Using *e*Science to calibrate our tools: parameterisation of quantum mechanical calculations with grid technologies

K. F. Austen[1], T. O. H. White[1], R. P. Bruin[1], M. T. Dove[1], E. Artacho[1], R. P. Tyer[2]
[1]Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge, CB2 3EQ
[2]CCLRC Daresbury Laboratory, Daresbury, Warrington, Cheshire, WA4 4AD

Abstract

A report is presented on the use of *e*Science tools to parameterise a quantum mechanical model of an environmentally important organic molecule. *e*Science tools are shown to enable better model parameterisation by facilitating broad parameter sweeps that would otherwise, were more conventional methods used, be prohibitive in both time required to set up, submit and evaluate the calculations, and in the volume of data storage required. In this case, the broad parameter sweeps performed highlighted the existence of a computational artefact that was not expected affect this system to such an extent, and which is unlikely to have been observed had fewer data points been taken. The better parameterisation of the model leads to more accurate results and the better identification of the applicability of aspects of the model to the system, such that great confidence can be put in the results of the research, which is of environmental importance.

1. Introduction

Polychlorinated biphenyls (PCBs) have long been known to have environmental significance due to their persistence within the environment, and their high toxicity. New information on the interaction of these chemicals with common soil minerals is continually being sought to facilitate meso-scale modelling of their movement through aquifers, through assessment of the retardation effects of different minerals and their adsorption isotherms.

The structure of biphenyl is shown in Figure 1. PCBs share this structure, and each hydrogen atom on the carbon rings can be substituted for chlorine atoms, in any combination. The number of possible PCB congeners, 209 different arrangements of chlorine atoms around the biphenyl rings, poses an arduous and time-consuming problem for the typical computational scientist. Were all of these congeners to be studied by hand, the time required to generate the starting structures alone would be prohibitive, without even considering the best parameterisation of the model.

This paper reports the work that has been carried out, using *e*Science tools, to refine the input parameters for the PCB calculations. A large number of multi-dimensional parameter sweeps have been performed; and through the generation of large data sets, the importance of further parameterisation has been realised, which previously might have gone unnoticed. As a consequence, the accuracy of the results is optimised beyond that reached *via* conventional methods.

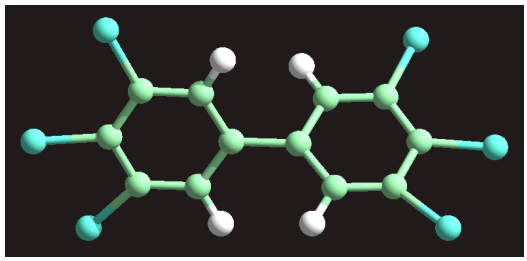

Figure 1 PCB structure (Cl=blue, C=green, H=white)

Investigations have already been made into the adsorption of the related molecules, polychlorinated dibenzodioxins (PCDDs) and polychlorinated dibenzofurans (PCDFs) [1], of which there are, respectively, 76, and 126, different congeners, onto the (001) surface of the clay mineral, pyrophyllite. The *e*Science tools, developed within the *e*Minerals project, generate each congener at various heights above the surface, and allow for relaxation of the geometries subject to certain constraints on the system [1]. The same tools will be used to study PCBs at the surface, but an additional complication occurs when investigating PCBs, as there is the possibility of rotation of the phenyl rings around the C–C bond that joins them. The ease of this rotation is expected to greatly influence the adsorption energies of these molecules onto the surface, and consequently it is extremely important that it is adequately described in the calculations. To this end,

an investigation has been made of the applicability of the current model to quantify of the energetics associated with the 360˚ rotation of the (Cl–)C–C–C–C(–Cl) torsion angle for 2,2'-dichloro biphenyl (hereafter 2-PCB). There has been work in the literature both experimentally [2] and computationally [3], which has shown that previous calculations of the groundstate equilibrium angle have not been able to reproduce the 75˚ angle for the molecule [3].

The role that *e*Science played in this work was integral to the methodology. The many tools that have been developed within the *e*Minerals project have been indispensable in enabling the individual scientist to use to quickly begin work and to perform initial, very detailed, exploration of the system, in a way that would not have otherwise been possible.

