# Deriving Trust Levels for Multi-Choice Data Analysis Workflows

**Daniel Speckhard[1, 2]†, C. Carbogno[2], M. Scheffler[1,2] and Claudia Draxl[1, 2]**

[1]Humboldt-Universität zu Berlin, Institut für Physik and IRIS Adlershof, Berlin, 12489, Germany

[2]Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin, 14195, Germany

Computational materials science simulations provide a discovery route to new novel materials away from the traditional path of expensive bench top experiments. Density functional theory (DFT) has become the workhorse of the materials science community to investigate, e.g., structural, elastic, electronic, and thermal, properties of materials. Materials databases currently host hundreds of millions of DFT simulations computed by different scientists around the world. In order to make use of these heterogeneous data sources beyond the scope of the investigations they were created for, we aim, in this study, to estimate trust levels that are related to the numerical settings used in the simulations. In order to do so, we will simulate hundreds of binary compounds with different DFT programs and settings to explore the dependence of the settings on the resulting material properties. With this dedicated data set we plan to train a machine learning model to estimate the deviation of a simulated material property with respect to the result we would expect from settings that correspond to a fully converged calculation. With a trained model in hand, we plan to apply it to materials simulation databases hosting DFT data.

## 1. Motivation and Goals

The products we use today are often limited by the materials used to build them. For example, the phone and automotive industry are constantly looking for new materials that might provide a breakthrough in energy density for batteries. Or policy makers, with an eye towards climate change, are wondering how to optimally invest research funds to find materials that can best turn waste heat into usable energy sources. As a result, a massive amount of research funds have been invested into a new age of intelligent materials discovery. The Materials Genome Initiative of the United States gave life to the computation of thousands of materials and their properties using various simulation programs. The rise of high-throughput screening of large material spaces, in turn, gave rise to a new question. Can the materials simulation community store and share these results so that researchers/engineers world wide can benefit from open and collaborative research efforts?

Materials databases such as the Novel Materials Discovery (NOMAD) Laboratory seek to facilitate storage and sharing of computed data (Draxl & Scheffler 2019). Through sharing data, we cannot only avoid repeating expensive simulations. Recent work has shown that through using larger computational materials data sets, like those hosted on NOMAD, researchers can discover new physical relationships and candidate materials for a wide variety of applications (Goldsmith *et al.* 2017). This study is related to two funded

† Email address for correspondence: speckhard@fhi-berlin.mpg.de

research projects: the DFG Collaborative Research Centre 1404 FONDA (Foundations of Workflows for Large-Scale Scientific Data Analysis), sub-project A3 (Link to FONDA website), and the NOMAD CoE which receives funding from the European Union's Horizon 2020 research and innovation program under the grant agreement number 951786.

Researchers would like to use the state of the art artificial intelligence technologies on combined data sets hosted online. The use of heterogeneous data is, however, a complex endeavour. The materials science community routinely uses about 40 different simulation programs (e.g. `exciting` (Gulans *et al.* 2014), FHI-aims (Blum *et al.* 2009), VASP (Hafner 2008), etc. . . ). These programs come with their own implementations, and thus sets of parameters and numerical settings that impact the precision of the results. For instance, a researcher may be looking into $TiO_2$ for application in heterogeneous catalysis and run a simulation program to return the band gap that is converged with respect to the range of settings/parameters specific to the program. From this simulation, the band gap result may be converged but another property like the dielectric constant may not be converged with respect to the same settings. Without knowing the potential precision error for the dielectric constant related to the used settings, re-purposing this simulation to obtain trends with respect to dielectric constants is fundamentally error-prone. This poses a challenge to researchers hoping to use data from different sources. To place data from different sources on the same footing, and thus improve interoperability of data in materials databases, normalizer programs are employed in practice. These normalizers standardize units, crystal structure definitions and common parameter definitions. They do not yet, however, give an estimate of how far a result may be from its converged value.

