Improvements to the data search and validation functionality in the Crystallography Open Database

Antanas Vaitkus[1], Andrius Merkys[1], Algirdas Grybauskas[1], Aleksandras Konovalovas[1], Miguel Quirós Olozábal[2] and Saulius Gražulis[1,3]

[1]Vilnius University, Life Sciences Center, Institute of Biotechnology, Saulėtekio av. 7, LT-10257 Vilnius, Lithuania
[2]Departamento de Química Inorgánica, Universidad de Granada, 18071, Granada, Spain
[3]Vilnius University, Faculty of Mathematics and Informatics, Naugarduko 24, LT-03225 Vilnius, Lithuania

Email: antanas.vaitkus@bti.vu.lt

Crystallography Open Database (COD)

https://www.crystallography.net/cod

- Open-access FAIR [1] repository of small molecule crystal structures.
- Data can be reused without any additional restrictions (CC0 license).
- Covers organic, inorganic, organometallic compounds and minerals.
- More than 475 000 entries and growing.

Validation using the CIF framework

The COD is an actively curated database that heavily utilises the CIF framework for its data maintenance tasks. Recent CIF-related innovations by the IUCr stipulated the development of several notable improvements to the COD software:

- **CIF 2.0 parser.** The COD ingests and disseminates crystallographic data using the CIF 1.1 [2] format. To aid in this purpose the COD team developed a specialised error-correcting CIF parser [3] which has now been updated to enable the processing of CIF 2.0 [4] files.
- **DDLm validator.** COD data are routinely validated against the official IUCr DDL1 dictionaries. With the introduction of the new generation DDLm language [5] that offers a more robust way of describing data, the COD validation software [6] was updated to handle both the DDL1 and the DDLm dictionaries.
- **DDL development tools.** Official deprecation of the DDL1 language has created the need to upgrade the existing DDL1 dictionaries. To ease this task the COD team has created a set of tools capable of migrating, comparing and checking DDL1 and DDLm dictionaries. Some of these tools are employed in the official IUCr dictionary development repositories.

Usage example:

- Validate a CIF file against a DDLm dictionary:

```
cif_validate --ddlm-add-dictionary cif_core.dic 1000000.cif
```

- Check a DDLm dictionary against a set of best practices:

```
cif_ddlm_dic_check cif_core.dic
```

The described CIF and DDL handling tools are distributed as part of the open-source `cod-tools` software package.

Validation issue database

CIF validation messages collected from the COD are stored in a publicly available relational database that can be accessed:

- By using a `RestfulDB` GUI at https://sql.crystallography.net/db/cod_validation.
- By connecting directly using a `MySQL` client:

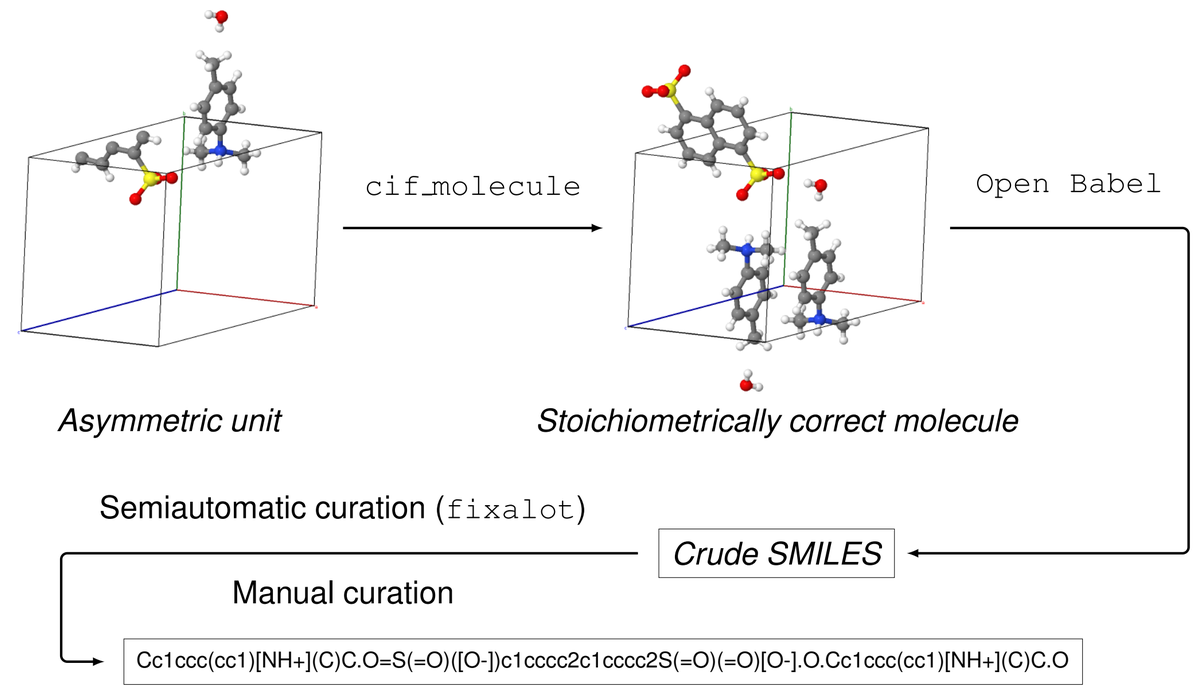
```
mysql -u cod_reader -h sql.crystallography.net cod_validation;
```

