Sharing scientific data: the crystallography experience

Saulius Gražulis

Vilnius, 2023

Vilnius University Institute of Biotechnology

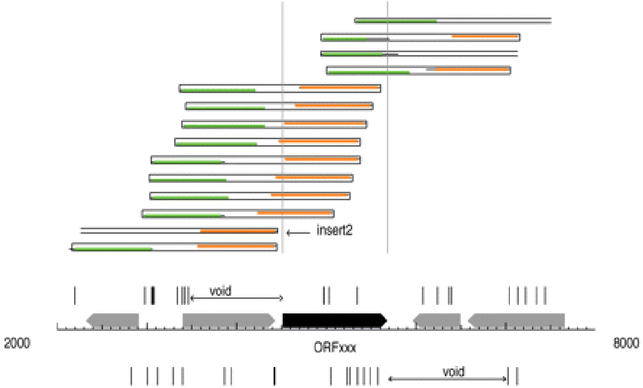
Id: slides.tex 2193 2023-05-25 06:23:18Z saulius May 25, 2023

Layout of the talk

1. Scientific data: volumes and uses
2. Crystallographic data(bases)
3. Data organisation principles

Discoveries in raw data

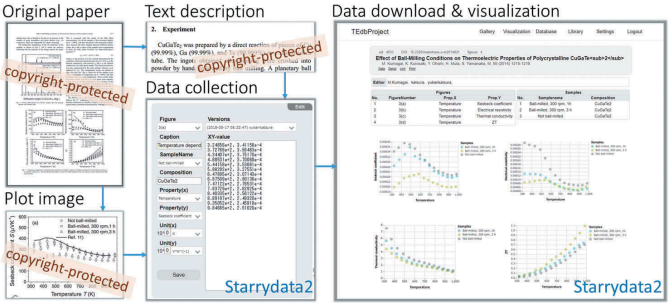
Zheng from the team of Roberts (NEB) use raw sequencing read data to discover *active* restriction-modification systems: [Zheng et al., 2008]:



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 ... milling. A planetary ball

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/

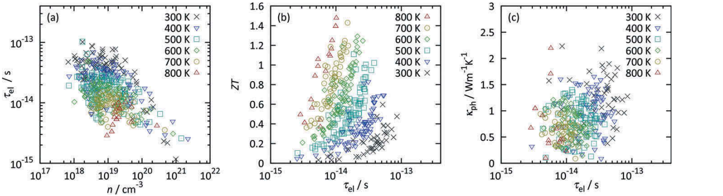


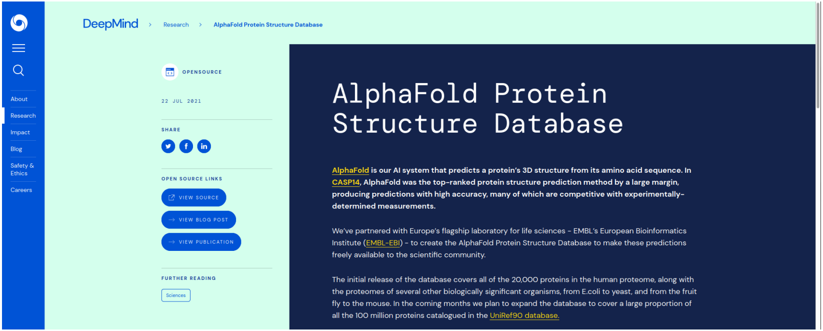
Figure 6. Relationship between (a) carrier doping level n and electron relaxation time τ_{el}, (b) τ_{el} and thermoelectric figure of merit ZT, and (c) τ_{el} and phonon thermal conductivity κ_{ph}, estimated for 207 experimental samples of n-type PbTe.

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

[Katsura et al., 2019]

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)

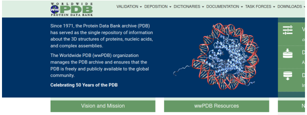
Crystallographic databases

Open Access:

Crystallographic databases

Open Access:

- Protein Data Bank;



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.