### 1.1. *Numerical Tools Used*

One of the bottlenecks for simulations is the basis set size. This parameter dictates the precision of a simulation. With an infinitely large basis set, the result of the simulation is as precise as possible for the chosen method. We call this limit of a result, the complete basis set (CBS) limit (Hill *et al.* 2009). Quantum chemists routinely perform a few simulations at different basis set sizes and then extrapolate results to the CBS limit using several different formulae with coefficients fitted to available data. In materials-science, convergence tests are typically routinely done, but extrapolation to the CBS not. The materials science community often employs density functional theory (DFT) codes with a variety of different basis sets. DFT calculates properties of a many-electron systems using functionals of the spatially dependent electron density. The benefit of using DFT is that the code scales $\mathcal{O}(N^3)$ where N is the basis set size. At the heart of a DFT run is an eigenvalue problem that is solved for every k-point of the Brillouin zone.

### 1.2. *Subject Specific Meaning, Expected Results and Use of Results from HLRN.*

For periodic, crystalline systems, no such calculator exists that would extrapolate DFT results to the CBS limit. This study will train a machine learning model to extrapolate fundamental physical properties like relative energies and derived properties routinely calculated by DFT such as the band gap in the Brillouin zone to the CBS limit. We aim in the worst case to estimate a confidence interval for results hosted NOMAD or a percentage error. In the best case, we aim to estimate the CBS limit result of the property with precision.

We will explore different artificial intelligence learning models to evaluate their performance in predicting the CBS values for material properties returned from DFT simulations. To do so, we will create data by two DFT codes, i.e., `exciting` which

employs linear augmented plane waves (LAPW) as its basis set and FHI-aims which employs numeric atomic orbitals (NAO) as its basis set. These two basis sets are very different from one another and are representatives of widely used DFT code families. For instance, popular LAPW codes are Wien2k (Blaha *et al.* 2001), Fleur, and Elk while popular NAO codes are Siesta and DMol3. Thus our results will be transferable to those codes. At the conclusion of this study, we will invite other DFT code developers to perform similar experiments to model the CBS limit for their code. We are already in contact with developers of two pseudopotential based DFT codes, ABINIT and GPAW, to push this plan foward. We plan to share our modeling efforts in python based Juptyer Notebooks while the raw data will be openly accessible on NOMAD. FHI-aims and `exciting` are both all-electron codes which means they are some of the most precise DFT codes that simulate the behaviour of each electron in a material. Both codes should give very similar physical results but since the basis sets are very different in nature we need to investigate their behavior separately.

We will use `exciting` and FHI-aims to simulate total energies, band gaps, elastic constants, and effective masses of binary compounds. These material properties are chosen since they are very useful to the community (e.g. band gaps indicate whether a material could be used in a solar cell while elastic constants tell us how a material behaves when being stretched). Knowing whether the properties in a DFT simulation database can be trusted could save previous computing and experimental lab hours. We choose to study a restricted class of materials, namely semiconductors composed of two different elements. Binary semiconductors have a wide variety of uses in fundamental science and in industry in LEDs, solar cells, power electronic components, computing processors, etc... At the same time they have small enough unit cells that allow for highthroughput calculations.

### 1.3. *Preparatory Work*

Research conducted by several of the applicants showed that simple models on small data sets 63 solid binary compounds) can be used to estimate the CBS total energies (energy errors) in three DFT codes (Carbogno *et al.* 2019). Recent work has demonstrated that these CBS total energy estimates resulting from unconverged settings can be improved using more sophisticated machine learning models as seen in Figure 1. The performance of this random forest model, giving mean percentage errors of around 20%, is not satisfactory yet but quite promising.

The applicants also looked into quantifying the precision of DFT code settings with respect to more complicated material properties like relative energies which measure the energy difference when expanding the unit cell (e.g. 5% expansion in all lattice vector directions) as a function of basis set size. The applicants found it difficult (R2 of best model was 0.82) to model CBS relative energies with a small data set. The learning curve of a random forest model fit to the relative energy data for these 63 binaries is shown in 2. We can extrapolate from the learning curve of the random forest model using an exponentially decaying function. We see that by using around 30 times as much training data we should reach our goal. This root-mean-squared error (RMSE) goal, similar to reported RMSE values of materials physics learning models (Xie & Grossman 2018), would be very useful to scientists and engineers who are browsing NOMAD for potential materials to use and would like an estimate of the precision of a simulation result.

