Raw data and external links in the Crystallography Open Database

Aušrinė Zeleckytė Andrius Merkys Antanas Vaitkus Saulius Gražulis

Poznań, 2025

Vilnius University, Life Sciences Centre, Institute of Biotechnology
Vilnius University, Faculty of Mathematics and Informatics, Institute of Informatics

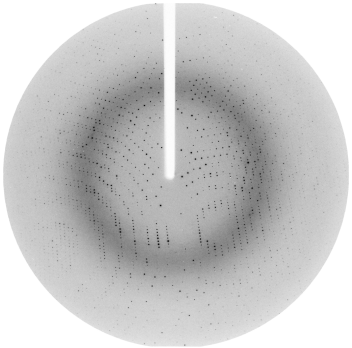
https://www.crystallography.net/cod/archives/2025/slides/2025-Poznan-ECM35/slides.pdf

Id: slides.tex 3115 2025-08-29 05:16:21Z saulius August 29, 2025

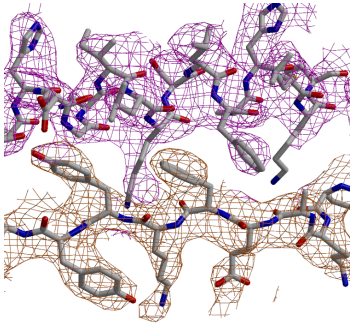
The crystallographic workflow

Data reduction steps

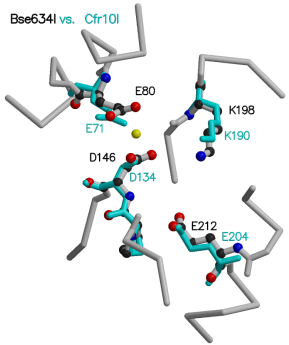
Detector signal → Iobs (Fobs) data → Atomic coordinates



⇒



⇒



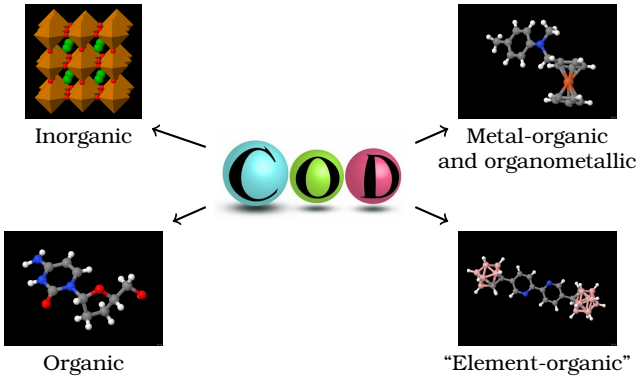
Most crystallographic databases store the last two. Today, we start linking with the first one.

What is the Crystallography Open Database (COD)

A brief intro

(Gražulis, Chateigner, et al. 2009; Gražulis, Daškevič, et al. 2012)

https://www.crystallography.net

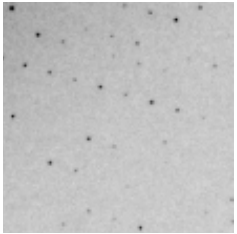


As of 2025-08-22:

527276 entries

Value of raw data

Verification, improvement, unintended reuse



- Verify and validate the findings;
- Ensure reproducible computations;
- Reuse raw data for (unintended) purposes (multiple lattices, sattelite reflections, non-Bragg scattering, machine studies, reflection shape...);
- Improve structures with better software/methdods;
- Teaching :)

Challenges with the raw data

Size, formats, net traffic

- Volume of raw data (orders of magnitude larger than even F_{obs})
- Many different data formats, some proprietary;
  - Good news: IUCr CBF
  - Good news: NeXUS & the Gold Standard (Bernstein et al. 2020)!
- Long transfer times;
- Political reasons...

Storages of XRD data

World wide repositories

- ProteinDiffraction;
- SBGrid;
- XRDa;
- Zenodo (?!) (e.g. https://zenodo.org/records/15992324, https://zenodo.org/records/12794355)
- National and institutional archives (?!)

The COD SQL data schema + COD CIF dictionary;

data

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

XRDa_x_COD

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

relations

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

Similar schemas: Star schema:
(Apeh 2024)

Example of the COD data record raw data link

XRDa as an example

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

COD Home
Home
What's new?

