MM 53: Topical Session (Symposium MM): Big Data in Materials Science - Managing and exploiting the raw material of the 21st century

Big Data IV

Time: Thursday 15:45–17:15 Location: H 0107

Topical Talk MM 53.1 *(30)* Thu 15:45 H 0107

Big Data of Materials Science: Interpretability of Machine Learning — Luca M. Ghiringhelli[1], •Jan Vybiral[2], Sergey V. Levchenko[1], Claudia Draxl[3], and Matthias Scheffler[1] — [1]Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin-Dahlem, Germany — [2]Czech Technical University, Dept. of Mathematics FNSPE, Prague, Czech Republic — [3]Humboldt-Universität zu Berlin, Institut für Physik and IRIS Adlershof, Berlin, Germany

An important part of every machine learning approach to statistical learning is the representation of the input data. This representation usually assumes an implicit (and largely unchallenged) choice of right descriptors. To allow for human interpretability of learned structures in the data, their choice is crucial. Trustful prediction of new promising materials, identification of anomalies, and scientific advancement are doubtful if the connection between the descriptor and the actuating mechanisms is unclear.

We use the techniques of compressed sensing and feature selection to analyze this issue, define requirements for useful descriptors, and propose a practical algorithm for their identification. It selects suitable descriptors from a large set of physically meaningful quantities, which is created in a human-guided process.

For a classical example, the energy difference of zincblende/wurtzite and rocksalt semiconductors, we demonstrate how a meaningful descriptor can be found systematically.

MM 53.2 *(301)* Thu 16:15 H 0107

Cluster expansions with CELL: a novel python package with a focus on complex alloys — •Santiago Rigamonti[1], Maria Troppenz[1], Martin Kuban[1], Axel Huebner[1], Christopher Sutton[2], Luca Ghiringhelli[2], Matthias Scheffler[2], and Claudia Draxl[1,2] — [1]Humboldt-Universität zu Berlin — [2]Fritz-Haber Institut

The cluster expansion (CE) technique allows for obtaining compact models of configuration-dependent properties in alloys with *ab initio* accuracy, by expanding these properties in terms of clusters (sets of crystal sites). In this work, we present the code CELL [1] which is an object-oriented, modular and user-friendly python package for building accurate CE models. Its focus is on complex surfaces and bulk alloys possessing large parent cells (>30 atoms), for which a full structure enumeration is impossible. We note, though, that CELL is not limited to complex alloys. The creation of training data-sets is incorporated into a user-friendly setup of parent- and supercell objects. For selecting the optimal set of clusters, compressed-sensing schemes, such as LASSO and the split Bregman method, and various cross-validation strategies are available. In addition, finite-temperature properties and the characterization of phase transitions are achieved by applying the Wang-Landau method and diffusive nested sampling. We will demonstrate CELL's capabilities by selected examples.

[1] S. Rigamonti, M. Troppenz, M. Kuban, A. Huebner, C. Sutton, L. Ghiringhelli, M. Stournara, M. Scheffler, and C. Draxl. *CELL: python package for cluster expansions with large parent cells.* In preparation.

MM 53.3 *(174)* Thu 16:30 H 0107

***Ab-initio* study of the clathrate $Ba_8Ni_xGe_{46-x-y}\Box_y$: Stability, structure, and electronic properties** — •Martin Kuban, Santiago Rigamonti, Maria Troppenz, and Claudia Draxl — Humboldt-Universität zu Berlin

The type-I clathrate $Ba_8Ni_xGe_{46-x-y}\Box_y$ is a prototypical example of the phonon-glass–electron-crystal concept and thus promising for thermoelectric applications. Recent experimental studies [1] show that either *n*- or *p*-type conductivity can be achieved by tuning its composition around the charge-compensation ($x+y=4$). This is important in view of building enhanced *n*-*p* thermoelectric junctions with the same base material. Upon the addition of Ni, the stable phases tend to form less vacancies (□), while the lattice constants stay almost unchanged. In this work we perform an *ab-initio* cluster-expansion [2] study in the composition range $0 \leq x \leq 6$ and $0 \leq y \leq 4$. This allows us to investigate the configurational ground-states (GS) at zero temperature. For these GS-structures, we observed the transition of a minimum in the electronic density of states (DOS) from below the Fermi level for $x+y<4$, to above the Fermi level for $x+y>4$, thus confirming the experimentally observed change from *n*- to *p*-type conductivity. From the calculation of the phase diagram, we find good agreement with experiment regarding the lattice constant and the formation of vacancies for increasing Ni content.

[1] U. Aydemir *et al.*; Dalton Trans **44**, 7524 (2015).

[2] S. Rigamonti *et al.*, in preparation.

MM 53.4 *(107)* Thu 16:45 H 0107

SISSO: a Compressed-Sensing Method for Systematically Identifying Efficient Physical Models of Materials Properties — •Runhai Ouyang[1], Stefano Curtarolo[2], Emre Ahmetcik[1], Matthias Scheffler[1], and Luca M. Ghiringhelli[1] — [1]Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin-Dahlem, Germany — [2]Materials Science, Duke University, Durham, NC, USA

We present a systematic data-driven approach for discovering physically interpretable descriptors and predictive models, within the framework of compressed sensing. SISSO (sure independence screening and sparsifying operator) tackles immense and correlated features spaces, and converges to the optimal solution from a combination of features relevant to the materials' property of interest. The methodology is benchmarked with the quantitative prediction of the ground-state enthalpies of octet binary materials (using *ab initio* data) and applied to the showcase example of predicting the metal-insulator classification (with experimental data). Accurate predictive models are found in both cases. For the metal-insulator classification model, the interpretability and predictive capability are tested beyond the training data: It perfectly rediscovers the available pressure-induced insulator to metal transitions and it allows for the prediction of yet unknown transition candidates, ripe for experimental validation.

MM 53.5 *(110)* Thu 17:00 H 0107

From autonomous subspace selection of material properties to physically meaningful predictions — •Benjamin Regler, Matthias Scheffler, and Luca M. Ghiringhelli — Fritz Haber Institute of the Max Planck Society, Berlin, Germany

In data-driven materials science, the discovery of functional relationships is an inverse problem on finding subsets of materials properties (features) which relate to physical observables.

The inverse problem can be solved by statistical-learning algorithms that require manual adjustment and are thus based on ongoing experience and the knowledge being built. This makes ensuring the physical interpretability of the generated data-based models a demanding task.

Therefore, we highlight an autonomous systematic feature-subspace construction and selection method using information-theoretic concepts. We use a generalization of the Shannon entropy to select physically meaningful subsets (sharing the same notion of uncertainty as the Gibbs entropy in statistical thermodynamics) and use compressed sensing to find the best approximate models without prior knowledge.

Then, we apply the framework and highlight key problems such as stability predictions of zinc-blende vs. rock-salt octet binary semiconductors and band-gap predictions of topological insulators. Finally, we discuss the physical interpretation of the generated models and identify the strongest correlated materials properties with the actuating mechanism.