ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY ACADEMY

RCSB PDB PROTEIN DATA BANK

Webinar:

Unlock Rapid Analyses Across the Whole PDB Using BinaryCIF

Sebastian Bittrich
Dennis W. Piehl

RCSB PDB
PROTEIN DATA BANK

rcsb.org

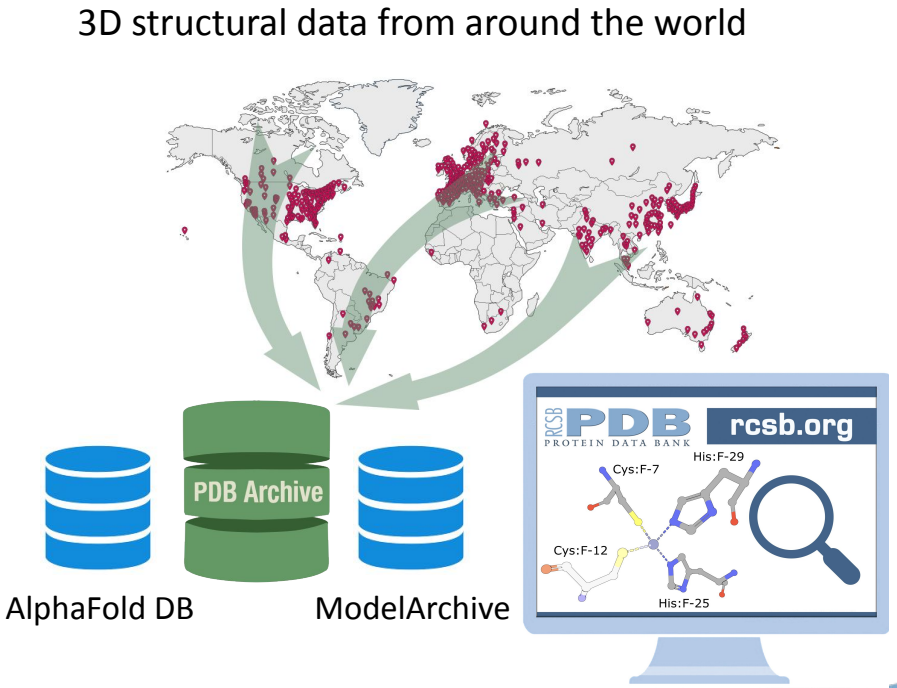
Welcome

Yana Rose, PhD

yana.rose@rcsb.org

RCSB.org: One-Stop-Shop for Public 3D Biostructure Data

- **RCSB.org:** Tools for searching, accessing, visualizing, analyzing, and downloading data
  - Open access to ~226,000 experimental PDB structures of macromolecules
  - >1 million Computed Structure Models (CSMs) predicted using AI/ML methods
- Provenance/reliability of both data types clearly identified



