

Sharing Computation Results about Solid Materials Using the Crystallographic Interchange Framework (CIF)

Saulius Gražulis

Lausanne, 2015

Vilnius University Institute of Biotechnology



Data Sharing and Reproducible Research

... the imperative

- in $< 1/2$ of the microarray publications, analyses are not reproducible due to lack of data/protocols/software [3]

Data Sharing in Crystallography

Started quite early

- **1948 Acta Cryst. (IUCr)** The Acta Crystallographica journal was launched, *all coordinates were printed in journal articles, and Acta Crystallographica published the structure factors as well* [2]
- **1965 CSD (CCDC)** The CCDC was established at the Department of Chemistry, Cambridge University /.../ about 2000 structures published before 1965 were gradually incorporated into the developing database [1]
- **1971 PDB** In June 1971, the two communities attended the Cold Spring Harbor Symposium on Quantitative Biology (Cold Spring Laboratory Press, 1972) [4, 2]

The CIF Framework

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 BaTiO3~
 ;
 _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_.
;
```

- To ensure high quality of deposited data;
- To offers ontologies in a form of CIF (Hall 1991) dictionaries for data description;
- To implement an automated pipeline that checks each submitted structure against a set of community-specified criteria for convergence, computation quality and reproducibility.



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



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](#).

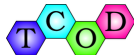
All data on this site have been placed in the public domain by the contributors.

Currently there are **314600** entries in the COD.
Latest deposited structure: [1519774](#) on **2015-05-20** at **08:04:11 UTC**



TCOD – a database for storing results of computations

DFT



Theoretical Crystallography Open Database

TCOD Home

[Home](#)
[What's new?](#)

Accessing TCOD Data

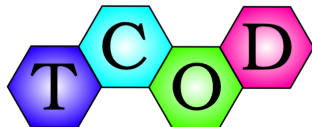
[Browse](#)
[Search](#)
[Search by structural formula](#)

Add Your Data

[Deposit your data](#)
[Manage depositions](#)
[Manage/release prepublications](#)

Documentation

[COD Wiki](#)
[Obtaining COD](#)
[Querying COD](#)
[Citing COD](#)
[COD Mirrors](#)
[Advices to donors](#)
[Useful links](#)

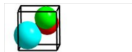


Open-access collection of theoretically calculated or refined crystal structures of organic, inorganic, metal-organic compounds and minerals, excluding biopolymers

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

Currently there are **180** entries in the TCOD.

Latest deposited structure: [10000001](#) on **2015-05-12** at **14:26:55 UTC**



CIFs Donators



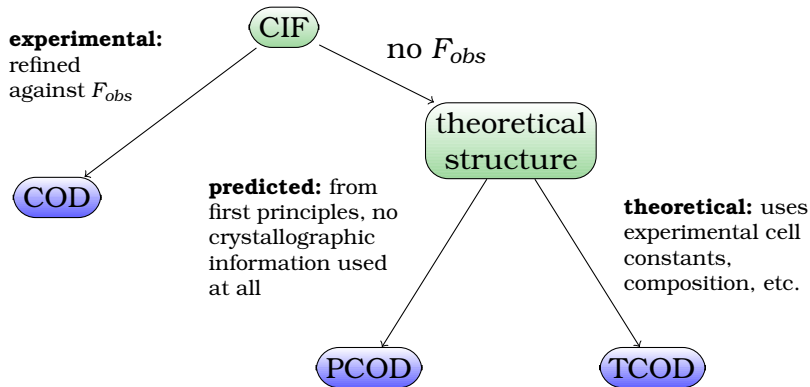
Accessing data

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'
```

# Structure classification

## COD sister databases



Dictionaries are available at:

<http://www.crystallography.net/tcod/cif/dictionaries/>:

## cif\_tcod.dic

```
data_tcode_structure_type
 _name '_tcode_structure_type'
 _type char
 loop__enumeration
 _enumeration_detail

 ground-state
 'refined crystal structure at ground state'
```