Materials physics tell us that the process of expanding/contracting a lattice is correlated to the Bravais lattice of the material. To model elastic coefficients, band gaps, and effective masses, properties which are very much linked to the crystal structure, we believe we will need a balanced data set of each Bravais lattice. Moreover, we expect to

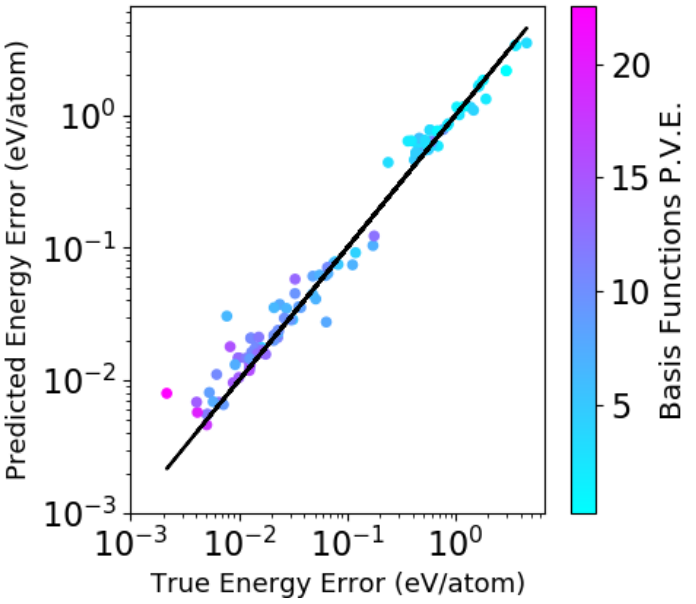


FIGURE 1. Random forest modeling on validation data of total energy-per-atom errors with respect to the converged basis set size for FHI-aims data from Carbogno *et al.* (2019). The color bar shows the number of basis functions per valence electron (abbreviated P.V.E.) used for each data point.

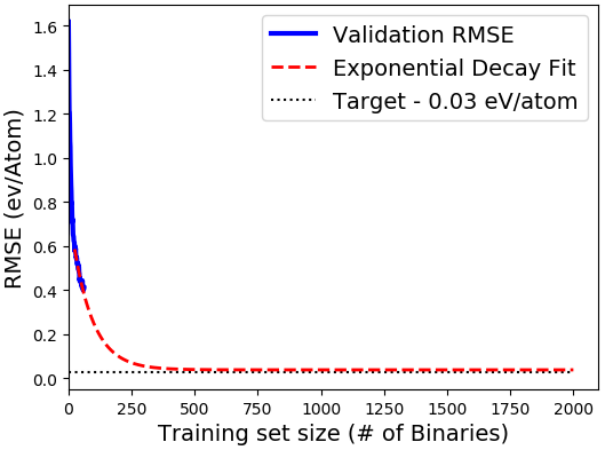


FIGURE 2. Learning curve of relative energy differences using a random forest model. The relative energy differences are calculated varying the basis set size with FHI-aims. The blue solid line shows the root mean squared error (RMSE) of the model on the validation data set. The red dashed line shows an exponentially decaying function fit to the learning curve of the model. The black dotted line shows the 0.03 RMSE eV/atom target for the model performance.

need a similar amount of data to learn these properties as calculated by extrapolating the learning curve of relative energy data.

The binary semiconductors are chosen from the NOMAD database. The total number of binary semiconductors hosted on NOMAD (not including variants of unit cells) is 1915. This includes structures from each of the 14 Bravais lattices. We choose to simulate these 1915 structures. This should provide us with sufficient training data to model the

| Task | Quarter/Year | | | |
|---|---|---|---|---|
| | Q1 2021 | Q2 2021 | Q3 2021 | Q4 2021 |
| Monoclinic Lattices | | 366 | | |
| Triclinic Lattices | | 97 | | |
| Orthorhombic Lattices | | | 478 | |
| Tetragonal Lattices | | | | 205 |
| Rhombohedral Lattices | 136 | | | |
| Hexagonal Lattices | | | | 281 |
| Cubic Lattices | 352 | | | |
| LDA Ground State Simulations for 500 Materials | | | | |

FIGURE 3. We plan to split up our computations by the lattice systems. The number of binary semiconductors of each lattice system is shown in the cell indicating which quarter we plan to simulate them. We also plan to perform ground state calculations using the LDA functional for five hundred randomly selected binary semiconductors.

properties mentioned above using traditional machine learning models such as random forests. Simulating these data will also provide a valuable data set serving as a benchmark to the community that will be hosted on NOMAD. As mentioned earlier, we will invite other DFT code developers to simulate these materials varying the numerical settings. In this manner, this study will provide the first step towards material databases containing predictions for how far a data entry is from its converged value for all DFT codes. Thereby, we envision to provide data quality metrics for heterogeneous DFT data hosted on open access platforms/databases like NOMAD.

