MM 54: Topical Session (Symposium MM): Big Data in Materials Science - Managing and exploiting the raw material of the 21st century

Big Data V

Time: Thursday 17:30–19:00 Location: H 0107

Topical Talk MM 54.1 *(29)* Thu 17:30 H 0107
Discovering Interpretable Patterns, Correlations, and Causality — •Jilles Vreeken — Max Planck Institute for Informatics, Saarbrücken, Germany — Saarland University, Saarbrücken, Germany

To gain non-trivial insight from data using machine learning, we need to be able to interpret what these results mean. This we can either do by staring long and hard at the highly complex and non-linear models that methods such as support vector machines or deep learning provide when we run them on our data. This most often ends in us throwing the towel, as these models are extremely difficult to understand. Alternatively, we can require the learning method to report in a language we can (much) (more) easily understand, instructing it to discover things beyond what we already know.

In this talk, I will give an introduction to this latter, interpretable approach. In particular, I will explain the power of local modeling, that of non-parametric correlation discovery, that of pattern languages, will give examples of recent discoveries we made on materials science data using a technique called subgroup discovery, and an outlook on very recent approach to discover causal dependencies in data without having to make (almost) any assumptions.

MM 54.2 *(4)* Thu 18:00 H 0107
Subgroup Discovery for Finding Local Patterns in Materials Data — •Mario Boley[1], Bryan R. Goldsmith[2], Christopher Sutton[3], Jilles Vreeken[1], Matthias Scheffler[3], and Luca M. Ghiringhelli[3] — [1]Max-Planck-Institut für Informatik, Saarbrücken — [2]University of Michigan, Ann Arbor — [3]Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin

We establish that subgroup discovery (SGD), a form of local pattern discovery for labeled data, can help find interpretable descriptors from materials-science data obtained by first-principles calculations. In contrast to global modelling algorithms, SGD finds descriptions of subpopulations in which, locally, the target property takes on an interesting distribution. First, we formulate the SGD algorithm for applications in scientific domains. Subsequently, SGD is applied to configurations of neutral gas-phase gold clusters to discern general and interesting patterns between their geometrical and physicochemical properties. For example, SGD uncovers that van der Waals interactions within gold clusters are linearly correlated with their radius of gyration and are weaker for planar clusters than for nonplanar clusters. Moreover, we explore SGD for finding descriptors that predict both the formation and bandgap energies of transparent conducting oxides as well as descriptors that classify the octet binary semiconductors as either rock salt or zincblende; in both settings using only information of their chemical composition. Lastly, an efficient optimal solver using branch-and-bound is developed for dispersion-corrected objective functions to facilitate the discovery of interpretable subgroups.

MM 54.3 *(168)* Thu 18:15 H 0107
Generation of ab initio datasets with predefined precision using uncertainty quantification — •Jan Janssen, Tilmann Hickel, and Jörg Neugebauer — Max-Planck-Institut für Eisenforschung GmbH, Düsseldorf, Germany

A major challenge in multiscale materials simulation is the ab initio prediction of phase stabilities in multi-phase materials. To extend the ab initio accuracy to larger length and time scales the fitting of machine learning potentials seems promising, but this approach is intrinsically limited to the accuracy of the input data. Therefore it is essential to quantify the different sources of uncertainty in ab initio calculation, including the systematical error of convergence, the statistical or numerical error and the model error for derived quantities. Already the determination of the equilibrium lattice constant and bulk modulus requires a careful analysis of the fitting of energy-volume curves, going beyond the consideration of standard convergence parameters like cutoff and k-points. In order to handle this delicate interplay of uncertainties, we introduce the concept of uncertainty phase diagrams. Based on the uncertainty phase diagrams we model the convergence gradients of the contributing errors, to automate the convergence process not only for the error in energy. The modelling of uncertainties in relation to the corresponding ab initio calculation is enabled by our recently developed Python based workbench pyiron. Our investigations revealed that commonly used rules of thumb for fitting ground state materials properties become invalid for high precision calculations.

MM 54.4 *(339)* Thu 18:30 H 0107
Numerical-Error Estimates for DFT Calculations and Materials Databases — C. Carbogno[1], K.S. Thygesen[2], B. Bieniek[1], C. Draxl[1,3], L. Ghiringhelli[1], A. Gulans[3], O.T. Hofmann[4], K.W. Jacobsen[2], •S. Lubeck[3], J.J. Mortensen[2], M. Strange[2], E. Wruss[4], and M. Scheffler[1] — [1]FHI Berlin, Germany — [2]DTU, Lyngby, Denmark — [3]HU Berlin, Germany — [4]TU Graz, Austria

Density-functional theory (DFT) has become an invaluable tool in materials science. Whereas the precision of different approaches has been scrutinized for the PBE functional using extremely accurate numerical settings [1], little is yet known about code- and method-specific errors that arise under more commonly used numerical settings. Recently, this has become a severe issue, since it prevents repurposing publicly available DFT data created using different settings and/or codes. To overcome this, we study the convergence of different properties (geometries, total and relative energies) in four conceptually-different DFT codes (exciting, FHI-aims, GPAW, VASP) for typical settings. Specifically, we discuss relative and absolute errors as a function of the numerical settings, e.g., basis sets and **k**-grids, for 71 elemental solids [1]. Using this data, we propose analytical models that allow for reliable error estimates for *any* compound, as we explicitly demonstrate for binary and ternary solids. We discuss the extensibility of our approach towards more complex materials properties and its applicability in computational materials databases.
[1] K. Lejaeghere *et al.*, *Science* **351**, aad3000 (2016).

MM 54.5 *(280)* Thu 18:45 H 0107
An Electronic Transport Properties Database From High-Throughput Ab-initio Computations — •Francesco Ricci[1], Wei Chen[2,4], Umut Aydemir[3,5], Jeffrey Snyder[3], Gian-Marco Rignanese[1], Anubhav Jain[2], and Geoffroy Hautier[1] — [1]Institute of Condensed Matter and Nanosciences (IMCN), Université catholique de Louvain, Louvain-la-Neuve, Belgium — [2]Lawrence Berkeley National Lab, Berkeley, CA, USA — [3]Department of Materials Science and Engineering, Northwestern University, Evanston, USA — [4]Department of Mechanical, Materials and Aerospace Engineering, Illinois Institute of Technology, Chicago, USE — [5]Koç University, Department of Chemistry, Turkey

Nowadays the state-of-the-art DFT codes and the high-throughput (HT) frameworks allow us to compute materials properties at a large scale. As recently made by Materials Project (MP) for elastic and piezoelectric tensors, we will present a large and freely accessible data set of transport properties as effective mass and Seebeck coefficient. This transport data has been computed on top of energy band structures available in MP, using the well-known BoltzTraP code inserted in a HT framework. Given the importance of electronic transport properties, the whole community of material science researcher will benefit from this database. We will present the work flow to obtain the data and the data set. Some correlations between the transport properties and some applications in the field of transparent conducting oxides and thermoelectric materials will be presented.