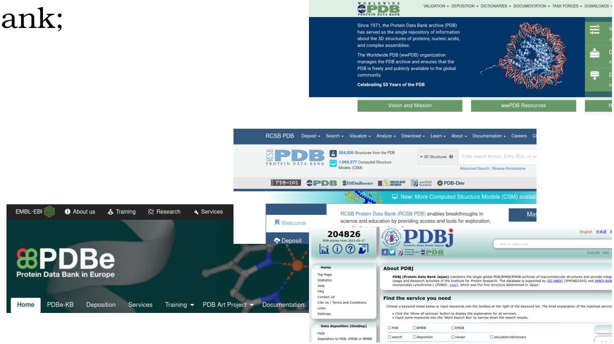
The Worldwide PDB (wwPDB) organization manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.

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
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 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.

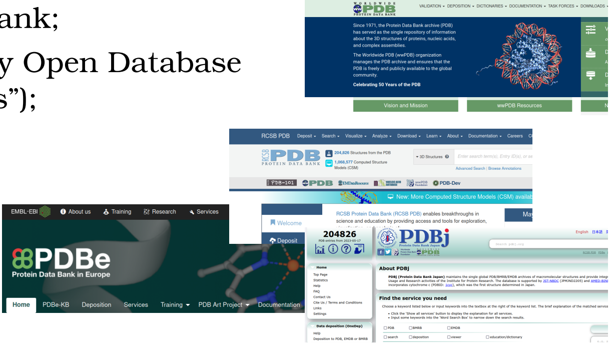
☐ PDB ☐ BMRB ☐ EMDB

☐ search ☐ deposition ☐ viewer ☐ education/dictionary

Crystallographic databases

Open Access:

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



WORLDWIDE PDB PROTEIN DATA BANK
VALIDATION ▾ DEPOSITION ▾ DICTIONARIES ▾ DOCUMENTATION ▾ TASK FORCES ▾ DOWNLOADS ▾

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

Vision and Mission
wwPDB Resources
RCSB PDB Deposit ▾ Search ▾ Visualize ▾ Analyze ▾ Download ▾ Learn ▾ About ▾ Documentation ▾ Careers
RCSB PDB PROTEIN DATA BANK

204,826 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

▾ 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
New: More Computed Structure Models (CSM) availab
Welcome
Deposit
May

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.

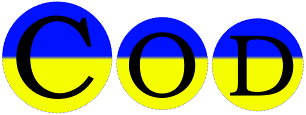
Find the service you need

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

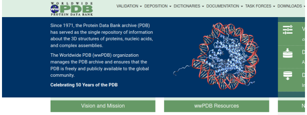
☐ search ☐ deposition ☐ viewer ☐ education/dictionary



Crystallographic databases

Open Access:

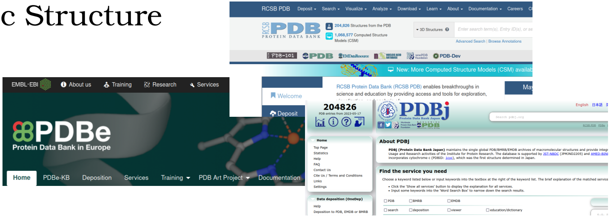
- Protein Data Bank;
- Crystallography Open Database (and its “sisters”);
- Bilbao Magnetic Structure Database



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.

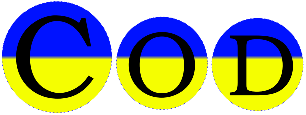
Find the service you need

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”);
- Bilbao Magnetic Structure Database

Proprietary:

- CCDC
- ICSD
- PDF
- Pauling File
- …

WORLDWIDE PDB PROTEIN DATA BANK
VALIDATION ▾ DEPOSITION ▾ DICTIONARIES ▾ DOCUMENTATION ▾ TASK FORCES ▾ DOWNLOADS ▾

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

Vision and Mission
wwPDB Resources
RCSB PDB Deposit ▾ Search ▾ Visualize ▾ Analyze ▾ Download ▾ Learn ▾ About ▾ Documentation ▾ Careers
RCSB PDB PROTEIN DATA BANK

204,826 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

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

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,

Welcome
Deposit
EMBL-EBI About us Training Research Services
PDBe Protein Data Bank in Europe
Home PDBe-KB Deposition Services Training ▾ PDB Art Project ▾ Documentation
204826
PDB entries from 2023-05-17
PDBj Protein Data Bank Japan
English 日本語
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

