Sixth FAIRmat Users Meeting

July 2-3, 2025

Ruhr-Universität Bochum

Bochum, Germany

Overview Schedule

Wednesday, July 2	Thursday, July 3
9:00-10:30 **Harnessing Data-Driven Methods in Materials Discovery**	*9:00-10:30* **AI & ML Activities Across NFDI Consortia**
10:30-11:00 Coffee Break	*10:30-11:00* Coffee Break
11:00-12:00 **Exploring AI & ML in Simulations, Databases, and Computational Workflows**	*11:00-12:00* **Language Models for Materials Science**
12:00-13:00 Lunch	*12:00-13:00* Lunch
13:00-14:30 **Hands-on Workshop: Role-Based NOMAD Usage and Development**	*13:00-16:00* **Hands-on Workshop: Large Language Models for Scientific Data Extraction**
14:30-15:00 Coffee Break	
15:00-17:00 **Poster Session & Meet Our Experts**	

The detailed program can be found below and on the **webpage**.

Harnessing Data-Driven Methods in Materials Discovery

FAIR Research Data Management with FAIRmat and NOMAD

Hampus Näsström

In this talk, I will present the latest developments of FAIRmat and the NOMAD ecosystem, emphasizing the critical aspects of FAIR research data management (RDM) in condensed-matter physics. The NOMAD ecosystem, which includes platforms such as NOMAD, NOMAD Oasis, and NOMAD CAMELS, provides a complete solution for an RDM framework tailored to the needs of the condensed-matter physics and materials-science communities. I will demonstrate how these platforms facilitate the systematic acquisition and management of research data from diverse sources, including materials synthesis, advanced characterization techniques, and simulations. The presentation will also cover how NOMAD ensures that research data is not only securely stored, but also remains accessible and interoperable across different research domains. This in turn supports collaborative research efforts and accelerates innovation by ensuring that data is AI-ready.

Time: 9:00-9:30

Quantitative XPS Analysis Using Convolutional Neural Networks

Lukas Pielsticker

Lukas Pielsticker[1],[2],Rachel L. Nicholls[1], Serena DeBeer[1], Walid Hetaba[1], Florian Dobener[2], Laurenz Rettig[3], José A. Márquez[2], Sandor Brockhauser[2], Heiko Weber[4], Claudia Draxl[2], Mark Greiner[1]

[1] Max Planck Institute for Chemical Energy Conversion, Mülheim an der Ruhr, Germany

[2] Physics Department and CSMB, Humboldt-Universität zu Berlin, Germany

[3] Department of Physical Chemistry, Fritz Haber Institute of the Max Planck Society, Berlin, Germany

[4] Chair of Applied Physics, FAU Erlangen-Nürnberg, Erlangen, Germany

X-ray photoelectron spectroscopy (XPS) is a powerful tool for studying the electronic structure and chemical composition of solid surfaces. Quantitative analysis of XP spectra typically relies on manual curve fitting by expert spectroscopists. However, recent advancements in the ease of use and reliability of XPS instruments have led to a growing number of (novice) users generating large datasets that are becoming difficult to analyze manually. Additionally, the expansion of publicly available XPS databases further increases the volume of data requiring efficient analysis. Reflecting these developments, more automated techniques are desirable to assist users in processing large XPS datasets.

Here we present a scalable framework for automated XPS quantification using convolutional neural networks (CNNs). By training CNN models on artificially generated XP spectra with known quantifications (i.e., for each spectrum, the concentration of each chemical species is known), it is possible to obtain models for auto-quantification of transition metal XP spectra [1]. CNNs are shown to be capable of quantitatively determining the presence of metallic and oxide phases, achieving accuracy comparable to or exceeding that of conventional data analysis methods. The models are flexible enough to handle spectra containing multiple chemical elements and acquired under varying experimental conditions. The use of dropout variational inference for the determination of quantification uncertainty is discussed. Finally, we demonstrate how these network models are integrated into NOMAD [2], enabling real-time analysis of newly generated data.

References

[1] Pielsticker, L.; Nicholls, R.; DeBeer, S.; Greiner, M., Analytica Chimica Acta 2023 1271, 341433.

[2] Scheidgen, M., et al., Journal of Open Source Software 2023 8.90, 5388.