Accessing COD Data
Browse
Search
Search by structural formula

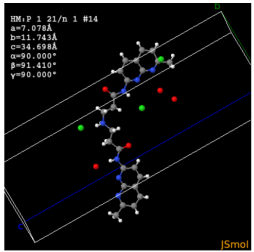
Add Your Data
Deposit your data
Manage depositions
Manage/release prepublications

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

Information card for entry 3000497

3000496 << **3000497** >> 3000498

Preview



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

▾ Structure parameters

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

(Sakurabayashi et al. 2025)

XRDa side

Links to the COD :)

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

Entry information

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

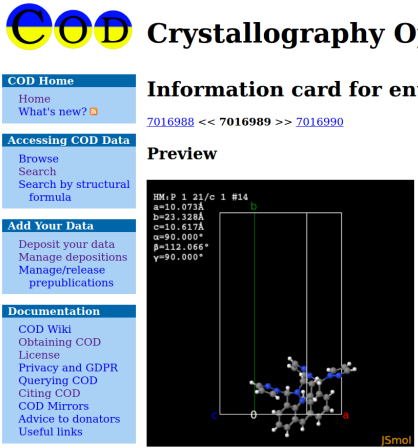
A side note: links with other data sources

Pubchem as an example

pubchem_x_cod

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

169 593 rows
(Vaitkus et al. 2023)



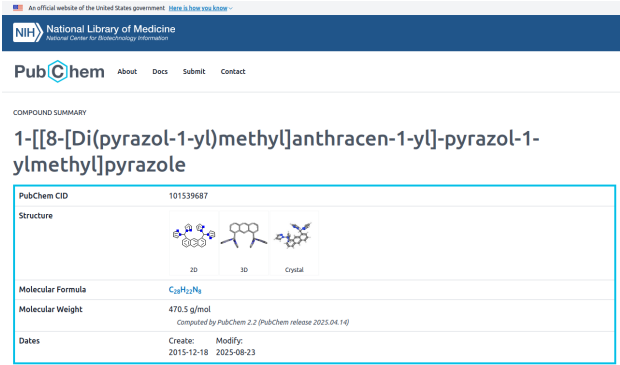
A side note: links with other data sources

Pubchem as an example

pubchem_x_cod

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

169 593 rows
(Vaitkus et al. 2023)



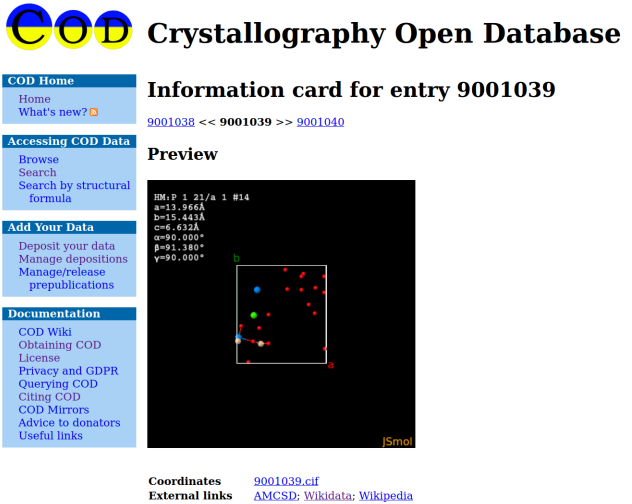
A side note: links with other data sources

Wikidata as an example

wikidata_x_cod

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

76 rows



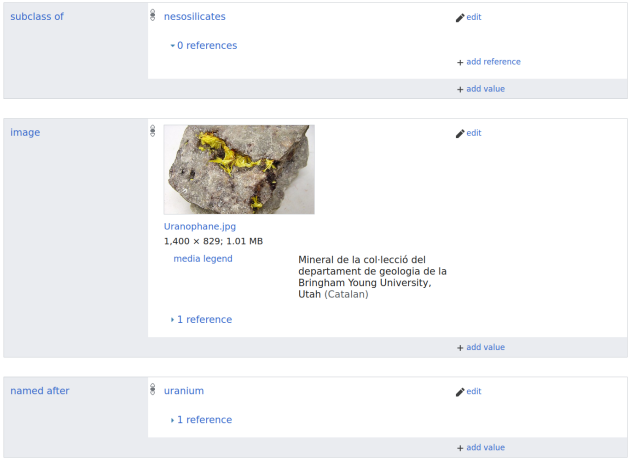
A side note: links with other data sources

Wikidata as an example

wikidata_x_cod

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

76 rows



Relation types

Different relations for different records

Relation “ontology”: the `relations` table;

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

CIF data source

Data flows from CIF to SQL (and never in reverse direction!)

