JPhys Materials

PURPOSE-LED PUBLISHING™

TECHNICAL NOTE • OPEN ACCESS

The NOMAD laboratory: from data sharing to artificial intelligence

To cite this article: Claudia Draxl and Matthias Scheffler 2019 *J. Phys. Mater.* **2** 036001

View the article online for updates and enhancements.

JPhys Materials

OPEN ACCESS

RECEIVED
13 February 2019

REVISED
16 March 2019

ACCEPTED FOR PUBLICATION
27 March 2019

PUBLISHED
13 May 2019

TECHNICAL NOTE

The NOMAD laboratory: from data sharing to artificial intelligence

Claudia Draxl[1,2] and Matthias Scheffler[1,2]

[1] IRIS Adlershof, Humboldt-Universität zu Berlin, Zum Großen Windkanal 6, D-12489 Berlin, Germany
[2] der Max-Planck-Gesellschaft, Faradayweg 4-6, D-14195 Berlin, Germany

E-mail: claudia.draxl@physik.hu-berlin.de **and** scheffler@fhi-berlin.mpg.de

Keywords: data science, data repository, data analytics, metadata, computational materials

Abstract

The Novel Materials Discovery (NOMAD) Laboratory is a user-driven platform for sharing and exploiting computational materials science data. It accounts for the various aspects of data being a crucial raw material and most relevant to accelerate materials research and engineering. NOMAD, with the NOMAD Repository, and its code-independent and normalized form, the NOMAD Archive, comprises the worldwide largest data collection of this field. Based on its findable accessible, interoperable, reusable data infrastructure, various services are offered, comprising advanced visualization, the NOMAD Encyclopedia, and artificial-intelligence tools. The latter are realized in the NOMAD Analytics Toolkit. Prerequisite for all this is the NOMAD metadata, a unique and thorough description of the data, that are produced by all important computer codes of the community. Uploaded data are tagged by a persistent identifier, and users can also request a digital object identifier to make data citable. Developments and advancements of parsers and metadata are organized jointly with users and code developers. In this work, we review the NOMAD concept and implementation, highlight its orthogonality to and synergistic interplay with other data collections, and provide an outlook regarding ongoing and future developments.

1. Introduction

Big-data driven materials science is emerging as a new research paradigm at the beginning of the 21st century. The amount of data that has been created in materials science in recent years and is being created every new day is immense, and the field is facing the 4*V challenges* of big data [1]. Bringing together big data with high-performance computing is a new, emerging field with a game-change potential for numerous applications. It promises breakthroughs in the improvement of the analysis and description of scientific phenomena and materials properties. Clearly, the future of computational materials science will combine simulations with datadriven and modeldriven techniques in an integrated manner. This will require interdisciplinary teams as well as next-generation researchers with different skill sets, mastering physics, chemistry or materials science simulations with data analysis and management as well as hardware aspects.

Also, to fully exploit the significant information and potential of the massive amounts of available data, requires to provide the community with proper infrastructures that allow for efficiently collecting, curating, and sharing data.

The most prominent early data collections of computational materials science were established to serve different special purposes rather than sharing and enabling contributions from the community. AFLOW [2] is the biggest single data base in *ab initio* computational materials science in the USA, with a focus on complex alloys. More recently, AFLOW also offers online applications for property predictions using machine learning. The Materials Project [3] may be called the initiator of materials genomics in the field. Its mission is to accelerate the discovery of new technological materials, e.g. for batteries, electrodes, etc, through advanced scientific computing and innovative design. The open quantum materials database (OQMD) [4] has a focus on energetics and thermodynamics. In all three the main underlying quantum engine is the VASP code [5], and the data were originally not open to the public, but—as will be described below—they are now open and part of the Novel

Table 1. Components of the NOMAD laboratory.

	Purpose	Data source/format
NOMAD Repository	Open-access platform for data sharing	Raw data from 40 different codes, provided under cc-by license
NOMAD Archive	Data source of NOMAD services	Normalized data in unified file format and units, provided under cc-by license
NOMAD Encyclopedia	Graphical user interface to characterize materials by computed properties	Based on archive data
Advanced visualization	Remote visualization tools, including virtual reality	Based on archive data
NOMAD Analytics Toolkit	Prototypical examples of artificial-intelligence tools	Archive data; also applicable to external data sources

Materials Discovery (NOMAD) data collection. For an overview of the current status of data-centric science for materials innovation and a more detailed account of the mentioned important initiatives we refer to a special issue of the Materials Research Society Bulletin [6].

Here, we describe how the NOMAD Laboratory has initiated data sharing in the computational materials-science community and give an overview about NOMAD's concept, its data collection, and its services based on it. Finally, an outlook is provided about the next steps and how to meet the big picture.

