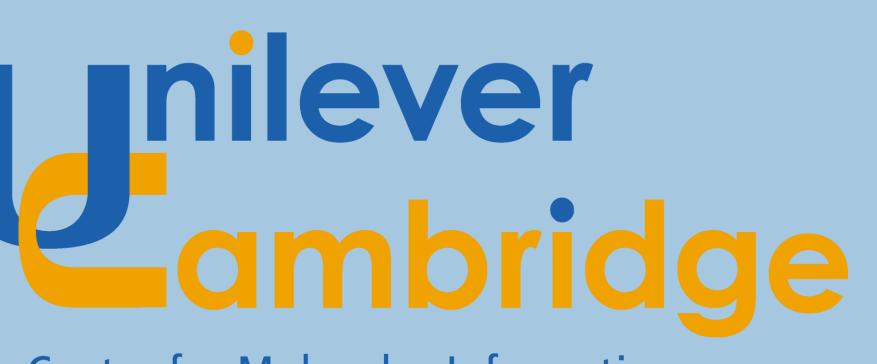


Workflows for High-throughput Computational Crystallography

Volker Thome¹, Peter Murray-Rust¹, Nick Day¹, Mary Heppenstall-Butler²

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



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 Automatic computation of bulk and surface properties of crystals

(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.

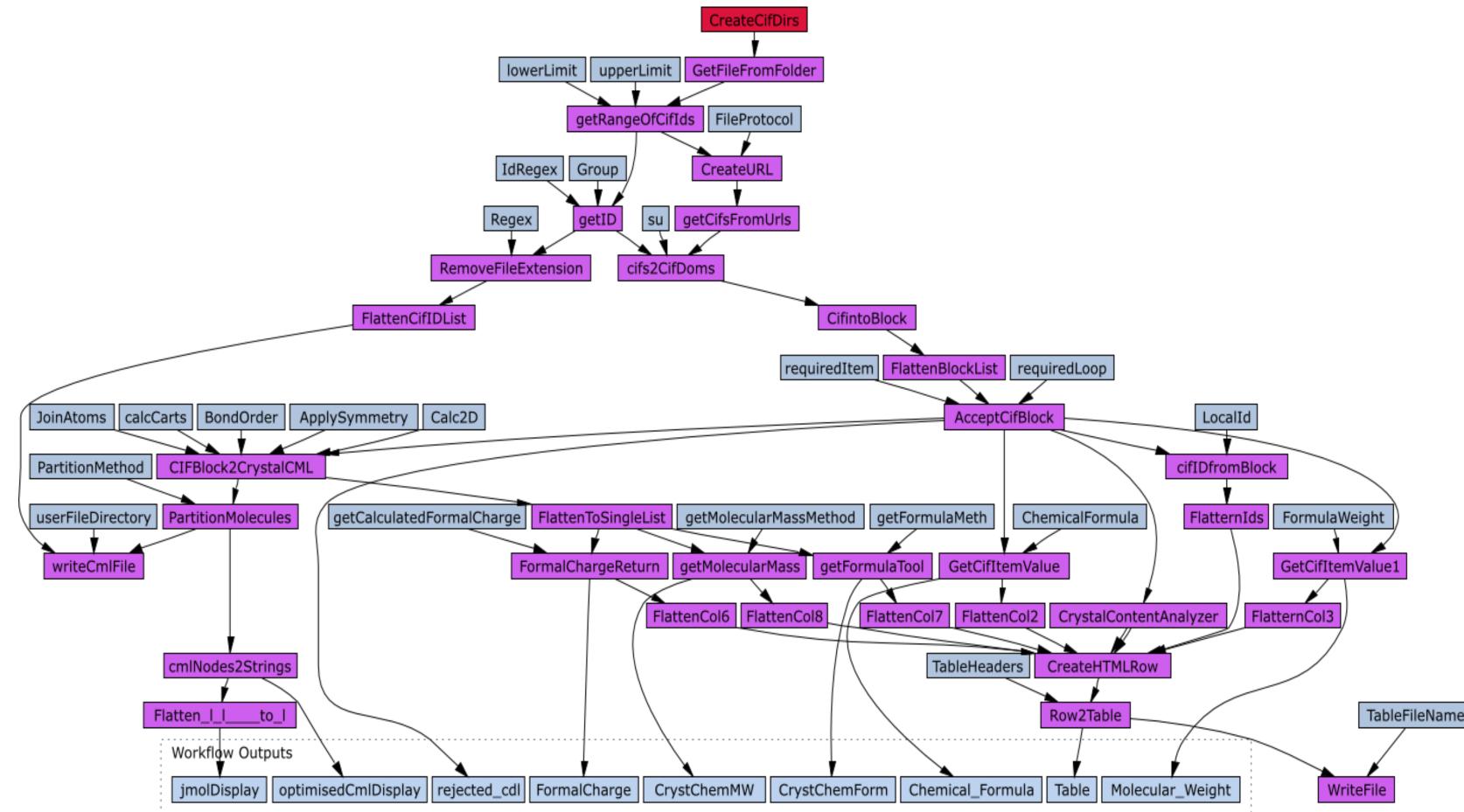
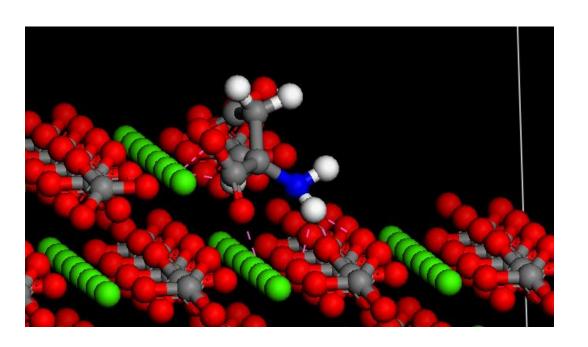


