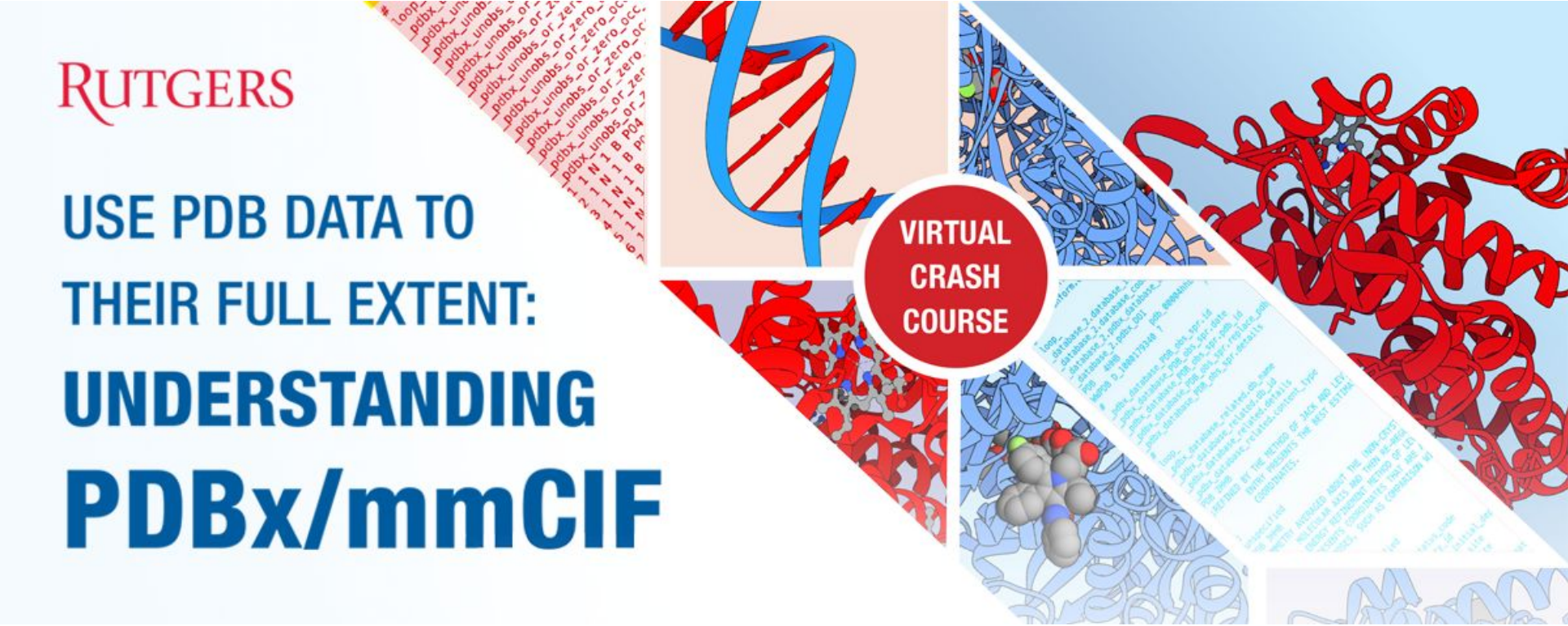
RUTGERS

USE PDB DATA TO THEIR FULL EXTENT: UNDERSTANDING PDBx/mmCIF



2023 IQB/RCSB PDB Spring Crash Course

RCSB PDB PROTEIN DATA BANK

IQB

Welcome and Introductions

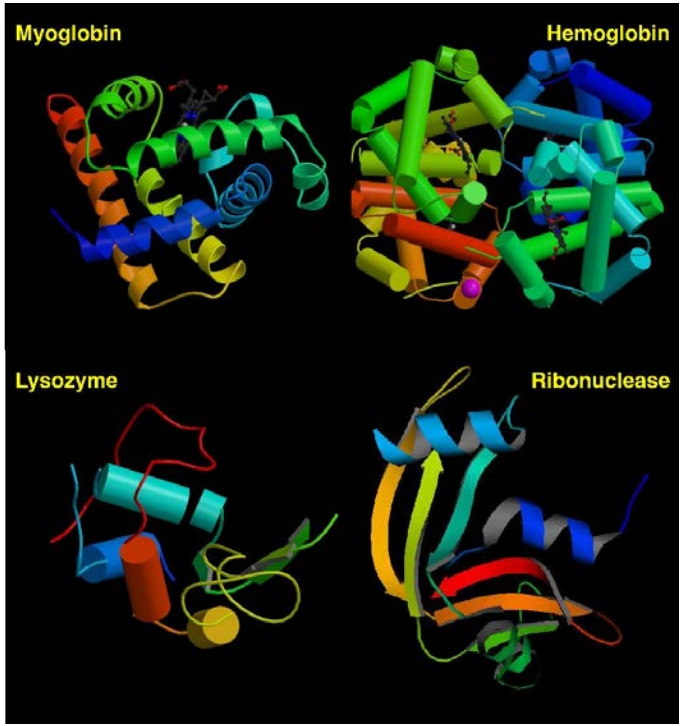
Stephen K. Burley, M.D., D.Phil

University Professor and Henry Rutgers Chair

Director, RCSB Protein Data Bank
Founding Director, Institute for Quantitative Biomedicine
Rutgers, The State University of New Jersey

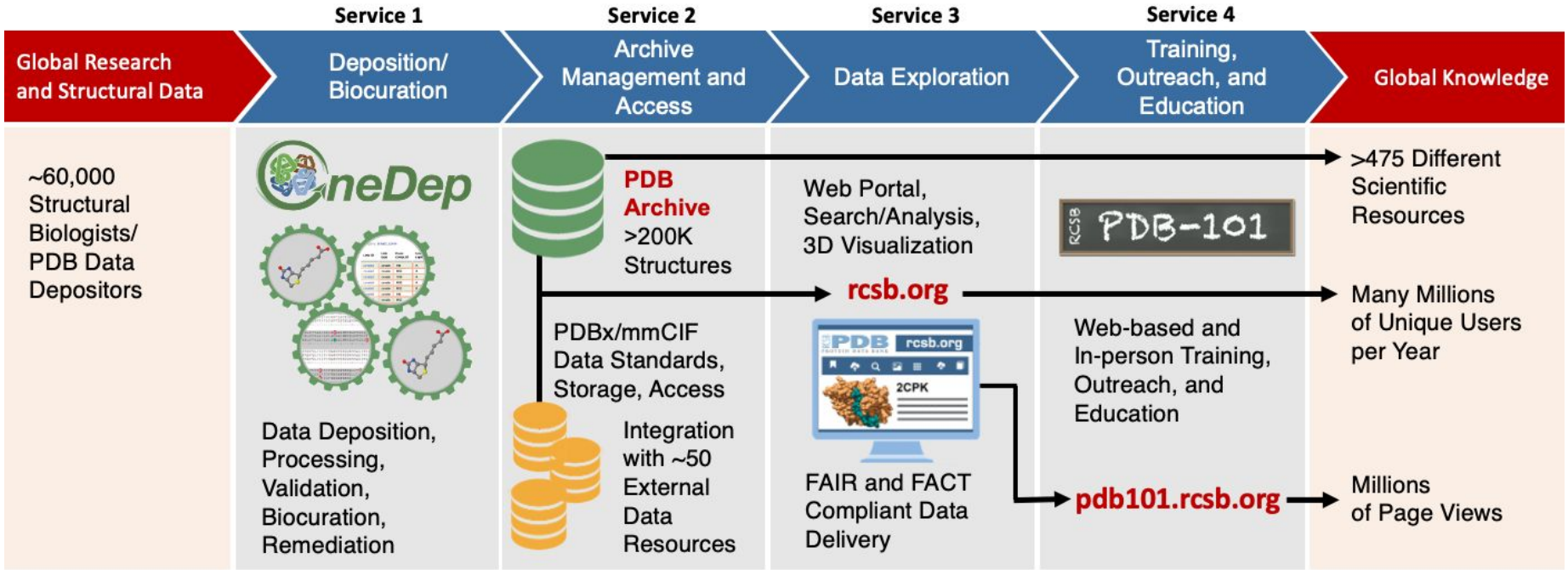
Protein Data Bank (Established 1971)

- PDB 1st Open Access digital data resource in all of biology
- Founded in 1971 with 7 protein structures
  - *Nature New Biology* 233, 223 (*1971*)
- Single global archive for protein and DNA/RNA experimental structures
- **Open Access to 204,104 structures!**
- wwPDB Collaboration: RCSB PDB (United States), PDBe (Europe), PDBj (Japan), and PDBc (China); and EMDB and BMRB
- Accredited by Core Trust Seal and Global Biodata Coalition

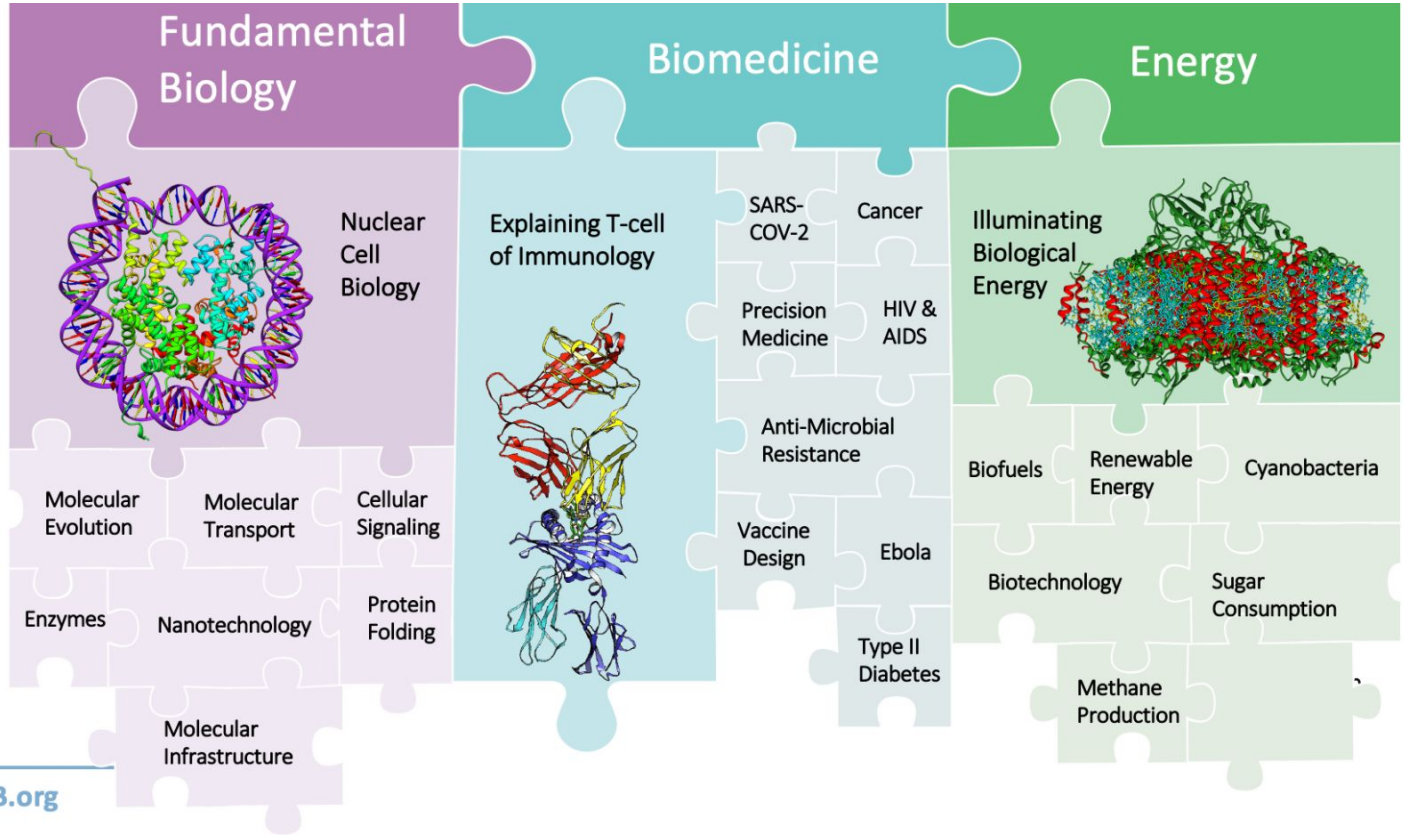


Structures that Inspired Launch of the PDB

RCSB PDB Converts Global Data into Knowledge



Impact: Research Funded by NSF, NIH, and DOE



RCSB.org

One-stop Shop for 3D Biostructure Data

RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning
(2023) *Nucleic Acids Research* 51: D488–D508 doi: 10.1093/nar/gkac1077

Introduction and Course Objectives

Gregg Crichlow, Ph.D.

RCSB PDB, Rutgers University

Today's Agenda (times in Eastern)

1:00–1:10 PM ***Welcome***
Stephen K. Burley, M.D., D.Phil.
Director, RCSB PDB and Founding Director,
Institute for Quantitative Biomedicine, Rutgers University

1:10–1:25 PM ***Introduction and course objectives***
Gregg Crichlow, Ph.D., RCSB PDB, Rutgers University

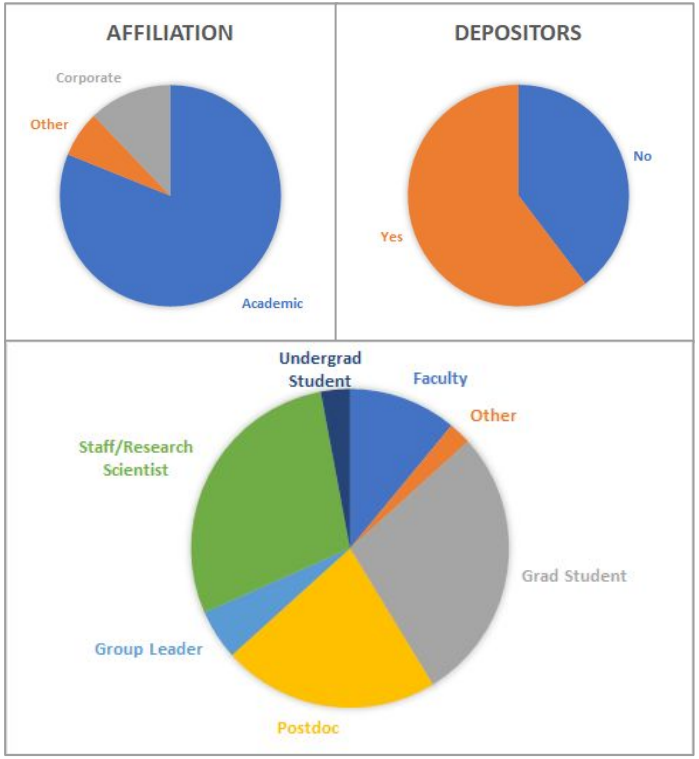
1:25–1:45 PM ***PDBx/mmCIF format - Not your parents' legacy PDB format***
Ezra Peisach, Ph.D., RCSB PDB, Rutgers University

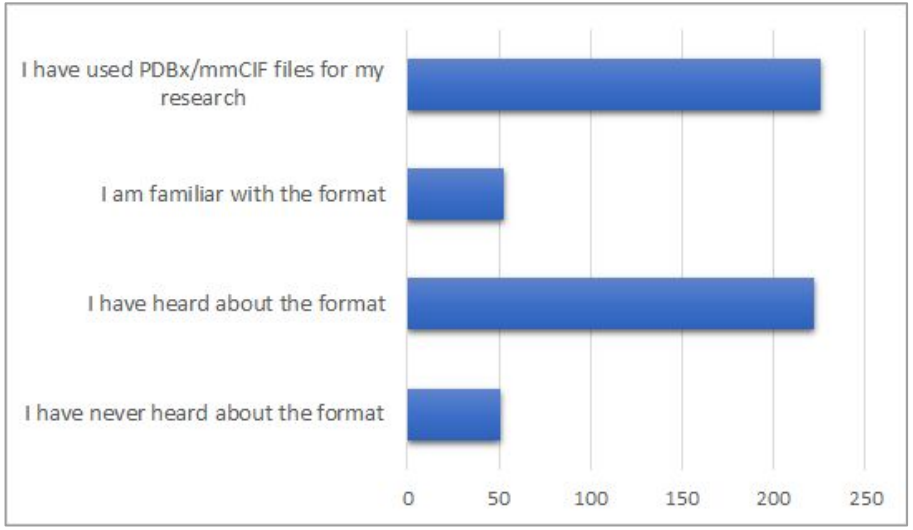
1:45–2:05 PM ***PDBx/mmCIF data files - Lifting the lid off the black box***
Brian Hudson, Ph.D., RCSB PDB, Rutgers University

2:05–2:50 PM ***Programmatic data access and analysis using PDBx/mmCIF files***
Irina Persikova, Ph.D. and Chenghua Shao, Ph.D.,
RCSB PDB, Rutgers University

2:50–3:00 PM ***Closing remarks and acknowledgements***
Stephen K. Burley, M.D., D.Phil.

