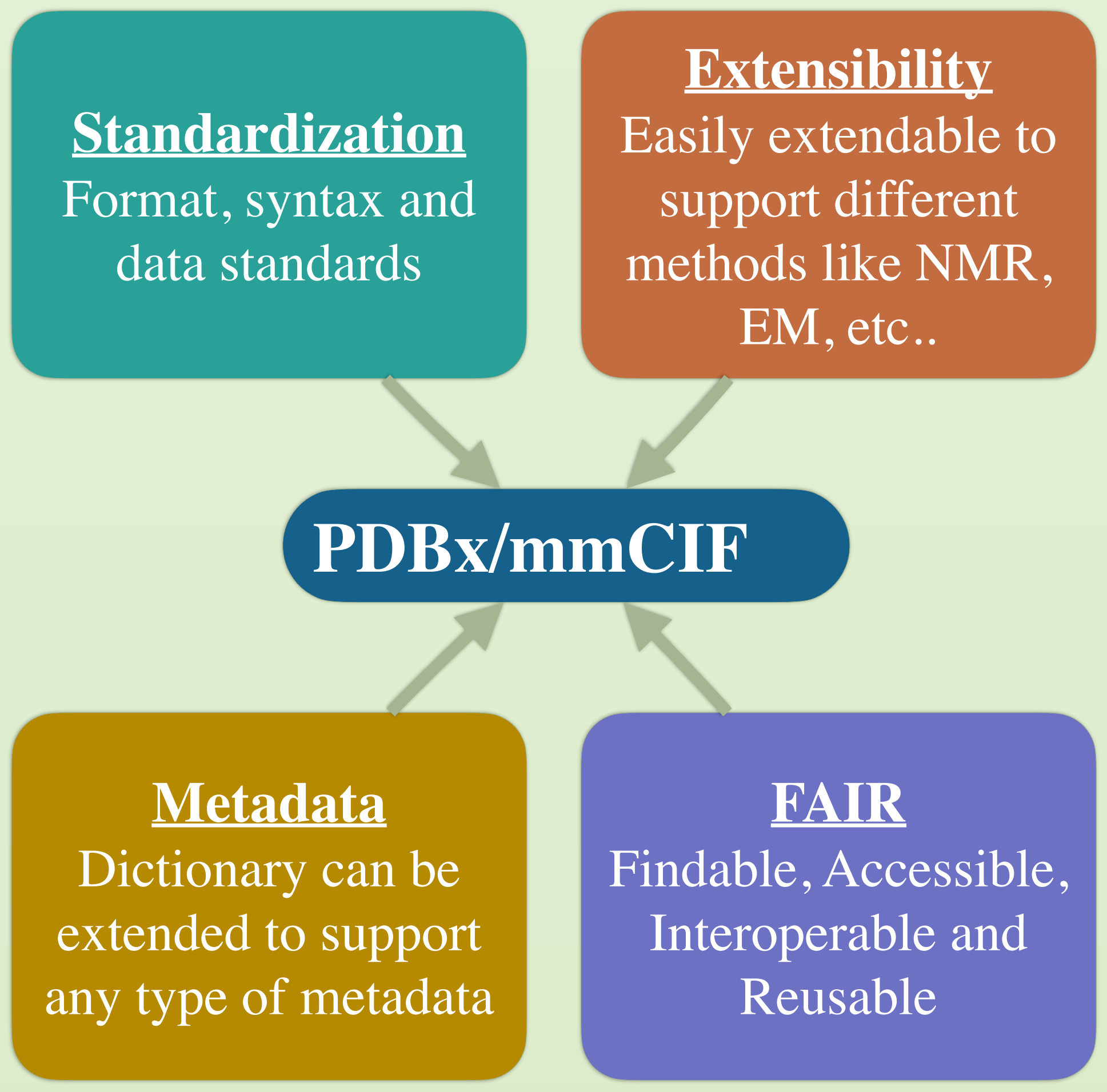
wwPDB data deposition

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

PDF version of the poster

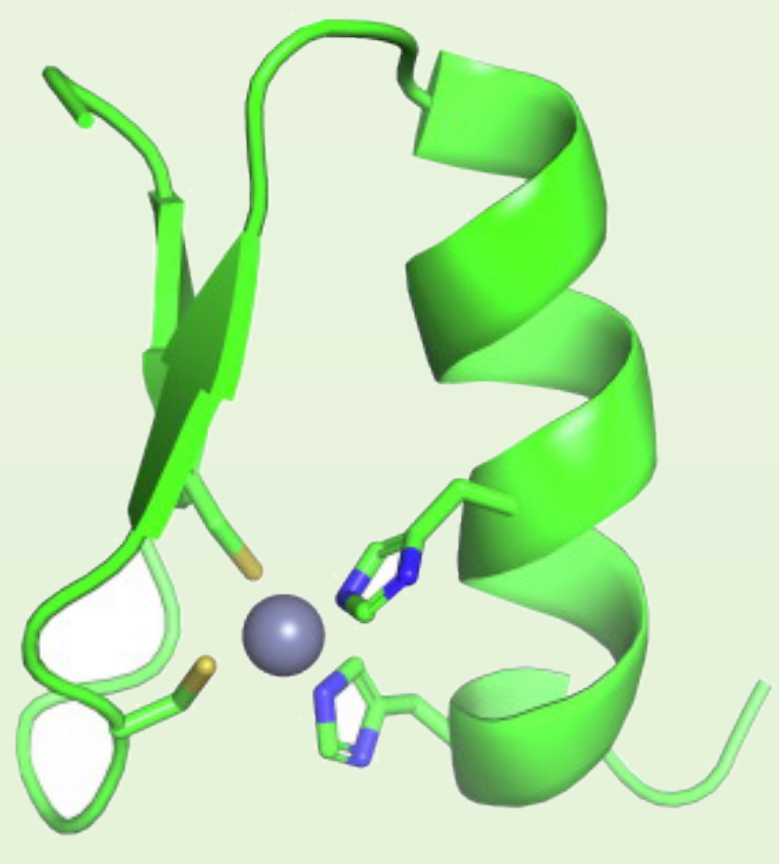
Limitations of legacy PDB format

- Fixed size columns
- Column informations hard coded
- Limited scope for meta data
- Can't be used represent large molecules like viruses
- No support to define biological assemblies
- No way define ligand interaction
- No way to define ensemble properties
- No way to represent conformers

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.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.ptnr1_auth_asym_id
_struct_conn.ptnr1_auth_comp_id