2. Game change by data sharing

The importance of an extensive data sharing for the advancement of science and engineering has been stated often, and it is a generally accepted assessment. Unfortunately, however, it has been mostly empty rhetoric of scientific societies and funding agencies [7]. For the field of computational materials science, the NOMAD Laboratory [8] has changed the scientific attitude towards a comprehensive and FAIR data sharing, opening new avenues for mining and refining big data of materials science [9]. (For the definition of FAIR; i.e. findable, accessible, interoperable, reusable see below and [10].) In November 2014, the principal investigators of the NOMAD Center of Excellence [8] launched an initiative at the Psi-k community, proposing 'a change in scientific culture' of computational materials science and engineering through extensive sharing of data—*all* data [11]. Clearly, only if the full input and output files of computations are shared, calculations do not need to be repeated again and again, and the wider community, or even the entire research field has access to big data which can be used in a totally new research manner, i.e. by artificial-intelligence (AI) methods. The vision of where this field should be going together with the most successful data-analytics concepts pursued in materials science and several examples thereof are described in [1].

Since going online in 2014, the number of calculations collected in the NOMAD Repository has exceeded even the most ambitious expectations. Thereby the NOMAD Center of Excellence [8] has assumed a pioneering role in data sharing and analytics, and in particular in all aspects of what is now called the FAIR handling of data [1, 10][3]. It has always been NOMAD's principle that data is *Findable* for anyone interested and stored in a way that make it easily *Accessible*; its representation follows accepted standards [12, 13], with open specifications, making data *Interoperable*; and finally, all of this makes the data *Re-purposable*[4], i.e. enables data to be used for research questions that could be different from their original purpose.

3. The NOMAD Laboratory concept

The NOMAD Laboratory processes, curates, and hosts computational materials science data, computed by all important materials-science codes available today and makes this data accessible by providing several related data services. More and more codes are added, jointly with the code developers or expert users. The big picture is to advance materials science by enabling researchers in basic science and engineering to understand and utilize materials data in order to identify improved, new or even novel materials and, in turn, pave the way to novel products and technologies. NOMAD also collaborates with many researchers and all other big databases, specifically AFLOW [2, 14–16], Materials Project [3, 17, 18], OQMD [4, 42]. In the following, NOMAD's cornerstones, as summarized in table 1, are described in some more detail.

[3] The concept of the NOMAD Repository and Archive was developed in 2014, independently and parallel to the 'FAIR Guiding Principles' [10]. Interestingly, the essence is practically identical.

[4] The NOMAD CoE uses the term re-purposable, while in the FAIR concept it was termed re-usable. Both obviously mean the same in this context.

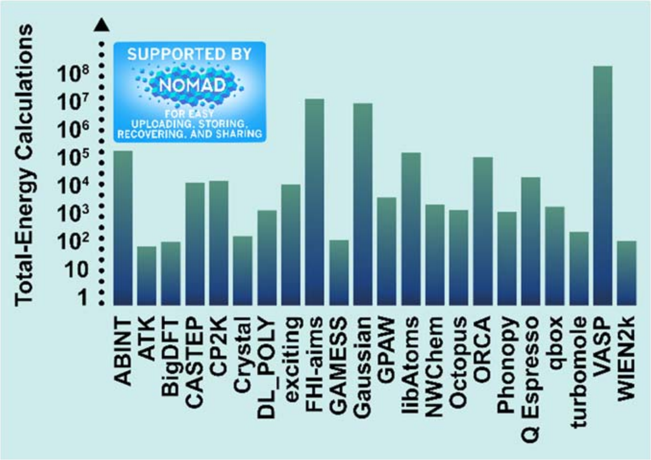

Figure 1. The NOMAD Laboratory supports all important codes in computational materials science. The figure shows the number of uploaded open-access total-energy calculations at the NOMAD Repository (NOMAD) as of 15 March 2018. The abscissa shows the various codes with more than 80 uploads. The total number of open-access total-energy calculations at the NOMAD Repository is more than 50 million, corresponding to billions of CPU-core hours. The stamp 'supported by NOMAD' can be found on the homepage of many *ab initio* software packages of computational materials science. Figure adapted with permission from [19]. © Materials Research Society 2018.

The *NOMAD Repository* contains the raw data as they are created by the employed computer code, and these data form the basis of all NOMAD services. Made publically available in early 2014, the NOMAD Repository now contains the input and output files from millions of calculations and has become the world's largest collection of computational materials science data. It comprises calculations that have been produced with any of the leading electronic-structure codes, and increasingly also with codes from quantum chemistry and force-field/molecular-mechanics simulations. Presently, NOMAD supports about 40 codes, and less-frequently used or newly established codes are being added on demand in close collaboration with the code developers or expert users. We point to the orthogonality to other databases, and emphasize that the NOMAD Repository is not restricted to selected computer codes or closed research teams but serves the entire community with its ecosystem of very different computational methods and tools.

