20 years of COD development: new data and features

Saulius Gražulis

Vilnius, 2024

Vilnius University Institute of Biotechnology

Id: slides.tex 2566 2024-01-11 12:49:10Z saulius January 11, 2024

Overview of the talk

- What is COD (with some history);
- COD contents and data curation principles;
- Recent developments with the COD;
- Future prospects.

https://www.crystallography.net/archives/2024/slides/20-years/slides.pdf

The Crystallography Open Database (COD)

https://www.crystallography.net/cod

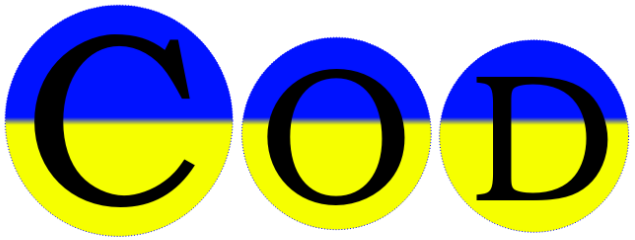
Crystallography Open Database

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
- Privacy and GDPR
- Querying COD
- Citing COD
- COD Mirrors
- Advice to donators
- Useful links



Open-access collection of crystal structures of organic, inorganic, metal-organic compounds and minerals, excluding biopolymers.

Including data and software from CrystalEye, developed by Nick Day at the department of Chemistry, the University of Cambridge under supervision of Peter Murray-Rust.

All data on this site have been placed in the public domain by the contributors.

Currently there are **509888** entries in the COD.
Latest deposited structure: 7159763 on **2024-01-11** at **01:32:14 UTC**

CIFs Donators

Advisory Board

Daniel Chateigner, Xiaolong Chen, Marco Ciriotti,
Robert T. Downs, Saulius Gražulis, Werner Kaminsky, Armel Le Bail, Luca Lutterotti,
Yoshitaka Matsushita, Andrius Merkys, Peter Moeck, Peter Murray-Rust, Miguel Quirós Olozábal,
Hareesh Rajan, Antanas Vaitkus, Alexandre F.T. Yokochi

If you find bugs in the COD or have any feedback, please contact us at
cod-bugs@ibt.lt

Top of the page

The COD project

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.

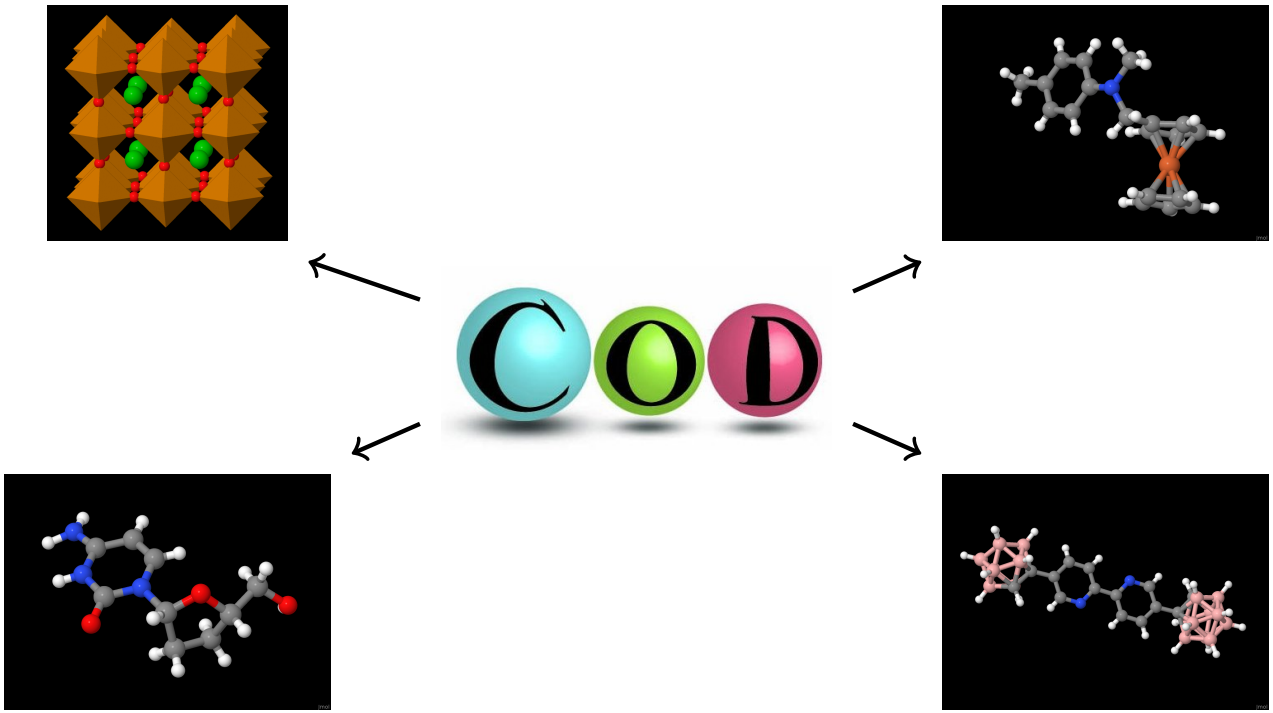
2. The authors (i.e. the scientific community = YOU) who provides the project with database entries (note, that if you have'nt sold your experimental results exclusively, you are free to distribute the data to such a database, even if they have already been part of a publication - and a lot of good data 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