Our Audience Today





PDBx/mmCIF - Community Origins

- The CIF (crystallographic information framework) format, used in small molecule crystallography, formed the basis of an expanded dictionary to make it suitable to describe macromolecules, resulting in PDBx/mmCIF (Protein Data Bank Exchange/macromolecular CIF)
- Engagement with the macromolecular crystallography (MX) community led to the formation of the mmCIF working group

mmCIF Working Group, 1993 and 2011

Westbrook, J.D. (1995) Thesis (Ph. D) Rutgers University

WORLDWIDE PDB PROTEIN DATA BANK

PDBx/mmCIF Data Standard/Format

- Since late 1990s, PDB archive has provided entries in PDBx/mmCIF format https://mmcif.wwpdb.org
- Since 2019, mandatory format for coordinate deposition for MX
- Crystallographic refinement software produce coordinate files in PDBx/mmCIF format
- Mandatory deposition for 3DEM and NMR coming soon
- NB: Today, we will use the terms CIF, mmCIF and PDBx/mmCIF interchangeably

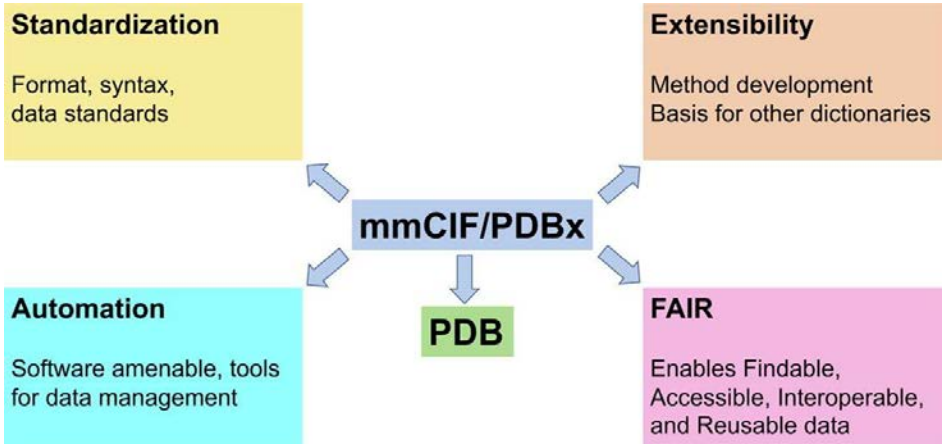
```
loop_
_database_2.database_id
_database_2.database_code
_database_2.pdbx_database_accession
_database_2.pdbx_DOI
PDB   8C5M         pdb_00008c5m 10.2210/pdb8c5m/pdb
WWPDB D_1292127815 ?            ?
#
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_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.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      . SER A  1 1   ? 83.18600  -46.80000 3.49300   1.000 68.99538 ? 6799 SER A N      1
ATOM   2    C  CA     . SER A  1 1   ? 82.67100  -47.53400 2.34100   1.000 61.66335 ? 6799 SER A CA     1
```

http://www.wwpdb.org/documentation/file-formats-and-the-pdb

Kremling, V., Sprenger, J., Oberthuer, D., Falke, S. (2023)
SARS-CoV-2 nsp10-16 methyltransferase in complex with MTA
doi: https://doi.org/10.2210/pdb8C5M/pdb

PDBx/mmCIF Data Standard/Format

- PDBx/mmCIF is the master archival format of the PDB since 2014
- No columns, no character length limitations
- Able to accommodate ribosomes and other large structures in a single file
- Allows successful realization of FAIR principles



Westbrook, *et al.*, (2022) *JMB* 434: 167599 doi: 10.1016/j.jmb.2022.167599

Attributes of PDBx/mmCIF

- Fully extensible, allowing for the accommodation of metadata from new increasingly complex experimental methods (SX, XFEL, 3DEM)
- Ensures new metadata items conform to data standards
- Founded upon the DDL2 dictionary, which is self-validating, allowing checking of PDBx/mmCIF files for self-consistency
- Human and machine readable

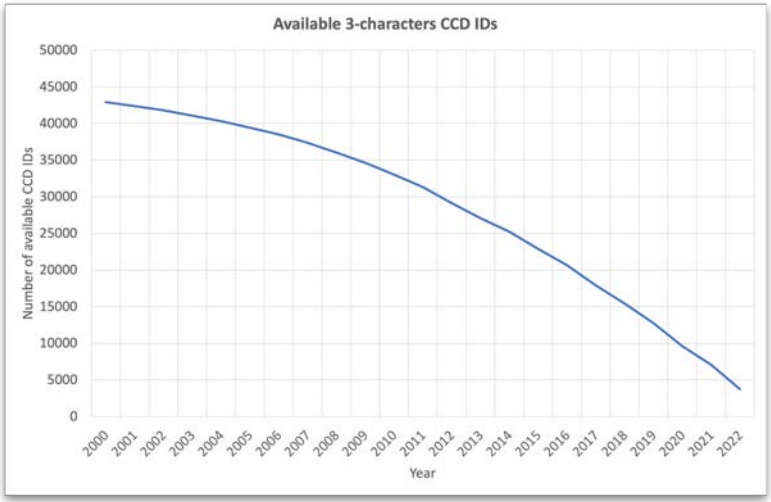
```
_pdbx_serial_crystallography_data_reduction.diffrn_id              1
_pdbx_serial_crystallography_data_reduction.frames_total           ?
_pdbx_serial_crystallography_data_reduction.xfel_pulse_events      ?
_pdbx_serial_crystallography_data_reduction.frame_hits             ?
_pdbx_serial_crystallography_data_reduction.crystal_hits           260392
_pdbx_serial_crystallography_data_reduction.droplet_hits           ?
_pdbx_serial_crystallography_data_reduction.frames_failed_index    ?
_pdbx_serial_crystallography_data_reduction.frames_indexed         160275
_pdbx_serial_crystallography_data_reduction.lattices_indexed       ?
_pdbx_serial_crystallography_data_reduction.xfel_run_numbers       ?
#
_pdbx_serial_crystallography_sample_delivery.diffrn_id     1
_pdbx_serial_crystallography_sample_delivery.description   ?
_pdbx_serial_crystallography_sample_delivery.method        injection
#
_pdbx_serial_crystallography_sample_delivery_injection.diffrn_id                1
_pdbx_serial_crystallography_sample_delivery_injection.description              ?
_pdbx_serial_crystallography_sample_delivery_injection.injector_diameter        75
_pdbx_serial_crystallography_sample_delivery_injection.injector_temperature     293
_pdbx_serial_crystallography_sample_delivery_injection.injector_pressure        ?
_pdbx_serial_crystallography_sample_delivery_injection.flow_rate                0.047
_pdbx_serial_crystallography_sample_delivery_injection.carrier_solvent          ?
_pdbx_serial_crystallography_sample_delivery_injection.crystal_concentration    ?
_pdbx_serial_crystallography_sample_delivery_injection.preparation              ?
_pdbx_serial_crystallography_sample_delivery_injection.power_by                 ?
_pdbx_serial_crystallography_sample_delivery_injection.injector_nozzle          ?
_pdbx_serial_crystallography_sample_delivery_injection.jet_diameter             ?
_pdbx_serial_crystallography_sample_delivery_injection.filter_size              ?
```

Westbrook, *et al.*, (2022) *JMB* 434: 167599 doi: 10.1016/j.jmb.2022.167599

https://doi.org/10.2210/pdb7Q7Q/pdb
Bath, *et al.*, (2022) Lipidic cubic phase serial femtosecond crystallography structure of a photosynthetic reaction centre. Acta Crystallogr D Struct Biol 78: 698-708 doi: https://doi.org/10.1107/S2059798322004144

Upcoming Exclusive Use of PDBx/mmCIF

- The available three-character chemical component IDs (CCIDs, small molecule ligand codes) will be exhausted in within approximately one year.
- ***This will necessitate that coordinate files for all PDB entries deposited after this point that contain novel ligands will be released exclusively in PDBx/mmCIF format.***
- PDB ID is limited to four characters, so wwPDB will soon need to expand PDB IDs to an eight-character format with prefix pdb_00001abc.
- ***Once this occurs, all newly deposited PDB entries will only be available in PDBx/mmCIF format.***



Summary/Course Objectives

- PDBx/mmCIF is the PDB archival data standard and file format
- Flexible/extensible format allows for accommodation of large structures and of new, exciting structural methodologies
- Human readable but also is produced and read by software
- Information may be parsed programmatically, but also visually and via web links
- Soon to be the exclusive distribution format for PDB coordinate data

PDBx/mmCIF Format

Ezra Peisach, Ph.D.

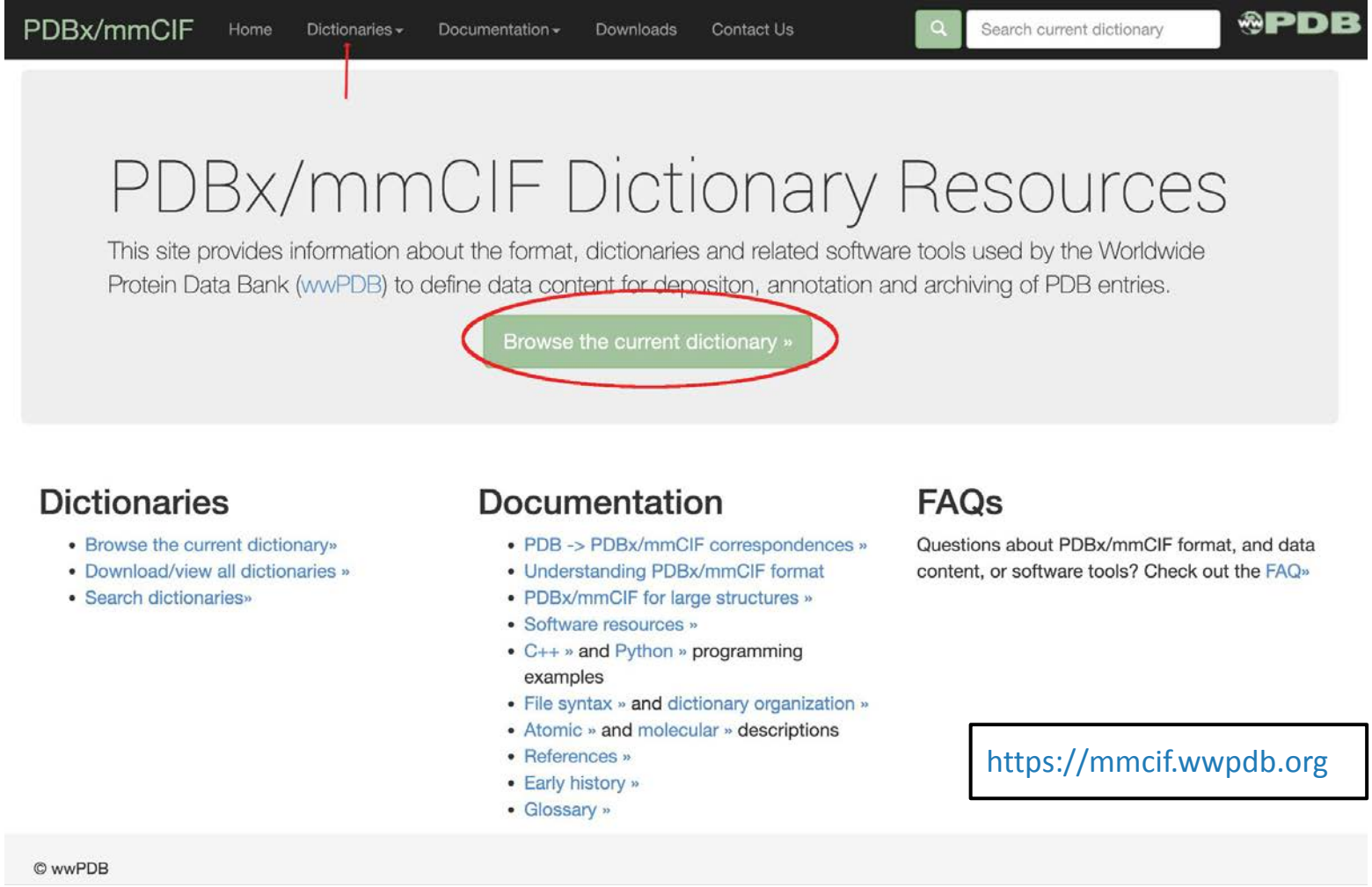
RCSB PDB, Rutgers University

Outline

- The PDBx/mmCIF Data Format & Dictionary
- Syntax
- Understanding the PDBx/mmCIF Data Model Organization

PDBx/mmCIF Data Standard/Format

- Flexible
- Human readable!
- No limits on width/precision of data fields
- Data controlled by a dictionary
  - Values must adhere to types (regular expressions), enumerations, range limits
  - Relationships between categories enforced by parent/child relationships
- Data organized as categories “tables” of keys and values
- Can be rapidly extended as scientific methodologies evolve
- Dictionary publicly available at https://mmcif.wwpdb.org
- Tools are available to validate data against the dictionary
- Basis of other dictionaries (ModelCIF, IhmCIF)


PDBx/mmCIF
Home
Dictionaries
Documentation
Downloads
Contact Us
Search current dictionary
PDB
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»
https://mmcif.wwpdb.org
© wwPDB

Management of PDBx/mmCIF Dictionary

- wwPDB
- PDBx/mmCIF Working Group
- Developers of extension dictionaries
- Feedback from users

https://github.com/wwpdb-dictionaries/mmcif_pdbx

Dictionary Range Limit Example

- pH - has physical limits
- Has advisory limits based on what is common in PDB
  - Lysozyme crystals grown at pH 3.0 exists. Allowed - but unusual.

Allowed Boundary Conditions

Minimum Value	Maximum Value
0.0	14.0

Advisory Boundary Conditions

Minimum Value	Maximum Value
3.5	10

Enumerations

_diffrn_source.type

APS BEAMLINE 31-ID
APS BEAMLINE 32-ID
APS BEAMLINE 34-ID
APS BEAMLINE 5ID-B
APS BEAMLINE 8-BM
AUSTRALIAN SYNCHROTRON BEAMLINE MX1
AUSTRALIAN SYNCHROTRON BEAMLINE MX2
Agilent SuperNova
AichiSR BEAMLINE BL2S1
BESSY BEAMLINE 14.1
BESSY BEAMLINE 14.2
BESSY BEAMLINE 14.3
BRUKER AXS MICROSTAR
BRUKER AXS MICROSTAR-H
BRUKER D8 QUEST

Syntax Overview

- Data presented in item/value pairs
- An item is made up of a *_category.attribute*
  - *_citation.id*
- Data can be represented as tables
- CIF rules for quoting text (https://www.iucr.org/resources/cif/spec/version1.1)

- The dictionary regulates the content of the archive!