RCSB PDB
PROTEIN DATA BANK

rcsb.org

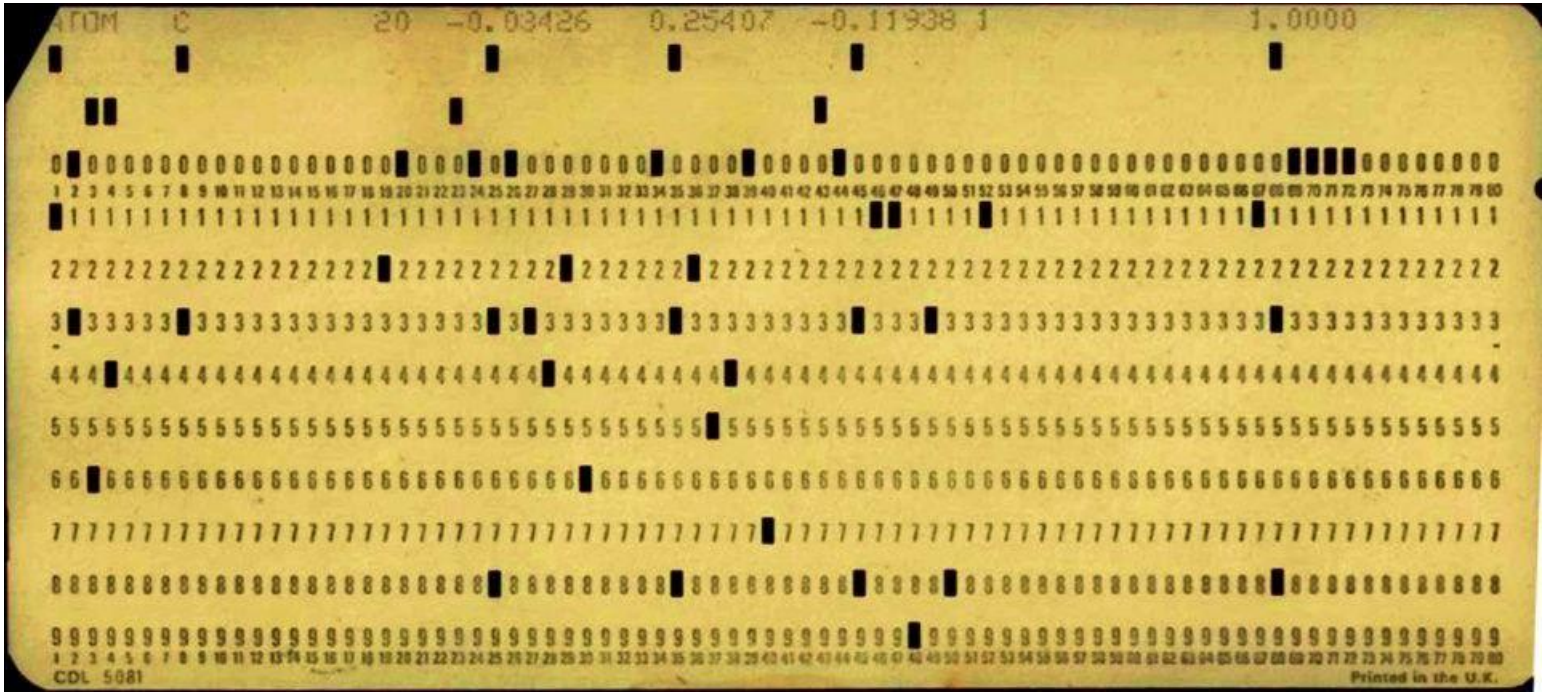
BinaryCIF

Yet *Another* File Format?

