Chemical annotation in the Crystallography Open Database

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

[1]Vilnius University Life Sciences Center, Saulėtekio 7, LT-10257 Vilnius, Lithuania
[2]Departamento de Química Inorgánica, Universidad de Granada, 18071, Granada, Spain

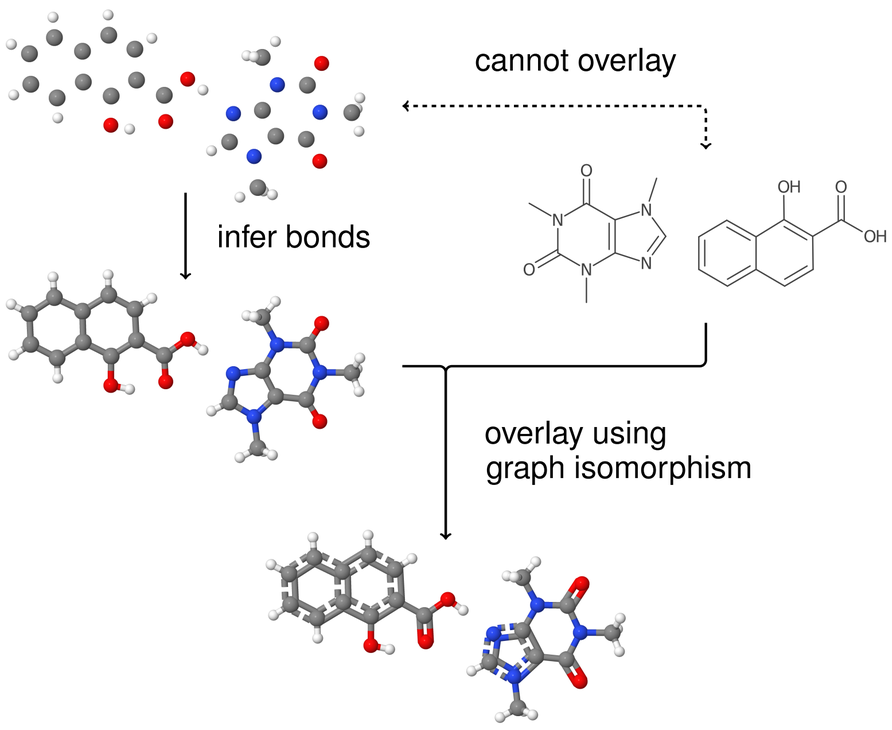
Abstract

Reliable knowledge about structure and properties of chemical compounds is essential for many branches of science. The most accurate data about the structure of molecules are obtained from X-ray crystallographic (XRC) analyses. These data are not immediately usable by chemists, as XRC does not detect atomic charges, bond types or the presence of lone electrons in radicals. All such information needs to be inferred from the crystallographic data, either manually [1], or using heuristics, implemented as computer programs [2, 3]. The existing programs rarely consider information other than the coordinates and their heuristics are usually specifically tailored for organic molecules. Thus the derivation of chemical annotations by these programs is not always reliable, especially for metal-organic complexes. Atomic coordinates in crystal structure reports are usually accompanied by additional chemical information, such as systematic chemical names and connectivity details, albeit mostly in forms not suitable for automated overlaying on the coordinate data. All this information could be employed to annotate crystallographic data with chemical details provided the mapping between different representations is known. The largest open access crystallographic database, the Crystallography Open Database (COD, [4]), contains computer readable chemical descriptions for nearly half of its entries [1]. Currently, these descriptions are not linked to particular atoms in crystals, thus studies that require the combined crystallographic and chemical information have to infer the correspondence on their own. Graph-based algorithms could be used to supplement the COD with the information about such missing links. Open-access nature of the COD allows dissemination of this information under FAIR (Findability, Accesibility, Interoperability and Reusability [5]) principles on the Web, immediately enabling numerous computational searches and research by pharmaceutical companies and academic groups.

Problem

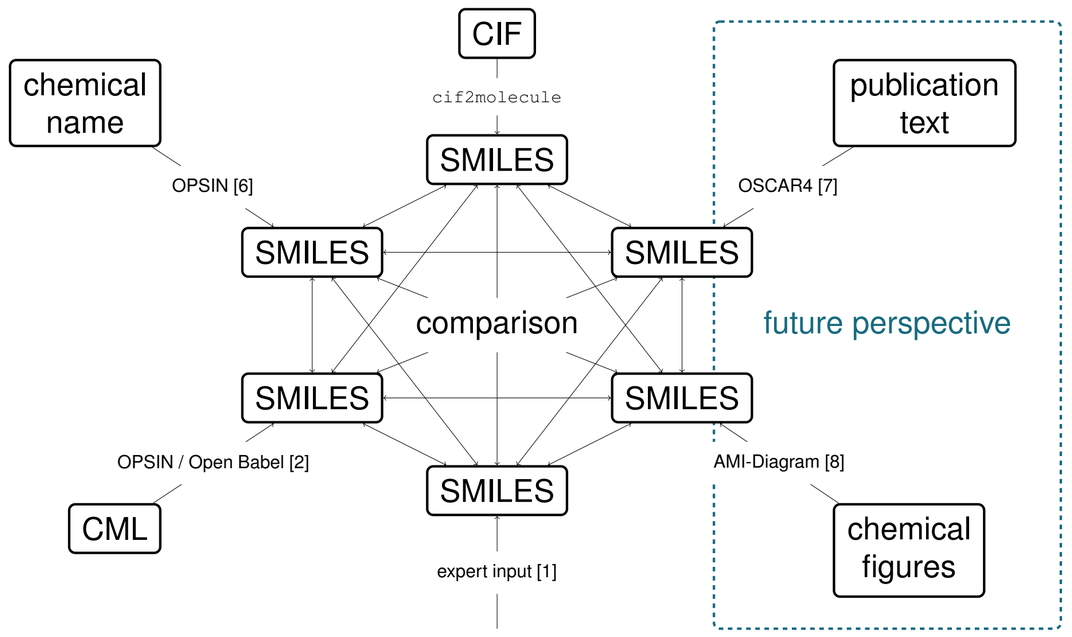
- Papers contain crystallographic and chemical descriptions of compounds;
- Information is scattered in items of different formats:
  - coordinates in CIF;
  - systematic chemical name in CIF;
  - systematic chemical name in publication title;
  - Chemical Markup Language (CML) files;
  - SMILES;
  - figures with chemical structures;
  - ...
- No automated methods exist to interrelate these descriptions.