Besides the basis set size we also plan to vary the k-points used in the simulation to ensure we have converged the density and total energies. This also allows us to statistically model how the k-points changes materials properties returned by a simulation.

### 1.4. *Planned Project Schedule*

In essence, we will vary basis set sizes and k-points calculating 1915 binary semiconductors. We also will run simulations to compute elastic constants. All these calculations will be carried out with the DFT functional PBE. Finally, we will run ground state (G.S.) simulations for a subset of materials, i.e., 500 binary semiconductors using the LDA functional to compare the performance between LDA and PBE functionals. Previous experience by the applicants did not show a large change in results when changing the functional. These five hundred simulations will give us an indication of whether changing the functional makes a noticeable difference on the results. We plan a budget of computation resources for four quarters and plan to divide our simulations by lattice system type. This plan in shown visually in a Gantt chart in Figure 3. For further details on the number of simulations planned, see section 4.

## 2. Theoretical methods and numerical procedures

The numerical calculations will be performed using the `exciting` and FHI-aims DFT codes. `exciting` is a full-potential all-electron package implementing DFT and methodology beyond. The code is based on the linearized augmented planewave method (LAPW), which is known as one of the most precise numerical schemes for solving the Kohn-Sham equations of DFT (accuracy up to $1\mu$ Hartree/atom (Gulans *et al.* 2018). With `exciting` a user is allowed to control physical parameters such as the muffin-tin

radius, the cutoff energy and the number of local orbitals. FHI-aims is an all electron DFT code that employs a numeric atom-centered orbital basis set. It has broad accuracy on par with best available benchmarks and scales to very large systems. The simulations we plan are ground-state calculations.

## 3. Parallelism in `exciting` and FHI-aims

### 3.1. *Parallelism in `exciting`*

The high level of accuracy afforded by the LAPW method implemented in `exciting`, requires more computational resources, in particular, in the amount of both RAM and data storage compared to less accurate methods. Due to these demands `exciting` has been developed to use state of the art methods to improve computational costs. For shared memory multiprocessing computing, the latest version of `exciting` uses a hybrid approach combining MPI-OpenMP library. Multithreading is used in `exciting` for compute time intensive tasks such as solving the Poisson's equation for the Coulomb potential or assembling the electron density.

### 3.2. *Scalability Tests*

We can try to model the computing resources used in a DFT simulation in both the `exciting` or FHI-aims code. The objective is to estimate the speedup of the code with respect to MPI and multithreading parallelization. Putting $T_s$ ($N_m$, $N_{thr}$) as the time required by the code to perform a single diagonalization of the Hamiltonian matrix, $N_m$ is the matrix size and $N_{thr}$ the threads per MPI process. The time of a complete SCF step in exciting ($T_{scf}$) would ideally scale as

$$T_{scf} = \frac{N_s}{N_{thr}} T_s$$

where $N_p$ is the number of MPI processes and $N_k$ the total number of k-points used in the calculation for sampling the reciprocal space, i.e. the Brillouin zone. From this we can extract the number of computation units ($N_c$) for GS calculations:

$$N_c = NPL * N_n * N_k * T_{scf}$$

where $T_{scf}$ is given in hours ($T_{scf} = T_s$ in case $N_k = N_p$), $N_n$ is the number of nodes, and NPL is a North-German Parallel-Computer Work Unit defined by the HLRN.

#### 3.2.1. *`exciting` Scalability Tests*

For exciting we can examine the model presented by looking at scalability tests on the HLRN servers. The results of which are shown in Figure 4. Perhaps equally relevant given this project's nature, We can look at how setting physical/numerical parameters affects computation time of a simulation. We've shown the matrix size affects the computation time of exciting calculations. Figure 5 shows computation time as a function of changing the rkmax parameter which (along with the fixed unit cell in this case) dictates the matrix size. Note the results are shown for a benchmark binary material $Cd_4O_4$. The rkmax parameter is a function of the muffin tin radius and the plane wave energy cutoff. We vary the rkmax parameter until it gives a energy value that changes by less than 10E-4 eV/atom. Unsurprisingly, as we expect simulations to scale $\mathcal{O}(N^3)$ where N is the basis set size, we see larger rkmax parameters take exponentially longer time to compute than smaller values.

