The impact of the Crystallography Open Database on Open Science

Saulius Gražulis

Calgary, 2026

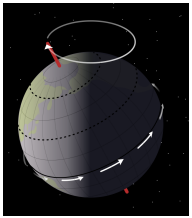
Vilnius University, Life Sciences Center, Institute of Biotechnology

Id: slides.tex 3459 2026-08-11 14:18:36Z saulius August 11, 2026

Data importance

Hipparchus (c. 190 – c. 120 BCE)

- measured the longitude of Spica and Regulus and other bright stars
- compared his measurements with data from his predecessors, Timocharis and Aristillus, who lived ≈**100** years before him,
- discovered what is now called *the precession of the equinoxes*



By NASA, Public Domain

(Wikipedia, see also articles on Timocharis and Aristyllus)

Crystallography – one of the earliest applications of computers in science :) (Shaffer et al. 1946);

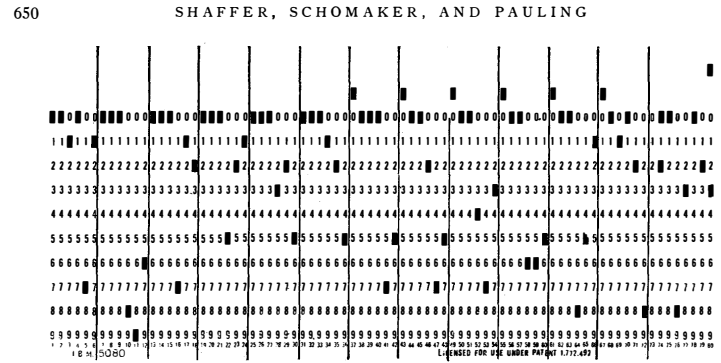


FIG. 1. A typical punched card. The vertical lines indicate the twelve function fields and the classification field for the card design described in the text.

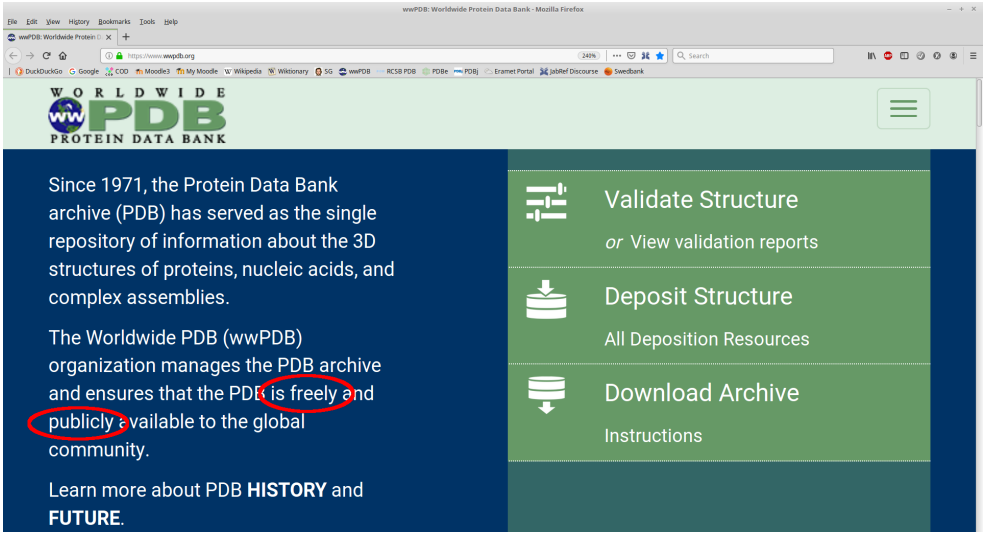
Distribution of data

US National Weather Records Center: 400 million punched cards (1960)

https://www.noaa.gov/heritage/resource-collections/managing-legacy-climate-data-in-20th-century

