Towards a double verified (experimental + DFT) reference structure database

M. Hušák, F. Fňukal

Department of Solid State Chemistry, University of Chemistry and Technology, Prague, Technická 5, Praha 6, 166 28, Czech Republic

husakm@vscht.cz

Any database used for structure verification or for machine learning (ML) force field development must be at first 100% reliable itself. In our work we focused on structure verification of pharmaceutical molecular crystals, so our main focus is the use of Cambridge Structure Database (CSD). The existing automatic validation tools like the checkCIF and PLATON software can verify the consistency of experimental data, but they cannot fully check the "chemical sense" of the structures [1]. From this point of view the data in the CSD cannot be used for data verification or for ML purposes directly without additional filtering and verification.

We suggest creating a double-verified CSD sub-database containing, in addition to experimental data, structures verified by comparison with their DFT minimized structures as already suggested in [2]. The advances in computer technology and functional development (e.g. meta-GGA r2SCAN) make it possible to do DFT calculations in acceptable time even in-house for new structures. Re-interpretation of experimental data by Hirshfeld atom refinement method (HAR) to get agreement between X-ray, neutron and DFT atomic positions will be required as well. The original method [2] based on RMSCD evaluation will need to be extended with more descriptors for experimental/DFT differences detection. Our algorithms are implemented in the form of the new software checkCIF-DFT. This tool utilizes existing DFT engines (CASTEP, Quantum Espresso) for calculations.

We already preliminarily tested the methodology on a semi-random sample of 100 structures from the CSD filtered by "published after 2013, non-disordered, pure organic structures, no errors detected by CSD, solved from single crystal" filter. From the 100 structures 30 ones were rejected by pre filtering (missing atoms, incorrect space group and non-interpreted voids). The most significant descriptors (maximal bond and maximal angle difference) are visualized in Fig. 1. Based on comparison with a set of data from neutron and other double verified structures the image indicates multiple structures from the set are for sure problematic.

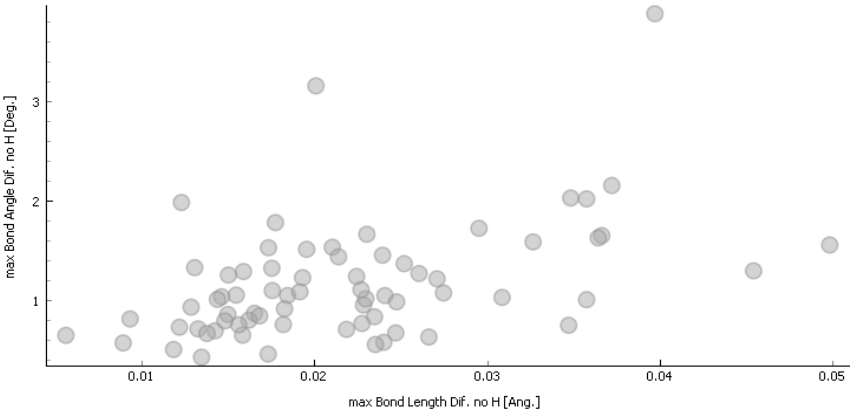

Figure 1. Maximal bond length difference versus maximal bond angle difference (hydrogen atoms excluded) for the 70 fully processed structures.

In this way, we plan to screen at least a few thousand more structures from the CSD. The main limit is the allocated supercomputer time. Our preliminary test shows that nowadays, the DFT method cannot be replaced by other less computationally demanding methods (DFTB, molecular mechanics) due to their low accuracy, so screening the whole CSD is impossible. The result of this work should be a reliable reference database of original structures, HAR re-interpreted data and DFT geometry-optimized structures suitable for both structure verification and ML force filed development purpose.

[1] Raymond, K. N., & Girolami, G. S. (2023). *Acta Cryst.* **C79**, 445.

[2] Streek, J., Neumann, M. A. (2010). *Acta Cryst.*, **B66**, 544.

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254). The authors gratefully acknowledge the Czech Science Foundation for financial support of the project no. 21-05926X.