CIF dictionaries for predicted structures

Saulius Gražulis Andrius Merkys Antanas Vaitkus

Vilnius, 2023

Vilnius University Institute of Biotechnology

Id: slides.tex 2507 2023-10-31 13:46:09Z saulius October 31, 2023

The increase of theoretically refined structures

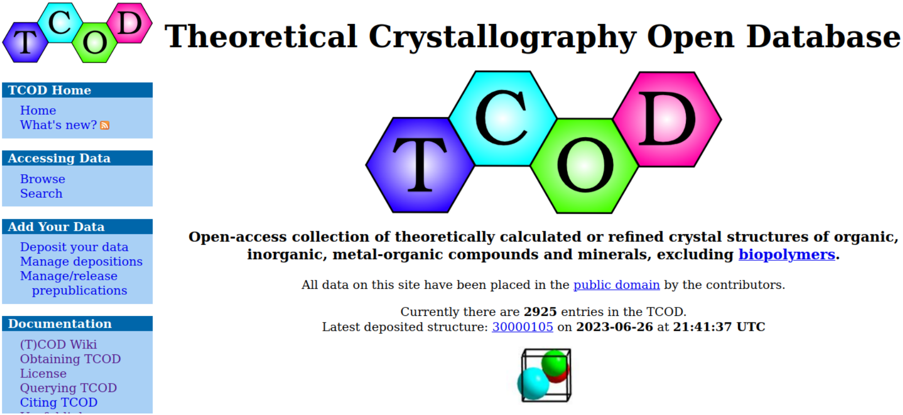
Theoretical methods.

- Rise of DFT methods allows to refine atomic coordinates in crystals, and to predict various crystal properties
- MM, MM/QM methods, MC methods allow to predict crystal structures.

TCOD – Theoretical Crystallography Open Database

Available at:

http://www.crystallography.net/tcod/



Theoretical Crystallography Open Database

TCOD Home
- Home
- What's new?

Accessing Data
- Browse
- Search

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

Documentation
- (T)COD Wiki
- Obtaining TCOD
- License
- Querying TCOD
- Citing TCOD

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 **2925** entries in the TCOD.
Latest deposited structure: 30000105 on **2023-06-26** at **21:41:37 UTC**

TCOD: recorded information, dictionaries, quality criteria

Contents:

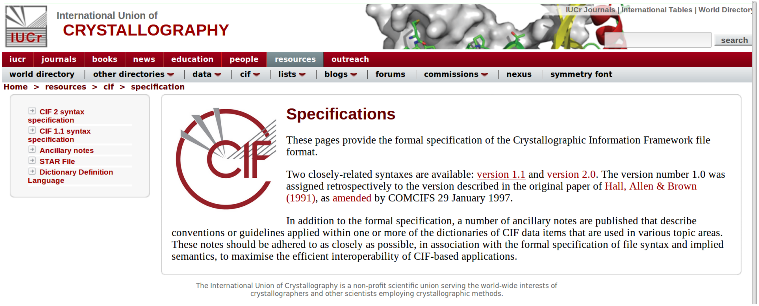
- Information needed to evaluate the quality of the theoretical structure;
- Information needed to reproduce the structure;
- Information needed to repeat computation verbatim;

Quality criteria:

(Join the discussion at http://lists.crystallography.net/cgi-bin/mailman/listinfo/tcod)

- DFT
  - residual forces on atoms;
  - residual charges at atomic positions
- MM
  - ...

The CIF framework



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.

- Is suitable for processing and archiving;
- Is maintained by the IUCr;
- Has > 1000 data names defined;
- Is independently extensible;

(Hall et al. 1991; Bernstein et al. 2016)

Expression of computations

Computation

Any computation can be expressed as a Unix (Linux ;) command line with specified:

- inputs (including STDIN)
- outputs (including STDOUT and STDERR)
- command to run
- environment (CPU, OS, libraries, ENV variables, etc.)

