MM 31: Data Driven Materials Science: Big Data and Work Flows – Machine Learning

Time: Wednesday 15:45–18:30 Location: SCH A 251

MM 31.1 Wed 15:45 SCH A 251

Neural networks trained on synthetically generated crystals can classify space groups of ICSD powder X-ray diffractograms — •Henrik Schopmans, Patrick Reiser, and Pascal Friederich — Institute of Theoretical Informatics, KIT, Karlsruhe, Germany

Machine learning techniques have successfully been used to extract structural information such as the crystal space group from powder X-ray diffraction (XRD) patterns. However, training directly on simulated patterns from databases like the ICSD is problematic due to its limited size, class-inhomogeneity, and bias toward certain structure types. We propose an alternative approach of generating random crystals with random coordinates by using the symmetry operations of each space group. Based on this approach, we present a high-performance distributed python framework to simultaneously generate structures, simulate patterns, and perform online learning. This allows training on millions of unique patterns per hour. For our chosen task of space group classification, we achieve a test accuracy of 77.4% on new ICSD structure types not included in the statistics dataset guiding the random generation. Instead of space group classification, the developed framework can also be used for other common tasks, such as augmentation and mixing of patterns for phase fraction determination. Our results demonstrate, using the domain of X-ray diffraction, how state-of-the-art models trained on large, fully synthetic datasets can be used to guide the analysis of physical experiments.

MM 31.2 Wed 16:00 SCH A 251

Critical Assessment of Uncertainty Estimates of Machine-Learning Potentials — •Shuaihua Lu[1,2], Luca M. Ghiringhelli[1], Christian Carbogno[1], and Matthias Scheffler[1] — [1]Novel Ma- terials Discovery at the FHI of the Max-Planck-Gesellschaft and IRIS- Adlershof of the Humboldt-Universität zu Berlin, Berlin, Germany — [2]School of Physics, Southeast University, Nanjing, China

Machine-learning potentials (MLP) trained on first-principles datasets are becoming increasingly popular since they enable the treatment of larger system sizes and longer time scales compared to straight ab initio techniques. A key aspect for the use of these MLPs is to reliably assess the accuracy viz. uncertainty of the predictions, e.g., by training an ensemble of models. Here, we critically examine the robustness of such uncertainty predictions using equivariant message-passing neural networks as an example [1]. We train an ensemble of models on liquid silicon simulated at the gradient-corrected density-functional-theory level and compare the predicted uncertainties with actual errors for various test sets, including liquid silicon at different temperatures and out-of-training-domain data such as solid phases with and without point defects as well as surfaces. These studies reveal that the predicted uncertainties are often overconfident. This is ascribed to the insufficient diversity in the members of the ensemble, as measured via error correlations. [1] S. Batzner et al., Nat. commun. 13, 2453 (2022).

MM 31.3 Wed 16:15 SCH A 251

Learning to Spell Materials - Coordinate-free Discovery with Natural Language Processing — •Konstantin Jakob, Karsten Reuter, and Johannes T. Margraf — Fritz Haber Institute, Berlin, Germany

Over the last decade, computational screening with structure-based machine learning models has led to some advances in the discovery of novel inorganic materials. Unfortunately, the overwhelmingly large space of possible compositions and atomic configurations together with the exceeding rarity of well-suited candidates ultimately poses a limit to the applicability of this approach. In contrast, purely composition-based representations neglect differences in the chemical properties of different crystal polymorphs and thus lack accuracy. A middle ground between full structural and simple compositional representations has been established for organic molecules using string representation such as SMILES. While these have proven highly advantageous for molecular discovery when combined with natural language processing models, analogous representations for the more complex class of inorganic materials are still missing. Bridging this gap, we investigate the performance of recurrent neural networks (RNNs) in predicting crystallographic properties by reading a materials composition element by element. Their striking accuracy suggests that symmetry- or prototype-based string representations could be generated with little computational effort at a large scale. The invertibility of these intermediate representations via restricted structure searches is investigated, paving the way to their application for conditional generative models.

MM 31.4 Wed 16:30 SCH A 251

Exploring materials dataspaces by combining supervised and unsupervised machine learning — •Andreas Leitherer[1], Angelo Ziletti[1], Christian H. Liebscher[2], Timofey Frolov[3], and Luca M. Ghiringhelli[1,4] — [1]NOMAD Laboratory at the FHI of the Max-Planck-Gesellschaft and IRIS-Adlershof of the Humboldt-Universität (HU) zu Berlin — [2]Max-Planck-Institut für Eisenforschung — [3]Lawrence Livermore National Laboratory — [4]Physics Department and IRIS-Adlershof of HU zu Berlin

To enable meaningful applications of AI to materials science, much of current efforts is concentrated on the creation of characterized datasets. In this talk, we discuss a rarely addressed topic - the development of automatic tools to explore available materials-science data. In particular, we go beyond purely supervised learning by combining unsupervised analysis with a recently developed crystal-structure recognition method [1]. This neural-network (NN) model automatically learns data representations that contain information on structurally diverse geometries. Using clustering, physically meaningful subgroups can be identified in the NN latent space, which are shown, e.g., to correspond to distinct, experimentally verified grain-boundary phases [2]. Moreover, dimension-reduction analysis allows us to create low-dimensional, interpretable materials charts that visualize complex structural data from both theoretical and experimental origin.

