MM 13: Data-driven Materials Science: Big Data and Workflows I

Time: Tuesday 10:15–12:45 Location: SCH/A251

MM 13.1 Tue 10:15 SCH/A251

Surface reconstruction via automated LEED analysis based on Bayesian optimization — •Xiankang Tang and Hongbin Zhang — Institute of Materials Science, Technische Universitat Darmstadt, 64287 Darmstadt, Germany

Recent advances in machine learning have enabled the automation of many material characterization processes, which are essential for realizing autonomous experimentation for solid-state materials in the near future. Low-energy electron diffraction (LEED) is a fundamental technique in surface science, providing structural reconstruction information encoded in the energy-dependent intensity modulation of diffracted beams. However, the complexity of data analysis and the computational demands have limited the broader adoption of quantitative LEED in routine surface structure determination. In this work, we implement a Bayesian optimization-based approach to automatize the LEED I(V) analysis, where the best matching structures can be obtained for the experimental I(V) curves by minimize the R-factor between the experimental and simulated I(V) data. This approach can bee combined with density functional calculations or atomisitic simulations to further accelerate the recommendation of plausible structures by minimizing total energies.

MM 13.2 Tue 10:30 SCH/A251

Structural relaxations for nonstoichiometric alloys without forces — •Luca Numrich and Hongbin Zhang — Insitute of Materials Science, Technische Univerität Darmstadt, Darmstadt, Germany

Advances in machine learning and first-principles electronic-structure methods are accelerating materials design, focusing on stoichiometric compounds. Coherent potential approximation (CPA) is a standard approach for alloys with chemical disorder but there is no compatible implementation for forces, hindering the investigation of structural relaxations for nonstoichiometric alloys. For example, in many Ni-based Heusler alloys, a high-temperature disordered B2 phase transforms into the fully ordered $L2_1$ structure upon cooling via a second-order B2-$L2_1$ order-disorder transition.

In this work, we integrate Bayesian optimization (BO) with CPA as implemented in the exact muffin-tin orbitals (EMTO) code to identify low-energy crystal structures for nonstoichiometric alloys, as demonstrated for Ni-Mn-Ga-X Heusler alloys. Symmetry analysis is used to reduce the number of independent Wyckoff coordinates, which serve as parameters for BO, while the energy per atom computed via EMTO-CPA defines the objective function to be minimized. Using a Gaussian process surrogate model with a parallel upper confidence bound acquisition function, the framework autonomously proposes candidate structures with progressively lower energies. Thus, instead of a relaxation via interatomic forces, an acquisition function is guiding the relaxation.

MM 13.3 Tue 10:45 SCH/A251

A Python-based workflow for phase identification and mapping via Raman spectroscopy — •Felix Drechsler[1], Mahnaz Mehdizadehlima[2], Cameliu Himcinschi[1], David Rafaja[2], and Jens Kortus[1] — [1]TU Bergakademie Freiberg, Institute of Theoretical Physics, D-09599 Freiberg, Germany — [2]TU Bergakademie Freiberg, Institute of Materials Science, D-09599 Freiberg, Germany

Raman spectroscopy is a powerful tool for identifying phases and compounds. It is highly sensitive to both chemical and structural variations and is particularly attractive due to its minimal sample preparation requirements. This makes it well suited for fast phase identification and the determination of spatial phase distributions, complementing established techniques such as X-ray diffraction, element mapping via X-ray spectroscopy, and phase mapping using electron backscatter diffraction. However, phase-resolved analysis of Raman mappings remains challenging, as it involves processing high-dimensional data matrices and robust identification relies on comparing the complete spectrum with references.

In this talk, we present a Python-based workflow developed to address this challenge. The approach employs multivariate analysis methods and integrates mathematical similarity metrics with reference datasets to enable robust and reproducible phase identification. The proposed workflow provides an efficient way to construct Raman phase maps and offers a valuable tool for material characterization.

