MM 33: Topical session (Symposium MM): Big Data Analytics in Materials Science

Sessions: Big Data Analytics in Materials Science I and II

Time: Thursday 10:15–13:15 Location: H43

Topical Talk MM 33.1 Thu 10:15 H43

Supervised and unsupervised learning from the large body of materials literature — •Gerbrand Ceder — University of California, Berkeley, CA, USA

The overwhelming majority of scientific knowledge is stored as unstructured text in millions of publications. I will show some results showing how knowledge can be extracted from such a large corpus of text using a combination of Natural Language Processes (NLP) combined with Machine Learning (ML). NLP is necessary to turn the unstructured information in the scientific literature into structured data on which ML can operate. As an example, I will demonstrate how a vector representation of words can capture inorganic materials science concepts from 3.3 million scientific abstracts without human labelling or supervision. Remarkably, such basic text-based methodology can be used to make predictions of new materials, the properties of which we verify with Density Functional Theory. An alternative and more complex example will be discussed whereby all materials synthesis information is extracted from several million papers. Constructing synthesis recipes from papers requires extremely high precision and recall of relevant chemicals and operational procedures. I will show how this can be achieved by combining various supervised and unsupervised machine learning methods to create the largest data set of solid-state synthesis reactions.

MM 33.2 Thu 10:45 H43

Reproducible massive calculations and data sharing with AiiDA and the Materials Cloud — •Giovanni Pizzi[1], Leopold Talirz[1], Snehal Kumbhar[1], Aliaksandr Yakutovich[1], Elsa Passaro[1], Marco Borelli[1], Sebastiaan P. Huber[1], Martin Uhrin[1], Spyros Zoupanos[1], Fernando Gargiulo[1], Ole Schuett[2], Joost VandeVondele[3], Thomas C. Schulthess[3], Berend Smit[1], and Nicola Marzari[1] — [1]NCCR MARVEL and EPFL, CH — [2]Empa, Switzerland — [3]CSCS and ETHZ, CH

We discuss the challenges and solutions to store data resulting from the modern, complex workflows of computational science, allowing for the search and dissemination of results according to the FAIR principles of sharing. We first show how a materials' informatics framework like AiiDA [1] allows to automate all calculations and store their entire provenance. By uploading all data to the Materials Cloud (materialscloud.org), results can be disseminated seamlessly, DOIs are assigned to the datasets, and interactive (online or local) browsing of the provenance makes it possible to explore any element of the workflow guaranteeing its full reproducibility and enabling reuse of the results. Materials Cloud also provides intuitive web-based simulation services based on AiiDA, reducing the access barrier to HPC simulation tools.

[1] G. Pizzi et al., Comp. Mat. Sci. 111, 218 (2016), www.aiida.net

MM 33.3 Thu 11:00 H43

The NOMAD 2018 Kaggle Competition: Tackling Materials-Science Challenges through Crowd Sourcing — •Christopher Sutton[1], Luca M. Ghiringhelli[1], Takenori Yamamoto[2], Xiangyue Liu[1], Angelo Ziletti[1], and Matthias Scheffler[1] — [1]Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin, Germany — [2]Institute for Mathematical and Computational Sciences, LLC Yokohama, Japan

Machine learning (ML) promises to accelerate the discovery of novel materials by screening candidate compounds at significantly lower computational cost than traditional electronic-structure approaches. However, it is often *a priori* unclear which ML models are suitable for a given problem and optimizing a model can be a time-consuming endeavor. Crowd sourcing allows for comparing several ML models by identifying a key problem and challenging the community to solve it. To this end, the Novel Materials Discovery (NOMAD) Centre of Excellence together with Kaggle - one of the most well known hosting platforms - organized an open data-science competition to predict two key properties of transparent conducting oxides (TCOs): band gap energy (for transparency) and formation energy (for stability). Although these materials are crucial for optoelectronic devices, only a small number TCOs are currently known. In this contribution, we present the winning model out of nearly 900 participants based on a novel crystal-graph representation and an analysis of the relative importance of representation vs regression model for the performance of several ML approaches.