Example of CIF computation

Computation using a local file:

```
cif_molecule \
    --preserve-stoichiometry \
    --covalent-sensitivity 0.75 \
    2200231.cif \
    > 2200231_molecular_entities.cif
```

Example of CIF computation

Computation using a local file:

```
cif_molecule \
    --preserve-stoichiometry \
    --covalent-sensitivity 0.75 \
    2200231.cif \
    > 2200231_molecular_entities.cif
```

Computation using a database record (specific version):

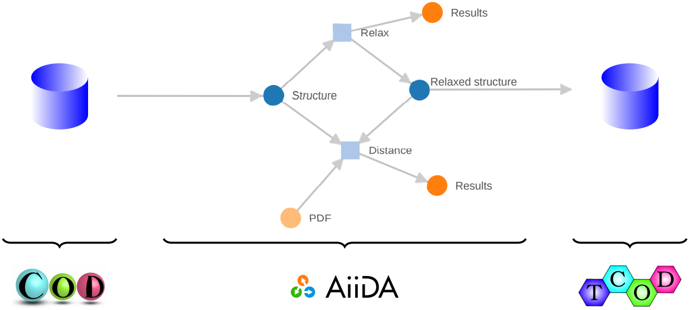
```
curl -ksSL \
    https://crystallography.net/cod/2200231.cif@1 | \
cif_molecule \
    --preserve-stoichiometry \
    --covalent-sensitivity 0.75 \
    > 2200231_molecular_entities.cif
```

Structure description levels

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

Level 0	**Level 1**	**Level 2**
	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

Synchronisation with AiiDA

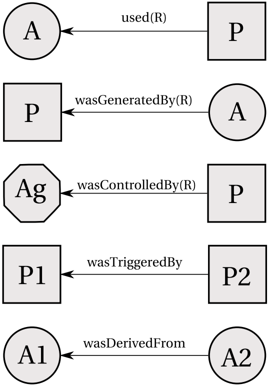


- TCOD + AiiDA:
  - Direct export of calculation results generated by any of the supported codes;
  - Automatic generation of level 2 structure descriptions.

(Merkys et al. 2017; Pizzi 2018)

Open Provenance Model

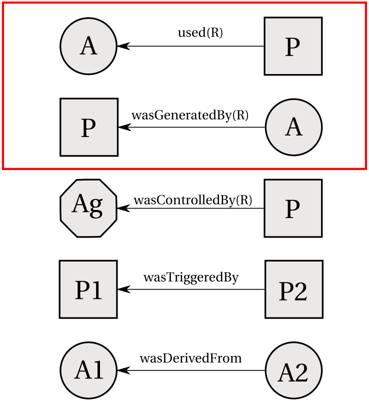
Essentially a subset of the OPM (the first two relations):



(Moreau et al. 2011)

Open Provenance Model

Essentially a subset of the OPM (the first two relations):



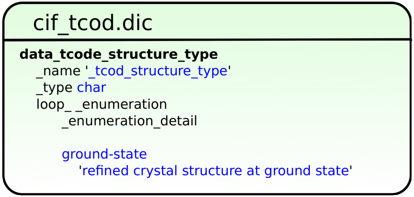
(Moreau et al. 2011)

Dictionaries

Dictionaries allow us:

- to give machine- and and human-readable definitions of data;
- to implement automatic conformace checks (validation).

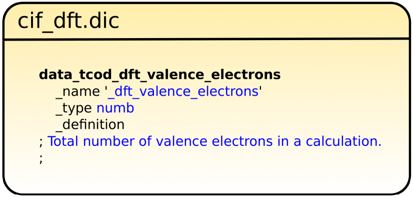
Dictionaries are available at:
http://www.crystallography.net/tcod/cif/dictionaries/:



cif_tcod.dic

```
data_tcode_structure_type
   _name '_tcod_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.
;
```

Data item overview

Category list

DFT data categories (cif_dft-wip.dic):

20	dft_BZ_integration	25	dft_calc_property
5	dft_BZ_integration_grid_IBZ_point	3	dft_cell_conv
8	dft_XC_functional	3	dft_cell_magn
3	dft_alloy	6	dft_cell_settings
4	dft_atom_basisset	10	dft_energy
14	dft_atom_type	3	dft_kinetic_energy_cutoff
9	dft_basisset	4	dft_pseudopotential

