chem-bla-ics

# Automatic Classification of thousands of Crystal Structures

**Egon Willighagen**

Published August 24, 2007

**Citation**
Willighagen, E. (n.d.). In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/4k6ht-k2z12

**Keywords**
Crystal

**Copyright**
Copyright © Egon Willighagen 2007. Distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

## chem-bla-ics

Clustering and classification of crystal structures is hot. Parkin hit the front cover of CrystEngComm with a story on *Comparing entire crystal structures: structural genetic fingerprinting* (DOI:10.1039/b704177b). Now, the story itself, while rather interesting and well written, has three major flaws:

1. the data set it way too small
2. the proposed proof-of-concept is not novel at all
3. they do not cite me

Well, the latter sounds a bit boohoo, and it is :) (BTW, I do like this paper.)

They propose the work as proof-of-concept, but use a very artificial data set of only 12 crystal structures (benzene and eleven polycyclic aromatic hydrocarbons, like naphtalene, anthracene, phenanthrene, triphenylene, pyrene, perylene, and coronene). While such a small set does make a nice example where you can still list all similarities (`0.5*N*(N-1)`), it is really too artificial.

Now, you may wonder if I am in the position to criticize this shortcoming, but I think I am. As part of my PhD work, I analyzed this problem myself, and published two years ago the paper *Method for the computational comparison of crystal structures* (DOI:10.1107/S0108768104028344). Apparently, Parkin was not aware of this publication and did not cite it. I should have went to a crystallography conference with a poster, and advertise my work more. In this paper, I analyzed a data set with 48 crystal structures, manually validated by visual inspection, resulting in having to compare 1128! crystal structure pairs. Took me two full weeks behind a Silicon Graphics. Yes, I really understand why they took only 12 structures :)

However, there is more prior art. While my approach was based on a new radial distibution function-based whole crystal structure descriptor, my supervisor (Ron) used the more common powder diffraction pattern and showed in *Representing Structural Databases in a Self-Organising Map* (DOI:10.1107/S0108768105020331) it to be a good enough descriptor for clustering of thousands of crystal structures using a self-organizing map (SOM).

Last week, my second paper in crystallography appeared: *Supervised Self-Organizing Maps in Crystal Property and Structure Prediction* (DOI:10.1021/cg060872y). In this paper, we show how supervised SOMs (see DOI:10.1016/j.chemolab.2006.02.003) can be used for supervised classification and even for property prediction. Note that these supervised SOMs are *truly* supervised, unlike many earlier modifications of the unsupervised SOMs: the training is supervised.

Finally, another advantage of this last work: the code is open source. The code for the unsupervised SOMs is available as R package: kohonen; and for powder diffraction patterns: wccsom. Details can be found in this R News issue. The first package is not actually limited to crystal structures, and can be used for any clustering problem. However, the articles mentioned here make use of simulated diffraction patters, and I am not sure there are open source tools to generate those.

BTW, I would still be interested in teaming up with CrystalEye in one way or another, and couple these data analysis methods to live streams of new crystal structures. Nick, let me know if you are interesting in idea exchange.

Getting back to Parkin's paper, I do like the work. Hirshfield surfaces are an interesting tool to visualize packing characteristics, and using them to describe a crystal structure sounds like an interesting idea indeed. I just hope that the method properly scales.