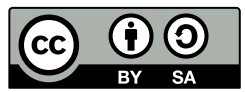
Sources of data in experimental structural biology: CryoEM, X-ray crystallography

Saulius Gražulis

Vilnius, 2024

Vilnius University Institute of Biotechnology

Id: slides.tex 11011 2024-01-03 08:01:39Z saulius January 3, 2024



Why databases?

How many crystal structures have been published each year? Search the COD database:

Why databases?

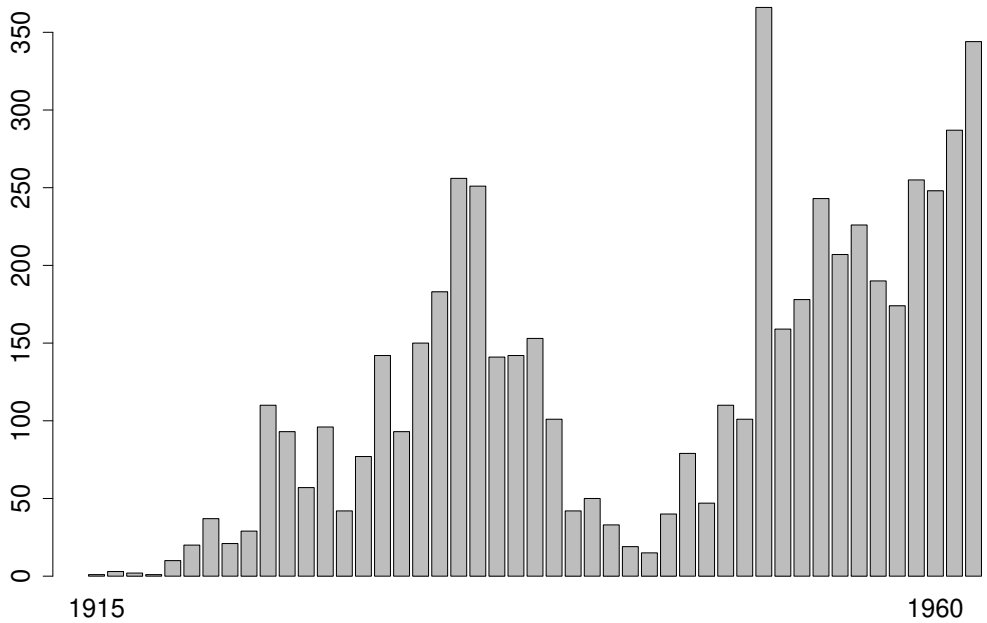
How many crystal structures have been published each year? Search the COD database:

```
SELECT count(*) AS nr, year FROM data
WHERE year IS NOT NULL AND
GROUP BY year ORDER BY year DESC
```

Why databases?

How many crystal structures have been published each year? Search the COD database:

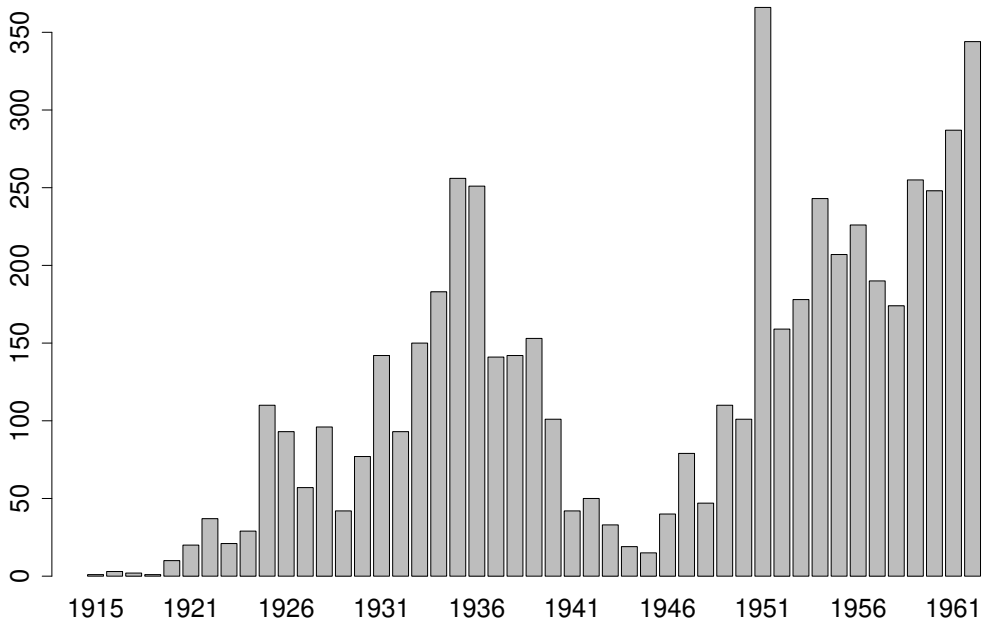
```
SELECT count(*) AS nr, year FROM data
WHERE year IS NOT NULL AND
GROUP BY year ORDER BY year DESC
```



Why databases?

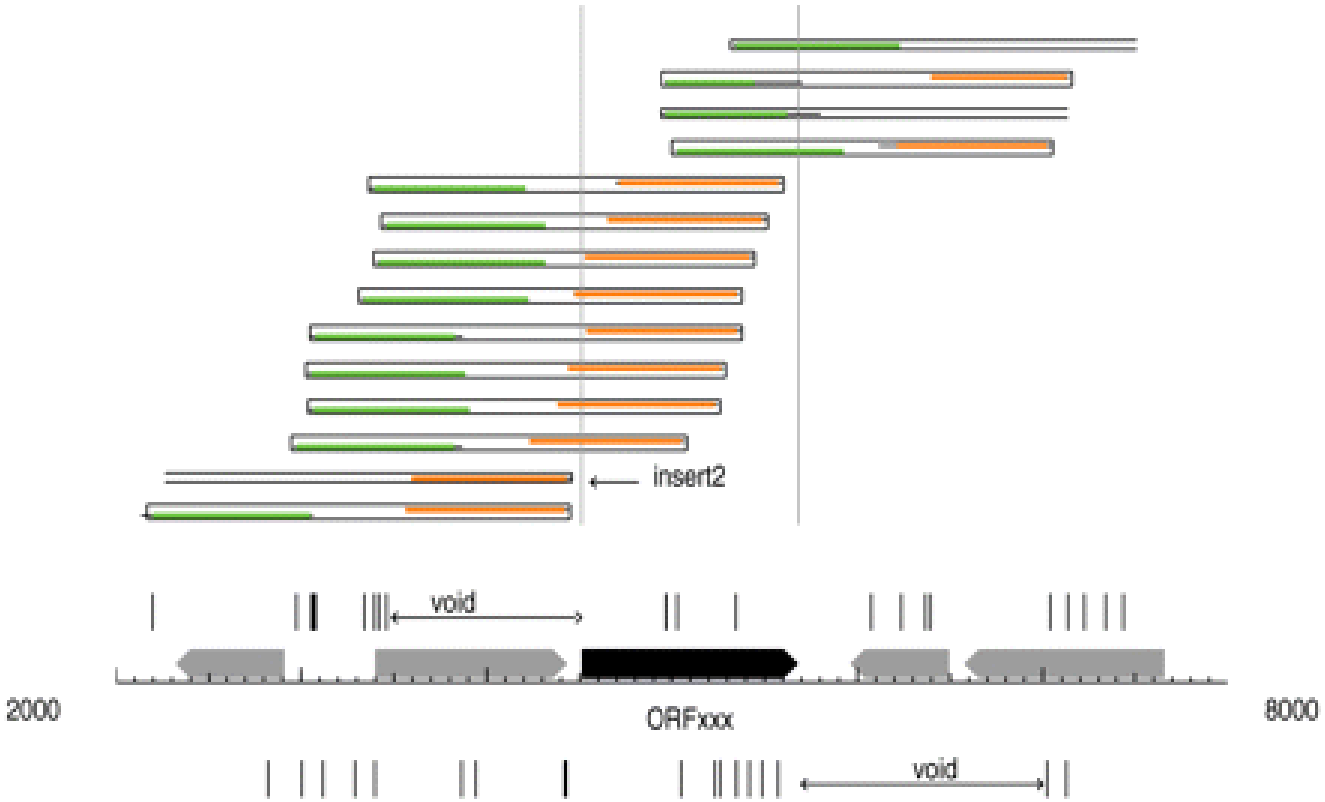
How many crystal structures have been published each year? Search the COD database:

```
SELECT count(*) AS nr, year FROM data
WHERE year IS NOT NULL AND
GROUP BY year ORDER BY year DESC
```