TCOD data categories (cif_tcod-wip.dic):

16	atom_site	5	tcod_initial_cell_param
12	atom_sites	5	tcod_initial_coordinate
1	citation	2	tcod_method
15	tcod_computation	17	tcod_software
3	tcod_content_encoding	11	tcod_software_library
2	tcod_data_source	3	tcod_source_database
1	tcod_database	4	tcod_source_structure_database
5	tcod_ff	6	tcod_total_energy
17	tcod_file		

Data item overview

Category list

DFT data categories (cif_dft-wip.dic):

20	dft_BZ_integration	25	dft_calc_property
5	dft_BZ_integration_grid_IBZ_point	3	dft_cell_conv
8	dft_XC_functional	3	dft_cell_magn
3	dft_alloy	6	dft_cell_settings
4	dft_atom_basisset	10	dft_energy
14	dft_atom_type	3	dft_kinetic_energy_cutoff
9	dft_basisset	4	dft_pseudopotential

TCOD data categories (cif_tcod-wip.dic):

16	atom_site	5	tcod_initial_cell_param
12	atom_sites	5	tcod_initial_coordinate
1	citation	2	tcod_method
15	tcod_computation	17	tcod_software
3	tcod_content_encoding	11	tcod_software_library
2	tcod_data_source	3	tcod_source_database
1	tcod_database	4	tcod_source_structure_database
5	tcod_ff	6	tcod_total_energy
17	tcod_file		

Data item overview

Category list

DFT data categories (cif_dft-wip.dic):

20	dft_BZ_integration	25	dft_calc_property
5	dft_BZ_integration_grid_IBZ_point	3	dft_cell_conv
8	dft_XC_functional	3	dft_cell_magn
3	dft_alloy	6	dft_cell_settings
4	dft_atom_basisset	10	dft_energy
14	dft_atom_type	3	dft_kinetic_energy_cutoff
9	dft_basisset	4	dft_pseudopotential

TCOD data categories (cif_tcod-wip.dic):

16	atom_site	5	tcod_initial_cell_param
12	atom_sites	5	tcod_initial_coordinate
1	citation	2	tcod_method
15	tcod_computation	17	tcod_software
3	tcod_content_encoding	11	tcod_software_library
2	tcod_data_source	3	tcod_source_database
1	tcod_database	4	tcod_source_structure_database
5	tcod_ff	6	tcod_total_energy
17	tcod_file		

Data item overview

Category list

DFT data categories (cif_dft-wip.dic):

20	dft_BZ_integration	25	dft_calc_property
5	dft_BZ_integration_grid_IBZ_point	3	dft_cell_conv
8	dft_XC_functional	3	dft_cell_magn
3	dft_alloy	6	dft_cell_settings
4	dft_atom_basisset	10	dft_energy
14	dft_atom_type	3	dft_kinetic_energy_cutoff
9	dft_basisset	4	dft_pseudopotential

TCOD data categories (cif_tcod-wip.dic):

16	atom_site	5	tcod_initial_cell_param
12	atom_sites	5	tcod_initial_coordinate
1	citation	2	tcod_method
15	tcod_computation	17	tcod_software
3	tcod_content_encoding	11	tcod_software_library
2	tcod_data_source	3	tcod_source_database
1	tcod_database	4	tcod_source_structure_database
5	tcod_ff	6	tcod_total_energy
17	tcod_file		

Example data names

TCOD provenance data names (cif_dft-wip.dic):

```
_tcod_computation_input_file
_tcod_computation_log_file
_tcod_computation_stdout
_tcod_computation_stderr
_tcod_computation_CPU_time
_tcod_computation_wallclock_time
_tcod_computation_command
_tcod_computation_step
_tcod_computation_environment
_tcod_computation_reference_uuid
_tcod_computation_reference_id
_tcod_computation_reference_URI
_tcod_computation_database_name
_tcod_computation_database_version
_tcod_computation_database_URI
```