Overlaying crystallographic and chemical annotations

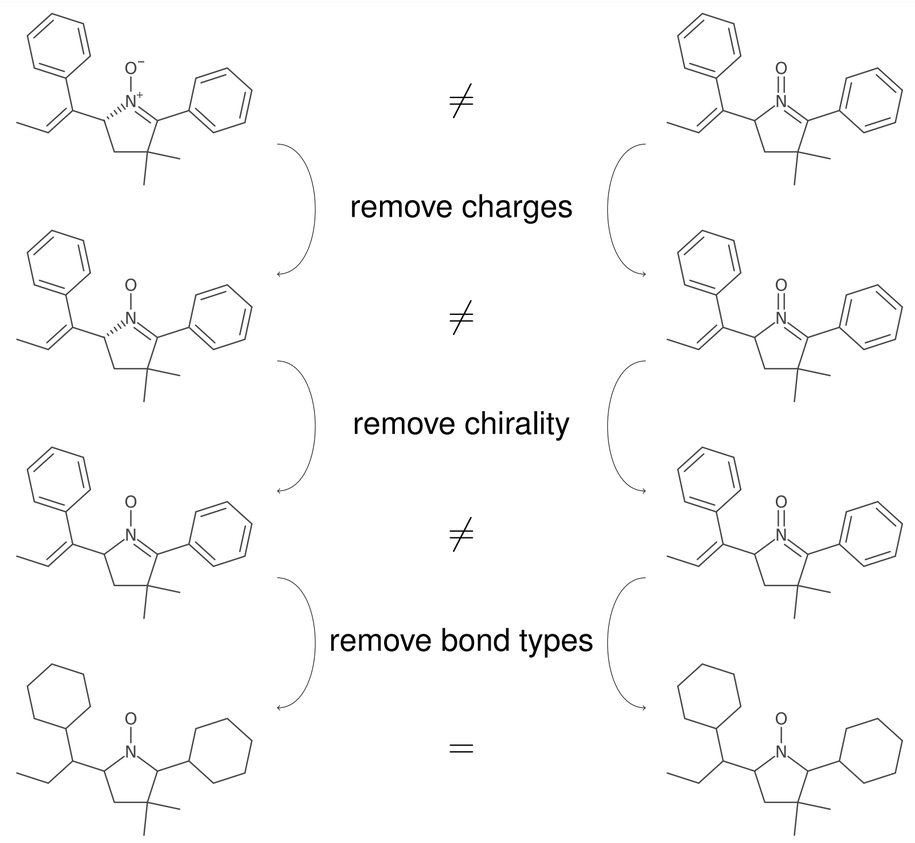


1. Connectivity is inferred from the coordinates;
2. Crystal contents are broken down into moieties;
3. Moieties of compared crystals are matched;
4. Corresponding moieties are overlayed.

Workflow



Comparison



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

Results

Source #1	Source #2	No. of pairs	Matches
Coordinate-derived	Chemical names	38 640	88%
Coordinate-derived	CML	1551	89%
Coordinate-derived	Expert-curated [1]	187 935	57%

- Analysis of a couple dozens of mismatches identified incomplete or incorrect published chemical annotations.

Challenges to address

- Polymer molecules are difficult to process;
- More interesting traits are dominated by differences in notation:
  - aromatic form vs. Kekulé form;
  - marked vs. unmarked metal coordination [9].

Bibliography

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

[2] O'Boyle et al. Open data, open source and open standards in chemistry: The Blue Obelisk five years on. *Journal of Cheminformatics*, 3:37, 2011.

[3] Willighagen et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of Cheminformatics*, 9(1), Jun 2017.

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

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

[6] Lowe et al. Chemical name to structure: OPSIN, an open source solution. *Journal of chemical information and modeling*, 51:739, 2011.

[7] Jessop et al. OSCAR4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, 3(1):41, Oct 2011.

[8] Murray-Rust et al. AMI-diagram: Mining facts from images. *D-Lib Magazine*, 20(11/12), Nov 2014.

[9] Clark. Accurate specification of molecular structures: The case for zero-order bonds and explicit hydrogen counting. *Journal of Chemical Information and Modeling*, 51(12):3149–3157, Dec 2011.

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

On-line version of the poster:
https://j.mp/3fWZHDS

andrius.merkys@gmc.vu.lt