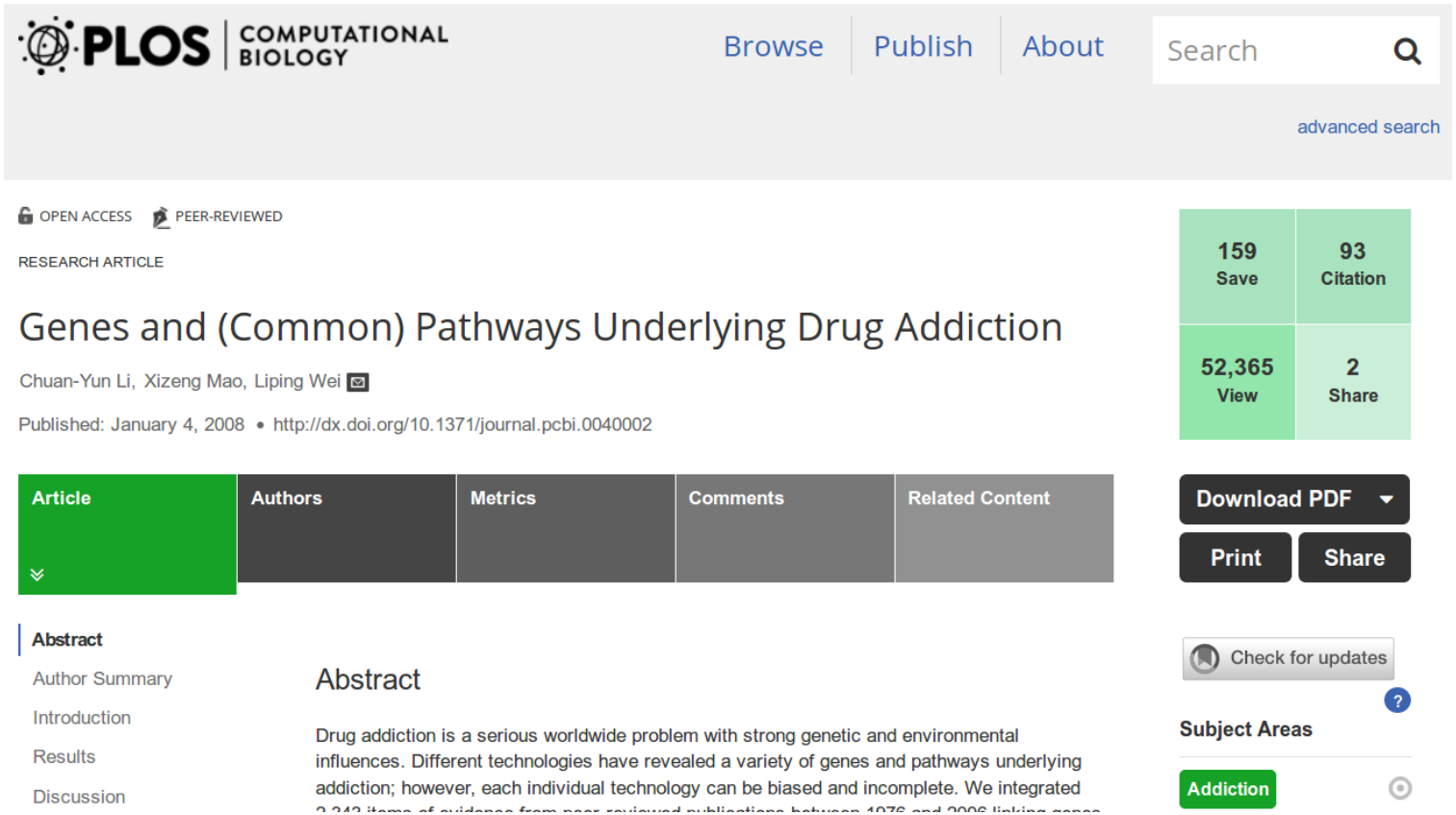
Discoveries in „raw“ data

Zheng from the NEB team lead by Roberts used the raw sequencing data to discover *active* restriction endonucleases (Zheng et al. 2008):



Importance of data

A group from China made a discovery biochemical pathways for drug addiction without even making and experiment (Li et al. 2008):



http://slidegur.com/doc/3077570/introducing-bioinformatics-databases

Amounts of data

High Energy Physics

- LHC generates 1 Terabyte per second.
  http://blogs.discovermagazine.com/cosmicvariance/2006/09/27/lhc-factoids/
  2009-12-07
- „When it starts in 2007 the LHC will ... produce 15 million Gigabytes of data a year“ (15 PB/year $\approx 15 \times 10^{15}$ bytes/year – S.G.)
  http://www.physorg.com/news10895.html
  2009-12-07

Amounts of data

Sequences

- „Release 57.11 of 24-Nov-09 of UniProtKB/Swiss-Prot contains 512994 sequence entries, comprising 180531504 amino acids abstracted from 184920 references.“ (1.81×10^8 a.r., 5.13×10^5 sekų – S.G.)

  http://www.expasy.ch/sprot/relnotes/relstat.html
  2009-12-07

- „As of Tuesday Dec 01, 2009 at 4 PM PST there are 61808 Structures“

  http://www.rcsb.org/pdb/statistics/holdings.do
  2009-12-07

Amounts of data

Crystallography

- „Cambridge Structural Database 1 January 2010 Total No. of structures 501857“
  http://www.ccdc.cam.ac.uk/products/csd/statistics/
  2010-12-20
- „Currently there are 126492 entries in the COD“
  http://www.crystallography.net/
  2010-12-20
- „Currently there are 509531 entry in the COD“
  http://www.crystallography.net/
  2024-01-03

Scientific data sources

1. Generic data archives (Data Dryad, FigShare, MIDAS, Zenodo ...);
   -
   -
2.
   -
   -
   -
   -

Scientific data sources

1. Generic data archives (Data Dryad, FigShare, MIDAS, Zenodo ...);
   - Zenodo https://doi.org/10.5281/zenodo.3560693
   - Zenodo https://zenodo.org/record/3841841
2.
   -
   -
   -
   -

Scientific data sources

1. Generic data archives (Data Dryad, FigShare, MIDAS, Zenodo ...);
   - Zenodo https://doi.org/10.5281/zenodo.3560693 – poster (PDF)...
   - Zenodo https://zenodo.org/record/3841841
2.
   -
   -
   -
   -

Scientific data sources

1. Generic data archives (Data Dryad, FigShare, MIDAS, Zenodo ...);
   - Zenodo https://doi.org/10.5281/zenodo.3560693 – poster (PDF)...
   - Zenodo https://zenodo.org/record/3841841 – linked COVID-19 data (Turtle .ttl)...
2.
   -
   -
   -
   -

Scientific data sources

1. Generic data archives (Data Dryad, FigShare, MIDAS, Zenodo ...);
   - Zenodo https://doi.org/10.5281/zenodo.3560693 – poster (PDF)...
   - Zenodo https://zenodo.org/record/3841841 – linked COVID-19 data (Turtle .ttl)...
2. Specialised data archives (PDB, COD, NCBI, SwissProt, EuropePMC, PubMed (!));
   -
   -
   -
   -

Scientific data sources

1. Generic data archives (Data Dryad, FigShare, MIDAS, Zenodo ...);
   - Zenodo https://doi.org/10.5281/zenodo.3560693 – poster (PDF)...
   - Zenodo https://zenodo.org/record/3841841 – linked COVID-19 data (Turtle .ttl)...
2. Specialised data archives (PDB, COD, NCBI, SwissProt, EuropePMC, PubMed (!));
   - https://www.crystallography.net/cod/1557684.cif
   - https://www.crystallography.net/cod/1544162.html
   - https://www.pdb.org/pdb/files/1KNV.cif
   - https://www.rcsb.org/structure/2IXS

