Environment from the molecular level: an escience testbed project

Martin T Dove[1], Mark Calleja[1], Jon Wakelin[1], Kostya Trachenko[1], Guillaume Ferlat[1], Peter Murray-Rust[2], Nora H de Leeuw[3], Zhimei Du[3], G David Price[4], Paul B Wilson[4], John P Brodholt[4], Maria Alfredsson[4], Arnaud Marmier[5], Richard P Tyer[6], Lisa J Blanshard[6], Robert J Allan[6], Kerstin Kleese van Dam[6], Ilian T Todorov[6], William Smith[6], Vassil N Alexandrov[7], Gareth J Lewis[7], Ashish Thandavan[7], S Mehmood Hasan[7]
www.eminerals.org

1. *Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EQ*
2. *Unilever Cambridge Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW*
3. *School of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX*
4. *Department of Earth Sciences, University College London, Gower Street, London WC1E 6BT*
5. *Department of Chemistry, University of Bath, Bath BA2 7AY*
6. *CCLRC eScience Centre, Daresbury Laboratory, Warrington WA4 4AD*
7. *Department of Computer Science, University of Reading, Reading RG6 6AY*

The testbed project has the ambition to push the practical possibilities of atomistic simulations forward to the point where we can perform realistic calculations on important environmental processes. The project has three components: the science driving the project, the development of the simulation codes, and setting up a grid infrastructure for this work. This paper describes these areas of work and gives a status report on each.

1. Introduction

The UK has a traditional strength in the area of simulations of matter at a molecular level. The types of simulations include those that use empirical functions to describe the interactions between atoms, and those that use quantum mechanics to describe the electronic structure. Both types of simulation have an important role to play across all the physical and biological sciences. The UK environmental and earth sciences communities have a strong background in the use of molecular simulation methods for a wide diversity of applications, ranging from the properties of natural materials in the inner Earth and at the Earth's surface, mineral–fluid interactions, crystal growth, adsorption of pollutants on mineral surfaces, waste storage etc.

The objective of our testbed project, funded by NERC, is to work towards an implementation of the vision of being able to run molecular simulations that are able to capture as much of the environmental situation as possible. This typically means being able to use realistically-large simulations, and capturing some of the details of realistic environmental fluids. Whatever technique is appropriate, there is a need for the tools to be optimised for large systems.

The testbed project team includes scientists, application developers, and grid experts (Fig 1). Because we will be performing simulations with greatly enhanced computational challenges, and will be generating data files that are, at least for this area of science, of unprecedented size and complexity, we recognised from the outset that we will be defining a completely new way of working. This will involve much greater inter-institute collaboration, with partners operating within the structure of a virtual organisation. Amongst the factors involved in setting up our infrastructure is to include a collaborative framework, a minigrid and unified portal for computing and data sharing, and the mechanisms for sharing data between simulation, analysis and visualisation tools.

2. Science drivers

2.1 Adsorption of pollutant molecules on mineral surfaces

In earlier less-careful times, a number of families of organic molecules were released into the environment. These were either by-products of industrial processes or pesticides. It is now known that they are persistent in the environment, and are now present in the human food chain. Examples are polychloro-biphenyl (PCB) and dioxine molecules.