_struct_conn.ptnr1_auth_seq_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
metalc1 metalc D ZN . ZN 1_555 C HIS 25 NE2 C ZN 201 C HIS 25 1_555 2.138
metalc2 metalc D ZN . ZN 1_555 C CYS 7  SG  C ZN 201 C CYS 7  1_555 2.232
metalc3 metalc D ZN . ZN 1_555 C CYS 12 SG  C ZN 201 C CYS 12 1_555 2.440
metalc4 metalc 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.Z_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
. . .
```

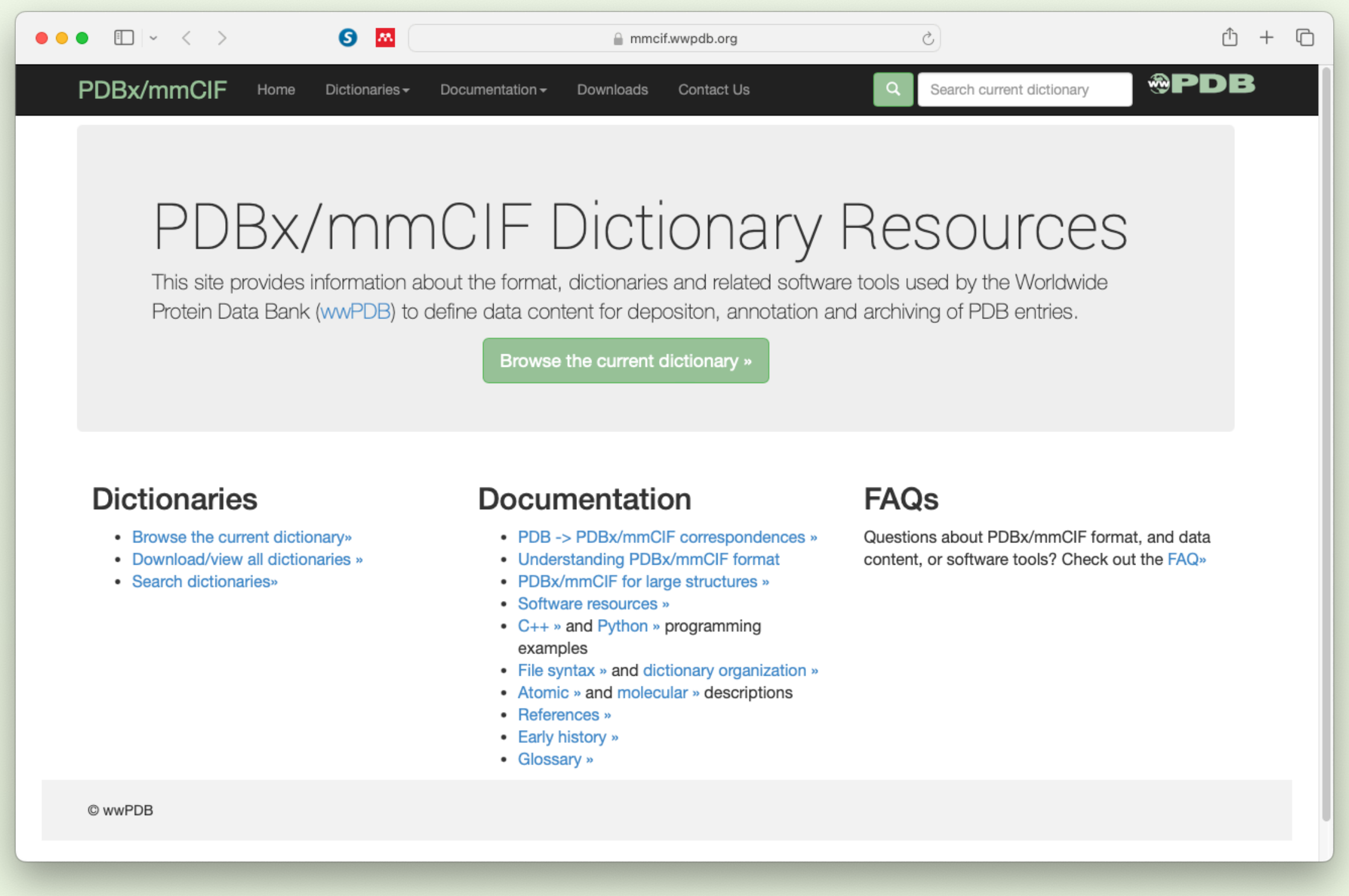
B

```
LINK        ZN    ZN C 201                NE2 HIS C  25     1555   1555  2.14
LINK        ZN    ZN C 201                SG  CYS C   7     1555   1555  2.23
LINK        ZN    ZN C 201                SG  CYS C  12     1555   1555  2.44
LINK        ZN    ZN C 201                NE2 HIS C  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        0.00000
SCALE2      0.000000  0.017794  0.000000        0.00000
SCALE3      0.000000  0.000000  0.007645        0.00000
```

(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/



mmcif.wwpdb.org

PDBx/mmCIF Home Dictionaries Documentation Downloads Contact Us Search current dictionary

PDBx/mmCIF Dictionary Resources

This site provides information about the format, dictionaries and related software tools used by the Worldwide Protein Data Bank (wwPDB) to define data content for depositon, annotation and archiving of PDB entries.