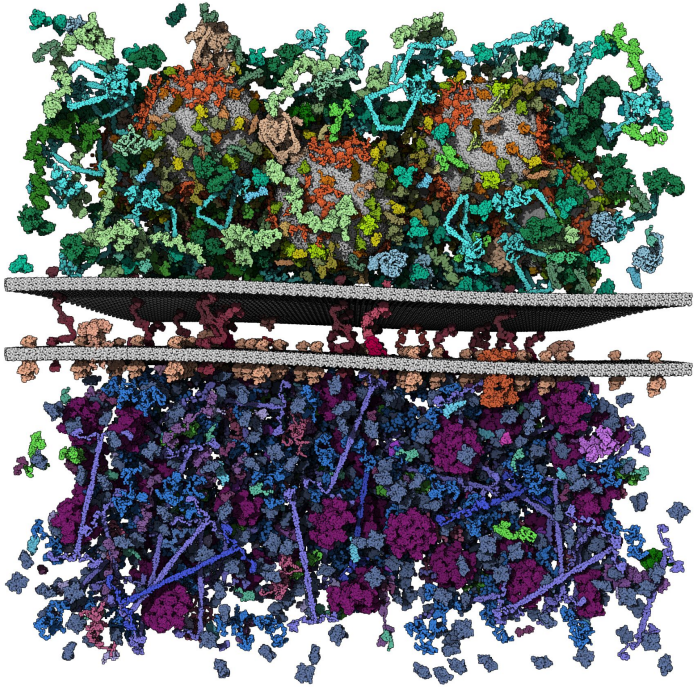
Sebastian Bittrich, PhD

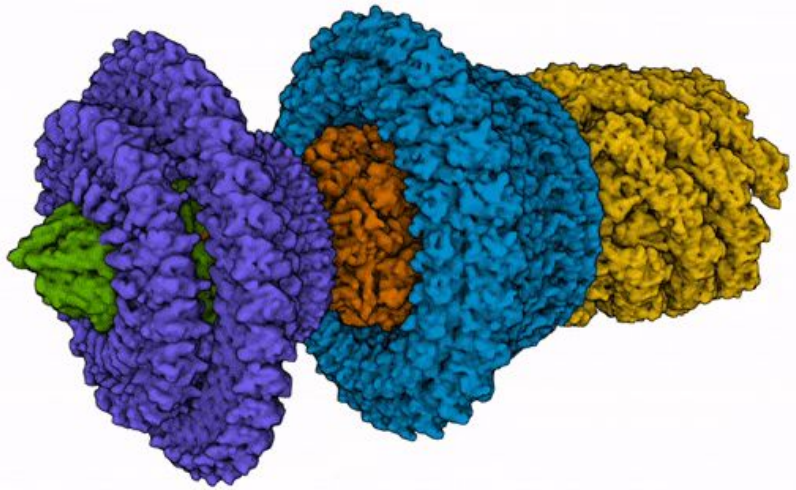
sebastian.bittrich@rcsb.org

How It Started in 1971



Today: The Full Wealth of the PDB in Your Pocket

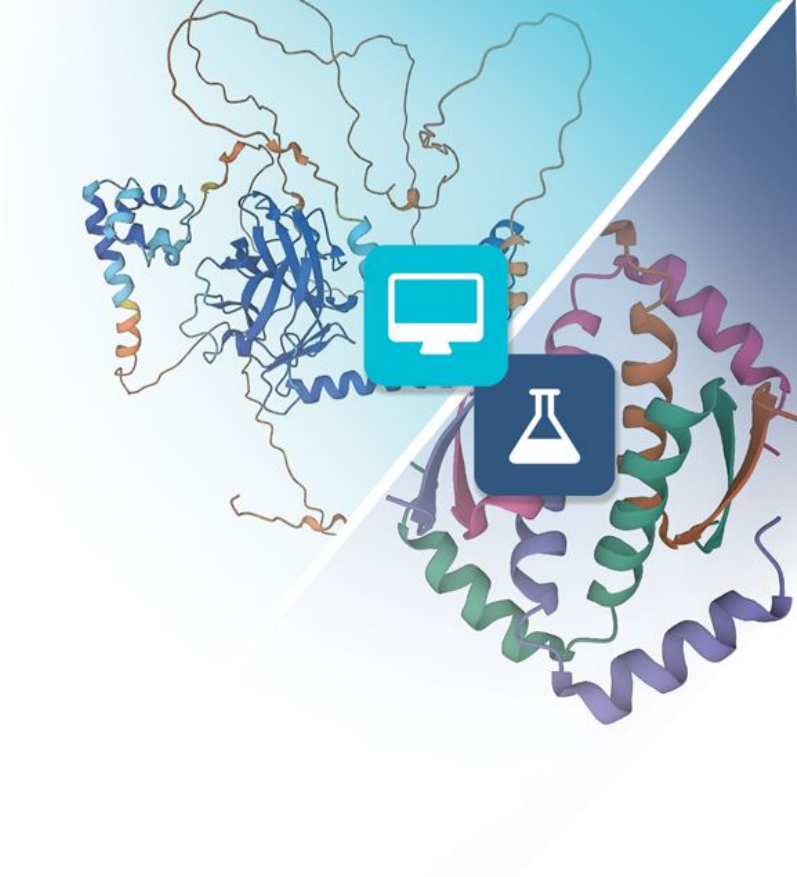




Left: petworld synaptic gap model by da Silva et al. (102.7 million atoms), right: flagellar motor-hook complex by Tan et al. (PDB ID 7cgo)

RCSB PDB
PROTEIN DATA BANK

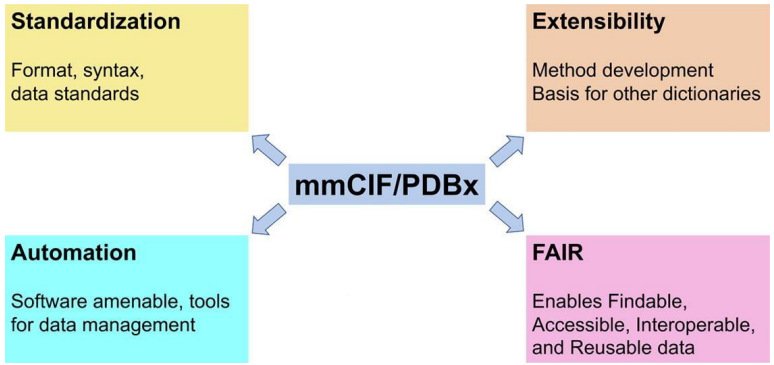
rcsb.org



CIF & Related Schemata

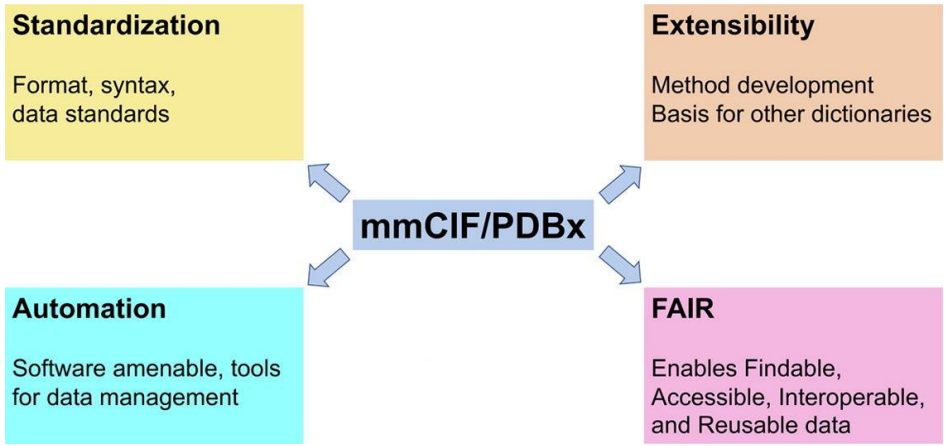
(Macromolecular) Crystallographic Information Files

```
1   interface File {
2       Block getBlock(int blockIndex);
3   }
4
5   interface Block {
6       String getHeader();
7       Category getCategory(String categoryName);
8   }
9
10  interface Category { // e.g. atom_site
11      String getCategoryName();
12      int getRowCount();
13      Column getColumn(String columnName);
14  }
15
16  interface Column { // e.g. atom_site.Cartn_X
17      String getColumnName();
18      int | float | String getValue(int rowIndex);
19  }
```

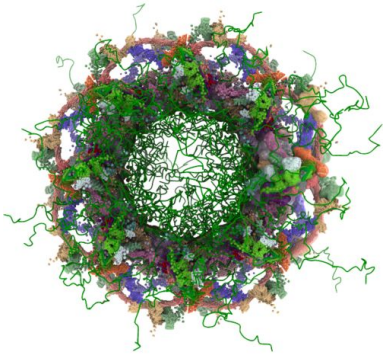


Westbrook et al. (2022). JMB, 10.1016/j.jmb.2022.167599. New PDBx/mmCIF User Guide: mmcif.wwpdb.org/docs/user-guide/guide.html.

