The Crystallography Open Database – new perspectives

Saulius Gražulis Andrius Merkys Antanas Vaitkus

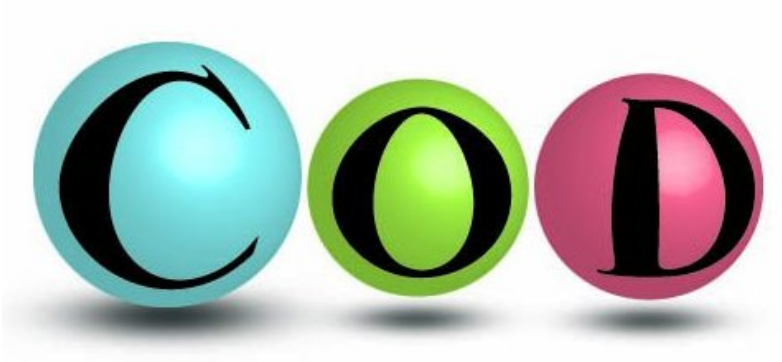
Leiden, CECAM 2016

Vilnius University Institute of Biotechnology

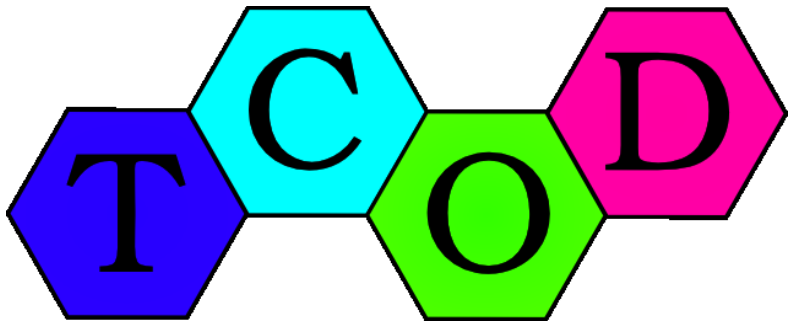
This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

Open Crystallographic Databases

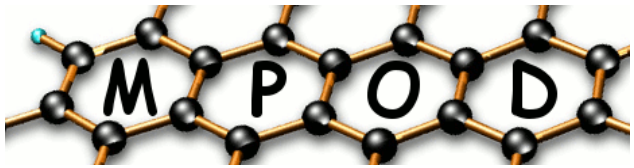
COD, TCOD, PCOD, MPOD, ...



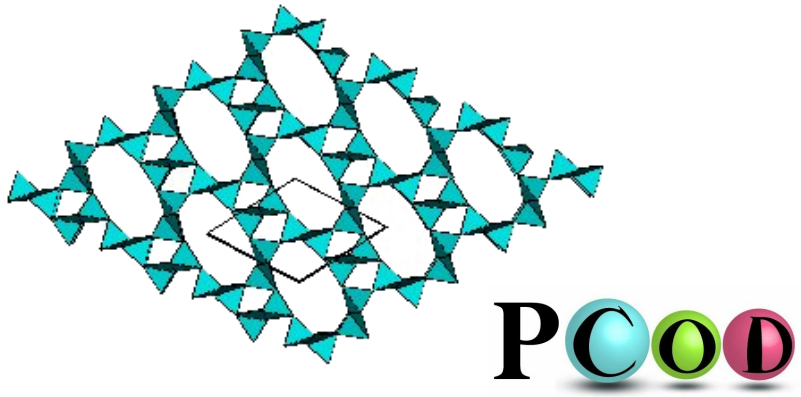
http://www.crystallography.net/cod
> 367 000 entries (ready to grow > 10^6?)



http://www.crystallography.net/tcod
> 2000 entries (ready to grow to > 350 000?)