COD contents

https://www.crystallography.net/cod



509 888 records as of 2024-01-11, available under **CC0 License**

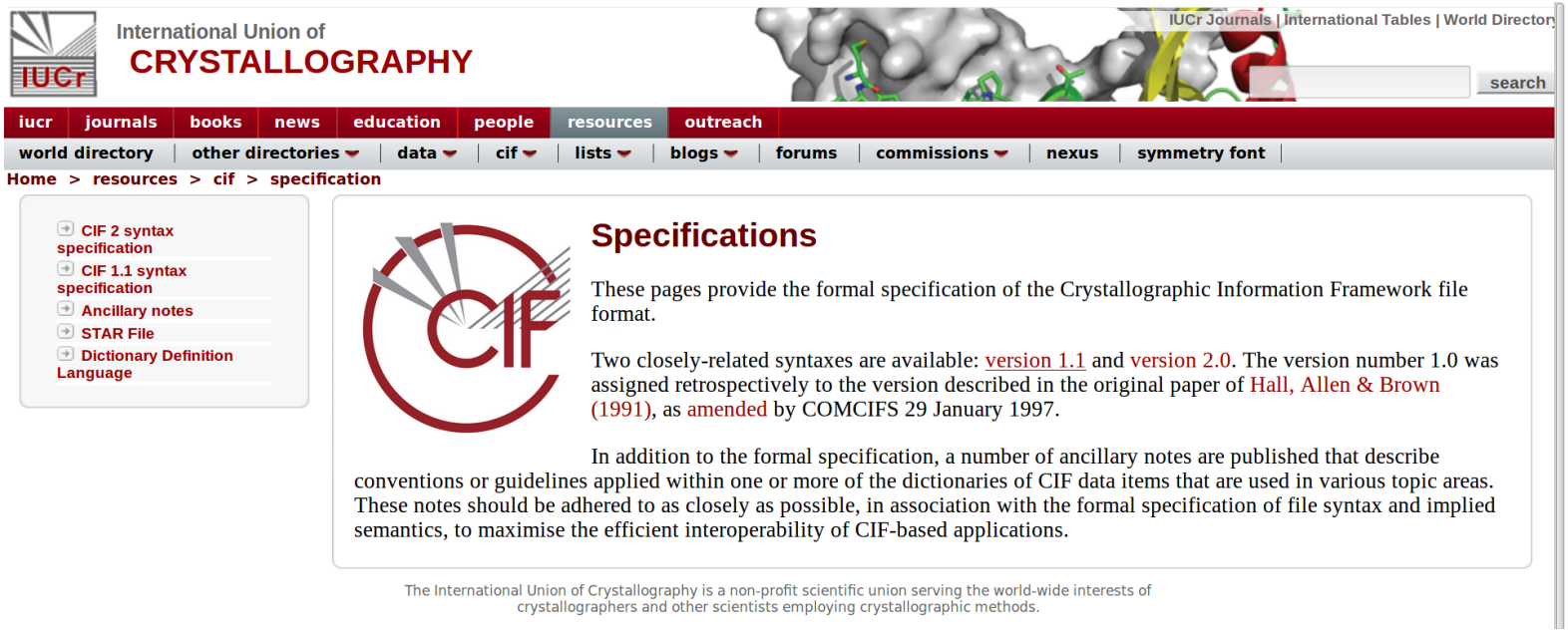
All data are presented in a standardised, machine-readable form (Gražulis et al. 2009; Gražulis et al. 2012).

COD data sources

- Peer-reviewed publications;
- Preprints, dissertations;
- Depositions by crystallographers (pers. comm., pre-publ.);
- Other databases; notably **AMCSD**, maintained by the group of Robert Downs (Downs et al. 2003; Rajan et al. 2006)

https://rruff.geo.arizona.edu/AMS/amcsd.php

CIF for the COD



(Hall et al. 1991; Bernstein et al. 2016)

The Crystallographic Interchange File/Framework (CIF):

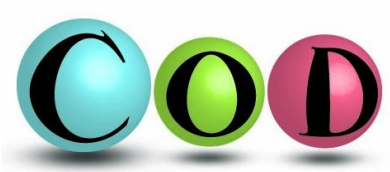
- Provides standard means for data publishing and exchange;
- Is suitable for archiving;
- Is maintained by the IUCr;

Accessing the COD

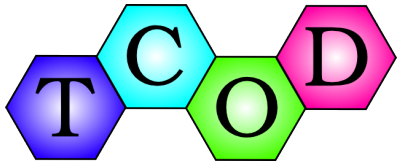
COD data can be accessed:

1. Via the Web page:
   https://www.crystallography.net/cod/7159763.html
2. Via the COD REST API:
   https://www.crystallography.net/cod/7159763.cif
   https://www.crystallography.net/cod/result?text=perovskite
3. Via the OPTIMADE API (Andersen et al. 2021):
   https://www.crystallography.net/cod/optimade/structures?filter=elements+HAS+"U"
4. Via SQL:
   ```
   mysql -u cod_reader -h sql.crystallography.net cod -e \
   'select file from data where formula = "- H2 O -"'
   ```
5. By downloading to your computer using Subversion, `rsync` or simple Web download:
   https://wiki.crystallography.net/howtoobtaincod

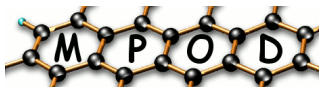
COD “sisters”



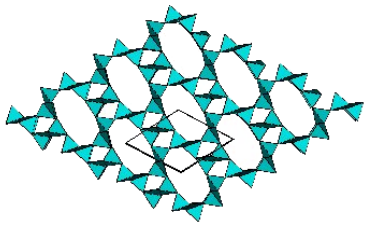
http://www.crystallography.net/cod
$>$ 500 000 entries



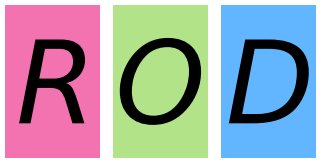
http://www.crystallography.net/tcod
$>$ 7400 entries (ready to grow to $>$ 10^7?)