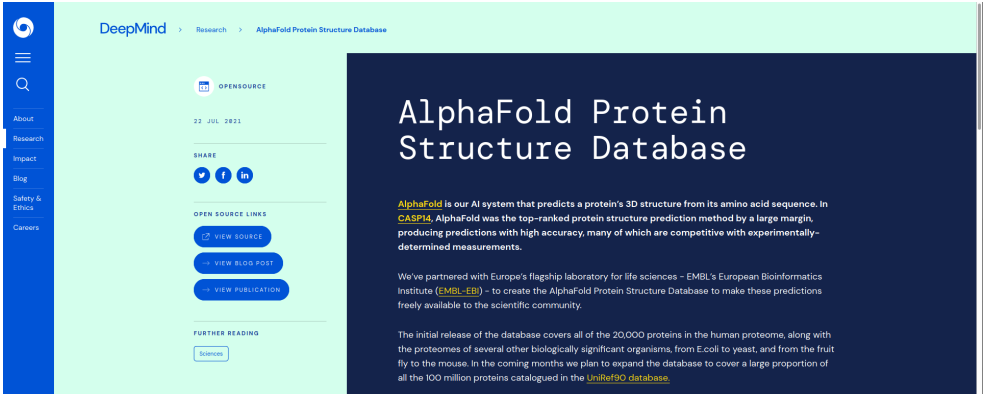
FAIR (Wilkinson et al. 2016), FAIR/O

PDB is excellent!



Consequences: AlphaFold

https://deepmind.com/research/open-source/alphafold-protein-structure-database[1]



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

[1] (accessed 2021-11-23)

The COD Project

Michaelio Berndto manifesto

But what if crystallographers work together to establish a public domain database with all relevant crystallographic data? This would not only overcome the current situation with 'fragmented' databases, it would also prevent for becoming dependent from monopolists.

What would be needed?

1. A small team of engaged scientists with some experience in database and software design to coordinate the project.

2. The authors (i.e. the scientific community = YOU) who provides the project with database entries (note, that if you have'nt sold your experimental results exclusively, you are free to distribute the data to such a database, even if they have already been part of a publication - and a lot of good data have never been published).

3. Free software a) for maintaining the database, b) for data evaluation and calculation of derived data (e.g. calculated powder pattern from crystal structures for search-match purposes), c) for browsing and retrieval.

gemstonede (Dr. Michael BERNDT) Fri Feb 14, 2003 1:26 pm

The COD Project

Michaelio Berndto manifesto

But what if crystallographers work together to establish a public domain database with all relevant crystallographic data? This would not only overcome the current situation with 'fragmented' databases, it would also prevent for becoming dependent from monopolists.

What would be needed?

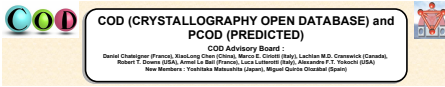
1. A small team of engaged scientists with some experience in database and software design to coordinate the project.

2. The authors (i.e. the scientific community = YOU) who provides the project with database entries (note, that if you have'nt sold your experimental results exclusively, you are free to distribute the data to such a database, even if they have already been part of a publication - and a lot of good data have never been published).

3. Free software a) for maintaining the database, b) for data evaluation and calculation of derived data (e.g. calculated powder pattern from crystal structures for search-match purposes), c) for browsing and retrieval.

gemstonede (Dr. Michael BERNDT) Fri Feb 14, 2003 1:26 pm

Establishment of COD (2003), Announcement at the IUCr (2005):



(Chateigner et al., 2005)

The COD Project

Michaelio Berndto manifesto

But what if crystallographers work together to establish a public domain database with all relevant crystallographic data? This would not only overcome the current situation with 'fragmented' databases, it would also prevent for becoming dependent from monopolists.

What would be needed?

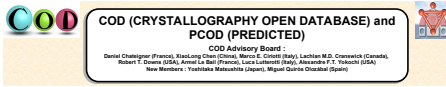
1. A small team of engaged scientists with some experience in database and software design to coordinate the project.

2. The authors (i.e. the scientific community = YOU) who provides the project with database entries (note, that if you have'nt sold your experimental results exclusively, you are free to distribute the data to such a database, even if they have already been part of a publication - and a lot of good data have never been published).

3. Free software a) for maintaining the database, b) for data evaluation and calculation of derived data (e.g. calculated powder pattern from crystal structures for search-match purposes), c) for browsing and retrieval.

