Quality control of the chemical information in the Crystallography Open Database

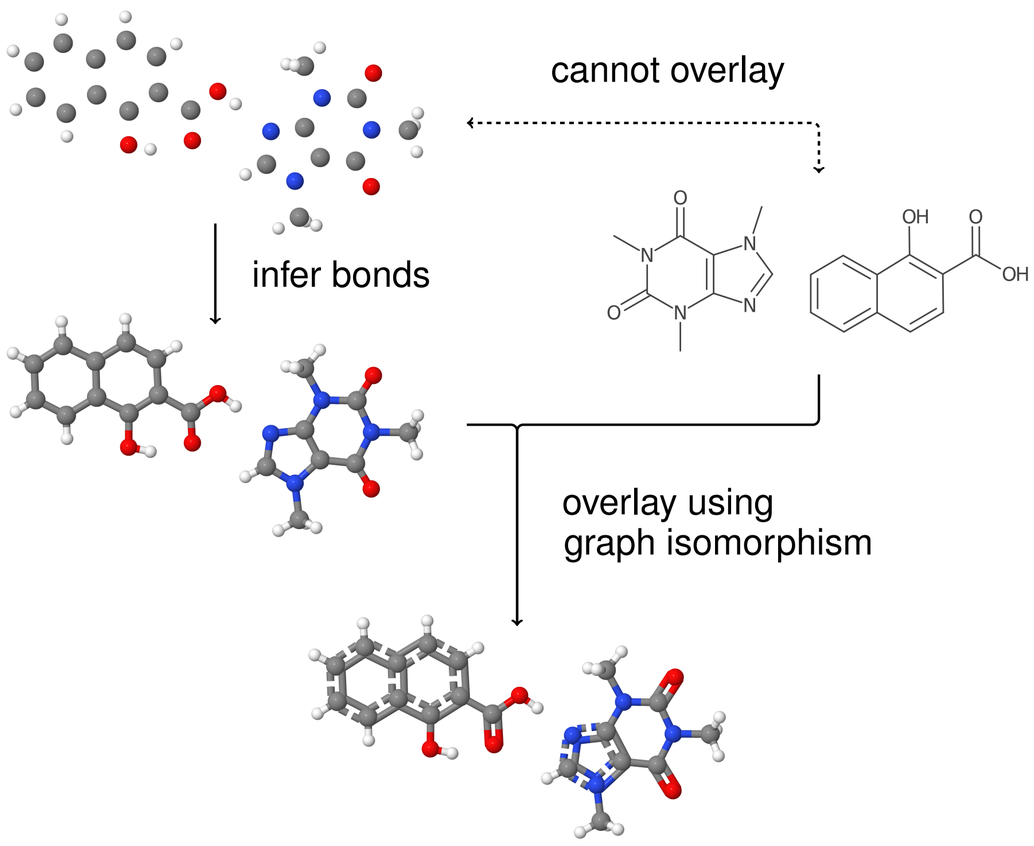
Andrius Merkys[1], Antanas Vaitkus[1], Eglė Šidlauskaitė[1], Miglė Urbonaitė[2], Algirdas Grybauskas[1] and Saulius Gražulis[1]

[1]Vilnius University Life Sciences Center, Saulėtekio 7, LT-10257 Vilnius, Lithuania
[2]Companial, Sporto g. 7A, LT-09238 Vilnius, Lithuania

Abstract

- Crystallography Open Database (COD, [1]) is the largest open access crystallographic database, housing over 500k crystal structure entries.
- However, with the advent of machine learning methods and the increased reliance on black box approaches [2] quality control becomes vital to ensure:
  - integrity and consistency of both chemical and crystallographic descriptions;
  - presence and correctness of chemical representations such as SMILES [3] and chemical names;
  - correctness of chemical connectivity.
- Solutions show the potential to improve the data quality in the COD:
  - cross-checking of crystallographic structures and chemical representations detects mismatches due to potential errors [4];
  - generated chemical names provide a reference point for name analysis and comparison;
  - derived covalent radii table provides insight into the choice of cutoff values [5].

Overlaying crystallographic and chemical annotations



1. Connectivity is inferred from the coordinates (CIF files);
2. Crystal contents are broken down into molecular entities;
3. Chemical descriptions are extracted from chemical names, CML files and SMILES;
4. Molecular entities of compared crystals are matched;
5. Corresponding molecular entities are overlaid.