Examples

```
_cell.entry_id              4HHB
_cell.length_a              63.150
_cell.length_b              83.590
_cell.length_c              53.800
_cell.angle_alpha           90.00
_cell.angle_beta            99.34
_cell.angle_gamma           90.00
_cell.Z_PDB                 4
```

```
loop_

_audit_author.name
_audit_author.pdbx_ordinal
'Fermi, G.'    1
'Perutz, M.F.' 2
```

Fermi et al., The crystal structure of human deoxyhaemoglobin at 1.74 angstroms resolution (1984) doi: 10.2210/pdb4HHB/pdb

Syntax: Four Equivalent Examples

```
_animal.id        1
_animal.owner     Mary
_animal.sizelittle
_animal.typelamb
_animal.details   'Its fleece was white as snow'
_animal.name?
```

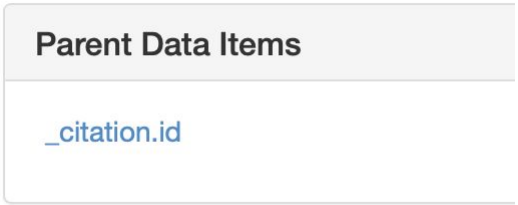
```
_animal.id        1
_animal.owner     Mary
_animal.sizelittle
_animal.typelamb
_animal.details
;Its fleece was white as snow
;
_animal.name?
```

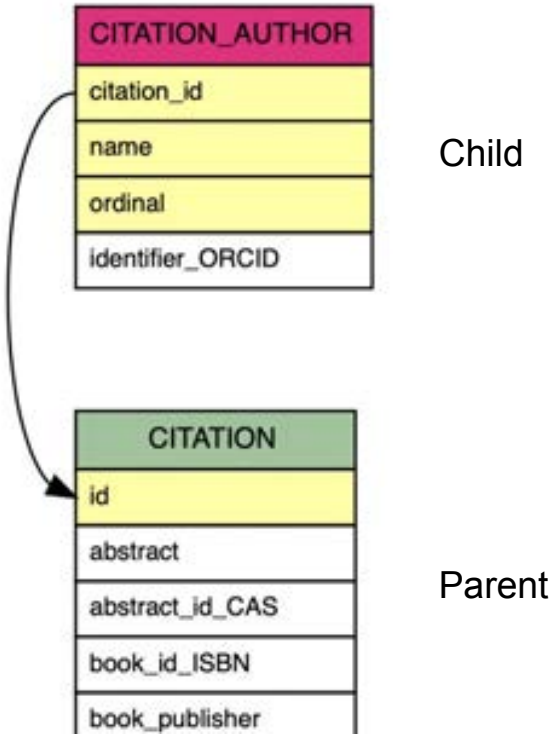
```
loop_
_animal.id
_animal.owner
_animal.size
_animal.type
_animal.details
_animal.name
1 Mary little lamb 'Its fleece was white as snow' ?
```

```
loop_
_animal.id
_animal.owner
_animal.size
_animal.type
_animal.details
_animal.name
1 Mary little lamb
;Its fleece was white as snow
;
?
```

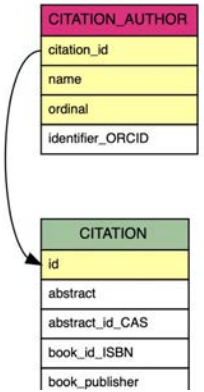
Basic Parent-child Relationships

- Data categories may be linked together
- Hierarchy relationship between categories
- Allows one to many relation
- In database terms a “foreign” key
- Defined in dictionary





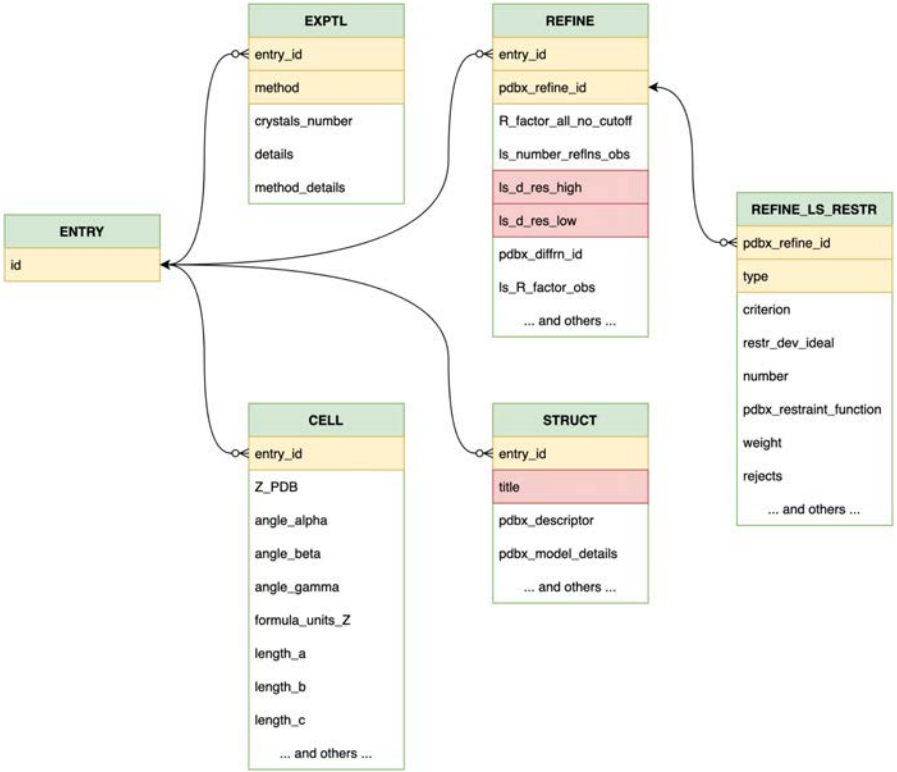
```
loop_
_citation.id
_citation.title
_citation.journal_abbrev
_citation.journal_volume
_citation.page_first
_citation.page_last
_citation.year
_citation.journal_id_ASTM
_citation.country
_citation.journal_id_ISSN
_citation.journal_id_CSD
_citation.book_publisher
_citation.pdbx_database_id_PubMed
_citation.pdbx_database_id_DOI
primary 'Constraints for Zinc Finger Linker Design as Inferred from X-ray Crystal Structure of Tandem Zif268-DNA Complexes'
J.Mol.Biol.             330 1    7    2003 JMOBAK UK 0022-2836 0070 ? 12818197 '10.1016/S0022-2836(03)00572-2'
1       'Zif268 protein-DNA complex refined at 1.6A: implications for understanding zinc finger DNA recognition'
Structure               6   451  464  1996 STRUE6 UK 0969-2126 2005 ? ?        '10.1016/S0969-2126(98)00047-1'
2       'Getting a handhold on DNA: design of poly-zinc finger proteins with femtomolar dissociation constants.'
Proc.Natl.Acad.Sci.USA 95  2812 2817 1998 PNASA6 US 0027-8424 0040 ? ?        10.1073/pnas.95.6.2812
#
loop_
_citation_author.citation_id
_citation_author.name
_citation_author.ordinal
primary 'Peisach, E.'        1
primary 'Pabo, C.O.'         2
1       'Elrod-Erickson, M.' 3
1       'Rould, M.A.'        4
1       'Nekludova, L.'      5
1       'Pabo, C.O.'         6
2       'Kim, J.S.'          7
2       'Pabo, C.O.'         8
#
```



Peisach et al. Zinc Finger-DNA recognition: Crystal Structure of tandem Zif268 molecules complexed to DNA (2003) doi: 10.2210/pdb1P47/pdb

Extensive Parent-child Relationships Exist

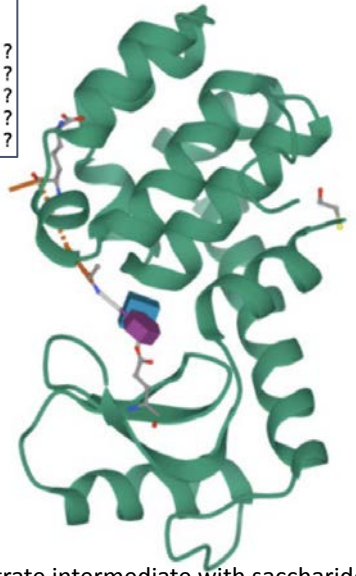
- Links together parts of an entry
- Hierarchy



Entity Ties Sections Together

```
loop_
_entity.id
_entity.type
_entity.src_method
_entity.pdbx_description
_entity.formula_weight
_entity.pdbx_number_of_molecules
_entity.pdbx_ec
_entity.pdbx_mutation
_entity.pdbx_fragment
_entity.details
1 polymer     man 'T4 LYSOZYME'                                                    18656.373 1   3.2.1.17 ? ? ?
2 polymer     nat 'SUBSTRATE CLEAVED FROM CELL WALL OF ESCHERICHIA COLI'           461.466   1   ?        ? ? ?
3 branched    man '2-acetamido-2-deoxy-beta-D-glucopyranose-(1-4)-N-acetyl-alpha-muramic acid' 496.463   1   ?        ? ? ?
4 non-polymer syn BETA-MERCAPTOETHANOL                                             78.133    1   ?        ? ? ?
5 water       nat water                                                            18.015    140 ?        ? ? ?
```

```
loop_
_entity_poly.entity_id
_entity_poly.type
_entity_poly.nstd_linkage
_entity_poly.nstd_monomer
_entity_poly.pdbx_seq_one_letter_code
_entity_poly.pdbx_seq_one_letter_code_can
_entity_poly.pdbx_strand_id
_entity_poly.pdbx_target_identifier
1 'polypeptide(L)' no no
;MNIFEMLRIDEGLRLKIYKDTEGYYEIGIGHLLTKSPSLNAAKSELDKAIGRNTNGVITKDEAEKLFNQDVDAAVRGILR
NAKLKPVYDSLDAVRRAALINMVFQMGETGVAGFTNSLRMLQQKRWDEAAVNLAKSRWYNQTPNRAKRVITTFRTGTWDA
YKNL
;
;MNIFEMLRIDEGLRLKIYKDTEGYYEIGIGHLLTKSPSLNAAKSELDKAIGRNTNGVITKDEAEKLFNQDVDAAVRGILR
NAKLKPVYDSLDAVRRAALINMVFQMGETGVAGFTNSLRMLQQKRWDEAAVNLAKSRWYNQTPNRAKRVITTFRTGTWDA
YKNL
;
E ?
2 'polypeptide(L)' no yes 'A(FGA)(API)(DAL)' AEKA S ?
```



Kuorko et al. A covalent enzyme-substrate intermediate with saccharide distortion in a mutant T4 Lysozyme (1994) doi: 10.2210/pdb148L/pdb

How to Read PDBx/mmCIF: Sequence

```
_entity_poly.entity_id                          1
_entity_poly.type                               'polypeptide(L)'
_entity_poly.nstd_linkage                       no
_entity_poly.nstd_monomer                       yes
_entity_poly.pdbx_seq_one_letter_code
;MEKKEFHIVAETGIHARPATLLVQTASKFNSDINLEYKGKSVNLK(SEP)IMGVMSLGVGQGSDVTITVDGADEAEGMAA
IVETLQKEGLA
;
_entity_poly.pdbx_seq_one_letter_code_can
;MEKKEFHIVAETGIHARPATLLVQTASKFNSDINLEYKGKSVNLKSIMGVMSLGVGQGSDVTITVDGADEAEGMAAIVET
LQKEGLA
;
_entity_poly.pdbx_strand_id                     A,B
_entity_poly.pdbx_target_identifier             ?
```

Audette et al. Crystal structure analysis of the phospho-serine 46 HPR from *Enterococcus faecalis* (2000) doi: 10.2210/pdb1FU0/pdb

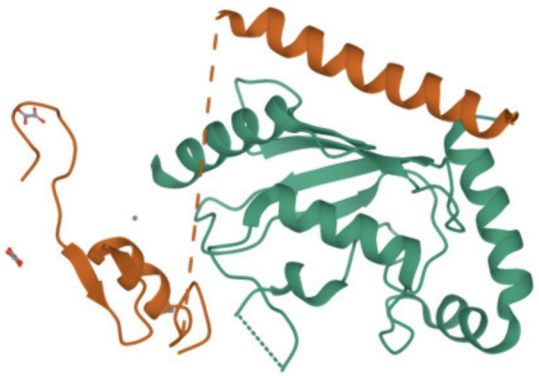
Scheme - Polymer

Maps ordinal numbering to author numbering

Sequential order

Author residue number

Unobserved

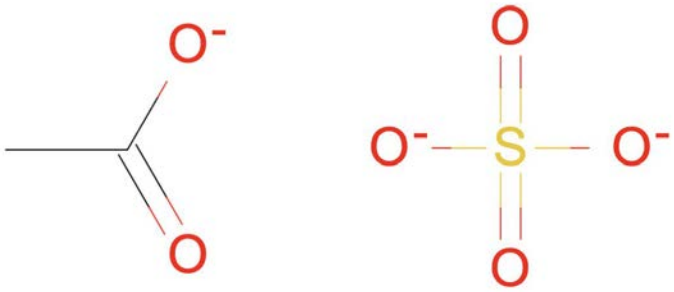


Liang et al. Allosteric regulation of E2:E3 interactions promote a processive ubiquitination machine (2013) doi: 10.2210/pdb4LAD/pdb

```
loop_
_pdbx_poly_seq_scheme.asym_id
_pdbx_poly_seq_scheme.entity_id
_pdbx_poly_seq_scheme.seq_id
_pdbx_poly_seq_scheme.mon_id
_pdbx_poly_seq_scheme.ndb_seq_num
_pdbx_poly_seq_scheme.pdb_seq_num
_pdbx_poly_seq_scheme.auth_seq_num
_pdbx_poly_seq_scheme.pdb_mon_id
_pdbx_poly_seq_scheme.auth_mon_id
_pdbx_poly_seq_scheme.pdb_strand_id
_pdbx_poly_seq_scheme.pdb_ins_code
_pdbx_poly_seq_scheme.hetero
...
B 2 1   HIS 1   311 ?   ?   ?   B . n
B 2 2   MET 2   312 ?   ?   ?   B . n
B 2 3   LYS 3   313 ?   ?   ?   B . n
...
B 2 29  ASP 29  339 339 ASP ASP B . n
B 2 30  ASP 30  340 340 ASP ASP B . n
B 2 31  CYS 31  341 341 CYS CYS B . n
B 2 32  ALA 32  342 342 ALA ALA B . n
...
B 2 70  MET 70  380 380 MET MET B . n
B 2 71  SER 71  521 ?   ?   ?   B . n
B 2 72  LEU 72  522 ?   ?   ?   B . n
B 2 73  ASN 73  523 ?   ?   ?   B . n
B 2 74  ILE 74  524 ?   ?   ?   B . n
...
B 2 121 GLY 121 571 ?   ?   ?   B . n
B 2 122 GLY 122 572 ?   ?   ?   B . n
B 2 123 GLY 123 573 573 GLY GLY B . n
B 2 124 SER 124 574 574 SER SER B . n
...
```

Scheme - Non-polymer

Provides mapping from author's original numbering and identifier to numbering in final atomic coordinate file




```
loop_
_pdbx_nonpoly_scheme.asym_id
_pdbx_nonpoly_scheme.entity_id
_pdbx_nonpoly_scheme.mon_id
_pdbx_nonpoly_scheme.ndb_seq_num
_pdbx_nonpoly_scheme.pdb_seq_num
_pdbx_nonpoly_scheme.auth_seq_num
_pdbx_nonpoly_scheme.pdb_mon_id
_pdbx_nonpoly_scheme.auth_mon_id
_pdbx_nonpoly_scheme.pdb_strand_id
_pdbx_nonpoly_scheme.pdb_ins_code
B 2 TB   1   154 1   TB  TB  A .
C 3 ACT  1   155 2   ACT ACT A .
D 3 ACT  1   156 3   ACT ACT A .
E 4 SO4  1   157 4   SO4 SO4 A .
F 5 HOH  1   158 158 HOH HOH A .
F 5 HOH  2   159 159 HOH HOH A .
F 5 HOH  3   160 1   HOH HOH A .
F 5 HOH  4   161 2   HOH HOH A .
F 5 HOH  5   162 3   HOH HOH A .
```