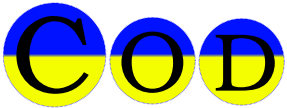
gemstonede (Dr. Michael BERNDT) Fri Feb 14, 2003 1:26 pm

Establishment of COD (2003), Announcement at the IUCr (2005):



(Chateigner et al., 2005)

23 years later...



there are 534673 entries in the COD https://crystallography.net

One of the first uses for the COD

„MOFomics: Computational pore characterization of metal–organic frameworks“ (First et al. 2013):

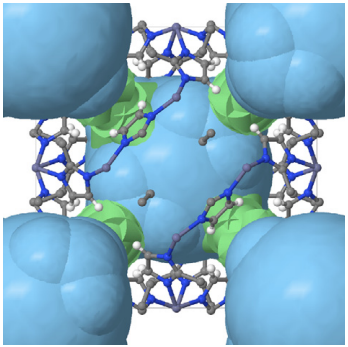
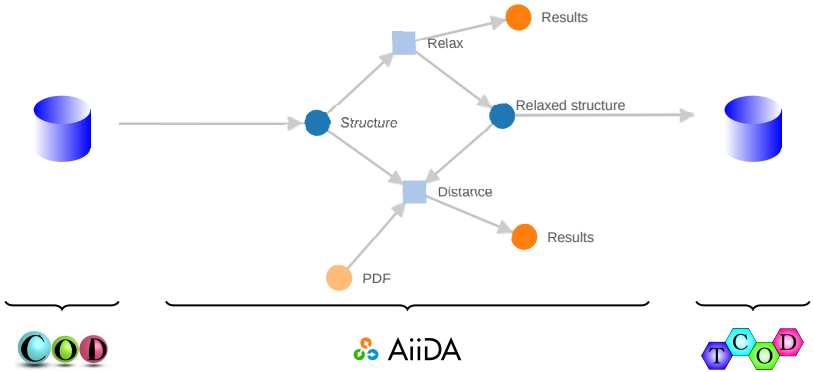


Table 2
Characterization results for selected structures. Volume and area are accessible to hydrogen (diameter 2.18 Å).

	LCD (Å)	PLD (Å)	Volume (cm^3/g)	Area (m^2/g)
GWMOF-3 (COD 4300942)	5.4	3.7	0.226	625
MIL-47 (CSD IDIWOH)	6.8	6.8	0.405	1781
MIL-53 (COD 4103388)	7.1	7.1	0.500	2203
MOF-5 (CSD SAHYIK)	15.0	7.8	1.186	2297
MOF-501 (COD 4300890)	11.0	5.1	0.747	2132
NOTT-401 (COD 7106668)	7.6	4.1	0.279	1080
ZIF-6 (CSD EQOCOC01)	9.5	6.7	0.749	1076
ZIF-8 (CSD VELVOY)	11.4	3.4	0.485	1531
Hyp. MOF 5082	9.8	4.0	0.850	2320
Hyp. MOF 18075	7.2	4.7	1.060	1435
Hyp. MOF 32532	5.9	3.2	0.345	1416

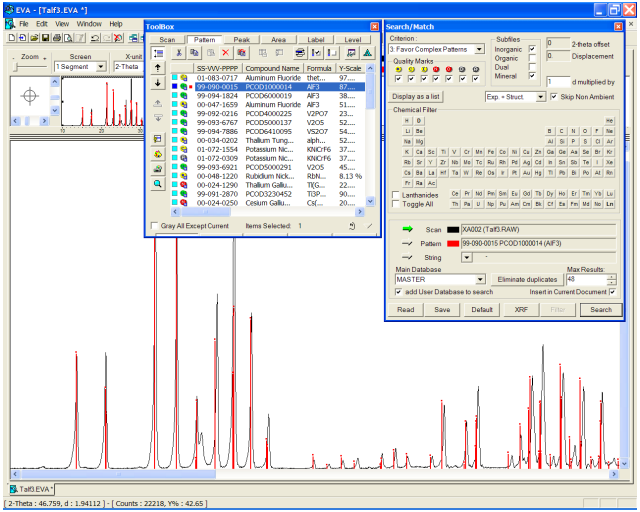
The use of COD



- Computational materials science

(Merkys et al. 2017; Pizzi 2018)

