Open Crystallographic Databases: COD, TCOD and the sisters

Saulius Gražulis

Vilnius, 2023

For the MIF++ seminar, University of Liverpool

Vilnius University Institute of Biotechnology

Id: slides.tex 2298 2023-09-28 12:41:15Z saulius September 28, 2023

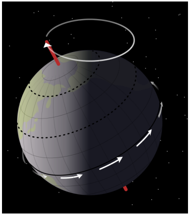
Layout of the talk

1. The value of crystallographic data
2. Crystallographic data(bases): COD, TCOD, PCOD, MPOD, ...
3. Applications of COD and sister databases
4. Mathematical considerations in crystal data processing

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*



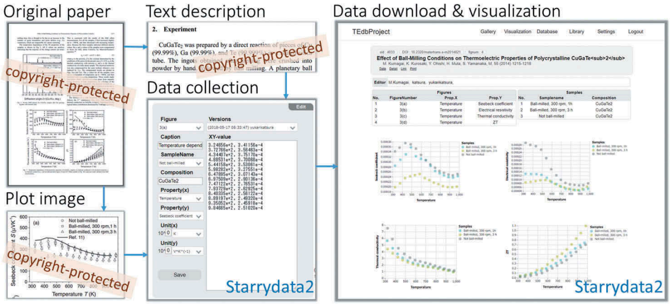
By NASA, Public Domain

(Wikipedia, see also articles on Timocharis and Aristyllus)

Publications are *not* data!

Starrydata2

Data need to be extracted (sometimes, manually...) from publications to make analyses.



2. Experiment

CuGaTe$_2$ was prepared by a direct reaction of pieces of Cu (99.99%), Ga (99.99%), and Te (99.999%) ... tube. The ingots obtained ... were crushed into powder by hand, followed by ball milling. A planetary ball

Edit

Figure: 3(a)

Versions: (2018-05-17 08:33:47) yukarikatsura

Caption: Temperature depend

SampleName: Not ball-milled

Composition: CuGaTe2

Property(x): Temperature

Property(y): Seebeck coefficient

Unit(x): 10^0 K

Unit(y): 10^0 V*K^(-1)

Save

XY-value

```
3.24656e+2, 3.41156e-4
3.72766e+2, 3.58463e-4
4.34407e+2, 3.75170e-4
4.88531e+2, 3.70068e-4
5.44158e+2, 3.53061e-4
5.98282e+2, 3.27551e-4
6.47895e+2, 3.07143e-4
6.97509e+2, 2.90136e-4
7.47122e+2, 2.76531e-4
7.93729e+2, 2.62925e-4
8.40335e+2, 2.56122e-4
8.89197e+2, 2.49320e-4
9.35052e+2, 2.45918e-4
9.84665e+2, 2.51020e-4
```

TEdbProject — Gallery · Visualization · Database · Library · Settings · Logout

sid : 4033 | DOI : 10.2320/matertrans.e-m2014821 | fignum : 4

Effect of Ball-Milling Conditions on Thermoelectric Properties of Polycrystalline CuGaTe₂

M. Kumagai, K. Kurosaki, Y. Ohishi, H. Muta, S. Yamanaka, M. 55 (2014) 1215-1218

Data Detail Link Print

Editor: M.Kumagai, katsura, yukarikatsura,

Figures			
No.	FigureNumber	Prop.X	Prop.Y
1	3(a)	Temperature	Seebeck coefficient
2	3(b)	Temperature	Electrical resistivity
3	3(c)	Temperature	Thermal conductivity
4	3(d)	Temperature	ZT

Samples		
No.	Samplename	Composition
1	Ball-milled, 300 rpm, 1h	CuGaTe2
2	Ball-milled, 300 rpm, 3 h	CuGaTe2
3	Not ball-milled	CuGaTe2

Figure 1. Concept of plot mining in the *Starrydata2* web system. An example paper [32] and the screenshots of *Starrydata2* web system are presented. Reproduced with permission from Thermoelectrics Society of Japan.

[Katsura et al. (2019)]

Publications are *not* data!

Starrydata2

But with data, new insights can be drawn from the aggregated publications: https://www.starrydata2.org/

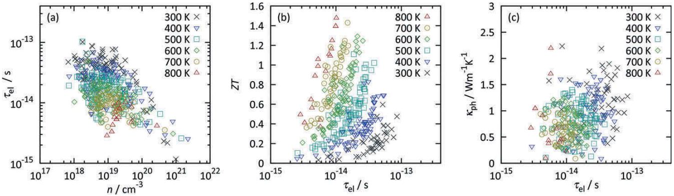


Figure 6. Relationship between (a) carrier doping level n and electron relaxation time τ_{el}, (b) τ_{el} and thermoelectric figure of merit ZT, and (c) τ_{el} and phonon thermal conductivity κ_{ph}, estimated for 207 experimental samples of n-type PbTe.

$$(\tau_{\text{el}} \in \left[10^{-15}..10^{-13}\right] \text{ vs. } \tau_{\text{el}} = 10^{-14} \text{ s})$$

[Katsura et al. (2019)]

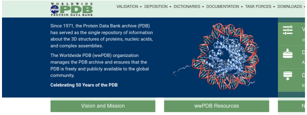
Crystallographic databases

Open Access:

Crystallographic databases

Open Access:

- Protein Data Bank;



WORLDWIDE PDB PROTEIN DATA BANK
VALIDATION ▾ DEPOSITION ▾ DICTIONARIES ▾ DOCUMENTATION ▾ TASK FORCES ▾ DOWNLOADS ▾
Vision and Mission
wwPDB Resources

Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

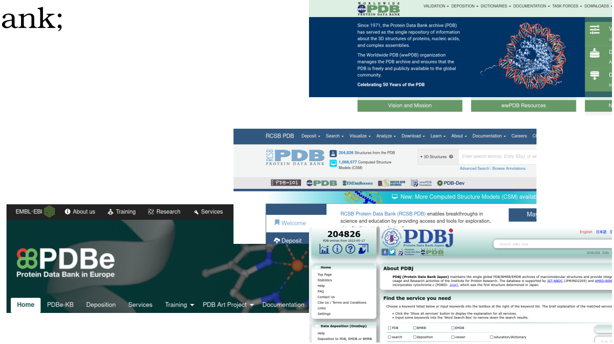
The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.

Celebrating 50 Years of the PDB

Crystallographic databases

Open Access:

- Protein Data Bank;


WORLDWIDE PDB PROTEIN DATA BANK
VALIDATION DEPOSITION DICTIONARIES DOCUMENTATION TASK FORCES DOWNLOADS
Vision and Mission
wwPDB Resources
RCSB PDB Deposit Search Visualize Analyze Download Learn About Documentation Careers
RCSB PDB PROTEIN DATA BANK
3D Structures
Enter search term(s), Entry ID(s), or se
Advanced Search | Browse Annotations
PDB-101 PDB EMDataResource NUCLEIC ACID DATABASE wwPDB Foundation PDB-Dev
Welcome
Deposit
May
EMBL-EBI About us Training Research Services
PDBe Protein Data Bank in Europe
Home PDBe-KB Deposition Services Training PDB Art Project Documentation
PDBj Protein Data Bank Japan
English 日本語
Search pdbj.org
Home
Top Page
Statistics
Help
FAQ
Contact Us
Cite Us / Terms and Conditions
Links
Settings
Data deposition (OneDep)
Help
Deposition to PDB, EMDB or BMRB

Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.