Fig. 1: Taverna Workflow for reading and processing CIFs to CML

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.

| | Acta | Crystallograp | hica - Summary | | | | | |
|---------------------|---------|---|---|---|---|---|---|------------|
| | and the | | (| | | | | |
| <u>wm6134sup1 I</u> | I | As4Cs4Se8 | (Cs) ₂ (As ₂ Se ₄) | 2 | 4 | Y | Y | |
| <u>wn2003sup1 I</u> | I | C7H13N2O3.50 | (C7H12N2O3)2(H2O) | 4 | 2 | Y | Y | |
| <u>wn6410sup1 I</u> | I | C ₁₁ H ₁₃ NO | C ₁₁ H ₁₃ NO | 4 | 4 | Y | Y | Br |
| <u>wn6416sup1 I</u> | I | C7H7N5O2S2 | C7H7N5O2S2 | 4 | 4 | Y | Y | |
| <u>ww6469sup1 I</u> | I | C ₁₅ H ₁₅ Cl ₂ N ₃ O ₄ | C ₁₅ H ₁₅ Cl ₂ N ₃ O ₄ | 2 | 2 | Y | Y | |
| <u>xu2001sup1 I</u> | I | C ₁₉ H ₁₁ Br ₂ NO | C ₁₉ H ₁₁ Br ₂ NO | 2 | 2 | Y | Y | |
| <u>xu6107sup1 I</u> | I | C ₂₄ H ₂₇ Cl ₂ MnN ₇ O ₈ | (C ₂₄ H ₂₇ MnN ₇)(ClO ₄) ₂ | 8 | 8 | Y | Y | |
| <u>xu6108sup1 I</u> | I | C ₂₁ H ₂₅ FeN | C ₂₁ H ₂₅ FeN | 2 | 2 | Y | Y |) <u> </u> |
| <u>ya6274sup1 I</u> | I | C ₂₄ H ₁₈ N ₄ O ₃ | C ₂₄ H ₁₈ N ₄ O ₃ | 4 | 4 | Y | Y | |
| <u>tk2002sup1 I</u> | I | C4H4FN3O,H2O | (C4H4FN3O)2(H2O)2 | 8 | 4 | Y | Y | |
| <u>tk6301sup1 I</u> | I | C ₁₂ H ₁₆ Cl ₂ N ₂ O | (Cl) ₂ (C ₁₂ H ₁₄ N ₂)(H ₂ O) | 2 | 2 | Y | Y | |

