Special Activities

MS-107: Robust programming for CIF, NeXus, and related file structures

The use of interconnected open data for material identification

Antanas Vaitkus[a], Andrius Merkys[a], Yassine El Mendili[b] and Saulius Gražulis[a]

[a]Vilnius University Institute of Biotechnology, Saulėtekio av. 7, LT-10257 Vilnius, Lithuania
[b]Normandie Université, CRISMAT-ENSICAEN, UMR CNRS 6508, Université de Caen Normandie, 14050 Caen, France

Introduction

Open access data repositories play an increasingly important role in the scientific community. The Crystallography Open Database (COD) [1] is one such resource – over the last 13 years it has become the largest curated and validated open access collection of small molecule crystal structures. Recently, a number of computer programs capable of automatically determining stoichiometrically [2] and chemically sound molecules from the crystallographic data have been developed; this, in turn, enabled the automated generation of structural formulae descriptors and eased the establishment of cross-links between the COD and other open access resources. New strides have also been made in relating spectral data to their corresponding crystal structures. The COD was chosen as the back-end database in the wide scale on-site sample analysis of the "Sonic Drilling coupled with Automated Mineralogy and chemistry On-Line-On-Mine-Real-Time" (SOLSA, `http://www.solsa-mining.eu`) project that focuses on developing highly efficient, cost-effective and sustainable exploration technologies. Since part of the sample analysis involves material identification via the means of Raman spectroscopy, reference spectra aggregation from various sources was carried out choosing CIF as the homogeneous data carrier format for both XRD and spectral data; this, in turn, stipulated the development of spectroscopy oriented CIF dictionary. These new developments of the CIF dictionaries will allow the SOLSA project to present various aspects of mineral characterization such as Raman spectra, XRD structures and fluorescence data in the COD database in a uniform, computer-readable way.

Crystallography Open Database (COD)