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

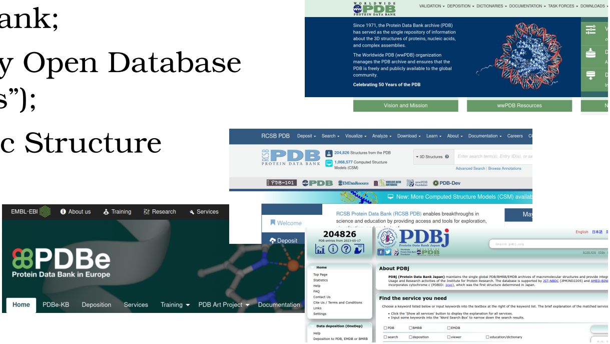
Crystallographic databases

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 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

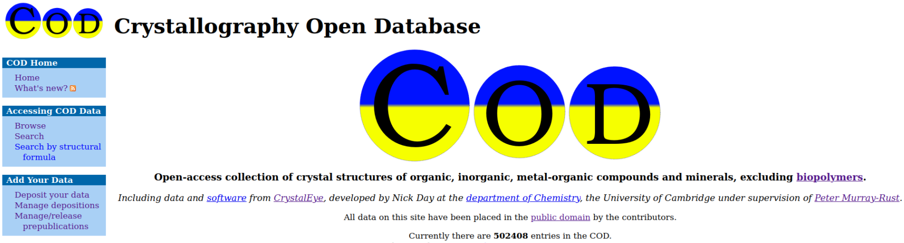
☐ search ☐ deposition ☐ viewer ☐ education/dictionary

COD

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

The Crystallography Open Database (COD)

https://www.crystallography.net
Online since 2003 :)



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

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.

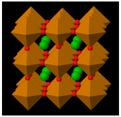
All data on this site have been placed in the public domain by the contributors.

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

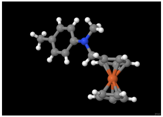
$>$ **500 000** records as of 2023-05-22, available under CC0 License

COD contents

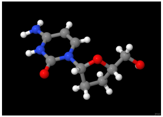
The Crystallography Open Database (COD)
https://www.crystallography.net



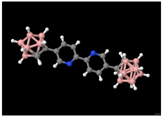
Inorganic



Metal-organic
and organometallic

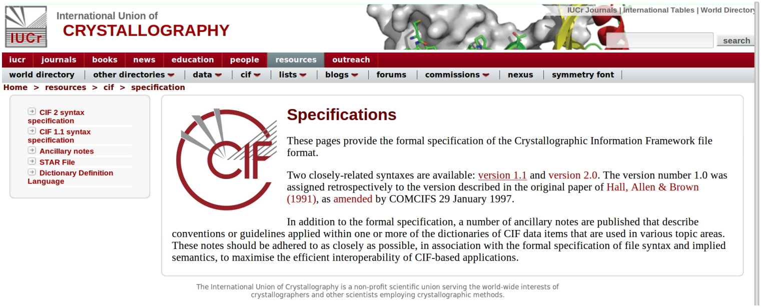


Organic



“Element-organic”

The CIF framework



International Union of
CRYSTALLOGRAPHY
IUCr
IUCr Journals | International Tables | World Directory
search
iucr journals books news education people resources outreach
world directory | other directories | data | cif | lists | blogs | forums | commissions | nexus | symmetry font |
Home > resources > cif > specification
CIF 2 syntax specification
CIF 1.1 syntax specification
Ancillary notes
STAR File
Dictionary Definition Language
CIF

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

CIF (Crystallographic Interchange Framework/Format)

```
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)
```

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);

Three levels of data validation

- Check of file syntax;
- Validation against dictionaries;
- Domain-specific checks:
  - internal consistency;
  - coherence with raw data;
  - scientific plausibility;

