Phenix workshop · 2026

From PDB to mmCIF

A short history of macromolecular structure file formats

Oleg Sobolev
Lawrence Berkeley National Laboratory

1971 — the PDB is born

1970-71 - Meyer, Cohen and others lobbied at ACA meetings in 1970–71 for a central coordinate repository, Cold Spring Harbor 1971: Walter Hamilton (Brookhaven) agreed on the spot, Olga Kennard (CCDC, Cambridge) — transatlantic partnership

October 1971 PDB Archive launched with **7 structures**, announced in *Nature New Biology*

1972 – First PDB format is 140 characters width

1976 – PDB format is narrowed the format to 80 columns to fit the 12-row × 80-column IBM/Hollerith punched card on which structures were physically mailed.

1998 – PDB is managed by Research Collaboratory for Structural Bioinformatics (RCSB)

2003 - Worldwide Protein Data Bank (wwPDB): RCSB, PDBe, PDBj, BMR, EMDB

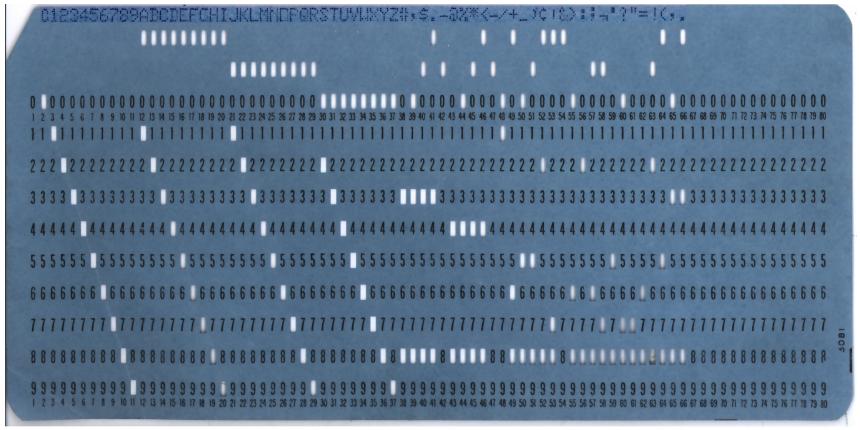
Berman, H. M. (2008). Acta Cryst. A64, 88-95.

Anatomy of a PDB ATOM record

```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
ATOM     32  N   AARG A  -3      11.281  86.699  94.383  0.50 35.88           N
```

Record name · Atom serial · Residue name · Altloc id · Residue name · Chain ID · Residue number · Atom coordinates · Occupancy · ADP · Element

- One atom, one line, 80 columns — every field has a fixed position
- 80 characters: the width of an IBM punched card
- 5-digit atom serial · 4-digit residue number · 1-char chain ID · 3-char residue name
- Every field width becomes a ceiling we will hit later



IBM 80-column Hollerith card

Hard limits, baked in

99,999 atoms per file (5-digit serial number)

9,999 residues per chain (4-digit sequence number)

~62 chains — A–Z, a–z, 0–9 in a single column

2 characters for element and formal charge — easy to corrupt

- No formal schema
- No machine-readable semantics
- No extensibility — every new field breaks every parser