2. Methodology

2.1 Simulation Details

All the calculations reported here were carried out at the density functional theory (DFT) level using the latest version of the SIESTA code[4], which includes CML output and z-matrix input. In the first instance, the calculations have been performed using the auto-generated double zeta polarized (DZP) basis sets within the code and the PBE functional[5].

The study has taken place in two parts. Initially, the box size surrounding an hexa–chlorinated PCB molecule was converged with respect to the total energy of the system. The aim of this was to determine the minimum box size necessary to surround the molecule without any self-interaction between periodic images. This is useful for two reasons: first, a smaller box size means less computational expense; second, the box size will determine the lower-limit of the surface size necessary when the interaction of the molecules with the surface is investigated.

The first part of the study was performed with the 3,4,5,3',4',5'-hexachloro biphenyl molecule (6-PCB), for reasons explained below.

Once the minimum box size was determined, a starting structure for the 2,2'-dichloro biphenyl molecule was generated and then described using the z-matrix format within the SIESTA code [4]. The z-matrix format allows the constraint of parameters within the molecule, such as, in this case, the torsion angle between the two phenyl rings (Figure 2). The torsion angle was varied over 360˚ at 5˚ intervals, requiring 72 calculations. Two basis sets were tested; the auto-generated DZP SIESTA basis set and a user-specified basis set, the latter of which was the more computationally expensive of the two.

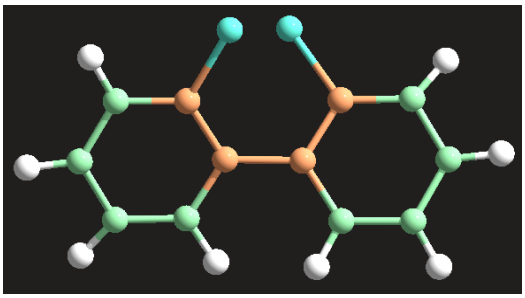

Figure 2 2,2'-dichloro biphenyl showing definition of torsion angle (red), the angle between the two planes defined by the pairs of carbons highlighted on each phenyl ring

2.2 Obtaining Starting Configurations

Starting configurations were required for both of the PCB molecules and these were obtained from experimental crystal structures stored in the Cambridge Crystallographic Database [6]. As the vacuum structure will differ from the crystal structure, further structural relaxation was required.

2.3 Calculating Box Size

Modelling of the molecule in vacuum was carried out using Periodic Boundary Conditions. The molecule is enclosed in a virtual box, the dimensions of which were varied to find the optimal size. The PCB chosen for the box size calculation was, as mentioned above, the 3,4,5,3',4',5'-hexachloro biphenyl (6-PCB) molecule. The chlorine atoms are larger, and the C–Cl bond lengths longer, than is the case for hydrogen. The fully chlorinated PCB molecule, where all 8 hydrogen atoms are replaced by chlorine atoms, posed difficulties in calculating the fixed planar geometry due to unfavourable steric repulsions between chlorine atoms in the 2 and 6 positions on opposite rings. 6-PCB was chosen, therefore, because it has the largest space-filling contributions of all the possible PCBs that can easily be calculated in a planar configuration. The molecule has the approximate dimensions of 10Å in length and 5Å across the chlorophenyl rings (Figure 1).

As such, an initial box size was taken to be 15Å x 10 Å x 10Å, and the molecule aligned so that it lay lengthwise along the long axis of the cuboid. The two shorter box–sides were kept equal to allow for free rotation around the C–C bond without interference between periodic images. The side lengths were sequentially increased up to 25 Å for the c parameter and 20 Å for a and b, at 0.5 Å intervals. This required the calculation of over 100 box sizes, each with different values of the lattice parameters. The generation and management of these calculations is detailed in Section 2.5.

The broad parameter sweeps, enabled by the use of *e*Minerals submission scripts, allowed such a