By hosting raw data and keeping scientific data for free for at least 10 years[5], the NOMAD Repository not only helps researchers to organize their results and make them available to others. Every calculation is uniquely identified by a persistent identifier, and digital object identifiers are issued on request. This makes data citable and helps to make connections between publications and data. It also supports the community to find out what kind of calculations have been performed for materials of interest, by whom and by which code. The NOMAD Repository was the first repository in materials science recommended by Scientific Data as stable and safe long-time storage.

Uploading of data is possible without any barrier. Results are requested in their raw format as produced by the underlying code. NOMAD implements an open data policy, where data are published according to the Creative Commons Attribution 3.0 License. Accessibility of data in NOMAD goes significantly further than meant in the FAIR Guiding Principles [10], as for searching and even downloading data from NOMAD, users do not even need to register. Nevertheless, uploaders can keep their data secret for a certain period that can be used for publishing the results and/or restricting the access to a selected group of colleagues (or referees). After a maximum period of three years though, all data become open access.

The *NOMAD Archive* hosts the normalized data, i.e. the open-access data of the NOMAD Repository converted into a common, code-independent format. In other words, numerous parsers have been developed that read out and translate all the information contained in in- and output files. This ensures the *I* in FAIR, namely that data from different sources can be compared and, hence, collectively operated upon by various NOMAD (and other) tools.

Figure 1 reflects the active participation of the entire community in the NOMAD initiative. It depicts a snapshot of the NOMAD Archive by 15 March 2018, showing the number of uploads (total-energy calculations) of data from various community codes.

Obviously, a clear metadata definition [20] is a prerequisite for the normalization step (and for all later steps) to make sure that the results/values obtained by atomistic/*ab initio* calculations are correctly interpreted by the

[5] NOMAD guarantees to keep data for at least ten years from the last upload. In practice this means that overall, data are kept much longer than ten years.

Figure 2. Virtual-reality setup as demonstrated during the Long Night of Sciences, Berlin 2017.

parsers. As such activity should be community driven, a first key workshop (others are already following) was organized at the CECAM Headquater, bringing together players from different codes and research areas. The goal and the outcome of this workshop was published in [12, 13]. The following development of an open, flexible, and hierarchical metadata classification system [20] was indeed challenging. The overall effort defining general and code-specific metadata and developing parsers for 40 codes amount to some ten or twenty person years. Obviously, as codes are continuously updated and extended, and new codes are being developed, this is an ongoing process, to which everybody is welcome to contribute.

The *NOMAD Visualization Tools* allow for remote visualization of the multi-dimensional NOMAD data through a dedicated infrastructure developed within the NOMAD Laboratory. A centralized service is provided that enables users to interactively perform comprehensive data visualization tasks on their computers without the need for specialized hardware or software installations. Interactive data exploration by virtual-reality (VR) tools is a special and most successful focus. Also for this branch of the toolbox, users have access to data and tools using standard devices (laptops, smartphones), independent of their location. Such VR enhances training and dissemination and even were a great success, e.g. when presented to the general public during the Long Night of Sciences, taking place in Berlin, in June 2017 (see figure 2). As an example, we note that 360° movies can be even watched with simple Google cardboard glasses as demonstrated, e.g. for CO_2 adsorption on CaO and excitons in LiF or hybrid materials [21]. Excitons, being six-dimensional objects, cannot easily be visualized in a standard way. Taking the position of an electron or a hole, VR allows for inspecting the 'space' (probability distribution) of its counterpart. This example indeed demonstrates how seeing helps understanding.

As the visualization tools already have evidenced, it is extremely useful to provide the data not only in a machine-readable but also in a human-accessible form in order to get a first insight into materials data. Therefore, the NOMAD CoE has created its data infrastructure not only for collecting and sharing data but also to let us see and explore what information all this data contain. The *NOMAD Encyclopedia* is a web-based public platform that gives a materials-oriented view on the Archive data that helps us to search for the properties of a large variety of materials. Here, unlike in the Repository, one does neither search for a single code nor for a specific contributor. The Encyclopedia serves the purpose of in-depth characterization and understanding of the materials of interest by providing knowledge of their various properties. This includes structural features, mechanical and thermal behavior, electronic and magnetic properties, the response to light, and more. Whatever property of a given material has been computed, the aim is to make this information easily accessible through a user-friendly graphical user interface. Having all this information in one place gives us an impression of the wealth of available materials data and even allows for comparing very different systems in terms of certain features. The NOMAD Encyclopedia allows us to explore and comprehend computations obtained with various tools and different methodology. We can directly assess the spread of results and, as for instance, measure the impact of a density functional on a given feature and materials class.