Structures that don't fit

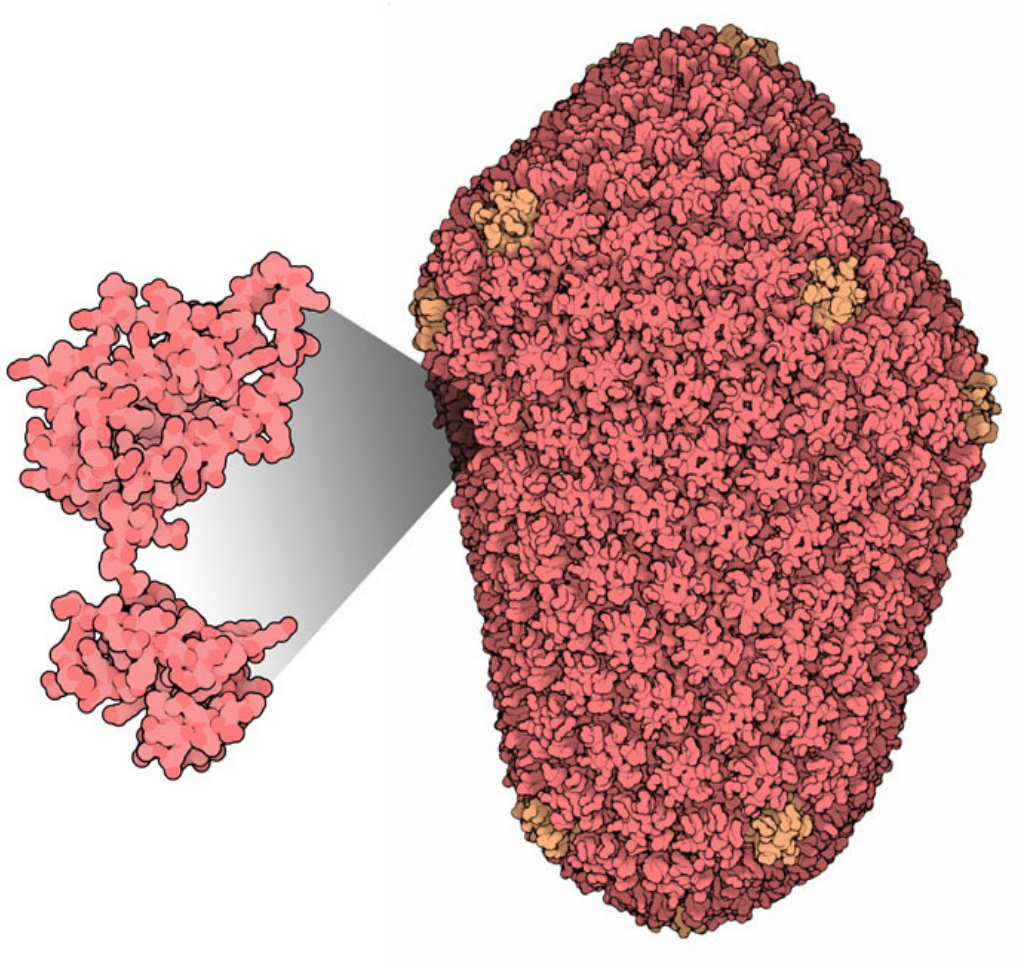


Illustration: David S. Goodsell, RCSB PDB-101 · CC BY 4.0

PDB **3J3Q** — HIV-1 capsid (cryo-EM + MD)

2,440,800 atoms — 24× the PDB ceiling

1,356 chains — 22× the legacy chain ID space

- Ribosomes, large viral capsids, big cryo-EM complexes — distributed in mmCIF only
- 3,254 PDB entries have no legacy PDB file at all (Oct 2022)

We tried to stretch the format

Hybrid-36 numbering (cctbx / Phenix)

Atom serial: 99,999 → **`87,440,031`**

Residues per chain: 9,999 → **`2,436,111`**

- Same 5 / 4 columns — switch to base-36 once base-10 overflows

Two-character chain IDs (cctbx / Phenix, MAXIT, GEMMI)

- Use the unused column 21 next to the chain ID column
- Still capped at 2 chars — ~1,036 PDB entries already exceed it

- mmCIF becomes unavoidable

Grosse-Kunstleve, cctbx, Feb 2007; https://cci.lbl.gov/hybrid_36/

1991 — CIF arrives (from a different room)

While we tried to squeeze in the 80-column box, new standard was proposed in 1991

STAR (Self-defining Text Archive and Retrieval)