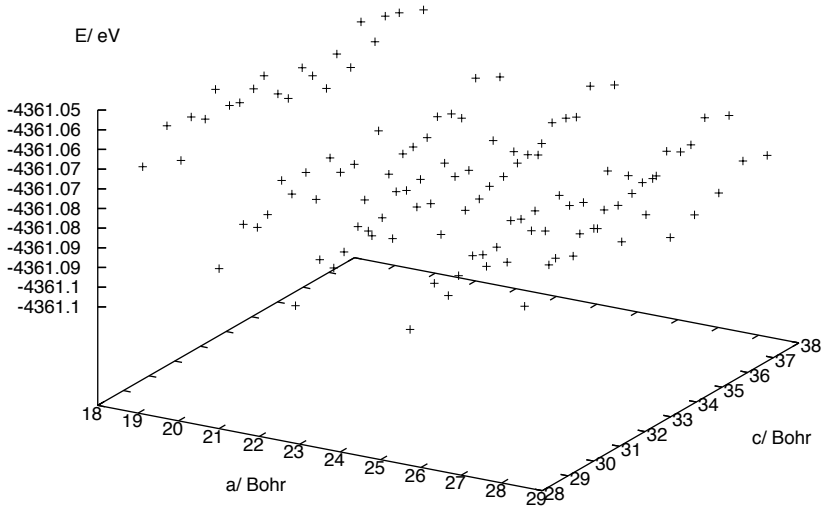

Figure 3 Energy v. box size (a=c) for calculations with a 100 Ry cutoff, showing periodic fluctuations in the energy with increase in box size, rather than the steady convergence expected

detailed sampling of the box dimensions that we were able to observe an unforeseen issue: it was found that the fineness of the grid over which the system's energy is calculated was insufficient for the PCB molecule, resulting in the periodic fluctuation in the energy corresponding to the distance between grid points that is shown in Figure 3. This is known as the 'eggbox' effect, and occurs when an inadequate grid fineness is used in calculations. While the mesh cutoff is always tested for convergence before starting production calculations, it was not expected that the eggbox effect would be observed so strongly in calculations of PCBs. The grid size was fine-tuned by running a number of suites of the box dimension parameter sweep calculations with different mesh cutoffs and searching for convergence.

For each of the box convergence runs it was only necessary to perform single point calculations on the system to determine the degree of interaction across the periodic boundary conditions, resulting in the large number of short calculations for which high-throughput *e*Science is particularly useful.

2.4 Calculating Torsional Energy

The structure of ortho-2,2'-dichloro biphenyl (Figure 2) was described in z-matrix format, so that the torsion angle could be constrained and, therefore, varied over 360˚. In the first instance it was considered adequate to sample every 5˚ over the rotation around the central carbon bond. These calculations were autogenerated and autosubmitted as previously described.

The only constraint on the system was the fixing of the torsion angle, so the positions of the other atoms were allowed to optimise around this fixed angle.

2.5 *e*Science Tools Used

The SIESTA input files for each suite of calculations were automatically generated using parameter sweep scripts which have been developed within the *e*Minerals project precisely to leverage the enhanced computing power made available through grid technology. These scripts are described in detail in [7], along with a detailed description of my_condor_submit (MCS), the submission script used in this work. However, a short description will be given here for clarity regarding this work.

The parameter sweep scripts use, as input, a template SIESTA input file, which can be modified to create all of the required input files, and a very simple configuration file. The configuration file is used to specify the input parameters that should be varied within the input file and the range and number of values over which they are to be varied. From this simple description of the problem space, running one command results in the creation of both a local and a logical Storage Resource Broker (SRB) directory structure, and the creation of the required SIESTA input files within this structure. Also, relevant MCS input files are created in the local directory structure.

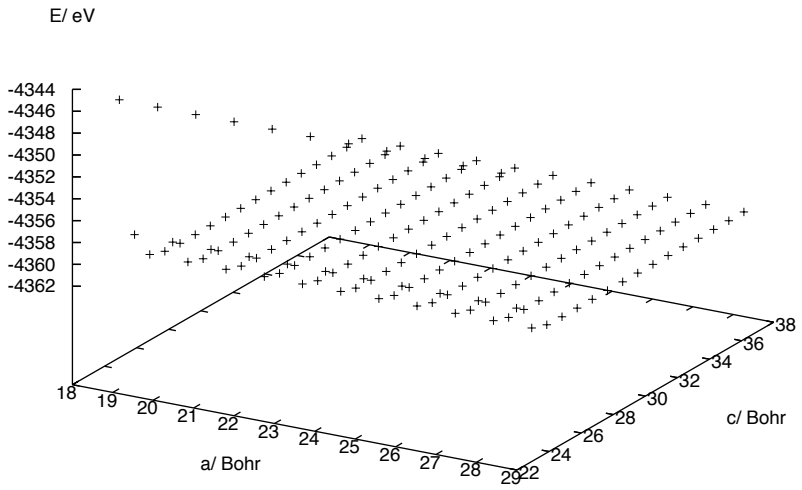

Figure 4 Energy v. box dimensions from calculations with a 300Ry cutoff and fcc + box centre grid sampling

The MCS files specify each of the jobs to be run, including the SRB location of the executable and input files, from where they should be downloaded and to where output files should be uploaded. In addition, the MCS files detail any relevant metadata to be collected. In the case of the box size convergence, the additional metadata requested were the lattice parameters, which were extracted from the XML SIESTA output file on completion of the calculation.