DFT calculated property data names (cif_tcod-wip.dic):

```
_dft_band_gap
_dft_bulk_modulus
_dft_lattice_energy
_dft_stiffness_tensor_ij
```

TCOD CIF example

http://www.crystallography.net/tcod/10000178.html

```
data_10000178
_publ_section_title
;
 Tutorial material of "Tutorial on high-throughput computations: General
 methods and applications using AiiDA, June-July, 2016"
;
_journal_name_full                'Personal communication to TCOD'
_journal_year                     2016
_chemical_formula_sum             'Cs O3 Ta'
_space_group_IT_number            221
_symmetry_Int_Tables_number       221
_symmetry_space_group_name_Hall   '-P 4 2 3'
_symmetry_space_group_name_H-M    'P m -3 m'
_audit_creation_method            'AiiDA version 0.7.0'
_cell_angle_alpha                 90.0
_cell_angle_beta                  90.0
_cell_angle_gamma                 90.0
```

...

```
loop_
_atom_site_label
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_type_symbol
Cs1 0.0 0.0 0.0 Cs
Ta1 0.5 0.5 0.5 Ta
O1 0.5 0.5 0.0 O
```

...

TCOD CIF example

http://www.crystallography.net/tcod/10000178.html

```
data_10000178
_publ_section_title
;
 Tutorial material of "Tutorial on high-throughput computations: General
 methods and applications using AiiDA, June-July, 2016"
;
_journal_name_full                 'Personal communication to TCOD'
_journal_year                      2016
_chemical_formula_sum              'Cs O3 Ta'
_space_group_IT_number             221
_symmetry_Int_Tables_number        221
_symmetry_space_group_name_Hall    '-P 4 2 3'
_symmetry_space_group_name_H-M     'P m -3 m'
_audit_creation_method             'AiiDA version 0.7.0'
_cell_angle_alpha                  90.0
_cell_angle_beta                   90.0
_cell_angle_gamma                  90.0
```

...

```
loop_
_tcod_computation_step
_tcod_computation_command
_tcod_computation_reference_uuid
_tcod_computation_environment
_tcod_computation_stdout
_tcod_computation_stderr
0 'cd 0; ./_aiidasubmit.sh' 09e00761-3128-414c-90e8-266490ba6e71
```

...

Data validation examples

COD data validation policies:

1. Syntactic checks:

```
$ cifparse 7234818.cif
```

   Syntax recently expanded to CIF2 (Bernstein et al. 2016; Merkys et al. 2016)

2. Semantic validation (against dictionaries)

```
$ cif_validate -D cif_core.dic 7234818.cif
```

   Validation capabilities recently expanded to DDLm (Vaitkus et al. 2021).

3. Database-specific checks

```
$ cif_cod_check 7234818.cif
```

Data validation capabilities

Detect automatically:

- Incorrect data types;
- Out of range values;
- (Some) broken loops (i.e. data tables);
- Missing or incorrect data keys (e.g. atom names);

COD validation examples

```
/usr/bin/cif_validate: 1506432.cif data_1506432:
NOTE, data item '_atom_site_aniso_label' contains value 'F40'
that was not found among the values of the parent data item
'_atom_site_label'.
```

COD validation examples

```
/usr/bin/cif_validate: 1506432.cif data_1506432:
NOTE, data item '_atom_site_aniso_label' contains value 'F40'
that was not found among the values of the parent data item
'_atom_site_label'.
```

```
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
# ... some data names omitted for brevity
>F40 F 0.21810(11) -1.5061(4) 0.7984(2) 0.0684(9) # ...
F41 F 0.29902(11) -1.4446(4) 0.8587(2) 0.0724(9)  # ...
```

COD validation examples

```
/usr/bin/cif_validate: 1506432.cif data_1506432:
NOTE, data item '_atom_site_aniso_label' contains value 'F40'
that was not found among the values of the parent data item
'_atom_site_label'.
```

```
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
# ... some data names omitted for brevity
>F40 F 0.21810(11) -1.5061(4) 0.7984(2) 0.0684(9) # ...
F41 F 0.29902(11) -1.4446(4) 0.8587(2) 0.0724(9)  # ...
```

Validation *might* help to catch data errors if applied consistently during the publication.