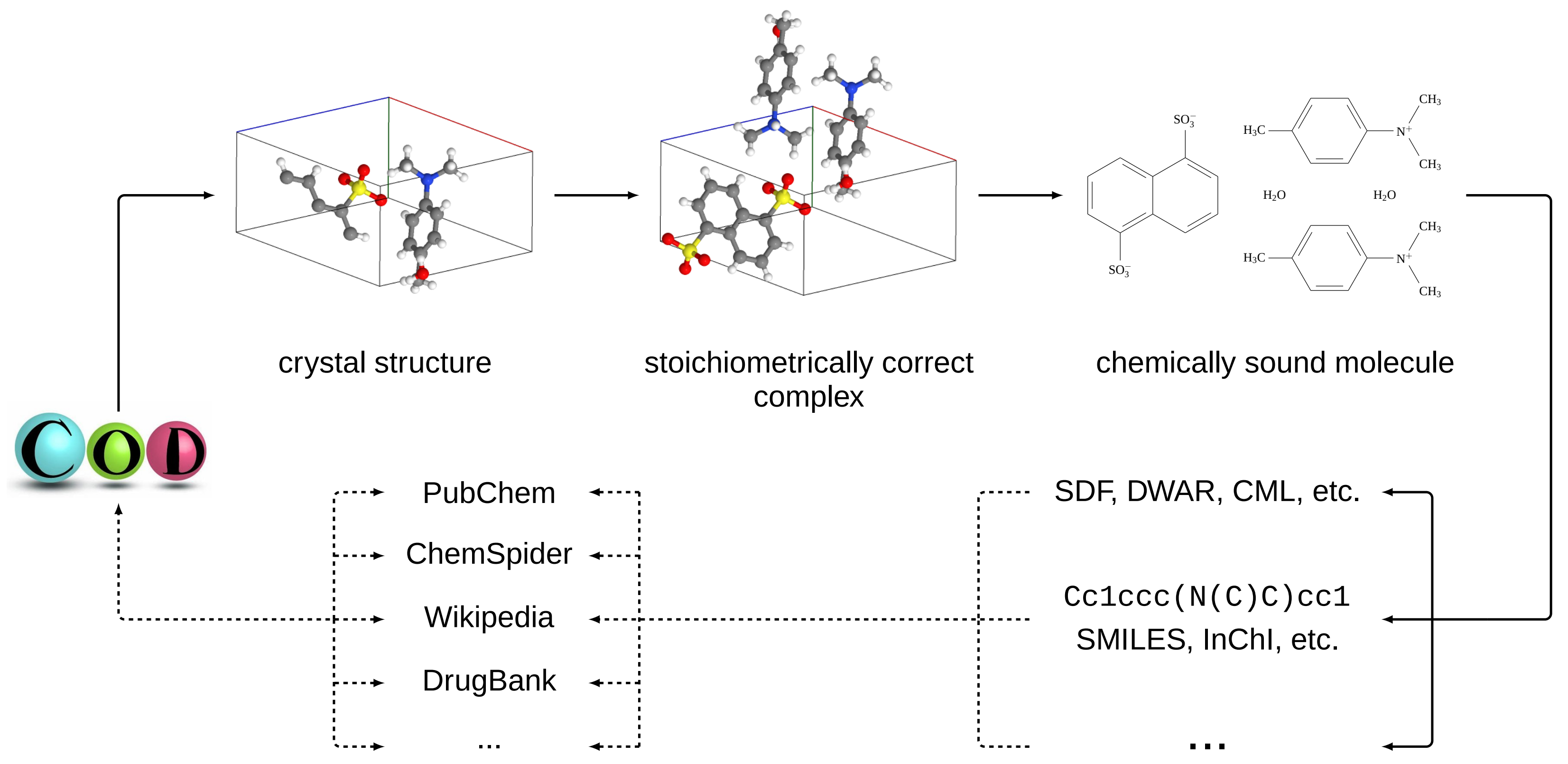
`http://www.crystallography.net/cod`

- ▶ Open-access;
- ▶ Contains small-molecule organic, inorganic, and metal-organic crystal structures;
- ▶ Uses CIF [3] as the carrier format;
- ▶ Over 380 000 entries.

Chemical information in the COD

- ▶ Periodically generated from crystallographic data in an automated way using open-source software [4, 2, 5];
- ▶ Enables a more efficient substructure search;
- ▶ Used to establish cross-links to other resources;
- ▶ Available in its entirety as a DataWarrior [6] file at `http://www.crystallography.net/dwar`