Time: 09:30-10:00

Small-Data Models for Materials Design

Luca Ghiringhelli

The modeling of macroscopic properties of materials often require to accurately evaluate physical quantities at several time and length scales. Here we show how symbolic inference, i.e., the machine learning of simple analytical expressions that explain and generalize the available data, can effectively bridge physical scales. The focus is on learning models that are as simple as possible (but not simpler), with as few as possible data points. I will demonstrate the application of the methods to the modeling of catalytic properties of materials, thermal conductivity, and more.

Time: 10:00-10:30

Exploring AI & ML in Simulations, Databases, and Computational Workflows

The Alexandria Database of Materials

Miguel Marques

The Alexandria database represents an important resource for materials science research, containing more than 5 million density-functional theory calculations for periodic three-, two-, and one-dimensional compounds. This open-access database addresses the critical challenge of data scarcity in materials science, where large, high-quality, and consistent datasets are rare. The Alexandria database serves as a foundation for training machine learning models in materials science, enabling the training of multiple material properties using both composition-based models, crystal-graph neural networks, or machine-learning interatomic potentials. This is also an important resource in accelerating materials discovery and in transforming high-throughput materials design. The Alexandria database exemplifies the power of open

science in materials research, providing the community with a robust platform for developing predictive models and discovering new materials with tailored properties.

Time: 11:00-11:30

Creating Data Analysis Pipelines Using the MADAS Framework

Martin Kuban

The NOMAD data infrastructure provides access to vast amounts of data that can be used for data analytics and machine learning (ML). Often, however, not all (meta)data are relevant for every task, making it necessary to apply filtering and processing steps to prepare input data for ML.

Here, we present MADAS, a Python framework that supports all steps of data analytics and machine learning, including automated download and storage of data, generation of material descriptors, and computing similarity metrics, and integrates well with established ML frameworks and libraries. MADAS allows to write robust, re-usable data analysis pipelines, while its modular structure allows to quickly extend the data processing with custom functions.

We demonstrate its capabilities and features by finding interoperable data within a large computational dataset hosted on NOMAD, and by finding distinct materials that exhibit similar electronic structures.

Time: 11:30-12:00

Hands-on Workshop: Role-Based NOMAD Usage and Development


Joseph F. Rudzinski and Hampus Näsström

NOMAD now provides a broad ecosystem of data infrastructure software and tools, enabling robust data management at the individual, research group, and institutional level. To navigate this ecosystem, it is useful to clearly identify your role and desired usage. In this tutorial, we will assist you in getting started down the right path, whether you are already an experienced user or are brand new to NOMAD.

You will choose from one of the following roles/topics to explore NOMAD's capabilities:

USER

- *Project and workflow management* (recommended for users with basic Python setup and minimal Python coding knowledge): organize, process, share, and publish datasets; interact programmatically via a simple Python API; customize entries and workflows.
- *Basic NOMAD usage via the GUI*: upload, share, and publish data; use NOMAD's electronic laboratory notebook (ELN) interface with built-in schemas; create ELNs with custom extensions.

APPLICATION ADMINISTRATOR (ADVANCED USER)

- *Plugin development*: create a plugin repository using a cookie-cutter template; transform custom YAML/JSON schemas into Python code; add automation and plotting features.

SYSTEM ADMINISTRATOR

- *Local infrastructure setup*: install and configure your own NOMAD Oasis; set up CI pipelines and create custom images with plugins.

Time: 13:00-14:30

Poster Session & Meet Our Experts

Our users will present their work highlighting the following topics:

- Use cases of the NOMAD platform, including novel workflows, machine learning applications, and integration into laboratory or simulation environments.
- Success stories with NOMAD Oasis that showcase how you have set up your local data infrastructure, enhanced team collaboration, or advanced reproducibility.
- Applications of large language models (LLMs) or other AI tools for mining and interpreting materials science literature or data.
- Tools, datasets, or community practices that improve the findability, accessibility, interoperability, and reusability of research data.
- Experimental research workflows enhanced by automation, robotics, or AI-driven discovery strategies.
- Machine learning applications in computational simulations and the development of reproducible, FAIR-compliant data workflows.

The participants will also be able to meet our experts and developers from various areas. Whether you need help with a specific question or need a consultation, our team will be around to help.