Schema Extensions

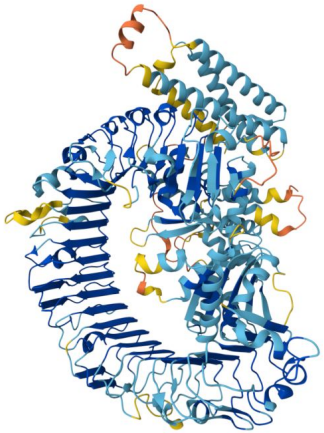


+ **IHMCIF** =



PDB-Dev 8zzc

+ **ModelCIF** =



AlphaFold DB Q8W3K0

Top: nuclear pore complex, bottom: potential plant disease resistance protein.
Vallat et al. (2023). JMB, 10.1016/j.jmb.2023.168021. Vallat et al. (2024). JMB, 10.1016/j.jmb.2024.168546.

RCSB PDB
PROTEIN DATA BANK

rcsb.org

How to Compress (Macromolecular) Data?

Examples of Compression Strategies

Media Type	Lossless Approach	Lossy Approach
Image	PNG	JPEG
Audio	FLAC	MP3
Text	GZIP	Language Models (“AI”)
Macromolecular Data	BinaryCIF	Foldcomp

Foldcomp: Kim et al. (2023). Bioinformatics, 10.1093/bioinformatics/btad153.

What Do We Know About the “atom_site” Category?

Array-Centric Encoding of Atom Site Records

```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
...
_atom_site.pdbx_PDB_model_num
```

ATOM	1	N	N	.	MET	A	1	1	?	43.914	-3.403	8.754	1.00	36.07	?	1	MET	A	N	1
ATOM	2	C	CA	.	MET	A	1	1	?	43.520	-2.018	8.895	1.00	20.51	?	1	MET	A	CA	1
ATOM	3	C	C	.	MET	A	1	1	?	42.097	-2.002	9.425	1.00	21.03	?	1	MET	A	C	1
ATOM	4	O	O	.	MET	A	1	1	?	41.435	-2.993	9.298	1.00	20.03	?	1	MET	A	O	1
ATOM	5	C	CB	.	MET	A	1	1	?	43.693	-1.264	7.552	1.00	17.30	?	1	MET	A	CB	1
ATOM	6	C	CG	.	MET	A	1	1	?	42.903	0.020	7.457	1.00	34.28	?	1	MET	A	CG	1
ATOM	7	S	SD	.	MET	A	1	1	?	43.932	1.487	7.661	1.00	39.71	?	1	MET	A	SD	1
ATOM	8	C	CE	.	MET	A	1	1	?	45.484	0.739	7.149	1.00	37.02	?	1	MET	A	CE	1
ATOM	9	N	N	.	ASN	A	1	2	?	41.661	-0.889	10.039	1.00	16.25	?	2	ASN	A	N	1
ATOM	10	C	CA	.	ASN	A	1	2	?	40.332	-0.753	10.606	1.00	11.56	?	2	ASN	A	CA	1
ATOM	11	C	C	.	ASN	A	1	2	?	39.944	0.729	10.691	1.00	17.42	?	2	ASN	A	C	1
ATOM	12	O	O	.	ASN	A	1	2	?	40.751	1.572	10.393	1.00	11.73	?	2	ASN	A	O	1
ATOM	13	C	CB	.	ASN	A	1	2	?	40.208	-1.491	11.989	1.00	15.25	?	2	ASN	A	CB	1
ATOM	14	C	CG	.	ASN	A	1	2	?	41.128	-0.942	13.041	1.00	16.24	?	2	ASN	A	CG	1
ATOM	15	O	OD1	.	ASN	A	1	2	?	41.131	0.268	13.300	1.00	18.64	?	2	ASN	A	OD1	1
ATOM	16	N	ND2	.	ASN	A	1	2	?	41.929	-1.811	13.632	1.00	14.71	?	2	ASN	A	ND2	1
ATOM	17	N	N	.	ILE	A	1	3	?	38.721	1.023	11.112	1.00	11.52	?	3	ILE	A	N	1
ATOM	18	C	CA	.	ILE	A	1	3	?	38.177	2.386	11.200	1.00	17.48	?	3	ILE	A	CA	1

Case 1: Constant Data

Repeat some *value* *n*-times.

Example: **Run Length encoding** (value = “ATOM”, repeats = 18)

Case 2: Slowly Changing Data

Store *delta* with respect to previous value.

Example: **Delta encoding** (start = 1, delta = 1)

Case 3: Recurring Text Data

Dictionary containing each unique string once.

