MM 22: Data-driven Materials Science: Big Data and Workflows III

Time: Wednesday 10:15–12:45 Location: SCH/A216

MM 22.1 Wed 10:15 SCH/A216

Hashing It Out: Overcoming the Duplicate Structure Filtering Bottleneck for Large Data Sets — •Julian Holland, Juan Manuel Lombardi, Chiara Panosetti, and Karsten Reuter — Fritz Haber Institute, Berlin, Germany

With the increasingly data-rich landscape of computational chemistry research, new bottlenecks to material property elucidation have emerged stemming from data processing. Duplicate detection is often an essential data processing step for active learning, global optimization, and general PES exploration algorithms to ensure efficiency and functionality. Such duplication checks typically scale unfavorably with the number of structures, potentially taking longer to perform than the data generation. Hashing-based methods, which have decoupled scaling with dataset size, circumvent this but are conventionally too rigid to reliably find duplicates. In this talk, we present a democratic hashing duplicate detection algorithm that is flexible enough to detect duplicate structures with arbitrarily similar, but distinct, global descriptors nearly instantly. The uniqueness of the structure can be determined by an ensemble of hash functions associated with a set of randomly perturbed global descriptors. We compare the performance of our duplicate detection algorithm against conventional distance-matrix-based methods and introduce a standardized suite of duplicate detection benchmarks. Our algorithm is not only faster but often significantly more robust at detecting known duplicates.

MM 22.2 Wed 10:30 SCH/A216

MC3D: The Materials Cloud FAIR and full-provenance materials database — •Michail Minotakis — PSI Center for Scientific Computing, Theory and Data, 5232 Villigen PSI, Switzerland

Carefully curated databases of materials and their properties have become invaluable resources for a range of applications, from property prediction using machine learning techniques to materials discovery. Here, we introduce MC3D, the Materials Cloud three-dimensional database, in which more than 95% of the available materials are, to date, classified as experimentally known. This database is derived from structures sourced from three major databases: the Pauling File, the Inorganic Crystal Structure Database, and the Crystallography Open Database. After careful curation, the final collection of 72,609 unique stoichiometric compounds is refined using density-functional theory calculations at the PBEsol level, executed in Quantum ESPRESSO and leveraging the SIRIUS library for optimized GPU performance. The AiiDA materials informatics infrastructure (http://aiida.net) manages each workflow stage, ensuring full traceability and preserving simulation provenance. The results are freely accessible in the MC3D section of Materials Cloud (https://mc3d.materialscloud.org) and are already being used as a starting point for materials discovery projects, such as novel thermoelectrics, electrides, superconductors, or materials displaying a large nonlinear Hall effect.

MM 22.3 Wed 10:45 SCH/A216

Building a FAIR Community around Parsing — •Nathan Daelman[1], Alvin N. Ladines[1], Esma Boydas[1], Martin Kuban[1], Bernadette Mohr[1], Sascha Klawohn[1], Rubel Mozumber[1], Christina Ertural[2], Silvana Botti[3], Joseph F. Rudzinski[1], Lauri Himanen[1], and FAIRmat Team[1] — [1]Inst. für Physik, Humboldt-Universität zu Berlin — [2]Department of Materials Chemistry, Federal Institute for Materials Research and Testing, Berlin — [3]RC-FEMS and Faculty of Physics, Ruhr University Bochum

NOMAD [nomad-lab.eu][1, 2] is an open-source data infrastructure for materials science data. One of its most praised features is how NOMAD allows for direct ingestion of various software output formats. This gives data producers access with minimal effort to the whole toolkit infrastructure system regardless of their choice of simulation code. As the NOMAD community extends into related scientific disciplines, parsing procedures should grow alongside and empower casual users to contribute too. To this end, I will be presenting two new parsing frameworks: (i) Mapping Annotation which connects code-specific formats to the NOMAD interoperable schema, while gracefully handling syntatic concerns; (ii) an agentic LLM interface for hooking up third-party parsers via the Model Context Protocol (MCP). Finally, I will highlight how both approaches fit into NOMAD Plugins and NOMAD Actions.