 $(C_{10}H_{13}NO_1)_2$

(Cl)(H₂O)(C₁₂H₁₈N₆Ni_{0.5})

(C9H17NO3)2

C₁₈H₂₅NO₁₀

(Li)(C₆H₅O₃Se)(H₂O)₃

(C₄H₁₆N₄Ni)(Br)₂

C14H12N3O2

(K0.047Li0.203)(As0.25Cr0.1667O)

(Cs)₂(As₂Se₄)

(C7H12N2O3)2(H2O)

C₁₁H₁₃NO

C7H7N5O2S2

C₁₅H₁₅Cl₂N₃O₄

C₁₉H₁₁Br₂NO

(C24H27MnN7)(ClO4)2

C₂₁H₂₅FeN

C24H18N4O3

8

2

2

| 2 |

2

4

4

2

2

4

4

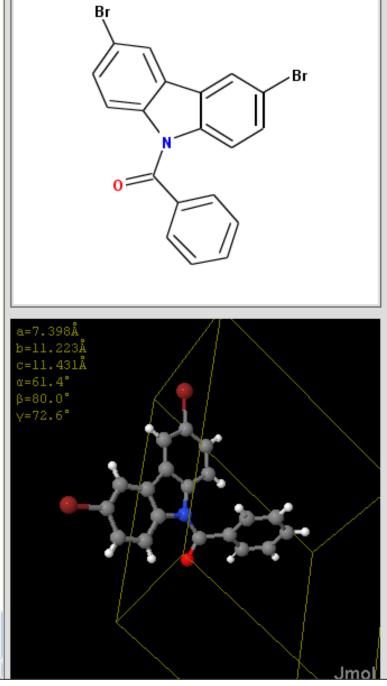
96

2

4

4

2



Y Y

Y Y

Y Y

Y Y

Y Y

Y Y

Y Y

Y Y

Y Y

Y Y

Y Y

Y Y

Y Y

Y Y

YL

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.