[1] A. Leitherer, A. Ziletti and L. M. Ghiringhelli. Nat. Commun. 12, 6234 (2021)

[2] T. Meiners et al. Nature 579, 375-378 (2020)

MM 31.5 Wed 16:45 SCH A 251

Minimizing data requirements by transfer learning for structure search on organic/inorganic interfaces — •Elias Fösleitner, Johannes Cartus, Lukas Hörmann, and Oliver T. Hofmann — Institute of Solid State Physics, Graz University of Technology, Graz, Austria

Performing structure search of organic molecules on metallic surfaces requires finding the structure with the lowest energy. However, calculating energies using conventional DFT codes proves to be a time-consuming task since single calculations are expensive and the number of configurations is large. To avoid the calculation of all possible structures, machine learning techniques such as Gaussian process regression have shown to be a useful tool in order to reduce the amount of DFT data needed. In our work we try to further reduce the amount of necessary data by using transfer learning from one substrate to another. To do this we include DFT data from structures of more than one substrate in our training set. In order to calculate the similarities between structures on different substrates we use the SOAP descriptor combined with an alchemical kernel which provides couplings between the different substrate elements. By optimizing these couplings, although molecule-substrate interactions differ notably (e.g. the interfacial charge transfer) between different substrates, we can save up to 50 % of the training data for one substrate A by also using the data of another substrate B. This serves as a stepping stone for the investigation of structures on computationally costly substrates.

15 min. break

MM 31.6 Wed 17:15 SCH A 251

Accelerating the Search for High-Performance, Novel Materials with Active Learning — •Thomas A. R. Purcell[1], Matthias Scheffler[1,2], Luca M. Ghiringhelli[1,2], and Christian Carbogno[1] — [1]The NOMAD Laboratory at the FHI of the Max-Planck-Gesellschaft and IRIS-Adlershof of the Humboldt-Universität zu Berlin — [2]Physics Department and IRIS-Adlershof at Humboldt Universität zu Berlin, Berlin, Germany.

Active-learning frameworks have the potential to greatly accelerate the search for new materials. By balancing exploitation and exploration, these approaches can efficiently search through materials space and

find the regions that are most likely to contain promising candidate materials [1]. Here we present an active learning framework, that uses an ensemble of expressions found by the sure-independence screening and sparsifying operator (SISSO) approach [2,3], and we demonstrate it for the example of discovering new thermal insulators. We statistically process the predictions of independent SISSO models to automatically select the most promising material candidates and then calculate their thermal conductivity κ_{L} using the *ab initio* Green Kubo method [4]. Using this approach we are able to find multiple new thermal insulators and gain insights into what is driving down their κ_{L}.

[1] A. G. Kusne, *et al.* Nat. Comm. **11**, 5966 (2020)
[2] R. Ouyang, *et al. Phys. Rev. Mater.* **2**, 083802 (2018)
[3] T. A. R. Purcell, *et al. J. Open Source. Softw.* **7**, 3960 (2022)
[4] F. Knoop, *et al.* arXiv:2209.12720

MM 31.7 Wed 17:30 SCH A 251

Machine learning discovery of new materials — •Jonathan Schmidt[1,2], Hai-Chen Wang[2], Noah Hoffman[2], Tiago Cerqueira[3], Pedro Borlido[3], Pedro Carrico[3], Love Pettersson[4], Claudio Verdozzi[4], Silvana Botti[1], and Miguel Marques[2] — [1]Friedrich-Schiller-University Jena, Germany — [2]Martin-Luther-University Halle-Wittenberg, Germany — [3]University of Coimbra, Portugal — [4]Lund University, Sweden

Graph neural networks for crystal structures typically use the atomic species and atomic positions as input. We construct crystal-graph attention networks replacing these precise bond distances with embeddings of graph distances. This allow us to perform high-throughput studies based on both compositions and crystal structure prototypes. Combining a newly curated dataset of 3M materials and the networks we have already scanned more than two thousand prototypes spanning a space of more than 5 billion materials and identified tens of thousands of theoretically stable compounds. We also demonstrate the effectiveness of transfer learning to adapt the networks to new domains such as of two dimensional structures.