So far, the NOMAD Encyclopedia processes structural, electronic, thermal properties (see figure 3 as an example), and more for bulk materials and low-dimensional systems. It is constantly extended in terms of new data, other system types, and properties. And soon, the Encyclopedia will handle molecules, surfaces and adsorbate systems (as shown in figure 4), the response to external excitations, elastic properties, Fermi surfaces, molecular dynamics, and more.

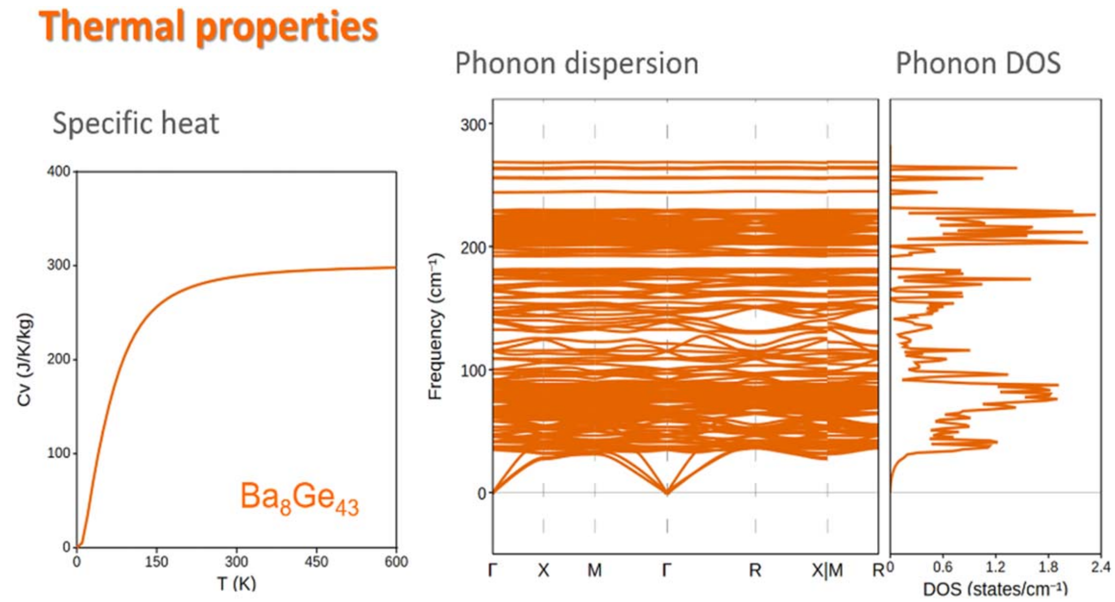

Figure 3. Collage showing as example thermal properties of Ba_8Ga_{43} from the NOMAD Encyclopedia.