Cross-check results

Source #1	Source #2	No. of pairs	Matches
Coordinate-derived	Chemical names	39 636	88%
Coordinate-derived	CML	1551	89%
Coordinate-derived	Expert-curated [3]	188 137	85%
Chemical names	CML	1533	97%
Chemical names	Expert-curated [3]	34 670	92%

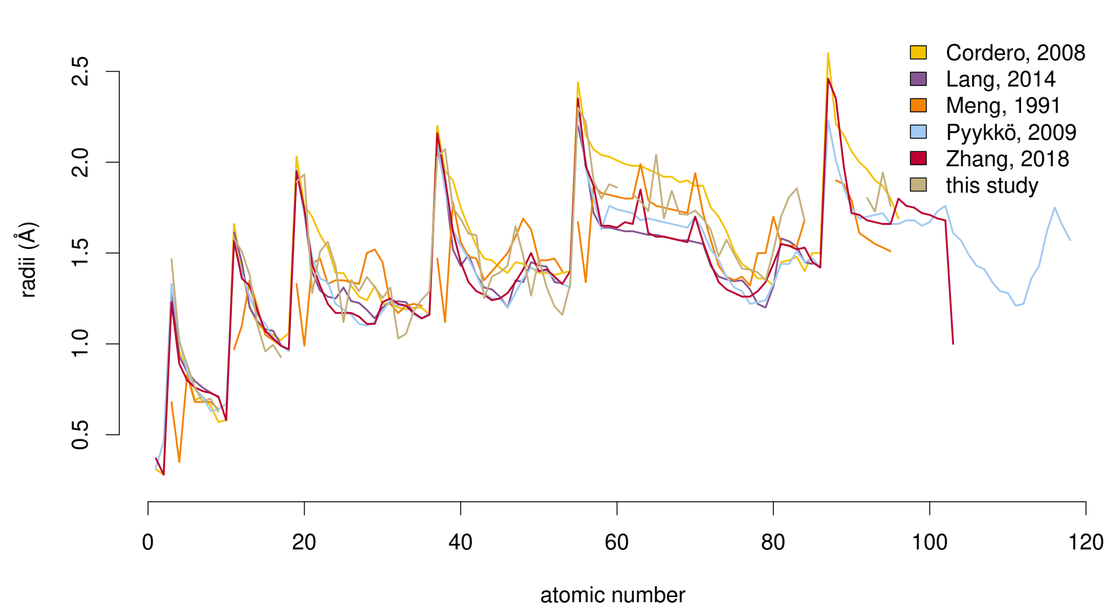
- Analysis of a couple dozens of mismatches identified incomplete or incorrect published chemical annotations [4].
- More interesting traits are dominated by differences in notation:
  - aromatic form vs. Kekulé form;
  - marked vs. unmarked metal coordination [6].

This research has received funding from the Research Council of Lithuania under Grant agreement No. MIP-23-87

On-line poster material:
https://bit.ly/3X4D8Di

Unsupervised method to derive a covalent radii table

- Uses Voronoi tessellation to find possible direct neighbours;
- Analyses distance distributions to locate van der Waals gap;
- Solves a system of equations to estimate covalent radii;
- Resulting radii table closely follows general trends of published tables.



Validation of the derived covalent radii table

Compared to expert-assigned connectivity [3], the derived radii:

- tend to miss bonds involving I, Mn and N;
- mark false-positive Mo-Mo, Ti-Ti, V-V and W-W bonds, similarly as the table by Cordero et al. [7].

Generation of preferred IUPAC names (PINs)

Attempt to reproduce 3696 PINs from the IUPAC Blue Book [8]

	ChemOnomatopist v0.10.0	STOUT v2.0 [9]
Correct PIN	1321	1132
Alternative name	781	1874
Incorrect	987	690
Refused	607	0
Time	≈ 4 min.	≈ 230 min.

- ChemOnomatopist performs best with:
  - saturated and unsaturated acyclic and cyclic hydrocarbons;
  - bicyclic compounds, including heterocycles;
  - noncarbon acids.
- ChemOnomatopist needs improvement to handle:
  - multiplication nomenclature;
  - amides, amidines, esters and ethers;
  - diazenes, hydrazines, hydrazides and urea compounds;
  - charges and stereochemistry.
- STOUT most likely has been trained on older generation PINs.

Bibliography

[1] 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.

[2] *Nature*, 617(7961):438–438, May 2023.

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

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

[5] Šidlauskaitė. Determination of atomic radii from small-molecule crystal structures. Master's thesis, Vilnius, 2023.

[6] 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.

[7] Cordero et al. Covalent radii revisited. *Dalton Transactions*, pages 2832–2838, 2008.

[8] IUPAC. *Nomenclature of Organic Chemistry. IUPAC Recommendations and Preferred Names 2013. v3*, volume 1. IUPAC, Feb 2013. URL: https://iupac.qmul.ac.uk/BlueBook/PDF/BlueBookV3.pdf WWW: https://iupac.qmul.ac.uk/BlueBook/PDF/.

[9] Rajan et al. STOUT: SMILES to IUPAC names using neural machine translation. *Journal of Cheminformatics*, 13(1):1–14, 2021.

andrius.merkys@gmc.vu.lt