Once the directory structure has been created, jobs are submitted by running a second command which walks the local directory structure and submits all of the input files using MCS which takes care of all appropriate data management and the meta-scheduling over the available computing resources. This use of MCS as the underlying submission system means that we can maximise the use of available resources and minimise the latency between job submission and results retrieval.

The calculations are metascheduled across the *e*Minerals minigrid [8] using Globus to check the machine availability, and Condor-G to submit the jobs, around which MCS is wrapped to enable communications with the SRB and metadata collection / storage . The user has the option of specifying the machine on which the calculation is run, or whether to run on a condor pool or a cluster.

Such a large number of calculations as encountered here quickly becomes unmanageable and it becomes difficult to trace individual jobs or suites of calculations on the SRB file structure. Consequently, the use of metadata, and its automatic extraction from XML files using AgentX [9], a library for logically based xml data handling, is of paramount importance in such a combinatorial study as this one.

In general terms metadata is stored in three tiers: the study, which is a self-contained piece of work; the dataset; and data objects. Each study can be labelled with various topics from a controlled taxonomy and can be annotated with a high level description of this work. The study level metadata can be searched at a later date either via keywords within these annotations or via the topic labels. Searches can be performed using the RCommands, which can be run from the command line or the Metadata Manager, a web interface to the metatdata database. There tools are fully discussed in [10], but a description of the specifics of their use in this study follows.

In this case, prior to commencing the runs, a study for PCB molecules was created on the metadata database, and within this study each suite of calculations constituted one dataset. Each data object within the dataset relates to one directory containing the files for one calculation. The data object is associated with the URI for the directory within the SRB. As with the study, it is possible to annotate the dataset and data objects with descriptions, which can later be searched for keywords. In addition to these free text annotations, dataset and data objects can have parameters associated with them. These are arbitrary name value pairs, which can be used to index the data using key parameters of interest. If these parameters are numerical, it is possible to search for data objects or datasets that have a specified value for a certain parameter.