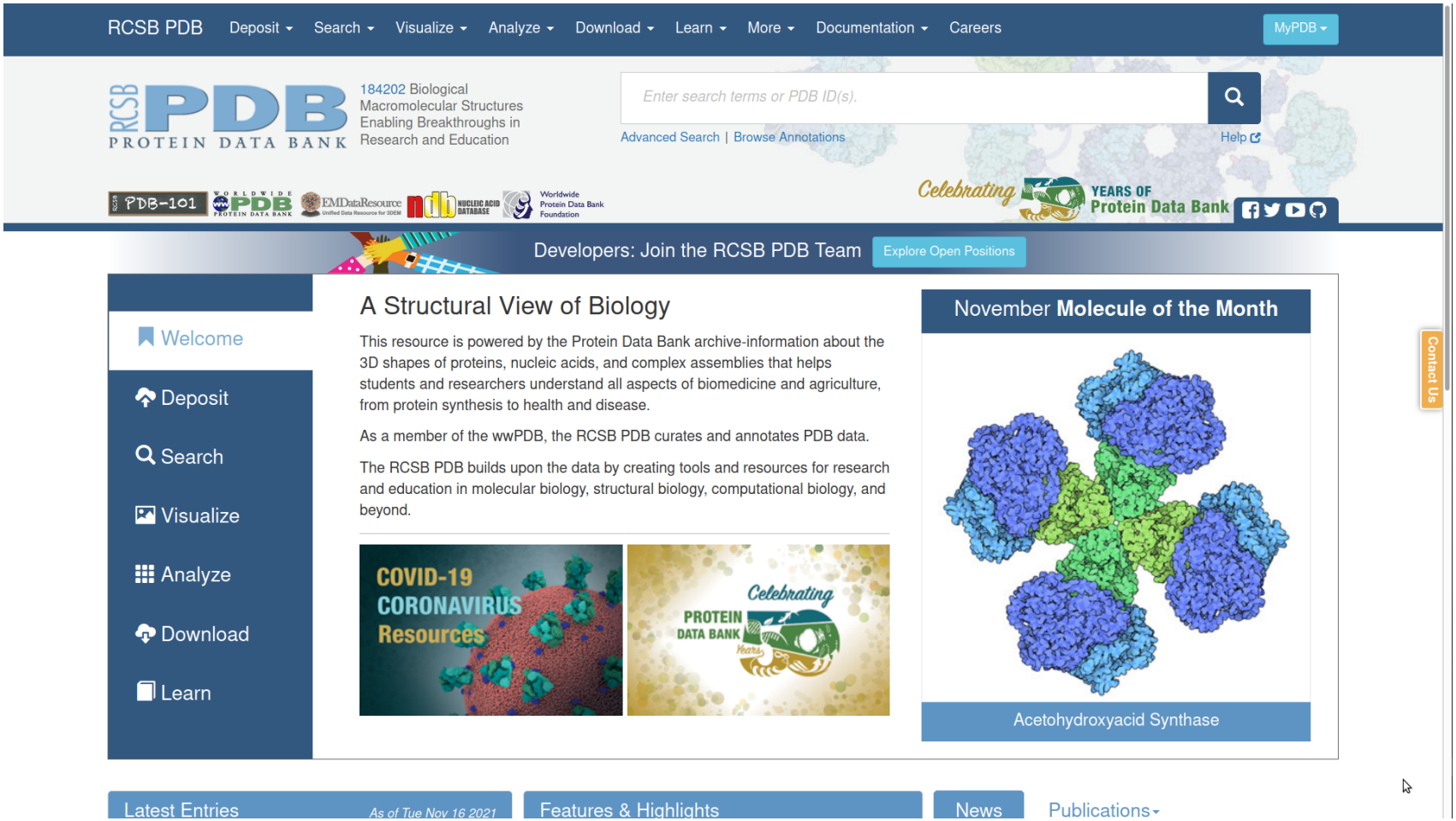
The Protein Data Bank

Three major repositories in different continents, governed by the wwPDB consortium:

https://www.wwpdb.org/

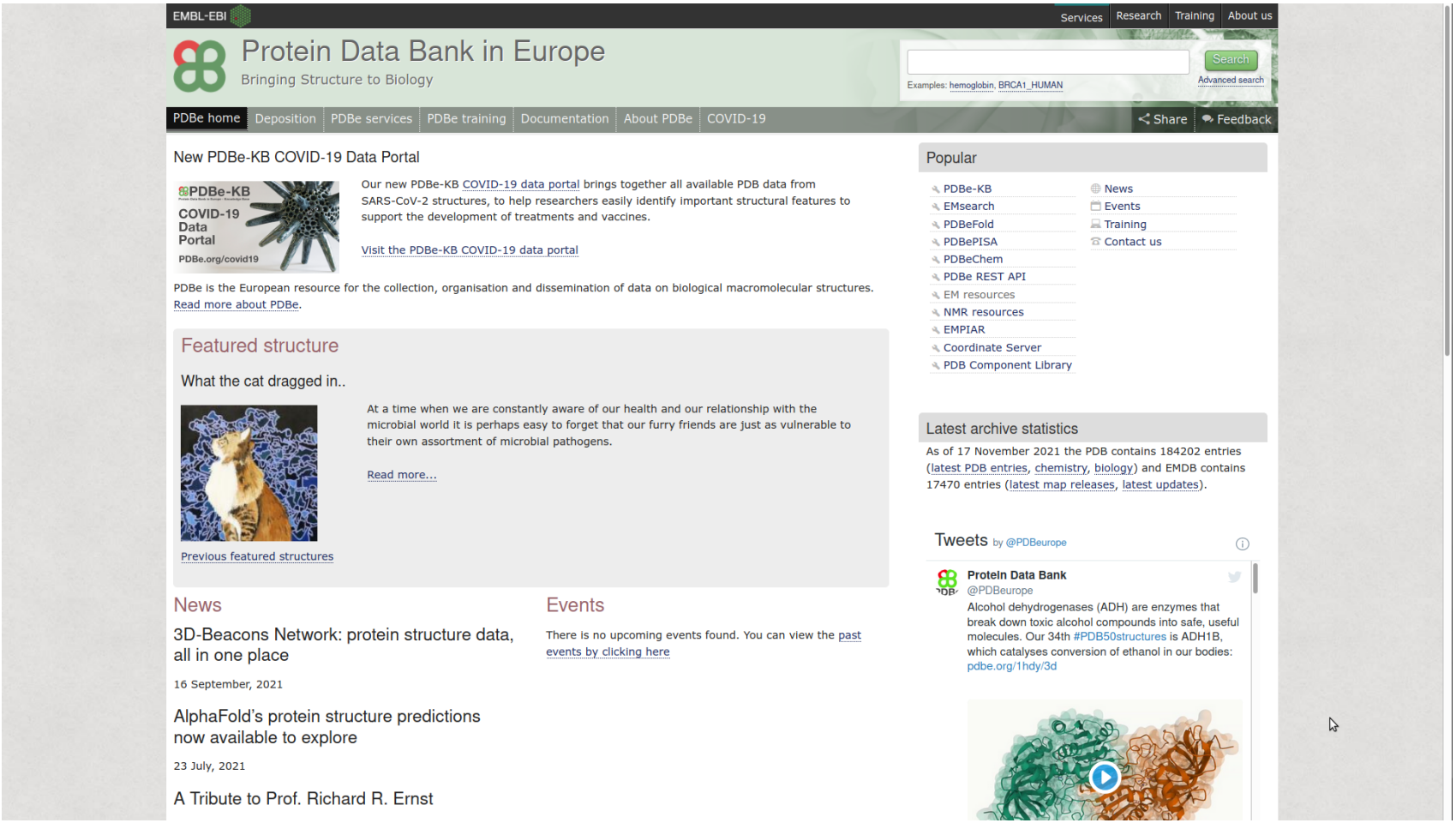
The Protein Data Bank: RCSD PDB

https://www.rcsb.org/



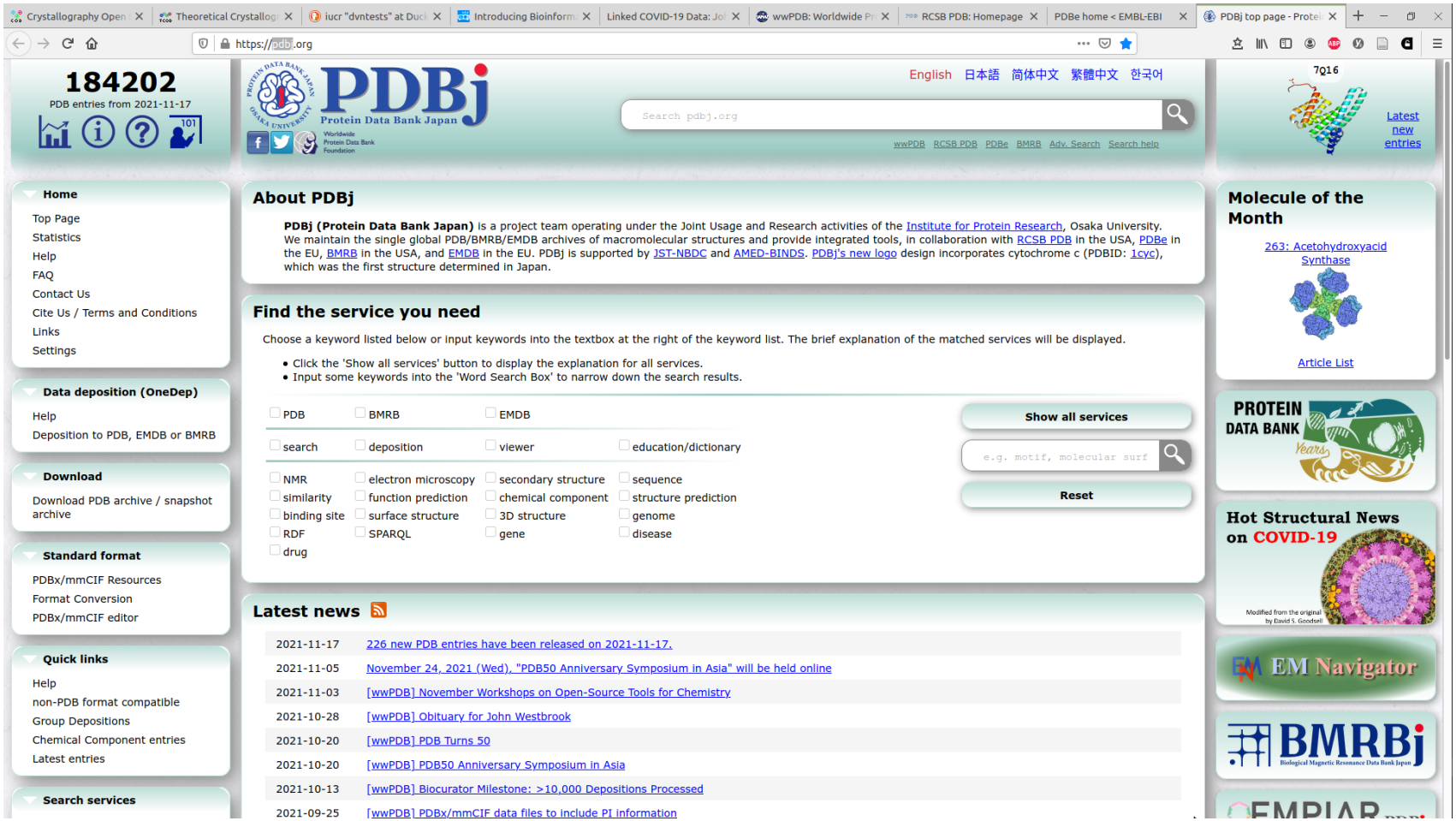
The Protein Data Bank: PDBe

https://www.ebi.ac.uk/pdbe/



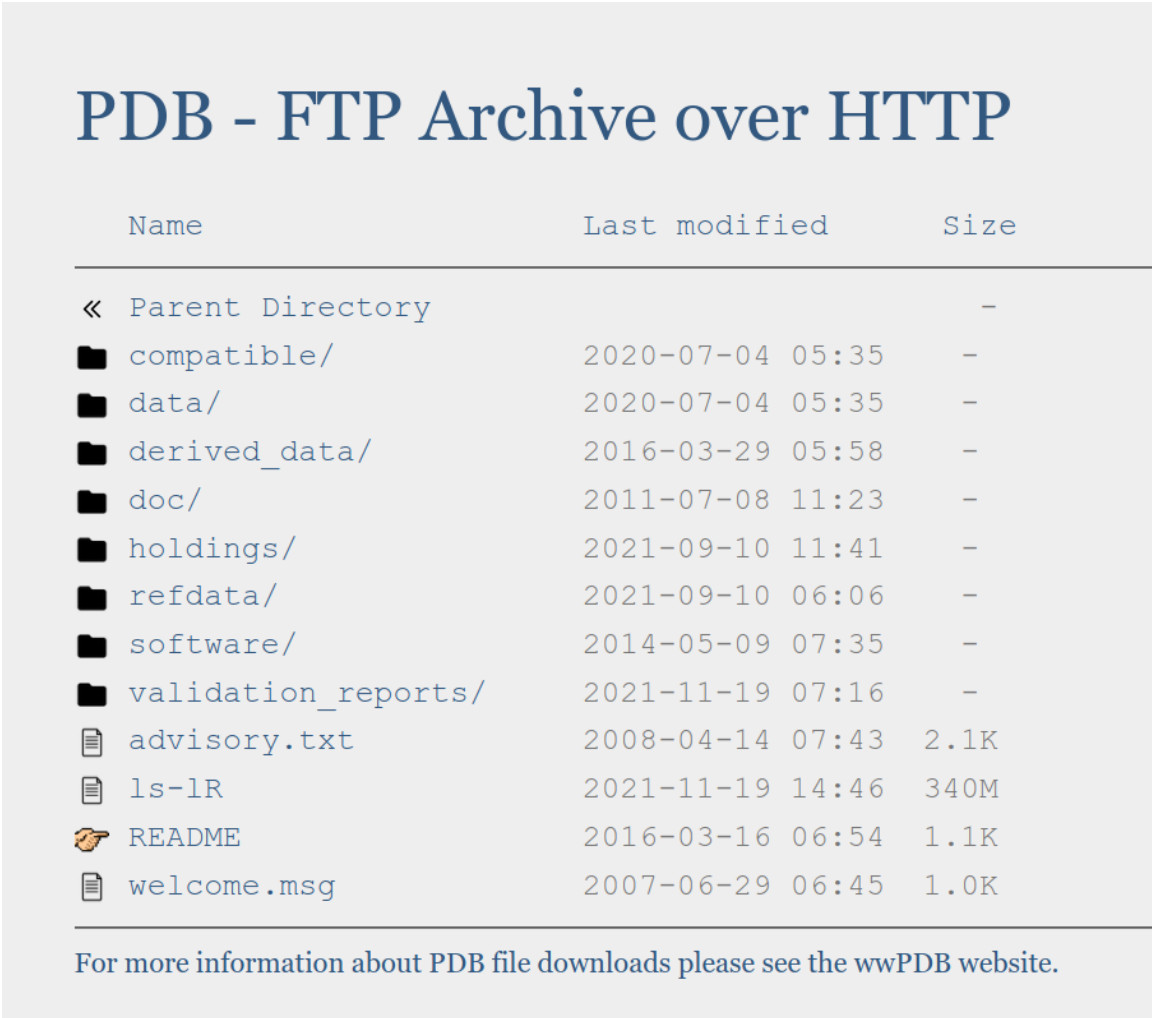
The Protein Data Bank: PDBj

https://pdbj.org/



The Protein Data Bank: the FTP site

http://ftp.wwpdb.org/pub/pdb/[accessed: 2021]
https://files.wwpdb.org/pub/pdb/[accessed: 2024-01-03T07:56+00:00]



PDB - FTP Archive over HTTP

Name	Last modified	Size
« Parent Directory		-
compatible/	2020-07-04 05:35	-
data/	2020-07-04 05:35	-
derived_data/	2016-03-29 05:58	-
doc/	2011-07-08 11:23	-
holdings/	2021-09-10 11:41	-
refdata/	2021-09-10 06:06	-
software/	2014-05-09 07:35	-
validation_reports/	2021-11-19 07:16	-
advisory.txt	2008-04-14 07:43	2.1K
ls-lR	2021-11-19 14:46	340M
README	2016-03-16 06:54	1.1K
welcome.msg	2007-06-29 06:45	1.0K

For more information about PDB file downloads please see the wwPDB website.

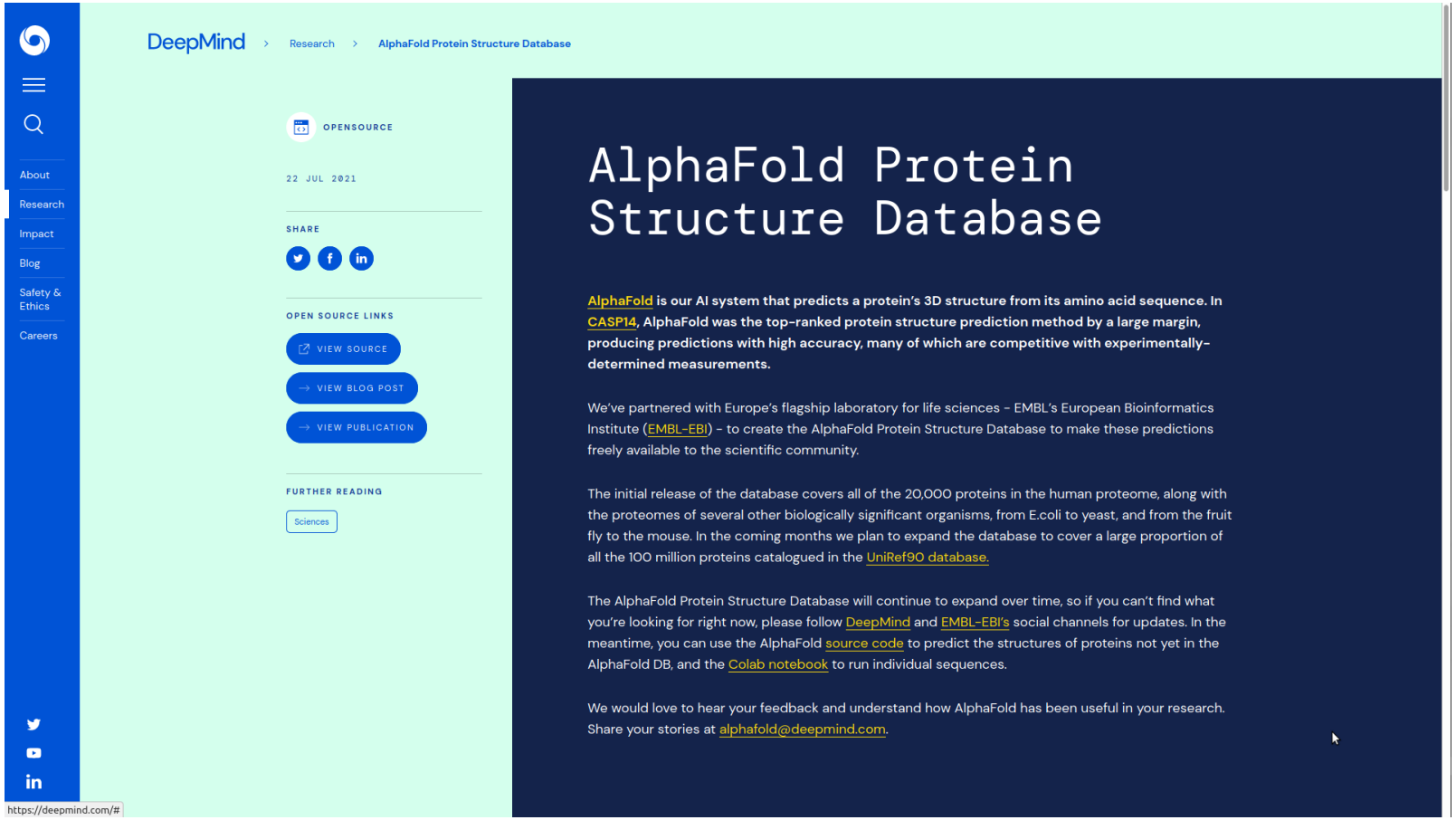
PDB contents

PDB contains macromolecular structures solved by:

- X-ray crystallography
- CryoEM
- NMR

The AlphaFold prediction database

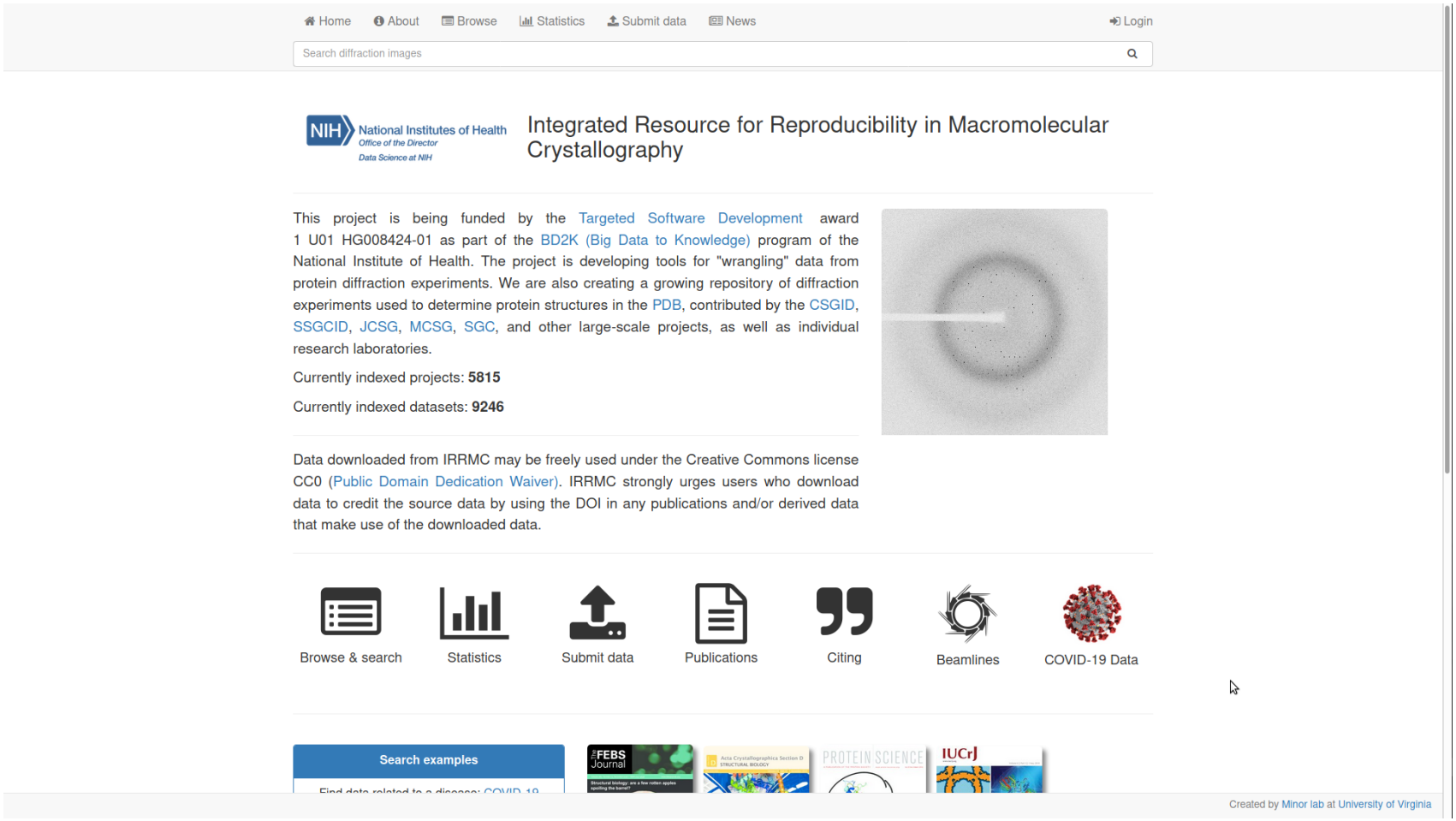
https://deepmind.com/research/open-source/alphafold-protein-structure-database



"Our models are trained on structures extracted from the PDB" (Senior et al. 2020).