Schmidt et al. Crystal graph attention networks for the prediction of stable materials, Sci. Adv. 7.49 (2021)

Schmidt et al., Large-scale machine-learning-assisted exploration of the whole materials space, arXiv:2210.00579 (2022)

Wang et al., Symmetry-based computational search for novel binary and ternary 2D materials, submitted (2022)

MM 31.8 Wed 17:45 SCH A 251

Data-driven magneto-elastic interatomic potentials for discovering novel phases of transition metal alloys — •Mani Lokamani[1], Kushal Ramakrishna[4], Julian Tranchida[3], Svetoslav Nikolov[2], Hossein Tahmasbi[4], Michael Wood[2], and Attila Cangi[4] — [1]HZDR Dresden, Germany — [2]SNL New Mexico, USA — [3]CEA Cadarache, France — [4]CASUS Görlitz, Germany

Structural prediction methods are used for identifying stable and metastable structures in a broad spectrum of materials. The presence of the electron spin degree of freedom in magnetic materials increases the complexity of finding such structures, constraining the analysis to the thermodynamically most relevant structures in a narrow range of temperatures and pressures. We achieve a search over much wider temperature and pressure conditions by utilizing machine-learning interatomic potentials based on the spectral neighbor analysis method within the coupled spin-molecular dynamics framework implemented in LAMMPS. This data-driven methodology enables predicting the properties of magnetic materials on much larger spatial, spin, and temporal domains and is parametrized by first-principles data. Leveraging this methodology, we predict the formation of metastable crystalline structures in transition metal alloys (FeNi, FeMn, FeCr, FeCo, FeGd) at high temperature-pressure conditions and assess their magnetic properties. This enables studying long-range spin structures in novel phases of transition metal alloys and complements the quest for permanent magnets for renewable energy applications that do not depend on rare-earth elements.

MM 31.9 Wed 18:00 SCH A 251

FAIR Modelling Recipes for High-Throughput Screening of Metal Hydrides — •Kai Sellschopp[1], Philipp Zschumme[2], Michael Selzer[2], Claudio Pistidda[1], and Paul Jerabek[1] — [1]Institute for Hydrogen Technology, Helmholtz-Zentrum Hereon, Geesthacht, Germany — [2]IAM - Microstructure Modelling and Simulation, Karlsruhe Institute of Technology, Karlsruhe, Germany

A simple modelling recipe for calculating the hydrogenation enthalpy of metal hydrides with ab-initio methods is presented. It consists of the everyday tasks of a computational materials scientist: relaxing a structure, optimising its volume, and calculating vibrational energies. The corresponding workflow is implemented in the framework of KadiStudio[1], where the scientific process is broken down into simple input-processing-output (IPO) tasks. The approach allows to track the inputs and outputs, to easily re-use the modular tasks in other workflows, and to share the workflow as a simple bash or python script. Therefore, not only the generated research data, but also the workflow itself fully comply with the FAIR data principles. As a first application, the recipe is employed to test how the different ingredients of an ab-initio calculation (e.g. xc-functional) affect the accuracy of predicting hydrogenation enthalpies. This helps to make better choices for studying this class of materials in the future and to judge the uncertainty in existing data. Furthermore, the standardized workflow enables a reliable high-throughput screening of new candidate materials for high-density hydrogen storage at near ambient conditions.

[1] L. Griem, *et al.*, Data Science Journal **21**, 16 (2022)

MM 31.10 Wed 18:15 SCH A 251

Take Two: Δ-Machine Learning for Molecular Co-Crystals — •Simon Wengert[1], Gábor Csányi[2], Karsten Reuter[1], and Johannes Theo Margraf[1] — [1]Fritz-Haber Institute, Berlin, Germany — [2]University of Cambridge, Cambridge, United Kingdom

Co-crystals are a highly interesting material class, as varying their components and stoichiometry in principle allows tuning supramolecular assemblies towards desired physical properties. The *in silico* prediction of co-crystal structures represents a daunting task, however, as they span a vast search space and usually feature large unit-cells. This requires theoretical models that are accurate and fast to evaluate, a combination that can in principle be accomplished by modern machine-learned (ML) potentials trained on first-principles data. Crucially, these ML potentials need to account for the description of long-range interactions, which are essential for the stability and structure of molecular crystals. In this contribution, we present a strategy for developing Δ-ML potentials for co-crystals, which use a physical baseline model to describe long-range interactions. The applicability of this approach is demonstrated for co-crystals of variable composition consisting of an active pharmaceutical ingredient and various co-formers. We find that the Δ-ML approach offers a strong and consistent improvement over the density-functional tight binding baseline. Importantly, this even holds true when extrapolating beyond the scope of the training set as demonstrated via molecular dynamics simulations at ambient conditions.