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

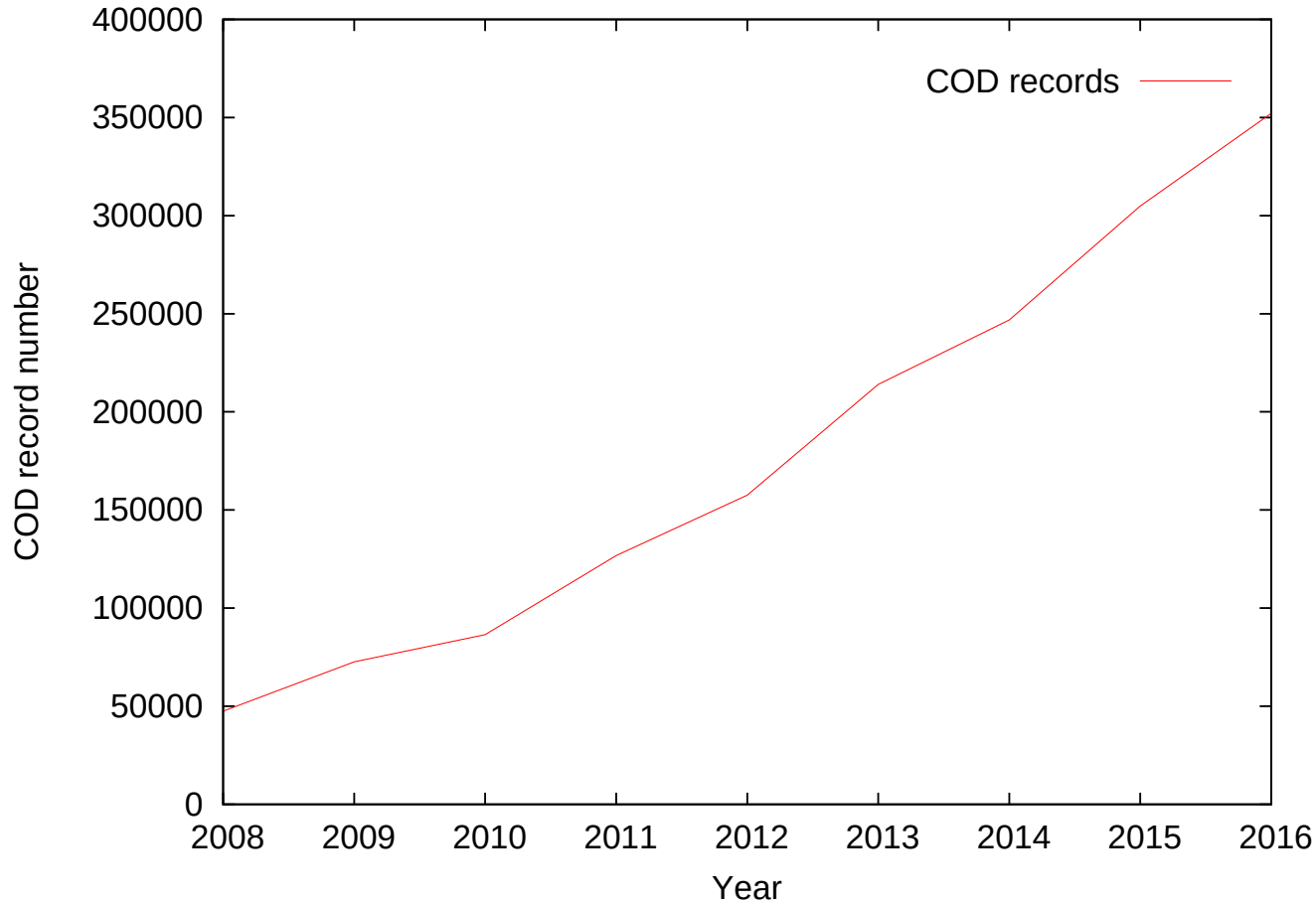


http://www.crystallography.net/pcod
> 10^6 entries (ready to grow to > 10^8?)

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

COD 13 years later

COD increased 7-fold; currently contains over 367000 records (Sept. 2016)



This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

Commong framework: the CIF

The Crystallographic Interchange Framework (CIF) is developed and curated by the International Union of Crystallography (IUCr).