MM 33.4 Thu 11:15 H43

Symmetry-invariant basis representations for machine- learning of electronic structure data beyond energies — •Michael Luya[1,2] and Reinhard J. Maurer[2] — [1]Department of Mathematics, University of Warwick, Coventry, United Kingdom — [2]Department of Chemistry, University of Warwick, Coventry, United Kingdom

Recent successes on the high-dimensional machine-learning-based (ML) interpolation of total energies and forces from ab-initio computations are extremely encouraging for the future role that machine-learning can play in condensed matter simulation and electronic structure theory. Beyond scalar energy fields, ML can be useful to find efficient representations of quantum mechanical interaction integrals and Hamiltonians in atomic orbital basis representations. These represent tensor fields, which, contrary to scalar fields, feature covariance properties and additional directional coordinate dependence that need to be addressed.

Here we present an approach based on a generalisation of Slater-Koster transformation and symmetry-adaptation to transform interaction integrals and Hamiltonians from electronic structure theory in atomic-orbital representation into rotationally invariant forms that are amenable to established machine learning methods. We validate our approach on a large set of training data on simple organic molecules of varying size.

15 min. break

Topical Talk MM 33.5 Thu 11:45 H43

Extending high-throughput materials discovery to finite temperatures: Concepts and application — •Tilmann Hickel[1], Janssen Jan[1], Halil Sözen[1], Fritz Körmann[1], Sudarsan Surendralal[1], Mira Todorova[1], Yury Lysogorskiy[2], Ralf Drautz[2], and Jörg Neugebauer[1] — [1]Max-Planck-Institut für Eisenforschung GmbH, Max-Planck-Str. 1, 40237 Düsseldorf, Germany — [2]Atomistic Modelling and Simulation, ICAMS, Ruhr-Universität Bochum, D-44801 Bochum, Germany

Present ab initio based high-throughput methods are commonly restricted to T=0 K calculations. For many technologically relevant materials, however, properties and thermodynamic stability drastically change for finite temperatures. Recent developments allow us to calculate thermodynamic quantities up to the melting point, but require complex simulation protocols that couple computer codes from various disciplines together with advanced mathematical algorithms. To provide a platform to develop, implement, test and apply such protocols we have created a Python based integrated development environment called pyiron. After highlighting the underlying algorithmic concepts, we use the example of the hard-magnetic material system Ce-Fe-Ti to demonstrate the materials scientific consequences. Using high throughput screening we study how adding further elements impacts relative phase stabilities at finite temperatures and thus partitioning. This yields design criteria that extend the chemical composition space to quaternary, more stable hard magnetic materials.

MM 33.6 Thu 12:15 H43

Crystal-structure identification in polycrystals via Bayesian deep learning — •Angelo Ziletti, Andreas Leitherer, Matthias Scheffler, and Luca Ghiringhelli — Fritz Haber Institute of the Max Planck Society Faradayweg 4-6 14195 Berlin, Germany

Thanks to open-access online computational repositories (e.g. http://nomad-coe.eu) and experiments (e.g. atom probe tomography), researchers have now access to a vast amount of three-dimensional structural data. To extract valuable information for materials characterization and analytics, computational methods that automatically and efficiently detect long-range order are needed. Current methods are either not stable with respect to defects, or base their representation on local atomic neighbourhoods, which in turn makes it difficult to detect "average" longe-range order. In the proposed approach, for

a given crystal structure, we simulate its (three-dimensional) diffraction pattern, and by means of a spherical-harmonics expansion, we compute a rotationally and translationally invariant representation. A convolutional neural network is then used to identify the correct crystal structure; in particular, we use a Bayesian neural network in order to obtain statistically-principled classification probabilities and model uncertainty. This methodology is used to classify grains in polycrystals, find coherent regions in amorphous solids, but also detect crystallographic defects such as twin boundaries, stacking faults, and edge dislocations in heavily defected crystal structures.

