CPP 68: Topical Session: Data Driven Materials Science - Descriptors (joint session MM/CPP)

Time: Wednesday 11:45–13:15 Location: BAR 205

CPP 68.1 Wed 11:45 BAR 205

Evaluating representations of atomistic systems for machine learning — •Marcel Langer[1] and Matthias Rupp[1,2] — [1]Fritz Haber Institute of the Max Planck Society, Berlin, Germany — [2]Citrine Informatics, Redwood City, CA, USA

Interpolating between computationally expensive first-principles calculations with fast machine-learning surrogate models increases the feasible scope of exploration when a large space of potentially similar structures is sampled, for instance in the search for novel materials or the exploration of phase diagrams.

The choice of representation of the atomistic systems under consideration is important for the accuracy of such surrogate models. We present a rigorous empirical comparison of the Many-Body Tensor Representation [1], Smooth Overlaps of Atomic Positions [2], and Symmetry Functions [3] for energy predictions of molecules and materials. In this, we control for data distribution, hyper-parameter optimization, and regression method. We also investigate the relationship between predictive performance and computational cost, and discuss how to assess predictions beyond mean errors, which cannot fully describe model behaviour in practice. [4,5]

[1] H. Huo and M. Rupp, *arXiv*, 1704.06439 (2017)
[2] A. Bartók, R. Kondor., G. Csányi, *Phys. Rev. B* **87**, 184115 (2013)
[3] J. Behler, *J. Chem. Phys.* **134**, 074106 (2011)
[4] C. Sutton *et al.*, *ChemRxiv*, 9778670 (2019)
[5] Z. del Rosario *et al.*, *arXiv*, 1911.03224 (2019)

CPP 68.2 Wed 12:00 BAR 205

Information-theory-driven identification of compact descriptors for accurate machine-learning predictions — •Benjamin Regler, Matthias Scheffler, and Luca M. Ghiringhelli — Fritz Haber Institute of the Max Planck Society, Berlin, Germany

Machine learning (ML) is useful for predicting materials behavior by relating physical and chemical properties (features) of known materials to the property of interest (target). Aiming at a rational, unbiased, and data-driven identification of relevant features, we use a combination of statistical and information-theoretical techniques to identify the subset of features that unequivocally represent each material in the data set and contribute most to predicting the target property. The novelty and power of our approach is that it does not assume any specific functional form of the "features → target" relationship. Based on the concept of cumulative mutual information, our framework assigns quantitative scores for the "strength" of the feature's contributions, ranks the features by their scores, and selects the most contributing features to be relevant prior to ensuing data analysis. The scoring and selection algorithm is then supplemented by a purely ML procedure built on the selected and compact feature subset. We identify compact feature subsets for predicting (i) the ground-state crystal-structure of octet-binary compound semiconductors and (ii) elastic properties of inorganic crystalline compounds. In each case, we show that only a few features are actually required to obtain accurate predictions, thereby reducing the complexity of the ML model and sensitivity to the availability of materials data.

CPP 68.3 Wed 12:15 BAR 205

Size-Extensive Molecular Machine Learning with Global Descriptors — •Johannes Margraf[1], Hyunwook Jung[2], Sina Stocker[1], Christian Kunkel[1], and Karsten Reuter[1] — [1]Technical University Munich, Germany — [2]Yonsei University, South Korea

Machine learning (ML) models are increasingly used to predict molecular properties in a high-throughput setting at a much lower computational cost than conventional electronic structure calculations. Such ML models require descriptors that encode the molecular structure in a vector. These descriptors are generally designed to respect the symmetries and invariances of the target property. However, size-extensivity is usually not guaranteed for so-called global descriptors. In this contribution, we show how extensivity can be build into ML models with global descriptors such as the Many-Body Tensor Representation. Properties of extensive and non-extensive models for the atomization energy are systematically explored by training on small molecules and testing on small, medium and large molecules. Our results show that the non-extensive model is only useful in the size-range of its training set, whereas the extensive models provide reasonable predictions across large size differences. Remaining sources of error for the extensive models are discussed.