Celebrating 50 Years of the PDB

204,826 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

New: More Computed Structure Models (CSM) availab

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration,

204826

PDB entries from 2023-05-17

About PDBj

PDBj (Protein Data Bank Japan) maintains the single global PDB/BMRB/EMDB archives of macromolecular structures and provide integ Usage and Research activities of the Institute for Protein Research. The database is supported by JST-NBDC (JPMJND2205) and AMED-BINI incorporates cytochrome c (PDBID: 1cyc), which was the first structure determined in Japan.

Find the service you need

Choose a keyword listed below or input keywords into the textbox at the right of the keyword list. The brief explanation of the matched service

- Click the 'Show all services' button to display the explanation for all services.
- Input some keywords into the 'Word Search Box' to narrow down the search results.

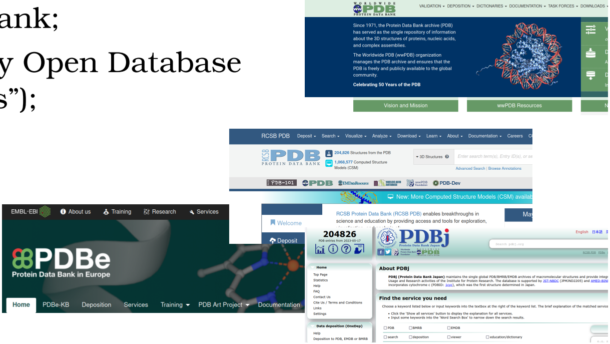
☐ PDB ☐ BMRB ☐ EMDB

☐ search ☐ deposition ☐ viewer ☐ education/dictionary

Crystallographic databases

Open Access:

- Protein Data Bank;
- Crystallography Open Database (and its “sisters”);



WORLDWIDE PDB PROTEIN DATA BANK
VALIDATION ▾ DEPOSITION ▾ DICTIONARIES ▾ DOCUMENTATION ▾ TASK FORCES ▾ DOWNLOADS ▾
Vision and Mission
wwPDB Resources
RCSB PDB Deposit ▾ Search ▾ Visualize ▾ Analyze ▾ Download ▾ Learn ▾ About ▾ Documentation ▾ Careers
RCSB PDB PROTEIN DATA BANK
▾ 3D Structures
Enter search term(s), Entry ID(s), or se
Advanced Search | Browse Annotations
PDB-101 PDB EMDataResource NUCLEIC ACID DATABASE wwPDB Foundation PDB-Dev
Welcome
Deposit
May
EMBL-EBI About us Training Research Services
PDBe Protein Data Bank in Europe
Home PDBe-KB Deposition Services Training ▾ PDB Art Project ▾ Documentation
PDBj Protein Data Bank Japan
English 日本語
Search pdbj.org
RCSB PDB PDBe
Home
Top Page
Statistics
Help
FAQ
Contact Us
Cite Us / Terms and Conditions
Links
Settings
Data deposition (OneDep)
Help
Deposition to PDB, EMDB or BMRB

Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.

Celebrating 50 Years of the PDB

204,826 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

New: More Computed Structure Models (CSM) availab

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration,

204826

PDB entries from 2023-05-17

About PDBj

PDBj (Protein Data Bank Japan) maintains the single global PDB/BMRB/EMDB archives of macromolecular structures and provide integ
Usage and Research activities of the Institute for Protein Research. The database is supported by JST-NBDC (JPMJND2205) and AMED-BIN
incorporates cytochrome c (PDBID: 1cyc), which was the first structure determined in Japan.

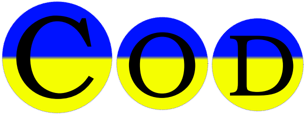
Find the service you need

Choose a keyword listed below or input keywords into the textbox at the right of the keyword list. The brief explanation of the matched service

- Click the 'Show all services' button to display the explanation for all services.
- Input some keywords into the 'Word Search Box' to narrow down the search results.

☐ PDB ☐ BMRB ☐ EMDB

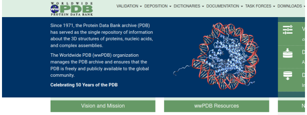
☐ search ☐ deposition ☐ viewer ☐ education/dictionary



Crystallographic databases

Open Access:

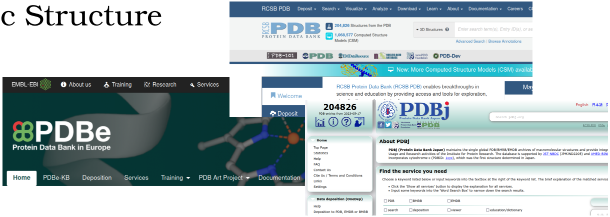
- Protein Data Bank;
- Crystallography Open Database (and its “sisters”);
- Bilbao Magnetic Structure Database



Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.

Celebrating 50 Years of the PDB



204,826 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

New: More Computed Structure Models (CSM) availab

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration,

204826

PDB entries from 2023-05-17

About PDBj

PDBj (Protein Data Bank Japan) maintains the single global PDB/BMRB/EMDB archives of macromolecular structures and provide integ Usage and Research activities of the Institute for Protein Research. The database is supported by JST-NBDC (JPMJND2205) and AMED-BINI incorporates cytochrome c (PDBID: 1cyc), which was the first structure determined in Japan.

Find the service you need

Choose a keyword listed below or input keywords into the textbox at the right of the keyword list. The brief explanation of the matched service

- Click the 'Show all services' button to display the explanation for all services.
- Input some keywords into the 'Word Search Box' to narrow down the search results.

□ PDB □ BMRB □ EMDB

□ search □ deposition □ viewer □ education/dictionary

COD

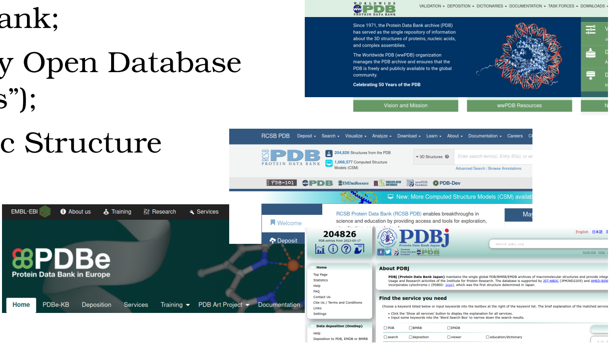
Crystallographic databases

Open Access:

- Protein Data Bank;
- Crystallography Open Database (and its “sisters”);
- Bilbao Magnetic Structure Database

Proprietary:

- CCDC
- ICSD
- PDF
- Pauling File
- ...



Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.

Celebrating 50 Years of the PDB

204,826 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

New: More Computed Structure Models (CSM) availab

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration,

204826

PDB entries from 2023-05-17

About PDBj

PDBj (Protein Data Bank Japan) maintains the single global PDB/BMRB/EMDB archives of macromolecular structures and provide integr Usage and Research activities of the Institute for Protein Research. The database is supported by JST-NBDC (JPMJND2205) and AMED-BINI incorporates cytochrome c (PDBID: 1cyc), which was the first structure determined in Japan.

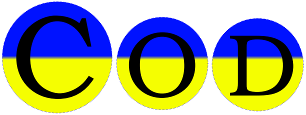
Find the service you need

Choose a keyword listed below or input keywords into the textbox at the right of the keyword list. The brief explanation of the matched service

- Click the 'Show all services' button to display the explanation for all services.
- Input some keywords into the 'Word Search Box' to narrow down the search results.

