Bridging experimental and theoretical data in crystallography

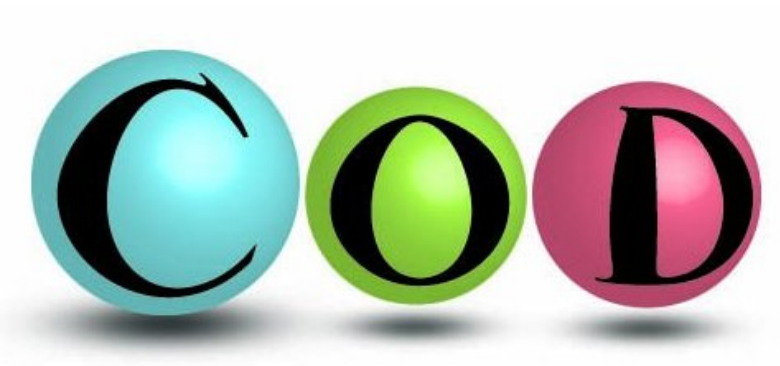
Saulius Gražulis

Lausanne, 2016

Vilnius University Institute of Biotechnology

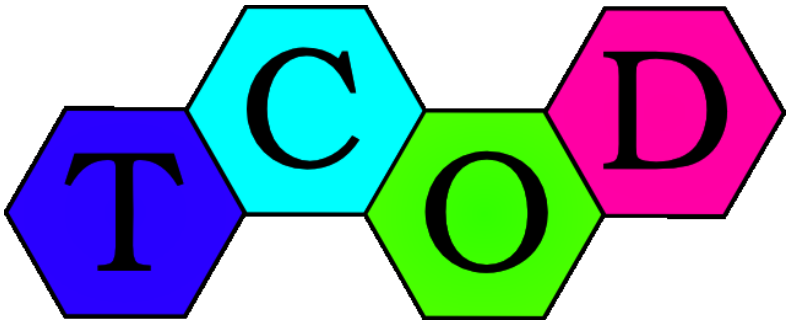
Open Crystallographic Databases

COD, TCOD, PCOD, MPOD, ...



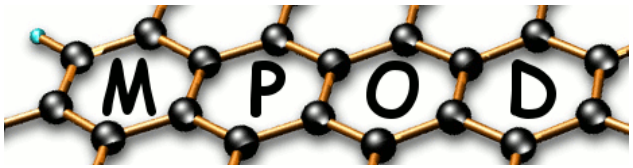
http://www.crystallography.net/cod

$>$ 350 000 entries



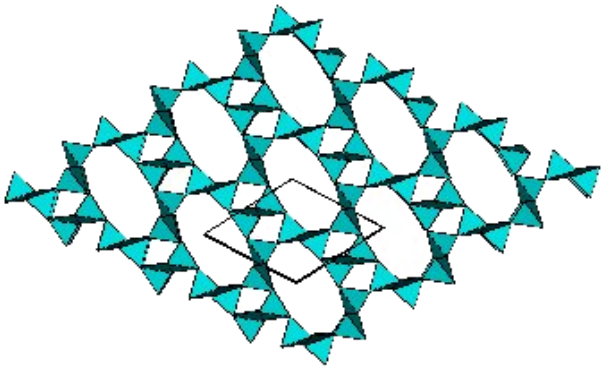
http://www.crystallography.net/tcod

$>$ 350 entries (ready to grow to $>$ 350 000?)