[1] Scheidgen, M. *et al.*, JOSS **8**, 5388 (2023).
[2] Scheffler, M. *et al.*, Nature **604**, 635-642 (2022).

MM 22.4 Wed 11:00 SCH/A216

Uncertainty Propagation in Machine-learned Interatomic Potentials — •Haitham Gaafer, Jan Janssen, and Jörg Neugebauer — Computational Materials Design, Max-Planck-Institute for Sustainable Materials, Düsseldorf

Accurate multiscale materials modeling requires that uncertainties be quantified and propagated consistently from the electronic-structure level to macroscopic property predictions. Machine-learned interatomic potentials (MLIPs), trained on density-functional theory (DFT) reference data, now routinely reach near-DFT accuracy at dramatically reduced computational cost. Yet the connection between fitting errors in an MLIP and uncertainties in derived physical properties, such as bulk moduli or phase stabilities, remains insufficiently understood. We present a data-driven pyiron workflow designed to analyze how uncertainties originating in MLIP training propagate into thermomechanical property predictions. As a case study, we construct diverse DFT training sets for Cu, Ag, and Au using the Automated Small SYmmetric Structure Training (ASSYST) workflow, and fit computationally efficient atomic cluster expansion (ACE) potentials employing a minimal basis optimized to reach a target root-mean-square error. These potentials are subsequently used to determine equations of state and to quantify uncertainties in key properties, including the equilibrium lattice constant, bulk modulus, and its pressure derivative. Our results provide a transparent link between MLIP fitting quality and property reliability, offering a systematic route for uncertainty-aware atomistic modeling.

MM 22.5 Wed 11:15 SCH/A216

Accurately predicting thermal conductivity using non-equilibrium molecular dynamics simulations and machine-learned force fields — •Florian Unterkofler[1], Lukas Legenstein[1], Sandro Wieser[2], and Egbert Zojer[1] — [1]Graz University of Technology, Austria — [2]TU Wien, Austria

With the rise of machine-learned interatomic potentials, simulations have become an even more crucial tool for predicting material properties. We previously achieved accurate predictions of experimentally observed thermal conductivity of acenes, using system-specific, machine-learned Moment Tensor Potentials (MTPs) within a lattice dynamics approach.[1] To obtain a complementary real-space perspective, we now investigate whether comparable accuracy can be achieved using non-equilibrium molecular dynamics (NEMD).

Here, we present the workflow required to obtain accurate and reliable predictions when applying MTPs in NEMD simulations. We show that, due to the inherently stochastic nature of both MD and MTP training, a thorough statistical analysis of multiple simulations with different initial conditions and different realizations of the MTP is necessary. Furthermore, we highlight the importance of selecting appropriate training data to generate robust MTPs. When these considerations are taken into account, we achieve an excellent agreement between experiments, lattice-dynamics, and NEMD results, with NEMD simulations providing tools to investigate heat-transport bottlenecks in real space.

[1] L. Legenstein et al., *npj Comput Mater* 11, 29 (2025)

15 min. break

MM 22.6 Wed 11:45 SCH/A216

Data-efficient training of interatomic potentials using finite-temperature DFT structures — •Martin Schlipf[1], Sudarshan Vijay[1,2], and Georg Kresse[1,3] — [1]VASP Software GmbH, Berggasse 21/14, 1090 Vienna, Austria — [2]Department of Chemical Engineering, Indian Institute of Technology Bombay, Powai, Mumbai, Maharashtra 400076 India — [3]Faculty of Physics and Center for Computational Materials Science, University of Vienna, Kolingasse 14-16, A-1090 Vienna, Austria

We successfully generated a database of 150,000 unique finite-temperature structures using VASP and a "one-shot" DFT method to systematically sample atomic environments across the periodic table. Despite the small size of our training set compared to the millions typically used for foundation models, our resulting interatomic potentials achieve a force prediction error of 72 meV/Å. This performance is of the same magnitude as current state-of-the-art foundation models when tested against the same high-quality dataset. This result demonstrates that focusing on data quality and chemical diversity at finite temperatures is as impactful as massive data quantity. Furthermore, we showcase the computational infrastructure that made it possible to integrate interatomic potentials into an ab-initio software and discuss necessary enhancements to electronic optimization methods to compute magnetic materials more reliably.