Example: **String Array encoding** { “N”: 0, “C”: 1, “O”: 2, “S”: 3 }

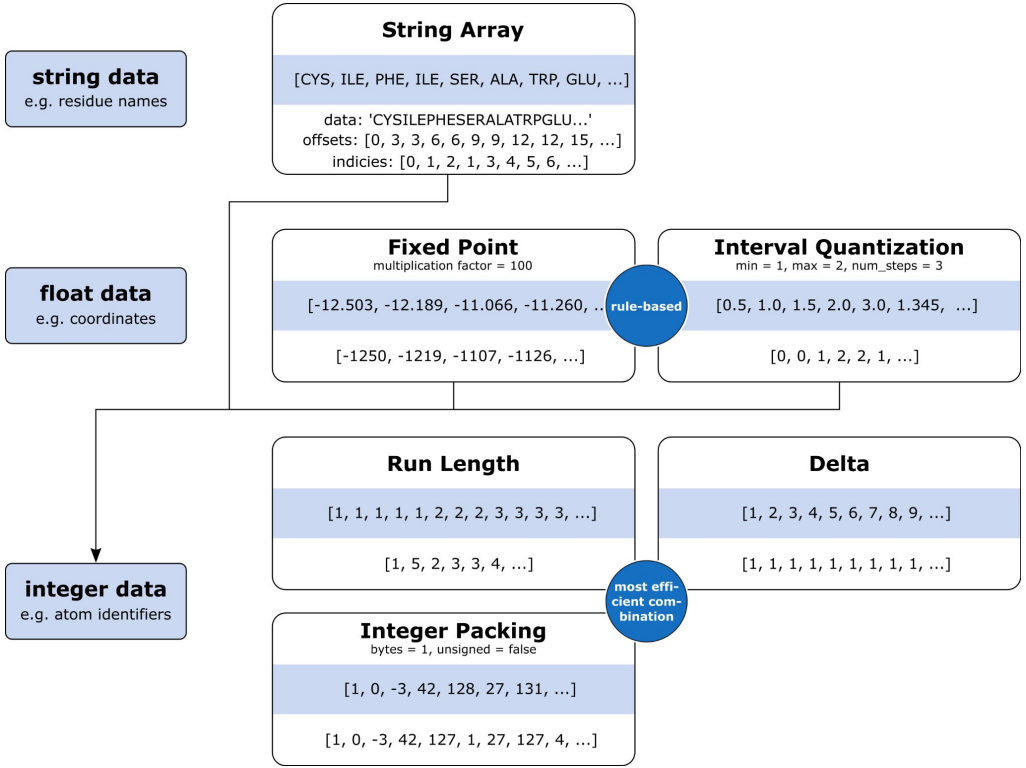
Case 4: Floating Point Data

Delta encoding (positional changes are small).

Example: **Fixed Point encoding** (factor = 1000)

Illustrative example, details in: Sehnal et al. (2020). PLoS CB, 10.1371/journal.pcbi.1008247.

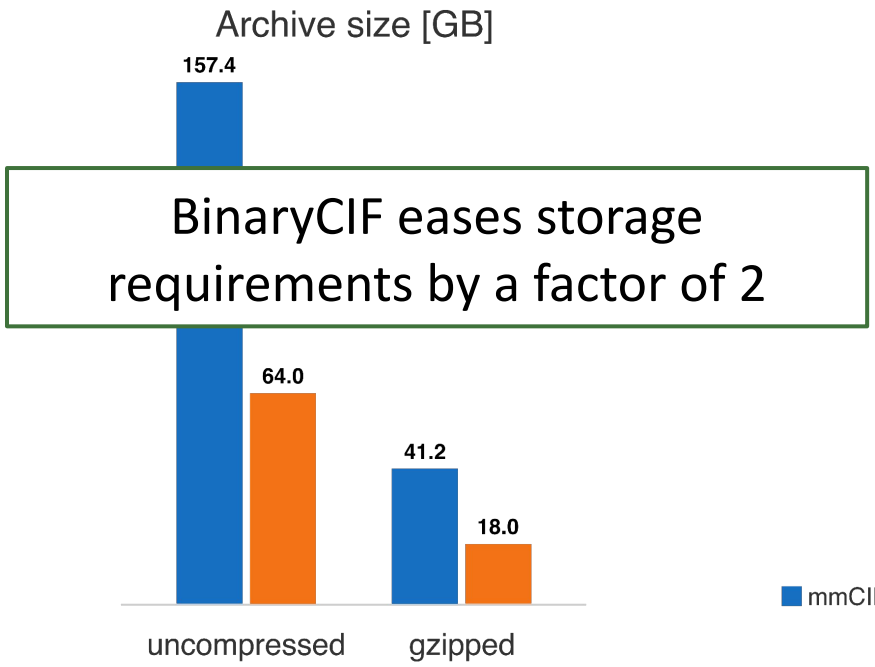
BinaryCIF Compression Strategies



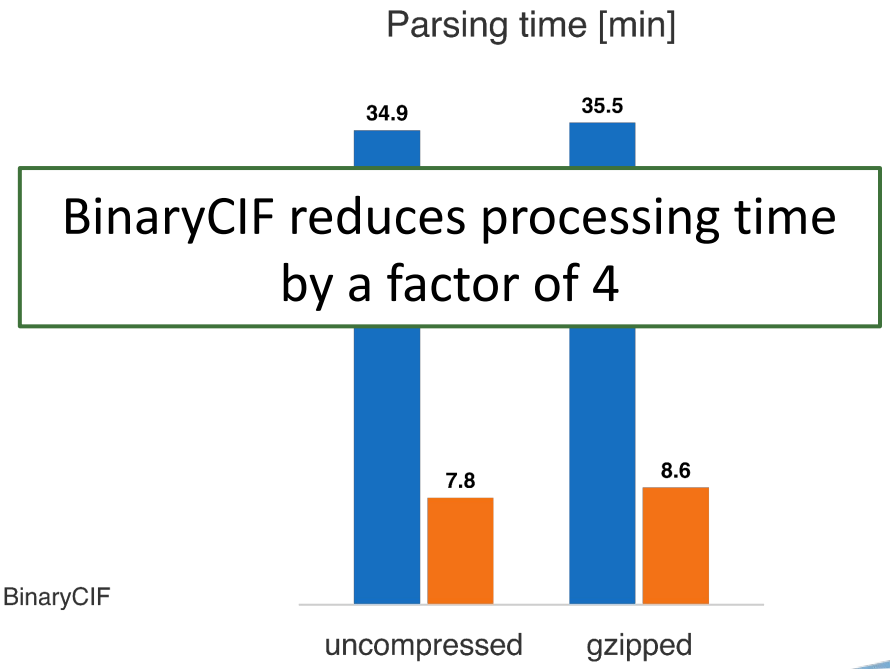
Sehnal et al. (2020). PLoS CB, 10.1371/journal.pcbi.1008247.