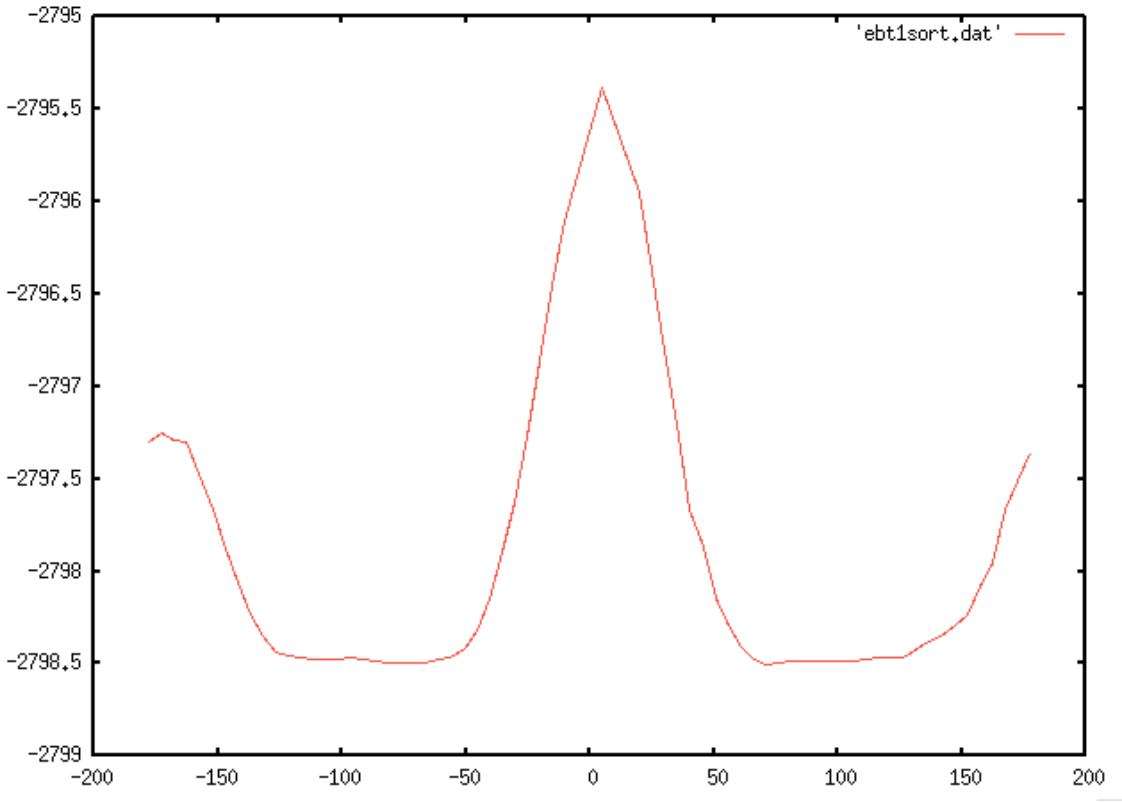

Figure 5 Energy (eV) v. torsion angle (degrees) with autogenerated basis set

In this work, each data object is associated with a number of parameters that are harvested on job completion by MCS, from both the MCS environment (including the name of the computer from which the jobs were submitted and the submission directory and the remote machine upon which the job has been run) and any metadata available from the simulation code (e.g.: the executable name and version). In addition, AgentX was used to harvest from the XML output file any metadata requested for extraction as specified in the configuration file used to generate the individual MCS scripts.

For the box convergence study, numerical values of the lattice vectors were stored as metadata, allowing the metadata to be searched for this parameter and a specified value. So, on completion of each calculation, the results are uploaded to the SRB and metadata is automatically updated. This removes the requirement from more traditional methods of logging into each remote machine to retrieve results, and making a record of each calculation in order to track the workflow. Obviously, with over one thousand jobs run in this study, such a method of working would be unmanageable, and prohibitive of this detailed type of parameterisation study.

With such a large number of files and suites as required for this study, and with the frequent updates of the simulation code used, it is of paramount importance to be able to trace the details of each calculation, both in order to process the results and to enable traceability of the workflow. This is important both for the individual scientist carrying out the calculations and for the facilitation of collaborations, which are integral to a virtual organisation as distributed as the *e*Minerals project.

The torsion angle calculations were constrained geometry optimisations. These calculations were generally well behaved; however, it was necessary to check the convergence of the calculations to ensure that the energy extracted from the XML output file would be the correct energy for the system. A number of python scripts and shell commands were used to check for convergence. Thereafter, an XSLT file was used to parse each XML file and extract the lattice vectors and total energy of the system, the data from which could then be plotted, and the images uploaded to the SRB for viewing by collaborators.

