Materials Discovery aided by Statistical Learning

Deriving Trust Levels for Multi-Choice Data Analysis Workflows

D. Speckhard[1, 2], C. Carbogno[2], C. Draxl[1, 2], *Humboldt-Universität zu Berlin[1], Fritz-Haber-Institut der Max-Planck-Gesellschaft[2]*

In Short

- Trust Level Estimation
- Materials Simulation Databases
- Novel Materials Discovery

The products we use today are often limited by the materials used to build them (e.g. phone display durability, electric car battery capacity, solar panel efficiency). Simulating materials properties to find good candidate materials has allowed researchers and engineers to avoid wasteful trial and error in costly laboratory settings. Materials simulation databases such as the Novel Materials Database (NOMAD, `https://nomad-coe.eu/`) allow researchers to upload their simulation results to an openly accessible database. As a result, researchers are able to train models on datasets from different sources to discover new candidate materials and physical relationships [1].

The simulated materials property data hosted on NOMAD (over 700 million simulations) come from a wide variety of simulation programs (40+) employing a wide variety of physical and numerical settings. For instance, a researcher may be looking into TiO_2 for heterogeneous catalysis applications and run a simulation program to return the bandgap that is converged with respect to the range of settings/parameters available in the program. In this simulation, the bandgap result may be converged but another property like the dielectric constant may not be converged with respect to the settings available. This work aims to assign trust levels to data simulated using a density functional theory (DFT) code. A trust level is assigned for a material property result from a simulation (e.g. elastic constants, bandgaps, effective masses, total energies) based on what material was simulated and what settings were used in the DFT simulation.

Research has shown that simple models on small datasets (65 solid binary compounds) can be used to estimate the total energy error resulting from employing unconverged settings in three density functional theory (DFT) codes [2]. Recent work has demonstrated that these total energy error estimates resulting from unconverged settings can be improved using more sophisticated statistical learning models as seen in Figure 1. The learning curve, seen in Figure 2, of using a random forest to model total energies on the data from [2], shows the model performing better on the validation dataset total as the size of dataset increases. The performance of modelling total energies with mean percentage errors of around 20% is quite promising. Although this model on data from [2] has shown great potential, a dataset of only 65 binary materials is far too small to provide predictions on the breadth of binary materials stored in material simulation databases.

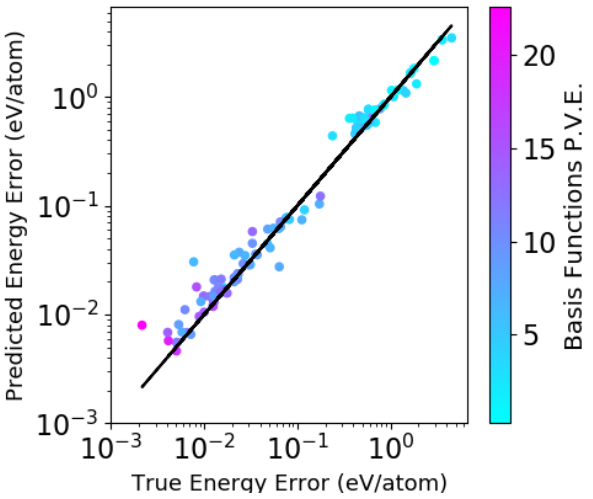

Figure 1: *Random forest modelling results of total energy-per-atom root mean square errors (RMSE) with respect to the basis set size for FHI-aims data. The color bar shows the basis functions per valence electron used for the binary compound in the simulation in the validation dataset.*

This project seeks to develop a dedicated data set of 4028 binary semiconductor compounds to improve the estimation of errors in material properties resulting from unconverged settings. Two DFT codes, FHI-aims and `exciting`, which use very different basis sets are employed to simulate total energies, bandgaps, elastic constants and effective masses of these 4028 binary compounds. For each binary compound we also vary the physical and numerical settings of the DFT codes. Statistical learning models trained on this dedicated data will enable us to derive trust level estimates for heterogeneous data of the materials science community, as stored in the NOMAD laboratory. This will allow data driven materials science researchers to select datasets from NOMAD that meet a required estimated trust level for a certain material property and bypass issues that result from using material properties that are not converged with respect to DFT

settings.

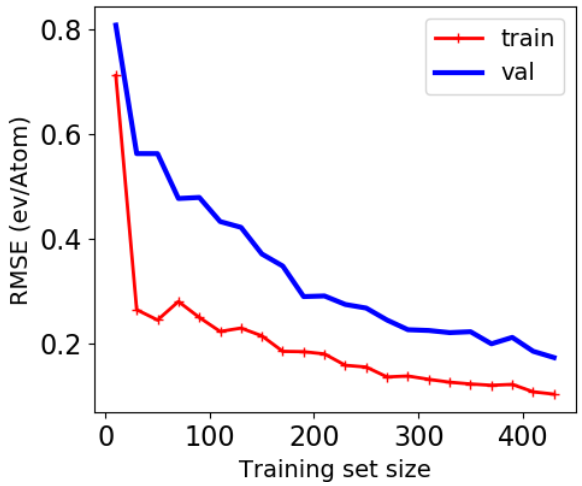

Figure 2: *Learning curve of random forest algorithm modelling energy differences with respect to numerical settings and basis set size for FHI-aims DFT data of [1]. The red line shows the root mean squared (RMSE) of the model on the training data. The blue line shows the RMSE of the model on the validation dataset.*

The research questions that this dedicated data set will enable us to answer are: 1) What models best estimate the material property error resulting from unconverged DFT settings. 2) What can we learn from modelling the materials class of binary semiconductors in order to extend this approach to wider material classes? 3) What models allow us to reduce the uncertainty in our estimates of material property errors so as to return narrow confidence intervals.

WWW

https://sol.physik.hu-berlin.de

More Information

[1] Draxl, C., Scheffler, M. . The NOMAD laboratory: from data sharing to artificial intelligence. *Journal of Physics: Materials*, **2019**, 2(3), 036001.

[2] Carbogno, C., Thygesen, K., Bieniek, B., Draxl, C., Ghiringhelli, L., Gulans, A., ... Strange, M. (2018). Quality Control of Numerical Settings for DFT Calculations and Materials Databases. *APS*, **2018**, P12-003.

Project Partners

Prof. Grunkse, Computer Science Department, Humboldt-Universität zu Berlin

Funding

FONDA: DFG SFB 1404, https://fonda.hu-berlin.de/.

NOMAD CoE which is funded from EU Horizon 2020 research and innovation program under grant agreement number 951786, https://nomad-coe.eu/