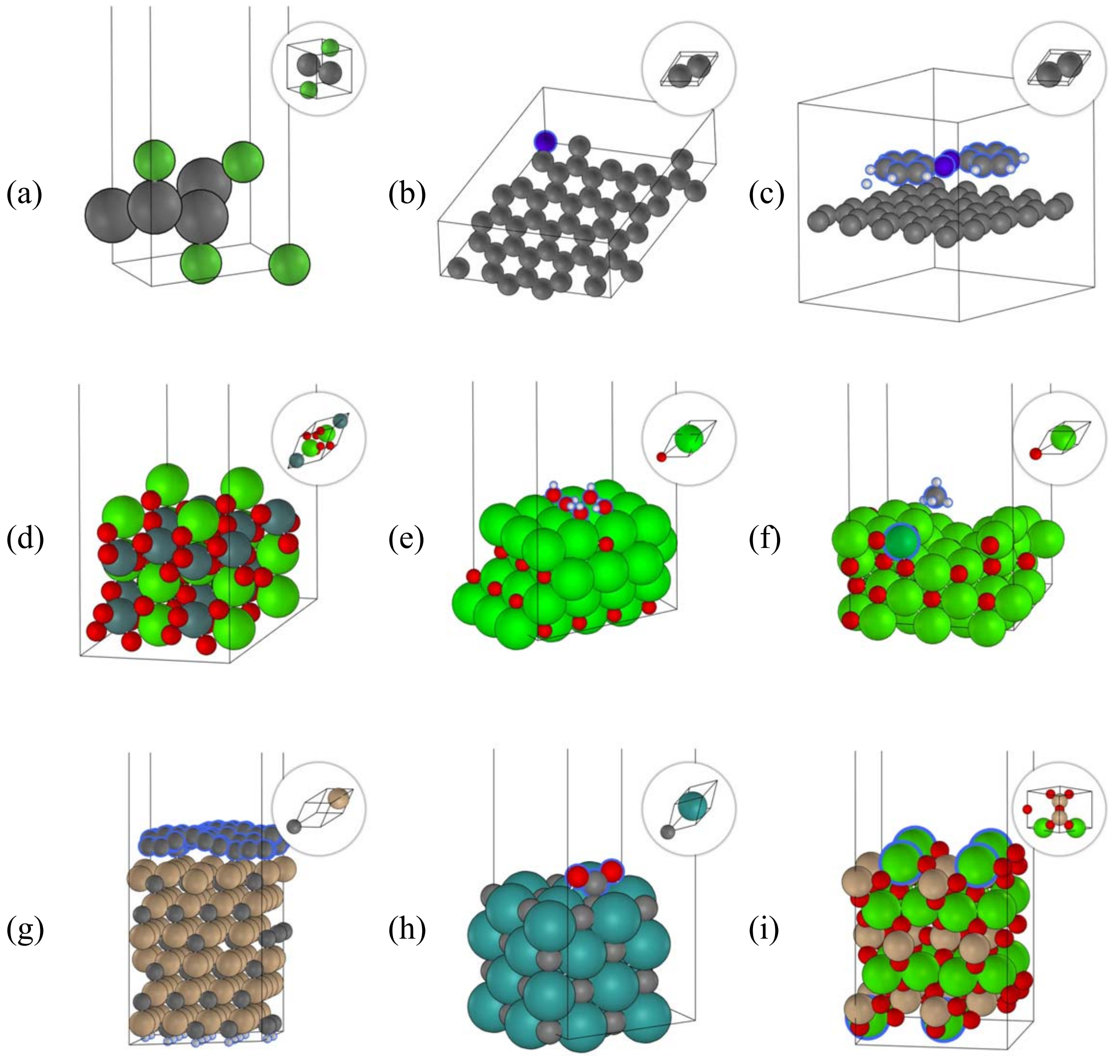

Figure 4. Unit cells of various surface systems with defects and adsorbed atoms and molecules. In the respective upper right corners, the irreducible unit cell of the pristine crystals are shown as detected by the Encyclopedia classification algorithm.

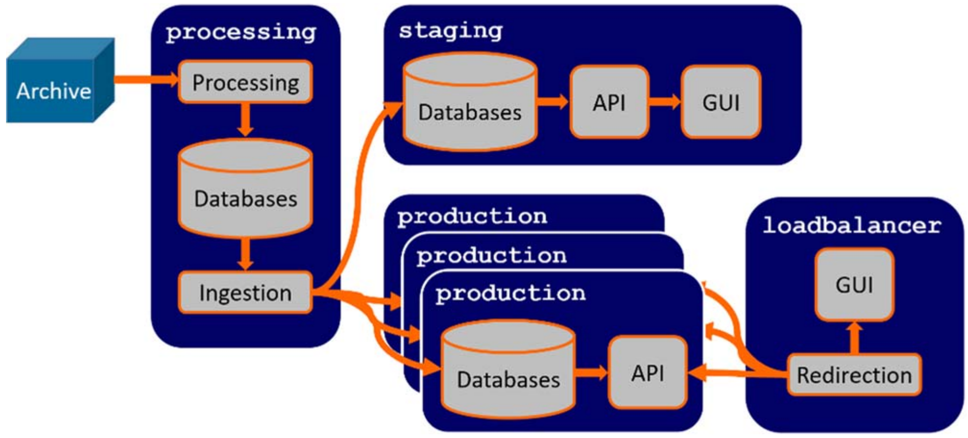

Figure 5. Current NOMAD Encyclopedia environment consisting of the processing, staging, and three production systems, each hosting several database- and web-services (API stands for Application Programmbale Interface, GUI for graphical user interface).

Furthermore, the Encyclopedia provides a material classification system, information on various levels, e.g. about computed quantities or the methodology behind the calculations, as well as links to external sources, like to useful Wikipedia pages. We also point to the error-reporting tool, implemented for the case of problematic data. Should there be a dataset or a graph that does not appear to be correct, making use of a simple pull-down menu, the user can let us know about it.

The Encyclopedia infrastructure currently consists of processing, stating and three production systems that are coordinated by a load balancer (see figure 5) to increase fault-tolerance, availability and handle dynamic loads.