The use of COD



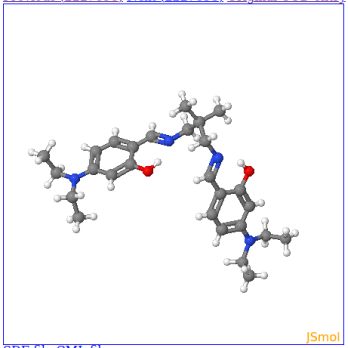
Use of PCOD and COD for materials identification

Courtesy Armel Le Bail (Le Bail 2008)

The use of COD

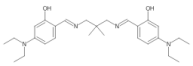
https://molecules.crystallography.net/

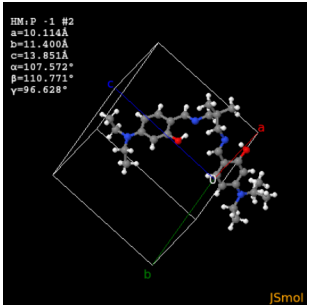
Previous (2227696) Next (2227698) Original COD entry



SDF file CML file

Reduced structural formula





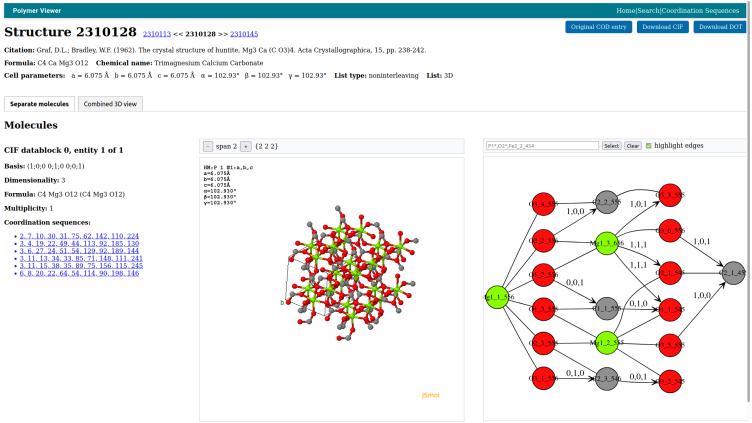
(Vaitkus et al. 2023)

Reduced canonical SMILES:

CCN(c1ccc(c(c1)O)/C=N/CC(C/N=C/c1ccc(cc1O)N(CC)CC)(C)C)CC **(x1)** PubChem

Searching for infinite nets

Building quotient graphs for covalent and coordination polymers.
http://polymers.crystallography.net/



Work of Yaroslav Rozdobudko

Restrain generation for macromolecular refinement

Statistics from the COD can serve as a basis for ligand geometry restrain generation: program AceDRG (Long et al. 2017a; Long et al. 2017b).

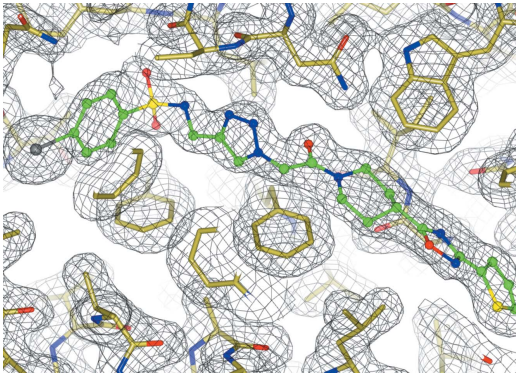
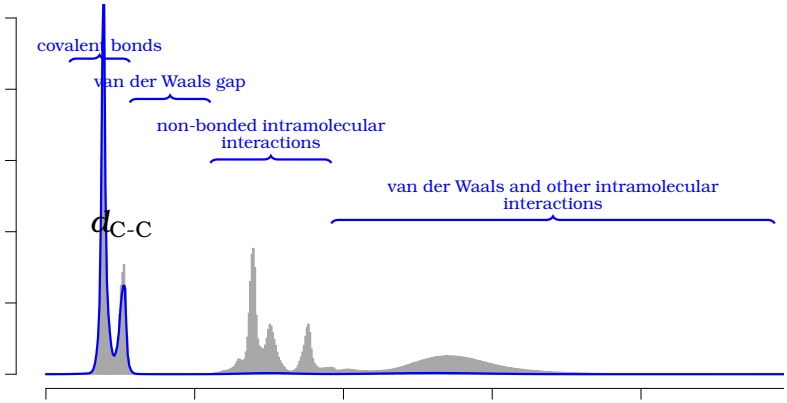