http://mpod.cimav.edu.mx/

$>$ 300 entries

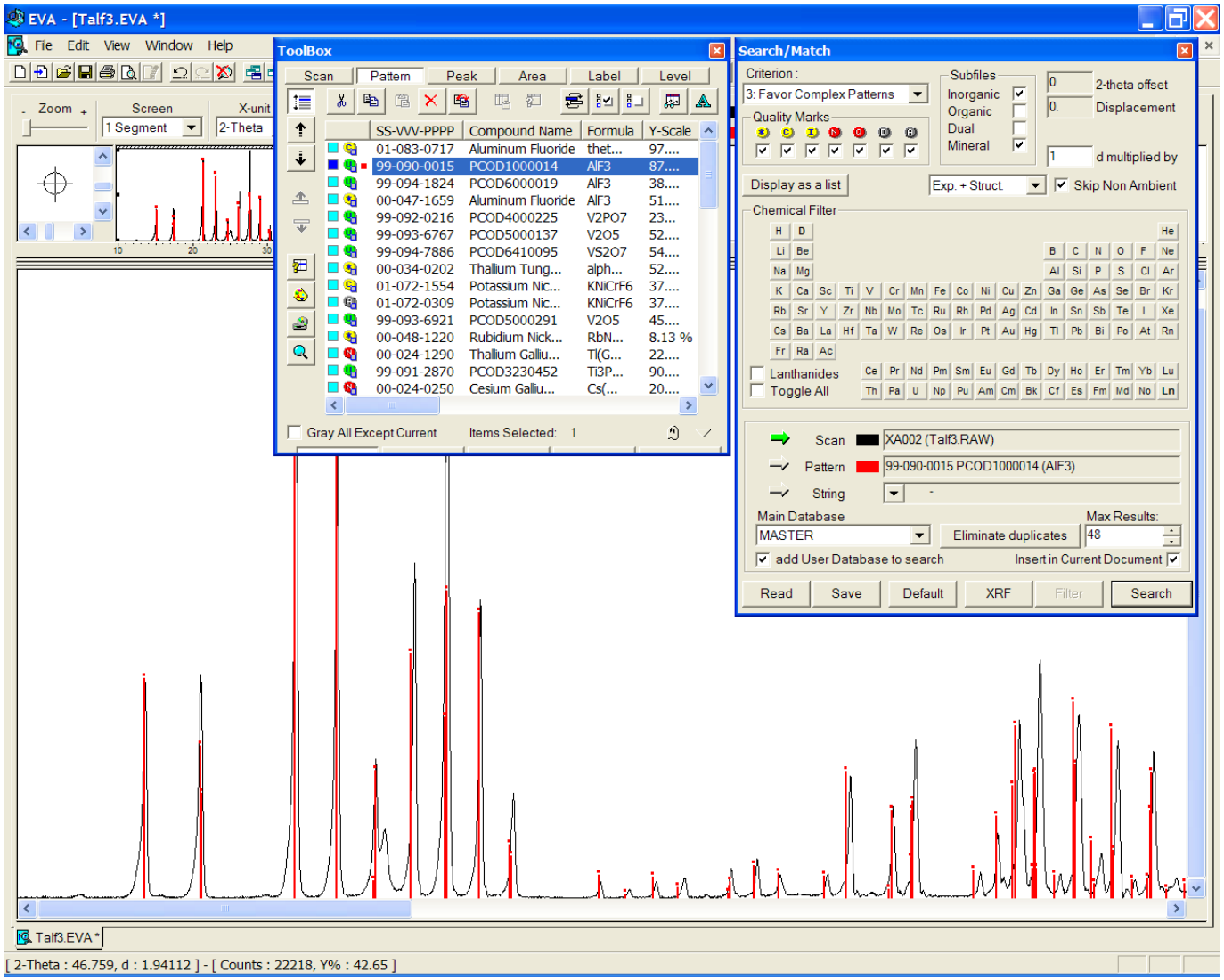


http://www.crystallography.net/pcod

$> 10^6$ entries (ready to grow to $> 10^8$?)