The NOMAD infrastructure provides the ideal platform for data-driven science. This topic is discussed in detail in [1]. Here, we focus on the *NOMAD Analytics Toolkit* which provides a collection of examples and tools to demonstrate how materials data can be turned into knowledge. When the results of AI methods are presented and discussed in scientific publications, a challenge is posed in terms of the complete and sound description of the provided model and of the way it was trained. This is difficult to fully report in a written text. One practical solution is to provide the model in the form of an 'interactive code' that can be inspected and run interactively by any interested researcher and even the wider public. Even better, one can provide the algorithm that trained the model, starting from the training data. Readers can then at once verify the procedure and learn new strategies in depth. The NOMAD Analytics Toolkit [22] was created as a platform for hosting such 'interactive codes'. Its infrastructure is composed of data-science notebooks, which run in containers. It allows developers of data-science notebooks related to their research work to be free in the choice of programming language, version, required libraries, etc. Presently, several notebooks prepared by the NOMAD team are already publicly available through the Analytics Toolkit and present the learning and/or the application of data-analytics models, applied to materials science. They concern topics like crystal-structure prediction, property prediction, error estimates, classification of materials, and more. Some of these notebooks correspond to peer-reviewed publications [23–34].

To explore these NOMAD analytics tools, there is no need to install any software, no need for registration, and no need for computational capacity (for not too demanding requests). Like the visualization tools, also this toolkit allows for remote usage. We emphasize that though initiated by the NOMAD CoE, the wider materials science community has been invited to use the NOMAD Analytics Toolkit platform to upload notebooks for any data-analytics publication through dissemination channels, such as NOMAD Summer 2018 [35], and has actively participating also in hackathons, contributing their ideas and tools to this platform.

Let us conclude this section by pointing to informative YouTube movies that describe the NOMAD project as a whole https://youtu.be/yawM2ThVlGw and https://youtu.be/sI2cPuIGNUU, the Repository https://youtu.be/UcnHGokl2Nc, and the Analytics Toolkit https://youtu.be/UcnHGokl2Nc, all of them also available at [21].

4. From data to novel tools by crowd sourcing

It is surprising to see what people do with data once it is made available in a fully characterized way. This was impressively demonstrated by a public data-analytics competition [36] set up by the NOMAD CoE. It was hosted by Kaggle—an online platform for data-science competitions. Specifically, this competition addressed the need to identify novel transparent and semiconducting materials using a dataset of $(Al_xGa_yIn_z)_2O_3$ compounds (with $x + y + z = 1$). The aim of this challenge was to identify the best machine-learning model for the prediction of two key physical properties that are relevant for optoelectronic applications: the electronic band-gap and the crystal-formation energy. These target properties were provided for 2400 fully relaxed crystalline systems. However, the participants were given only an idealized (unrelaxed) geometry of the atomic positions. These unrelaxed structures were generated without an extensive calculation but just using the concentration-weighted averages of the pure binary crystalline systems (Al_2O_3, Ga_2O_3, and In_2O_3). The reason for only providing unrelaxed geometries was that the exact geometry is typically unknown *a priori*, and a model using features from the optimized structure would thus require first calculating the relaxed geometry for any new machine-learning prediction. In this case, the band gap and formation energy would also be known for the relaxed structure, and the prediction of these properties by machine learning would be of little value.

The test set for the competition was composed of an additional 600 systems where only the idealized geometries and no other materials property was provided. By including data from 7 different crystal space groups and unit cells ranging from 10 to 80 atoms, both data sets (those for learning and for testing) were constructed to ensure that the proposed models would be generalizable to diverse structures and various sizes.

The competition was launched on 18 December 2017, and when it finished on 15 February 2018, an impressive total of 883 solutions had been submitted by researchers or research teams from around the world, employing many different methods. A recent publication [32] presents a summary of the top three crowd-sourced machine-learning approaches from the competition. Interestingly, the top three approaches adopted completely different descriptors and regression models. The winning solution employed a crystal-graph representation to convert the crystal structure into features by counting the contiguous sequences of unique atomic sites of various lengths (called *n*-grams), and combined this with kernel ridge regression [37]. The 2nd-place model represented each system using many candidate descriptors from a set of compositional, atomic environment-based, and average structural properties, which was combined with the light gradient-boosting machine regression model [38]. The 3rd-place solution employed the smooth overlap of atomic positions representation [39] combined with a neural network.

