FAIR data in the Crystallography Open Database: curation, validation and (re)use

Andrius Merkys[1], Antanas Vaitkus[1], Algirdas Grybauskas[1] and Saulius Gražulis[1]

[1]Vilnius University, Life Sciences Center, Institute of Biotechnology, Saulėtekio av. 7, LT-10257 Vilnius, Lithuania

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 500 000 entries and growing.

Validation using the CIF framework

The COD is an actively curated database that heavily utilises the CIF framework [2] 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 format. To aid in this purpose the COD team developed a specialised error-correcting CIF parser [3] which is now also able to process CIF 2.0 files.
- **DDLm validator.** COD data are routinely validated against the official IUCr DDL1 dictionaries. With the introduction of the new generation DDLm language, the COD validation software [4] 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. The COD team has created a set of tools for 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.

Data query and access

Query methods:

- Cell parameters and bibliography (Web form)
- Chemical (sub)structure search (Web form) in a set of high-quality manually curated SMILES strings that:
  - Covers more than 44% 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 [5].
- OPTIMADE interface (common API for structural databases [6])
  - Example: select ternary structures having at least one of C, Si, Ge or Sn, but not Pb:

```
curl -L https://www.crystallography.net/cod/optimade/structures \
   -F 'filter=elements HAS ANY "C", "Si", "Ge", "Sn" AND
   NOT elements HAS "Pb" AND elements LENGTH 3'
```

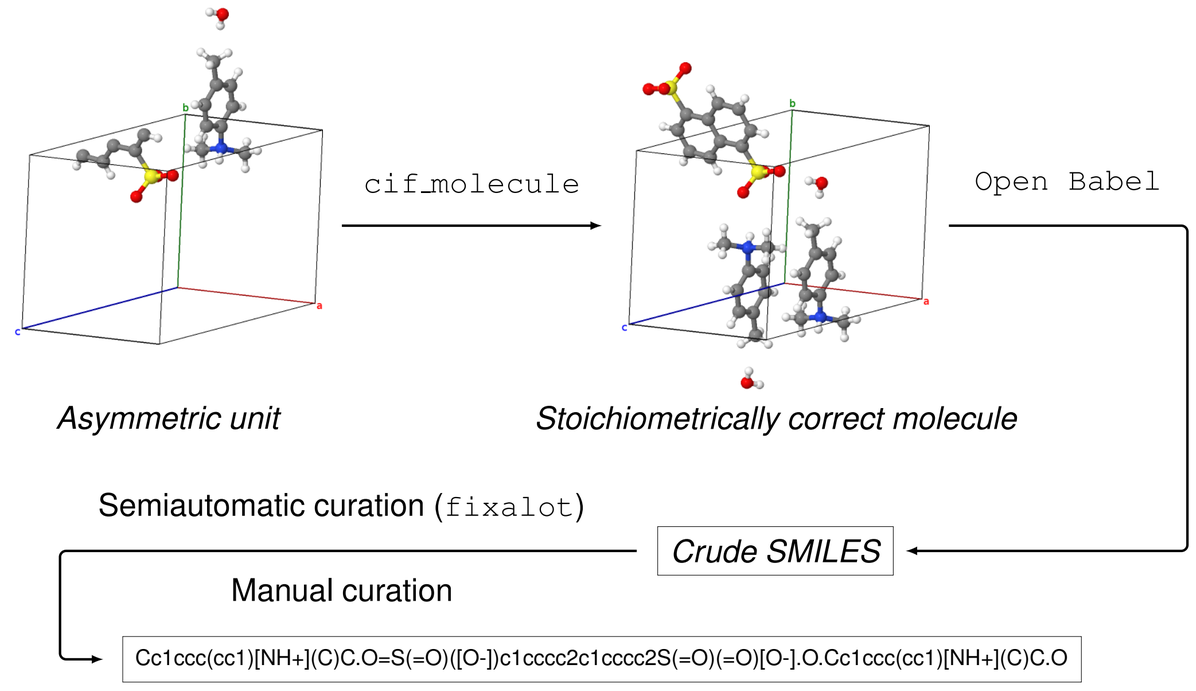
- SQL access
  - Example: count records in the COD:

```
mysql -u cod_reader -h sql.crystallography.net cod -e 'select count(*) from data'
```