A Crystallography Perspective

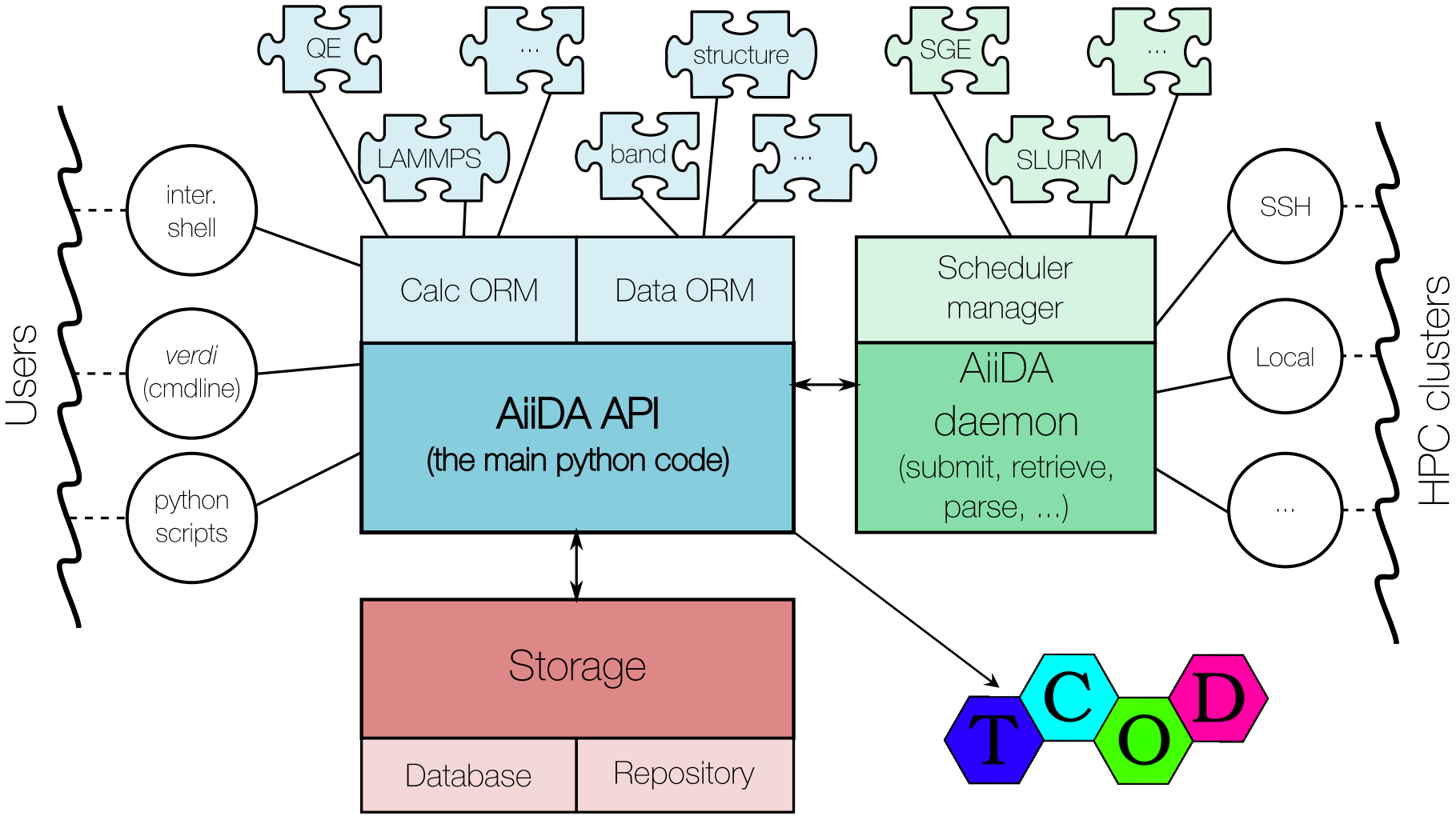
Why crystallographers are interested in theoretical structures?



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

Courtesy Armel Le Bail
[Le Bail, 2008]

TCOD and AiiDA link



Courtesy AiiDA developers [Pizzi et al., 2016]

Crystallographic Interchange Framework (CIF)

CIF, CIF2

1. CIF1,2 are **extendable** in a centralised and **decentralised** ways:
   - The COMCIFS committee of the IUCr manages standard dictionaries;
   - Users can register their unique prefixes;
   - Special data names (`_[local]_name`) can be used privately;
2. CIF is **evolving**: new, more precise names can be introduced (without breaking old code);
3. CIF is an **text based, human readable**
4. CIF is **(open and useful)**! Provided and accepted by:
   - programs (Jmol, Openbabel, Coot, parsers for Perl, Python, C [Merkys et al., 2016], ...);
   - journals;
   - databases;

The CIF Example

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)
```

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_.
;
```