Coordinates

Asym id

Residue ordinal

Author residue number

Author chain id

Peisach et al. Structure of Interleukin 1B solved by SAD using an inserted Lanthanide Binding Tag (2011) doi: 10.2210/pdb3LTQ/pdb

```
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.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    . PRO A 1 3   ? 3.410  49.306 56.552 1.00 59.36 ? 2   PRO A N    1
ATOM   2    C  CA   . PRO A 1 3   ? 4.560  48.828 55.773 1.00 58.74 ? 2   PRO A CA   1
ATOM   3    C  C    . PRO A 1 3   ? 4.324  48.946 54.260 1.00 54.92 ? 2   PRO A C    1
ATOM   4    O  O    . PRO A 1 3   ? 3.241  49.354 53.842 1.00 52.52 ? 2   PRO A O    1
ATOM   5    C  CB   . PRO A 1 3   ? 4.684  47.360 56.193 1.00 54.77 ? 2   PRO A CB   1
ATOM   6    C  CG   . PRO A 1 3   ? 3.293  46.962 56.563 1.00 58.10 ? 2   PRO A CG   1
ATOM   7    C  CD   . PRO A 1 3   ? 2.651  48.198 57.162 1.00 59.84 ? 2   PRO A CD   1
ATOM   8    N  N    . VAL A 1 4   ? 5.326  48.580 53.461 1.00 52.99 ? 3   VAL A N    1
ATOM   9    C  CA   . VAL A 1 4   ? 5.320  48.858 52.021 1.00 50.84 ? 3   VAL A CA   1
ATOM   10   C  C    . VAL A 1 4   ? 4.133  48.248 51.257 1.00 45.68 ? 3   VAL A C    1
ATOM   11   O  O    . VAL A 1 4   ? 3.880  47.044 51.314 1.00 44.98 ? 3   VAL A O    1
ATOM   12   C  CB   . VAL A 1 4   ? 6.668  48.461 51.362 1.00 51.83 ? 3   VAL A CB   1
ATOM   13   C  CG1  . VAL A 1 4   ? 6.879  46.959 51.415 1.00 50.21 ? 3   VAL A CG1  1
```

Author numbering

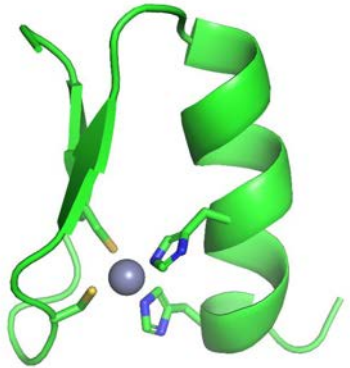
auth_xxx items map to PDB formatted file

Protein-Ligand Interaction

- Intermolecular links recorded
- Covalent and metallo

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
. . .
```



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
. . .
```

Pavletich et al. Zinc Finger-DNA recognition: Crystal structure of a ZIF268-DNA Complex at 2.1 Angstroms (1993) doi: 10.2210/pdb1ZAA/pdb

Summary

- Dictionary and tools ensure data well regulated archive
- PDBx/mmCIF extensible as experimental methods evolve
- Basic syntax covered
- High level PDBx/mmCIF model described

Any questions?

The PDBx/mmCIF Formatted File:

Lifting the Lid Off the Black Box

Brian Hudson, Ph.D.
RCSB PDB,
Rutgers University

The PDBx/mmCIF Formatted File:

Lifting the Lid Off the Black Box

Brian Hudson, Ph.D.
RCSB PDB,
Rutgers University

“Tools and tips to simplify preparing PDBx/mmCIF files for deposition, visualization, and more”

“How will you use PDBx/mmCIF formatted files?”

“Black box, please.”

“I’d like to be able to edit them safely.”

“I want as much control as possible.”

Visualizing PDBx/mmCIF Formatted *Atomic Coordinate* Files

- Mol*
- PyMol
- UCSF Chimera
- UCSF ChimeraX
- RasMol
- Jmol
- VMD
- CCP4mg
- …

Creating PDBx/mmCIF Formatted Files

- Refinement software packages
  - PHENIX
    - mmCIF output by default
  - Buster
    - mmCIF output by default
  - REFMAC
    - mmCIF output is an available option
    - instructions are available for adding mmCIF output to existing refinements
- Visualization programs
  - Mol*
  - PyMol
  - UCSF ChimeraX
  - CCP4mg
- **pdb_extract**
- **PDBj CIF Editor**

pdb_extract Integrates Data and Metadata

- pdb_extract is a data harvesting tool that parses a coordinate file, data processing log files, and the author's metadata (template) file, and merges the data into a more complete pdbx/mmCIF output file for PDB deposition
- https://pdb-extract.wwpdb.org
- pdb_extract extracts the following categories from
  - PDB formatted files:
    - *atom_site* (XYZ atomic coordinates)
    - *cell* (cell parameters)
    - *symmetry* (space group)
    - *refine* (refinement statistics)
  - PDBx/mmCIF formatted files:
    - **EVERYTHING**
- pdb_extract combines content…
  - …from multiple sources into a PDBx/mmCIF file
  - …on a category-by-category basis
    - will not mix content from multiple sources within a single category
- pdb_extract prioritizes content….

pdb_extract Data/Metadata Integration Priority

Highest → **Lowest**

Atomic Coordinate File (required)
- PDB format
  -or-
- PDBx/mmCIF format

Reflection Data Files (X-ray, optional)
- Final refinement
- Other reflection data files

Polymer Sequences (optional)
- Entered in pdb_extract interface

Processing Log Files (X-ray, optional)
- Data collection / index
- Reflection data reduction
- Molecular replacement
- Final structure refinement

PDBx/mmCIF Formatted "Input Template File" (optional)
- Contact author information
- Entry title and author list
- Citation information
- Keywords
- Macromolecule information
- Sample / crystal information
- Data collection details
- *et cetera*

PDBx/mmCIF Formatted Atomic Coordinate File (+ PDBx/mmCIF Formatted Structure Factor File)

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 chosen

[Run] [Reset]

Note: If the file size is too large (e.g. >100 MB), you can upload gzipped (***.gz**) or compressed (*.Z) file for faster loading.
Reference: Huanwang Yang, Vladimir Guranovic, Shuchismita Dutta, Zukang Feng, Helen M. Berman and John D. Westbrook (2004), Acta Cryst. D60, 1833-18399

Questions, comments, and suggestions should be sent to deposit@deposit.rcsb.org.

pdb_extract Page #2, Part I (All Methods)

WORLDWIDE PDB PROTEIN DATA BANK

pdb_extract

Home Version Documentation

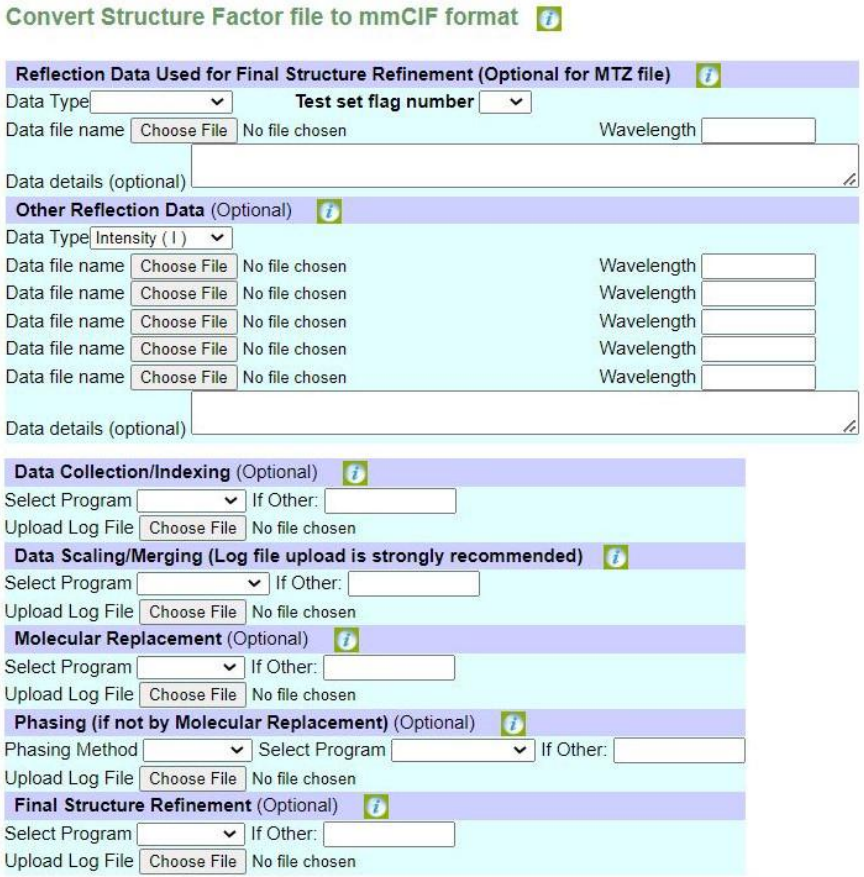
EXTRACTING INFORMATION FOR PDB DEPOSITION help in each step

- MTZ Structure Factor file does not need to be converted to mmCIF for wwPDB deposition. You may skip the conversion, and upload MTZ file directly at wwPDB OneDep deposition system.
- All other information asked on this page is also optional. However, the information extracted from your template and log files will save time during your subsequent wwPDB deposition.
- Click the Run button at the bottom to complete the pdb_extract session.

Information about Authors, Structure Description, and Experiments ...

Click here to edit and download a re-usable template via PDBj's mmCIF Editor if you haven't done so (*Note: click Editor's upper-left gear icon to save/download or access help)
Upload completed mmCIF-format template: Choose File No file chosen (*Note: legacy template not in mmCIF format is no longer accepted)
Or you can skip this step and fill in the information when you deposit your structure through wwPDB OneDep deposition system.

pdb_extract Page #2, Part 2 (MX)



pdb_extract Page #2, Part 3 (All Methods)

Macromolecular Sequence Information (polymer entity)

Provide sample sequence of each molecule in the empty text box, using standard one-letter codes. Please include the complete sequence of all residues used in the experiment including expression tags, linkers, mutations, and unobserved residues due to disorder. Non-standard residues should be input using the three-letter code in parenthesis, e.g. (MSE).

Entity identifier	1
Sequence in Model	MQKEPRKVREFRRREQEILDTALKLFLEQGEDSVTVEMIADAVGIGKGTIYKHFKSKAEIYLRLMLDYERDLAALFHSED VARDKEALSRAYFEFRMRDPQRYRLFDRLEEKVVKTSQVPEMVEELHKIRASNFERLTQLIKERIADGKLENVPPYFHYC AAWALVHGAVALYHSPFWREVLEDQEGFFHFLMDIGVRMGNKRKREGDAPSA
Provide Full Sequence in Sample	
Chain ID	A,B
Polymer Type	polypeptide(L)

Run Reset

PDBj (PDBx/mm)CIF Editor

- Two general modes of access
  - Via pdb_extract
    - For creation / editing of input template files
    - Link is at the top of pdb_extract pages #1 and #2
  - Standalone Editor at https://pdbj.org/cif-editor/
- Functions
  - Build new PDBx/mmCIF formatted files
  - Edit existing PDBx/mmCIF formatted files
  - Check PDBx/mmCIF formatted file content against an mmCIF dictionary
- Applications
  - Manipulate PDBx/mmCIF *safely*
  - Prepare content for pdb_extract or for direct EMDB deposition

CIF Editor

CIF Editor

Interactive mmCIF-file editor

Access

https://pdbj.org/cif-editor/

Load a file that is accessible via the web

https://pdbj.org/cif-editor/#https://data.pdbj.org/pub/pdb/data/structures/divided/mmCIF/cr/1crn.cif

Load a file from your local HDD

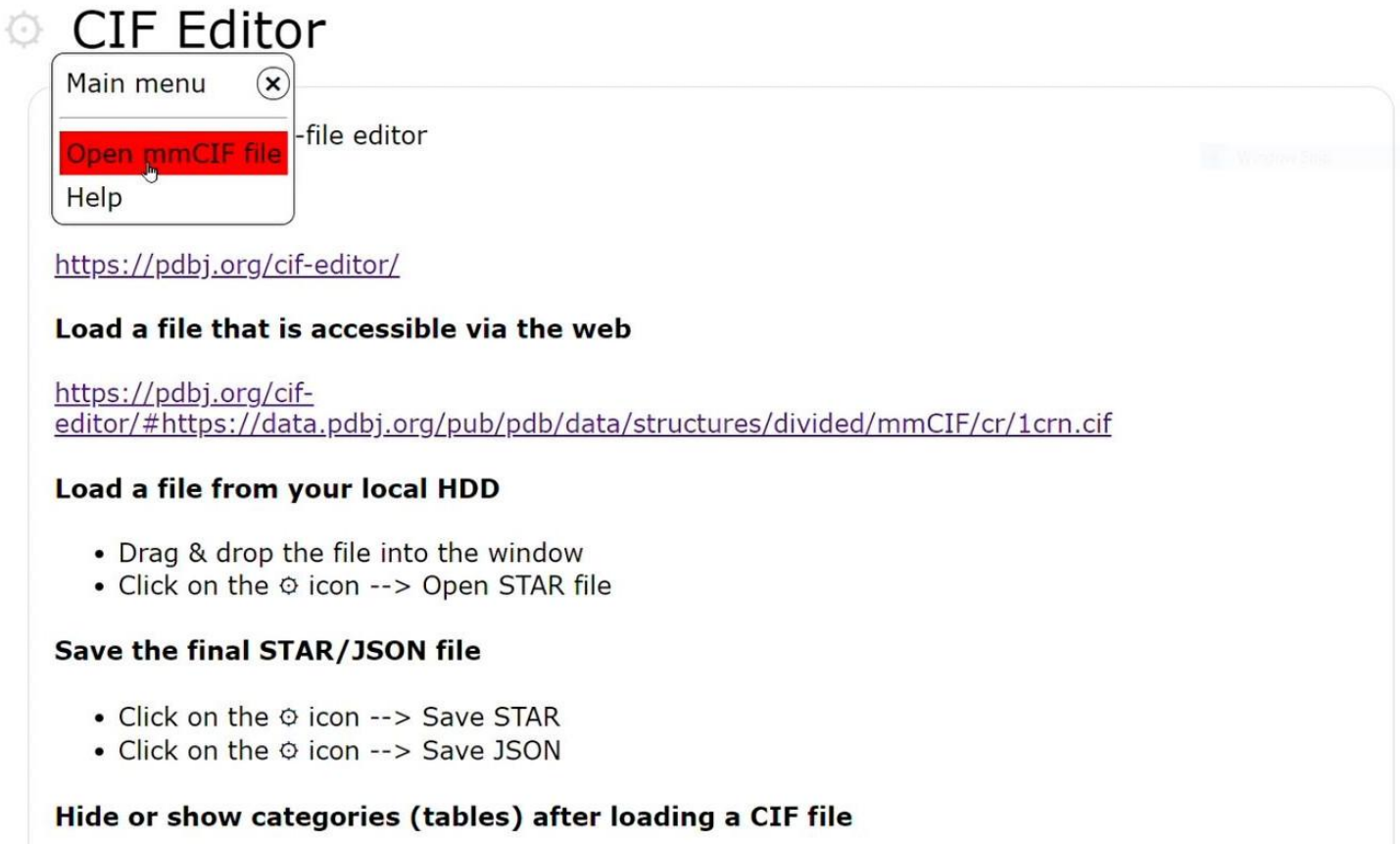
- Drag & drop the file into the window
- Click on the ⚙ icon --> Open STAR file

