Nature Reviews Physics | https://doi.org/10.1038/s42254-021-00373-8 | Published online 9 September 2021

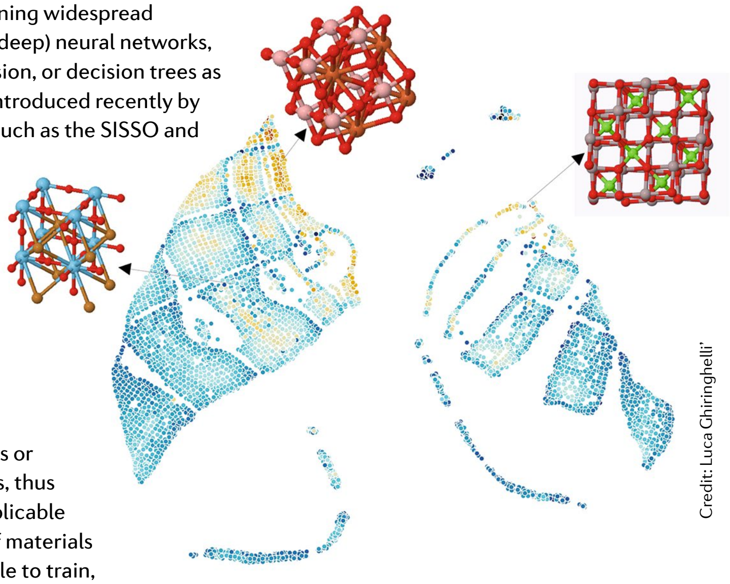
TOOLS OF THE TRADE

An AI-toolkit to develop and share research into new materials

Probably the biggest challenge in materials science is the discovery or design of new materials that exhibit exceptional performance for a desired function, or uncovering new properties of known materials. AI methods can be used to identify patterns and trends from big data to these ends. In materials science, these big data are a complex, hierarchical structure of experimental measures and theoretical estimates. Since 2014, the Novel Materials Discovery (NOMAD) Laboratory has established a materials data infrastructure, based on a large repository of materials data, and provides AI tools and training for researchers to freely access this resource, in compliance with the FAIR principles — that data should be findable, accessible, interoperable and reusable (or recyclable).

The NOMAD Lab manages a computational-materials-science repository that contains the raw input and output of more than 100 million calculations, performed by hundreds of groups in the world, using tens of supported atomistic-simulations codes. These raw data are mapped onto a standardized metadata structure, called NOMAD MetaInfo, which ensures that the raw data satisfy the FAIR guidelines. The data can be accessed, explored and analysed via the tools provided by the NOMAD AI-toolkit, which can be run on a web browser. The AI-toolkit gives access to tens (so far) of Jupyter notebooks, which are interacting web pages that allow for editing and running code (in languages such as Python) as well as annotating and/or commenting the code and its result.

By running or modifying demonstrative Jupyter notebooks (or creating and sharing newly created ones), users have direct access to the data in the NOMAD Archive via the NOMAD API and can perform AI tasks such as data mining and machine learning. This is possible through a rich library of tools spanning widespread algorithms such as (deep) neural networks, kernel-ridge regression, or decision trees as well as algorithms introduced recently by the NOMAD team, such as the SISSO and subgroup discovery. The latter are AI methods (symbolic regression) that select the relevant input parameters out of many candidates and build interpretable models in terms of analytic equations or Boolean expressions, thus transparent and applicable to fast screenings of materials spaces. Users are able to train, optimize, and test AI models as well as utilize visualization tools for an intuitive access to the data. For newcomers to the field, introductory tutorials are provided so that they can learn by imitation and exercise widespread as well as newly introduced AI tools.



Credit: Luca Ghiringhelli

The possibility for all users to produce and publish notebooks enables researchers to share the data and code used in their published studies. In this way, one can understand where each single data point in a published plot comes from, down to the input file of the atomistic code that generated it. Many scientific journals have realized that, for the sake of scientific reproducibility, data used for each publication must be made available to the community, ideally fulfilling FAIR requirements. With the AI-toolkit it is possible to share the whole analysis workflow, from the raw data to the published results, with the community.

The AI-toolkit is a community-oriented and increasingly community-driven software infrastructure for the interactive retrieval and AI-based analysis of the big data of materials science. Developers of new analysis methods are welcome to contribute and add their own tutorials to the AI-toolkit. Discovery and design of new materials are made possible by sharing the access to an enormous (and growing) amount of data as well as AI algorithms and visualization tools. In this sense, we use a forward looking interpretation of the FAIR acronym as Findable & AI-Ready data.

Luca M. Ghiringhelli

The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society and Humboldt University, Berlin, Germany

e-mail: ghiringhelli@fhi-berlin.mpg.de

Acknowledgements
The author would like to acknowledge the other key players in developing the NOMAD-Laboratory infrastructure: Fawzi Mohamed (concept and implementation of MetaInfo); Markus Scheidgen (design and implementation of the Repository & Archive); Matthias Scheffler, Angelo Ziletti, Luigi Sbailò and Ádám Fekete (concept, design and implementation of the AI-toolkit) . Finally, Matthias Scheffler and Claudia Draxl for the conception and leadership of the NOMAD Laboratory.

Competing interests
The author declares no competing interests.