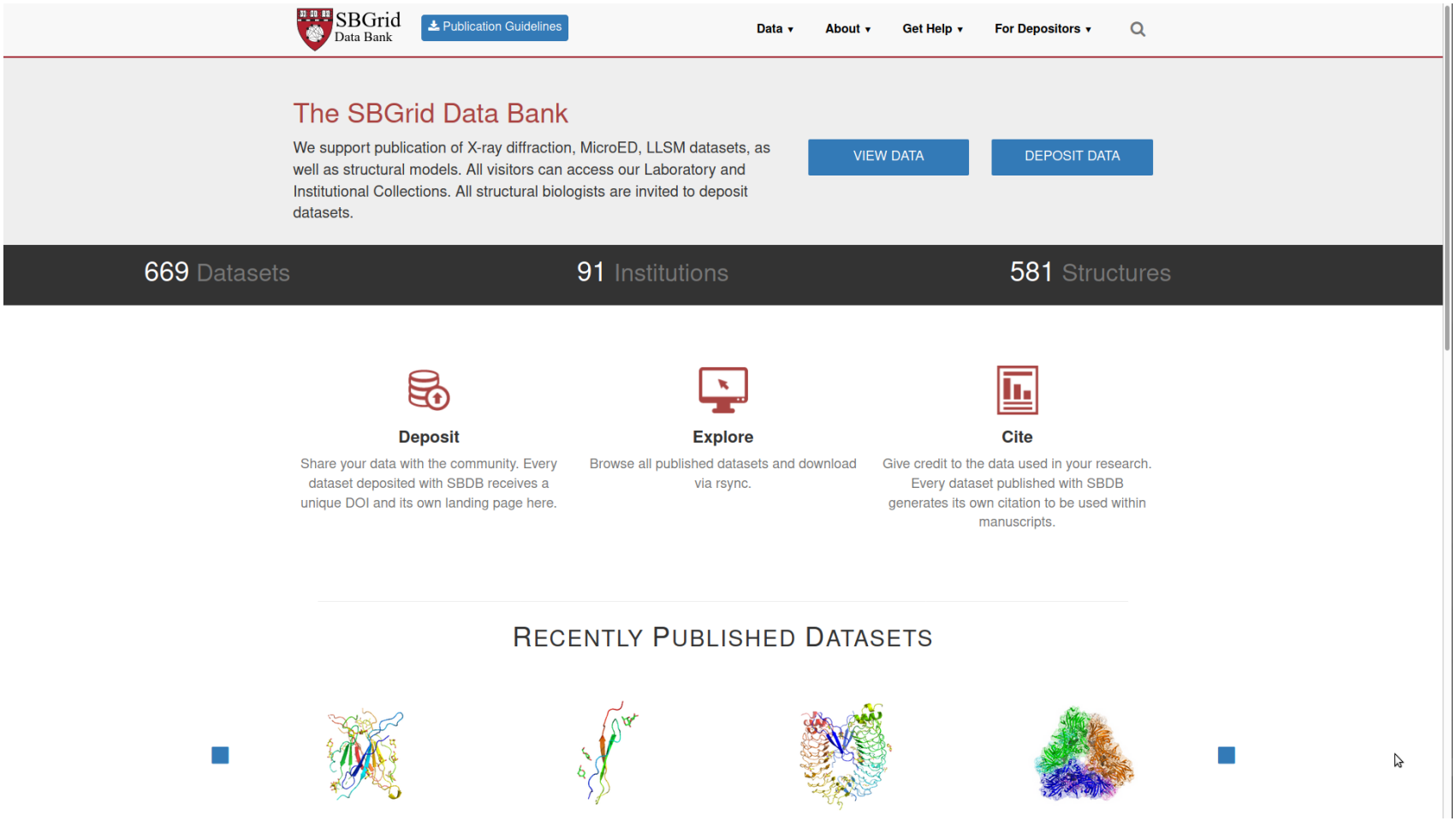
Raw diffraction data

https://proteindiffraction.org/



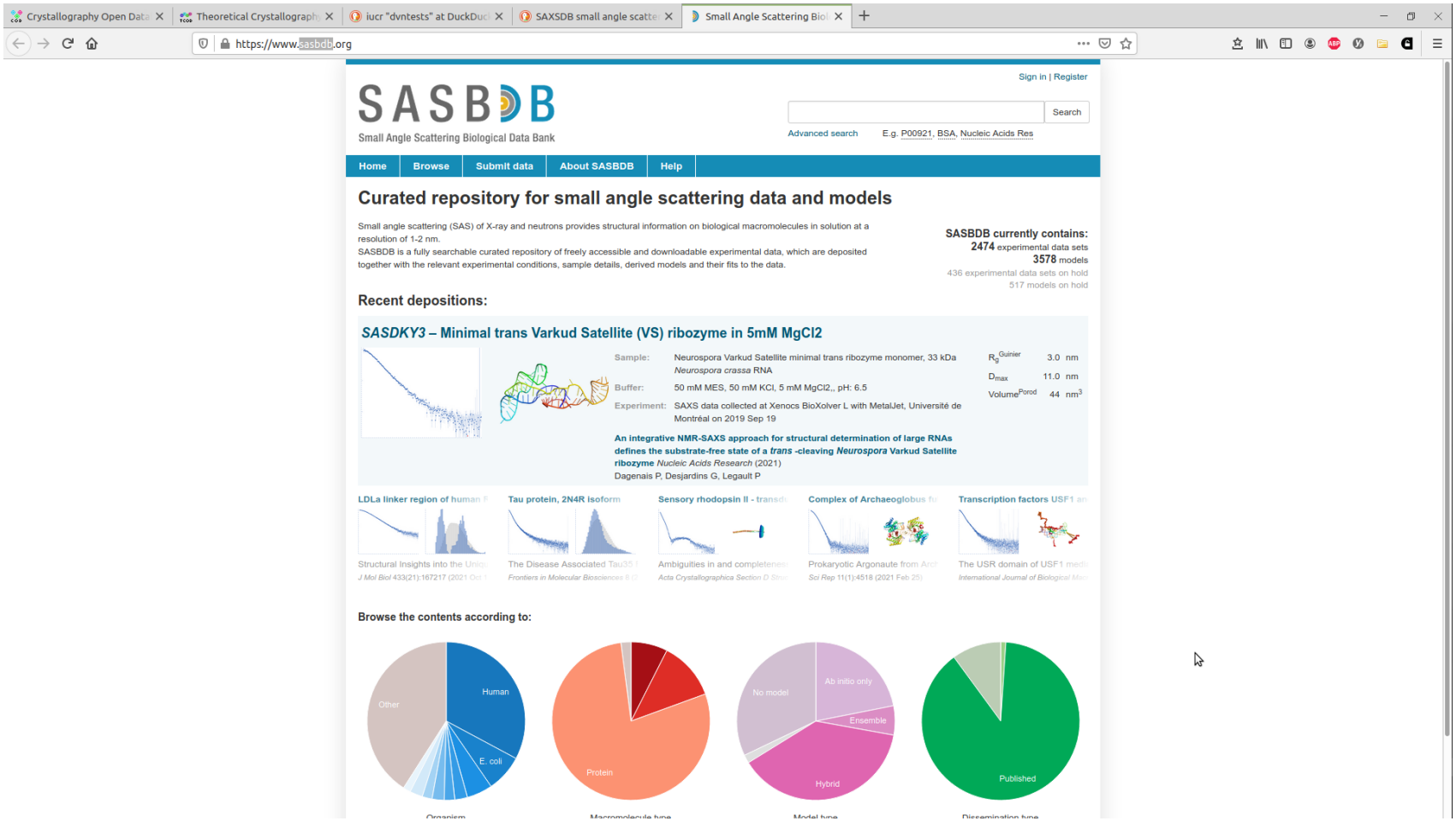
Raw diffraction data

https://data.sbgrid.org/



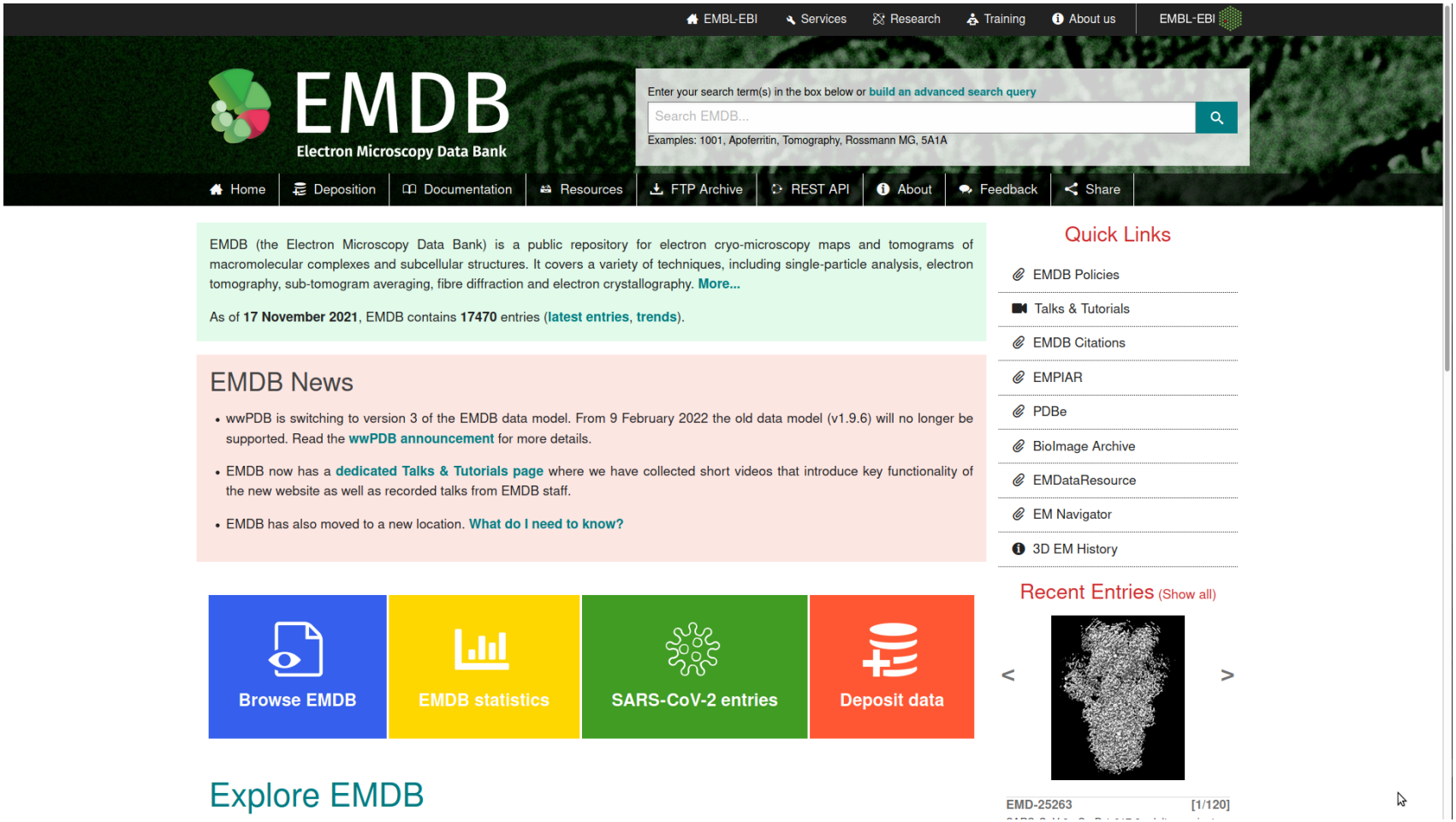
Small angle scattering database

https://www.sasbdb.org/



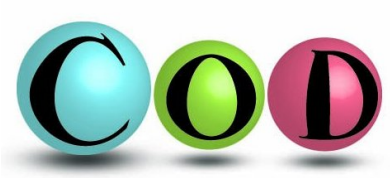
Electron microscopy densities

https://www.ebi.ac.uk/emdb/

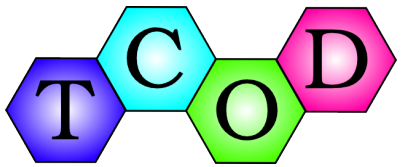


COD and TCOD databases

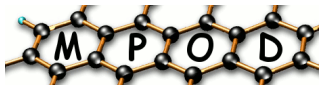
What is COD and TCOD?



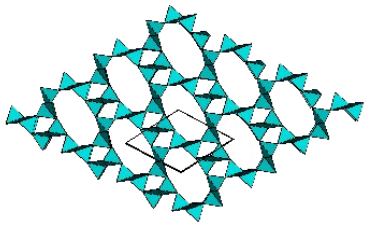
http://www.crystallography.net/cod
$>$ 480 000 entries



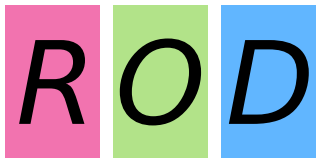
http://www.crystallography.net/tcod
$>$ 2900 entries (ready to grow to $> 10^7$?)



http://mpod.cimav.edu.mx/
$>$ 300 entries



http://www.crystallography.net/pcod
$>$ 10^6 entries (ready to grow to $> 10^8$?)



http://solsa.crystallography.net/rod/
$>$ 1100 entries

(Gražulis et al. 2009; Gražulis et al. 2012; Pepponi et al. 2012; Fuentes-Cobas et al. 2017; Mendili et al. 2019)

The COD database

The COD (https://www.crystallography.net/) is the largest to-date open access database of *experimental* crystal structures (Gražulis et al. 2009; Gražulis et al. 2012)

- Contains crystal data for compounds:
  - organic
  - metal-organic and organometallic
  - inorganic (including minerals, also from AMCSD)
- experimentally determined structures
- covers most peer-reviewed journals, updated daily
- over **500 000** records at the moment
- curated, both automatically and manually
- versioned

A typical COD record

http://www.crystallography.net/cod/1559914.html (deposited 2021-11-21)

Crystallography Open Database

COD Home
- Home
- What's new?