Save the final STAR/JSON file

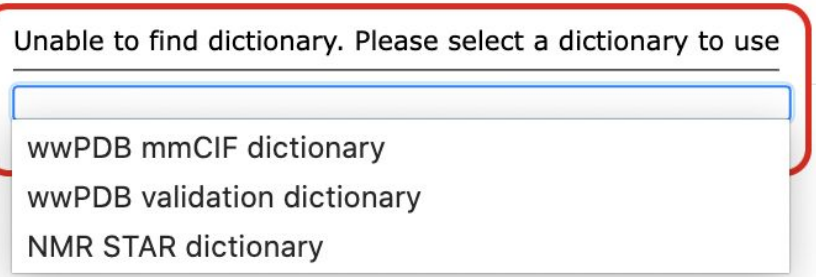
- Click on the ⚙ icon --> Save STAR
- Click on the ⚙ icon --> Save JSON

Hide or show categories (tables) after loading a CIF file

CIF Editor


CIF Editor
Main menu
Open mmCIF file
Help
-file editor
https://pdbj.org/cif-editor/
Load a file that is accessible via the web
https://pdbj.org/cif-editor/#https://data.pdbj.org/pub/pdb/data/structures/divided/mmCIF/cr/1crn.cif
Load a file from your local HDD
• Drag & drop the file into the window
• Click on the ⚙ icon --> Open STAR file
Save the final STAR/JSON file
• Click on the ⚙ icon --> Save STAR
• Click on the ⚙ icon --> Save JSON
Hide or show categories (tables) after loading a CIF file

CIF Editor



Interactive mmCIF-file editor

Access

https://pdbj.org/cif-editor/

Load a file that is accessible via the web

https://pdbj.org/cif-editor/#https://data.pdbj.org/pub/pdb/data/structures/divided/mmCIF/cr/1crn.cif

Load a file from your local HDD

- Drag & drop the file into the window
- Click on the ⚙ icon --> Open STAR file

Save the final STAR/JSON file

- Click on the ⚙ icon --> Save STAR
- Click on the ⚙ icon --> Save JSON

Hide or show categories (tables) after loading a CIF file

CIF Editor (disp

Problems have been identified with the supplied CIF data

Invalid data values:
entity_src_nat.pdbx_src_id

Close

audit_author

	name*!	pdbx_ordinal*
×	Laurel, S.	1
×	Hardy, O.	2

cell

	angle_alpha	angle_beta	angle_gamma	entry_id*	length_a	length_b	length_c
×	90.00	90.00	90.00	D_8000240802	50.0	40.0	60.0

citation

	id*	journal_abbrev	title!
×	primary	To be Published	This is a citation title

CIF Editor (disp

Problems have been identified with the supplied CIF data

Invalid data values:
entity_src_nat.pdbx_src_id

Close

audit_author

	name*!	pdbx_ordinal*
×	Laurel, S.	1
×	Hardy, O.	2

cell

	angle_alpha	angle_beta	angle_gamma	entry_id*	length_a	length_b	length_c
×	90.00	90.00	90.00	D_8000240802	50.0	40.0	60.0

citation

Click to add a row

	id*	journal_abbrev	title!
×	primary	To be Published	This is a citation title

Click to delete row

CIF Editor (display.cif)

Main menu

- View dictionary
- Toggle tables
- Merge additional mmCIF file
- Re-validate CIF
- Save mmCIF
- Save mmJSON
- Help

	angle_alpha	angle_beta	angle_gamma	entry_id*	length_a	length_b	length_c
×	90.00	90.00	90.00	D_8000240802	50.0	40.0	60.0

citation

	id*	journal_abbrev	title!
×	primary	To be Published	This is a citation title

CIF Editor (display.cif)

Main menu

- View dictionary
- Toggle tables
- Merge additional mmCIF file
- Re-validate CIF
- Save mmCIF
- Save mmJSON
- Help

	angle_alpha	angle_beta	angle_gamma	entry_id*	length_a	length_b	length_c
×	90.00	90.00	90.00	D_8000240802	50.0	40.0	60.0

citation

	id*	journal_abbrev	title!
×	primary	To be Published	This is a citation title

CIF Editor (display.cif)

Entry editor: select tables

- [ ] entry: There is only one item in the ENTRY category, _entry.id. This data item gives a name to this entry and is indirectly a key to the categories (such as CELL, GEOM, EXPTL) that describe information pertinent to the entire data block.
- [ ] entry_link: Data items in the ENTRY_LINK category record the relationships between the current data block identified by _entry.id and other data blocks within the current file which may be referenced in the current data block.
- [ ] exptl: Data items in the EXPTL category record details about the experimental work prior to the intensity measurements and details about the absorption-correction technique employed.
- [ ] exptl_crystal: Data items in the EXPTL_CRYSTAL category record the results of experimental measurements on the crystal or crystals used, such as shape, size or density.
- [ ] exptl_crystal_face: Data items in the EXPTL_CRYSTAL_FACE category record details of the crystal faces.
- [x] exptl_crystal_grow: Data items in the EXPTL_CRYSTAL_GROW category record details about the conditions and methods used to grow the crystal.
- [ ] exptl_crystal_grow_comp: Data items in the EXPTL_CRYSTAL_GROW_COMP category record details about the components of the solutions that were 'mixed' (by whatever means) to produce the crystal. In general, solution 1 is the solution that contains the molecule to be crystallized and solution 2 is the solution that contains the precipitant. However, the number of solutions required to describe the crystallization protocol is not limited to 2. Details of the crystallization protocol should be given in _exptl_crystal_grow_comp.details using the solutions described in EXPTL_CRYSTAL_GROW_COMP.
- [ ] geom: Data items in the GEOM and related (GEOM_ANGLE, GEOM_BOND, GEOM_CONTACT, GEOM_HBOND and GEOM_TORSION) categories record details about the molecular geometry as calculated from the contents of the ATOM, CELL and SYMMETRY data. Geometry data are therefore redundant, in that they can be calculated from other more fundamental quantities in the data block. However, they provide a check on the correctness of both sets of data and enable the most important geometric data to be identified for publication by setting the appropriate publication flag.
- [ ] geom_angle: Data items in the GEOM_ANGLE category record details about the bond angles as calculated from the contents of the ATOM, CELL and SYMMETRY data.

CIF Editor (display.cif)

Main menu

- View dictionary
- Toggle tables
- Merge additional mmCIF file
- Re-validate CIF
- Save mmCIF
- Save mmJSON
- Help

	angle_alpha	angle_beta	angle_gamma	entry_id*	length_a	length_b	length_c
×	90.00	90.00	90.00	D_8000240802	50.0	40.0	60.0

citation

	id*	journal_abbrev	title!
×	primary	To be Published	This is a citation title

If You Are Preparing a PDB Deposition

Which PDBx/mmCIF formatted atomic coordinate file content is required for OneDep upload?

MX

- Data Block Label
- Category: **atom_site**
  - XYZ Atomic Coordinates
  - Atom Labels
  - Residue Numbers and Labels
  - Chain IDs
- Category: **cell**
  - Unit Cell Lengths (a,b,c)
  - Unit Cell Angles (α,β,ɣ)
- Category: **symmetry**
  - Space Group
- Category: **refine**
  - High Resolution Limit

3DEM

- Data Block Label
- Category: **atom_site**
  - XYZ Atomic Coordinates
  - Atom Labels
  - Residue Numbers and Labels
  - Chain IDs

NMR

- Data Block Label
- Category: **atom_site**
  - XYZ Atomic Coordinates
  - Atom Labels
  - Residue Numbers and Labels
  - Chain IDs
  - **Model Numbers**

If You Are Preparing a PDB Deposition

Which PDBx/mmCIF formatted atomic coordinate file content is required for OneDep upload?

MX

- Data Block Label
- Category: **atom_site**
  - XYZ Atomic Coordinates
  - Atom Labels
  - Residue Numbers and Labels
  - Chain IDs
- Category: **cell**
  - Unit Cell Lengths (a,b,c)
  - Unit Cell Angles (α,β,ɣ)
- Category: **symmetry**
  - Space Group
- Category: **refine**
  - High Resolution Limit

3DEM

- Data Block Label
- Category: **atom_site**
  - XYZ Atomic Coordinates
  - Atom Labels
  - Residue Numbers and Labels
  - Chain IDs

NMR

- Data Block Label
- Category: **atom_site**
  - XYZ Atomic Coordinates
  - Atom Labels
  - Residue Numbers and Labels
  - Chain IDs
  - **Model Numbers**

OneDep will require ***lots*** of additional metadata! Use pdb_extract and the PDBj CIF editor to prepare PDBx/mmCIF formatted atomic coordinate files that are more useful for deposition.

Any Questions?

Programmatic File Access and Data Parsing Using PDBx/mmCIF Files - Part I

Irina Persikova Ph.D.

RCSB PDB, Rutgers University

Data Access - Files and Services Overview

- What structure data files are available in PDBx/mmCIF format
- What is their content
- How to access the files

Atomic Coordinates

Atomic coordinates

- xyz
- residue names, numbers
- atom names, types
- chain ids, asym ids
- temperature factors
- occupancy

Owen *et al*., (2021) Science **374**: 1586-1593 doi: 10.1126/science.abl4784
PDB doi: 10.2210/pdb7RFS/pdb

```
…
_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   . SER A 1 1   ? -2.625 4.852   -17.094 1    40.95 ? 1   SER A N   1
ATOM   2       C CA  . SER A 1 1   ? -2.258 6.069   -16.373 1    42.2  ? 1   SER A CA  1
ATOM   3       C C   . SER A 1 1   ? -2.529 5.901   -14.874 1    42.02 ? 1   SER A C   1
ATOM   4       O O   . SER A 1 1   ? -3.149 4.925   -14.455 1    42.27 ? 1   SER A O   1
ATOM   5       C CB  . SER A 1 1   ? -3.045 7.274   -16.891 1    45.05 ? 1   SER A CB  1
ATOM   6       O OG  . SER A 1 1   ? -3.105 7.319   -18.307 1    50.43 ? 1   SER A OG  1
```

Metadata In Atomic Coordinate Files

Metadata

- Primary
  - Title, authorship, citation
  - Polymer, non-polymer components, polymer sequences
  - Molecule names, source, expression system
  - Experimental conditions, data collection instruments
  - Data collection & refinement statistics, software
  - Structure determination method
- Derived/Added
  - UNP sequence alignment and discrepancy annotation (mutation, tag…)
  - Secondary structure
  - Assembly info
  - Unobserved residues, atoms
  - Links (covalent, metal, disulfide)

```
_diffrn_source.source                          SYNCHROTRON
_diffrn_source.target                          ?
_diffrn_source.type                            'SPRING-8 BEAMLINE BL26B2'
_diffrn_source.voltage                         ?
_diffrn_source.pdbx_wavelength                 1.0
_diffrn_source.pdbx_synchrotron_beamline       BL26B2
_diffrn_source.pdbx_synchrotron_site           SPring-8
…
```

```
…
_struct_ref_seq_dif.pdbx_seq_db_seq_num
_struct_ref_seq_dif.details
_struct_ref_seq_dif.pdbx_auth_seq_num
_struct_ref_seq_dif.pdbx_ordinal
1 8I5W GLY A 1 ? UNP Q8BZN6 ? ? 'expression tag' 1869 1
1 8I5W SER A 2 ? UNP Q8BZN6 ? ? 'expression tag' 1870 2
1 8I5W SER A 3 ? UNP Q8BZN6 ? ? 'expression tag' 1871 3
```

Chemical Component Files

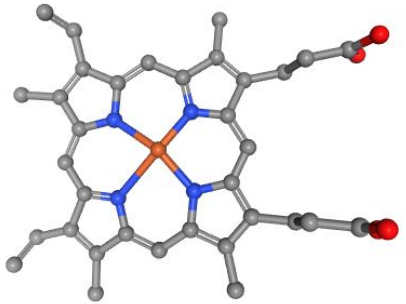
External reference file describing all residue and small molecule components found in PDB entries.

```
_chem_comp.id                                   HEM
_chem_comp.name                                 "PROTOPORPHYRIN IX CONTAINING FE"
_chem_comp.type                                 NON-POLYMER
_chem_comp.formula                              "C34 H32 Fe N4 O4"
_chem_comp.pdbx_synonyms                        HEME
_chem_comp.pdbx_formal_charge                   0
_chem_comp.pdbx_initial_date                    1999-07-08
_chem_comp.pdbx_modified_date                   2020-06-17
_chem_comp.formula_weight                       616.487
_chem_comp.pdbx_model_coordinates_db_code       3IA3
…
```

```
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
HEM  CHA  C1A   SING  N  N  1
HEM  CHA  C4D   DOUB  N  N  2
HEM  CHA  HHA   SING  N  N  3
HEM  CHB  C4A   SING  N  N  4
…
```

```
…
_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
…
HEM  CHA   CHA   C  0  1  N  N  N   2.748  -19.531  39.896  -2.161  -0.125   0.490  CHA  HEM  1
HEM  CHB   CHB   C  0  1  N  N  N   3.258  -17.744  35.477   1.458  -3.419   0.306  CHB  HEM  2
HEM  CHC   CHC   C  0  1  N  N  N   1.703  -21.900  33.637   4.701   0.169  -0.069  CHC  HEM  3
HEM  CHD   CHD   C  0  1  N  N  N   1.149  -23.677  38.059   1.075   3.460   0.018  CHD  HEM  4
HEM  C1A   C1A   C  0  1  Y  N  N   3.031  -18.673  38.872  -1.436  -1.305   0.380  C1A  HEM  5
...
```

- Name, formula, weight
- Coordinates (model, ideal)
- Connectivity
- Descriptors (SMILES, InChI)
- Synonyms
- Additional identifiers



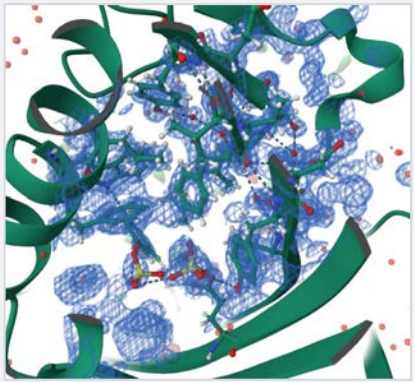
https://www.wwpdb.org/data/ccd

Structure Factor Files

- Reflection indices, the measured value of the structure factors and intensities and their sigma values, free flags, metadata
- Might contain data blocks for refinement, phasing, unmerged reflection data, anomalous diffraction data

```
_cell.entry_id           8abt
_cell.length_a           78.045
_cell.length_b           78.045
_cell.length_c           159.541
_cell.angle_alpha        90.000
_cell.angle_beta         90.000
_cell.angle_gamma        120.000
#
_diffrn_radiation_wavelength.id           1
_diffrn_radiation_wavelength.wavelength   0.97625
#
_symmetry.entry_id                 8abt
_symmetry.space_group_name_H-M     "P 63"
_symmetry.Int_Tables_number        173
```

```
loop_
_refln.crystal_id
_refln.wavelength_id
_refln.scale_group_code
_refln.index_h
_refln.index_k
_refln.index_l
_refln.status
_refln.pdbx_r_free_flag
_refln.F_meas_au
_refln.F_meas_sigma_au
_refln.intensity_meas
_refln.intensity_sigma
_refln.pdbx_FWT
_refln.pdbx_PHWT
_refln.pdbx_DELFWT
_refln.pdbx_DELPHWT
1 1 1 0  0  4  o 6  25.16   17.07 -1.16   2.66  9.59    180.00 1179.49 -0.00
1 1 1 0  0  6  o 10 472.40  5.15  240.26  5.23  564.09  0.00   237.72  0.00
1 1 1 0  0  8  o 8  836.55  6.55  753.44  11.80 1087.77 180.00 289.19  180.00
1 1 1 0  0  10 o 17 539.03  5.57  312.83  6.46  664.78  180.00 190.19  180.00
1 1 1 0  0  12 f 0  665.20  6.07  476.43  8.69  825.01  0.00   209.50  180.00
1 1 1 0  0  16 o 18 831.07  6.84  743.74  12.22 1004.64 0.00   210.97  -0.00
```