CPP 68.4 Wed 12:30 BAR 205

Hierarchical SISSO: predicting complex materials properties building on simpler ones — •Lucas Foppa[1], Sergey V. Levchenko[2,1], Matthias Scheffler[1], and Luca M. Ghiringhelli[1] — [1]Fritz-Haber-Institut der MPG, Berlin, DE — [2]Skolkovo Institute of Science and Technology, Moscow, RU

Symbolic regression is a promising tool to identify analytical models (descriptors) for predicting materials properties that are otherwise accessed via rather expensive *ab initio* calculations. In this context, the sure-independence screening and sparsifying operator (SISSO),[1] which combines the systematic generation of large feature spaces with compressed sensing, has been successfully applied, e.g., to the prediction of the (meta)stability of binary systems and perovskites from atomic properties only. However, if the relationship between the features and the target property is too complex, the descriptor search can become very inefficient. Here, we tackle this issue via a hierarchical approach: features that are easily computed (e.g., atomic properties) are used for predicting simple properties (e.g., lattice constant) and the resulting descriptors are in turn used as candidate features for modeling more complex properties (e.g., bulk modulus, position of band centers or band gaps). We demonstrate the hierarchical approach by analyzing a dataset of >700 cubic simple (ABO_3) and double ($A_2BB'O_6$) perovskites for predicting mechanical and electronic properties. The learned models require only atomic features as inputs and are therefore suitable for high-throughput screening of such materials.

[1] R. Ouyang, *et al.*, *Phys. Rev. Mater.* **2**, 083802 (2018).

CPP 68.5 Wed 12:45 BAR 205

Similarity descriptors for data-driven materials science — •Martin Kuban, Santiago Rigamonti, and Claudia Draxl — Humboldt-Universität zu Berlin

Learning from materials data is a topic of increasing importance in materials science. This task is supported by the availability of data through large online databases, like NOMAD [1]. For the application of artificial-intelligence (AI) methodology, materials must be characterized by a set of features that together build up *descriptors*. The success of AI tasks depends heavily on the quality of these descriptors, since they must contain all relevant information to map the input data onto the target property. Recent advances in the development of high-quality descriptors have allowed for both accurate predictions of material properties as well as highly interpretable models [2]. In this work, we develop a new type of descriptors based on the similarity of materials. To achieve this goal, we use both existing and newly developed descriptors to establish metrics that serve as quantitative similarity measures. These measures are combined into "similarity descriptors", which are then used for the construction of AI models. The performance of these models is optimized with respect to their predictive power. We demonstrate the applicability of our approach by predicting target properties for different classes of materials, including oxides and 2D systems.

[1] C. Draxl and M. Scheffler, MRS Bulletin, 43, 676, (2018).
[2] L. Ghiringhelli *et al.*, PRL, 114, 105503, (2015).

CPP 68.6 Wed 13:00 BAR 205

Machine-learning descriptors with domain knowledge of the interatomic bond — •Thomas Hammerschmidt, Jan Jenke, Aparna P.A. Subramanyam, Jörg Kossmann, Yury Lysogorskiy, and Ralf Drautz — ICAMS, Ruhr-Universität Bochum, Germany

The performance of machine-learning depends critically on the quality of the descriptors. In the case of learning atomic-scale properties, like formation energies obtained from density-functional theory (DFT) calculations, the descriptors typically measure the atomistic geometry and the distribution of chemical elements. Here, we construct descriptors that additionally include prior knowledge of the interatomic bond from a hierarchy of coarse-grained electronic-structure methods. In particular, we use tight-binding (TB) and analytic bond-order potentials (BOPs) that are derived from a second-order expansion of DFT. We demonstrate that a recursive solution of the TB problem and the closely related moments of the electronic density-of-states at the BOP level establish a smooth structure-energy relation. This first level of domain knowledge of the interatomic bond shows highly descriptive power in machine-learning applications already with simple, qualitative TB models. As second level of domain knowledge we include the bond chemistry in terms of bond-specific TB Hamiltonians that are obtained from downfolding the DFT eigenspectrum of molecular dimers. In the third level of domain knowledge we include the role of the valence electrons by determining non-selfconsistent bond energies with the bond-specific TB Hamiltonians.