□ PDB □ BMRB □ EMDB

□ search □ deposition □ viewer □ education/dictionary



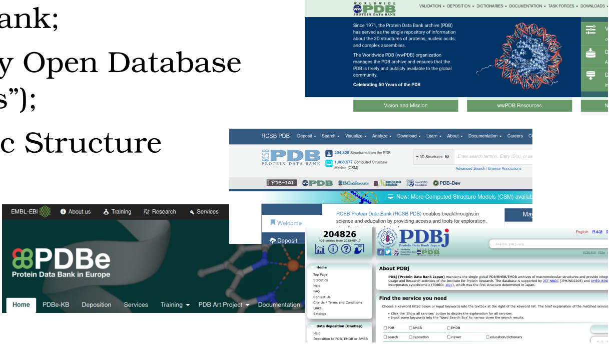
Crystallographic databases

Open Access:

- Protein Data Bank;
- Crystallography Open Database (and its “sisters”);
- Bilbao Magnetic Structure Database

Proprietary:

- CCDC
- ICSD
- PDF
- Pauling File
- …



Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.

Celebrating 50 Years of the PDB

204,826 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

New: More Computed Structure Models (CSM) availab

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration,

204826

PDB entries from 2023-05-17

About PDBj

PDBj (Protein Data Bank Japan) maintains the single global PDB/BMRB/EMDB archives of macromolecular structures and provide integ Usage and Research activities of the Institute for Protein Research. The database is supported by JST-NBDC (JPMJND2205) and AMED-BINI incorporates cytochrome c (PDBID: 1cyc), which was the first structure determined in Japan.

Find the service you need

Choose a keyword listed below or input keywords into the textbox at the right of the keyword list. The brief explanation of the matched service

- Click the 'Show all services' button to display the explanation for all services.
- Input some keywords into the 'Word Search Box' to narrow down the search results.

☐ PDB ☐ BMRB ☐ EMDB

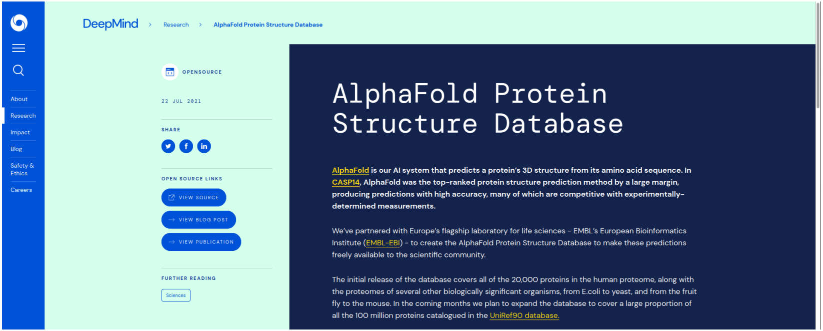
☐ search ☐ deposition ☐ viewer ☐ education/dictionary

COD

About $\mathbf{10^6}$ – $\mathbf{10^7}$ crystallographic records are available.

Consequences: AplhaFold

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



AlphaFold Protein Structure Database

AlphaFold is our AI system that predicts a protein's 3D structure from its amino acid sequence. In CASP14, AlphaFold was the top-ranked protein structure prediction method by a large margin, producing predictions with high accuracy, many of which are competitive with experimentally-determined measurements.

We've partnered with Europe's flagship laboratory for life sciences – EMBL's European Bioinformatics Institute (EMBL-EBI) – to create the AlphaFold Protein Structure Database to make these predictions freely available to the scientific community.

The initial release of the database covers all of the 20,000 proteins in the human proteome, along with the proteomes of several other biologically significant organisms, from E.coli to yeast, and from the fruit fly to the mouse. In the coming months we plan to expand the database to cover a large proportion of all the 100 million proteins catalogued in the UniRef90 database.

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

[1] (accessed 2021-11-23)

Contents

1. The value of crystallographic data
2. Crystallographic data(bases): COD, TCOD, PCOD, MPOD, ...
3. Applications of COD and sister databases
4. Mathematical considerations in crystal data processing

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).

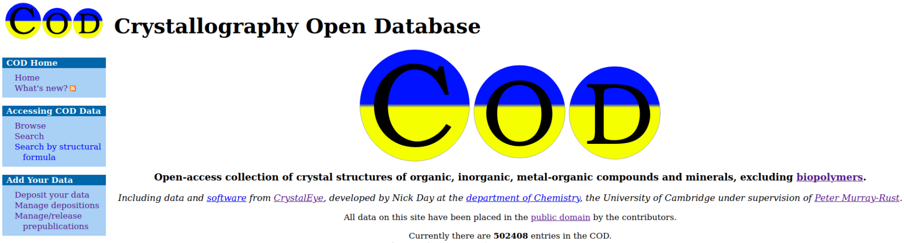
3. Free software a) for maintaining the database, b) for data evaluation and calculation of derived data (e.g. calculated powder pattern from crystal structures for search-match purposes), c) for browsing and retrieval.

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

The Crystallography Open Database (COD)

https://www.crystallography.net

Online since 2003 :)



Crystallography Open Database

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.

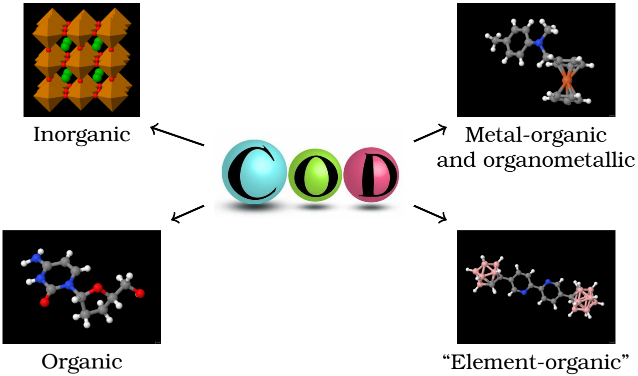
All data on this site have been placed in the public domain by the contributors.

Currently there are **502408** entries in the COD.

> **500 000** records as of 2023-05-22, available under CC0 License

COD contents

The Crystallography Open Database (COD)
https://www.crystallography.net



Inorganic

Metal-organic
and organometallic

Organic

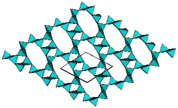
“Element-organic”

Open Crystallographic Databases

COD, TCOD, PCOD, MPOD, ROD, HOD ...

http://www.crystallography.net/cod
$>$ 479 000 entries

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



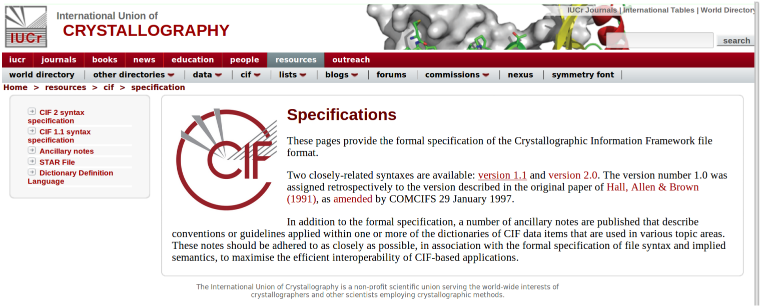
http://mpod.cimav.edu.mx/
$>$ 300 entries

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 CIF framework



Specifications

These pages provide the formal specification of the Crystallographic Information Framework file format.

Two closely-related syntaxes are available: version 1.1 and version 2.0. The version number 1.0 was assigned retrospectively to the version described in the original paper of Hall, Allen & Brown (1991), as amended by COMCIFS 29 January 1997.