Validation Reports

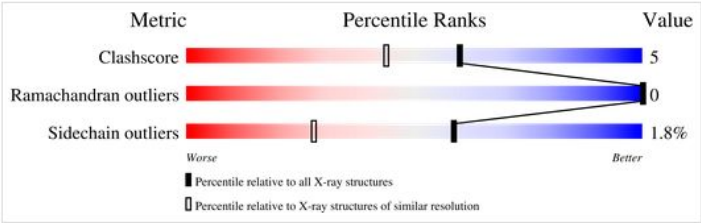
All (2) bond angle outliers are listed below:

Mol	Chain	Res	Type	Atoms	Z	Observed(^{o})	Ideal(^{o})
2	A	200	REA	C11-C10-C9	-2.40	123.89	127.31
2	A	200	REA	C18-C5-C6	2.05	126.83	124.53

PDF

mmCIF

```
loop_
_pdbx_vrpt_instance_mogul_angle_outliers.ordinal
_pdbx_vrpt_instance_mogul_angle_outliers.instance_id
_pdbx_vrpt_instance_mogul_angle_outliers.atom_1
_pdbx_vrpt_instance_mogul_angle_outliers.atom_2
_pdbx_vrpt_instance_mogul_angle_outliers.atom_3
_pdbx_vrpt_instance_mogul_angle_outliers.obsval
_pdbx_vrpt_instance_mogul_angle_outliers.mean
_pdbx_vrpt_instance_mogul_angle_outliers.numobs
_pdbx_vrpt_instance_mogul_angle_outliers.stdev
_pdbx_vrpt_instance_mogul_angle_outliers.Zscore
_pdbx_vrpt_instance_mogul_angle_outliers.mindiff
1  138  C18  C5   C6  126.83  124.53   89  1.12  2.05  0.06
2  138  C11  C10  C9  123.89  127.31  139  1.43  2.40  0.65
```



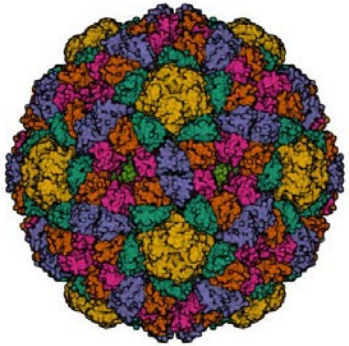
- Contain the same info as in PDF and XML files
- Well organized, easy to parse
- Interoperable with the PDB archival format
- Entity, chain and residue level information

mmcif.wwpdb.org/dictionaries/mmcif_pdbx_vrpt.dic/Index

Biological Assembly Files

1AL0 PROCAPSID OF BACTERIOPHAGE PHIX174

Asymmetric Unit



Biological Assembly

- Macromolecular assembly that has either been shown or is believed to be the functional form of the molecule
- Include all symmetry generated copies of each chain within a single model, with distinct chain IDs
- Check if your visualization software (e.g., PyMol, ChimeraX, etc.) supports instantiation of assemblies directly from atomic coordinate files (for efficiency)

Dokland *et al*., (1997) Nature **389**: 308-313 doi: 10.1038/38537
PDB doi: 10.2210/pdb1AL0/pdb

```
...
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
ATOM 188088 N N   . TYR E-20 2 353 ? -22.804  20.710   -109.287 1.00 20.00 ? ? ? ? ? ? 353 TYR F-20 N  1
ATOM 188089 C CA  . TYR E-20 2 353 ? -24.203  20.346   -109.349 1.00 20.00 ? ? ? ? ? ? 353 TYR F-20 CA 1
ATOM 188090 C C   . TYR E-20 2 353 ? -24.417  19.132   -110.204 1.00 20.00 ? ? ? ? ? ? 353 TYR F-20 C  1
ATOM 188091 O O   . TYR E-20 2 353 ? -23.674  18.875   -111.121 1.00 20.00 ? ? ? ? ? ? 353 TYR F-20 O  1
ATOM 188092 C CB  . TYR E-20 2 353 ? -25.038  21.510   -109.869 1.00 20.00 ? ? ? ? ? ? 353 TYR F-20 CB 1
```

Accessing Data Files from the PDB Archive

- Maintained by the wwPDB at files.wwpdb.org
- All data are available via the HTTPS protocol
- The FTP protocol will be phased out on November 1st, 2024

Atomic coordinates, experimental data and validation reports are grouped by the n-1, n-2 characters of the 4-character PDB identifier:
../**pub**/**pdb**/**data**/**structures**/**divided**/**mmCIF**/**01**/1**01**d.cif.gz

CCD: ../pub/pdb/refdata/chem_comp/**M**/HE**M**/HEM.cif

../pub/pdb/data/monomers/components.cif

- **pub**
  - **pdb**
    - refdata
      - chem_comp
    - **data**
      - assemblies
        - mmcif
          - divided
      - **structures**
        - **divided**
          - **mmcif**
          - structure_factors
      - monomers

Accessing Files from the Versioned Archive

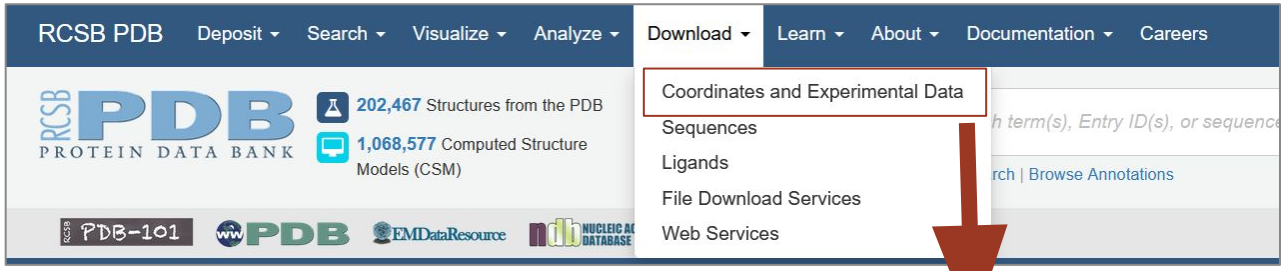
- Maintained by the wwPDB at files-versioned.wwpdb.org
- Contains latest and prior versions of each entry
  - Latest minor of each major version
- Uses the extended PDB codes in file naming ../pdb_versioned/data/entries/ab/pdb_00001abc/
- PDB archive and versioned archive are updated Wednesdays at 00:00 UTC

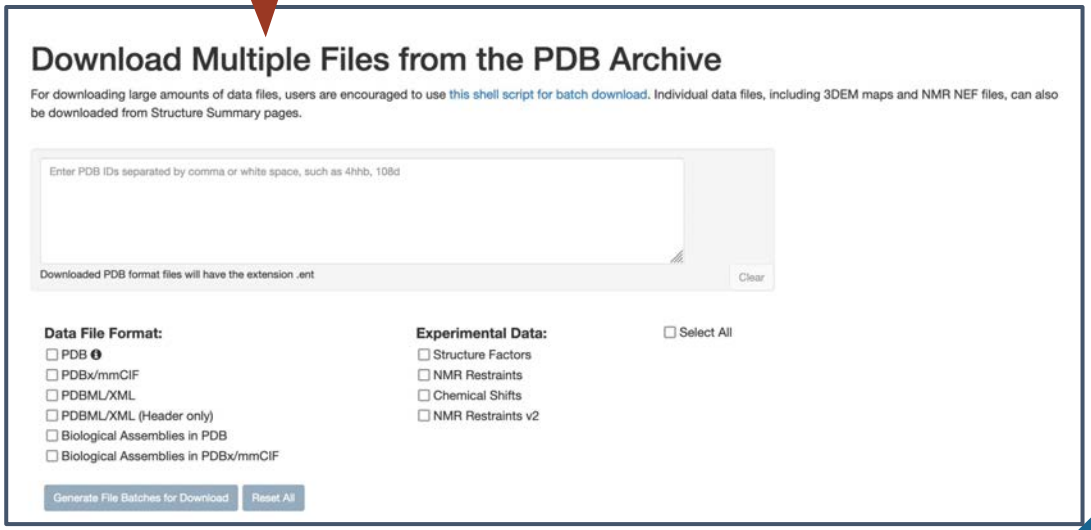
PDB - FTP Archive over HTTP

Name	Last modified	Size
« Parent Directory		-
pdb_00006lu7_xyz_v1-0.cif.gz	2020-01-31 12:22	72K
pdb_00006lu7_xyz_v1-0.xml.gz	2020-01-31 12:22	93K
pdb_00006lu7_xyz_v2-9.cif.gz	2020-06-19 12:49	74K
pdb_00006lu7_xyz_v2-9.xml.gz	2020-06-19 12:49	96K
pdb_00006lu7_xyz_v3-1.cif.gz	2021-03-05 12:50	74K
pdb_00006lu7_xyz_v3-1.xml.gz	2021-03-05 12:50	96K

- Major version - updates to atomic coordinates, polymer sequence, or chemical description
- Minor version - other changes to the metadata

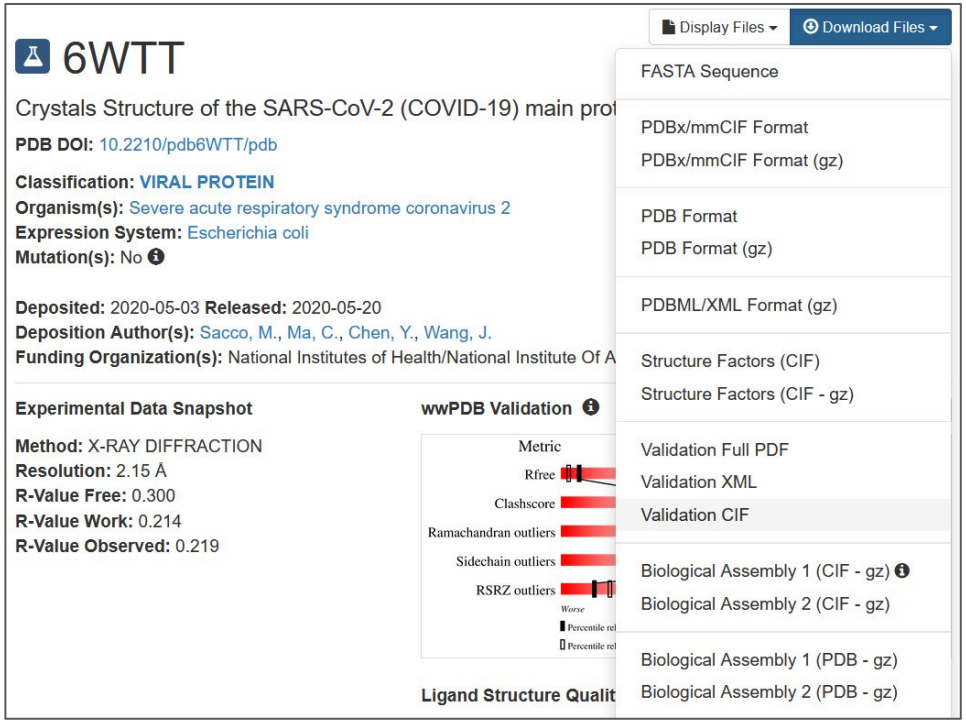
RCSB.org Offers Access and Download Options





File Access - *via* RCSB.org

- Using download option on the SSP:



- Direct download using short style links:

https://files.rcsb.org/download/6wtt.cif
https://files.rcsb.org/download/6wtt-sf.cif
https://files.rcsb.org/download/6wtt-assembly1.cif
https://files.rcsb.org/download/6wtt-assembly2.cif

- Compressed, binary and header only files are also available.

- CCD:

https://files.rcsb.org/ligands/download/HEM.cif

File Access - Download a Set of Files with Python

- Direct URLs are convenient to use for programmatic file download
- For example using Python **urllib** module

```
import urllib.request

filepath_local = '/tmp/4HHB.cif'

pdb_url ='https://files.rcsb.org/download/4HHB.cif'

urllib.request.urlretrieve(pdb_url, filepath_local)
```

File Access - RSYNC

RCSB PDB:

https://www.wwpdb.org/ftp/pdb-ftp-sites

USING RSYNC PROTOCOL:

```
rsync --port=33444 rsync.rcsb.org::
ftp             Top level of PDB ftp tree ( /pub/pdb )
ftp_data        Data directory within PDB ftp archive ( /pub/pdb/data )
ftp_derived     Derived data directory within PDB ftp archive ( /pub/pdb/derived_data )
ftp_doc         Doc directory within PDB ftp archive ( /pub/pdb/doc )
emdb            Top level of EMDB ftp tree ( /pub/emdb )
```

Download coordinate files in PDB Exchange Format (mmCIF):

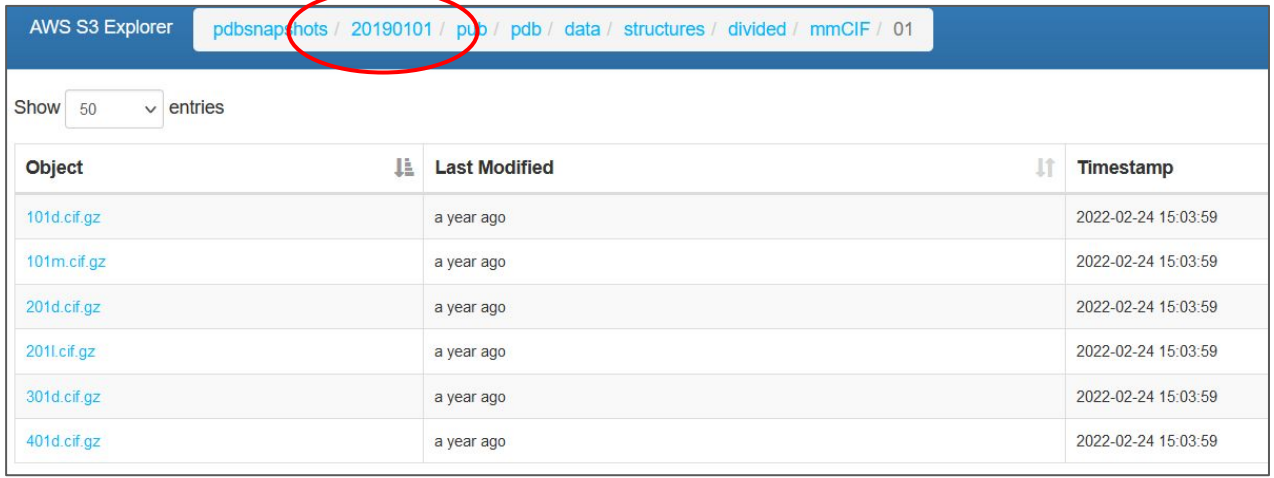
```
rsync -rlpt -v -z --delete --port=33444 \
rsync.rcsb.org::ftp_data/structures/divided/mmCIF/ ./mmCIF
```

File Access - Archive Snapshots

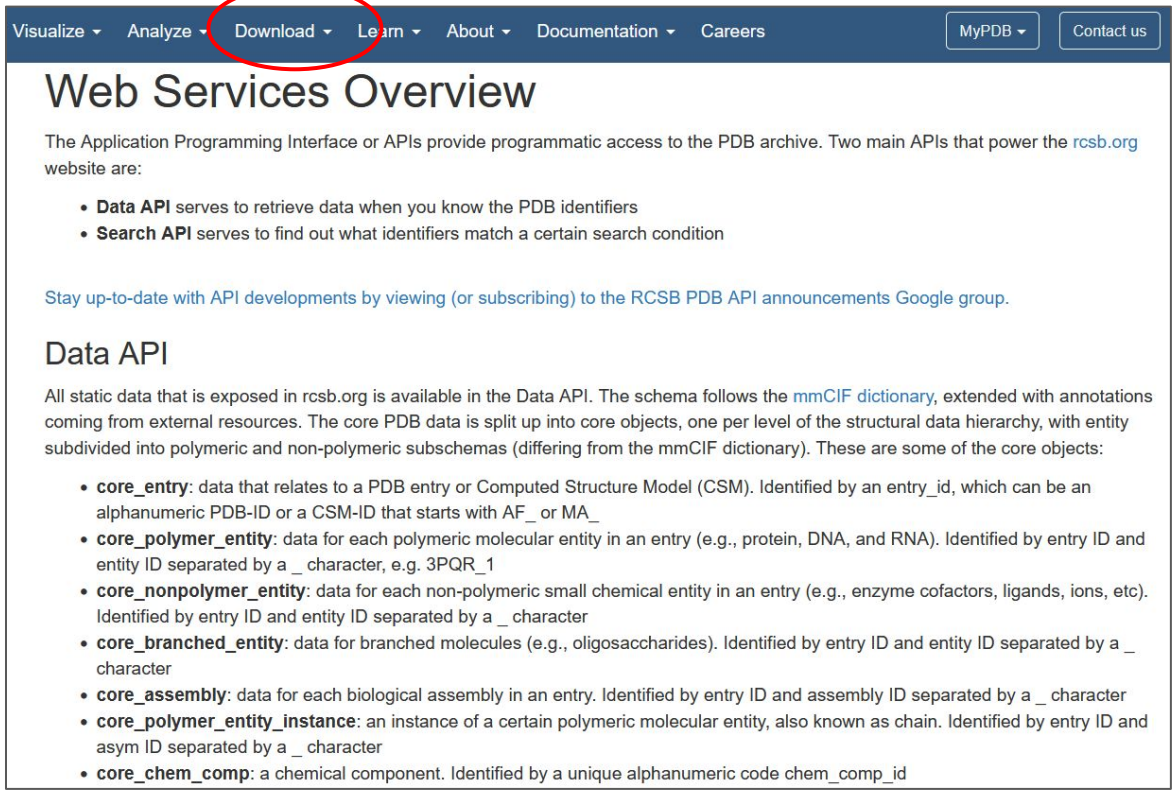
Snapshots have been archived annually since 2005.