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



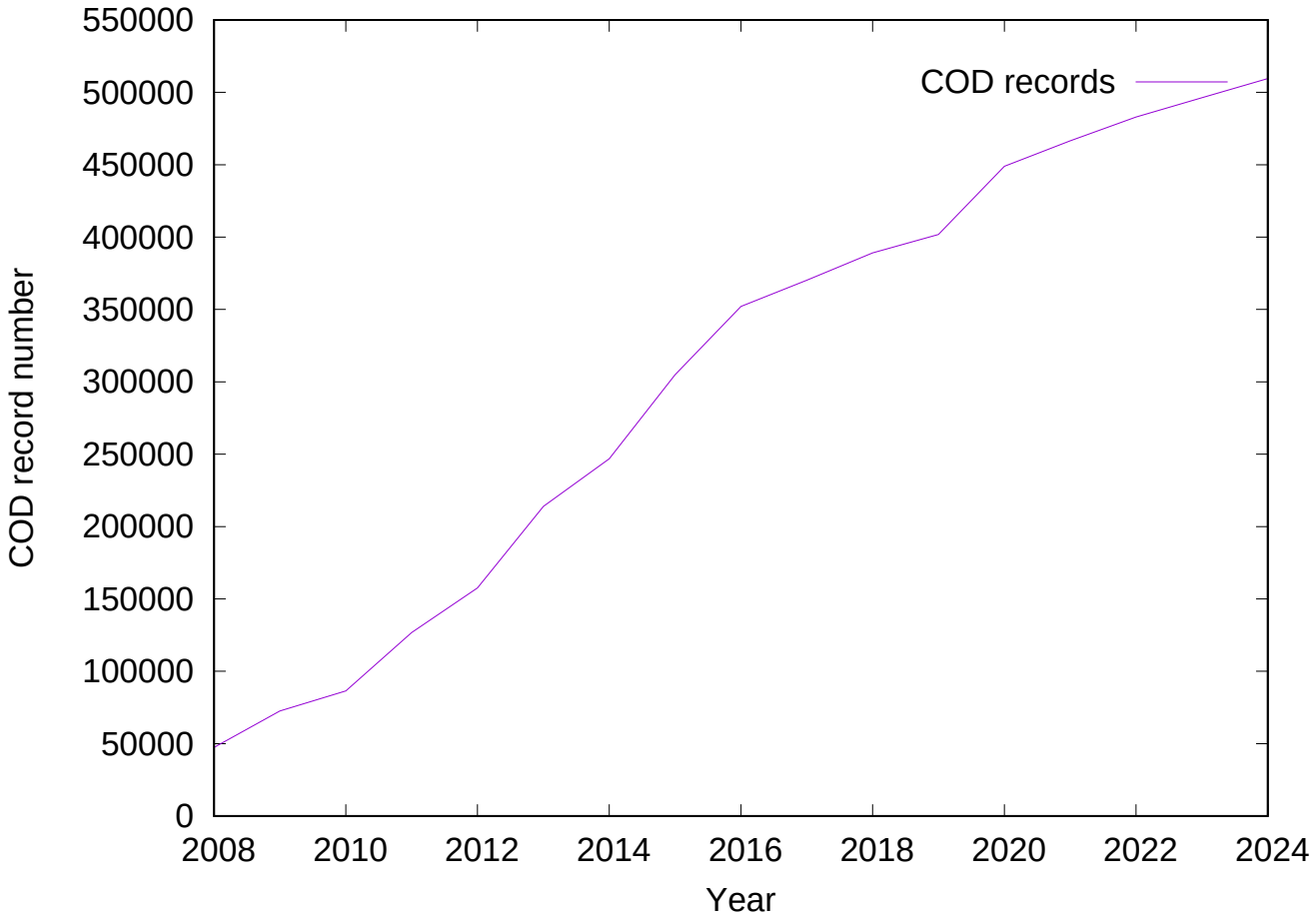
http://www.crystallography.net/pcod
$>$ 10^6 entries (ready to grow to $>$ 10^8?)



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)

COD growth


COD records
COD record number
Year
550000
500000
450000
400000
350000
300000
250000
200000
150000
100000
50000
0
2008
2010
2012
2014
2016
2018
2020
2022
2024

COD data curation

Inputs from COD users

Thomas Dortmann (2013), PANalytical, "`COD-minerals.xlsx`":

> *The entries that now have the mineral name are minerals, the rest are not.*

$> 3\,500$ unique mineral names assigned 104 "atypical" names[1].

Update (2024):

$> 4\,257$ unique mineral names, 566 "atypical" names

[1] Not matching the RE `/^[A-Z][-a-zA-Z ()]+$/`

COD versioning

Essential for reproducibility

All COD changes are tracked in a Subversion repository.