Browse the current dictionary »

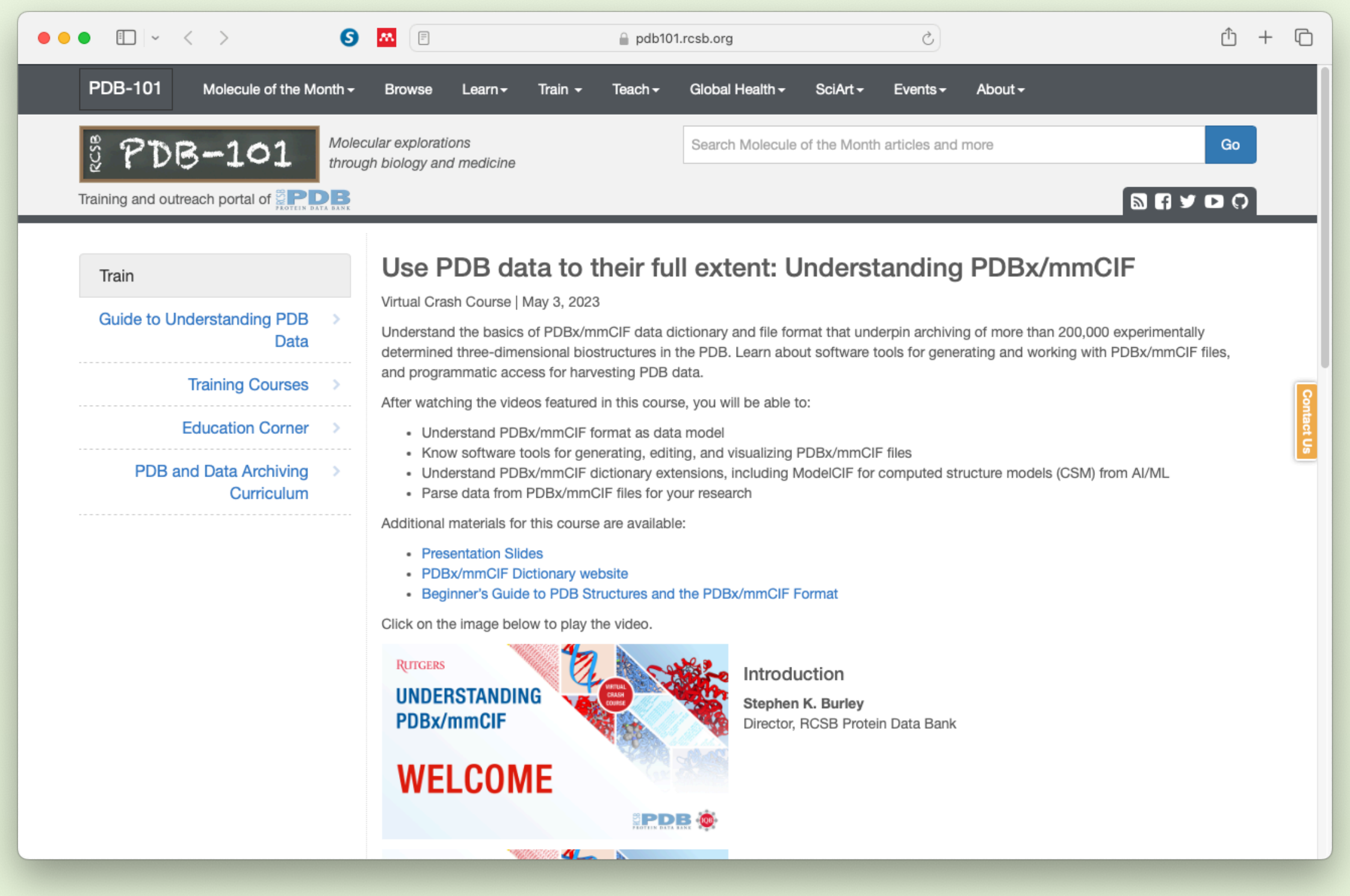
Dictionaries
- Browse the current dictionary»
- Download/view all dictionaries »
- Search dictionaries»

Documentation
- PDB -> PDBx/mmCIF correspondences »
- Understanding PDBx/mmCIF format
- PDBx/mmCIF for large structures »
- Software resources »
- C++ » and Python » programming examples
- File syntax » and dictionary organization »
- Atomic » and molecular » descriptions
- References »
- Early history »
- Glossary »

FAQs
Questions about PDBx/mmCIF format, and data content, or software tools? Check out the FAQ»

© wwPDB

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



pdb101.rcsb.org

PDB-101 Molecule of the Month Browse Learn Train Teach Global Health SciArt Events About

PDB-101 Molecular explorations through biology and medicine
Training and outreach portal of RCSB PDB
Search Molecule of the Month articles and more Go

Train
Guide to Understanding PDB Data
Training Courses
Education Corner
PDB and Data Archiving Curriculum

Use PDB data to their full extent: Understanding PDBx/mmCIF

Virtual Crash Course | May 3, 2023

Understand the basics of PDBx/mmCIF data dictionary and file format that underpin archiving of more than 200,000 experimentally determined three-dimensional biostructures in the PDB. Learn about software tools for generating and working with PDBx/mmCIF files, and programmatic access for harvesting PDB data.