In addition to the formal specification, a number of ancillary notes are published that describe conventions or guidelines applied within one or more of the dictionaries of CIF data items that are used in various topic areas. These notes should be adhered to as closely as possible, in association with the formal specification of file syntax and implied semantics, to maximise the efficient interoperability of CIF-based applications.

The International Union of Crystallography is a non-profit scientific union serving the world-wide interests of crystallographers and other scientists employing crystallographic methods.

[Hall et al. (1991)]

The Crystallographic Interchange File/Framework (CIF):

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

Example of a CIF file

examples/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_angle_alpha                90.0
_cell_angle_beta                 90.0
_cell_angle_gamma                90.0
_cell_formula_units_Z            1
_cell_length_a                   3.9998(8)
_cell_length_b                   3.9998(8)
_cell_length_c                   4.0180(8)
```

CIF atomic coordinates

examples/2100858-coordinates.cif:

```
loop_
_atom_site_label
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
Ba 0.0 0.0 0.0 ?
Ti 0.5 0.5 0.4820(10) ?
O1 0.5 0.5 0.016(5) ?
O2 0.5 0.0 0.515(3) ?
```

Controlled vocabularies, ontologies

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_.
;
```

COD data management principles

- Strictly stick to IUCr standards (CIF syntax, dictionaries);

COD data management principles

- Strictly stick to IUCr standards (CIF syntax, dictionaries);
- Do not invent data;

COD data management principles

- Strictly stick to IUCr standards (CIF syntax, dictionaries);
- Do not invent data;
- Better to have no data than wrong data;

COD data management principles

- Strictly stick to IUCr standards (CIF syntax, dictionaries);
- Do not invent data;
- Better to have no data than wrong data;
- Consult original papers or authors themselves if in doubt;

COD data management principles

- Strictly stick to IUCr standards (CIF syntax, dictionaries);
- Do not invent data;
- Better to have no data than wrong data;
- Consult original papers or authors themselves if in doubt;
- Document: record and explain (justify) all changes;

COD data management principles

- Strictly stick to IUCr standards (CIF syntax, dictionaries);
- Do not invent data;
- Better to have no data than wrong data;
- Consult original papers or authors themselves if in doubt;
- Document: record and explain (justify) all changes;
- Keep track of all changes in a version control system;

COD data management principles

- Strictly stick to IUCr standards (CIF syntax, dictionaries);
- Do not invent data;
- Better to have no data than wrong data;
- Consult original papers or authors themselves if in doubt;
- Document: record and explain (justify) all changes;
- Keep track of all changes in a version control system;
- Keep data provenance (original file names);

COD data validation

COD data validation policies:

1. Syntactic checks [Merkys et al. (2016)]:

   ```
   $ cifparse 7234818.cif
   ```

2. Semantic validation (against dictionaries) [Vaitkus et al. (2021)]:

   ```
   $ cif_validate -D cif_core.dic 7234818.cif
   ```

3. Database-specific checks [Gražulis et al. (2009)]:

   ```
   $ cif_cod_check 7234818.cif
   ```

COD data curation

Data curation in the COD:

```
svn log -r283960 --diff svn://www.crystallography.net/cod/cif/9
```

```
--- 00/15/9001556.cif (revision 283959)
+++ 00/15/9001556.cif (revision 283960)
@@ -68,8 +68,24 @@
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
{+_atom_site_type_symbol+}
{+_atom_site_attached_hydrogens+}
Fe 0.25000 0.25000 0.25000 0.00490 {+Fe 0+}
O-H1 0.50000 0.17800 0.30800 0.00100 {+O 1+}
O-H2 0.19500 0.19000 0.50000 0.00100 {+O 1+}
O-H3 0.31800 0.50000 0.32300 0.00100 {+O 1+}
Wat 0.00000 0.50000 0.50000 0.00640 {+O 2+}
/.../
```