▾ Version history

Revision	Date	Message	Files
277834 (current)	2022-09-14	cif/ Added space group information derived from the space group operation list using the 'cif_filter' program.	2000000.cif
199748	2017-08-14	cif/2/00/00/ (antanas@echidna.ibt.lt) Removing 43 symmetrically equivalent atoms in entry 2000000.	2000000.cif

- The latest revision has a stable URI:
  https://www.crystallography.net/cod/2000000.cif
- A URI with a specific revision allows to reconstruct the *specific byte stream*:
  https://www.crystallography.net/cod/2000000.cif@199748

COD data validation

COD data validation policies:

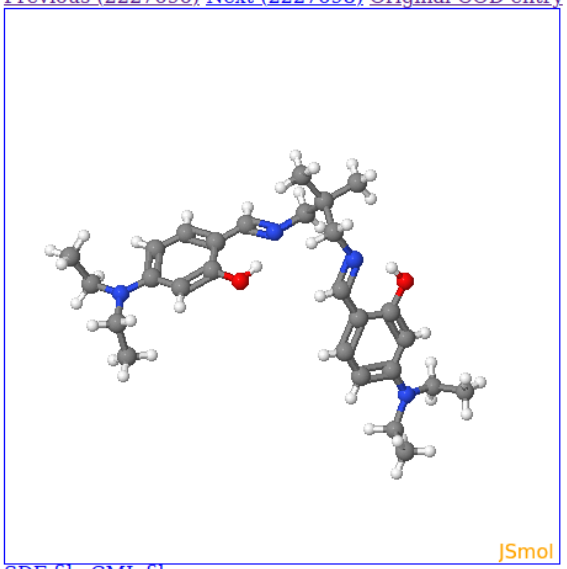
1. Syntactic checks:
   `$ cifparse 7234818.cif`
   Syntax recently expanded to CIF2 (Bernstein et al. 2016; Merkys et al. 2016)
2. Semantic validation (against dictionaries)
   `$ cif_validate -D cif_core.dic 7234818.cif`
   Validation capabilities recently expanded to DDLm (Vaitkus et al. 2021).
3. Database-specific checks
   `$ cif_cod_check 7234818.cif`

Commands from the `cod-tools` package:
svn://cod.ibt.lt/cod-tools
https://github.com/cod-developers/cod-tools

COD chemical repertoire

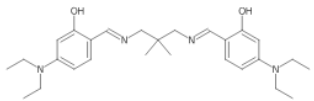
http://molecules.crystallography.net/

Previous (2227696) Next (2227698) Original COD entry



SDF file CML file

Reduced structural formula



Reduced canonical SMILES:

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

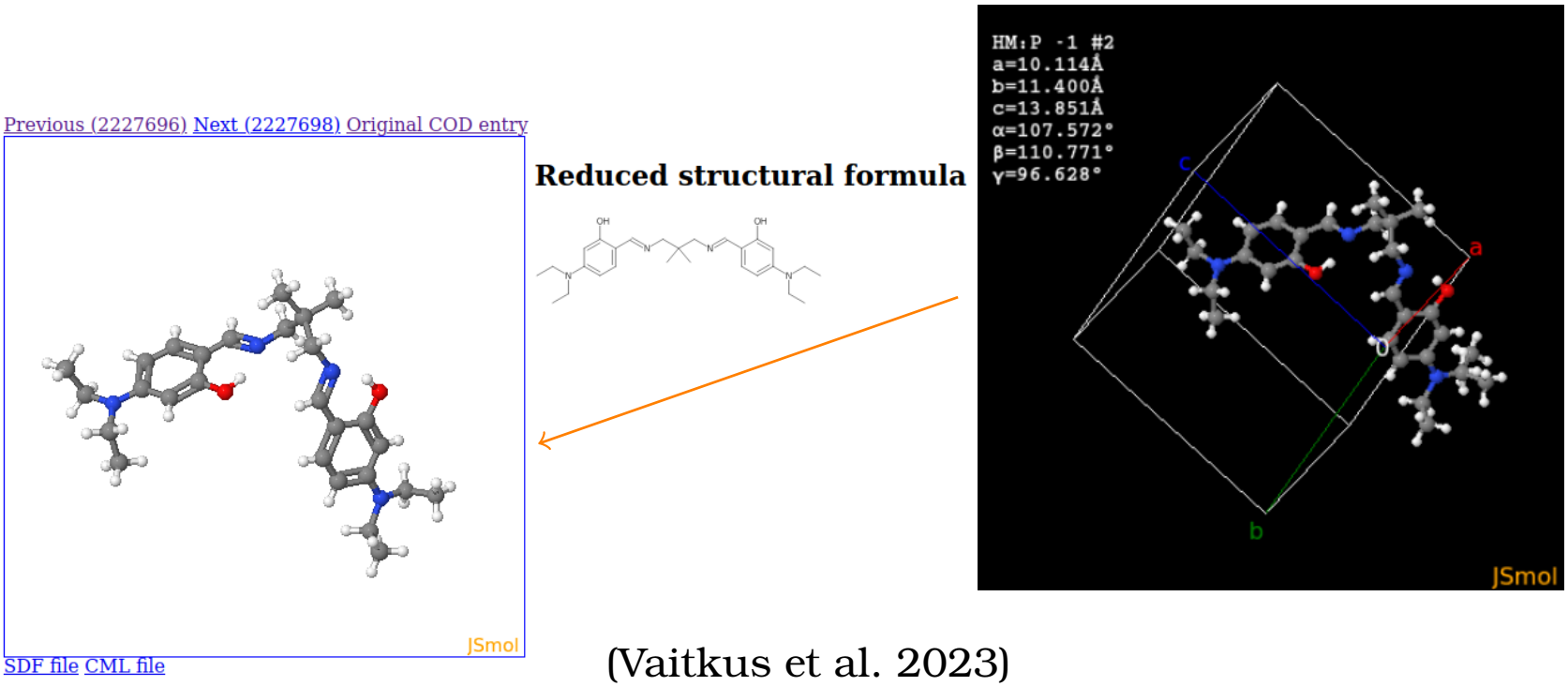
Unique components

SMILES	InChI
CCN(c1ccc(c(c1)O)/C=N/CC(C/N=C /c1ccc(cc1O)N(CC)CC)(C)C)CC	InChI=1S/C27H40N4O2/c1-7-30(8-2)23-13-11-21(25(32)15-23)17-28-19-27(5,6)2 /h11-18,32-33H,7-10,19-20H2,1-6H3/b28-17+,29-18+

See also poster by Merkys et al. (https://bit.ly/3BKZ5vG)

COD chemical repertoire

http://molecules.crystallography.net/



(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

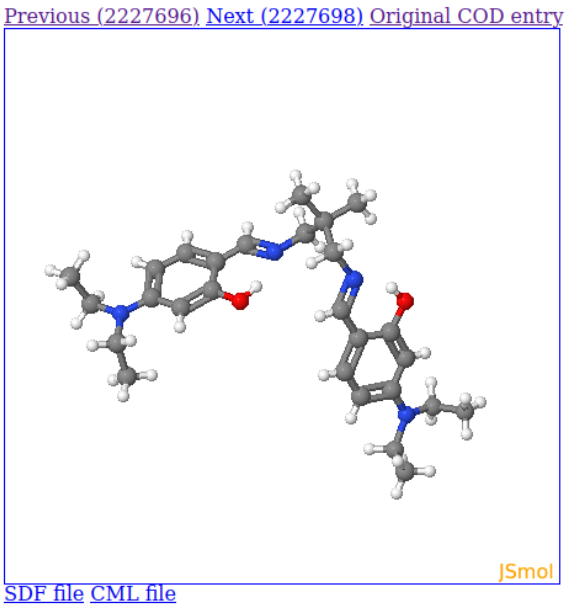
Unique components

SMILES	InChI
CCN(c1ccc(c(c1)O)/C=N/CC(C/N=C /c1ccc(cc1O)N(CC)CC)(C)C)CC	InChI=1S/C27H40N4O2/c1-7-30(8-2)23-13-11-21(25(32)15-23)17-28-19-27(5,6)2(/h11-18,32-33H,7-10,19-20H2,1-6H3/b28-17+,29-18+

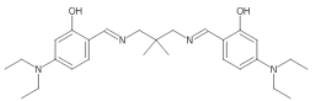
See also poster by Merkys et al. (https://bit.ly/3BKZ5vG)