After watching the videos featured in this course, you will be able to:
- Understand PDBx/mmCIF format as data model
- Know software tools for generating, editing, and visualizing PDBx/mmCIF files
- Understand PDBx/mmCIF dictionary extensions, including ModelCIF for computed structure models (CSM) from AI/ML
- Parse data from PDBx/mmCIF files for your research

Additional materials for this course are available:
- Presentation Slides
- PDBx/mmCIF Dictionary website
- Beginner's Guide to PDB Structures and the PDBx/mmCIF Format

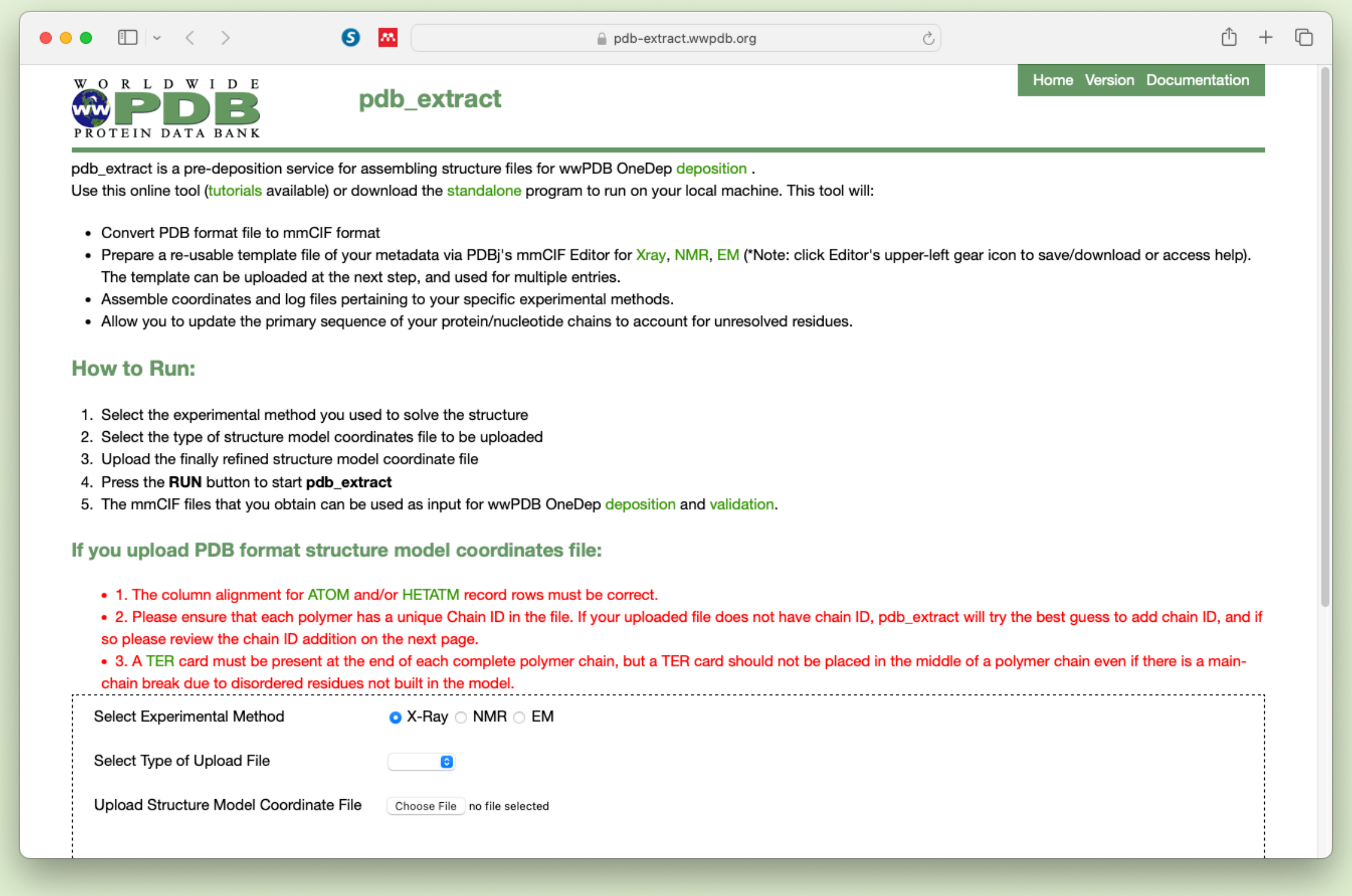
Click on the image below to play the video.