TCOD dictionary contents

The most basic data names

- `cif_tcod.dic`: ver. 0.008, last update 2015-06-16, 106 data names;
- `cif_dft.dic`: ver. 0.015, last update 2016-01-22, 84 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
...
;
```

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

Limitations of CIF

Not really limitations:

- large size (text files); but – can be compressed efficiently;
- not seekable; but – easy to map into relational databases;
- awkward for binary data; but – CBF (CIF Binary Format) exists for 2D image data;

- Not suitable for **very large** files (100 GB – $\sim$ TB scale datasets); interoperability of CBF with HDF5 is being developed.

Other possibilities XML and CML

The Chemical Modelling Language, Dictionary for quantum mechanical computations; developed by Peter Murray-Rust and his team.

- XML-based;
- used in the Quixote project;
- supported by multiple Java packages;
- Defines CML *Conventions* and *Dictionaries*: http://www.xml-cml.org/dictionary/

Comparison of CIF, XML and JSON

XML	CIF	JSON
text based	text based	text based
easy to parse	easy to parse	easy to parse
extendable	extendable	extendable
noisy?	frugal	frugal
verifiable	verifiable	verifiable?
eof-verifiable	eof-open	eof-verifiable
not cat-able	cat-able	cat-able
XML-in-XML?	CIF-in-CIF OK	JSON-in-JSON OK

Harmonisation of TCOD dictionaries

Are we all nomads? :)

- Import new dictionary definitions (from Nomad, other communities, etc.)
- Rename or link existing TCOD dictionary definitions if they are different from those in other ontologies (Nomad, etc.);
- Offer our definitions for other ontologies (we are Open :);
- Make a round-trip CIF$\leftrightarrow$XML possible!

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).


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


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).


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.

Acknowledgements

VU Institute of Biotechnology

Virginijus Siksnys (*head of the dept.*)

Andrius Merkys
Antanas Vaitkus

QM community

Björkman Torbjörn
Stefaan Cottenier
Nicola Marzari
Giovanni Pizzi
Lubomir Smrcok
Linas Vilčiauskas
Chris Wolverton

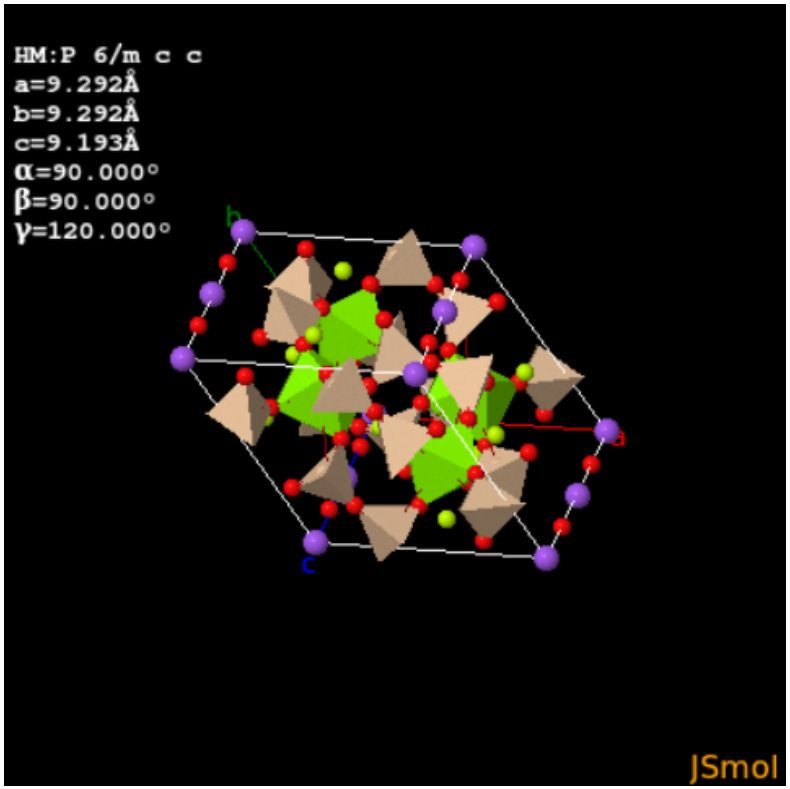
COD Advisory board

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

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

Thank you!

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



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