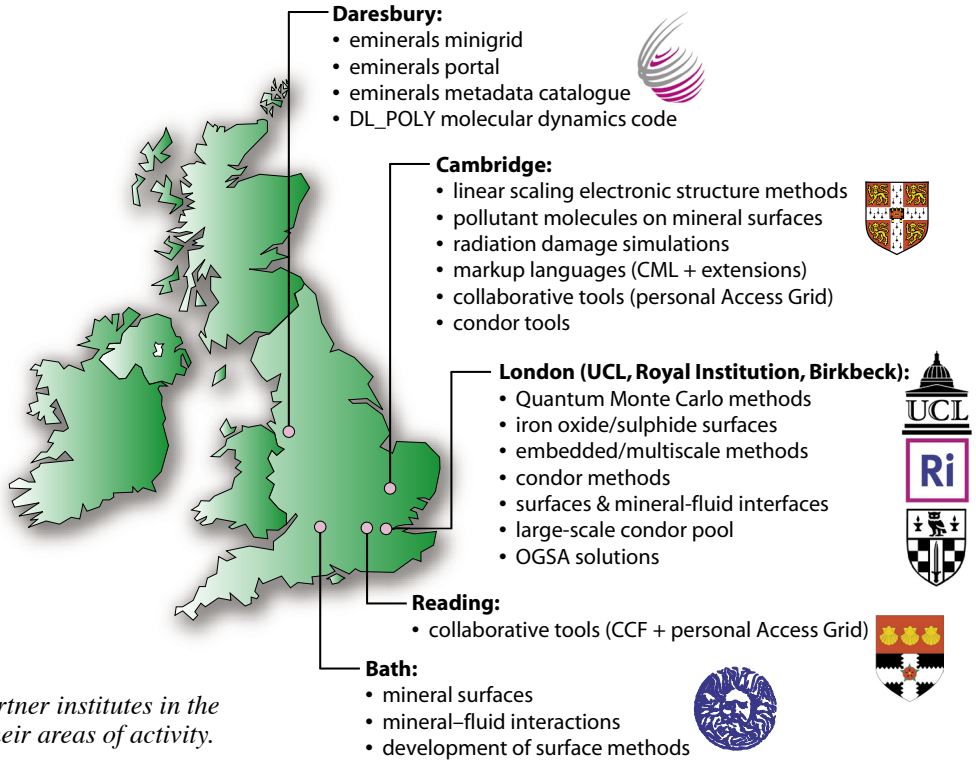


Fig 1. The partner institutes in the project and their areas of activity.

We are using quantum mechanics calculations (§3.3) to investigate the binding of these molecules to mineral surfaces. Initially we are working with an artificial vacuum above the mineral surface, but will extend this to include an appropriate fluid environment, and other organic molecules to reflect natural organic matter.

At the heart of this problem is a serious combinatorial problem, namely that there are many members of each family of molecules depending on the number of attached atoms and their position in the molecule. For example, there are 209 members of the PCB family. The problem of running similar calculations for all members of a family of molecules lends itself to computing within a grid infrastructure.

2.2 *Materials for encapsulation of radioactive waste*

The question of how to store high-level radioactive waste, particularly spent nuclear fuel, is one of the most pressing issues facing industrial societies. Among the materials being considered as storage media are silicate glasses and cements. Another option is to use crystalline ceramics. This approach is an example of "learning from nature", since it is known that some minerals, such as zircon, have contained radioactive atoms for geological timescales.

The challenge is to simulate the response of any proposed ceramic when one of the radioactive ions undergoes a decay process. In a high-energy alpha-decay process, the damage to the structure is caused by the recoiling atom. We are performing simulations of the damage caused by recoil atoms on a range of potential ceramic matrix materials (Fig 2). These are giving insights into the origin of volume swelling of damaged materials, and the resistance of materials to amorphisation[1].

2.3 *Mineral–fluid interactions*

Our interest in mineral–fluid interactions

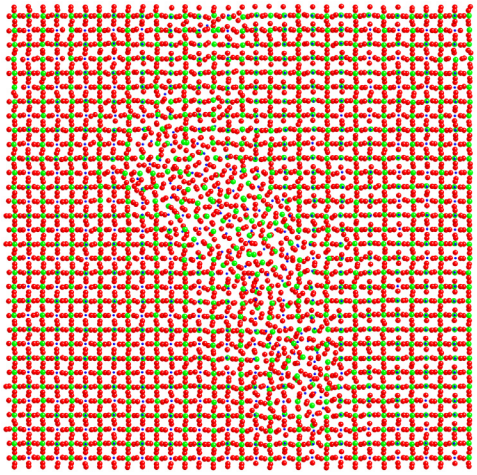


Fig 2. Simulation of radiation damage in the mineral zircon.

is two-fold: to consider processes that lead to dissolution of minerals in realistic environments (weathering), and the processes that cause precipitation of minerals from solutions. The latter is a particular issue in the oil industry, where inhibitors have to be introduced to prevent scaling of pipes.