MM 13.4 Tue 11:00 SCH/A251

Automated Prediction of Phase Stability with ab-initio Accuracy — •Prabhath Chilakalapudi, Marvin Poul, Jan Janssen, and Jörg Neugebauer — Computational Materials Design, Max-Planck-Institute for Sustainable Materials, Düsseldorf

Developing sustainable metallic alloys-free of toxic elements and compatible with circular synthesis-requires novel and efficient ways to explore large composition spaces. A key bottleneck is the automated, ab-initio-accurate prediction of temperature-composition phase diagrams, where experimental phase data is limited or difficult to obtain.

We present a reproducible, automated workflow that uses Machine-Learned Interatomic Potentials (MLIPs) such as Atomic Cluster Expansion (ACE)[1] and non-equilibrium thermodynamic integration (Calphy[2]) to compute free energies and phase stabilities. By analysing key approximations including point-defect models, different entropic contributions, and free-energy interpolation schemes, we quantify the reliability of the calculated phase boundaries and provide meaningful "error bars" on the diagram. The workflow is demonstrated for representative binary alloys and is structured for gradual scaling to multicomponent systems. We leverage the pyiron[3] workflow framework for reproducible and efficient automation, to accelerate the discovery of sustainable materials.

[1] R. Drautz, Phys. Rev. B 100, 249901 (2019).
[2] S. Menon et al., npj Comput. Mater. 10, 261 (2024).
[3] J. Janssen et al., Comput. Mater. Sci. 163, 24-36 (2019).

15 min. break

MM 13.5 Tue 11:30 SCH/A251

Towards Disorder-Aware Materials Discovery - Recognizing and Modeling Crystallographic Disorder — •Konstantin S. Jakob[1], Aron Walsh[2], Karsten Reuter[1], and Johannes T. Margraf[1,3] — [1]Fritz-Haber-Institut der MPG, Berlin — [2]Imperial College London — [3]Universität Bayreuth

Recent computational materials discovery efforts have led to an enormous number of predictions of previously unknown, potentially stable inorganic, crystalline materials. However, these efforts are currently limited to predicting perfectly crystalline materials. As a consequence, many of these predictions cannot be verified in experiments, where kinetic effects, defects, and crystallographic disorder can be crucial. Here, we discuss disorder as a current frontier in materials discovery. To this end, we show that machine learning classification models can reliably recognize disordered materials and demonstrate that a significant fraction of computationally predicted materials are likely disordered [1]. On the example of compositionally complex transition metal ferrite spinels, we then demonstrate how machine learning interatomic potentials and Monte Carlo sampling can be used to tackle such disordered systems efficiently.

[1] K.S. Jakob, A. Walsh, K. Reuter, and J.T. Margraf, Adv. Mater. e14226 (2025).

MM 13.6 Tue 11:45 SCH/A251

Efficient Exploration of the Unknown: Distance-Based Active Learning with SISSO Descriptors and Mendeleev Similarities for Materials Discovery — •Sreejani Karmakar[1], Akhil S. Nair[1,2], Lucas Foppa[1], and Matthias Scheffler[1] — [1]Fritz Haber Institute of the Max Planck Society, Berlin, Germany — [2]Freie Universität Berlin, Berlin, Germany

The performance of AI models depends strongly on the distribution of their training data, which ideally should be independent and identically distributed. Materials-science datasets often violate this condition, containing redundancy and bias that hinder the discovery of statistically rare high-performance materials. Active learning (AL) helps by building concise, diverse training sets, introducing underrepresented materials classes. Commonly, AL relies on uncertainty estimates derived from the variance of model ensembles[1], but these are frequently overconfident and limit AL efficiency. We introduce an alternative strategy that selects candidate materials based on their distance from the existing training set in a low-dimensional descriptor space[2]. These descriptors, derived via SISSO (sure independence screening and sparsifying operator) symbolic regression approach. This distance-guided approach outperforms ensemble-based uncertainty AL, success-

fully identifying perovskites with exceptional properties. Adding the Mendeleev similarity metric further improves dataset diversity and supports efficient navigation of unexplored material space.

