Workflows for High-throughput Computational Crystallography

Volker Thome[1], Peter Murray-Rust[1], Nick Day[1], Mary Heppenstall-Butler[2]

[1] *Unilever Centre for Molecular Science Informatics, Chemistry Department, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, UK.*
[2] *Unilever Research Centre, Colworth, UK*

Unilever Cambridge
Centre for Molecular Informatics

The computation of crystal geometries, energies and physical properties usually requires several programs and systematic variation of parameters. To automate this we have created workflows with the OpenSource Taverna and Condor/DAGMan architectures. The components are engineered to accept and emit XML, specifically CML (Chemical Markup Language), and include legacy conversion (e.g. to and from CIF). We can support codes such as GULP, CASTEP, SIESTA, DL_POLY, MOPAC and GAMESS-US and display the results in SVG or the Jmol visualiser. We give two examples of current use.

Automatic download of CIFs from *Acta Cryst. E* and computation of molecular properties by high-level QM

(a) The Taverna architecture was configured to locate and download CIFs from the *Acta Cryst. E* site and convert them to CML. 25,000 CIFs can be downloaded [1] and converted in a few hours, including extraction of chemistry from atomic coordinates.

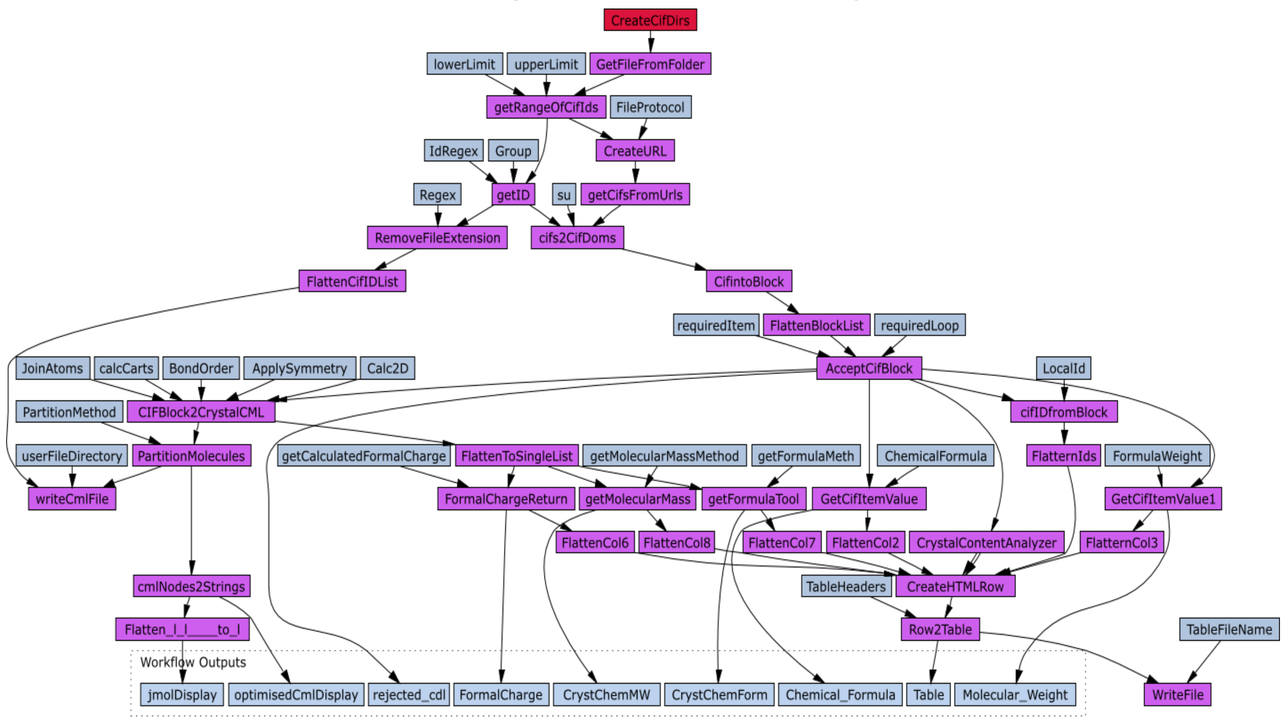


Fig. 1: Taverna Workflow for reading and processing CIFs to CML

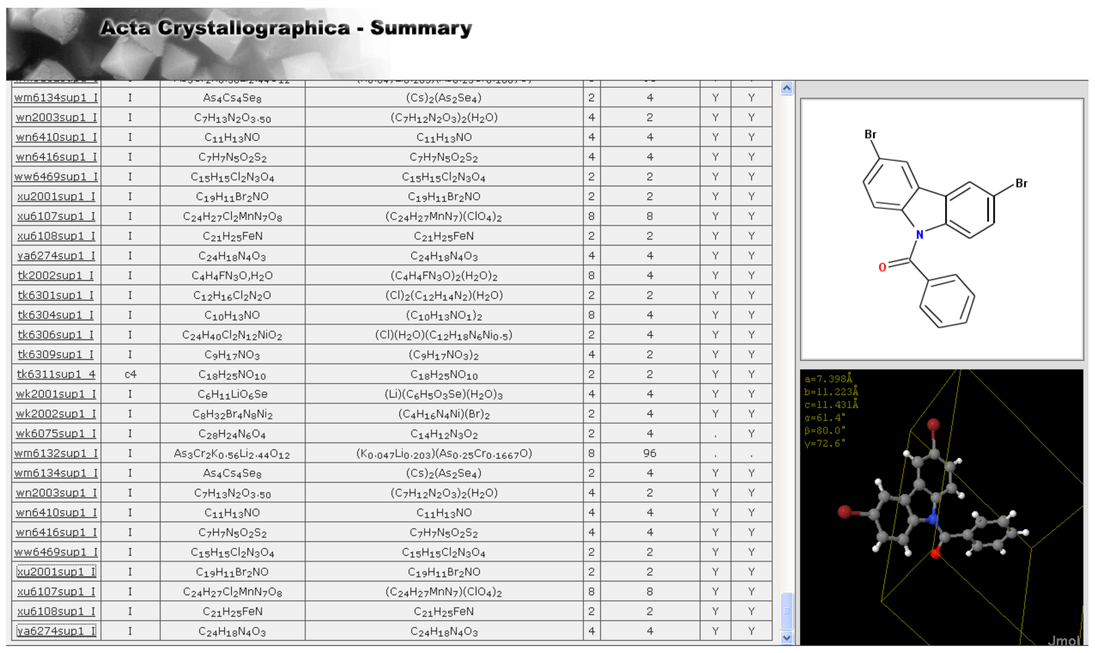


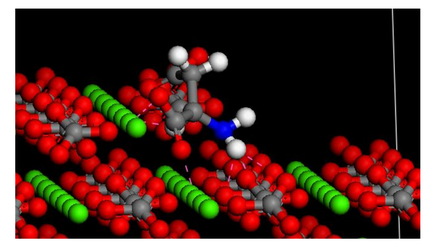
