Developing Experimental & Theoretical Crystallography Open Databases

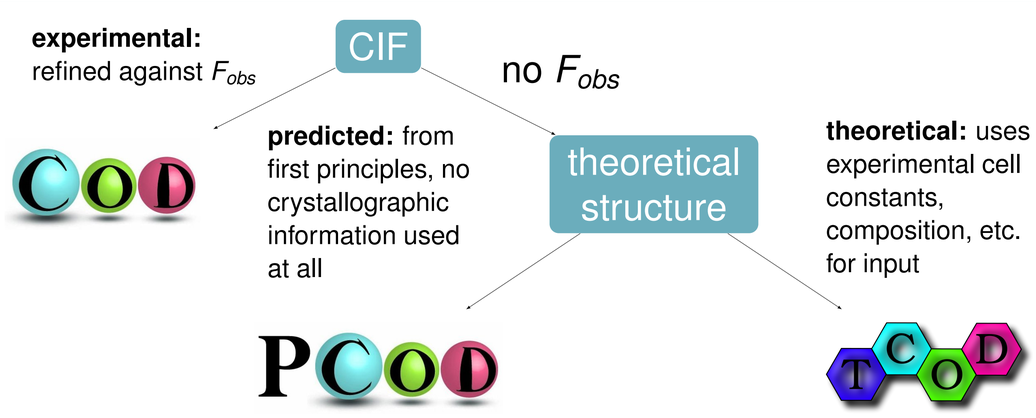
Andrius Merkys[a,b], Giovanni Pizzi[a], Andrea Cepellotti[a], Nicolas Mounet[a], Saulius Gražulis[b] and Nicola Marzari[a]

[a]Theory and Simulation of Materials (THEOS) and National Center for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland; [b]Department of Protein-DNA Interactions, Vilnius University Institute of Biotechnology, V. A. Graičiūno 8, LT-02241 Vilnius, Lithuania

Abstract

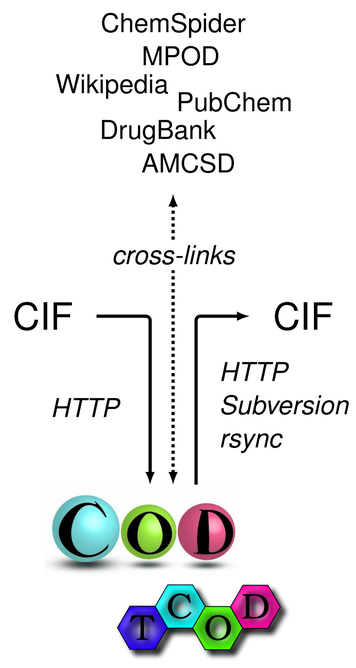
The Crystallography Open Database (COD) [1], launched as a grass-root initiative by an international group of scientists, has become the largest open-access resource to date for experimentally determined small-molecule crystal structures and is ready to be used as a source for large-scale automated analyses in various fields of computational chemistry, such as drug design and material research. A variety of data access and selection options, cross-links with other resources are made possible thanks to the open-access nature of the COD. Recently, a similar effort – the Theoretical Crystallography Open Database (TCOD) – was launched alongside the COD, aimed to collect the results of atomistic simulations using the unified Crystallographic Interchange Framework/Format (CIF) [2].

Family of COD databases



COD & TCOD

- COD, http://www.crystallography.net/cod/
  - Contains ∼ 300 000 entries (as of May 2015);
  - Stores supplementary material of published research as well as prepublication and personal communication material;
  - Harvests data from open journals, accepts depositions via automatic data submission site;
  - Accepts single crystal as well as powder diffraction experiment data;
  - Performs routine automatic quality checks on all incoming structures.
- TCOD, http://www.crystallography.net/tcod/
  - An open-access resource of theoretical computation results;
  - Based on the infrastructure of the COD;
  - Aims to store the metadata for the full replication of computation results.

