MM 11: Data Driven Material Science: Big Data and Workflows II

Time: Monday 15:45–18:00 Location: C 243

MM 11.1 Mon 15:45 C 243

Leveraging Multi-Fidelity Data In AI-Driven Sequential Learning of Materials Properties: Identifying Stable Water-Splitting Catalysts — •Akhil S. Nair, Lucas Foppa, and Matthias Scheffler — The NOMAD Laboratory at the FHI of the Max-Planck-Gesellschaft and IRIS-Adlershof of the Humboldt-Universität zu Berlin, Germany

The sequential learning of materials properties can enable a cost-effective materials discovery by iteratively extending the training data guided by an AI model [1]. Such an approach balances the exploitation of the model and the exploration of unvisited regions of the materials space. However, the efficiency of sequential learning relies on the performance of the AI model and on the quality of the data used to train the models. In material science, high-quality data is typically scarce. To address this challenge, we develop a sequential learning framework which utilizes low-fidelity data to improve the performance of the AI models for high-fidelity materials properties. In particular, we employ the symbolic regression based sure-independence screening and sparsifying operator (SISSO) method, which is suitable for small data sets and can better capture the behaviour of unseen materials compared to widely used AI methods [2]. Our approach is demonstrated for the discovery of stable oxide catalysts for water splitting, a process of significant importance in sustainable hydrogen production. For this, low and high-fidelity data are obtained from DFT-PBE and DFT-HSE calculations, respectively.

MM 11.2 Mon 16:00 C 243

From ab-initio to scattering experiments using neuroevolution potentials — •Eric Lindgren[1], Adam Jackson[2], Zheyong Fan[3], Christian Müller[4], Jan Swenson[1], Thomas Holm-Rod[5], and Paul Erhart[1] — [1]Department of Physics, Chalmers University of Technology, Gothenburg, Sweden — [2]Centre for Sustainable Chemical Technologies and Department of Chemistry, University of Bath, United Kingdom — [3]College of Physical Science and Technology, Bohai University, Jinzhou, People's Republic of China — [4]Department of Chemistry and Chemical Engineering, Chalmers University of Technology, Gothenburg, Sweden — [5]ESS Data Management and Software Center, Copenhagen, Denmark

Machine-learned interaction potentials have in recent years emerged as an appealing alternative to traditional methods for obtaining forces for molecular dynamics simulations, combining the computational efficiency of semi-empiricial potentials with the accuracy of ab-inito methods. In particular, Neuroevolution potential (NEP) models, as implemented in the GPUMD package, are highly accurate and computationally efficient, enabling large scale MD simulations with system sizes up to millions of atoms with ab-initio level accuracy. In this work, we present a workflow for constructing and sampling NEPs using the 'calorine' package, and how the resulting trajectories can be analysed with the 'dynasor' package to predict observables from scattering experiments. We focus on our recent work on crystalline benzene as an example system, but the approach is readily extendable to other systems.

MM 11.3 Mon 16:15 C 243

Multi-Objective Optimization of Subgroups for the Discovery of Exceptional Materials — •Lucas Foppa and Matthias Scheffler — The NOMAD Laboratory at the FHI of the MPG and IRIS-Adlershof of the HU Berlin, Germany

Artificial intelligence (AI) can accelerate the design of materials by identifying correlations and complex patterns in data. However, AI methods commonly attempt to describe the entire, practically infinite materials space with a single model, whereas different mechanisms typically govern the materials behaviors in different regions of materials space. The subgroup-discovery (SGD) approach identifies local rules describing exceptional subsets of data with respect to a given target of interest. Thus, SGD can focus on mechanisms leading to exceptional performance.[1] However, the identification of appropriate SG rules requires a careful consideration of the generality-exceptionality tradeoff. Here, we analyse the tradeoff between exceptionality and generality of rules based on a Pareto front of SGD solutions.[2]

[1] B.R. Goldsmith, *et al.*, *New. J. Phys.* **19**, 013031 (2017).
[2] L. Foppa and M. Scheffler, arXiv.2311.10381 (2023).

MM 11.4 Mon 16:30 C 243

From Prediction to Action: Critical Role of Performance Estimation for Machine-Learning-Driven Materials Discovery — •Lucas Foppa[1], Mario Boley[2], Felix Luong[2], Simon Teshuva[2], Daniel Schmidt[2], and Matthias Scheffler[2] — [1]The NOMAD Laboratory at the FHI of the MPG and IRIS-Adlershof of the HU Berlin, Germany — [2]Department of Data Science and AI, Monash University, Australia

The development of machine-learning models for materials properties focuses on improving the average predictive performance of the models with respect to some training-data distribution. However, a good performance in average might not translate into an efficient discovery of materials via model-driven blackbox optimization (e.g., Bayesian). In these iterative materials-discovery approaches, the training data is extended based on a model-informed acquisition function whose goal is to maximize a cumulative *reward* over iterations, such as the maximum property value discovered so far. Crucially, the rewards might be decoupled from the average predictive performance, as they can be dictated by the model performance for the few exceptional materials of interest. Here, we illustrate this problem for the example of bulk-modulus maximization in perovskites and propose an estimator that recovers qualitative aspects of the actual rewards and can be computed using the intial training data.[1]

[1] M. Boley, *et al.*, arXiv:2311.15549 (2023).

