Theoretical Crystallography Open Database – open-access repository of theoretically computed crystal structures

Andrius Merkys[a,b], Giovanni Pizzi[a], Nicola Marzari[a] and Saulius Gražulis[b]

[a]Theory and Simulation of Materials (THEOS) and National Center for Computational Design and Discovery of Novel Materials (MARVEL), Ècole 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

Unprecedented developments of electronic structure methods, the increase of computer power and the decrease of the price/performance ratio brought about the era of high-throughput crystal structure optimization and property calculation. The Theoretical Crystallography Open Database (TCOD) [1] has been launched in order to collect the results of calculations, performed by the plethora of theoretical calculation groups, into an open-access resource. TCOD, together with the large set of experimental structures in the Crystallography Open Database (COD) [2], opens the possibility for experimental-theoretical data cross-validation. To ensure the homogeneity of calculation results, ontologies are offered in the form of Crystallographic Information Framework (CIF) format [3] dictionaries. Here we present the structure of the TCOD, its CIF dictionaries, and discuss the ongoing efforts for the integration of TCOD with the AiiDA framework [4, 5].

COD & TCOD

- COD, http://www.crystallography.net/cod/
  - Largest to date curated open-access collection of small to medium sized unit cell crystal structures;
  - Harvests data from open journals, accepts depositions via automatic data submission site;
  - Performs routine automatic quality checks on all incoming structures.
- TCOD, http://www.crystallography.net/tcod/
  - An open-access resource of theoretical calculation results;
  - Based on the infrastructure of the COD;
  - Stores supplementary material of published research as well as prepublication and personal communication material;
  - Aims to save metadata for data provenance and reproducibility.

Crystallography Open Database

Theoretical Crystallography Open Database

Dictionaries

- Aim at automated checks for convergence, computational quality and reproducibility;
- Enable automated deposition and data mining;
- Accessible at:
  - Main TCOD dictionary (all `_tcod_*` tags):
    - http://www.crystallography.net/tcod/cif/dictionaries/cif_tcod.dic
    - svn://www.crystallography.net/tcod/cif/dictionaries/cif_tcod.dic
  - DFT dictionary (all `_dft_*` tags):
    - http://www.crystallography.net/tcod/cif/dictionaries/cif_dft.dic
    - svn://www.crystallography.net/tcod/cif/dictionaries/cif_dft.dic
- Open mailing list for discussions:
  - http://lists.crystallography.net/cgi-bin/mailman/listinfo/tcod

Levels of structure description

Level 2

Level 1

Level 0

- lattice and symmetry
  `_cell_length_{a,b,c}`
  `_cell_angle_{alpha,beta,gamma}`
  `_symmetry_space_group_name_{H-M,Hall}`
  ...
- atomic coordinates
  `_atom_site_fract_{x,y,z}`
  ...
- bibliography reference
  `_journal_{name_full,paper_doi}`
  ...
- computational setup & parameters
  `_tcod_software_{package,compiler,library}`
  `_tcod_software_runtime_{CPU,OS}`
  `_dft_{XC_functional,basisset,pseudopotential}_type`
  ...
- residual forces on atoms and cell
  `_tcod_atom_site_residual_force_fract_{x,y,z}`
  ...
- code-specific convergence criteria
  `_dft_cell_{energy,density,potential}_conv`
  ...
- input scripts and files
  `_tcod_file_{name,URI,contents,role,interpreter}`
  ...
- command line
  `_tcod_computation_{command,environment}`
  ...
- output logs of the code
  `_tcod_computation_{log_file,stdout,stderr}`
  ...

Levels of structure description: rationale

Level 0 | Level 1 | Level 2

Quantity of machine-readable data
Reproducibility

- Software for converting calculation output files to CIF:
  - AiiDA (http://www.aiida.net) – level 0, is being extended to 2
  - vasp2cif (http://github.com/egplar/vasp2cif) – level 0

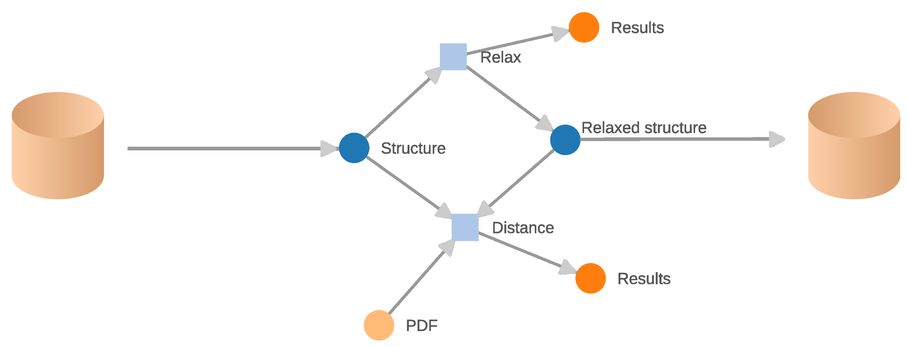
AiiDA

- AiiDA – Automated Interactive Infrastructure and Database for Atomistic simulations, http://www.aiida.net
- An engine for automation of calculations and storage of full data provenance;
- Employs a high-level plugin interface;
- Support extendable to all command line interface-based codes;
- Four pillars of AiiDA infrastructure:

ADES

Automation	Data	Environment	Sharing
Remote management Repository pipelines High-throughput	Storage Database Provenance	High-level workspace Scientific workflows Data analytics	Social Reusability Standards
Abstract away the low-level tasks to manage preparation, submission and processing of calculations	Management and persistence of the data produced by the simulations; database search and querying; simulation reproducibility	Natural, high-level environment for computational research; usable platform with shallow learning curve	Social ecosystem to foster interaction, share codes, data and scientific workflows, using standardized data formats

Integration of TCOD and AiiDA



experimental structure database — calculation environment — theoretical structure database

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

Conclusions

- CIF format proves to be flexible for description of theoretically calculated structures at any level of detail;
- Integration with AiiDA makes collection of metadata for preserving the data provenance straightforward.

Bibliography

[1] Gražulis et al. Theoretical crystallography open database, 2013. http://www.crystallography.net/tcod/.

[2] 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.

[3] 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.

[4] Marzari et al. AiiDA: Automated Interactive Infrastructure and Database for Atomistic simulations, 2012. http://www.aiida.net.

[5] Pizzi et al. AiiDA: an Infrastructure for Computational Material Science. To be published.

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

Online version of the poster: http://j.mp/1GBo8kb