Where to go from here?

- It would be great to reuse the TCOD and DFT dictionaries;
- Full computational provenance is *a must*!
- New data items can be added;
- New dictionaries (CSP specific?) can be created;
- Validation of other formats (JSON, XML) can be done (?);

Acknowledgements

DFT/MM community

Stefaan Cottenier
Björkman Torbjörn
Linas Vilciauskas
Lubomir Smrcok
Chris Wolverton
Peter Murray-Rust

COD Advisory board

Daniel Chateigner
Robert T. Downs
Armel Le Bail
Luca Lutterotti
Peter Moeck
Miguel Quirós

VU Institute of Biotechnology

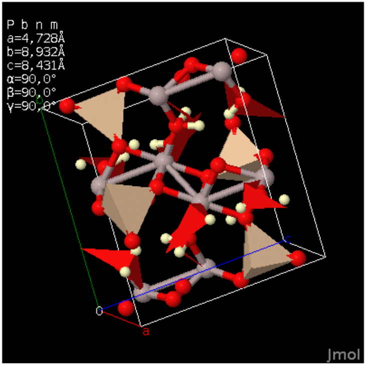
Algirdas Grybauskas
Andrius Merkys
Antanas Vaitkus

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

Academic financing: Research Council of Lithuania (2010–2011, 2013–2015, 2021–2023), COST, Vilnius University, VU Institute of Biotechnology.

Thank you!

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



Coordinates	2207377.cif
Original IUCr paper	HTML

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

http://www.crystallography.net/cod/archives/2023/slides/TCOD-for-CSP/slides.pdf

References I

Bernstein, Herbert J. et al. (Feb. 2016). “Specification of the Crystallographic Information File format, version 2.0”. In: *Journal of Applied Crystallography* 49.1, pp. 277–284. ISSN: 1600-5767. DOI: 10.1107/s1600576715021871. URL: http://dx.doi.org/10.1107/S1600576715021871.

Hall, S. R. et al. (1991). “The crystallographic information file (CIF): a new standard archive file for crystallography”. In: *Acta Crystallographica Section A* 47, pp. 655–685. DOI: 10.1107/S010876739101067X. URL: http://dx.doi.org/10.1107/S010876739101067X.

Merkys, Andrius et al. (Feb. 2016). “*COD::CIF::Parser*: an error-correcting CIF parser for the Perl language”. In: *Journal of Applied Crystallography* 49.1, pp. 292–301. DOI: 10.1107/S1600576715022396. URL: http://dx.doi.org/10.1107/S1600576715022396.

Merkys, Andrius et al. (Nov. 2017). “A posteriori metadata from automated provenance tracking: Integration of AiiDA and TCOD”. In: *Journal of Cheminformatics* 9.1, p. 56. DOI: 10.1186/s13321-017-0242-y. arXiv: 1706.08704v3 [cond-mat.mtrl-sci]. URL: https://jcheminf.springeropen.com/articles/10.1186/s13321-017-0242-y.

Moreau, Luc et al. (June 2011). “The Open Provenance Model core specification (v1.1)”. In: *Future Generation Computer Systems* 27.6, pp. 743–756. DOI: 10.1016/j.future.2010.07.005. eprint: https://eprints.soton.ac.uk/268332/1/opm.pdf.

References II

Pizzi, Giovanni (2018). “Open-Science Platform for Computational Materials Science: AiiDA and the Materials Cloud”. In: *Handbook of Materials Modeling*. Springer International Publishing, p. 1. DOI: 10.1007/978-3-319-42913-7_64-1.

Vaitkus, Antanas et al. (Feb. 2021). “Validation of the Crystallography Open Database using the Crystallographic Information Framework”. In: *Journal of Applied Crystallography* 54.2, pp. 1–12. ISSN: 1600-5767. DOI: 10.1107/s1600576720016532. URL: https://doi.org/10.1107/S1600576720016532.