Time: 15:00-17:00

AI & ML Activities Across NFDI Consortia

Artificial Neural Network Inference on FPGAs in PUNCH4NFDI

Johann C. Voigt

Upcoming experiments in particle physics and astrophysics like the High-Luminosity LHC and the Square Kilometre Array will be producing data at a rate of multiple TBit/s, pushing beyond the limits of what can reasonably be stored permanently. This drives the need for a fast readout and real-time reconstruction of the data as part of a trigger system, that selects interesting events to store, while rejecting background events. Machine learning algorithms provide promising tools to increase the signal sensitivity. FPGAs offer a very high data throughput and good computing capabilities at a very low latency, making them uniquely attractive for use in readout and trigger systems. Within the Task area 5 of PUNCH4NFDI we evaluated different approaches to deploy neural networks on FPGAs. Hls4ml is a framework offering easy conversion of neural networks into FPGA firmware with a focus on low latency and ease of use. This can be used to very quickly get estimates on the feasibility of deploying a certain network architecture on FPGAs. For a particular use-case of the readout of the ATLAS experiment, we also developed a more low level, but configurable, implementation of the inference code of 1D convolutional neural networks in the VHDL hardware description language. Some FPGA models offer a connection of regular FPGA fabric with machine learning accelerator units on the same chip. We evaluated the possibility to directly program these in the context of a low level trigger system. We also collected recommendations for users who want to explore machine learning on FPGAs.

Time: 9:00-9:30

Towards FAIR workflows in computational materials science

Sarath Menon[1],[2]

Computational materials science workflows often involve numerous steps, integrate diverse software tools, span multiple length and time scales, and cover a broad range of material compositions, structures, and thermodynamic conditions. Therefore, ensuring reproducibility, data reusability, and meaningful interpretation requires not only detailed descriptions of data and metadata at each stage of the workflow but also complete provenance. This includes precise documentation of software versions, computational environments, and the execution order of workflow steps. These requirements motivate the development of FAIR computational workflows, which extend the FAIR data and FAIR4RS principles to address the specific needs of computational reproducibility and reusability.

We present design concepts for a set of automated, reproducible, and FAIR workflows implemented within the pyiron workflow environment. Pyiron provides both a graphical user interface and a Jupyter notebook-based interface, lowering the barrier for composing and managing workflows and facilitating the integration of diverse software tools. Workflow outputs are annotated using open formats, and provenance information is recorded throughout. Each step functions as a self-contained computational node with version control and automatic resolution of software dependencies. These nodes are enriched with metadata, including references to the underlying computational methods. As an example, we present a set of workflows for computing phase diagrams with near ab initio accuracy, employing machine learning interatomic potentials. These automated workflows include all steps from the generation of ab initio reference datasets to the parameterisation of the interatomic potentials, calculation of the phase

diagram, and comparison with the CALPHAD approach. This example demonstrates how complex simulation protocols can be combined in a reproducible manner to create workflows that follow the FAIR principles.

[1] Ruhr University Bochum, Germany
[2] Max Planck Institute for Sustainable Materials, Dusseldorf, Germany

Time: 9:30-10:00

Neuro-Symbolic Organization of Research Contributions with Knowledge Graphs and Large Language Models

Sören Auer

In the evolving landscape of scientific knowledge management, the integration of neuro-symbolic AI approaches offers opportunities to enhance the organization, discovery, and synthesis of research contributions. We explore how knowledge graphs and large language models (LLMs) can be synergistically combined to advance the representation and accessibility of scholarly knowledge. At the heart of this approach is the Open Research Knowledge Graph (ORKG) – a platform that structures scientific knowledge into machine-readable representations, enabling comparative analyses, automated reasoning, and contextualized exploration of research findings. Extending this vision, ORKG ASK introduces a novel query and synthesis system, combining symbolic knowledge with neural AI capabilities to provide precise, explainable, and interactive responses to complex scientific inquiries. By bridging the gap between symbolic representations and neural models, this approach aims to make scientific knowledge more accessible, transparent, and actionable – paving the way for a new era of AI-driven research assistance.