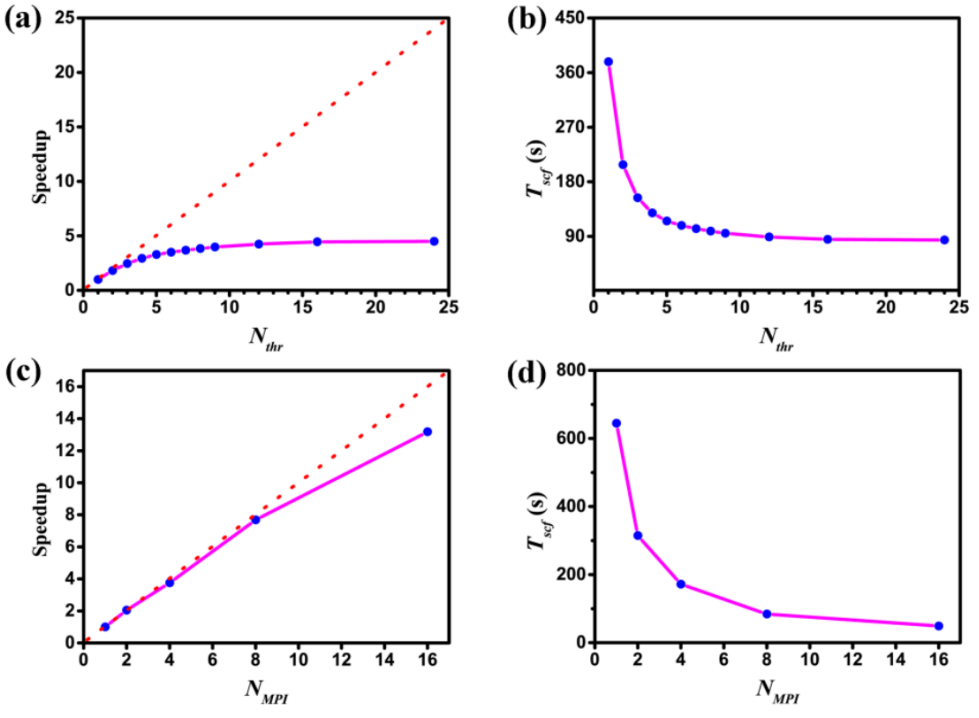


FIGURE 4. Speedup and (b) time for $T_{scf}$, as a function of the number of threads $N_{thr}$ in the GS calculation of GaN (10-10) surface with 16 layers performed by Cecilia Vona on HLRN servers using `exciting`. (c) Speedup and (d) time for $T_{scf}$, with respect to MPI parallelization. The red dashed line corresponds to the ideal speed up. Tests are made in the queue mpp2testq. Although the unit cell here is much larger than the ones we plan to simulate it serves as a benchmark for how `exciting` scales on HLRN servers.

#### 3.2.2. *FHI-aims Scalability Tests*

We can also estimate how long each FHI-aims simulation will take by looking at the effect of changing the basis set size and numerical setting parameters on the computation time. Figure 5 shows computation time for different settings for the model binary system of $Cd_4O_4$. These tests were run on the Max Planck Computing Data Facility COBRA machine using 32 nodes. For comparison with HLRN servers, we repeated the minimal-light simulation using the standard96 queue and got a computation time of 7.8 seconds (compare this with the 32 seconds used to run on MPCDF). This equates to a speedup of roughly three times in terms of computation time with three times as many processors.

We can see from 5 that as we increase the basis size we see an exponentially longer computation time in line with our expectation of the eigenvalue problem scaling as $\mathcal{O}(N^3)$ where N is the basis set size.