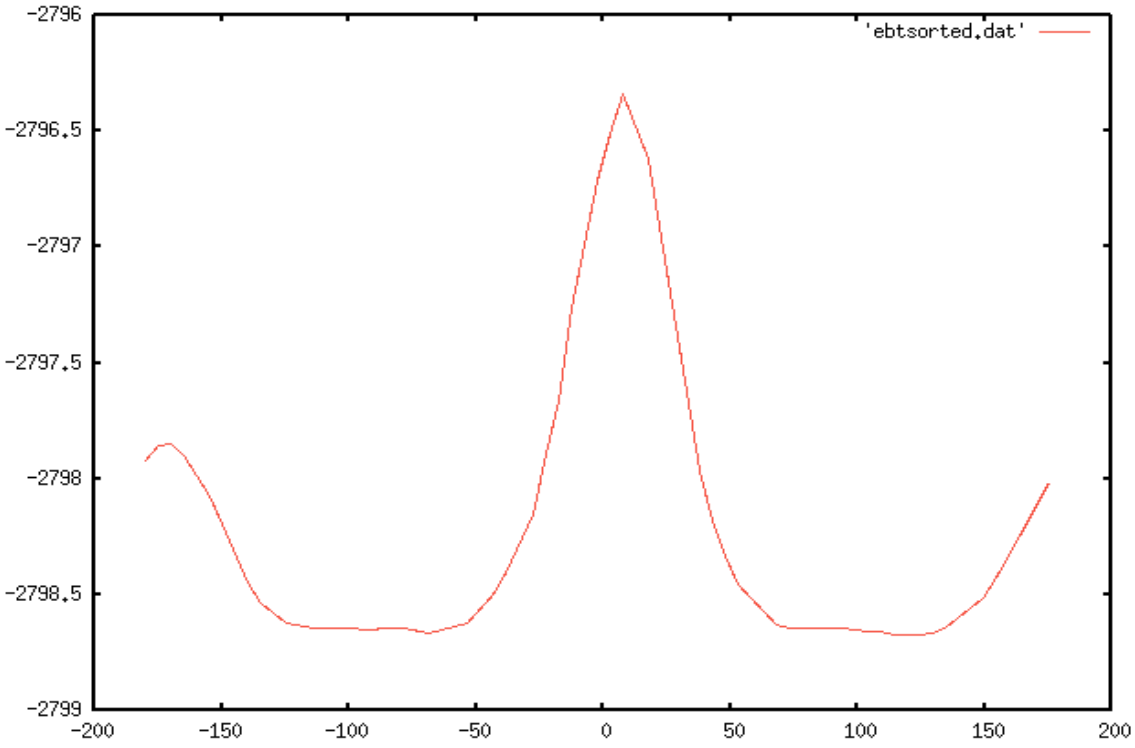

Figure 6 Energy (eV) v. torsion angle (degrees) with user-specified basis set

3. Results

3.1 Box Size Calculations

The first dimension sweep performed on the box containing the 6-PCB molecule showed wave-like perturbations in the energy with increasing box size (Figure 3). Further analysis revealed that energy minima coincide with grid points, such that the perturbations are synchronised with grid periodicity. Consequently, further suites of calculations were performed, 15 in total, in order to determine the least computationally expensive method to minimise this 'eggbox' effect. In addition to increasing the fineness of the grid, the grid sampling method was used in the hope of reducing the mesh cutoff needed to minimise these fluctuations. The graph showing the optimal combination of these factors, plotting box size against energy, is shown in Figure 4. The parameters decided upon were a 300 Ry cutoff and a combination of fcc and box centre grid cell sampling. It can be seen from the plot that a box size of 24 x 30 Bohr (roughly 13 x 16 Å) proved sufficient for removing size effects. The total number of calculations carried out to parameterise the model to such a high level of accuracy was 1815. Using conventional methods, this level of exploration of the parameter space would be prohibited by the hours required to generate and submit the calculations, and to process the results. It is probable that the 'eggbox' effect, so easily identified with the three-dimensional plots obtained *via* these calculations, would have gone unobserved, were traditional parameterisation methods used in the study. Usually, for such a study, the box size would be increased in fairly large increments, until it appears that the energy is near convergence. Thereafter, a few calculations that vary the lattice parameters with smaller increments would be carried out to tune the values, and convergence would be considered to have been achieved. It is conceivable that, using this method, the investigator could happen upon a well in the eggbox, and believe the parameters to be converged without seeing the larger picture and hence without realising that the mesh cutoff was insufficient for this study.

3.2 Torsional Energy

The energy of each calculation is plotted against the torsion angle in Figure 5. The calculations were carried out with an autogenerated DZP basis set as the first pass. The box size and mesh cutoff obtained from the box size parameterisation were used in the study to ensure that the calculations were of the highest accuracy in this respect.

The effect of different basis sets was to be determined in this study, in order to find the least computationally expensive, adequate description of the torsional rotation around the central C–C bond. As such, the autogeneration of the suites of calculations was extremely useful, as a change in the

basis set in the SIESTA input file was all that was required in order to generate and submit all the required jobs. At present, two basis sets have been tested. The autogenerated SIESTA DZP basis set, and a set of basis sets taken from previous parameterisation calculations for the PCDDs[1].

Inspection of Figure 5 shows that the barrier to rotation is highest at 0º, and that there is a minimum in the energy at around 70º, as expected, although this is only slightly lower in energy than the broad minimum within which it sits, between 70º and 130º. The kinetic barrier to rotation is around 1.25 eV at 180º and 3.25 at 0º, although a more detailed sampling is needed at 0º in order to determine this value accurately.