Figure 6
An example of a ligand refined using *AceDRG* restraints: PDB entry 2o8h with the ligand O8H (Willand *et al.*, 2010). Although the bond lengths are significantly different, the overlaid structures are visually almost identical. This figure was produced using *CCP4mg* (McNicholas *et al.*, 2011).

Updated atomic radii from open sources

Example: carbon-carbon distances. Data presented with *full provenance*.



((Šidlauskaitė et al. 2026), in press.)

The COD approach

Linking tables

The COD SQL data schema + COD CIF dictionary;

data

Field	Type	Key
file	int(7) unsigned	PRI
a	double unsigned	MUL
siga	float unsigned	
b	double unsigned	MUL
sigb	float unsigned	
c	double unsigned	MUL
sigc	float unsigned	
...	...	...

XRDa_x_COD

Field	Type	Key
id	int(11)	PRI
cod_id	int(7) unsigned	MUL
ext_id	int(11)	MUL
relation_id	int(11)	MUL

relations

Field	Type	Key
id	int(11)	PRI
name	varchar(255)	
description	text	

Similar schemas: Star schema: (Apeh 2024)

Example of the COD data record raw data link

XRDa as an example

https://www.crystallography.net/cod/3000497.html

COD Home
Home
What's new?

Accessing COD Data
Browse
Search
Search by structural formula

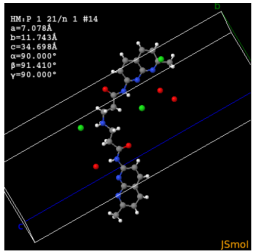
Add Your Data
Deposit your data
Manage depositions
Manage/release prepublications

Documentation
COD Wiki
Obtaining COD
License
Privacy and GDPR
Querying COD
Citing COD
COD Mirrors
Advice to donators
Useful links

Information card for entry 3000497

3000496 << **3000497** >> 3000498

Preview



Coordinates	3000497.cif
Original paper (by DOI)	HTML
External links	XRDa

▾ Structure parameters

Formula	C24 H26 Cl3 N7 O5
Calculated formula	C24 H26 Cl3 N7 O5
Authors of publication	Kawamoto, Akihiro; Kojima Chojiro; Kurisu, Genji; Nakane, Takanori; Nakatani Kazuhiko; Sakurabayashi Shuhei
Journal of publication	Journal of the American Chemical Society
Year of publication	2025
a	7.078 Å
b	11.743 Å
c	34.698 Å
α	90°
β	91.41°
γ	90°

(Sakurabayashi et al. 2025)

XRDa side

Links to the COD :)

https://xrda.pdbj.org/entry/335

Entry information

Entry ID:	XRD-00335
Entry title:	MicroED dataset of ND
Entry type:	Electron Diffraction
XRD entry authors:	Shuhei Sakurabayashi; Kyoko Furuita; Takeshi Yamada; Noriaki Sugiura; Makoto Nomura; Takanori Nakane; Akihiro Kawamoto; Genji Kurisu; Yohei Miyanoiri; Toshimichi Fujiwara; Kazuhiko Nakatani; Chojiro Kojima
Entry:	Download (141.85 GB)
XRDa entry DOI:	10.51093/xrd-00335
CCDC entry:	2350096
Crystallography Open Database entry:	3000497
Citation DOI:	10.1021/jacs.4c17538