Chemical information extraction

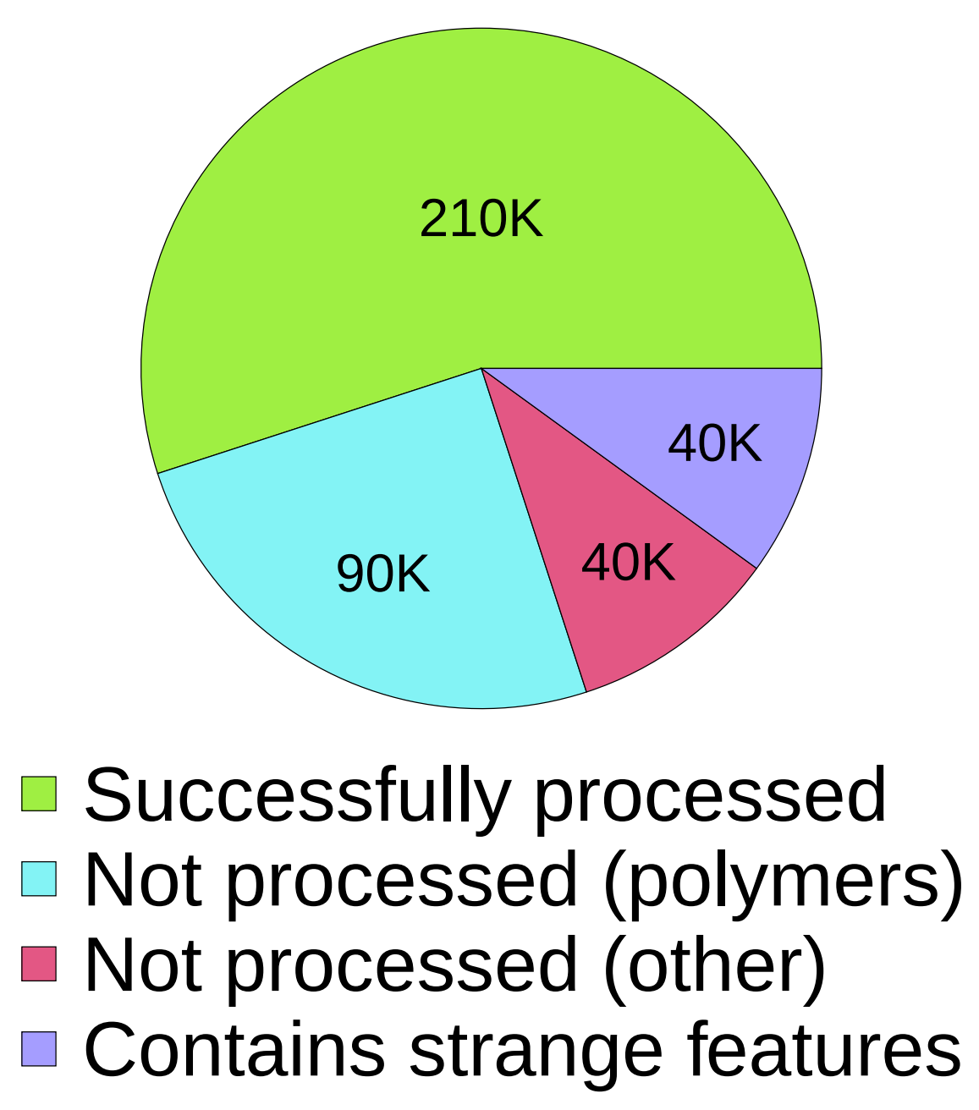


From crystals to molecules

Challenges:

- ▶ Lack of an unambiguous way to divide polymeric molecules into monomers;
- ▶ The core CIF dictionary does not provide sufficient means of detailing chemical properties;
- ▶ Certain discrepancies in the input crystallographic data require manual curation (e. g. unmarked disorder sites).

Overview of the COD processing



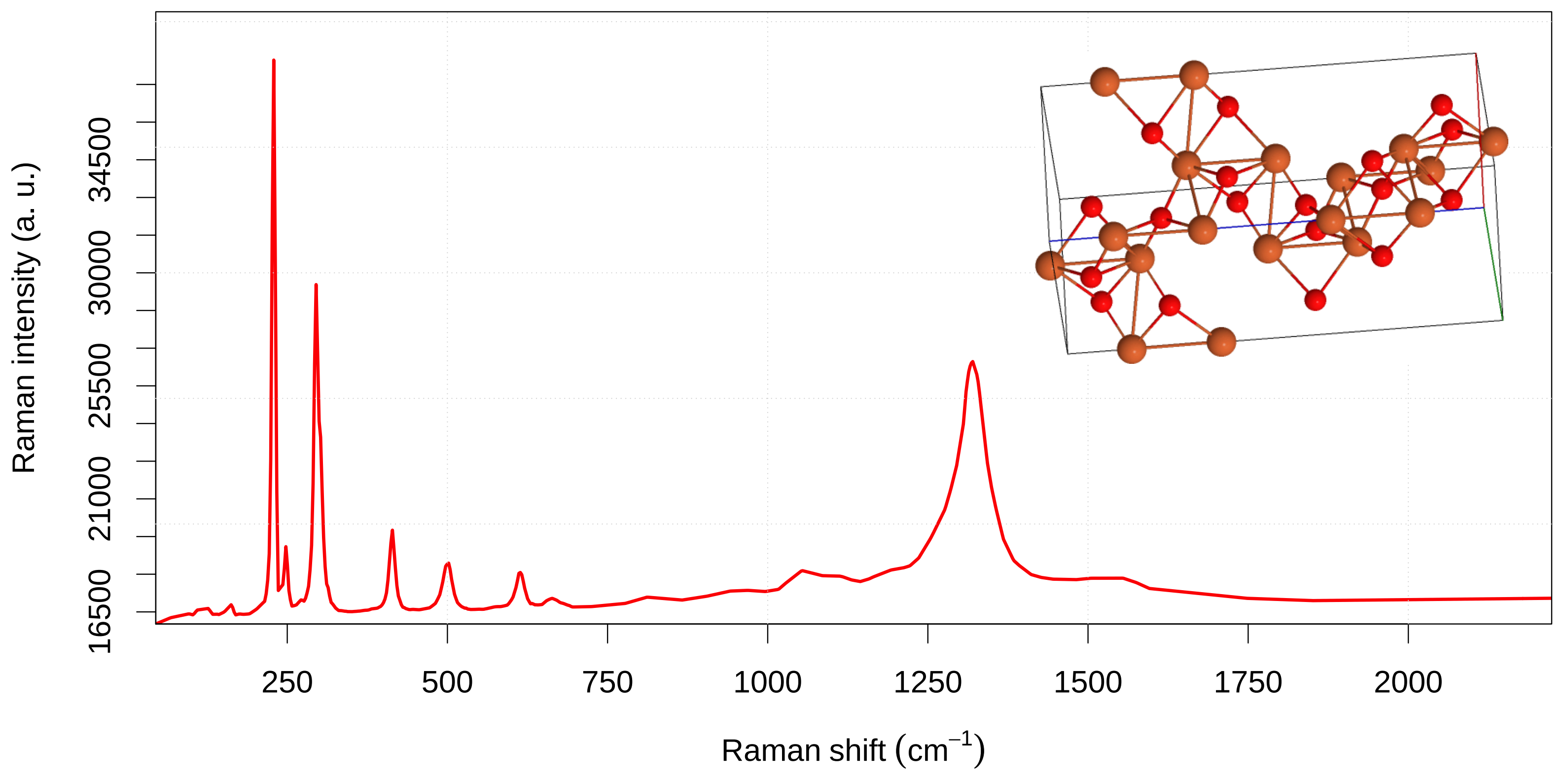
from a total of over a 380 000 structures.