RCSB PDB (US/AWS): https://s3snapshots.rcsb.org

Archival snapshot from 2018



File Access - API Services



https://www.rcsb.org/docs/programmatic-access/web-services-overview

Data API:

- Retrieve data when you know the PDB identifiers
- Follows the mmCIF dictionary
- Extended with annotations from external resources
- Two interfaces:
  - REST API
  - GraphQL interface
- Output in JSON format
- Tutorial
  - https://data.rcsb.org/index.html#data-api

To Summarize

Flexible and extensible PDBx/mmCIF format can accommodate a diverse structural information available as

- Atomic coordinate files
- Structure factor files
- Chemical component files (CCD)
- Biological assembly files
- Validation report files

wwPDB provides

- Individual and batch file download options
- *Via* FTP/HTTPS/RSYNC protocols

questions?

Programmatic File Access and Data Parsing Using PDBx/mmCIF Formatted Files - Part II

Chenghua Shao Ph.D.

RCSB PDB, Rutgers University

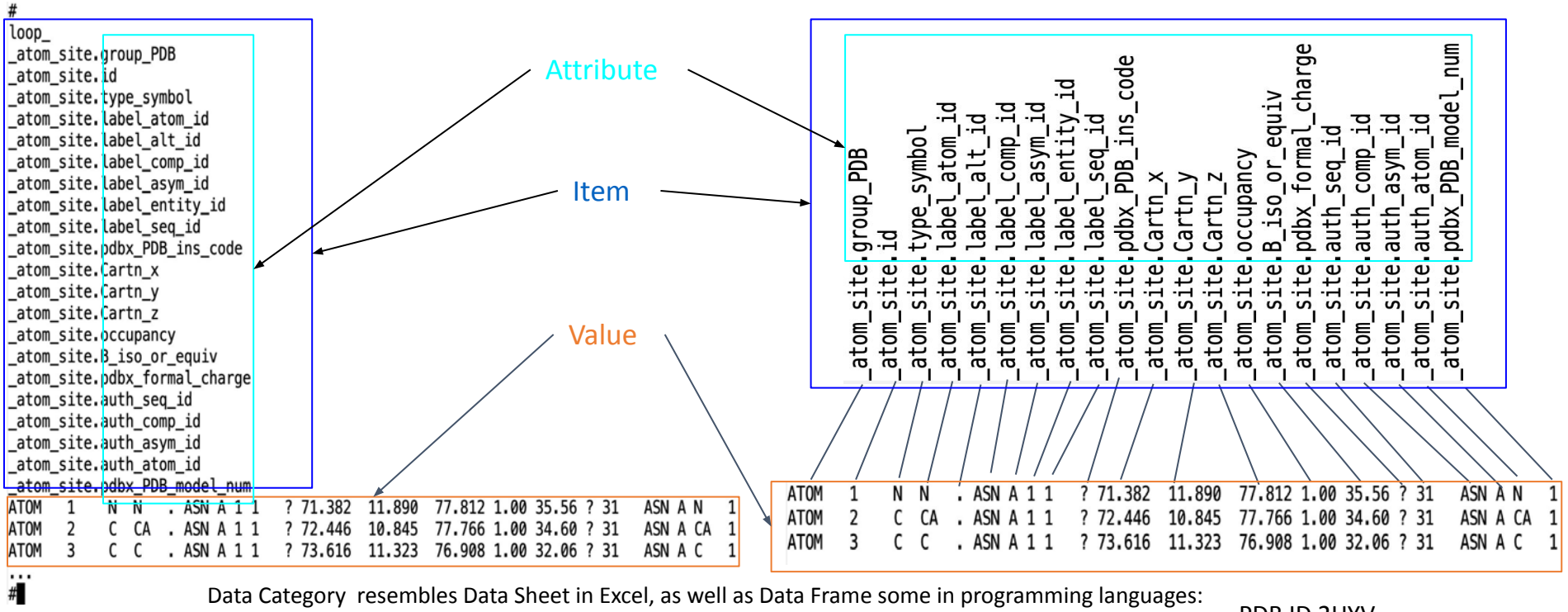
Data Parsing - PDBx/mmCIF Parsers

- PDBx/mmCIF files can be parsed by program easily
- Various PDBx/mmCIF parsers available
  - https://mmcif.wwpdb.org/docs/software-resources.html
- “mmcif” parser is the primary parser used by wwPDB
  - A Python PDBx/mmCIF API wrapping the PDB C++ Core PDBx/mmCIF Library
  - PDBx/mmCIF dictionary API included
  - Fast and reliable
  - Freely available through PyPI and GitHub
    - pip install mmcif
    - https://github.com/rcsb/py-mmcif

Data Parsing - PDBx/mmCIF Data Hierarchy

PDB ID 2HYV

Data Parsing - Focusing on Data Category



Data Category resembles Data Sheet in Excel, as well as Data Frame some in programming languages:

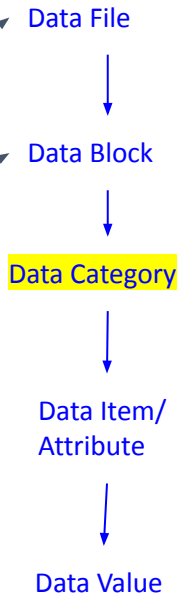
- Items/Attributes as column headers
- Values as rows of each record

Data Category is intrinsically similar to Python Dictionary data type, with resembling key/value pairs

PDB ID 2HYV

Data Parsing - “mmcif” Python Package Usage

- Import mmcif python package
- Create a parser class instance
- Read PDBx/mmCIF file
- Parse data content into data objects
  - Level 1: File-level object: List of Data Containers
    - Operate data blocks in the file
  - Level 2: Block-level object: Data Container
    - Operate data categories in the data block
  - Level 3: Category-level object: Data Category
    - Operate data items/attributes and values within a data category



Data Parsing - “mmcif” Data Object Hierarchy

Import package

```
from mmcif.io.IoAdapterCore import IoAdapterCore
# File IO class, read and write mmCIF file
```

Create parser instance

```
io = IoAdapterCore()
```

File-level obj

```
list_data_container = io.readFile("2hyv.cif")
# read file, generate list of data containers
```

Block-level obj

```
data_container = list_data_container[0]
# select the 1st data container which contains the 1st block
```

Category-level obj

```
entity = data_container.getObj('entity')
# obtain entity category data
```

Examples codes in this and the following slides can be found at: https://github.com/rcsb/py-mmcif_demo

Data Parsing Example - Molecular Entity

```
entity = data_container.getObj('entity')
print(entity.data)
# get all data in the entity category
[['1', 'polymer', 'man', 'Annexin A2', '35352.344',...],
['2', 'branched', 'man', '4-deoxy…', '1411.128',...],...]
print(entity.getAttributeValueList("type"))
# get list of values by attr
['polymer', 'branched', 'non-polymer', 'water']
print(entity.getRowAttributeDict(0))
# get 1st data row as dictionary
{'id': '1', 'type': 'polymer', 'src_method': 'man',
'pdbx_description': 'Annexin A2',...}
print(entity.getValue("pdbx_description",0))
# get value of a data cell by attr and index
Annexin A2
```

```
data_2HYV
#
...
#
loop_
_entity.id
_entity.type
_entity.src_method
_entity.pdbx_description
_entity.formula_weight
_entity.pdbx_number_of_molecules
_entity.pdbx_ec
_entity.pdbx_mutation
_entity.pdbx_fragment
_entity.details
1 polymer     man 'Annexin A2' 35352.344 1   ? ? ? ?
2 branched    man
;4-deoxy-2-O-sulfo-alpha-L-threo-hex-4-enopyranuronic acid-(1-4)-2-deoxy-6-O-sulfo-
so-2-(sulfoamino)-alpha-D-glucopyranose-(1-4)-2-O-sulfo-alpha-L-idopyranuronic acid-
d-(1-4)-2-deoxy-6-O-sulfo-2-(sulfoamino)-alpha-D-glucopyranose-(1-4)-2-O-sulfo-al-
pha-L-idopyranuronic acid
;
1411.128  1   ? ? ? ?
3 non-polymer syn 'CALCIUM ION' 40.078    5   ? ? ? ?
4 water       nat water 18.015    518 ? ? ? ?
#
```

PDB ID 2HYV

Data Parsing Example - Ligand Entity

```
loi = data_container.getObj("pdbx_entity_instance_feature")
# select the LOI category from data container

l_index_doi = loi.selectIndices("SUBJECT OF INVESTIGATION",
"feature_type")
# get list of indices by value/attribute pair, i.e. LOI
KQF

ccd_id_1 = loi.getValue("comp_id", l_index_doi[0])
print(ccd_id_1)
# get LOI ligand CCD ID

nonpoly = data_container.getObj("pdbx_entity_nonpoly")
# select the nonpoly category from data container

l_index_nonpoly = nonpoly.selectIndices(ccd_id_1, "comp_id")
# get list of indices by value/attribute pair, i.e. LOI CCD ID

print(nonpoly.getValue("name", l_index_nonpoly[0]))
# get LOI ligand name
4-(furan-2-yl)benzoic acid
```

```
#
_pdbx_entity_instance_feature.ordinal          1
_pdbx_entity_instance_feature.comp_id          KQF
_pdbx_entity_instance_feature.asym_id          ?
_pdbx_entity_instance_feature.seq_num          ?
_pdbx_entity_instance_feature.auth_comp_id     KQF
_pdbx_entity_instance_feature.auth_asym_id     ?
_pdbx_entity_instance_feature.auth_seq_num     ?
_pdbx_entity_instance_feature.feature_type     'SUBJECT OF INVESTIGATION'
_pdbx_entity_instance_feature.details          ?
#
loop_
_pdbx_entity_nonpoly.entity_id
_pdbx_entity_nonpoly.name
_pdbx_entity_nonpoly.comp_id
2 'PROTOPORPHYRIN IX CONTAINING FE'  HEM
3 '4-(furan-2-yl)benzoic acid'       KQF
4 'CHLORIDE ION'                     CL
5 'MAGNESIUM ION'                    MG
6 water                              HOH
#
```

PDB ID 7TRT

Data Parsing Example - Atomic Coordinates

```
coordinates = data_container.getObj('atom_site')
# obtain data category from data container

l_index = coordinates.selectIndices("KQF", "auth_comp_id")
for i in l_index:
    d_row = coordinates.getRowAttributeDict(i)
    l_value = [d_row["auth_comp_id"], d_row["auth_asym_id"],
               d_row["auth_seq_id"], d_row["auth_atom_id"],
               d_row["Cartn_x"], d_row["Cartn_y"],
               d_row["Cartn_z"]]
    print('\t'.join(l_value))
# obtain coordinates of the ligand KQF
```

```
KQF     A       502     C02     4.593   5.330   17.196
KQF     A       502     C04     4.406   3.963   17.802
KQF     A       502     C05     5.170   3.614   18.912
KQF     A       502     C06     4.924   2.417   19.549
KQF     A       502     C07     3.992   1.539   19.040
……
```

```
#
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.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
......
HETATM 3169 C  C02  . KQF C 3 .  ? 4.593   5.330   17.196  0.88 6.99  ?  502  KQF A C02  1
HETATM 3170 C  C04  . KQF C 3 .  ? 4.406   3.963   17.802  0.88 6.72  ?  502  KQF A C04  1
HETATM 3171 C  C05  . KQF C 3 .  ? 5.170   3.614   18.912  0.88 8.45  ?  502  KQF A C05  1
HETATM 3172 C  C06  . KQF C 3 .  ? 4.924   2.417   19.549  0.88 8.01  ?  502  KQF A C06  1
HETATM 3173 C  C07  . KQF C 3 .  ? 3.992   1.539   19.040  0.88 8.54  ?  502  KQF A C07  1
HETATM 3174 C  C08  . KQF C 3 .  ? 3.581   0.296   19.752  0.88 11.66 ?  502  KQF A C08  1
HETATM 3175 C  C09  . KQF C 3 .  ? 3.222   0.273   21.106  0.88 10.96 ?  502  KQF A C09  1
HETATM 3176 C  C10  . KQF C 3 .  ? 2.751   -1.070  21.322  0.88 11.48 ?  502  KQF A C10  1
HETATM 3177 C  C11  . KQF C 3 .  ? 2.859   -1.662  20.070  0.88 11.13 ?  502  KQF A C11  1
HETATM 3178 C  C13  . KQF C 3 .  ? 3.226   1.856   17.935  0.88 7.26  ?  502  KQF A C13  1
HETATM 3179 C  C14  . KQF C 3 .  ? 3.457   3.069   17.304  0.88 7.15  ?  502  KQF A C14  1
HETATM 3180 O  O01  . KQF C 3 .  ? 5.544   6.063   17.546  0.88 8.61  ?  502  KQF A O01  1
HETATM 3181 O  O03  . KQF C 3 .  ? 3.614   5.731   16.488  0.88 8.34  -1 502  KQF A O03  1
HETATM 3182 O  O12  . KQF C 3 .  ? 3.336   -0.818  19.170  0.88 11.23 ?  502  KQF A O12  1
```

PDB ID 7TRT

Data Parsing Example - Ligand Definition File

```
from mmcif.io.IoAdapterCore import IoAdapterCore
```

```
# read file into data container
filepath = "KQF.cif"
io = IoAdapterCore()
list_data_container = io.readFile(filepath)
data_container = list_data_container[0]
```

```
# extract chemical descriptor from the corresponding category
descriptor = data_container.getObj('pdbx_chem_comp_descriptor')
for i in range(n_rows = descriptor.getRowCount()):
   d_row = descriptor.getRowAttributeDict(i)
   if d_row["type"] == "InChIKey":
       print(d_row["descriptor"])
```

```
ZKHQWZAMYRWXGA-KQYNXXCUSA-N
# change the descriptor type you can get SMILES and InChI
```

```
data_KQF
#
_chem_comp.id                                    KQF
_chem_comp.name                                  "4-(furan-2-yl)benzoic acid"
_chem_comp.type                                  non-polymer
_chem_comp.pdbx_type                             HETAIN
_chem_comp.formula                               "C11 H8 O3"
_chem_comp.mon_nstd_parent_comp_id               ?
_chem_comp.pdbx_synonyms                         ?
_chem_comp.pdbx_formal_charge                    0
_chem_comp.pdbx_initial_date                     2022-02-10
_chem_comp.pdbx_modified_date                    2023-03-10
_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                        188.179
_chem_comp.one_letter_code                       ?
_chem_comp.three_letter_code                     KQF
_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        7TRT
_chem_comp.pdbx_subcomponent_list                ?
_chem_comp.pdbx_processing_site                  RCSB
#
......
#
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
KQF SMILES           ACDLabs              12.01 "O=C(O)c1ccc(cc1)c1ccco1"
KQF InChI            InChI                1.03  "InChI=1S/C11H8O3/c12-
KQF InChIKey         InChI                1.03  FOJYVBSPOBUCMV-UHFFFAOYSA-N
KQF SMILES_CANONICAL CACTVS               3.385 "OC(=O)c1ccc(cc1)c2occc2"
KQF SMILES           CACTVS               3.385 "OC(=O)c1ccc(cc1)c2occc2"
KQF SMILES_CANONICAL "OpenEye OEToolkits" 2.0.7 "c1cc(oc1)c2ccc(cc2)C(=O)O"
KQF SMILES           "OpenEye OEToolkits" 2.0.7 "c1cc(oc1)c2ccc(cc2)C(=O)O"
#
```