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

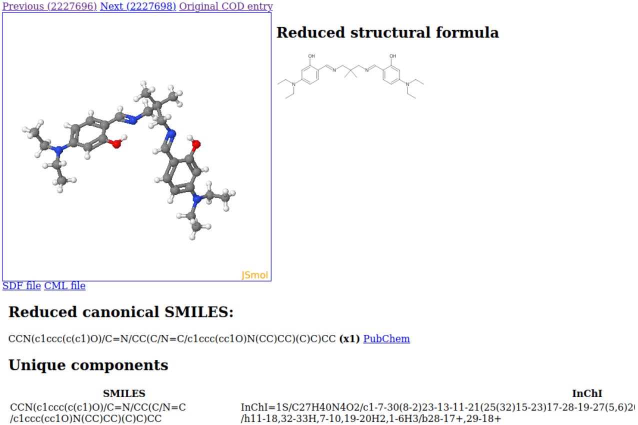
Data curation in the COD:

```
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 chemical repertoire

https://molecules.crystallography.net/~saulius/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

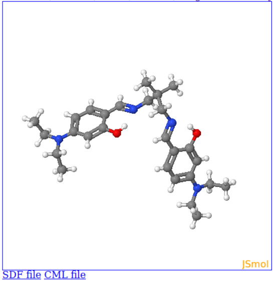
SMILES	InChI
CCN(c1ccc(c(c1)O)/C=N/CC(C/N=C /c1ccc(cc1O)N(CC)CC)(C)C)CC	InChI=1S/C27H40N4O2/c1-7-30(8-2)23-13-11-21(25(32)15-23)17-28-19-27(5,6)2 /h11-18,32-33H,7-10,19-20H2,1-6H3/b28-17+,29-18+

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

COD chemical repertoire

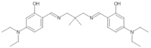
https://molecules.crystallography.net/~saulius/cod-molecules/cod/2227697.html

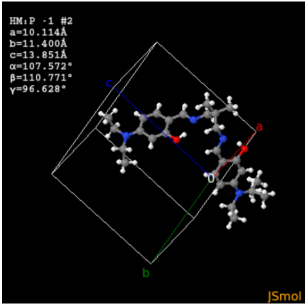
Previous (2227696) Next (2227698) Original COD entry



SDF file CML file

Reduced structural formula





A. Vaitkus
ms. in
preparation

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

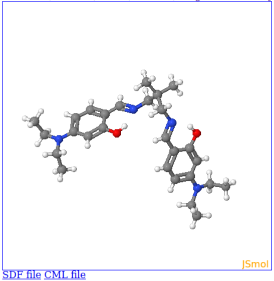
SMILES	InChI
CCN(c1ccc(c(c1)O)/C=N/CC(C/N=C /c1ccc(cc1O)N(CC)CC)(C)C)CC	InChI=1S/C27H40N4O2/c1-7-30(8-2)23-13-11-21(25(32)15-23)17-28-19-27(5,6)2 /h11-18,32-33H,7-10,19-20H2,1-6H3/b28-17+,29-18+

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

COD chemical repertoire

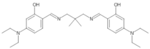
https://molecules.crystallography.net/~saulius/cod-molecules/cod/2227697.html

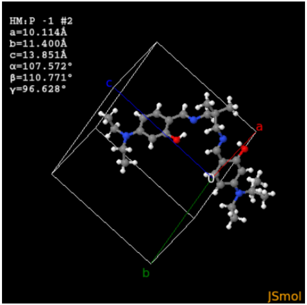
Previous (2227696) Next (2227698) Original COD entry



SDF file CML file

Reduced structural formula





A. Vaitkus
ms. in
preparation

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

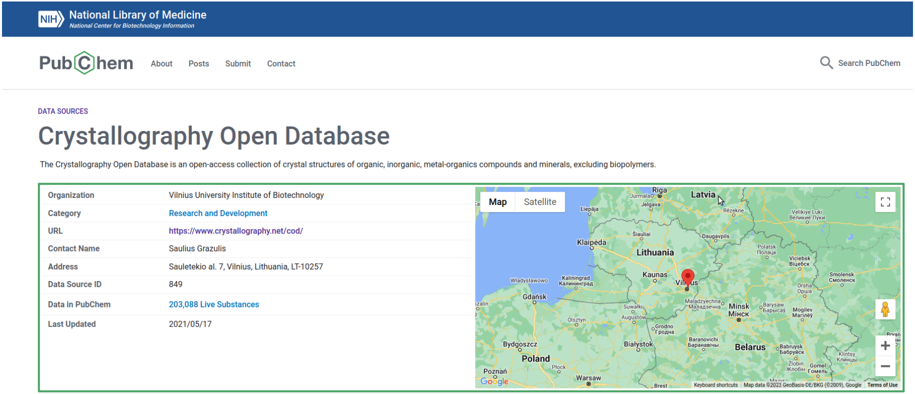
SMILES	InChI
CCN(c1ccc(c(c1)O)/C=N/CC(C/N=C /c1ccc(cc1O)N(CC)CC)(C)C)CC	InChI=1S/C27H40N4O2/c1-7-30(8-2)23-13-11-21(25(32)15-23)17-28-19-27(5,6)2 /h11-18,32-33H,7-10,19-20H2,1-6H3/b28-17+,29-18+

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

COD use cases

COD and PubChem

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



National Library of Medicine
National Center for Biotechnology Information

PubChem About Posts Submit Contact Search PubChem

DATA SOURCES

Crystallography Open Database

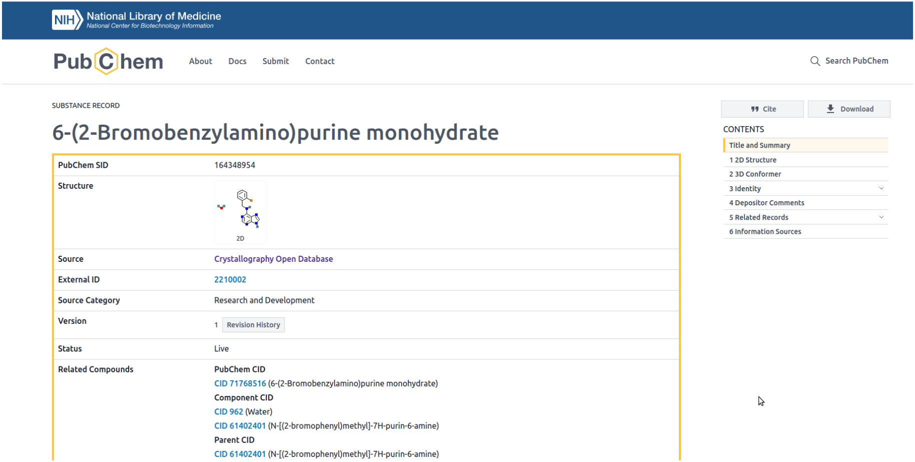
The Crystallography Open Database is an open-access collection of crystal structures of organic, inorganic, metal-organics compounds and minerals, excluding biopolymers.