## 4. Estimation of Operating Resources Requested

We plan to run several different settings for both `exciting` and FHI-aims while simulating in the worst case 1,915 binary semiconductor compounds and in the best case only 210 binaries. We also plan to deviate the unit cell around 8 times to uncover details about the elastic constant and effective mass of the structure. This is done in

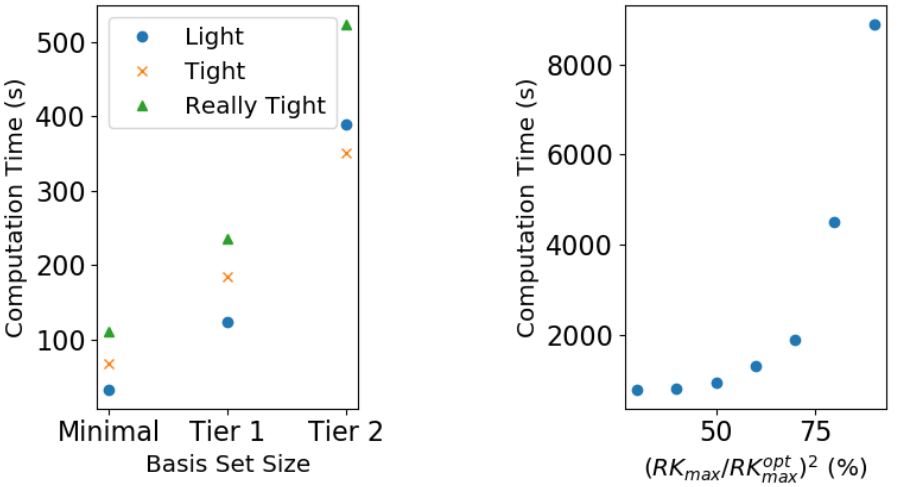


FIGURE 5. Computation time benchmarking for different settings in FHI-aims (left) and `exciting` (right). The basis set size for FHI-aims data was varied as was the numerical settings (different markers in the plot show light, tight and really tight settings). The exchange-correlation functional was PBE and the k point density was fixed as 8. These tests were performed for FHI-aims on a single node 32 core COBRA machine of the Max Planck Data Computing Facility (MPCDF). The numbers shown here compare well with the standard96 queue when tested. The `exciting` scalability tests are performed as a function of the $\mathrm{RK}_max$ parameter. The x-axis parameter $(RK_{max}/RK_{max}^{opt})^2$ encodes how far the rkmax parameter used in the simulation is from the $\mathrm{RK}_max$ value that gives a converged ground state energy, $RK_{max}^{opt}$. The `exciting` simulations were performed with a single 4 core node on DUNE servers at HU-Berlin.

`exciting` and FHI-aims using the `ElaStic` package (Golesorkhtabar *et al.* 2013). To ensure that that densities and total energies are converged with respect to the k-points we also plan to run four different k-points settings (equally spaced) for each simulation.

If we include an estimate of eight different ground state calculations using `ElaStic`, we multiply 48 settings of basis set size by 4 different k-points settings by 1,915 binary (Table 1) to find the total number of simulations for `exciting`. This adds up to 283,008 simulations. We add another 500 ground state simulations of binaries chosen randomly from our set and simulated with the LDA functional. Recall, these LDA simulations on a small portion of the binary set are performed to see if changing the functional makes a noticeable difference. We recall from the previous section not all `exciting` simulations run will take equal amount of time and depend largely on the settings used. From Figure 5 we can see an average computation time of roughly 2740 seconds (0.76 hours) using a single node 4 core server over the different $\mathrm{RK}_max$ parameters. We also noted from Figure 4 that we see a less than linear scaling of computation time with the number of threads/cores. If we use the charge rate for the SKL medium40 (phase 1) queue with 40 cores and a charge rate of 6 NPL per node hour, and use our estimate of 0.76 hours per simulation and multiply by a rough estimate of 4/40 to predict the speedup for the increased number of cores we end up with a rough estimate of required resources of $6 * 0.76 * 1/10 * 2,941,940 = 1,341,525$ NPL (1,341 kNPL).

We can see from Table 2 the estimate of a total of 735,860 simulations to be run with FHI-aims. From our benchmark binary of $Cd_4O_4$ in Figure 5 a simulation averaged over basis set size and numerical settings on a 32 core single node takes 224 seconds (0.06

| Setting | Number of Different Settings to Use |
|---|---|
| Muffin Tin Radius | 4 |
| Plane Wave Cutoff Energy | 4 |
| Number of Local Orbitals | 3 |
| Number of k-points | 4 |
| Ground State Calcs in `ElaStic` | 8 |
| Binary Semiconductors | 1915 |
| Total Sims + 500 Ground State Calcs | 2,941,940 |

TABLE 1. Table detailing calculations to be done in `exciting`

| Setting | Number of Different Settings to Use |
|---|---|
| Basis Set Size | 4 |
| Numerical Settings | 3 |
| Number of k-points | 4 |
| Ground State Calcs in `ElaStic` | 8 |
| Binary Semiconductors | 1915 |
| Total Sims + 500 Ground State Calcs | 735,860 |

TABLE 2. Table detailing calculations to be done in FHI-aims

hours). Therefore for similar single node 40 core charge rate for the SKL medium40 (phase 1) of 6 NPL per hour, we estimate $0.06 * 6 * 735,860 = 264,910$ NPL (265 kNPL).