COD query examples

Web, REST, SQL

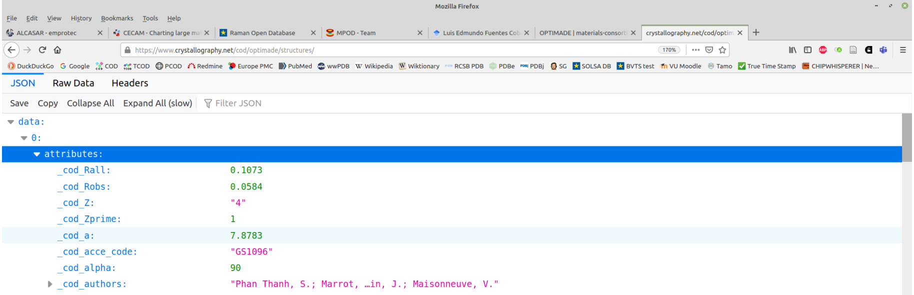
- Via the WWW interface – go for “search” in:
  - http://www.crystallography.net/cod
  - http://www.crystallography.net/tcod
  - http://www.crystallography.net/pcod
- Via the **stable** URLs (REST):
  - http://www.crystallography.net/cod/2000000.cif
  - http://www.crystallography.net/cod/2000000.html
  - http://www.crystallography.net/cod/result?text=perovskite
- Via the **views** of the SQL database:
  -
    ```
    mysql -u cod_reader cod -h sql.crystallography.net\
        -e 'select file, a, b, c, vol, formula
            from data where
                year between 2013 and
                             2014 and
                formula regexp " C[0-9]* "
                order by vol desc limit 10'
    ```

Yes, we OPTIMADE!

http://optimade.org/ [Andersen et al. (2021)]

OPTIMADE
Open Databases Integration
for Materials Design

http://www.crystallography.net/cod/optimade/v1/structures/


Mozilla Firefox
File Edit View History Bookmarks Tools Help
ALCASAR - emprotec
CECAM - Charting large ma
Raman Open Database
MPOD - Team
Luis Edmundo Fuentes Cob
OPTIMADE | materials-consort
crystallography.net/cod/optim
https://www.crystallography.net/cod/optimade/structures/
170%
DuckDuckGo Google COD TCOD PCOD Redmine Europe PMC PubMed wwPDB Wikipedia Wiktionary RCSB PDB PDBe PDBj SG SOLSA DB BVTS test VU Moodle Tamo True Time Stamp CHIPWHISPERER | Ne...
JSON Raw Data Headers
Save Copy Collapse All Expand All (slow) Filter JSON

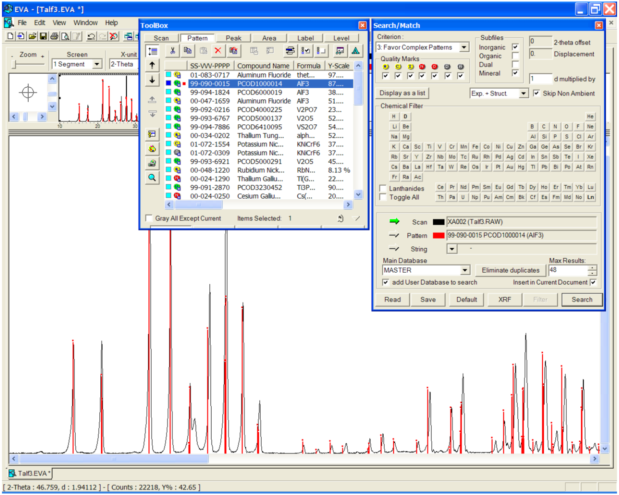
```
▾ data:
  ▾ 0:
    ▾ attributes:
        _cod_Rall:          0.1073
        _cod_Robs:          0.0584
        _cod_Z:             "4"
        _cod_Zprime:        1
        _cod_a:             7.8783
        _cod_acce_code:     "GS1096"
        _cod_alpha:         90
      ▸ _cod_authors:       "Phan Thanh, S.; Marrot, …in, J.; Maisonneuve, V."
```

Contents

1. The value of crystallographic data
2. Crystallographic data(bases): COD, TCOD, PCOD, MPOD, ...
3. Applications of COD and sister databases
4. Mathematical considerations in crystal data processing

Use of COD and PCOD databases

Search-match identification of the materials



	SS-VVV-PPPP	Compound Name	Formula	Y-Scale
C	01-083-0717	Aluminum Fluoride	thet...	97....
U	99-090-0015	PCOD1000014	AlF3	87....
U	99-094-1824	PCOD6000019	AlF3	38....
*	00-047-1659	Aluminum Fluoride	AlF3	51....
U	99-092-0216	PCOD4000225	V2PO7	23....
U	99-093-6767	PCOD5000137	V2O5	52....
U	99-094-7886	PCOD6410095	VS2O7	54....
*	00-034-0202	Thallium Tung...	alph...	52....
C	01-072-1554	Potassium Nic...	KNiCrF6	37....
B	01-072-0309	Potassium Nic...	KNiCrF6	37....
U	99-093-6921	PCOD5000291	V2O5	45....
*	00-048-1220	Rubidium Nick...	RbN...	8.13 %
N	00-024-1290	Thallium Galliu...	Tl(G...	22....
U	99-091-2870	PCOD3230452	Ti3P...	90....
N	00-024-0250	Cesium Galliu...	Cs(...	20....

☐ Gray All Except Current   Items Selected: 1

Criterion : 3: Favor Complex Patterns

Quality Marks: ☑ * ☑ C ☑ I ☑ N ☑ Q ☑ D ☑ B

Subfiles: Inorganic ☑ Organic ☐ Dual ☐ Mineral ☑

0 2-theta offset

0. Displacement

1 d multiplied by

Exp. + Struct.   ☑ Skip Non Ambient

☐ Lanthanides

☐ Toggle All

Scan XA002 (Talf3.RAW)

Pattern 99-090-0015 PCOD1000014 (AlF3)

String -

Main Database: MASTER

Max Results: 48

☑ add User Database to search

Insert in Current Document ☑

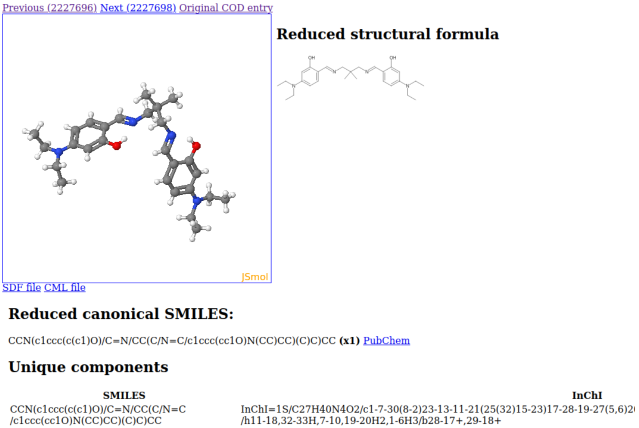
[2-Theta : 46.759, d : 1.94112] - [Counts : 22218, Y% : 42.65]

A **predicted** phase from PCOD could be identified in experimental data.

Courtesy Armel Le Bail [Le Bail(2008)]

COD chemical repertoire

https://molecules.crystallography.net/cod-molecules/cod/2227697.html



Previous (2227696) Next (2227698) Original COD entry

Reduced structural formula

SDF file CML file

Reduced canonical SMILES:

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

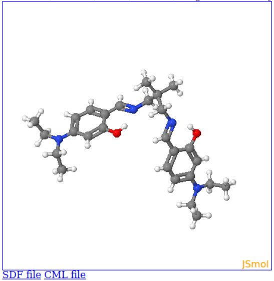