Chemical structure search

Chemical (sub)structure search in the COD is enabled by a set of high-quality manually curated SMILES strings that:

- Covers more than 40% of COD entries and is continuously updated.
- Is available under the same license as the COD CIF files (CC0).
- Follows additional conventions that are extensively described in a peer-reviewed publication [7].

Simplified SMILES generation workflow



Search using the OPTIMADE API

It is extremely beneficial to be able to access information from multiple materials databases as they often differ in fidelity and focus across material classes and properties. However, retrieving data from multiple databases is difficult as each database has its own specific application programming interface (API). The OPTIMADE consortium aims to simplify such federated queries by defining a common RESTful OPTIMADE API [8, 9] based on the JSON:API specification [10].

COD team is a member of the OPTIMADE consortium and participates in the development of the OPTIMADE API. COD is the first experimental database that implements the OPTIMADE API to query records created using experimental and hybrid techniques (XRD, EM).

OPTIMADE API query examples:

- Retrieve entries of materials that contain Ag or Au atoms:

  https://www.crystallography.net/cod/optimade/v1/structures?filter=elements HAS ANY "Ag", "Au"

- Retrieve entries of binary materials consisting of Ag and Au atoms:

  https://www.crystallography.net/cod/optimade/v1/structures?filter=elements HAS ALL "Ag", "Au" AND nelements=2

Conclusions

- The COD team develops open-source software that can be used to manipulate and validate CIF files.
- The COD enhances crystallographic information with expert-curated SMILES strings.
- The COD team is one of the developers and early adopters of the OPTIMADE API.

Acknowledgements

This research has received funding from the Research Council of Lithuania under grant agreement No. MIP-20-21.

References

[1] M. D. Wilkinson et al. *Scientific Data*, 3(1), 2016. doi:10.1038/sdata.2016.18.
[2] S. R. Hall et al. *Acta Crystallographica Section A*, 47(6):655–685, 1991. doi:10.1107/S010876739101067X.
[3] A. Merkys et al. *Journal of Applied Crystallography*, 49(1):292–301, 2016. doi:10.1107/S1600576715022396.
[4] H. J. Bernstein et al. *Journal of Applied Crystallography*, 49(1):277–284, 2016. doi:10.1107/S1600576715021871.
[5] N. Spadaccini et al. *Journal of Chemical Information and Modeling*, 52(8):1907–1916, 2012. doi:10.1021/ci300075z.
[6] A. Vaitkus et al. *Journal of Applied Crystallography*, 54(2):661–672, 2021. doi:10.1107/S1600576720016532.
[7] M. Quirós et al. *Journal of Cheminformatics*, 10(1), 2018. doi:10.1186/s13321-018-0279-6.
[8] C. Andersen et al. The OPTIMADE specification, 2020. https://doi.org/10.5281/zenodo.4251947.
[9] C. W. Andersen et al. *Scientific Data*, 8(1), 2021. doi:10.1038/s41597-021-00974-z.
[10] JSON:API v1.0. https://jsonapi.org/format/1.0/.

Antanas Vaitkus has no conflict of interest.
Andrius Merkys has no conflict of interest.
Algirdas Grybauskas has no conflict of interest.
Aleksandras Konovalovas has no conflict of interest.
Miguel Quirós Olozábal has no conflict of interest.
Saulius Gražulis has no conflict of interest.

On-line version of the poster:
https://bit.ly/3lTbnLI