15 min. break

MM 11.5 Mon 17:00 C 243

A generic Bayesian Optimization framework for the inverse design of materials — •Zhiyuan Li, Yixuan Zhang, and Hongbin Zhang — Institute of Materials Science, TU Darmstadt, 64287 Darmstadt Germany

The traditional approach to develop materials relies on the time- and resource-costly trial-and-error experiments, as well as phenomenological theory with limited predictivity. Despite recent advances in high-throughput density functional theory calculations and statistical machine learning techniques, it is still a big challenge to efficiently explore a vast chemical space with a small number of initial samples to identify materials with optimized properties.

In this study, we propose and implement a comprehensive inverse design framework based on Bayesian optimization, integrating feature engineering, surrogate models, and acquisition functions, aiming to expedite the process of materials discovery. Focusing on the intrinsic physical properties such as formation energy, hardness, band gaps, and magnetization, it is demonstrated how such a framework can be applied to recommend optimal compositions in a vast chemical space exhibiting desired properties.

MM 11.6 Mon 17:15 C 243

Uncertainty quantification by shallow ensemble propagation — •Matthias Kellner and Michele Ceriotti — École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

Statistical learning algorithms provide a generally-applicable framework to sidestep time-consuming experiments, or accurate physics-based modeling, but they introduce a further source of error on top of the intrinsic limitations of the experimental or theoretical setup. One way to estimate this error is uncertainty estimation which make application of data-centric approaches more trustworthy. To ensure that uncertainty quantification is used widely, one should aim for algorithms that are reasonably accurate, but also easy to implement and apply. In particular, including uncertainty quantification on top of an existing model should be straightforward, and add minimal computational overhead. Furthermore, it should be easy to process the outputs of one or more machine-learning models, propagating uncertainty over further computational steps. We compare several well-established uncertainty quantification frameworks against these requirements, and propose a practical approach, which we dub shallow ensemble propagation, that provides a good compromise between ease of use and accuracy. We present applications to the field of atomistic machine learning for chemistry and materials, which provides striking examples of the importance of using a formulation that allows to propagate errors without making strong assumptions on the correlations between

different predictions of the model.

MM 11.7 Mon 17:30 C 243

Extracting physics with feature selection: How much data do we need and what can we really learn? — Guido Gaggl, Johannes C. Cartus, and •Oliver T. Hofmann — Institute of Solid State Physics, TU Graz

Feature selection algorithms such as SISSO allow a quick and automated analysis of data with the aim to find an equation that relates a target property of a system with properties of its constituent. Ideally, this equation coincides with the correct underlying physics. Unfortunately, this is often not the case, but even then, analyzing which constituent properties appear is often used to identify promising features. In this work, we analyze how well SISSO performs this task adverse circumstances. First, we demonstrate that given enough high-quality data and a sufficiently large feature space, it is indeed able to recover the correct physical equation. This is also surprisingly robust when reducing the number of available data points, even when including random or systematic bias into it. Conversely, adding even relatively small amount of noise to the data quickly deteriorates the performance. Finally, we discuss that in situations where two physical effects are superimposed, SISSO is intrinsically unable to find either, even when including multiple rungs.

MM 11.8 Mon 17:45 C 243

Adaptive-precision potentials for large-scale atomistic simulations — •David Immel[1,2], Ralf Drautz[1], and Godehard Sutmann[1,2] — [1]ICAMS, Ruhr-Universität Bochum, Bochum, Germany — [2]JSC, Forschungszentrum Jülich, Jülich, Germany

Large-scale atomistic simulations rely on interatomic potentials providing an efficient representation of atomic energies and forces. Modern machine learning (ML) potentials provide the most precise representation compared to electronic structure calculations while traditional potentials provide a less precise, but computationally much faster representation and thus allow simulations of larger systems.

We combine a traditional and a ML potential to a multi-resolution description, leading to an adaptive-precision potential with an optimum of performance and precision in large complex atomistic systems. The required precision is determined per atom by a local structure analysis and updated automatically during a simulation. We use Copper as demonstrator material with an embedded atom model (EAM) as traditional and an atomic cluster expansion (ACE) as ML potential, but any material and potential combination can be used for an adaptive-precision potential. The approach is developed for the molecular dynamics simulator LAMMPS and includes a load-balancer to prevent problems due to the atom dependent force-calculation times, which makes it suitable for large-scale atomistic simulations.

In this contribution strategies for the creation of an adaptive-precision potential are discussed. First results from Copper nanoindentations are reported and further improvements are outlined.