Automating the derivation of chemical information from crystallographic data

Antanas Vaitkus[a], Andrius Merkys[a,b] and Saulius Gražulis[a,b]

[a]Vilnius University Institute of Biotechnology, Saulėtekio av. 7, LT-10257 Vilnius, Lithuania; [b]Vilnius University Faculty of Mathematics and Informatics, Naugarduko 24, LT-03225 Vilnius, Lithuania

Abstract

The emergence of new interdisciplinary fields stipulates the ever growing need of greater connectivity between scientific data from a diverse range of research areas. The process of establishing such relationships differs from field to field with the need of creating an identifier common for both fields remaining a constant. In the case of crystallography, generating a chemical descriptor of a molecule from its crystallographic structure opens up the possibility of identifying chemical compounds and relating the crystal structure to the properties of the compounds it encompasses. Carrying out this task manually for larger datasets might have been viable a century ago, but with the arrival of big data it is no surprise that there have been multiple attempts at automating the process. However, none of the approaches were sufficient enough for the processing of the open-access Crystallography Open Database (COD) [1] due to incompatible licenses or the lack of functionality. As a result, we have developed an automated approach of extracting the chemical data such as atom connectivity, bond orders and atom charges from the crystallographic atom coordinates and thus enabling the generation of chemical descriptors and the cross-linking of the COD with other open resources. Our approach strictly adheres to the principles of open science by making all of the data open-access and all of the developed programs [2, 3] open-source.

Data Source

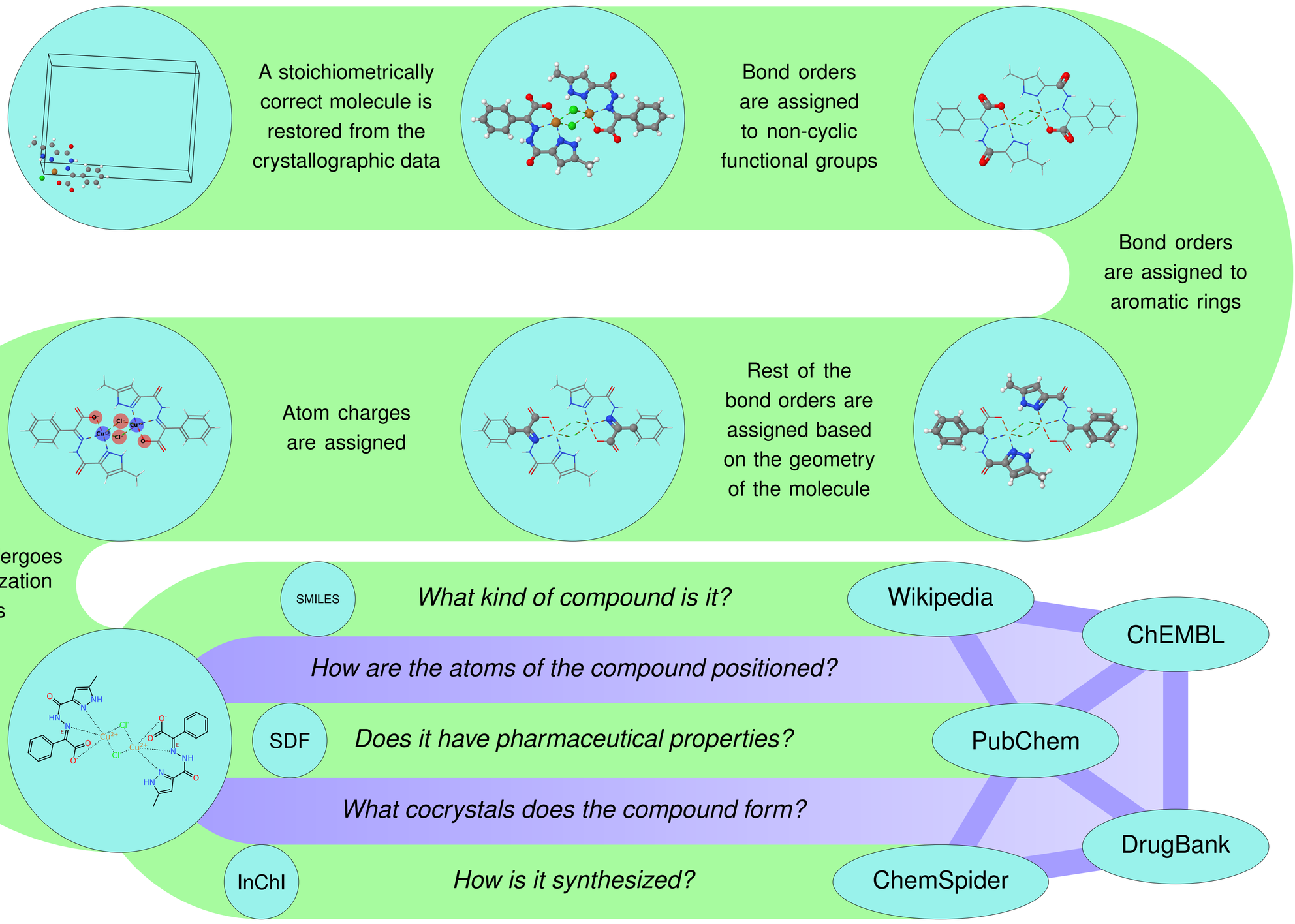
- Open-access;
- Contains small-molecule organic, inorganic, and metal-organic crystal structures;
- Over 375 000 entries;
- Uses CIF as the carrier format.

www.crystallography.net

The Perks of Linked Open Data

- Increased data applicability;
- Increased user base;
- Federated search.

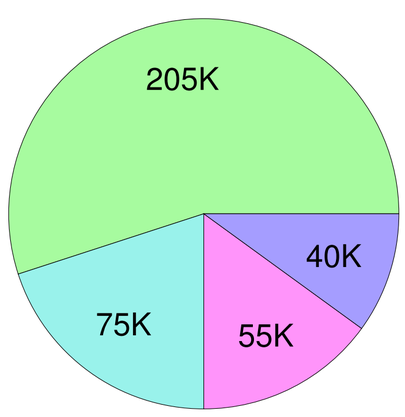
Workflow



Challenges and Results

- Lack of an unambiguous way to divide polymeric molecules into monomers;
- The core CIF dictionary does not provide sufficient means of detailing chemical properties;
- Certain discrepancies in the input crystallographic data require manual curation (e. g. unmarked disorder sites).

Overview of the COD processing



from a total of 375 000 structures.

Conclusions

- Chemically sound molecules can be automatically generated from crystallographic data;
- The COD can be linked to other open data resources;
- Linked open data benefits researchers from all fields of science.

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(D1):D420–D427, Jan 2012.

[2] Gražulis et al. Computing stoichiometric molecular composition from crystal structures. *Journal of Applied Crystallography*, 48:85–91, 2015.

[3] Merkys et al. COD::CIF::Parser: an error-correcting cif parser for the perl language. *Journal of Applied Crystallography*, 49:293–301, 2016.

Antanas Vaitkus has no conflict of interest.
Andrius Merkys has no conflict of interest.
Saulius Gražulis is a member of the COD Advisory Board.

On-line version of the poster:
http://j.mp/2mazMuI

antanas.vaitkus90@gmail.com