Fig. 2: Chemical enhancement (Jumbo, CDK, JChempaint, Jmol) of *Acta Cryst. E* For each structure the chemical and crystallographic contents are compared.

(b) The semi-empirical program MOPAC running in a CONDOR architecture can process ca. 1 million jobs [2]. We are now investigating the agreement in geometry between molecules in crystals and the results of high-level QM gas-phase calculations (GAMESS-US at B3LYP/6-31G*). From the *Acta Cryst. E* structures those with disorder, inconsistent formulae, heavy atoms (Z > 18) and > 15 non-H atoms were rejected. So far about 1000 discrete organic moieties have been automatically computed, using over 1000 days CPU time.

Automatic computation of bulk and surface properties of crystals

We are currently using atomistic simulations to compute the bulk and surface properties of minerals, especially calcium carbonate polymorphs, and their interaction with small organic molecules. This requires systematic computational sweeps over the following variables:

- Polymorph
- Surface and area of interest
- Small molecule identity and conformation
- Level of theory and parameterisation
- (e.g. Force Field, basis set)



A simple example involves the calculation of surface properties on a rectangular grid, with one job per point. In the example above an organic anion is systematically moved over a calcite surface. At each point the z-coordinate is optimised and properties calculated (e.g. with GULP). When all jobs have finished, the results are fed to a second program to analyse and display the properties.