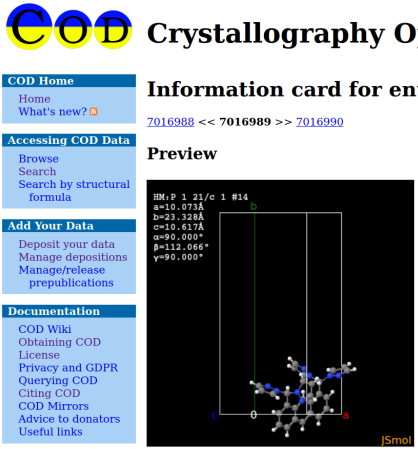
A side note: links with other data sources

Pubchem as an example

pubchem_x_cod

id	ext_id	cod_id
120396	101539687	7016989
99284	168336540	4135255
18081	168304792	1554891
147524	25198875	7155506
134464	168350383	7104999

169 593 rows
(Vaitkus et al. 2023)



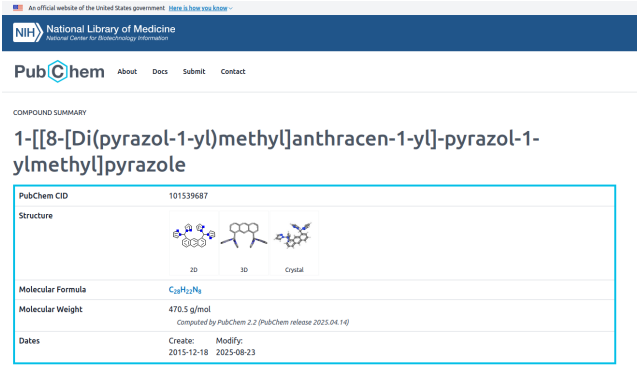
A side note: links with other data sources

Pubchem as an example

pubchem_x_cod

id	ext_id	cod_id
120396	101539687	7016989
99284	168336540	4135255
18081	168304792	1554891
147524	25198875	7155506
134464	168350383	7104999

169 593 rows
(Vaitkus et al. 2023)



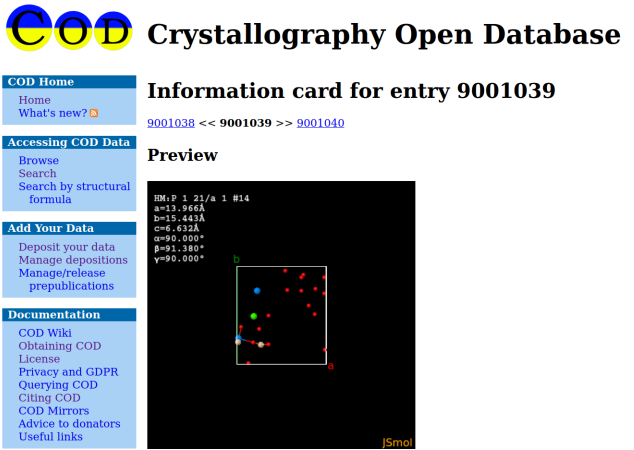
A side note: links with other data sources

Wikidata as an example

wikidata_x_cod

id	ext_id	cod_id
145	Q423494	9001039
2	Q1067912	9017917
150	Q182264	9001030
14	Q3649812	9000094
111	Q19861259	2105051

76 rows



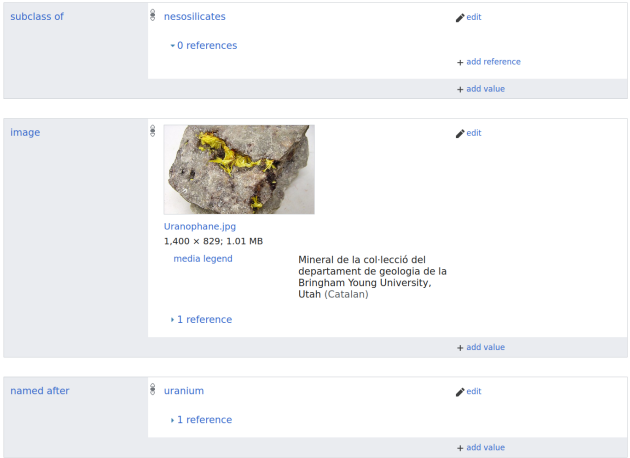
A side note: links with other data sources

