**Poster**

# Quality control of the chemical information in the Crystallography Open Database

**A. Merkys[1], A. Vaitkus[1], E. Šidlauskaitė[1], M. Urbonaitė[2], A. Grybauskas[1], S. Gražulis[1]**

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

*andrius.merkys@gmc.vu.lt*

Crystallographic databases, such as the Crystallography Open Database (COD [1], https://www.crystallography.net/cod/), are important resources for structural knowledge of chemical compounds. With the advent of machine learning methods, quality control in datasets becomes more and more important due to the increased reliance on black box approaches [2]. Here we present the recent developments in increasing the data quality in the COD.

Publications in crystallographic journals usually are accompanied by machine-readable representations of chemical comprehension of the described structures. In order to spot irregularities between these representations and crystallographic structures we have developed a cross-checking method using canonical representations and isomorphism tests for molecular graphs [3]. This method not only detects mismatches, but as well attempts to resolve them by gradually relaxing the strictness of matching. Using this tool we have scanned almost 200k entries in the COD and detected over 31k entries with outright consistent descriptions and another 144k needing relaxing the matching strictness.

In the fields of both organic and inorganic chemistry, IUPAC nomenclature is used widely to identify chemical compounds. Out of over 500k COD entries, nearly 50k have author-provided systematic chemical names. In order to both validate the already existing chemical names and to present ones for the remaining COD entries, we have started to develop an open-source tool, *ChemOnomatopist*, capable of deriving preferred IUPAC names from chemical structures. The tool is already capable of naming branched acyclic compounds and some cyclic compounds.

While X-ray crystallography provides accurate 3D structure of molecules, it does not usually obtain the chemical connectivity. Most of the time distance-based heuristics are applied to identify bonded atoms, using published tables of covalent radii [4-6]. In order to evaluate the existing tables, we have developed an unsupervised workflow deriving covalent radii values from observed inter-atom distances between Voronoi neighbours in crystal structures in the COD. The radii values derived by this approach closely follow the trends of previously published tables [4-6] and can be updated with the advent of new input structural data.

All the described developments – cross-checking of chemical and crystallographic descriptions, validation and generation of the preferred IUPAC names and derivation of covalent radii tables – have the potential to improve the quality of data in the COD. What is more, the described workflows and tools are not tied to the COD and can be applied to any other body of chemical and crystallographic data.

[1] Gražulis, S., Daškevič, A., Merkys, A., Chateigner, D., Lutterotti, L., Quirós, M., Serebryanaya, N. R., Moeck, P., Downs, R. T. & Le Bail, A. (2012). *Nucleic Acids Res.* **40**, D420-D427.

[2] Nature (2023). *Nat.* **617**, 438.

[3] Merkys, A., Vaitkus, A., Grybauskas, A., Konovalovas, A., Quirós & M., Gražulis, S. (2023). *J. Cheminf.* **15**, 1.

[4] Meng, E. C. & Lewis, R. A. (1991). *J. Comp. Chem*. **12**, 891-898.

[5] Cordero, B., Gómez, V., Platero-Prats, A. E., Revés, M., Echeverría, J., Cremades, E., Barragán, F. & Alvarez, S. (2008). *Dalton Trans.* 2832-2838.

[6] Pyykkö, P. & Atsumi, M. (2009). *Chem. – Eur. J*. **15**, 186-197.

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