5. Outlook

The by now well established NOMAD Laboratory has recently been incorporated into a non-profit association, FAIR Data Infrastructure for Physics, Chemistry, Materials Science, and Astronomy e.V. (FAIR-DI) [40]. This initiative is dedicated to providing a stable and sustainable data infrastructure that hosts and further develops the NOMAD Repository, Archive, and Encyclopedia such to guarantee long-term viability, independent of individual research projects currently driving it. It also forms an ideal basis for advancing data-driven science, for example, also addressing experimental studies. The proof of concept regarding Artificial Intelligence for materials science has been successfully demonstrated, as discussed in detail in [1], with examples shown in the Analytics Toolkit [22]. Now it is time for the next steps towards the ambitious goals of rigorous materials classification and NOMAD. NOMAD's long-term vision is to provide (high-dimensional) materials maps that tell where in the composition and compound space one can find materials for a certain application, be it thermoelectrics, solar-cell materials, superconductors, topological insulators, or other systems of interest.

What FAIRDI's name already suggests is another crucial and most timely issue, i.e. making the link to experimental materials science. We note that this is a very demanding task. The input file in computations is a comprehensive characterization of the sample used in the experiment. In the best case this includes the whole workflow of the sample preparation and the sample's history. Furthermore, it may also be difficult to describe the experimental setup in every potentially relevant detail. First steps, building on NOMAD's experience, are already taken together with the BiGmax network [41], where computational data-driven materials science meets data-intensive experiments. Thus far, the search for new materials enabling new applications (or improved performance) was limited to educated guesses mostly based on selective experiments. Recent advances in data mining will allow pattern recognition and pattern prediction in an unprecedented way. The outcome of big-data-driven materials science approaches will then impact the way experiments and data analyses will be done in the future.

On the computational side, there are also big challenges to tackle. Despite the fact that millions of CPU core hours are spent every day on supercomputers worldwide, thereby tremendously increasing the amount of available data, we still may ask the question whether we have enough data for trustful materials discovery? In fact, current state-of-the-art electronic-structure methods, i.e. DFT with semi-local functionals, often face severe deficiencies when describing modern materials. The approximations made for exchange-correlation effects are often too crude to properly capture, e.g. the energetics of defects and polaronic distortions, and often completely fail to predict the energy-level alignment at interfaces between materials of different character, like organicinorganic hybrid materials. Hence, higher-level methodologies, ranging from advanced density functionals to coupled-cluster (CC) theory, are essential to enable computational materials science to meet the needs of modern technology. However, on present-day computers such higher-level calculations are not feasible. While DFT, with most popular semi-local functionals can handle nowadays 1000 atoms and more, the higher-level hybrid functionals are limited to the order of 100 atoms. The same holds for the random-phase approximation and the *GW* approach of many-body perturbation theory. CC theory, the gold-standard of quantum chemistry, is currently out of reach for system sizes beyond some 10 atoms. Thus reliably predicting novel materials with desired properties by quantum-theory based tools can currently not be realized because highly precise methods are computationally too demanding to be applied to 'real' materials. Obviously, our aim must be to push the limits of *ab initio* computational science in terms of speed, accuracy, and precision, to finally serve the scientific community, industry, and thus society.

Acknowledgments

This work received funding from the European Union's Horizon 2020 Research and Innovation Programme, Grant Agreement No. 676580, the NOMAD Laboratory CoE and No. 740233, ERC TEC1P. The initial phase of the NOMAD design and development was supported by a donation from the Supporting Members of the MPG and by the Einstein Foundation Berlin through the project ETERNAL. The work profited from programs and discussions at the Institute of Pure and Applied Mathematics (IPAM) at UCLA, supported by the NFS, and from BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials-Science. We thank the whole NOMAD team for the invaluable effort to build the entire NOMAD infrastructure and its services.

ORCID iDs

Claudia Draxl https://orcid.org/0000-0003-3523-6657

References

[1] Draxl C and Scheffler M 2019 Big-data-driven materials science and its FAIR data infrastructure *Handbook of Materials Modeling* ed S Yip *et al* (Basel: Springer International Publishing) submitted (arXiv:1904.05859)

[2] AFLOW. Automatic FLOW for materials discovery http://aflowlib.org; see [14–16]

[3] Materials Project https://materialsproject.org

[4] OQMD. Open Quantum Materials Database http://oqmd.org

[5] VASP https://vasp.at

[6] Tanaka I, Rajan K and Wolverton C 2018 *MRS Bull.* **43** 662

[7] Nature Editorial 2017 *Nature* **546** 327

[8] NOMAD https://nomad-coe.eu

[9] Draxl C, Illas F and Scheffler M 2017 *Nature* **548** 523

[10] Wilkinson M D *et al* 2016 *Sci. Data* **3** 160018

[11] Why Sharing? https://repository.nomad-coe.eu/

[12] Ghiringhelli L M, Carbogno C, Levchenko S, Mohamed F, Huhs G, Lüder M, Oliveira M and Scheffler M 2016 Psi-k Scientific Highlight of the Month No.131 http://psi-k.net/download/highlights/Highlight_131.pdf