[1] A. Nair et al., npj Comput. Mater., 11, 150, 2025. [2] D. Wu et al., Inf. Sci., 474, 90, 105, 2019.

MM 13.7 Tue 12:00 SCH/A251

Where Are Large Language Models Actually Useful for Materials Design? — •Hedda Oschinski, Maximilian L. Ach, David Greten, Konstantin S. Jakob, Christian Carbogno, and Karsten Reuter — Fritz-Haber-Institut der MPG, Berlin

The rapid development of large language models (LLMs) and LLM-based agents has opened new possibilities for accelerating materials discovery and design. In this work, we explore their potential in the context of solar cell materials, a class of systems requiring complex, multi-property optimization across chemistry and materials science. By systematically evaluating a range of tasks for a well-known test set of Elpasolites - from context preparation and descriptor prioritization to design hypothesis generation and autonomous validation within an agent framework - we identify where current LLMs provide genuine utility and where critical limitations remain. Our findings offer a grounded perspective on how these tools can be integrated into materials discovery workflows, and what developments are needed to expand their impact in the future.

MM 13.8 Tue 12:15 SCH/A251

Predictive and interpretable machine learning models for thermodynamics tuning of metal hydrides for hydrogen storage — •Sinan S. Faouri[1], Kai Sellschopp[2,3], Claudio Pistidda[3], and Paul Jerabek[3] — [1]Mechanical and Industrial Engineering Department, Applied Science Private University, Amman, Jordan — [2]Department of Chemical and Process Engineering, University of Canterbury, Christchurch, New Zealand — [3]Institute of Hydrogen Technology, Helmholtz-Centre Hereon, Geesthacht, Germany

Metal hydrides remain among the most promising materials for solid-state hydrogen storage due to their tunable thermodynamic behavior. However, predicting key properties such as equilibrium pressure and hydrogenation enthalpy remains challenging, especially across diverse alloy systems. In this work, we explore feature-based machine learning strategies to model these thermodynamic quantities from elemental descriptors and derived structural features. The study combines experimental and computational data to identify the most relevant predictors governing hydrogen absorption thermodynamics. Particular attention is given to the relationship between atomic-scale size parameters, electronic features, and their collective influence on pressure-enthalpy correlations. The results demonstrate that data-driven approaches can reveal non-obvious structure-property relationships and guide the search for alloys with optimized storage performance. The presented framework offers a step toward integrating machine learning with physical insights for accelerated discovery of functional hydrides.

MM 13.9 Tue 12:30 SCH/A251

Score-based diffusion models for accurate crystal structure inpainting and reconstruction of hydrogen positions — •Timo Reents[1], Arianna Cantarella[2], Marnik Bercx[1], Pietro Bonfà[2,3], and Giovanni Pizzi[1] — [1]PSI Center for Scientific Computing, Theory and Data, CH-5232 Villigen PSI, Switzerland — [2]Department of Physics and Earth Sciences, University of Parma, IT-43124 Parma, Italy — [3]Dipartimento di Scienze Fisiche, Informatiche e Matematiche, University of Modena and Reggio Emilia, IT-41125 Modena, Italy

Generative AI methods are rapidly evolving to speed up and improve materials discovery. Diffusion based models can not only be adopted to generate new materials with desired properties but also to reconstruct crystal structures for which structural information is only partially available. In this work, we use Microsoft's mattergen [1], a diffusion based model originally designed to generate new stable crystal structures, and extend it to reconstruct missing hydrogen sites in crystal structures reported in experimental databases. This is particularly useful as the experimental measurement of hydrogen sites with standard XRD is typically challenging due to weak scattering of hydrogen. We show how to leverage image inpainting approaches known from computer vision, combined with universal machine learning interatomic potentials, to improve the success rate of correctly identifying the missing sites or finding lower energy configurations while significantly lowering the computational cost with respect to a direct DFT approach.

[1] Zeni, C. et al., Nature 639, 624-632 (2025)