`examples/data/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)
```

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

Description of semantics

CIF dictionaries

```
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_.
;
```

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

TCOD dictionary contents

The most basic data names

- `cif_tcod.dic`: ver. 0.008, last update 2015-06-16, 107 data names;
- `cif_dft.dic`: ver. 0.019, last update 2016-04-13, 87 data names.

e.g. (same as NOMAD atom_forces?):

```
data_tcod_atom_site_residual_force
loop_ _name
'_tcod_atom_site_resid_force_Cartn_x'
'_tcod_atom_site_resid_force_Cartn_y'
'_tcod_atom_site_resid_force_Cartn_z'
# ... some names omitted for brevity
_type numb
_units eV/\%A
_units_detail 'electronvolts per Angstroem'
_definition
;
These data items describe residual forces on atoms in the final
structure. For a converged computation of a stable structure these
...
;
```

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

New developments: CIF2

- Support of Unicode (UTF-8) [Bernstein et al., 2016];
- Array data (including multidimensional arrays);
- Data hashes (key–value pairs);
- Computer readable semantics definitions (in a multiparadigm language `dREL`):

```
    _units.code                         angstroms_cubed
    _method.expression
;
    With v as cell_vector
        _cell.volume = v.a * ( v.b ^ v.c )
;
```

http://oldwww.iucr.org/iucr-top/cif/ddlm/dREL_spec_20071013.html

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

COD accessibility

COD is a **fully open-access database**. All records are available under public domain designation.

Provided access methods are:

- Web search
- URLs constructed from stable identifiers
- RESTful interfaces
- Full data download

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

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/tcod/10000002.cif
  - http://www.crystallography.net/cod/result?text=perovskite
- Via the **views** of the SQL database:
  - ```
    mysql -u cod_reader cod -h www.crystallography.net\
        -e 'select file, a, b, c, vol, formula
            from data where
                date between "2013-01-01" and
                             "2014-12-31" and
                formula regexp " C[0-9]* "
                order by vol desc limit 10'
    ```

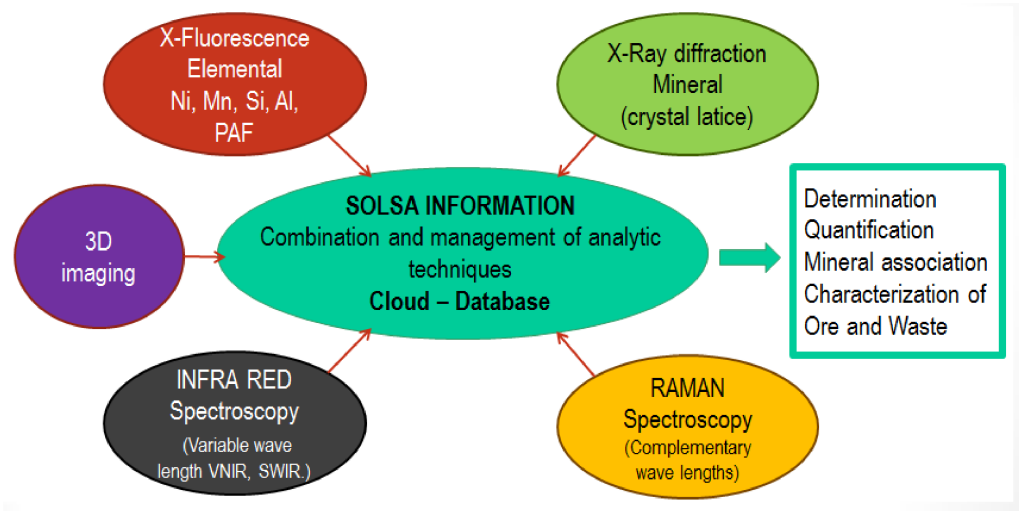
This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.
    ```

COD applications

- SOLSA
  - http://www.solsa-mining.eu/
- AiiDA [Pizzi et al., 2016]
  - http://www.aiida.net/
- COSMOS [Sadowski and Baldi, 2013]
  - http://cdb.ics.uci.edu/
- FPSM [Boullay et al., 2014], MAUD [Boullay et al., 2012]
  - http://fpsm.radiographema.com/
  - http://maud.radiographema.eu/
- DataWarrior
  - http://www.openmolecules.org/datawarrior/
- MolView
  - http://molview.org/
- search-match (Bruker, PANalytical, Rigaku)
- ... and more!