## cif\_dft.dic

```
data_tcod_dft_valence_electrons
 _name '_dft_valence_electrons'
 _type numb
 _definition
; Total number of valence electrons in a calculation.
;
```

# TCOD dictionary contents

The most basic data names

- `cif_tcod.dic`: ver. 0.005, last update 2015-05-21, 102 data names;
- `cif_dft.dic`: ver. 0.005, last update 2015-05-07, 71 data name.

e.g.:

```
data_dft_core_electrons
 _name '_dft_core_electrons'
 _type numb
 _enumeration_range 1:
 _definition
; Total number of core electrons in calculation
;
```

# Structure description levels


Structures may be described at different level of detail in TCOD:

| <b>Level 0</b>           | <b>Level 1</b>                       | <b>Level 2</b>            |
|--------------------------|--------------------------------------|---------------------------|
|                          | Level 0, plus:                       | Level 1, plus:            |
| ① lattice and symmetry   | ① computational setup & parameters   | ① input scripts and files |
| ② atomic coordinates     | ② residual forces on atoms and cell  | ② command line            |
| ③ bibliography reference | ③ code-specific convergence criteria | ③ output logs of the code |

# Our first Level 2 structure in TCOD'e

Relaxed cod/1507756 entry

## TCOD Home

[Home](#)  
[What's new?](#) 

## Accessing TCOD Data

[Browse](#)  
[Search](#)  
[Search by structural formula](#)

## Add Your Data

[Deposit your data](#)  
[Manage depositions](#)  
[Manage/release prepublications](#)

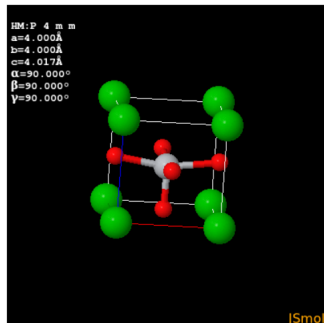
## Documentation

[COD Wiki](#)  
[Obtaining COD](#)  
[Querying COD](#)  
[Citing COD](#)  
[COD Mirrors](#)  
[Advices to donators](#)  
[Useful links](#)

## Information card for 10000001

[20000179](#) << **10000001** >> [20000001](#)

## Preview



[Display in Jmol](#)

**Coordinates** [10000001.cif](#)

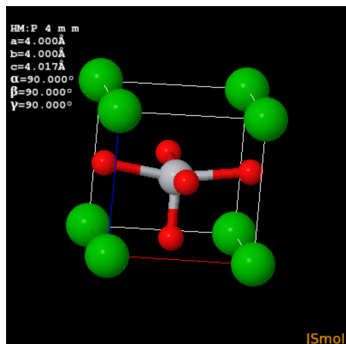
## Structure parameters

|                        |                                                        |
|------------------------|--------------------------------------------------------|
| Formula                | Ba O3 Ti                                               |
| Calculated formula     | Ba O3 Ti                                               |
| Title of publication   | Relaxation of COD entry 1507756 using Quantum ESPRESSO |
| Authors of publication | Andrius Merkys                                         |
| Journal of publication | Personal communication to TCOD                         |
| Year of publication    | 2015                                                   |

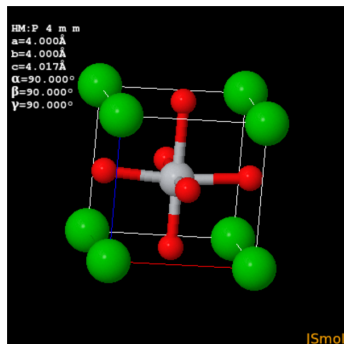
# Comparison of theory and experiment

Relaxed and initial cod/1507756 structure

In theory, there should be no difference between the theory and the experiment, but in practice...