Unique components

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+

COD chemical repertoire

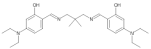
https://molecules.crystallography.net/cod-molecules/cod/2227697.html

Previous (2227696) Next (2227698) Original COD entry

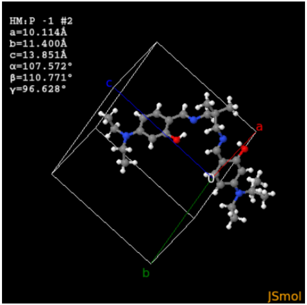


SDF file CML file

Reduced structural formula



A. Vaitkus
ms. ~~in preparation~~
under review



Reduced canonical SMILES:

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

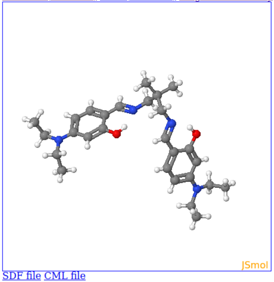
Unique components

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+

COD chemical repertoire

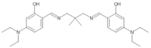
https://molecules.crystallography.net/cod-molecules/cod/2227697.html

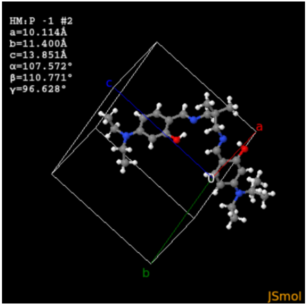
Previous (2227696) Next (2227698) Original COD entry



SDF file CML file

Reduced structural formula





A. Vaitkus
ms. ~~in preparation~~ under review

Reduced canonical SMILES:

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

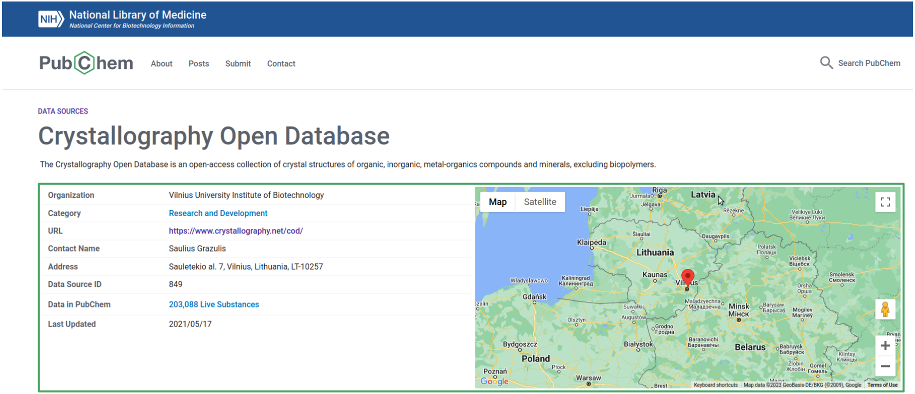
Unique components

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+

COD use cases

COD and PubChem

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



NIH National Library of Medicine
National Center for Biotechnology Information

PubChem About Posts Submit Contact Search PubChem

DATA SOURCES

Crystallography Open Database

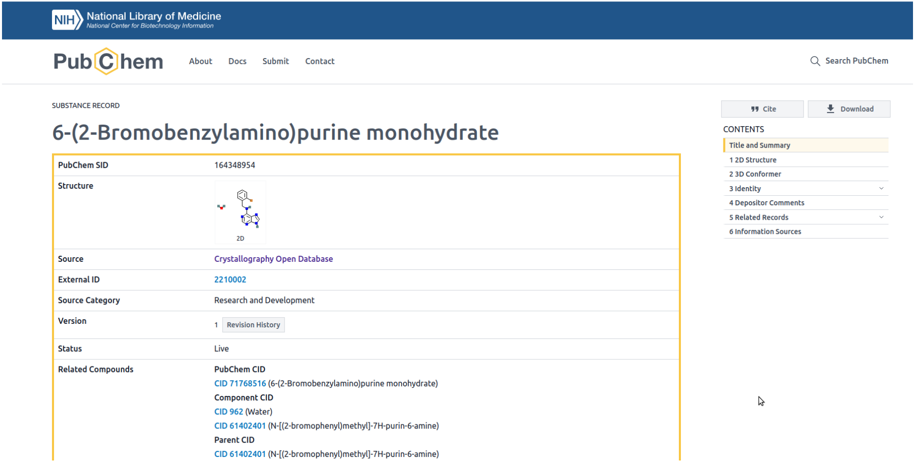
The Crystallography Open Database is an open-access collection of crystal structures of organic, inorganic, metal-organics compounds and minerals, excluding biopolymers.

Organization	Vilnius University Institute of Biotechnology
Category	Research and Development
URL	https://www.crystallography.net/cod/
Contact Name	Saulius Grazulis
Address	Sauletekio al. 7, Vilnius, Lithuania, LT-10257
Data Source ID	849
Data in PubChem	203,088 Live Substances
Last Updated	2021/05/17

COD use cases

COD and PubChem

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



NIH National Library of Medicine
National Center for Biotechnology Information
PubChem
About
Docs
Submit
Contact
Search PubChem
Cite
Download

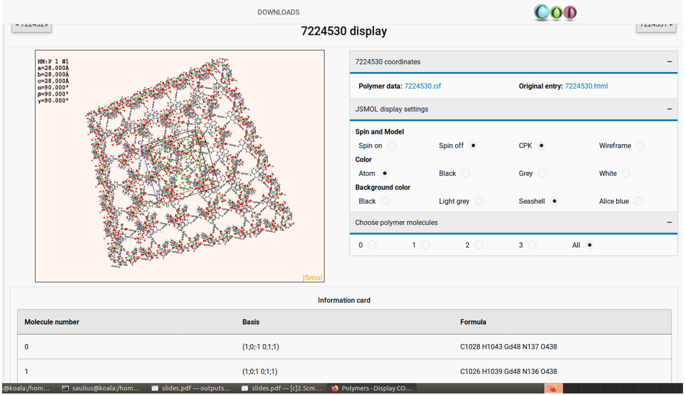
SUBSTANCE RECORD

6-(2-Bromobenzylamino)purine monohydrate

PubChem SID	164348954
Structure	2D
Source	Crystallography Open Database
External ID	2210002
Source Category	Research and Development
Version	1 Revision History
Status	Live
Related Compounds	**PubChem CID** CID 71768516 (6-(2-Bromobenzylamino)purine monohydrate) **Component CID** CID 962 (Water) CID 61402401 (N-[(2-bromophenyl)methyl]-7H-purin-6-amine) **Parent CID** CID 61402401 (N-[(2-bromophenyl)methyl]-7H-purin-6-amine)

COD data applications: polymer search

- polymers-in-COD: ≈ 400 000 COD records processed
- polymers of different dimensionality (1D, 2D, 3D, 1D-2D and so on) detected, ≈ 93 000 polymer records in total.



7224530 display

7224530 coordinates –

Polymer data: 7224530.cif **Original entry:** 7224530.html

JSMOL display settings –

Spin and Model

Spin on ○ Spin off ● CPK ● Wireframe ○

Color

Atom ● Black ○ Grey ○ White ○

Background color

Black ○ Light grey ○ Seashell ● Alice blue ○

Choose polymer molecules –

0 ○ 1 ○ 2 ○ 3 ○ All ●

Information card

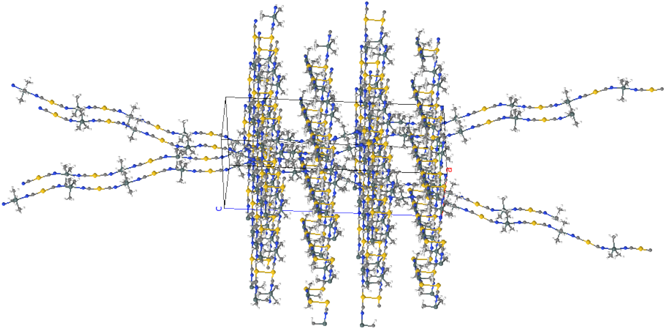
Molecule number	Basis	Formula
0	(1;0;-1 0;1;1)	C1028 H1043 Gd48 N137 O438
1	(1;0;1 0;1;1)	C1026 H1039 Gd48 N136 O438

http://crystallography.net/cod/7224530.html

results of A. Belova

COD data analysis: polymers

Find interpenetrating chains (crystal nets of covalent bonds):

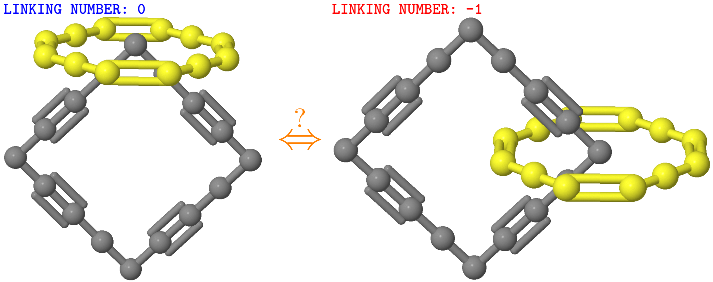


http://crystallography.net/cod/4103983.html

results of A. Belova, 2019

COD data analysis: search of knots and links

- Compute knot invariants, such as:
  - linking number;
  - Alexander and/or Conway polynomials;
  - etc. ...
- Use the set of invariants to distinguish links and knots.