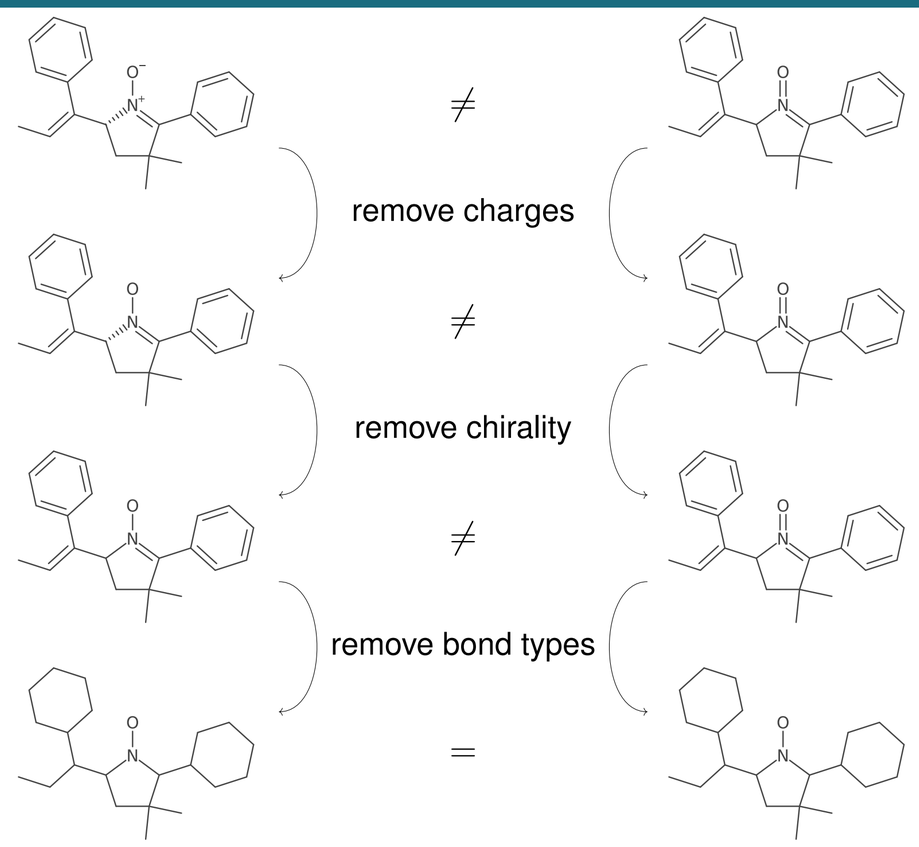
File access methods:

- Single entry/bulk download via HTTP(S)
- rsync
- FTP
- Subversion

Simplified SMILES generation workflow



Workflow of chemical comparison



- Stripping chemical attributes until match is found.
- Marking nonmatching structures for further review.

Results of chemical comparison

Source #1	Source #2	No. of pairs	Matches
Coordinate-derived	Chemical names	39 636	92%
Chemical names	Expert-curated [5]	34 670	94%
Coordinate-derived	Expert-curated [5]	188 137	92%

- Analysis of several mismatches helped to identify incomplete or incorrect published chemical annotations [7].

Conclusions

- The COD team develops open-source software that can be used to manipulate and validate CIF files.
- There are multiple ways to query and obtain the data from the COD.
- The COD team enhances the COD data with chemical information.

References

[1] Wilkinson et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), 2016.

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

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

[4] Vaitkus et al. Validation of the Crystallography Open Database using the Crystallographic Information Framework. *Journal of Applied Crystallography*, 54(2):661–672, 2021.

[5] Quirós et al. Using SMILES strings for the description of chemical connectivity in the Crystallography Open Database. *Journal of Cheminformatics*, 10(1), 2018.

[6] Andersen et al. OPTIMADE, an API for exchanging materials data. *Scientific Data*, 8(1), 2021.

[7] Merkys et al. Graph isomorphism-based algorithm for cross-checking chemical and crystallographic descriptions. *Journal of Cheminformatics*, 15(1), feb 2023.

This research has received funding from the Research Council of Lithuania under grant agreements MIP-20-21 and MIP-23-87.

On-line version of the poster: https://bit.ly/3BKZ5vG

andrius.merkys@gmc.vu.lt