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

SOLSA project and COD



$\Leftrightarrow$

COD will be used in SOLSA for:

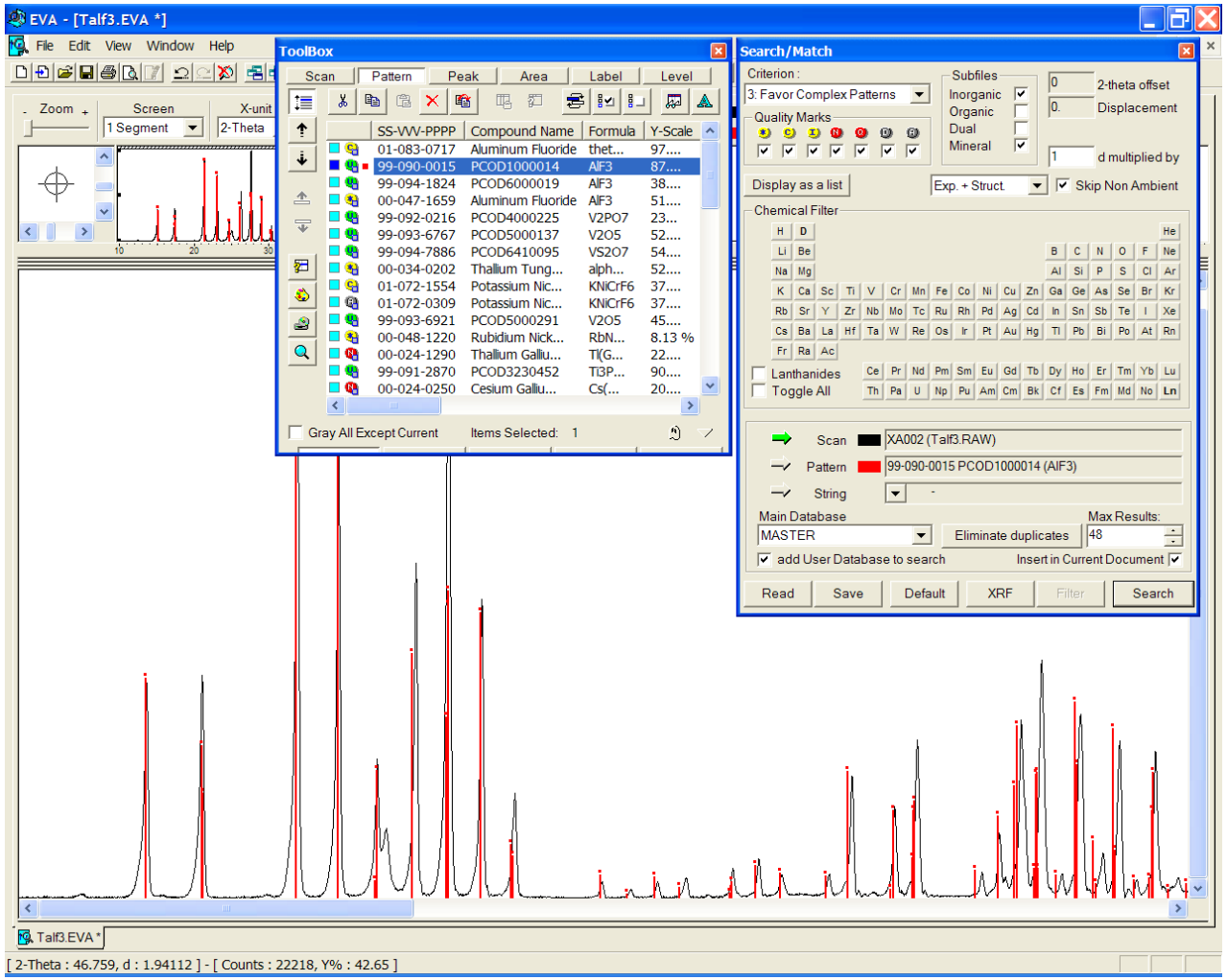
- mineral identification;
- subsequent data dissemination.