```
# $Id: slides.tex 2298 2023-09-28 12:41:15Z saulius $
#@kw label b1_comp b1_a1 b1_a2 b2_comp b2_a1 b2_a2 sign filename
CROSS X1 C10 a26 a25 C11 a8 a9 -1 hopf-link-integer.cml
CROSS X2 C10 a20 a19 C11 a8 a9 -1 hopf-link-integer.cml
# COMPONENTS: hopf-link-integer.cml 2
# LINKING NUMBER: -1
```

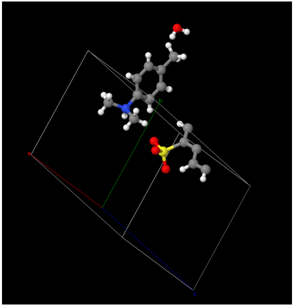
results of A. Belova, 2019

Contents

1. The value of crystallographic data
2. Crystallographic data(bases): COD, TCOD, PCOD, MPOD, …
3. Applications of COD and sister databases
4. Mathematical considerations in crystal data processing

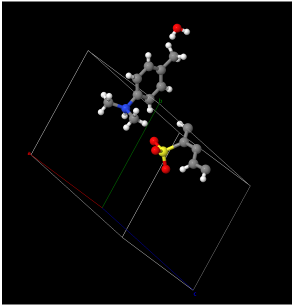
Contents of a crystallographic file

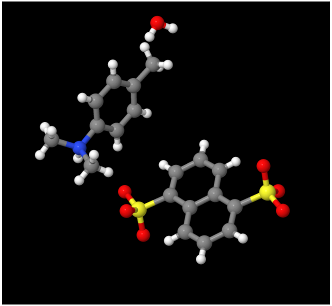
http://www.crystallography.net/cod/2231955.html



Contents of a crystallographic file

http://www.crystallography.net/cod/2231955.html





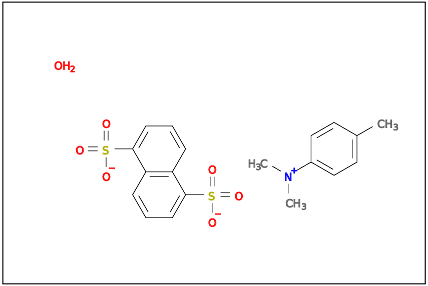
Reconstructing stoichiometric molecular ensemble

1. find a symmetry group S of each molecule;
2. find a symmetry group H the whole molecular ensemble;
3. find (left) coset decomposition of the crystal space group H by S, $S \trianglelefteq H$;
4. to each molecule, apply *one* symmetry element from each coset;
5. each choice of symmetry operations from the cosets (transversal) generates a *crystallographically identical* atom set present in the crystal;

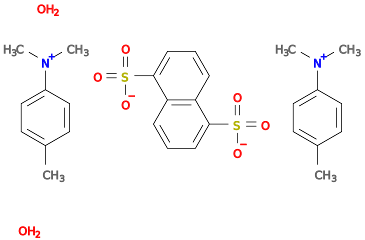
Reconstructing molecules from the COD

http://www.crystallography.net/cod/2231955.html

Usual algorithms:



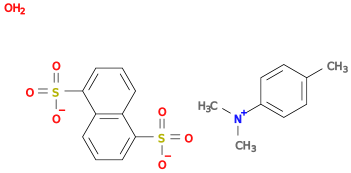
The new algorithm:



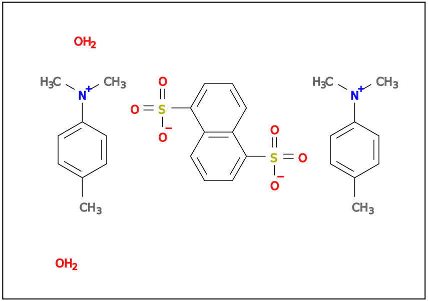
Reconstructing molecules from the COD

http://www.crystallography.net/cod/2231955.html

Usual algorithms:



The new algorithm:

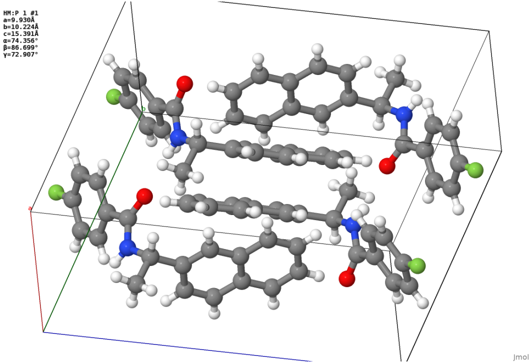


[Gražulis et al. (2015)]

A. Vaitkus, cif-perceive-chemistry (formerly cif2molecule) + OpenChemLib

Predicates on computed data

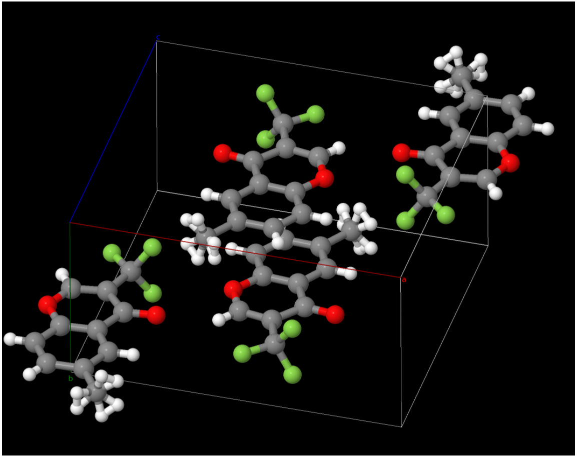
COD 7241595



$$\mathit{IsChiral}(M_1) \wedge \mathit{SymopMapsTo}(S, M_1, M_2) \wedge \det(S) = -1$$
$$\Rightarrow M_1 \ \mathit{IsEnantiomerOf} \ M_2$$

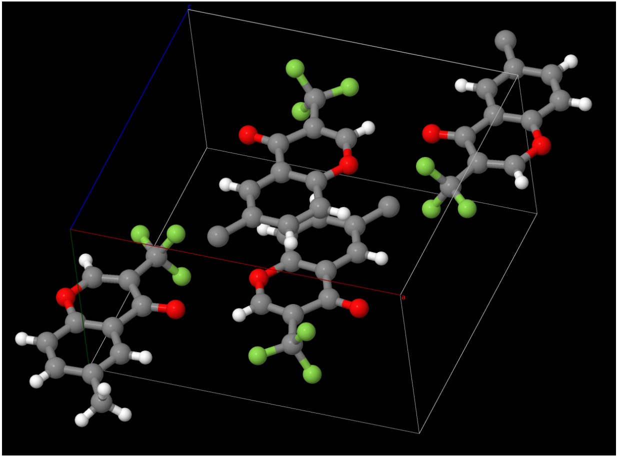
Disorder around a special position

COD 1544968 [Xiang et al. (2016)]



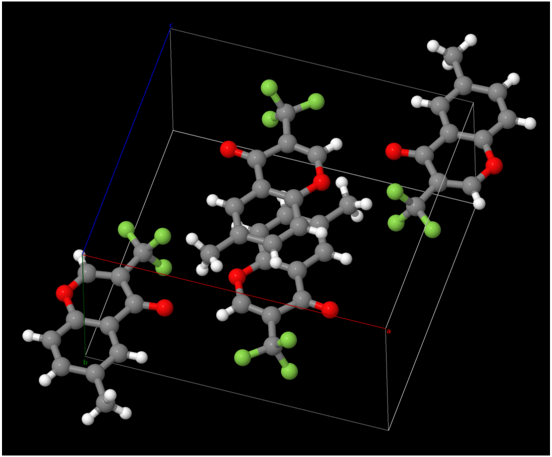
Disorder around a special position

COD 1544968 [Xiang et al. (2016)]



Disorder around a special position

COD 1544968 [Xiang et al. (2016)]



Generating a representative structure

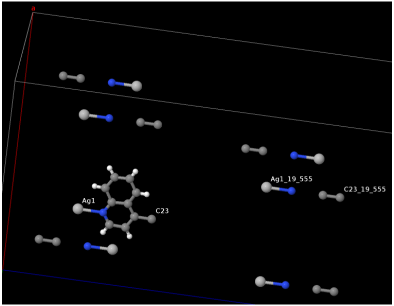
1. find a symmetry group S of a special position (Stabiliser);
2. find (left) coset decomposition of the crystal space group G by S, $S \trianglelefteq G$;
3. take *one* symmetry element from each coset and apply it to the disordered group;
4. each choice of symmetry operations from the cosets (transversal) generates a *distinct* atom set present in the crystal;

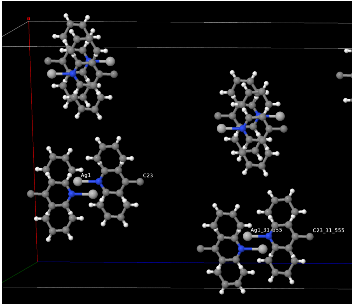
Disorder around a special position

COD entry 4111132

Original entry:

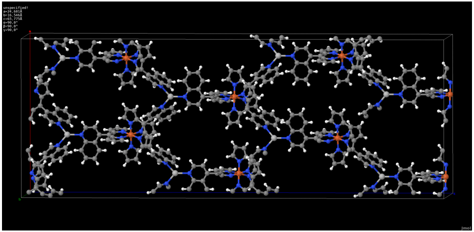
http://crystallography.net/cod/4111132.html



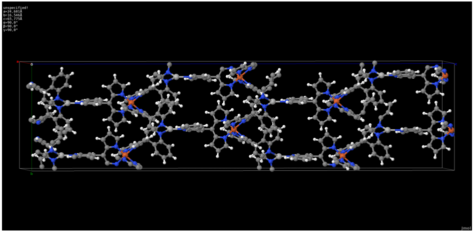


Disorder around a special position in polymers

COD 4111132 [Halper et al. (2006)]



video



video

Conclusions

- Data publication is as important as papers!
- Aggregated data allows new discoveries...
- ... but for this data need to be properly organised.
- COD, TCOD and the sister databases offer open data in crystallography.
- Mathematical insights are of paramount importance to understand crystal structures.
- Sharing data gives benefits to all.

Acknowledgements

VU LSC IBT (KICIS)

Andrius Merkys
Antanas Vaitkus
Algirdas Grybauskas

VU LSC IBT (BVTS)

Daumantas Matulis
Vytautas Petrauskas
Darius Lingė
Marius Gedgaudas

VU LSC IBT (BNSTS)

Mindaugas Zaremba
Elena Manakova

QM community

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

VU MIF II (FMG)

Linas Laibinis
Karolis Petrauskas
Haroldas Giedra

COD Advisory board

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

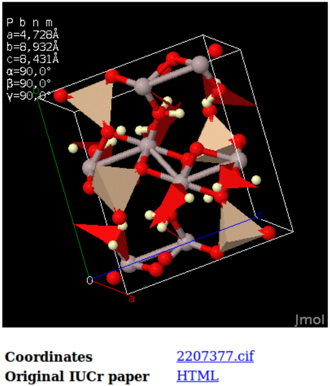
Cheminf community

Evan Bolton
Paul Thiessen
Thomas Sander

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

Thank you!

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



Coordinates	2207377.cif
Original IUCr paper	HTML

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

Slides available at:
https://www.crystallography.net/cod/archives/2023/slides/MIF++/slides.pdf