Wikidata as an example

wikidata_x_cod

id	ext_id	cod_id
145	Q423494	9001039
2	Q1067912	9017917
150	Q182264	9001030
14	Q3649812	9000094
111	Q19861259	2105051

76 rows



Relation types

Different relations for different records

Relation “ontology”: the `relations` table;

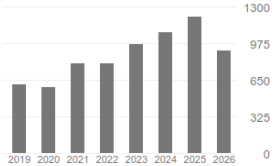
```
+----+--------------------+------------------------------------------ ...
| id | name               | description                                ...
+----+--------------------+------------------------------------------ ...
|  1 | RELATED            | THE CIF is related in some way to the ext ...
|  2 | DIRECTLY_DERIVED   | The CIF file was generated directly from  ...
|  3 | INDIRECTLY_DERIVED | Related by name, but not confirmed as the ...
|  4 | SIMILAR_TO         | It shares some characteristics but wasn't ...
+----+--------------------+------------------------------------------ ...
```

COD citations

Citations according to Google Scholar (https://scholar.google.lt/citations?user=BswX10sAAAAJ&hl=en, 2026-08-08 21:00 UTC):

TITLE	CITED BY	YEAR
Crystallography Open Database–an open-access collection of crystal structures S Gražulis, D Chateigner, RT Downs, AFT Yokochi, M Quirós, L Lutterotti, ... Applied Crystallography 42 (4), 726-729	2349	2009
Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration S Gražulis, A Daškevič, A Merkys, D Chateigner, L Lutterotti, M Quirós, ... Nucleic Acids Res 40, D420-D427	1710	2012
AceDRG: a stereochemical description generator for ligands F Long, RA Nicholls, P Emsley, S Gražulis, A Merkys, A Vaitkus, ... Biological Crystallography 73 (2), 112-122	511	2017
Validation of the crystallography open database using the crystallographic information framework A Vaitkus, A Merkys, S Gražulis Applied Crystallography 54 (2), 661-672	479	2021
Using SMILES strings for the description of chemical connectivity in the Crystallography Open	350	2018

Citations	9612	5785
h-index	39	24
i10-index	65	42



Public access VIEW ALL

8 articles 38 articles

New ways of publishing science

Suggesting a new type of scholarly publication:

Peer reviewed database record!

COD data can be **cited**:

- Referencing machine readable data:

  Srivastava, R. C.; Klooster, W. T.; Koetzle, T. F. “Neutron Structures of Ammonium Fluoroberyllate” (1999) *The Crystallography Open Database*, rev. 176759, the COD Advisory Board (eds.), https://www.crystallography.net/cod/2002926.cif@176759. [Retrieved 2016-09-21 16:48 EEST]

- Referencing a human-readable page:

  Srivastava, R. C.; Klooster, W. T.; Koetzle, T. F. “Neutron Structures of Ammonium Fluoroberyllate” (1999) *The Crystallography Open Database*, rev. 176759, the COD Advisory Board (eds.), https://www.crystallography.net/cod/2002926. [Retrieved 2016-09-21 16:48 EEST]

The summing up

- Open data and software allows us to provide *full* provenance to our calculations;
- Open data drive scientific innovation;
- Open data can make linked knowledge networks;
- Open data can increase awareness of general public about science, demonstrate justifications;

Acknowledgements

VU LSC IBT (KICIS)

Anastasija Kudrevceva
Andrius Merkys
Antanas Vaitkus
Algirdas Grybauskas

VU MIF II (FMG)

Linas Laibinis
Karolis Petrauskas
Irus Grinis
Haroldas Giedra

QM community

Linas Vilčiauskas
Vytautas Žalandauskas
Lukas Razinkovas
Nicola Marzari
Giovanni Pizzi
Matthew Evans
Rickard Armiento