SOLSA data flow diagram courtesy Monique Le Guen, ERAMET.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

Use of *COD databases

Search-match identification of the materials

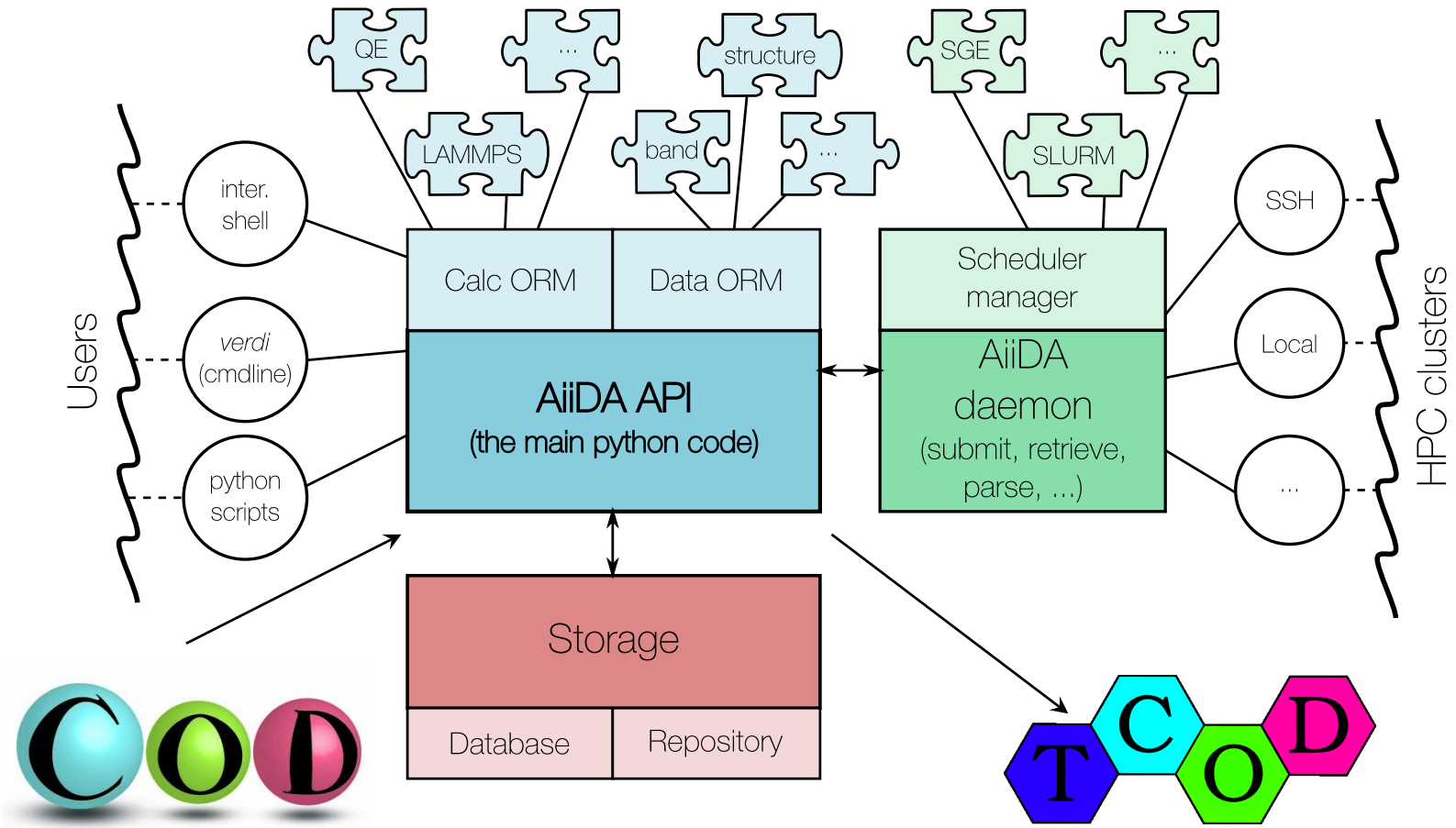


A **predicted** phase from PCOD could be identified in experimental data.

Courtesy Armel Le Bail
[Le Bail, 2008]

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

COD, TCOD and AiiDA link



Courtesy AiiDA developers [Pizzi et al., 2016]

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

*COD data citation

The Research Data Alliance has just published and endorsed **recommendations** from the RDA Working Group on Data Citation:
https://www.rd-alliance.org/groups/data-citation-wg.html

COD data can be cited in several ways:

- Using a data reference URI:

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

- Using a "landing page" URI:

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

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

*COD data citation (2)

COD data can be cited in several ways:

- Using a data reference URI **with explicit revision**:

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

- Using a content-negotiable URI (with or without explicit revision):