MM 22.7 Wed 12:00 SCH/A216

MACE-based Machine Learning Interatomic Potentials for Iron-Nickel Alloys: Validation Across Composition and Pressure Ranges — •Kushal Ramakrishna[1], Mani Lokamani[1], and Attila Cangi[1,2] — [1]Helmholtz-Zentrum Dresden-Rossendorf (HZDR), D-01328 Dresden, Germany — [2]Center for Advanced Systems Understanding (CASUS), D-02826 Görlitz, Germany

Machine-learned interatomic potentials have emerged as powerful tools bridging quantum-level accuracy with mesoscale simulations in computational materials science. We present a comprehensive evaluation of MACE models for iron-nickel alloys across a wide range of compositions and pressures, with direct relevance to Earth's core modeling and industrial applications. We construct special quasirandom structures (SQS) to simulate random iron-nickel alloy configurations and train MACE models on density functional theory datasets combined with experimental validation data. Extensive short-range order analysis confirms improved chemical randomness for larger supercells, critical for faithful property sampling. Multiple MACE flavors are systematically compared against experimental measurements for structural and elastic properties in both body-centered cubic and face-centered cubic phases. Our results demonstrate that fine-tuned MACE models achieve remarkable predictive accuracy for equation-of-state behavior and elastic properties across all compositions. This approach successfully bridges computational predictions with experimental observations, enabling accelerated materials discovery for technologically relevant transition metal alloys.

MM 22.8 Wed 12:15 SCH/A216

Benchmarking the MACE Foundation Model for Solid-State Ion Conductors — •Takeru Miyagawa, Yufeng Xu, Levon Satzger, Waldemar Kaiser, and David A. Egger — Physics Department, TUM School of Natural Sciences, Technical University of Munich, 85748 Garching, Germany

Recent progress in foundation model machine learning potentials (MLPs) has demonstrated promising transferability and accuracy across diverse material classes [1, 2]. Instead of being trained from scratch for each new system, these large pretrained models aim to provide broadly accurate force and energy predictions that can be refined for new chemistries with comparatively small datasets. This offers a complementary route to traditional system-specific MLPs and may reduce the cost of studying complex ionic materials.

Here, we benchmark the MACE foundation model [2] on representative solid-state ion conductors (SSICs) through direct comparison with first-principles calculations. We assess its accuracy for phonons and vibrational properties, characterize temperature-driven structural and phase transitions, and analyze ion transport across different phases. We then explore data-efficient DFT-based fine-tuning strategies to improve the foundation model's accuracy for SSICs and clarify the limits and strengths of pretrained representations in the context of ionic transport. References [1] Batatia, I. et al., Adv. Neural Inf. Process. Syst. 35, 11423-11436, 2022, [2] Batatia, I. et al., J. Chem. Phys. 163, 184110, 2025

MM 22.9 Wed 12:30 SCH/A216

MACE-μ-α: A Foundation Model for Molecular Dipole Moments and Polarizabilities — •Nils Gönnheimer[1,2], Venkat Kapil[3], Karsten Reuter[2], and Johannes T. Margraf[1,2] — [1]Universität Bayreuth — [2]Fritz-Haber-Institut der MPG — [3]University College London

Machine-learning interatomic potentials (MLIPs) have had a strong impact on computational chemistry, physics, and materials science in recent years by filling the accuracy gap between first-principles methods and classical force fields, at a fraction of the computational cost of the former. MLIPs are so far typically limited to predicting energies and forces, however, while other properties traditionally obtained from first-principles calculations have remained less accessible. Here, equivariant neural network architectures have led to enormous progress, as they allow the prediction of vectorial and tensorial properties on the same footing as energies and forces.

Here, we present the MACE-μ-α architecture for predicting dielectric properties based on the MACE MLIP framework. Trained on over 1.6 million organic systems, the corresponding foundation model allows the accurate prediction of molecular dipole moments and polarizabilities, as well as Raman and IR spectra (when combined with an MLIP). Notably, despite being trained on gas-phase molecules and clusters, the model also shows transferability to condensed systems such as molecular crystals.