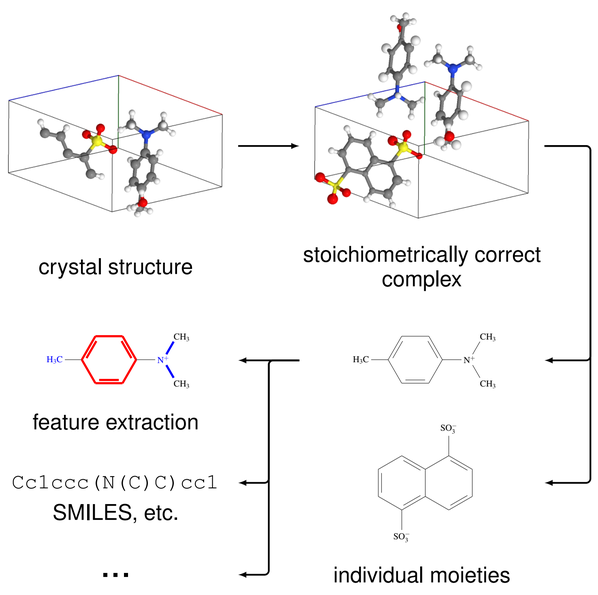


CIF dictionaries for (T)COD

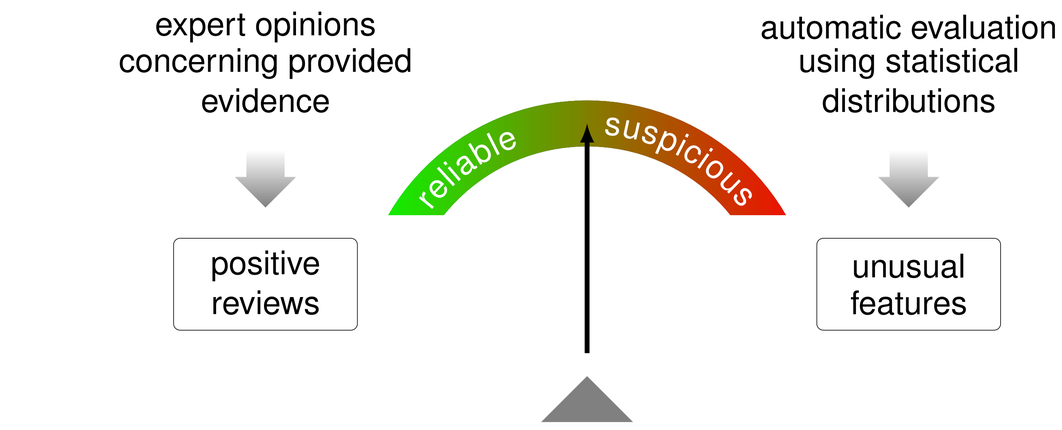
- Offer ontologies for data description;
- Aim at automated checks for convergence, computational quality and reproducibility;
- Enable automated deposition and data mining;
- Accessible at:
  - COD CIF dictionary:
    - http://www.crystallography.net/cod/cif/dictionaries/cif_cod.dic
  - TCOD CIF dictionaries:
    - http://www.crystallography.net/tcod/cif/dictionaries/cif_tcod.dic
    - http://www.crystallography.net/tcod/cif/dictionaries/cif_dft.dic
- Open mailing lists for discussions:
  - http://lists.crystallography.net/cgi-bin/mailman/listinfo/cod-dev
  - http://lists.crystallography.net/cgi-bin/mailman/listinfo/tcod

COD: extraction of the chemical information

- Fully automatic pipeline is devised;
- Software from CrystalEye [3] is employed:
  - heuristics for calculation of partial charges;
  - heuristics for determination of bond orders;
  - algorithm to isolate individual moieties;
  - algorithms to extract ring and chain nuclei.
- Input and output use common file formats (CIF, CML and SDF).



COD & TCOD: platform for data reviews



- "Unusual" is not necessarily "wrong"
  - Automated checks spot unusual geometric features (bond lengths, valence and dihedral angles, voids);
  - The most unusual structures will be forwarded to a (T)COD reviewer Web forum for verification;
  - Convincing evidence confirms validity of unusual structures.
- The set of usual and verified unusual structures should be used for reliable scientific inferences, unusual structures requiring attention.

Integration of (T)COD and AiiDA

- AiiDA, http://www.aiida.net
  - Automated interactive infrastructure and database for atomistic simulations [4];
  - An engine for automation of computations and storage of full data provenance;
  - Employs a high-level plugin interface;
  - Support extendable to all command line interface-based codes;
  - Seamless integration with high-performance computing clusters.

- COD + AiiDA + TCOD:
  - Direct download of input data and storage of computation results;
  - Full provenance of computations is recorded in CIF format and stored in TCOD together with results.
- Example: http://www.crystallography.net/tcod/10000001.html
  - Describes $BaTiO_3$ structure, relaxed with *Quantum ESPRESSO* [5];
  - Contains input and output files of the computation as well as an importable subset of AiiDA database.

Data selection options

- Bibliography, cell parameters and composition;
- Queries for substructure formulae can be submitted by drawing substructures with Web browser applet or entering SMILES [6];
- Resource Description Framework (RDF) descriptors are present for structures to facilitate SPARQL queries [7].

Conclusions

- COD and TCOD open a possibility for cross-validation of experimental-theoretical data;
- CIF format proves to be flexible for description of theoretically computed structures together with input data and code;
- Integration with AiiDA makes automatic collection of metadata for preserving the data provenance straightforward.

Bibliography

[1] Gražulis et al. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research*, 40(D1):D420–D427, Jan 2012.

[2] Hall et al. The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A*, 47(6):655–685, Nov 1991.

[3] Day. *Automated Analysis and Validation of Open Chemical Data*. PhD thesis, University of Cambridge, nov 2008.

[4] Pizzi et al. AiiDA: Automated Interactive Infrastructure and Database for Computational Science. arXiv:1504.01163.

[5] Giannozzi et al. Quantum ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of Physics: Condensed Matter*, 21(39):395502, 2009.

[6] Anderson et al. SMILES: A line notation and computerized interpreter for chemical structures. Technical report, Environmental Research Laboratory-Duluth, 1987.

[7] Prud'hommeaux et al. SPARQL Query Language for RDF. Technical report, W3C, 2008.

This research is funded by the SCIEX Fellowship grant No. 13.169

On-line version of the poster: http://j.mp/1Ag7Wm0

andrius.merkys@gmail.com