[13] Ghiringhelli L M, Carbogno C, Levchenko S, Mohamed F, Hus G, Lüder M, Oliveira M and Scheffler M 2017 *npj Comput. Mater.* **3** 46

[14] Toher C *et al* 2018 The AFLOWFleet for materials discovery *Handbook of Materials Modeling* ed S Yip *et al* (Basel: Springer International Publishing)

[15] Curtarolo S, Setyawan W, Hart G L W, Jahnatek M and Chepulskii R V 2012 *Comput. Mater. Sci.* **58** 218

[16] Calderon C E *et al* 2015 *Comput. Mater. Sci.* **108** 233

[17] Jain A *et al* 2013 *APL Mater.* **1** 011002

[18] Jain A, Montoya J, Dwaraknath S, Zimmermann N E R, Dagdelen J, Horton M, Huck P, Winston D, Cholia S, Ong S P and Persson K 2018 The materials project: accelerating materials design through theory-driven data and tools *Handbook of Materials Modeling* ed S Yip *et al* (Basel: Springer International Publishing)

[19] Draxl C and Scheffler M 2018 *MRS Bull.* **43** 676

[20] NOMAD Meta Info. The metadata structure defined for the NOMAD Laboratory (called NOMAD Meta Info) is a conceptual model to store results from *ab initio* and force-field atomistic calculations https://nomad-coe.eu/the-project/nomad-archive/archive-meta-info/

[21] YouTube channel of the NOMAD CoE: https://www.youtube.com/channel/UCG1B2E82zuJ-MChCuofOBaA

[22] NOMAD Analytics Toolkit https://analytics-toolkit.nomad-coe.eu

[23] Saad Y, Gao D, Ngo T, Bobbitt S, Chelikowsky J R and Andreoni W 2012 *Phys. Rev.* B **85** 104104

[24] Ghiringhelli L M, Vybiral J, Levchenko S V, Draxl C and Scheffler M 2015 *Phys. Rev. Lett.* **114** 105503

[25] Ghiringhelli L M, Vybiral J, Ahmetcik E, Ouyang R, Levchenko S V, Draxl C and Scheffler M 2017 *New J. Phys.* **19** 023017

[26] Goldsmith B R, Boley M, Vreeken J, Scheffler M and Ghiringhelli L M 2017 *New J. Phys.* **19** 013031

[27] Glielmo A, Sollich P and De Vita A 2017 *Phys. Rev.* B **95** 214302

[28] Ouyang R, Curtarolo S, Ahmetcik E, Scheffler M and Ghiringhelli L M 2018 *Phys. Rev. Mater.* **2** 083802

[29] Lambert H, Fekete A, Kermode J R and De Vita A 2018 *Comput. Phys. Commun.* **232** 256

[30] Ziletti A, Kumar D, Scheffler M and Ghiringhelli L M 2018 *Nat. Commun.* **9** 2775

[31] Bartel C J, Sutton C, Goldsmith B R, Ouyang R, Musgrave C B, Ghiringhelli L M and Scheffler M 2019 *Sci Adv.* **5** eaav0693

[32] Sutton C, Ghiringhelli L M, Yamamoto T, Lysogorskiy Y, Blumenthal L, Hammerschmidt T, Golebiowski J, Liu X, Ziletti A and Scheffler M 2019 *npj Comput. Mater.* accepted

[33] Acosta C M, Ouyang R, Fazzio A, Scheffler M, Ghiringhelli L M and Carbogno C 2019 *Nat. Mater.* in preparation

[34] Carbogno C *et al* 2019 in preparation

[35] NOMAD Summer 2018 http://meetings.nomad-coe.eu/nomad-summer-2018/index.php?n=Meeting.Scope

[36] Kaggle/Nomad 2018 Predicting Transparent Conductors - Predict the key properties of novel transparent semiconductors https://kaggle.com/c/nomad2018-predict-transparent-conductors

[37] Broder A Z, Glassman S C, Manasse M S and Zweig G 1997 *Comput. Netw. ISDN Syst.* **29** 1157

[38] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye O and Liu T-Y 2017 *Adv. Neural Inf. Process. Syst.* **30** 3146–54

[39] Bartók A P, Kondor R and Csányi G 2013 *Phys. Rev.* B **87** 184115

[40] https://fairdi.eu

[41] https://bigmax.mpg.de/

[42] Saal J, Kirklin S, Aykol M, Meredig B and Wolverton C 2013 *JOM* **65** 1501–9