References I

Andersen CW, Armiento R, Blokhin E, Conduit GJ, Dwaraknath S, Evans ML, et al. (2021) OPTIMADE, an API for exchanging materials data. Scientific Data 8(1):1–10, DOI 10.1038/s41597-021-00974-z, URL https://doi.org/10.1038/s41597-021-00974-z

Fuentes-Cobas LE, Chateigner D, Fuentes-Montero ME, Pepponi G, Grazulis S (2017) The representation of coupling interactions in the Material Properties Open Database (MPOD). Advances in Applied Ceramics 116(8):428–433, DOI 10.1080/17436753.2017.1343782, URL https://doi.org/10.1080/17436753.2017.1343782

Gražulis S, Chateigner D, Downs RT, Yokochi AFT, Quirós M, Lutterotti L, et al. (2009) Crystallography Open Database – an open-access collection of crystal structures. Journal of Applied Crystallography 42:726–729, DOI 10.1107/S0021889809016690, URL http://dx.doi.org/10.1107/S0021889809016690

Gražulis S, Daškevič A, Merkys A, Chateigner D, Lutterotti L, Quirós M, et al. (2012) Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. Nucleic Acids Research 40:D420–D427, DOI 10.1093/nar/gkr900, URL http://nar.oxfordjournals.org/content/40/D1/D420.abstract

References II

- Gražulis S, Merkys A, Vaitkus A, Okulič-Kazarinas M (2015) Computing stoichiometric molecular composition from crystal structures. Journal of Applied Crystallography 48:85–91, DOI 10.1107/S1600576714025904, URL http://dx.doi.org/10.1107/S1600576714025904

- Hall SR, Allen FH, Brown ID (1991) The crystallographic information file (CIF): a new standard archive file for crystallography. Acta Crystallographica Section A 47:655–685, DOI 10.1107/S010876739101067X, URL http://dx.doi.org/10.1107/S010876739101067X

- Halper SR, Do L, Stork JR, Cohen SM (2006) Topological control in heterometallic metal-organic frameworks by anion templating and metalloligand design. Journal of the American Chemical Society 128(47):15,255–15,268, DOI 10.1021/ja0645483, URL https://doi.org/10.1021/ja0645483

- Katsura Y, Kumagai M, Kodani T, Kaneshige M, Ando Y, Gunji S, et al. (2019) Data-driven analysis of electron relaxation times in PbTe-type thermoelectric materials. Science and Technology of Advanced Materials 20(1):511–520, DOI 10.1080/14686996.2019.1603885, URL https://doi.org/10.1080/14686996.2019.1603885

- Le Bail A (2008) Frontiers between crystal-structure prediction and determination by powder diffractometry. Powder Diffraction Suppl pp S5–S12, DOI 10.1154/1.2903488, URL https://doi.org/10.1154/1.2903488

References III

- Mendili YE, Vaitkus A, Merkys A, Gražulis S, Chateigner D, Mathevet F, et al. (2019) Raman Open Database: first interconnected Raman–X-ray diffraction open-access resource for material identification. Journal of Applied Crystallography 52(3):618–625, DOI 10.1107/s1600576719004229, URL `https://doi.org/10.1107/s1600576719004229`

- Merkys A, Vaitkus A, Butkus J, Okulič-Kazarinas M, Kairys V, Gražulis S (2016) *COD::CIF::Parser*: an error-correcting CIF parser for the Perl language. Journal of Applied Crystallography 49(1):292–301, DOI 10.1107/S1600576715022396, URL `http://dx.doi.org/10.1107/S1600576715022396`

- Pepponi G, Gražulis S, Chateigner D (2012) MPOD: A Material Property Open Database linked to structural information. Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms 284(0):10–14, DOI 10.1016/j.nimb.2011.08.070, URL `http://www.sciencedirect.com/science/article/pii/S0168583X11008639`, e-MRS 2011 Spring Meeting, Symposium M: X-ray techniques for materials research-from laboratory sources to free electron lasers

References IV

- Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. (2020) Improved protein structure prediction using potentials from deep learning. Nature 577(7792):706–710, DOI 10.1038/s41586-019-1923-7, URL `https://doi.org/10.1038/s41586-019-1923-7`, `https://doi.org/10.1038/s41586-019-1923-7`

- Vaitkus A, Merkys A, Gražulis S (2021) Validation of the Crystallography Open Database using the Crystallographic Information Framework. Journal of Applied Crystallography 54(2):1–12, DOI 10.1107/s1600576720016532, URL `https://doi.org/10.1107/S1600576720016532`

- Xiang H, Zhao Q, Tang Z, Xiao J, Xia P, Wang C, et al. (2016) Visible-light-driven, radical-triggered tandem cyclization of o-hydroxyaryl enaminones: Facile access to 3-CF2 /CF3-containing chromones. Organic Letters 19(1):146–149, DOI 10.1021/acs.orglett.6b03441, URL `https://doi.org/10.1021/acs.orglett.6b03441`