Therefore, in total for both FHI-aims and `exciting` we estimate a required budget of 1,341 kNPL + 265 kNPL = 1,606 kNPL. We request this budget over 4 quarters. Dividing the total budget over 4 quarters, we request 401.5 kNPL per quarter.

### 4.1. *Storage Requirements*

Input/output in `exciting` and FHI-aims are carried out using the standard Fortran utilities. Small size formatted text files are printed to monitor the working progress. The large size data is stored into binaries. For ground state calculations, at each SCF cycle, eigenvalues, eigenfunctions and occupation numbers are stored into MPI-rank dependent output files, and are correspondingly read when assembling the total electron density. From previous work, we estimate an average of 1 Mb of space per simulation. For the combined 3,677,800 simulations we plan to compute over four quarters, we will generate 3.7 TB of data. We plan however to upload this data each quarter to NOMAD where it

will be hosted and accessible to the community. Therefore we request a budget of 1 TB to host a quarters worth of simulations.

## 5. References


REFERENCES

Blaha, Peter, Schwarz, Karlheinz, Madsen, Georg KH, Kvasnicka, Dieter, Luitz, Joachim & others 2001 wien2k. *An augmented plane wave+ local orbitals program for calculating crystal properties* .

Blum, Volker, Gehrke, Ralf, Hanke, Felix, Havu, Paula, Havu, Ville, Ren, Xinguo, Reuter, Karsten & Scheffler, Matthias 2009 Ab initio molecular simulations with numeric atom-centered orbitals. *Computer Physics Communications* **180** (11), 2175–2196.

Carbogno, Christian, Thygesen, KS, Bieniek, B, Drax, C, Ghiringhelli, LM, Gulans, A, Hofmann, OT, Jacobsen, KW, Lubeck, S, Mortensen, JJ & others 2019 Numerical quality control for dft-based materials databases .

Draxl, Claudia & Scheffler, Matthias 2019 The nomad laboratory: from data sharing to artificial intelligence. *Journal of Physics: Materials* **2** (3), 036001.

Goldsmith, Bryan R, Boley, Mario, Vreeken, Jilles, Scheffler, Matthias & Ghiringhelli, Luca M 2017 Uncovering structure-property relationships of materials by subgroup discovery. *New Journal of Physics* **19** (1), 013031.

Golesorkhtabar, Rostam, Pavone, Pasquale, Spitaler, Jürgen, Puschnig, Peter & Draxl, Claudia 2013 Elastic: A tool for calculating second-order elastic constants from first principles. *Computer Physics Communications* **184** (8), 1861–1873.

Gulans, Andris, Kontur, Stefan, Meisenbichler, Christian, Nabok, Dmitrii, Pavone, Pasquale, Rigamonti, Santiago, Sagmeister, Stephan, Werner, Ute & Draxl, Claudia 2014 Exciting: a full-potential all-electron package implementing density-functional theory and many-body perturbation theory. *Journal of Physics: Condensed Matter* **26** (36), 363202.

Gulans, Andris, Kozhevnikov, Anton & Draxl, Claudia 2018 Microhartree precision in density functional theory calculations. *Physical Review B* **97** (16), 161105.

Hafner, Jürgen 2008 Ab-initio simulations of materials using vasp: Density-functional theory and beyond. *Journal of computational chemistry* **29** (13), 2044–2078.

Hill, J Grant, Peterson, Kirk A, Knizia, Gerald & Werner, Hans-Joachim 2009 Extrapolating mp2 and ccsd explicitly correlated correlation energies to the complete basis set limit with first and second row correlation consistent basis sets. *The Journal of chemical physics* **131** (19), 194105.

Xie, Tian & Grossman, Jeffrey C 2018 Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters* **120** (14), 145301.