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

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)

Cite Download

CONTENTS

- Title and Summary
- 1 2D Structure
- 2 3D Conformer
- 3 Identity
- 4 Depositor Comments
- 5 Related Records
- 6 Information Sources

Conclusions

- Data publication is as important as papers!
- Aggregated data allows new discoveries...
- ... but for this data need to be properly organised.
- Sharing data gives benefits to all.
- **Your contribution is important!**

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

$\boxed{\text{Audrius Alkauskas}}$
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

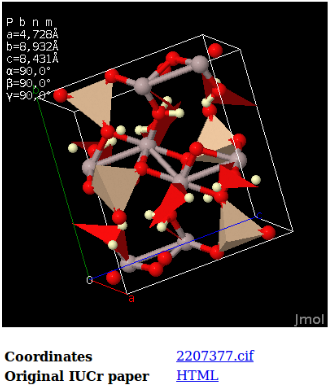
Cheminf community

Evan Bolton

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/JSMC-2023/slides.pdf

References I

Gražulis, S., Chateigner, D., Downs, R. T., Yokochi, A. F. T., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P., and Le Bail, A. (2009).
Crystallography Open Database – an open-access collection of crystal structures.
Journal of Applied Crystallography, 42:726–729.

Hall, S. R., Allen, F. H., and Brown, I. D. (1991).
The crystallographic information file (CIF): a new standard archive file for crystallography.
Acta Crystallographica Section A, 47:655–685.

Katsura, Y., Kumagai, M., Kodani, T., Kaneshige, M., Ando, Y., Gunji, S., Imai, Y., Ouchi, H., Tobita, K., Kimura, K., and Tsuda, K. (2019).
Data-driven analysis of electron relaxation times in PbTe-type thermoelectric materials.
Science and Technology of Advanced Materials, 20(1):511–520.

Merkys, A., Vaitkus, A., Butkus, J., Okulič-Kazarinas, M., Kairys, V., and Gražulis, S. (2016).
COD::CIF::Parser: an error-correcting CIF parser for the Perl language.
Journal of Applied Crystallography, 49(1):292–301.

References II

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., and Hassabis, D. (2020).
Improved protein structure prediction using potentials from deep learning.
Nature, 577(7792):706–710.

Vaitkus, A., Merkys, A., and Gražulis, S. (2021).
Validation of the Crystallography Open Database using the Crystallographic Information Framework.
Journal of Applied Crystallography, 54(2):1–12.

Zheng, Y., Posfai, J., Morgan, R. D., Vincze, T., and Roberts, R. J. (2008).
Using shotgun sequence data to find active restriction enzyme genes.
Nucleic Acids Research, 37(1):e1–e1.