Materials Discovery aided by Statistical Learning

Deriving Trust Levels for Multi-Choice Data Analysis Workflows

D. Speckhard[1, 2], C. Carbogno[2], M. Scheffler[1, 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 Discovery (NOMAD, `https://nomad-coe.eu/`) Repository, allow researchers to upload their simulation results to an openly accessible database. As a result, researchers are able to train models on data sets from different sources to discover new physical relationships and candidate materials [1].

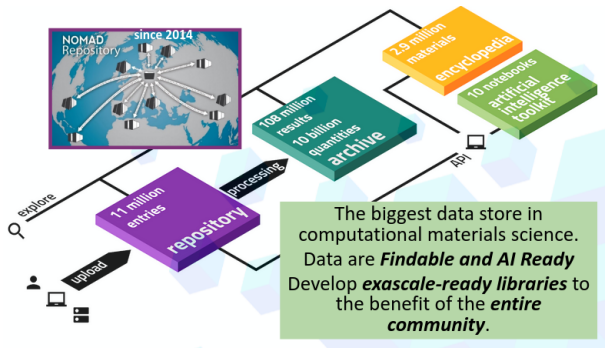

Figure 1: *Concept of the NOMAD Laboratory, the biggest data store of computational materials-science data.*

The simulated materials data hosted on NOMAD (over 100 million simulations) are simulated from a wide variety of density functional theory (DFT) programs (40+) employing a wide variety of physical and numerical settings (more information seen in Figure 1). For instance, a researcher may be looking into TiO_2 for heterogeneous catalysis applications and run a simulation program to return the band gap that is converged with respect to the range of settings/parameters available in the program. In 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. This work aims to assign a trust level to a material property (e.g. elastic constant, band gap, effective mass, total energy) resulting from a simulation based on what material was simulated and what settings were used in the DFT simulation.

Research conducted by the applicants has shown that simple models on small data sets (63 solid binary compounds) can be used to estimate the total energy error resulting from employing unconverged settings in three 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 (e.g. random forests) as seen in Figure 2.

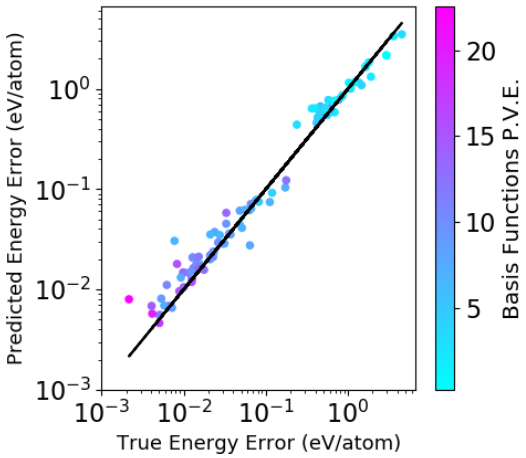

Figure 2: *Random forest modeling on validation data of total energy-per-atom errors with respect to the converged basis set size. The color bar shows the number of basis functions per valence electron (abbreviated P.V.E.) used for each data point.*

To model useful quantities such as elastic coefficients, band gaps and effective masses of materials, we will need a much larger data set. Since these properties may be sensitive to the structural configuration, we need to have representative data for all Bravias lattices. We choose to study binary semiconductors due to their simplicity and wide spread use in fundamental science and industry. The applicants plan to use HLRN servers to perform DFT simulations of 1915 binary semiconductors, a number that should provide sufficient training data to accurately model the effect of DFT settings on the convergence of these properties. The binary semiconductors and correspond structures are chosen from the NOMAD database. In Table 1, we see the number of unique binary semiconductors hosted on NOMAD for each Bravais lattice.

Bravais Lattice	Number of Binaries
Face-centered cubic	185
Body-centered cubic	54
Primitive cubic	113
Primitive hexagonal	281
Primitive tetragonal	116
Body-centered tetragonal	89
Primitive rhombohedral	136
Primitive triclinic	97
Primitive monoclinic	195
Base-centered monoclinic	171
Primitive orthorhombic	337
Base-centered orthorhombic	84
Body-centered orthorhombic	24
Face-centered orthorhombic	33

Table 1: *Number of simulated unique semiconductor binaries hosted on NOMAD for each Bravais lattice.*

We plan to consider these binary semiconductors using two DFT codes, FHI-aims and `exciting` to compute total energies, band gaps, elastic constants and effective masses. These material properties are chosen since they they characterize materials from different viewpoints. For each of the 1915 binary semiconductors, we vary the numerical settings of the DFT codes. This will allow us to get information on how sensitive different quantities are with respect to the numerical parameters. Statistical learning models trained on this dedicated data will then enable us to derive trust level estimates for heterogeneous data of the materials science community, as stored in the NOMAD Repository. This will allow data driven materials science researchers to select data sets from NOMAD that meet a required estimated trust level for a certain material property and bypass issues that result from using data that are not converged with respect to DFT settings.

We emphasize that this work will provide at the same time a comprehensive set of benchmark data that is urgently needed in the community. In a next step, we plan to invite other DFT code developers to perform similar experiments with their code. The ultimate aim is that any simulation using any DFT code on a materials database will have a model to predict how far a material property is from its converged value.

The research questions that this dedicated data set will enable us to answer are: 1) What can we learn from modeling the material class of binary semiconductors in order to extend this approach to wider material classes? 2) What models best estimate the material property error resulting from unconverged DFT settings. 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, Claudia, and Matthias Scheffler. "The NOMAD laboratory: from data sharing to artificial intelligence." Journal of Physics: Materials 2, no. 3 (2019): 036001.

[2] Carbogno, Christian, Kristian Sommer Thygesen, Björn Bieniek, Claudia Draxl, Luca M. Ghiringhelli, Andris Gulans, Oliver T. Hofmann et al. "Numerical quality control for DFT-based materials databases." arXiv preprint arXiv:2008.10402 (2020).

Project Partners

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

Funding

FONDA: DFG SFB 1404, `https://fonda.hu-berlin.de/`.
NOMAD CoE funded from EU Horizon 2020 research and innovation program under grant agreement number 951786, `https://nomad-coe.eu/`