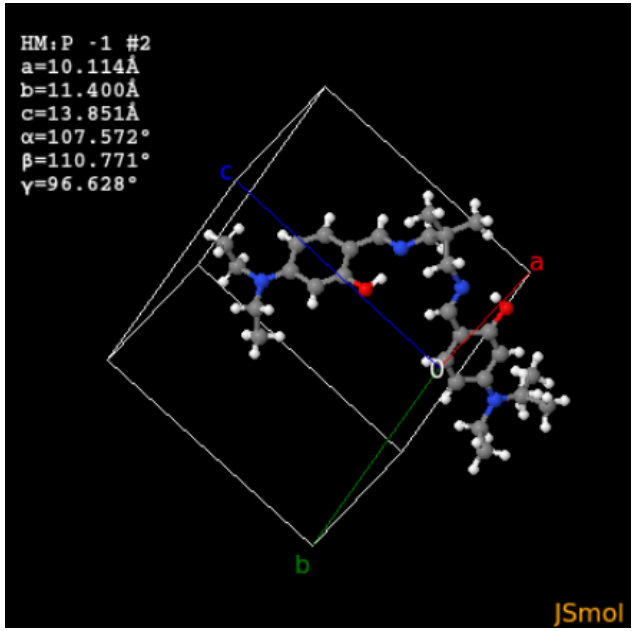
COD chemical repertoire

http://molecules.crystallography.net/



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

Unique components

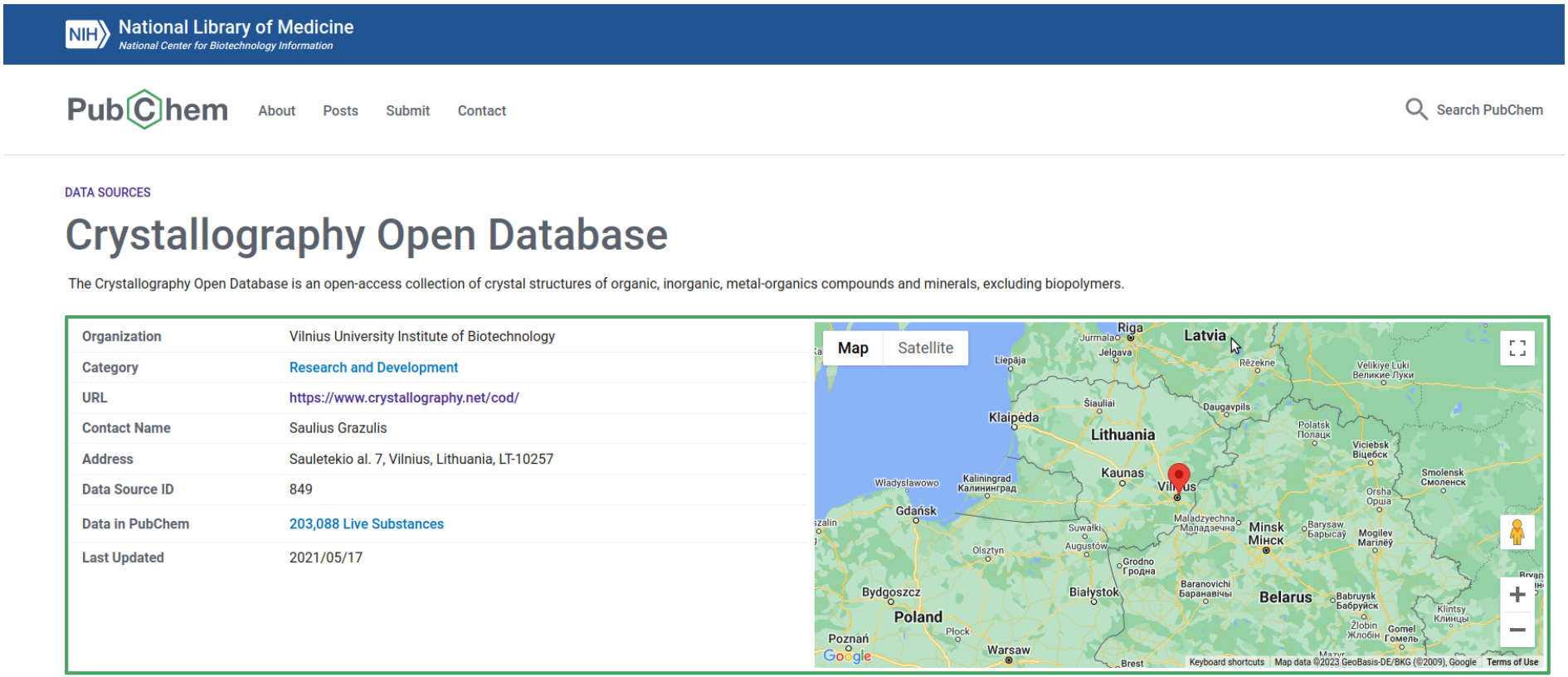
SMILES	InChI
CCN(c1ccc(c(c1)O)/C=N/CC(C/N=C/c1ccc(cc1O)N(CC)CC)(C)C)CC	InChI=1S/C27H40N4O2/c1-7-30(8-2)23-13-11-21(25(32)15-23)17-28-19-27(5,6)2(h11-18,32-33H,7-10,19-20H2,1-6H3/b28-17+,29-18+

See also poster by Merkys et al. (https://bit.ly/3BKZ5vG)

COD use cases

COD and PubChem

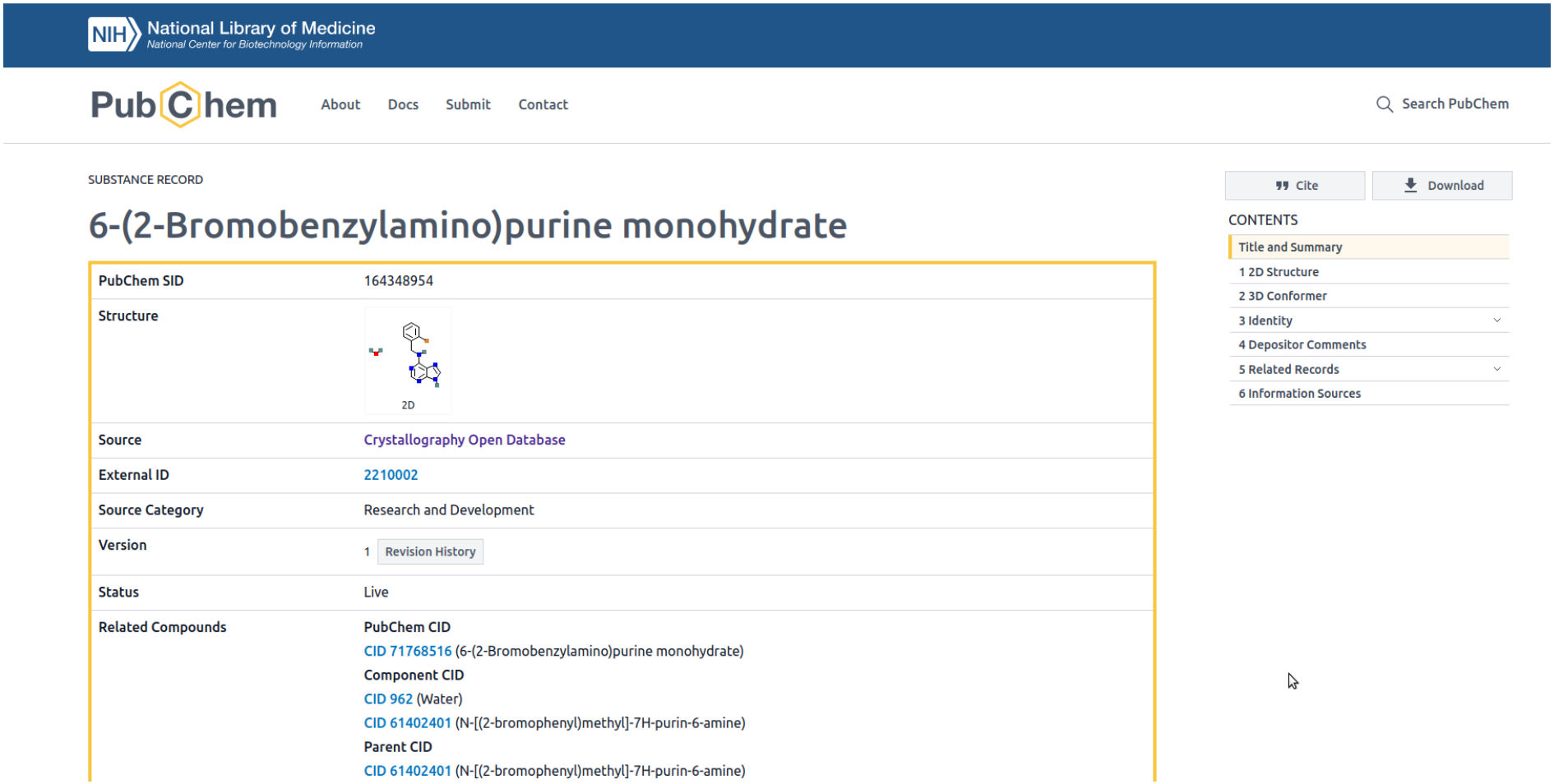
https://pubchem.ncbi.nlm.nih.gov/source/849