We are carrying out a number of studies of mineral surfaces with hydration layers. Examples include alumina[2] and quartz[3]. We use both quantum mechanical and empirical models; the former are essential to describe dissociation of water molecules. We are extending this work to study a wide range of oxide minerals in contact with water.

3. Code developments

3.1 Molecular dynamics

The key UK code for molecular dynamics simulations is DL_POLY, developed by one of the project partners. The current release version is not optimised for large systems, although it can handle up to 300,000 atoms. For the work on radiation damage (§2.2) we will eventually need to be able to work with several millions of atoms.

As part of our project we are developing a new version of DL_POLY. This involves several major changes in the basic algorithms, including how to handle the long-range Coulomb interactions, and how to parallelise the simulation to make use of high-performance computers such as the UK HPCx facility.

3.2 Surface codes

Surface energies can be calculated using the METADISE code. This is being extended to use Monte Carlo techniques to scan a wide range of possible surface states, which will include many states that are present experimentally but which are often not taken into account in theoretical studies. This is particularly appropriate for computing in a grid environment.

3.3 Linear-scaling electronic structure calculations

Electronic structure calculations for periodic systems have now reached the point where quite complicated studies are routine. Most calculations now use density functional theory (DFT) as the means of describing the electronic Hamiltonian, and the computational problem is made less severe by treating the inner electrons of the atoms by an effective potential energy function, called the 'pseudopotential'. The challenge that is now being faced is how to develop methods to allow the size of a calculation to scale linearly with the size of the system, the "linear scaling" problem.

We are helping to develop a linear scaling DFT code called SIESTA. The approach to implementing a linear scaling algorithm is to describe the electron density in terms of localised atomic orbitals (many of the current leading codes use a superposition of waves, representing the Fourier components of the electron wave functions, but computations using this approach typical scale as the cube power of the system size). Current work is being carried out for calculations on water and in the study of organic molecules on mineral surfaces (§2.1).

3.4 Quantum Monte Carlo methods

It is known that DFT has a number of limitations, including problems handling elements such as iron. These arise from some of the key approximations in the method, particularly how electron exchange and correlation energies are handled. One alternative is the Quantum Monte Carlo method. This is much more challenging from a computational perspective. However, the method involves integrating over many configurations, which can be set up in parallel. Thus the problem is ideally suited for computing in a distributed grid environment.

4. Grid areas and the eminerals virtual organisation

4.1 The eminerals minigrid

Several partners of the project are bringing compute and data resources to the project (including the CONDOR pool described in §4.3). These are being integrated to form a minigrid using the Globus Toolkit 2 in analogy to the UK *e*Science Grid. This is described in more detail elsewhere in these proceedings[4].

4.2 The eminerals portal

The *e*minerals minigrid is accessible from the unified portal that has been developed at the Daresbury Laboratory, bringing together the components of the HPC and data portals. This will provide an interface to the compute resources on the minigrid for running simulations, and will also interface to a database server which holds the project's metadata catalogue describing archived results from previous simulation runs.