BinaryCIF automatically selects suitable encoding strategies

Performance: PDBx/mmCIF vs. BinaryCIF



BinaryCIF eases storage requirements by a factor of 2



BinaryCIF reduces processing time by a factor of 4

mmCIF BinaryCIF

Sehnal et al. (2020). PLoS CB, 10.1371/journal.pcbi.1008247.

Key Advantages of BinaryCIF

- **Interchangeable with PDBx/mmCIF**
  - Develop and debug using human-readable PDBx/mmCIF
  - Switch to BinaryCIF for **better performance** when ready
- **Flexible, schema-independent**
  - Customizable by omitting data that's irrelevant to your use case
  - Add custom data easily—no need for workarounds (*i.e.*, no more misusing the B-factors field to store data)
- **Resource Efficiency**
  - Save on storage and bandwidth
  - Reduce compute overhead for faster data processing

BinaryCIF Ecosystem

RCSB PDB PROTEIN DATA BANK

PDBe Protein Data Bank in Europe

AlphaFold DB

Languages	TypeScript/JavaScript	Python	Java
Libraries	Mol*	py-mmcif py-rcsb_utils_io biotite	ciftools-java BioJava

Resources & Publications

- Access the PDB archive as BinaryCIF: models.rcsb.org/4hhb.bcif.gz
- ModelServer API (manipulate data and select *e.g.* an individual chain): models.rcsb.org/v1/4hhb/atoms?label_asym_id=C

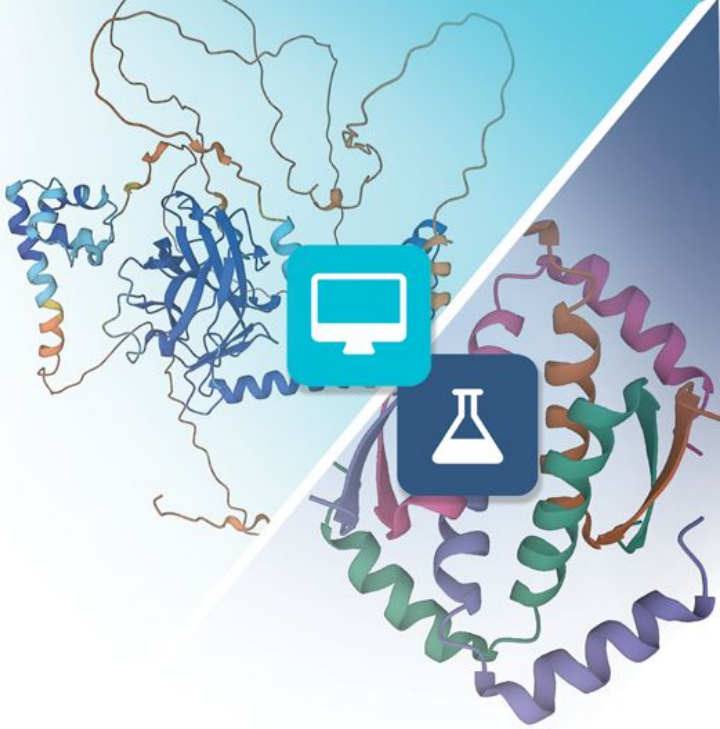
Discussed software packages

- github.com/rcsb/py-mmcif (Python)
- github.com/rcsb/py-rcsb_utils_io (high-level Python wrapper)
- github.com/rcsb/ciftools-java (Java)

Sehnal et al. (2020). BinaryCIF and CIFTools—Lightweight, efficient and extensible macromolecular data management. PLoS CB, 16(10), e1008247, doi: 10.1371/journal.pcbi.1008247.

RCSB PDB
PROTEIN DATA BANK

rcsb.org



Compute Archive-Wide Statistics in <10 Minutes

Computing Archive-Wide Statistics

Query — Fetch — Parse — Study

1. Collect All 225,000 PDB Entry Identifiers

Retrieve the set of all archive entries from the RCSB Search API.

2. Access Structure Data as BinaryCIF

For each entry: Transfer its source file to your machine.

BinaryCIF saves time and bandwidth.