RUTGERS UNDERSTANDING PDBx/mmCIF WELCOME

Introduction
Stephen K. Burley
Director, RCSB Protein Data Bank

Contact Us

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



pdb-extract.wwpdb.org

WORLDWIDE PDB PROTEIN DATA BANK pdb_extract Home Version Documentation

pdb_extract is a pre-deposition service for assembling structure files for wwPDB OneDep deposition .
Use this online tool (tutorials available) or download the standalone program to run on your local machine. This tool will:

- Convert PDB format file to mmCIF format
- Prepare a re-usable template file of your metadata via PDBj's mmCIF Editor for Xray, NMR, EM (*Note: click Editor's upper-left gear icon to save/download or access help). The template can be uploaded at the next step, and used for multiple entries.
- Assemble coordinates and log files pertaining to your specific experimental methods.
- Allow you to update the primary sequence of your protein/nucleotide chains to account for unresolved residues.

How to Run:

1. Select the experimental method you used to solve the structure
2. Select the type of structure model coordinates file to be uploaded
3. Upload the finally refined structure model coordinate file
4. Press the **RUN** button to start **pdb_extract**
5. The mmCIF files that you obtain can be used as input for wwPDB OneDep deposition and validation.

If you upload PDB format structure model coordinates file:

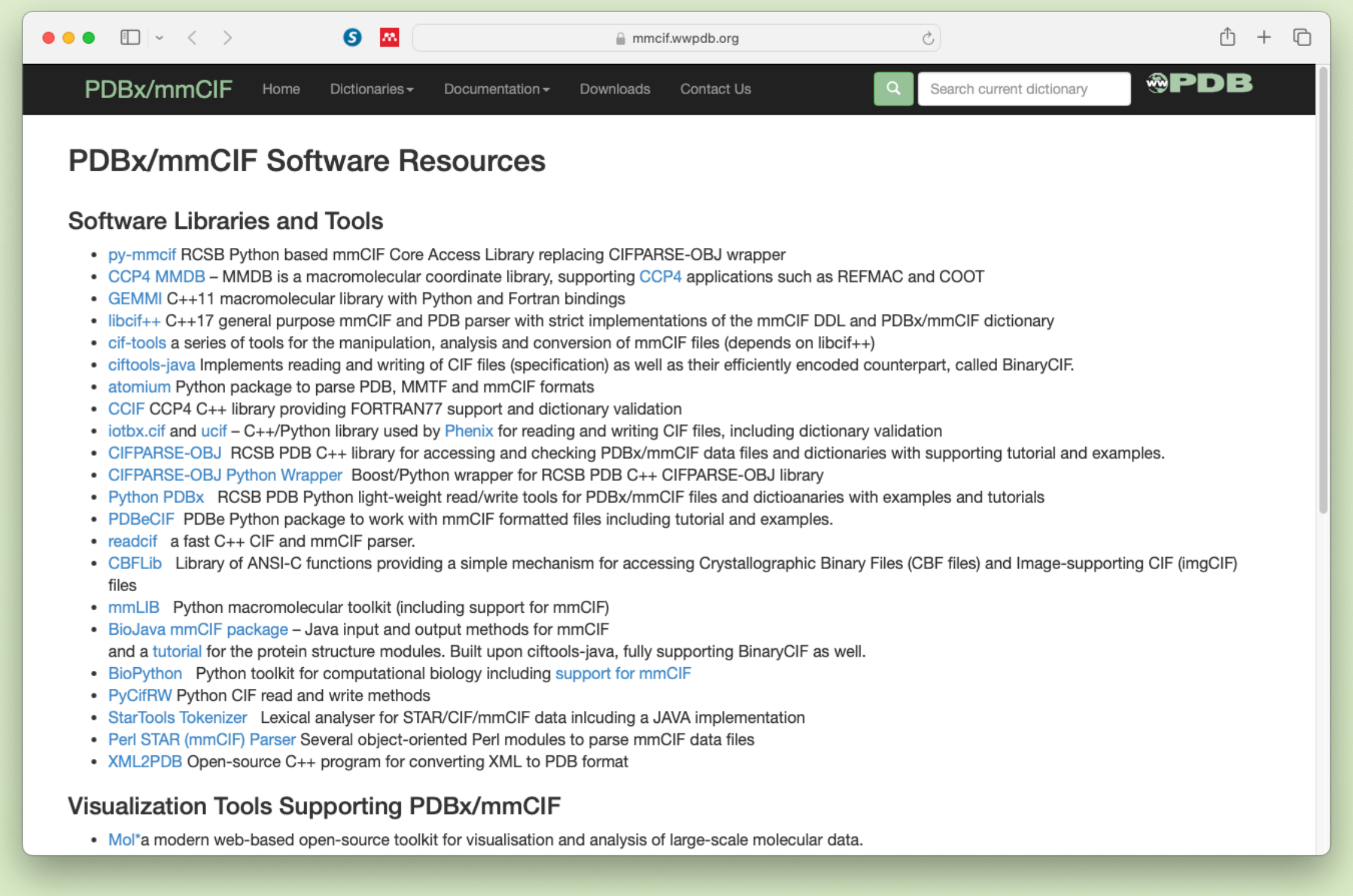
- 1. The column alignment for ATOM and/or HETATM record rows must be correct.