Raman Open Database (ROD)

`http://solsa.crystallography.net/rod`

- ▶ Open access;
- ▶ Contains Raman spectroscopy data;
- ▶ Uses CIF2 [7] as the carrier format;
- ▶ Validates input files upon deposition;
- ▶ Spectral data is cross-linked with the XRD data in the COD;
- ▶ Used in the SOLSA project for material identification.

COD on a ROD



The Raman spectra and the related crystal structure of hematite. ROD ID 1000001, COD ID 1546383.

Raman Spectroscopy Ontology

- ▶ Developed and maintained by an international team of Raman spectroscopy experts;
- ▶ Expressed as a DDLm [8] conforming CIF dictionary;
- ▶ Latest version available at `http://solsa.crystallography.net/rod/cif/dictionaries/cif_raman.dic`

Conclusions

- ▶ The COD is now enhanced with chemical data and interconnected with other open access resources;
- ▶ There is a need for a curated set of Raman spectroscopy data;
- ▶ CIF2 is suitable format for storing scientific data of all sorts;
- ▶ Open access data repositories are a viable alternative to proprietary databases.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 689868.

References

[1] S. 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, 2012.

[2] S. Gražulis et al. Computing stoichiometric molecular composition from crystal structures. *Journal of Applied Crystallography*, 48:85–91, 2015.

[3] S. R. Hall et al. The Crystallographic Information File (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A*, 47(6):655–685, 1991.

[4] A. Merkys et al. COD::CIF::Parser: an error-correcting CIF parser for the Perl language. *Journal of Applied Crystallography*, 49(1):292–301, 2016.

[5] OpenChemLib. Open source Java-based chemistry library. `https://github.com/Actelion/openchemlib`.

[6] T. Sander et al. DataWarrior: An open-source program for chemistry aware data visualization and analysis. *Journal of Chemical Information and Modeling*, 55(2):460–473, 2015.

[7] H. J. Bernstein et al. Specification of the Crystallographic Information File format, version 2.0. *Journal of Applied Crystallography*, 49(1):277–284, 2016.

[8] N. Spadaccini et al. DDLm: A new dictionary definition language. *Journal of Chemical Information and Modeling*, 52(8):1907–1916, 2012.

Antanas Vaitkus has no conflict of interest.
Andrius Merkys has no conflict of interest.
Saulius Gražulis is a member of the COD Advisory Board.

On-line version of the poster:
`http://j.mp/2vXMfXP`