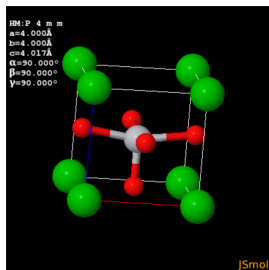
Theory (tcod/10000001)



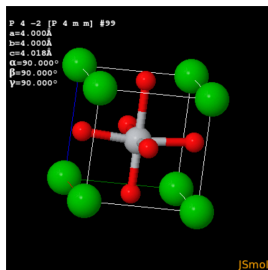
Experiment (cod/1507756)

# Comparison of theory and experiment (2)

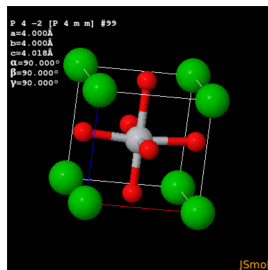
## More experimental structures



Theory (tcod/1000001)



Experiment (cod/2100858)



Experiment (cod/2100859)



# Quantitative structure comparison

Bilbao Crystallographic Server

<http://www.cryst.ehu.es/cryst/compstru.html>

Maximum distance ( $d_{max}$ , Å)/Arithmetic mean ( $d_{av}$ , Å)

|          | TCOD<br>10000001 | COD<br>1507756 | COD<br>1513252 | COD<br>2100858 | COD<br>2100859 |
|----------|------------------|----------------|----------------|----------------|----------------|
| 10000001 | -                | Err.           | 0.0360/0.0144  | 0.1059/0.0574  | 0.1259/0.0607  |
| 1507756  |                  | -              | Err.           | Err.           | Err.           |
| 1513252  |                  |                | -              | 0.0703/0.0466  | 0.0905/0.0498  |
| 2100858  |                  |                |                | -              | 0.0201/0.0080  |

# Conclusions

- Having COD and TCOD in uniform format, in same setting of the unit cell enables immediate comparisons;
- DFT methods are accurate enough to validate experimental structures;
- Can we also validate DFT methods?
- Should work much more to populate TCOD and make it comprehensive computation archiving tool;

# References



Frank H. Allen.

The cambridge structural database: a quarter of a million crystal structures and rising.

*Acta Crystallographica Section B*, 58(3 Part 1):380–388, Jun 2002.



Helen M. Berman, Philip E. Bourne, and John Westbrook.

The protein data bank: A case study in management of community data.

*Current Proteomics*, pages 49–57, 2004.



John P. A. Ioannidis, David B. Allison, Catherine A. Ball, Issa Coulibaly, Xiangqin Cui, Aedín C. Culhane, Mario Falchi, Cesare Furlanello, Laurence Game, Giuseppe Jurman, Jon Mangion, Tapan Mehta, Michael Nitzberg, Grier P. Page, Enrico Petretto, and Vera van Noort.

Repeatability of published microarray gene expression analyses.

*Nat Genet*, 41(2):149–155, 2009.



Protein Data Bank.

Protein Data Bank.

*Nature New Biology*, 233(42):223, Oct 1971.

## **VU Biotechnologijos institutas**

Virginijus Siksnys  
(*skyriaus vadovas*)

Andrius Merkys  
Antanas Vaitkus

## **COD Advisory Board**

Daniel Chateigner

Robert T. Downs

Armel Le Bail

Luca Lutterotti

Peter Moeck

Peter Murray-Rust

Miguel Quirós

## **DFT Experts**

Nicola Marzari  
Chris Wolverton

Stefaan Cottenier  
Björkman Torbjörn

Linas Vilčiauskas  
Lubomir Smrcok

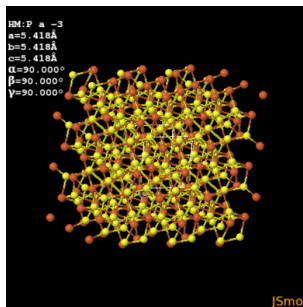
**Many thanks to our commercial users and supporters: Bruker, Crystal Impact, PANalytical, Rigaku**

Financing: Research Council of Lithuania (2010–2011, 2013–2015), Vilnius University, VU Institute of Biotechnology.

# Thank you!



<http://en.wikipedia.org/wiki/Pyrite>  
"2780M-pyrite1" by CarlesMillan –  
Own work. Licensed under CC  
BY-SA 3.0 via [Wikimedia Commons](#)



<http://www.crystallography.net/cod/5000115.html>