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

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

*COD metadata

COD metadata are available in RDF format:
http://www.crystallography.net/cod/2002926.rdf
examples/data/2002926-example.rdf:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:cod="http://www.crystallography.net/cod/doc/rdf/">
  <rdf:Description rdf:about=
        "http://www.crystallography.net/cod/2002926.html">
    <cod:Rall>0.0476</cod:Rall>
    <cod:Robs>0.0476</cod:Robs>
    <cod:Z>8</cod:Z>
    <cod:Zprime>2</cod:Zprime>
    <cod:a>15.017</cod:a>
    <cod:acce_code>BK0051</cod:acce_code>
    <cod:alpha>90</cod:alpha>
    <cod:author>Srivastava, R. C.</cod:author>
    <cod:author>Klooster, W. T.</cod:author>
    <cod:author>Koetzle, T. F.</cod:author>
    <cod:b>5.876</cod:b>
    <cod:beta>90</cod:beta>
    <cod:c>10.418</cod:c>
    <cod:calcformula>Be F4 H8 N2</cod:calcformula>
    <cod:celltemp>163</cod:celltemp>
    <cod:chemname>ammonium tetrafluoroberyllate</cod:chemname>
    <!-- Some content omitted for brevity ... -->
  </rdf:Description>
</rdf:RDF>
```

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

Database **request** citation

- Data requests (searches) should get their persistent identifiers to hide the underlying mechanism:

  http://www.crystallography.net/cod/query/123456
  $\rightarrow$
  http://www.crystallography.net/cod/result?text=perovskite

- Data requests (searches) should be re-runnable on new and old versions of the database:

  http://www.crystallography.net/cod/query/567890
  $\rightarrow$
  http://www.crystallography.net/cod/result?text=perovskite&dbrev=112233

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

Software citations?

FORCE11 recommendations:

Should software be cited? (Yes!) Recommendations published in [Smith et al., 2016]

- TCOD dictionaries contain data items for versions and names of programs and libraries – additional data items for unique identifiers should be provided;
- How deep do we need to cite software? Compiler? OS? CPU? How reproducible will these be?

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

Interlinked data in COD

Crystallography Open Database

COD Home
Home
What's new?

Accessing COD Data
Browse
Search
Search by structural formula

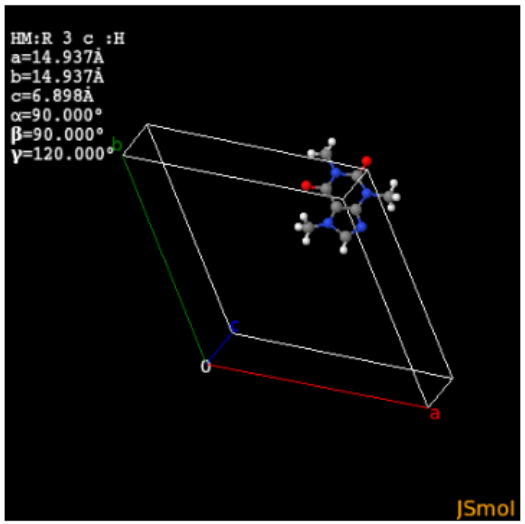
Add Your Data
Deposit your data
Manage depositions
Manage/release prepublications

Documentation
COD Wiki
Obtaining COD
Querying COD
Citing COD
COD Mirrors
Advices to donators
Useful links

Information card for 2100202

2100201 << **2100202** >> 2100203

Preview



Display in Jmol