From the COD 3000497 entry:

```
loop_
_cod_related_entry_id
_cod_related_entry_database
_cod_related_entry_code
_cod_related_entry_description
1 XRDA 00335 10.51093/xrd-00335
```

CIF data source

Data flows from CIF to SQL (and never in reverse direction!)

From the COD 3000497 entry:

```
loop_
_cod_related_entry_id
_cod_related_entry_database
_cod_related_entry_code
_cod_related_entry_description
1 XRDA 00335 10.51093/xrd-00335
```

From the COD 7016989 entry:

```
loop_
_cod_related_entry_id
_cod_related_entry_database
_cod_related_entry_code
1 AMCSD 0001061
2 Wikidata Q423494
3 Wikipedia Uranophane
```

Encoding in CIFs

Using COD CIF dictionary

```
data_cod_related_entry_code
_name                                  '_cod_related_entry_code'
_category                              cod_related_entry
_type                                  char
_list                                  both
# ...
_definition
;
 The unique identifier of the related entry assigned by the
 related database (see _cod_related_entry_database).
;
```

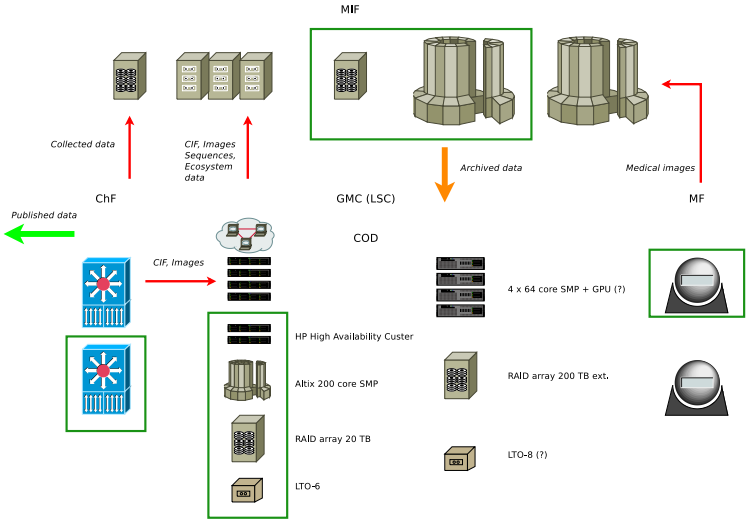
IUCr ImgCIF

Direct links to frames

https://github.com/COMCIFS/imgCIF.git

- Should in principle permit direct links to X-ray frames;
- Still work in progress (recent commit: `Date: Mon Apr 28 09:55:40 2025 +1000`);
- No tags in Git ...;
- Still in flux (already has deprecated items);

Infrastructure



Conclusions

- Simple link tables let us find Raw diffraction images;
- The method is general and suitable for other data as well;
- Predicates enable more precise descritpion of the links;
- More advanced methods that are under development (e.g. ImgCIF) are compatible with the proposed links;

Acknowledgements

VU LSC IBT (KICIS)

Andrius Merkys
Antanas Vaitkus
Algirdas Grybauskas

VU MIF II (FMG)

Linas Laibinis
Karolis Petrauskas
Irus Grinis
Haroldas Giedra

QM community

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

COD Advisory board

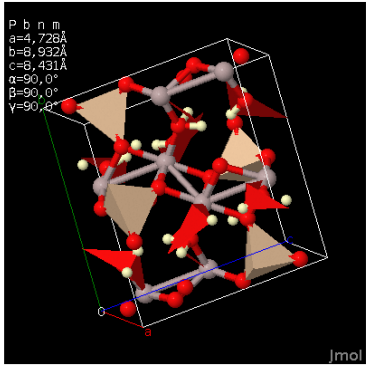
Daniel Chateigner
Robert T. Downs
Werner Kaminsky
Armel Le Bail
Luca Lutterotti
Peter Moeck
Peter Murray-Rust
Miguel Quirós

Funding:

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

Project IN-CHARGE (project number **191218**), funded by **Nordic Energy research** in the The joint Baltic-Nordic Energy Research programme.

Thank you!



Coordinates	2207377.cif
Original IUCr paper	HTML

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

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

References I

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

Bernstein, Herbert J. et al. (July 2020). "Gold Standard for macromolecular crystallography diffraction data". In: *IUCrJ* 7.5, pp. 784–792. ISSN: 2052-2525. DOI: 10.1107/s2052252520008672. URL: https://doi.org/10.1107/S2052252520008672.

Gražulis, Saulius, Daniel Chateigner, 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, Adriana Daškevič, 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.

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

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.