- A generic plain-text format — domain-agnostic, machine-parseable, human-readable
- the *syntax* (how tokens are arranged: _name value, loop_, data_, save frames)
- Same syntax later powers **NMR-STAR** (BMRB), **NEF** (NMR Exchange Format), others

IUCr publishes **CIF** = STAR syntax + a crystallography dictionary

- The archival format for every crystal structure in an IUCr journal

Hall, S.R. (1991). *The STAR File: A New Format for Electronic Data Transfer and Archiving.* **J. Chem. Inf. Comput. Sci.** 31, 326–333. doi:10.1021/ci00002a020

Hall, S.R., Allen, F.H., Brown, I.D. (1991). *The Crystallographic Information File (CIF): a new standard archive file for crystallography.* **Acta Cryst.** A47, 655–685. doi:10.1107/S010876739101067X

Anatomy of mmCIF atom_site loop

```
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   . GLY A 1 1 ? -9.009  4.612  6.102  1.00 16.77 ? 1  GLY A N   1
```

- Every column has a name
- Arbitrary width and order
- White-space separated

PDB vs mmCIF

PDB

- Fixed 80-column record
- 99,999 atom ceiling
- 62-chain ceiling
- No schema
- Cannot extend without breaking parsers
- Position is meaning

mmCIF

- Named tags
- No limit
- Multi-character label_asym_id
- Machine-readable dictionary
- Extensible — unknown tags ignored
- Name is meaning

mmCIF — built for macromolecules

1990 — IUCr forms the mmCIF working group

- Chair: Paula Fitzgerald (Merck)
- Berman, Bourne, Dodson, Westbrook, Watenpaugh and others

1991 — DDL (Dictionary Definition Language)

1993 — first in-person meeting at York (hosted by Eleanor Dodson)

1995 — DDL2 (Dictionary Definition Language) – support relations between items

Philip E. Bourne, Helen M. Berman, Brian McMahon, Keith D. Watenpaugh, John D. Westbrook, Paula M.D. Fitzgerald (1997). Macromolecular crystallographic information file. *Methods in Enzymology*, **277**, 571-590. https://doi.org/10.1016/S0076-6879(97)77032-0

The wwPDB transition

2014 — PDBx/mmCIF becomes the wwPDB master archive format

- OneDep deposition system launches

2016 — OneDep extended to NMR and cryo-EM (3DEM) depositions

1 July 2019 — mmCIF is the ***only*** accepted deposition format for macromolecular crystallography

Announcing mandatory submission of PDBx/mmCIF format files for crystallographic depositions to the Protein Data Bank (PDB). P.D. Adams, P.V. Afonine, <…>, and J.Y. Young. *Acta Cryst.* D**75**, 451-454 (2019).

- *"We now announce that as of 1 July 2019, PDBx/mmCIF will be the only format allowed for deposition of the atomic coordinates for PDB structures resulting from macromolecular crystallography (MX), including X-ray, neutron, fiber and electron diffraction methods, via OneDep"*

And the ceilings keep falling

12 Dec 2023 — wwPDB: PDB three-character Chemical Component IDs are consumed

Every new ligand gets a 5-character CCD code (e.g. **`A1LU6`**)

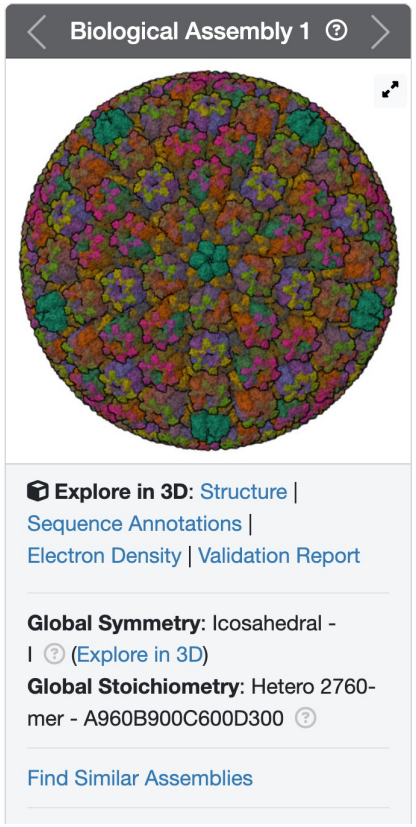
5-char names **do not fit** in the 3-column residue name field — new-ligand entries are mmCIF only