COD use cases

COD and PubChem

https://pubchem.ncbi.nlm.nih.gov/substance/164348954

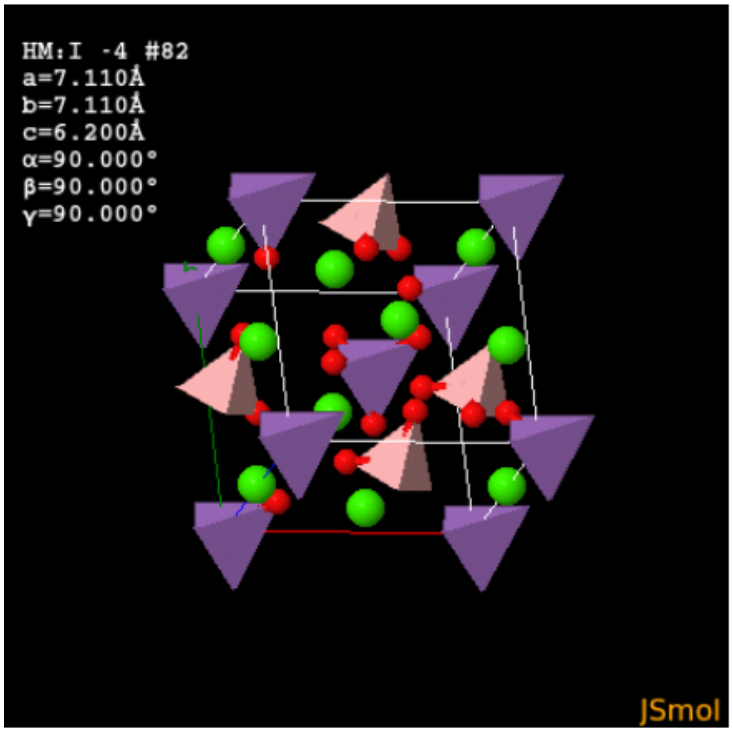


Data cross-referencing

External links

Links to external databases are implemented and populated:

- Implemented: AMCSD, Wikidata, Wikipedia, MPOD, ChemSpider;
- Planned: PubChem, **raw diffraction data**;



Coordinates 9016740.cif
External links AMCSD; Wikidata; Wikipedia

Group theory in Ada/SPARK

examples/group_theory.ads

```
pragma Spark_Mode (On);

generic
   type Element is private;
   Identity : Element;
   with function "*" (E, F: Element) return Element is <>;
```

```
   function Is_Closed_On_Multiplication (G : Group) return Boolean
   is (for all E of G =>
         (for all F of G => (Belongs_To (E*F, G))))
     with Ghost;

   function All_Elements_Have_Inverses (G : Group) return Boolean
   is (for all E of G => Has_Inverse (E, G))
     with Ghost;

   function Is_Group (G : Group) return Boolean
   is (Has_Identity (G) and then
         All_Elements_Have_Inverses (G) and then
         Is_Closed_On_Multiplication (G)
      )
     with Ghost;
```

(Petrauskas et al. 2022)

Automatic compilation of proven code

Ada and SPARK

examples/make_group.ads

```
8     type Ring_Element is mod 37;
```

```
29    function Build_Group (E : Ring_Element) return Group
30      with
31      Post => Is_Group (Build_Group'Result);
```

gnatprove -P main.gpr --report=all make_group.adb

```
make_group.ads:23:14: info: postcondition proved
make_group.ads:27:14: info: postcondition proved
make_group.ads:31:14: info: postcondition proved
group_theory.ads:16:15: info: postcondition proved, in instantiation at make_group.ads:16
```

```
saulius@tasmanijos-velnias spacegroups/ $ ./run_make_group 8
(1, 8, 27, 31, 26, 23, 36, 29, 10, 6, 11, 14)

saulius@tasmanijos-velnias spacegroups/ $ ./run_make_group 7
(1, 7, 12, 10, 33, 9, 26, 34, 16)
```

Where to go further?