3. Access Information for Each Entry

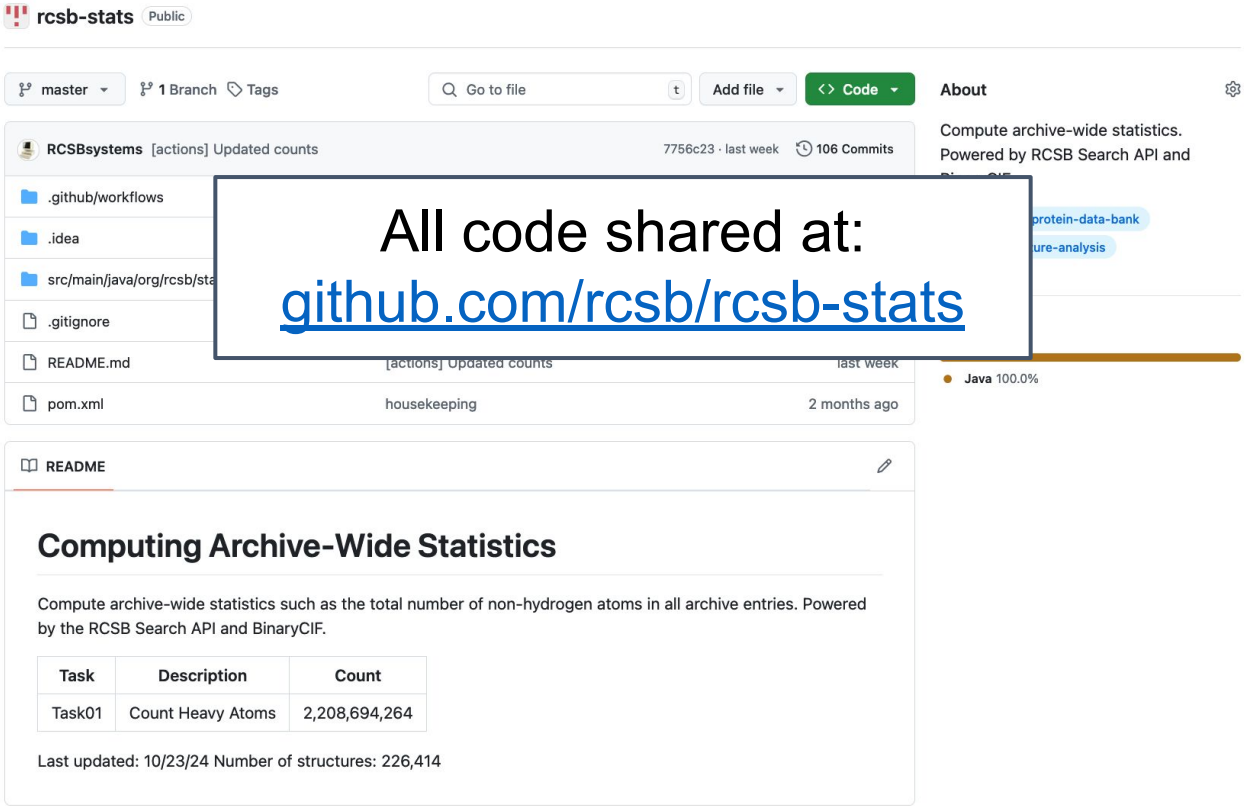
Navigate to data of interest using the typed mmCIF schema.

BinaryCIF saves time and compute resources.

4. Perform Your Analysis

Process data how you see fit. We will aggregate the number of non-hydrogen atoms in all entries.

Live demo



All code shared at:
github.com/rcsb/rcsb-stats

Key functionality: Reading from URL – Accessing data by category, column, and row – Recent GitHub workflow run

RCSB PDB
PROTEIN DATA BANK

rcsb.org

Python

Dennis W. Piehl, PhD

dennis.piehl@rcsb.org

Working with mmCIF and BCIF in Python

RCSB PDB Python Packages

- **py-mmcif:** contains the core code for working with CIF data
  - https://github.com/rcsb/py-mmcif
- **rcsb.utils.io:** simple wrapper for py-mmcif, with additional tooling
  - https://github.com/rcsb/py-rcsb_utils_io

*Be sure to follow along these repositories for future updates and enhancements!

Working with mmCIF and BCIF in Python

Quick review of PDBx/mmCIF structure and terminology:

- **Data Container:** entire structure file data
- **Dictionary:** Definitions of categories and attributes (mmcif.wwpdb.org)

Category: `database_2`

Attribute: `database_id`

```
data_4HHB

_entry.id  4HHB
#
_audit_conform.dict_name       mmcif_pdbx.dic
_audit_conform.dict_version    5.392
_audit_conform.dict_location   http://mmcif.pdb.org/dictionaries/ascii/mmcif_pdbx.dic
#
loop_
_database_2.database_id
_database_2.database_code
_database_2.pdbx_database_accession
_database_2.pdbx_DOI
PDB     4HHB          pdb_00004hhb  10.2210/pdb4hhb/pdb
WWPDB   D_1000179340  ?             ?
#
```