Accessing COD Data
- Browse
- Search
- Search by structural formula

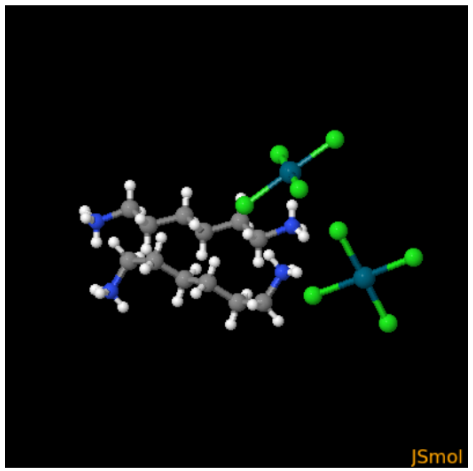
Add Your Data
- Deposit your data
- Manage depositions
- Manage/release prepublications

Documentation
- COD Wiki
- Obtaining COD
- License
- Querying COD
- Citing COD
- COD Mirrors
- Advice to donators
- Useful links

Information card for entry 1559914

1559913 << **1559914** >> 1559915

Preview



Chemical name	hexane-1,6-diaminium tetrachloro-palladium(ii)
Formula	C6 H18 Cl4 N2 Pd
Calculated formula	C6 H18 Cl4 N2 Pd
Title of publication	Experimental and Theoretical Evidence of Attractive Interactions between Dianions: [PdCl4]2-...[PdCl4]2-
Authors of publication	Zierkiewicz, Wiktor; Michalczyk, Mariusz; Maris, Thierry; Wysokinski, Rafal; Scheiner, Steve
Journal of publication	Chemical Communications
Year of publication	2021
a	7.2281 ± 0.0007 Å
b	8.1281 ± 0.0005 Å
c	11.7212 ± 0.0012 Å

CIF – Crystallographic Interchange Framework

Created and maintained by the International Union of Crystallography, IUCr (Hall et al. 1991).

`examples/data/2100858-head.cif`:

```
data_2100858
loop_
_publ_author_name
'Buttner, R. H.'
'Maslen, E. N.'
_publ_section_title
;
 Structural parameters and electron difference density in BaTiO~3~
;
_journal_issue                   6
_journal_name_full               'Acta Crystallographica Section B'
_journal_page_first              764
_journal_page_last               769
_journal_volume                  48
_journal_year                    1992
_chemical_compound_source        'synthetic, from a mixture of KF:KMoO4:BaTiO3'
_chemical_formula_sum            'Ba O3 Ti'
_chemical_formula_weight         233.24
_symmetry_cell_setting           tetragonal
_symmetry_space_group_name_Hall  'P 4 -2'
_symmetry_space_group_name_H-M   'P 4 m m'
_cell_length_a                   3.9998(8)
_cell_length_b                   3.9998(8)
_cell_length_c                   4.0180(8)
```

Semantic descriptions

- CIF – dictionaries
- XML, JSON – schemas
- SQL – schemas
- Semantic networks (RDF);

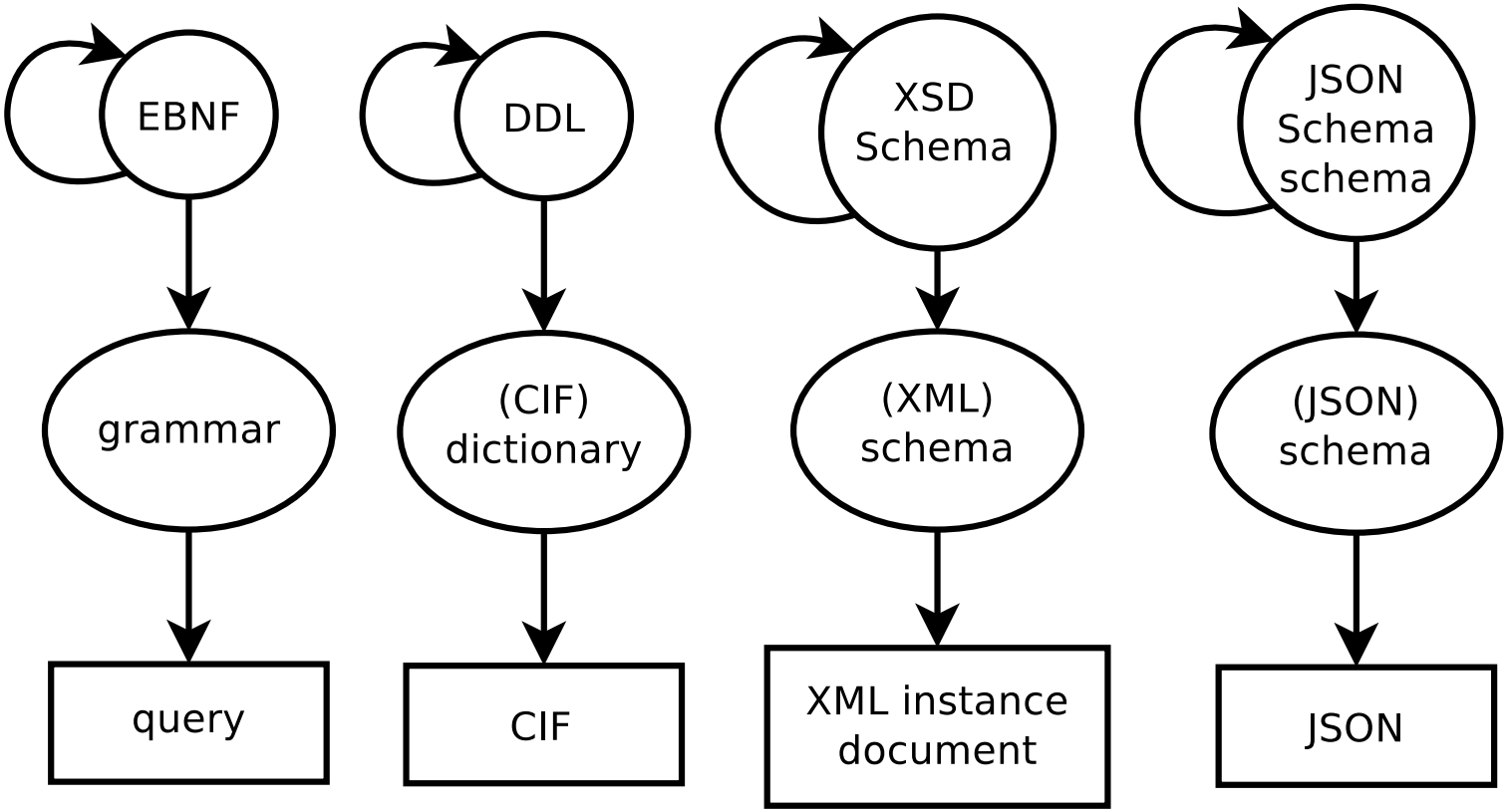
Controlled vocabularies

examples/dictionaries/cif-core-example.cif:

```
data_cell_length_
    loop_ _name                  '_cell_length_a'
                                 '_cell_length_b'
                                 '_cell_length_c'
    _category                     cell
    _type                         numb
    _type_conditions              esd
    _enumeration_range            0.0:
    _units                        A
    _units_detail                'angstroms'
    _definition
;               Unit-cell lengths in angstroms corresponding to the structure
                reported. The values of _refln_index_h, *_k, *_l must
                correspond to the cell defined by these values and _cell_angle_
                values. The values of _diffrn_refln_index_h, *_k, *_l may not
                correspond to these values if a cell transformation took place
                following the measurement of the diffraction intensities. See
                also _diffrn_reflns_transf_matrix_.
;
```

Self-describing data

Three levels are always enough!