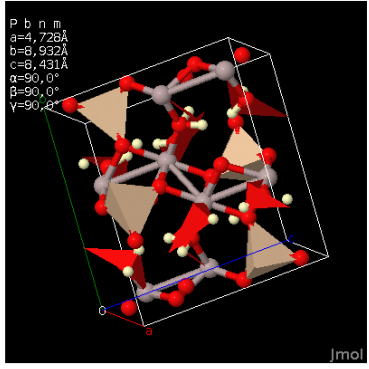
COD Advisory board

Daniel Chateigner
Robert T. Downs
Werner Kaminsky
Armel Le Bail
Luca Lutterotti
Peter Moeck
Peter Murray-Rust
Miguel Quirós

Funding:

Data Center for Machine Learning and Quantum Computing in Natural and Biomedical Sciences, Research Council of Lithuania Project No.: S-A-UEI-23-11

Thank you!



Coordinates	2207377.cif
Original IUCr paper	HTML

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

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

References I

Apeh, Kenneth (2024). *What is a Star Schema in Power BI*. URL: https://fp20analytics.com/wp-content/uploads/2024/10/What-is-a-Star-Schema-in-Power-BI.pdf.

First, Eric L. et al. (2013). "MOFomics: Computational pore characterization of metal–organic frameworks". In: *Microporous and Mesoporous Materials* 165, pp. 32–39. ISSN: 1387-1811. DOI: 10.1016/j.micromeso.2012.07.049. URL: http://www.sciencedirect.com/science/article/pii/S1387181112004817.

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.

Le Bail, Armel (2008). "Frontiers Between Crystal-structure Prediction and Determination by Powder Diffractometry". In: *Powder Diffraction Suppl.*, S5–S12. DOI: 10.1154/1.2903488.

Long, Fei et al. (Feb. 2017a). "ACEDRG: A stereo-chemical description generator for ligands". In: *Acta Crystallographica Section D* 73.2, pp. 112–122. DOI: 10.1107/S2059798317000067.

— (Feb. 2017b). "Validation and extraction of stereochemical information from small molecular databases". In: *Acta Crystallographica Section D* 73.2, pp. 103–111. DOI: 10.1107/S2059798317000079.

References II

Merkys, Andrius et al. (Nov. 15, 2017). “A posteriori metadata from automated provenance tracking: Integration of AiiDA and TCOD”. In: *Journal of Cheminformatics* 9.1, p. 56. DOI: 10.1186/s13321-017-0242-y. arXiv: 1706.08704v3 [cond-mat.mtrl-sci]. URL: https://jcheminf.springeropen.com/articles/10.1186/s13321-017-0242-y.

Pizzi, Giovanni (2018). “Open-Science Platform for Computational Materials Science: AiiDA and the Materials Cloud”. In: *Handbook of Materials Modeling*. Springer International Publishing, p. 1. DOI: 10.1007/978-3-319-42913-7_64-1.

Sakurabayashi, Shuhei et al. (Apr. 2025). “NMR-based rational drug design of G:G mismatch DNA binding ligand trapping transient complex via disruption of a key allosteric interaction”. In: *Journal of the American Chemical Society*, p. 4c17538. ISSN: 1520-5126. DOI: 10.1021/jacs.4c17538.

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. eprint: https://doi.org/10.1038/s41586-019-1923-7.

Shaffer, P. A. et al. (Nov. 1946). “The use of punched cards in molecular structure determinations I. Crystal structure calculations”. In: *The Journal of Chemical Physics* 14.11, pp. 648–658. DOI: 10.1063/1.1724081. URL: https://authors.library.caltech.edu/45530/1/SCHjcp46a.pdf.

Šidlauskaitė, Eglė et al. (2026). “Determination of atomic radii from small-moleculecrystal structures”. In: *Journal of Cheminformatics*. in press.

References III

Vaitkus, Antanas et al. (Dec. 2023). "A workflow for deriving chemical entities from crystallographic data and its application to the Crystallography Open Database". In: *Journal of Cheminformatics* 15.1. ISSN: 1758-2946. DOI: 10.1186/s13321-023-00780-2.

Wilkinson, Mark D. et al. (Mar. 2016). "The FAIR guiding principles for scientific data management and stewardship". In: *Scientific Data* 3.1. DOI: 10.1038/sdata.2016.18.