Live demo

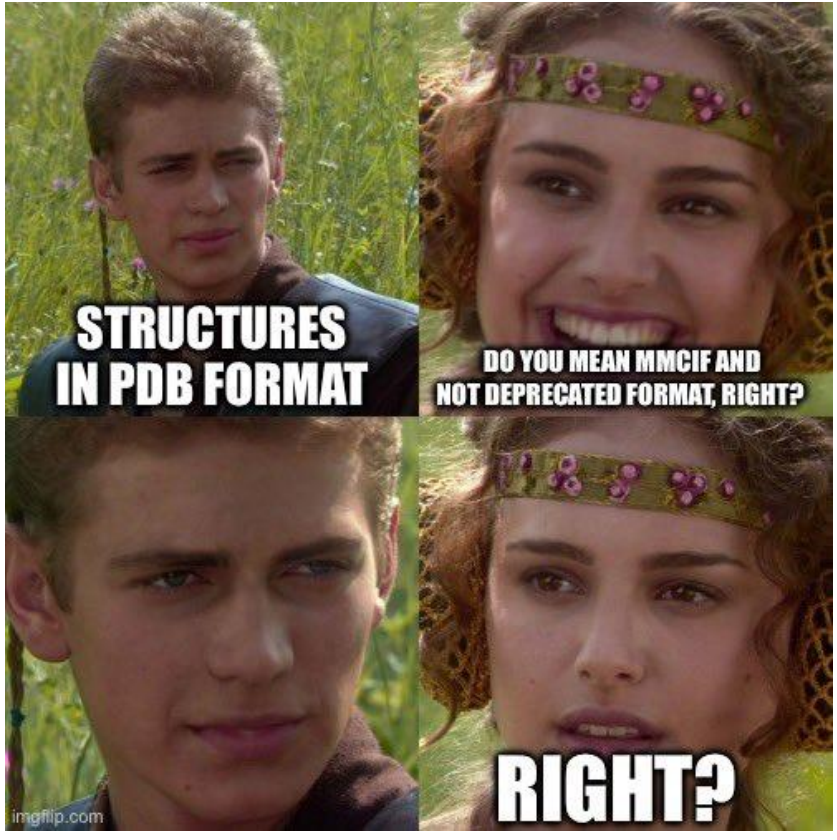
https://github.com/rcsb/rcsb-training-resources/blob/master/training-events/2024/utilizing-binary-cif

Short URL: https://t.ly/XCKbp

Open notebook on GitHub, then click "*Open in Colab*" badge at top of page (to open in Google Colab)

PDB IDs are Changing (pdb_00001abc)

- wwPDB anticipates four-character PDB accession codes (PDB IDs) will be consumed by 2028
- Entries issued with extended PDB IDs are not compatible with the legacy PDB file format
- PDB users (including scientific journals) should transition to PDBx/mmCIF format and new PDB ID format as early as possible



RCSB PDB Team



RCSB PDB PROTEIN DATA BANK

RCSB.ORG

info@rcsb.org

Core Operations Funding

US National Science Foundation (DBI-2321666),
National Institute of General Medical Sciences,
National Institute of Allergy and Infectious Disease,
and
National Cancer Institute (NIH R01GM157729), and
the US Department of Energy (DE-SC0019749)

Management

RUTGERS THE STATE UNIVERSITY OF NEW JERSEY

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

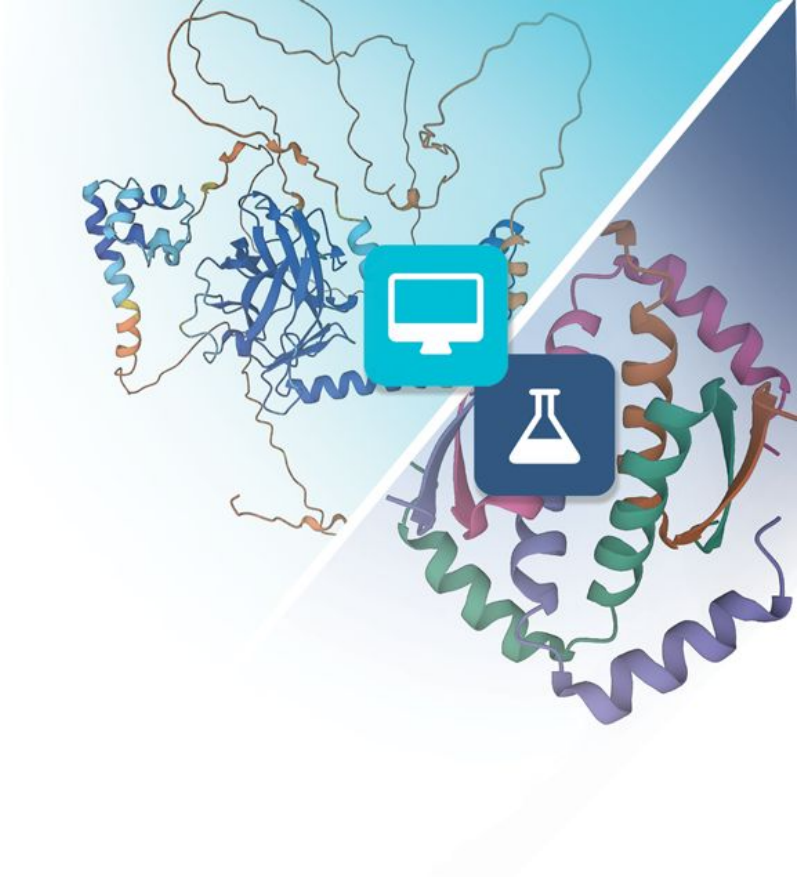
David Sehnal

Alexander Rose

John D. Westbrook

RCSB PDB
PROTEIN DATA BANK

rcsb.org



Questions?