MM 33.7 Thu 12:30 H43

Neural-network representation of materials for robust crystal-structure recognition — •Andreas Leitherer, Angelo Ziletti, Matthias Scheffler, and Luca M. Ghiringhelli — Fritz Haber Institute of the Max Planck Society, Berlin, Germany

Assigning the crystal structure to local regions of large atomic structures can reveal hidden patterns and thus interesting material properties. Available computational methods either support a large number of space groups but show critically limited robustness, or are very robust but can treat only a handful of classes. We use neural networks to robustly assign the correct crystal-structure type to a given material while being able to treat numerous space groups and chemical species. To capture information about the local chemical environments, we apply the smooth-overlap-of-atomic-positions (SOAP) descriptor, serving as input to the deep-learning model. Since the neural network provides an intrinsic similarity metric, we are able to investigate structural transitions such as the Bain path between face-centered cubic and body-centered cubic structures. We also discuss the application of our framework to detect precipitates in Ni-based superalloys (materials used in aircraft engines), whose structure is usually experimentally investigated via atom probe tomography. Finally, we show that the neural network automatically learns how to map crystal structures to a meaningful low-dimensional manifold, an ability which we exploit by building easily interpretable structural-similarity maps.

MM 33.8 Thu 12:45 H43

Artificial intelligence in materials science: towards optimal descriptors — •Benedikt Hoock[1,2], Santiago Rigamonti[1], Luca Ghiringhelli[2], Matthias Scheffler[1,2], and Claudia Draxl[1,2] — [1]Humboldt-Universität zu Berlin, Berlin, DE — [2]Fritz-Haber-Institut der MPG, Berlin, DE

Materials data contained in repositories like NOMAD [1] can be exploited in many useful ways, such as to better understand existing materials or to discover new materials with desired properties. A crucial step towards these goals is to find a set of meaningful descriptors, i.e. parameters based on computationally cheap input data that capture the physical mechanisms underlying certain material properties. In this work, we develop principles for constructing up to millions of candidate descriptors from simple physical properties. These principles involve mathematical operations [2] and different averaging procedures considering the local ordering. We compare two compressed sensing methods, LASSO+ℓ_0 [2] and SISSO [3], at identifying optimal descriptors out of all the candidates. Likewise, we introduce and compare cross-validation based model-selection strategies that use either the average training or the average test error as a criterion, aiming at increasing the descriptors' generalizability. We use two ab initio data sets, comprising group-IV zincblende ternaries and transparent conducting oxides, to test this methodological approaches.

[1]: C. Draxl & M. Scheffler, MRS Bulletin, 43, 676 (2018).
[2]: L. M. Ghiringhelli, et. al., Phys. Rev. Lett. 114, 105503 (2015).
[3]: R. Ouyang, et. al., Phys. Rev. Mater. 2, 083802 (2018).

MM 33.9 Thu 13:00 H43

Global sensitivity analysis and surrogate modeling for materials models with rapid local variations — Juan Lorenzi[2], Sandra Döpking[1], and •Sebastian Matera[1] — [1]Institut f. Mathematik, Freie Universität Berlin — [2]Lehrstuhl F. Theoretische Chemie, Technische Universität München

Most material models depend on a number of input parameters which carry some uncertainty. Quantifying the impact of these errors on the model output is the purpose of global uncertainty and sensitivity analysis. This requires some kind of sampling of the parameter space and surrogate modeling has become a popular tool to lift the problem of repetitive, computationally expensive model evaluations. Surrogate modeling becomes challenging when the underlying model shows locally rapid variations, e.g. if a materials model exhibits a phase transition within the parameter domain. We present a modification of the classical Shepard interpolation, which has been designed for such problems. This approach employs a local, node specific distance metric instead of a global metric and uses error estimates for the superposition of different local linear models at a query point. We demonstrate the approach on the global sensitivity analysis of a stochastic model for CO oxidation, which has been parametrized using Density Functional Theory. We find that we can obtain reasonably accurate estimates of the sensitivity indices already at a modest number of evaluations of the original high-fidelity model.