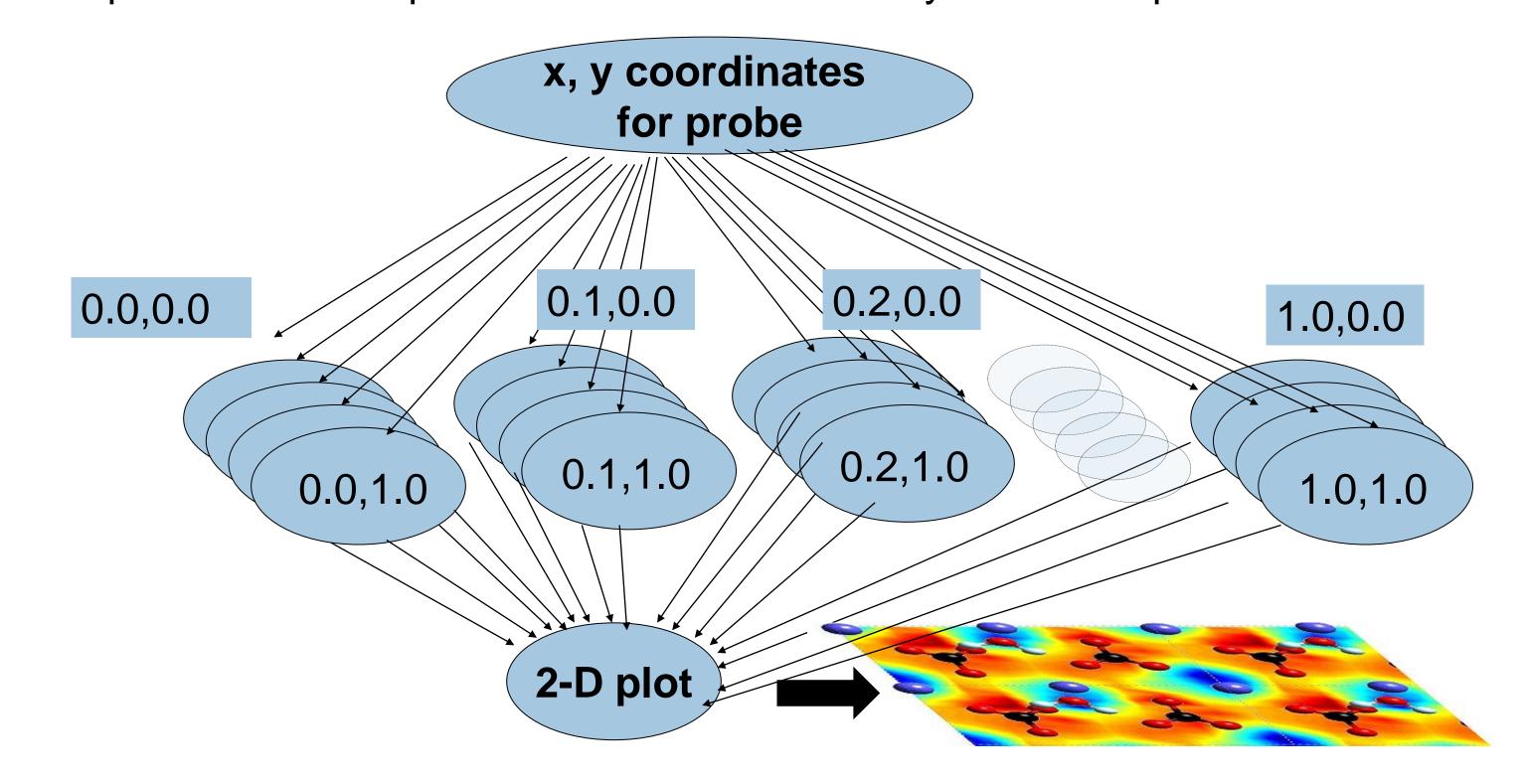


Fig. 3: Typical DAGMan-CONDOR architecture for systematic computation of properties over a surface

Conclusion

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

References

<u>tk6304sup1</u>

<u>tk6306sup1</u>

<u>tk6309sup1 (</u>

<u>tk6311sup1 4</u>

<u>wk2001sup1</u>

<u>wk2002sup1</u>

<u>wk6075sup1</u>

<u>vm6132sup1</u>

<u>/m6134sup1</u>

wn2003sup1

<u>wn6410sup1</u>

<u>wn6416sup1</u>

<u>vw6469sup1</u>

<u>ku2001sup1</u>

<u>xu6107sup1</u>

<u>4u6108sup1 </u>

<u>ya6274sup1</u>]

c4

C10H13NO

C24H40Cl2N12NiO2

C9H17NO3

C₁₈H₂₅NO₁₀

C₆H₁₁LiO₆Se

C₈H₃₂Br₄N₈Ni₂

C₂₈H₂₄N₆O₄

As3Cr2K0.56Li2.44O12

As4Cs4Se8

C7H13N2O3.50

C₁₁H₁₃NO

C7H7N5O2S2

 $C_{15}H_{15}Cl_2N_3O_4$

C₁₉H₁₁Br₂NO

C24H27Cl2MnN7O8

C₂₁H₂₅FeN

C24H18N4O3

- (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

- 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: <u>http://wwmm.ch.cam.ac.uk</u> Email: vt228@cam.ac.uk, ned24@cam.ac.uk, pm286@cam.ac.uk

Acknowledgements

•The International Union of Crystallography for student support

•Toby White, Emilio Artacho, Tom Archer (Dep. of Earth Sciences, Cambridge) for advice