```
select * from wikipedia_x_cod
+----+--------------+---------+-------------+
| id | ext_id       | cod_id  | relation_id |
+----+--------------+---------+-------------+
|  1 | Ibuprofen    | 2006278 |           1 |
|  2 | Caffeine     | 2100202 |           1 |
|  3 | Serotonin    | 2019147 |           1 |
|  4 | Pristinamycin| 1000001 |           1 |
|  5 | Cucurbituril | 1516465 |           1 |
|  6 | Rubrene      | 1516682 |           1 |
+----+--------------+---------+-------------+
```

Coordinates	2100202.cif
Original IUCr paper	HTML
External links	ChemSpider; DrugBank; PubChem; Wikipedia

▾ **Structure parameters**

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

Acknowledgements

VU Institute of Biotechnology	**QM community**	**COD Advisory board**
	Björkman	Daniel Chateigner
Virginijus Siksnys	Torbjörn	Robert T. Downs
(*head of the dept.*)	Stefaan Cottenier	Werner Kaminsky
	Nicola Marzari	Armel Le Bail
Andrius Merkys	Giovanni Pizzi	Luca Lutterotti
Antanas Vaitkus	Lubomir Smrcok	Peter Moeck
	Linas Vilčiauskas	Peter Murray-Rust
	Chris Wolverton	Miguel Quirós

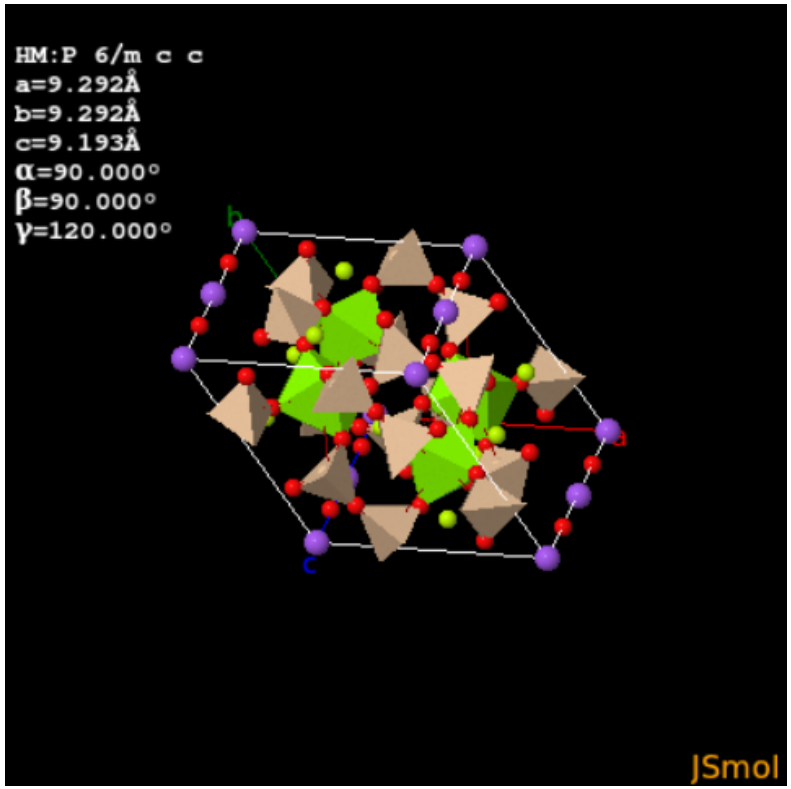
Thanks to the Lorentz Center for supporting participation of the Vilnius group

Thanks to commercial COD users and supporters – Bruker, PANalytical, Rigaku; thanks to IUCr for support and consultations.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

Thank you!

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



http://www.crystallography.net/5000095.html

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.

References

Bernstein, H. J., Bollinger, J. C., Brown, I. D., Gražulis, S., Hester, J. R., McMahon, B., Spadaccini, N., Westbrook, J. D., and Westrip, S. P. (2016).
Specification of the Crystallographic Information File format, version 2.0.
Journal of Applied Crystallography, 49(1).

Boullay, P., Lutterotti, L., and Chateigner, D. (2012).
Quantitative analysis of electron diffraction ring patterns using the MAUD program.

Boullay, P., Lutterotti, L., Chateigner, D., and Sicard, L. (2014).
Fast microstructure and phase analyses of nanopowders using combined analysis of transmission electron microscopy scattering patterns.
Acta Crystallographica Section A, 70:448–456.

Le Bail, A. (2008).
Frontiers between crystal-structure prediction and determination by powder diffractometry.
Powder Diffraction Suppl., pages S5–S12.

Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N., and Kozinsky, B. (2016).
AiiDA: automated interactive infrastructure and database for computational science.
Computational Materials Science, 111:218–230.

Sadowski, P. and Baldi, P. (2013).
Small-molecule 3d structure prediction using open crystallography data.
Journal of Chemical Information and Modeling, 53:3127–3130.

Smith, A. M., Katz, D. S., and Niemeyer, K. E. (2016).
Software citation principles.
PeerJ Computer Science, 2:e86.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.