- Derive chemical names;
- Collect more structures;
- Find all papers with crystal structures;
- Apply machine learning;
- Expand the community – **your contributions are invaluable!**

Acknowledgements

VU LSC IBT (KICIS)

Andrius Merkys
Antanas Vaitkus
Algirdas Grybauskas

QM community

Audrius Alkauskas
Vytautas Žalandauskas
Lukas Razinkovas
Nicola Marzari
Giovanni Pizzi
Lubomir Smrcok
Linas Vilčiauskas
Rickard Armiento

VU MIF II (FMG)

Linas Laibinis
Karolis Petrauskas

COD Advisory board

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

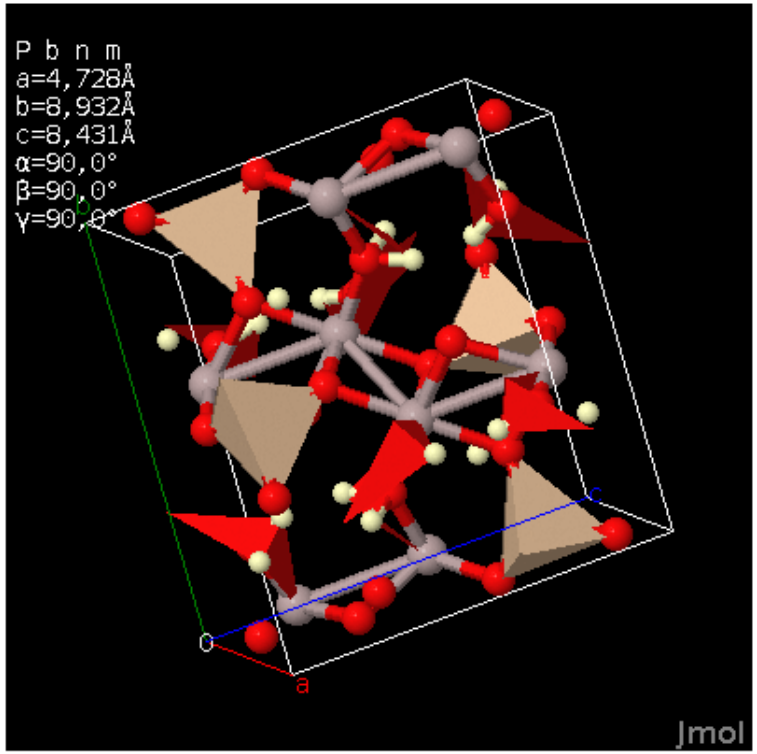
Cheminf community

Evan Bolton
Paul Thiessen
Thomas Sander

Enormous thanks for our commercial users and supporters: PANalytical, Rigaku, Bruker

Funding:
Lithuanian-French Program “Gilibert” (RCoL grant S-LZ-23-3); CECAM; RCoL grants S-MIP-20-21, S-MIP-23-87, VU Intramural funding.

Thank you!



Coordinates	2207377.cif
Original IUCr paper	HTML

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

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

https://www.crystallography.net/archives/2024/slides/20-years/slides.pdf

References I

Andersen, Casper W. et al. (Aug. 2021). "OPTIMADE, an API for exchanging materials data". In: *Scientific Data* 8.1, pp. 1–10. DOI: 10.1038/s41597-021-00974-z.

Bernstein, Herbert J. et al. (Feb. 2016). "Specification of the Crystallographic Information File format, version 2.0". In: *Journal of Applied Crystallography* 49.1, pp. 277–284. ISSN: 1600-5767. DOI: 10.1107/s1600576715021871. URL: http://dx.doi.org/10.1107/S1600576715021871.

Downs, Robert T. et al. (2003). "The American Mineralogist crystal structure database". In: *American Mineralogist* 88, pp. 247–250. URL: http://geo.arizona.edu/xtal/group/pdf/am88_247.pdf.

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.

References II

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.

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.

Merkys, Andrius et al. (Feb. 2016). "*COD::CIF::Parser*: an error-correcting CIF parser for the Perl language". In: *Journal of Applied Crystallography* 49.1, pp. 292–301. DOI: 10.1107/S1600576715022396. URL: http://dx.doi.org/10.1107/S1600576715022396.

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.

References III

Petrauskas, Karolis et al. (May 2022). "Proving the correctness of the algorithm for building a crystallographic space group". In: *Journal of Applied Crystallography* 55.3, pp. 515–525. DOI: 10.1107/s1600576722003107.

Rajan, H. et al. (2006). "Building the American Mineralogist Crystal Structure Database: A recipe for construction of a small Internet database". In: *Geoinformatics: Data to Knowledge*. Ed. by A.K. Sinha. Vol. 397. Geological Society of America Special Papers. Boulder, CO, United States: Geological Society of America, pp. 73–80. DOI: 10.1130/2006.2397(06).

Vaitkus, Antanas et al. (Feb. 2021). "Validation of the Crystallography Open Database using the Crystallographic Information Framework". In: *Journal of Applied Crystallography* 54.2, pp. 1–12. ISSN: 1600-5767. DOI: 10.1107/s1600576720016532. URL: https://doi.org/10.1107/S1600576720016532.

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.