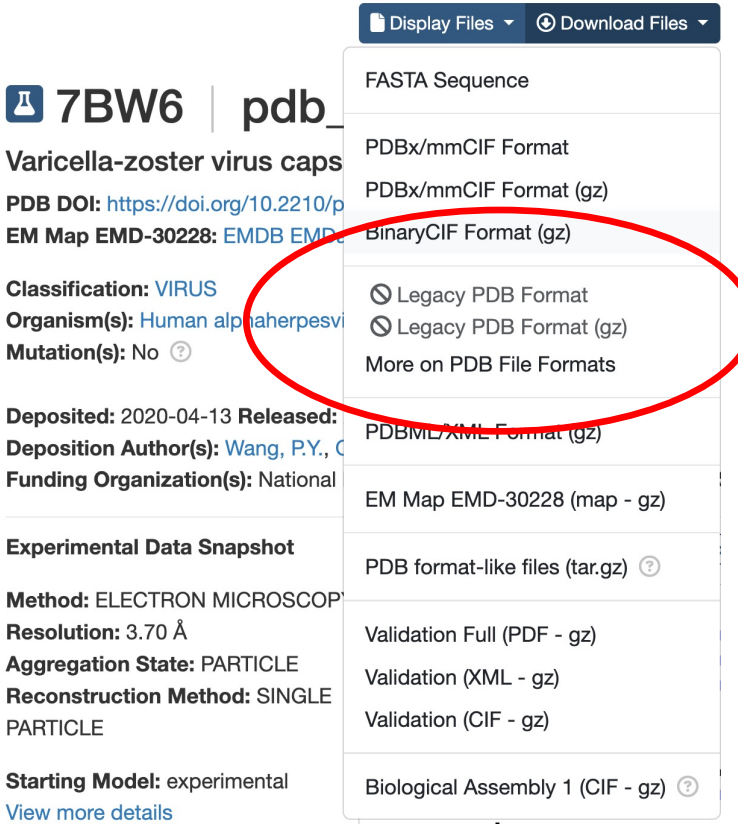
21 July 2027 — 4-char PDB IDs projected to fill before 2028

Replacement: 12-character extended IDs (**`pdb_00001abc`**)

- New entries get extended IDs only, distributed in PDBx/mmCIF and PDBML

April 2025: 3.7% of the PDB archive (>8,000 entries) has no legacy PDB file at all[1]





[1] https://www.rcsb.org/docs/general-help/structures-without-legacy-pdb-format-files

CIF is a family, not a file

Same STAR syntax, completely different content

- **Coordinate mmCIF** — the model (`_atom_site`, `_entity`, `_struct_*`); what replaces .pdb
- **Reflection mmCIF** — `_refln` category; what replaces .mtz for deposition
- **Restraint / monomer CIFs** — CCP4 monomer library, Phenix GeoStd, eLBOW
- **Chemical Component Dictionary (CCD)** — every ligand and residue the PDB knows about
- **Validation mmCIF** — wwPDB validation reports (PDF, XML, PDBx/mmCIF since 2 June 2021)
- **imgCIF / CBF** — raw diffraction images (IUCr standard for detectors)

Phenix is mmCIF-native

- Refinement, validation, map work, ligand fitting — all accept mmCIF directly
- Refinement programs write mmCIF (and PDB when it fits) by default
- You should not have to think about file formats — the tools already did

Questions?

Then let's go hands-on.