CCD ID KQF

Data Parsing Example - Validation Report

```
from mmcif.io.IoAdapterCore import IoAdapterCore
filepath = "7r2y_validation.cif"
io = IoAdapterCore()
list_data_container = io.readFile(filepath)
data_container = list_data_container[0]

summary =
data_container.getObj('pdbx_vrpt_summary_geometry')
d_row = summary.getRowAttributeDict(0)

print("percent_ramachandran_outliers: %s%%" %
     d_row["percent_ramachandran_outliers"])
```

```
percent_ramachandran_outliers: 1.71%
```

```
print("clashscore: %s" % d_row["clashscore"])
```

```
clashscore: 8.26
```

```
#
_pdbx_vrpt_summary_geometry.ordinal                    1
_pdbx_vrpt_summary_geometry.percent_ramachandran_outliers 1.71
_pdbx_vrpt_summary_geometry.percent_rotamer_outliers   1.65
_pdbx_vrpt_summary_geometry.clashscore                 8.26
_pdbx_vrpt_summary_geometry.num_H_reduce               2402
_pdbx_vrpt_summary_geometry.num_bonds_RMSZ             2375
_pdbx_vrpt_summary_geometry.bonds_RMSZ                 0.83
_pdbx_vrpt_summary_geometry.num_angles_RMSZ            3192
_pdbx_vrpt_summary_geometry.angles_RMSZ                0.85
#
```

Validation Report for PDB ID 7R2Y

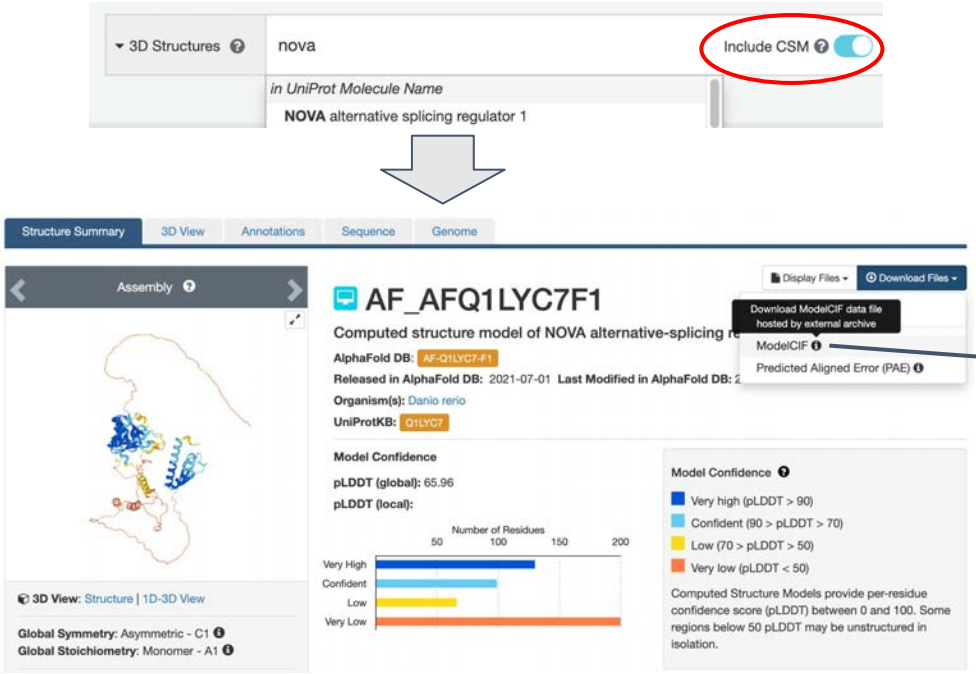
ModelCIF and Computed Structure Models

- AlphaFold2 stimulated significant interest in Computed Structure Models (CSMs)
- ModelCIF (https://github.com/ihmwg/ModelCIF) provides the data representation for describing CSM, developed and maintained by the ModelCIF Working Group.
- ModelCIF data standards is used by the ModelArchive and MODBASE repositories.
- ModelCIF is an extension of the PDBx/mmCIF dictionary, and thus can be parsed by “mmcif” parser.

J Mol Biol . 2023 Feb 23;168021. doi: 10.1016/j.jmb.2023.168021

CSM Search and Example

Turn on CSM Search at RCSB.org



ModelCIF can be parsed by “mmcif” parser as well

```
#
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.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_site.pdbx_sifts_xref_db_acc
_atom_site.pdbx_sifts_xref_db_name
_atom_site.pdbx_sifts_xref_db_num
_atom_site.pdbx_sifts_xref_db_res
ATOM 1    N N   . MET A 1 1   ? -3.846  51.891  -59.686 1.0 44.50 ? 1   MET A N   1 Q1LYC7 UNP 1   M
ATOM 2    C CA  . MET A 1 1   ? -4.197  50.651  -60.415 1.0 44.50 ? 1   MET A CA  1 Q1LYC7 UNP 1   M
ATOM 3    C C   . MET A 1 1   ? -4.723  49.680  -59.362 1.0 44.50 ? 1   MET A C   1 Q1LYC7 UNP 1   M
```

Summary of PDBx Resources

- PDBx/mmCIF dictionary https://mmcif.wwpdb.org/
  - Software resources https://mmcif.wwpdb.org/docs/software-resources.html

- ‘mmcif’ parser for PDBx/mmCIF formatted file parsing https://github.com/rcsb/py-mmcif

- ‘mmcif’ parser demo https://github.com/rcsb/py-mmcif_demo

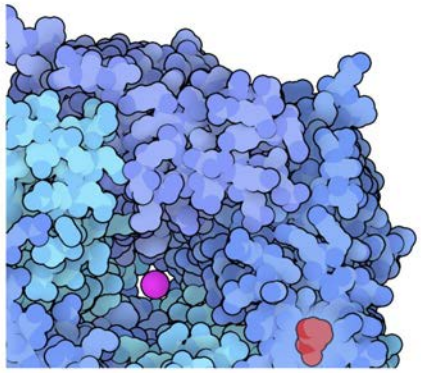
Parsing Demonstration

- Code: https://github.com/rcsb/py-mmcif_demo
- Case study #1: *Parsing data from multiple files*
  - Ligand coordinates from coordinates file
  - Chemical descriptors from ligand definition file
  - Ligand quality metrics in co-crystal structure from validation report file
- Case study #2: *Separate a set of PDB structures based on species*

Today We Have Learned

- How to understand the PDBx/mmCIF format and its dictionary
- How to read PDBx/mmCIF formatted files
- How to edit PDBx/mmCIF formatted files
- Ways to prepare PDBx/mmCIF formatted files
- How to access/parse PDBx/mmCIF data programmatically

Questions



Closing Remarks and Acknowledgements

Stephen K. Burley, M.D., D.Phil

University Professor and Henry Rutgers Chair

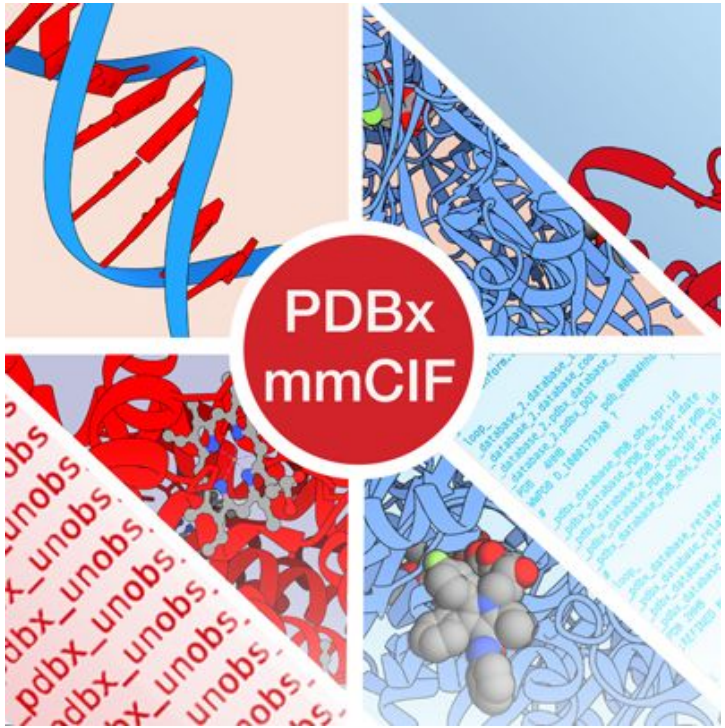
Director, RCSB Protein Data Bank
Founding Director, Institute for Quantitative Biomedicine
Rutgers, The State University of New Jersey

Thanks for attending this event

We'd love to get your feedback so that we can make the next webinar even better

Take the Zoom exit survey and receive a link to the crash course presentation slides

Thank you for your time!



Training Resources on PDB-101

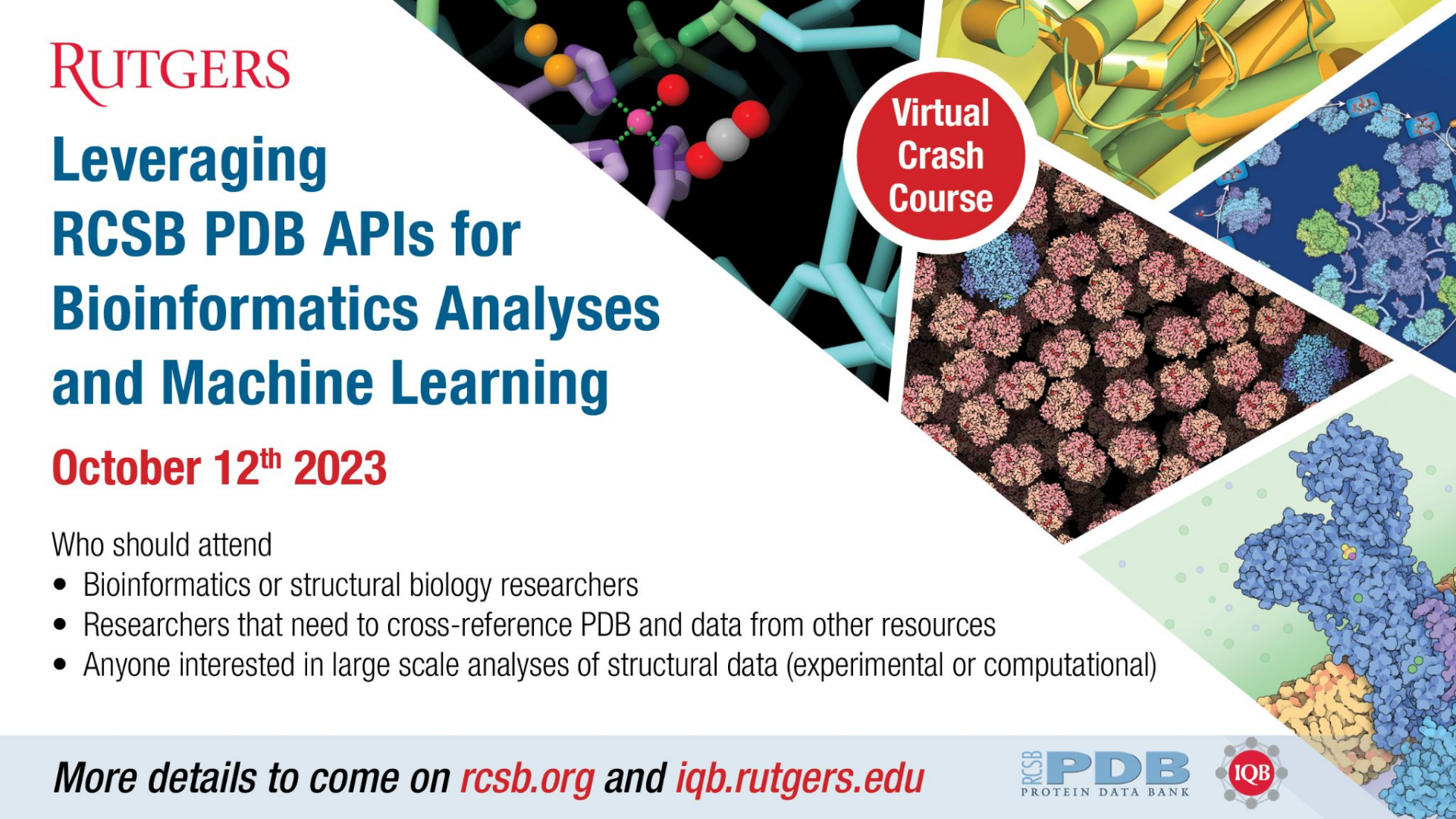
pdb101.rcsb.org > *Train*

Materials to help effectively use **RCSB.org** tools for searching, visualizing, and analyzing 3D biostructure data

- Training Courses
  Videos from today will be added
- Guide to Understanding PDB Data
- Education Corner
- PDB & Data Archiving Curriculum

RUTGERS

Leveraging RCSB PDB APIs for Bioinformatics Analyses and Machine Learning



October 12th 2023

Who should attend

- Bioinformatics or structural biology researchers
- Researchers that need to cross-reference PDB and data from other resources
- Anyone interested in large scale analyses of structural data (experimental or computational)

More details to come on rcsb.org and iqb.rutgers.edu

RCSB PDB PROTEIN DATA BANK

IQB

OPPORTUNITIES for
SCIENTIFIC SOFTWARE DEVELOPER
Graduates and Undergraduates

Develop innovative analysis, integration, query, and visualization tools for 3D biomolecular structures at **RCSB.org** to help accelerate research and training in biology, medicine, and related disciplines. Design, develop, and deploy modern web and data applications and complex interactive graphical user interfaces. Visit www.rcsb.org/pages/jobs for more information

- Database Administrator (Rutgers)
- Undergraduate Summer Research (Rutgers)
- Gap Year Opportunities (Rutgers)
- Postdoctoral Researchers
    - Metalloproteins (Rutgers)
    - Bioinformatics (UCSD)

Summer Scholars Emma and Jordi beta testing the Bound! card game

RUTGERS UC San Diego SDSC San Diego Supercomputer Center UCSF University of California San Francisco

Acknowledgements

wwPDB Members:

EMDB Electron Microscopy Data Bank

BMRB Biological Magnetic Resonance Data Bank

PDBj Protein Data Bank Japan

PDBe Protein Data Bank in Europe

RCSB PDB PROTEIN DATA BANK

RCSB PDB Team

RCSB PDB PROTEIN DATA BANK

RCSB.ORG

info@rcsb.org

Core Operations Funding

National Science Foundation (DBI-1832184), National Institute of General Medical Sciences, National Institute of Allergy and Infectious Disease, and National Cancer Institute (NIH R01GM133198), and the US Department of Energy (DE-SC0019749)

Management

RUTGERS

UC San Diego

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UCSF University of California San Francisco

WORLDWIDE PDB PROTEIN DATA BANK

Member of the Worldwide Protein Data Bank (wwPDB; **wwpdb.org**)

Follow us

John D. Westbrook
In memoriam
1957-2021