An analogous plot is shown for the explicit basis sets in Figure 6. It can be seen that, once again, the barrier to rotation is at 0º, and another at 180º. Once again an energy minimum is found to be at 70º, but a second minimum is apparent at roughly 120º. The barrier to rotation in this case, however, can be seen to be considerably lower at 2.5 eV. Further investigation is obviously necessary to ascertain the best description of the molecule, which will include the repetition of the suites of torsion angle calculations using different basis sets. Additionally, the *e*Science tools that have already been used can further be applied with different codes that are not limited to DFT calculations. Investigations into the applicability of DFT to the problem will be made by the study of the system using higher-level quantum mechanical calculations.

4. Conclusions

Quantum mechanical calculations have been carried out using *e*Science tools developed during the *e*Minerals project. These tools change the way that computational chemists are able to carry out scientific research for a number of reasons. First, the tools allow for facile generation of many files, removing the necessity for lengthy set-up times. Secondly, the metascheduling aspect of the setup and the use of the SRB for data storage, along with the use of digital certificates for security, removes the need for monitoring of the machines upon which the calculations are running, or the need to directly log into the machines. Searchable metadata, through the use of the RCommands and Metadata Manager, ensures that the vast numbers of calculations do not get confused and thereby remain manageable. The use of tools that test convergence in the jobs, along with the structure of XML output, which allows for the easy extraction of relevant data, means that the processing of the large number of results does not pose a problem to the research scientist.

The consequence of using all of these tools is that it is much easier to properly parameterise the models used, as a comprehensive sweep of parameter space is neither lengthy nor computationally too expensive when using high-throughput computing. In this particular case, the parameterisation led to the identification of a phenomenon that was not expected to be observed in this system, which might have gone unobserved using traditional methods.

5. Acknowledgments

We are grateful for funding from NERC (grant reference numbers NER/T/S/2001/00855, NE/C515698/1 and NE/C515704/1).

References:

1. TOH White, RP Bruin, J Wakelin, C Chapman, D Osborn, P Murray-Rust, E Artacho, MT Dove, M Calleja, *e*Science methods for the combinatorial chemistry problem of adsorption of pollutant organic molecules on mineral surfaces. *Proceedings of the All Hands Meeting, Nottingham, 773-780* (2005).
2. C Romming, HM Seip, and Aaneseno.Im, Structure of Gaseous and Crystalline 2,2'-Dichlorobiphenyl. *Acta Chemica Scandinavica Series A-Physical and Inorganic Chemistry A 28 (5), 507-514* (1974).
3. R Zimmermann, C Weickhardt, U Boesl, EW Schlag, Influence of chlorine substituent positions on the molecular structure and the torsional potentials of dichlorinated bipheyls: R2P1 spectra of the first singlet transition and AM1 calculations. *Journal of Molecular Structure 327, 81-997* (1994).
4. JM Soler, E Artacho, JD Gale, A Garcia, J Junquera, P Ordejon, D Anachez-Portal, *T*he SIESTA method for ab initio order-N materials simulation. *Journal Of Physics-Condensed Matter 14 (11), 2745-2779* (2002).
5. JP Perdew, K Burke, and M Ernzerhof, Generalized Gradient Approximation Made Simple. *Physical Review Letters 77, 3865-3868* (1996).
6. FH Allen, The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallographica Section B-Structural Science 58, 380-388* (2002).
7. RP Bruin, TOH White, AM Walker, KF Austen, MT Dove, et al., Job submission to grid computing environments. *Submitted to All Hands Meeting 2006* (2006).

8. M Calleja, R Bruin, MG Tucker, MT Dove, R Tyer, L Blanshard, K Kleese Van Dam, RJ Allan, C Chapman, W Emmerich, P Wilson, J Brodholt, A Thandavan and VN Alexandrov, Collaborative grid infrastructure for molecular simulations: The *e*Minerals minigrid as a prototype integrated compute and data grid. *Molecular Simulation, 31, 5, 303-313 (2005)*
9. PA Couch, P Sherwood, S Sufi, IT Todorov, RJ Allan, PJ Knowles, RP Bruin, MT Dove and P Murray-Rust, Towards Data Integration for Computational Chemistry. *Proceedings of the All Hands Meeting*, Nottingham (2005).
10. RP Tyer, PA Couch, K Kleese van Dam, IT Todorov, RP Bruin, TOH White, AM Walker, KF Austen, MT Dove, MO Blanchard, Automatic metadata capture and grid computing. *Submitted to All Hands Meeting 2006* (2006)