- 2. Please ensure that each polymer has a unique Chain ID in the file. If your uploaded file does not have chain ID, pdb_extract will try the best guess to add chain ID, and if so please review the chain ID addition on the next page.
- 3. A TER card must be present at the end of each complete polymer chain, but a TER card should not be placed in the middle of a polymer chain even if there is a main-chain break due to disordered residues not built in the model.

Select Experimental Method ● X-Ray ○ NMR ○ EM

Select Type of Upload File

Upload Structure Model Coordinate File Choose File no file selected

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



mmcif.wwpdb.org

PDBx/mmCIF Home Dictionaries Documentation Downloads Contact Us Search current dictionary

PDBx/mmCIF Software Resources

Software Libraries and Tools
- py-mmcif RCSB Python based mmCIF Core Access Library replacing CIFPARSE-OBJ wrapper
- CCP4 MMDB – MMDB is a macromolecular coordinate library, supporting CCP4 applications such as REFMAC and COOT
- GEMMI C++11 macromolecular library with Python and Fortran bindings
- libcif++ C++17 general purpose mmCIF and PDB parser with strict implementations of the mmCIF DDL and PDBx/mmCIF dictionary
- cif-tools a series of tools for the manipulation, analysis and conversion of mmCIF files (depends on libcif++)
- ciftools-java Implements reading and writing of CIF files (specification) as well as their efficiently encoded counterpart, called BinaryCIF.
- atomium Python package to parse PDB, MMTF and mmCIF formats
- CCIF CCP4 C++ library providing FORTRAN77 support and dictionary validation
- iotbx.cif and ucif – C++/Python library used by Phenix for reading and writing CIF files, including dictionary validation
- CIFPARSE-OBJ RCSB PDB C++ library for accessing and checking PDBx/mmCIF data files and dictionaries with supporting tutorial and examples.
- CIFPARSE-OBJ Python Wrapper Boost/Python wrapper for RCSB PDB C++ CIFPARSE-OBJ library
- Python PDBx RCSB PDB Python light-weight read/write tools for PDBx/mmCIF files and dictioanaries with examples and tutorials
- PDBeCIF PDBe Python package to work with mmCIF formatted files including tutorial and examples.
- readcif a fast C++ CIF and mmCIF parser.
- CBFLib Library of ANSI-C functions providing a simple mechanism for accessing Crystallographic Binary Files (CBF files) and Image-supporting CIF (imgCIF) files
- mmLIB Python macromolecular toolkit (including support for mmCIF)
- BioJava mmCIF package – Java input and output methods for mmCIF and a tutorial for the protein structure modules. Built upon ciftools-java, fully supporting BinaryCIF as well.
- BioPython Python toolkit for computational biology including support for mmCIF
- PyCifRW Python CIF read and write methods
- StarTools Tokenizer Lexical analyser for STAR/CIF/mmCIF data inlcuding a JAVA implementation
- Perl STAR (mmCIF) Parser Several object-oriented Perl modules to parse mmCIF data files
- XML2PDB Open-source C++ program for converting XML to PDB format

Visualization Tools Supporting PDBx/mmCIF
- Mol*a modern web-based open-source toolkit for visualisation and analysis of large-scale molecular data.

NMR-STAR data model

BMRB uses NMR-STAR[2] 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 interested to contribute to the data modeling efforts and become a NMR PDBx/mmCIF working group member, feel free to contact BMRB at help@bmrb.io

BMRB is supported by NIGMS through grant R24GM150793

[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 Biomol NMR 73, 5–9 (2019). doi: 10.1007/s10858-018-0220-3