The workflow strategy for this has been developed in the eMinerals project using the CONDOR system to take advantage of unused cycles on teaching and other machines [3]. DAGMan ensures that jobs are run in the correct order and can detect failed jobs and resubmit them. Only when all jobs are completed does the plot routine collect and analyse the complete area.

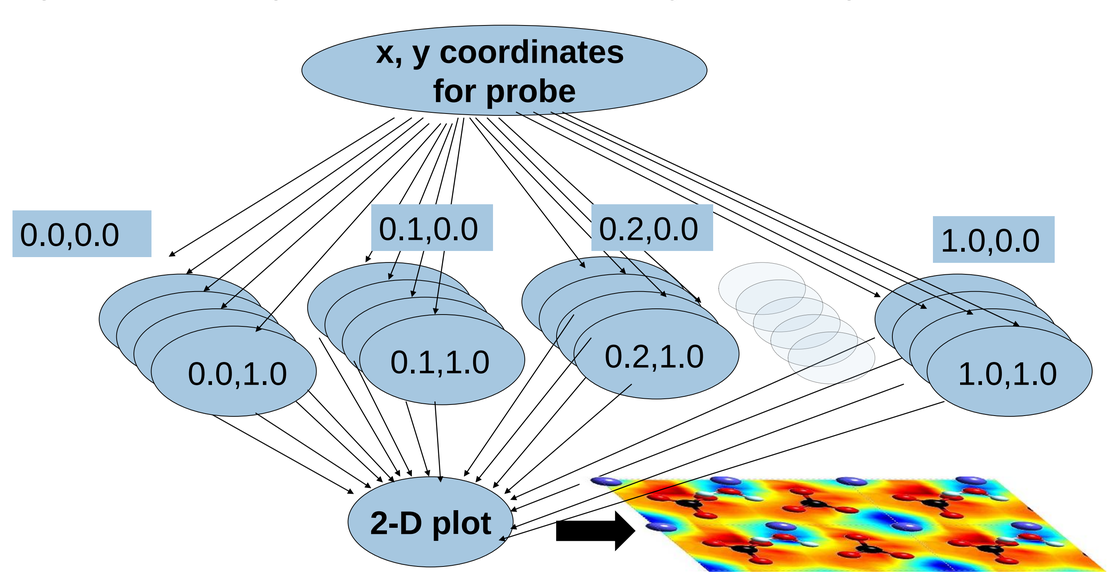


Fig. 3: Typical DAGMan-CONDOR architecture for systematic computation of properties over a surface

Conclusion

We can provide an infrastructure of distributable components where robots can:

- read journals, extract molecules and compute their properties
- publish them to newsfeeds and Open repositories
- perform multi-step simulation processes, e. g. for time consuming QM calculations

Services and tools

All tools are open and downloadable
Home page: http://wwmm.ch.cam.ac.uk
Email: vt228@cam.ac.uk, ned24@cam.ac.uk, pm286@cam.ac.uk

References

(1) Peter Murray-Rust, Robert C. Glen, Henry S. Rzepa, James J. P. Stewart, Joe A Townsend, Egon L. Willighagen, Yong Zhang *"A semantic GRID for molecular science",* Proceedings of the 2003 UK e-Science All Hands Meeting

2) Peter Murray-Rust, Henry S. Rzepa, James J. P. Stewart and Yong Zhang *"A global resource for computational chemistry"* Journal of Molecular Modeling, Publisher: Springer Berlin / Heidelberg, Vol. No. 6 (2005) p: 532 - 541

(3) J. Wakelin, P. Murray-Rust, S. Tyrrell, Y. Zhang, H. S. Rzepa, A. García *"CML tools and information flow in atomic scale simulations",* Molecular Simulation Vol. 31, No. 5 (2005) p: 315 - 322

Acknowledgements

•The International Union of Crystallography for student support
•Toby White, Emilio Artacho, Tom Archer (Dep. of Earth Sciences, Cambridge) for advice