Quality criteria

- IUCr quality criteria
  - IUCr list of data validation criteria
    ftp://ftp.iucr.org/pub/dvntests
  - IUCr requirements for publications
- COD quality criteria
  - ✓✓ Correct syntax;
  - ✓ Validation using data dictionaries;
  - ✓ Validation using data statistics;
  - ✓ Validation using first physical principles;

IUCr data validation criteria

IUCr criteria example (ftp://ftp.iucr.org/pub/dvntests):

```
------------------------------------------------------------------------
data validation criteria

AUTO CHECK LIST-Version: 2000.06.09

Full list of validation algorithms
------------------------------------------------------------------------

CHECK: [CHEMS] _chemical_formula_sum

TESTS:

  1. in the order C, H followed by alphabetic [CHEMS_01]
  2. classify as organic, inorganic, metal-organic [CHEMS_02]

EXAMPLE(S): 'C18 H19 N7 O8 S'

------------------------------------------------------------------------

CHECK: [CHEMW] _chemical_formula_weight

FORMAT: F.2

TESTS:


  1. agrees with _chemical_formula_sum [CHEMW_01]
  2. agrees with _atom_site_ data [CHEMW_03]
```

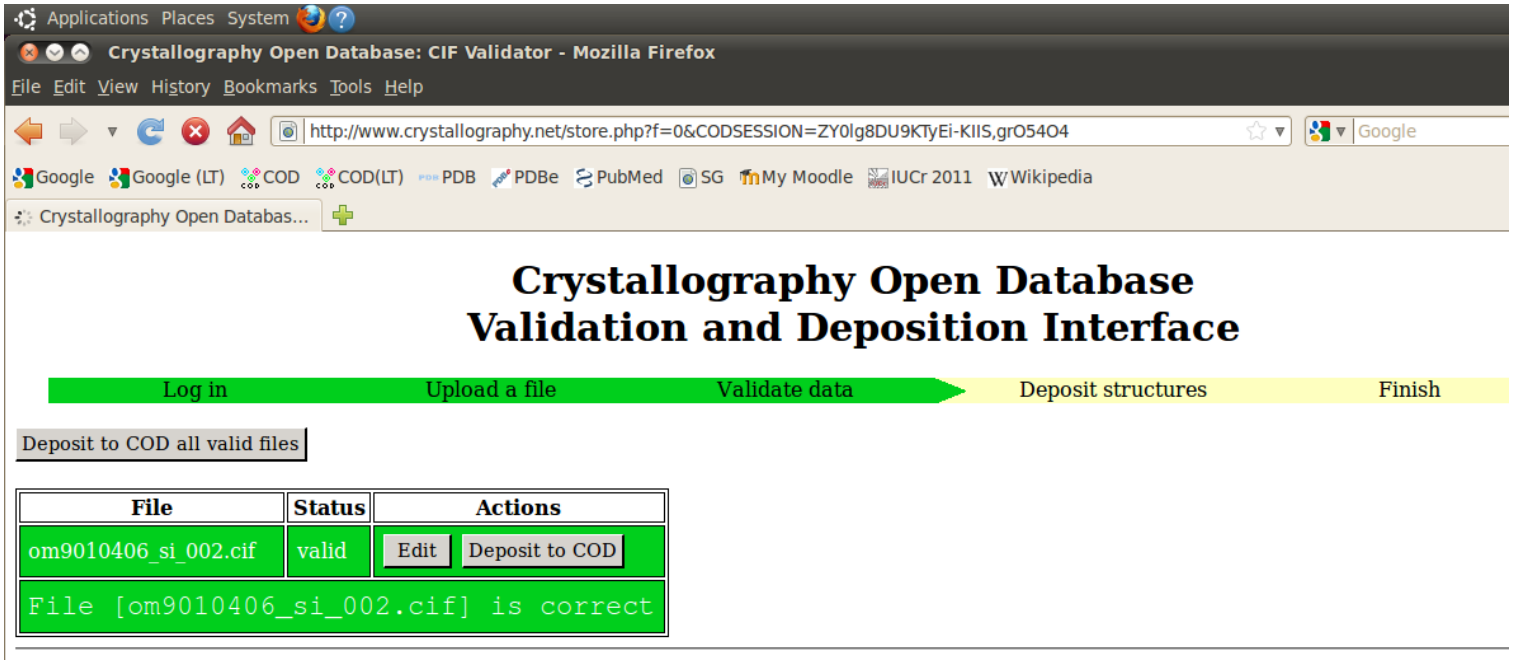
COD data checks

COD data validation policy:

1. Syntax checks:
   ```
   $ cifparse 7234818.cif
   ```
2. Semantic (dictionary) checks:
   ```
   $ cif_validate -D cif_core.dic 7234818.cif
   ```
3. Database specific tests:
   ```
   $ cif_cod_check 7234818.cif
   ```

COD data deposition Web site

COD data deposition Web site



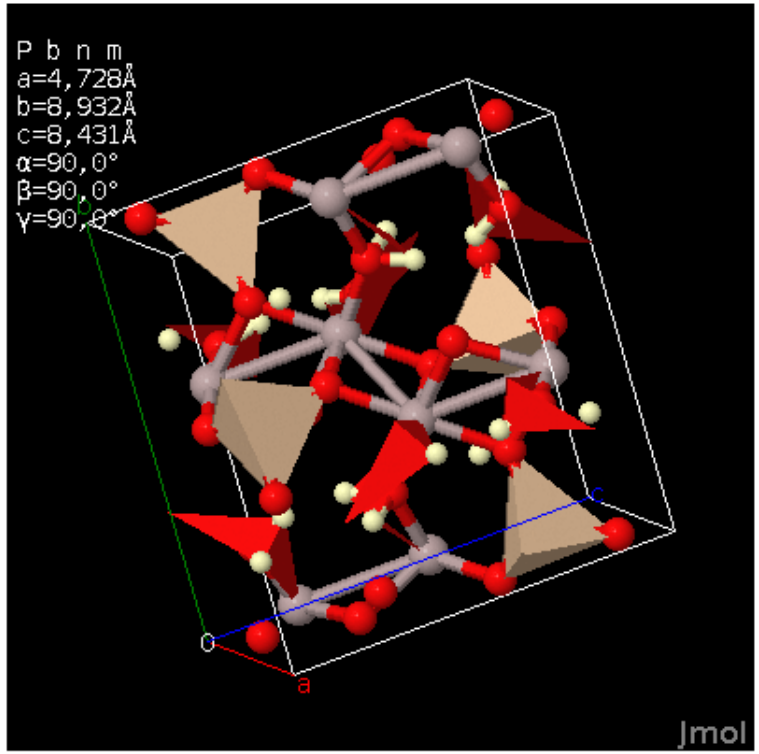
Types of identifiers

1. A centralised registry:
   - COD identifiers (e.g. COD 2000000);
   - PDB identifiers (e.g. PDB 1KNV);
   - DOI (e.g. 10.1093/nar/gkn883);
   - ARK (Archival Resource Key, e.g. ark:/53355/cl010066723)
   - URI (e.g. https://www.w3.org/Provider/Style/URI.html)
   - ISSN, ISBN, PMID, PMCID, ...
2. Randomised identifiers
   - UUID (e.g. 90376010-a315-11ea-adba-6bb1c61159af)
   - Cryptographic checksums (e.g. git commit 42a03a255612b8d43ecd77bb0acc02def888f688, 42a03a2);

Take home messages

1. Data provenance is important for reproducible research;
2. Unique stables identifiers help tracking data;
3. Comprehensive metadata are needed for reproducibility;
4. Let's use standard well-defined data formats;
5. Document your computation work flow.

Thank you!



Coordinates	2207377.cif
Original IUCr paper	HTML

http://en.wikipedia.org/wiki/Topaz

http://www.crystallography.net/2207377.html

References I

Fuentes-Cobas, Luis E. et al. (Aug. 2017). "The representation of coupling interactions in the Material Properties Open Database (MPOD)". In: *Advances in Applied Ceramics* 116.8, pp. 428–433. DOI: 10.1080/17436753.2017.1343782.

Gražulis, Saulius et al. (2009). "Crystallography Open Database – an open-access collection of crystal structures". In: *Journal of Applied Crystallography* 42, pp. 726–729. DOI: 10.1107/S0021889809016690. URL: http://dx.doi.org/10.1107/S0021889809016690.

Gražulis, Saulius et al. (2012). "Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration". In: *Nucleic Acids Research* 40, pp. D420–D427. DOI: 10.1093/nar/gkr900. URL: http://nar.oxfordjournals.org/content/40/D1/D420.abstract.

Hall, S. R. et al. (1991). "The crystallographic information file (CIF): a new standard archive file for crystallography". In: *Acta Crystallographica Section A* 47, pp. 655–685. DOI: 10.1107/S010876739101067X. URL: http://dx.doi.org/10.1107/S010876739101067X.

Li, Chuan-Yun et al. (Jan. 2008). "Genes and (common) pathways underlying drug addiction". In: *PLoS Computational Biology* 4.1. Ed. by Peter D. Karp, e2. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.0040002. URL: http://dx.doi.org/10.1371/journal.pcbi.0040002.

References II

Mendili, Yassine El et al. (May 2019). "Raman Open Database: first interconnected Raman–X-ray diffraction open-access resource for material identification". In: *Journal of Applied Crystallography* 52.3, pp. 618–625. DOI: 10.1107/s1600576719004229.

Pepponi, Giancarlo et al. (2012). "MPOD: A Material Property Open Database linked to structural information". In: *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* 284.0. E-MRS 2011 Spring Meeting, Symposium M: X-ray techniques for materials research-from laboratory sources to free electron lasers, pp. 10–14. ISSN: 0168-583X. DOI: 10.1016/j.nimb.2011.08.070. URL: http://www.sciencedirect.com/science/article/pii/S0168583X11008639.

Senior, Andrew W. et al. (Jan. 2020). "Improved protein structure prediction using potentials from deep learning". In: *Nature* 577.7792, pp. 706–710. DOI: 10.1038/s41586-019-1923-7.

Zheng, Yu et al. (Nov. 2008). "Using shotgun sequence data to find active restriction enzyme genes". In: *Nucleic Acids Research* 37.1, e1–e1. DOI: 10.1093/nar/gkn883. URL: https://doi.org/10.1093/nar/gkn883.