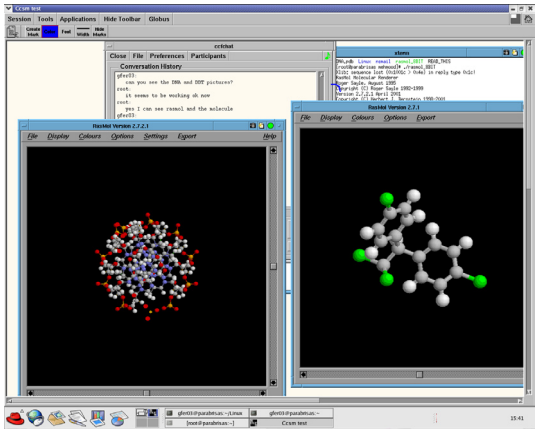


Fig 3. Screen shot of the CCF tool running a shared molecular visualisation application.

4.3 The UCL/eminerals CONDOR pools

The department of Information Services at UCL manages around 800 PCs, which are primarily used for teaching. These are set up to access services on a central server, and this means that, in effect, each PC typically uses only 5% of its individual processing power. We have linked these machines into one large CONDOR pool, which is now being used by members of the project team for production runs. Although the individual processors are no longer top range (1.4 GHz P3 processors with either 256 or 512 MB RAM), we are finding that the throughput on this system is comparable to that on national high-performance facilities for appropriate tasks.

We have set up other computing pools using CONDOR, which are providing additional computing resources. One pool has machines with much higher memory. All pools are connected to the *e*minerals minigrid.

One of the initial problems with the CONDOR pools was that users needed tools to keep track of the progress of their jobs, particularly when there are possibilities of algorithms failing to converge. We have developed a set of tools (using PERL) that are accessed through a web browser with a simple password security. These mean that users do not need to locate and then log in to the machines running specific tasks, which would clearly be breaking the spirit of the CONDOR approach to distributed computing.

4.4 Data transfer via CML

We have developed a mark-up language for use in condensed matter sciences. This is a superset of the well-established Chemical Mark-up Language (CML)[5], with essential new elements such as `latticeVector` and `particle`. A complementary FORTRAN90 library, called JUMBO90, has been developed from an earlier FORTRAN77 version. This can be incorporated in new or existing FORTRAN simulation codes, such as SIESTA, DL_POLY and METADISE (§3), allowing them to easily generate marked-up data. We have tested the idea with JUMBO77 and the CMLcore language, and now intend to use JUMBO90 in conjunction with CMLsolid, refining both the language and the parser as we go. Partly because of a lack of FORTRAN-based XML tools and the difficulty of implementing some aspects of XML (such as validation, namespace, etc) in FORTRAN, we have decided to handle program 'input' externally using XSLT and/or scripting languages such as PERL and PYTHON. A collection of XSLT stylesheets have been developed to convert marked-up chemical information into text input files for the programs we use in this project.

Our objective in using CML is to facilitate data transfer between simulation programs, particularly within the framework of the *e*minerals portal (§4.2). and into analysis/visualisation tools.

4.5 Collaborative tools

The final component of the eminerals virtual organisation is the need for collaborative tools. For this work we use a combination of tools. The personal version of the Access Grid ('Personal Interface to the Grid') is used in both its Windows NT and Linux versions to provide desktop communications. This is coupled with the Collaborative Computing Framework (CCF) tools for applications sharing and white board tools. One example is that we have linked molecular visualisation tools to CCF for shared viewing of simulation configurations (Fig 3).

Acknowledgements

The major share of funding for this project has been given by NERC. Additional support has been provided by EPSRC and CMI. High-performance computing has been carried out on the UK HPCx and Cambridge HPCF facilities.

References

1. Trachenko K, Dove MT & Salje EKH, *J Phys Cond Matt* **15**, L1, 2003
2. Marmier A & Parker SC, *Phys Rev B* (submitted)
3. de Leeuw NH, Du Z, Li J, Yip S & Zhu T. *Nanoletters* (in press)
4. Tyer RP, Blanshard LJ, Allan RJ, Kleese-van Dam K & Dove MT, *Proceedings of All Hands 2003*
5. Murray-Rust P, Glen RC, Rzepa HS, Townsend JA & Zhang Y, *Proceedings of All Hands 2003*