Time: 10:00-10:30

Language Models for Materials Science

Text Mining in Materials Science

Markus Stricker

Lei Zhang, Doaa Mohamed, Sepideh Baghaee Ravari, Markus Stricker

Beyond the direct raw data sources experiments and simulations, scientific publication are an underused resource at scale. The content of scientific publications can be converted into high-dimensional vector representations to gain access to the underlying correlations. Raw text can be converted to word embeddings (word2vec) and combined vision-language models can be used to extract structured datasets from scientific publications. These high-dimensional representations can then be used for data mining. I will demonstrate the potential of text mining using two examples: (1) how correlations in word embedding space can accelerate active learning loops in materials discovery, and (2) workflows for converting unstructured scientific publications to structured. However, robust and standardized pipelines for these methods are still a work in progress but these result, among others, already demonstrate useful applications.

Photo: (C) Marquard, RUB

Time: 11:00-11:30

Transforming Materials Science with Transformers

Nawaf Alampara

The advent of large-scale transformer models has opened new frontiers in scientific discovery, with materials science poised for a significant transformation. These models promise to accelerate research by automating tasks from literature review to property prediction. This talk will provide a critical evaluation of the current state of transformer-based models in chemistry and materials

science. This talk will synthesize these findings to argue that the path to transforming materials science lies not with a single monolithic model, but in a hybrid approach, where LLMs would be used to automate research task coupling with specialized models or experiments.

Time: 11:30-12:00

Hands-on Workshop: Large Language Models for Scientific Data Extraction

Mara Schilling-Wilhelmi and Sharat Patil

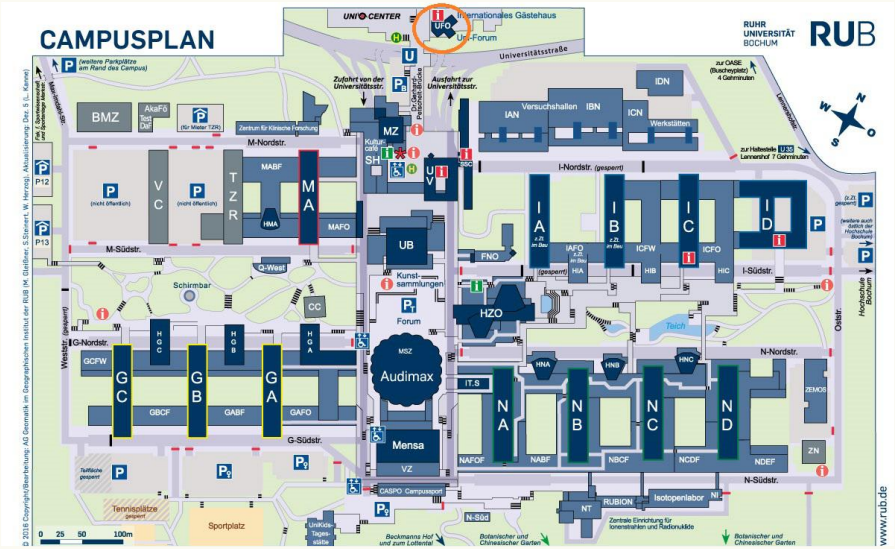
Structured data is key to machine learning – but much of the world's information is unstructured. This tutorial introduces how Large Language Models (LLMs) can be used to extract structured data from scientific publications. After a 30-minute introduction talk, participants will dive into a hands-on session exploring practical steps of the extraction workflow.

Time: 13:00-16:00

Venue

Uni-Forum Ost (UFO)

Ruhr-Universität Bochum
Querenburger Höhe 283,
44801 Bochum, Germany



Arriving by public transport: from the 'Ruhr-Universität' stop on the U35, leave the stop and turn left towards the UNI-Center via the Universitätsbrücke bridge. The UFO is located on the right at the end of the Universitätsbrücke bridge (entrance between Pizzeria and Druckhaus Bochum).

Arriving by car: approach the UFO via Universitätsstr. 150, 44801 Bochum and NOT via Querenburger Höhe 283, 44801 Bochum.

Parking at Uni-Mitte: Leave Universitätsstraße via the Uni-Mitte exit and keep left. Follow the signs for P4-9 in the middle lane (we recommend choosing a parking space in the P4-6 parking areas, as otherwise the walk will be unnecessarily long). Walk from Forumsplatz towards the

university library and university administration. Cross the Universitätsbrücke bridge towards the U35 stop and Uni-Center. The UFO is located on the right at the end of the Universitätsbrücke bridge (entrance between Pizzeria and Druckhaus Bochum).

Stay in Touch

General organizational questions:

- Natalia Slepenko: natalia.slepenko@physik.hu-berlin.de