OPEN ACCESS
IOP Publishing
Modelling and Simulation in Materials Science and Engineering

Modelling Simul. Mater. Sci. Eng. **32** (2024) 063301 (90pp) https://doi.org/10.1088/1361-651X/ad4d0d

Roadmap

Roadmap on data-centric materials science

Stefan Bauer[1], Peter Benner[2], Tristan Bereau[3], Volker Blum[4], Mario Boley[5], Christian Carbogno[6], C Richard A Catlow[7,8,9,10], Gerhard Dehm[11], Sebastian Eibl[12], Ralph Ernstorfer[13], Ádám Fekete[14], Lucas Foppa[6], Peter Fratzl[15], Christoph Freysoldt[11], Baptiste Gault[11,16], Luca M Ghiringhelli[6,17], Sajal K Giri[18], Anton Gladyshev[14], Pawan Goyal[2], Jason Hattrick-Simpers[19], Lara Kabalan[7,20,21], Petr Karpov[12], Mohammad S Khorrami[11], Christoph T. Koch[14], Sebastian Kokott[6,22], Thomas Kosch[23], Igor Kowalec[7,8], Kurt Kremer[24], Andreas Leitherer[6,25], Yue Li[11], Christian H Liebscher[11], Andrew J Logsdail[7,8], Zhongwei Lu[7,8], Felix Luong[5], Andreas Marek[12], Florian Merz[26], Jaber R Mianroodi[11], Jörg Neugebauer[11], Zongrui Pei[27], Thomas A R Purcell[6,28], Dierk Raabe[11], Markus Rampp[12], Mariana Rossi[29], Jan-Michael Rost[30], James Saal[31], Ulf Saalmann[30], Kasturi Narasimha Sasidhar[11], Alaukik Saxena[11], Luigi Sbailò[14], Markus Scheidgen[14], Marcel Schloz[14], Daniel F Schmidt[5], Simon Teshuva[5], Annette Trunschke[32], Ye Wei[33], Gerhard Weikum[34], R Patrick Xian[35], Yi Yao[6], Junqi Yin[36], Meng Zhao[14] and Matthias Scheffler[6,*]

[1] School of Computation, Information and Technology, Technical University of Munich & Helmholtz AI, Munich, Germany
[2] Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany
[3] Institute for Theoretical Physics, Heidelberg University, Heidelberg, Germany
[4] Thomas Lord Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC, United States of America
[5] Department of Data Science and AI, Monash University, Melbourne, Australia
[6] The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society, Berlin, Germany

[*] Author to whom any correspondence should be addressed.

[7] Max Planck Centre on the Fundamentals of Heterogeneous Catalysis (FUNCAT), School of Chemistry, Cardiff University, Cardiff, United Kingdom
[8] Cardiff Catalysis Institute, School of Chemistry, Cardiff University, Cardiff, United Kingdom
[9] UK Catalysis Hub, Research Complex at Harwell, RAL, Oxford, United Kingdom
[10] Department of Chemistry, University College London, London, United Kingdom
[11] Max Planck Institute for Sustainable Materials, Düsseldorf, Germany
[12] Max Planck Computing and Data Facility, Garching, Germany
[13] Institute for Optics and Atomic Physics, Technical University of Berlin, Berlin, Germany
[14] Department of Physics & CSMB, Humboldt-Universität zu Berlin, Berlin, Germany
[15] Max Planck Institute of Colloids and Interfaces, Potsdam, Germany
[16] Department of Materials, Imperial College, London, United Kingdom
[17] Department of Materials Science and Engineering, Friedrich-Alexander Universität, Erlangen-Nürnberg, Germany
[18] Department of Chemistry, Northwestern University, Evanston, IL, United States of America
[19] Department of Materials Science and Engineering, University of Toronto, Toronto, Canada
[20] STFC Hartree Centre, Daresbury Laboratory, Daresbury, Warrington, United Kingdom
[21] STFC Hartree Centre, Daresbury Laboratory, Daresbury, Warrington, United Kingdom
[22] Molecular Simulations from First Principles e.V., Berlin, Germany
[23] Department of Computer Science, Humboldt-Universität zu Berlin, Berlin, Germany
[24] Max Planck Institute for Polymer Research, Mainz, Germany
[25] ICFO-Institut de Ciencies Fotoniques, The Barcelona Institute of Science and Technology, Castelldefels, Barcelona, Spain
[26] Lenovo HPC Innovation Center, Stuttgart, Germany
[27] New York University, New York, NY 10012, United States of America
[28] The Department of Chemistry and Biochemistry, University of Arizona, Tucson, AZ, United States of America
[29] Max Planck Institute for the Structure and Dynamics of Matter, Hamburg, Germany
[30] Max Planck Institute for the Physics of Complex Systems, Dresden, Germany
[31] Citrine Informatics, Inc., Redwood City, CA, United States of America
[32] Department of Inorganic Chemistry, Fritz Haber Institute of the Max Planck Society, Berlin, Germany
[33] Ecole Polytechnique Fédérale de Lausanne, School of Engineering, Lausanne, Switzerland
[34] Max Planck Institute for Informatics, Saarbrücken, Germany
[35] Department of Statistical Sciences, University of Toronto, Toronto, Canada
[36] Oak Ridge National Laboratory, Oak Ridge, TN, United States of America

E-mail: scheffler@fhi-berlin.mpg.de

Received 24 January 2024; revised 1 May 2024
Accepted for publication 17 May 2024
Published 3 July 2024

Abstract

Science is and always has been based on data, but the terms 'data-centric' and the '4th paradigm' of materials research indicate a radical change in how information is retrieved, handled and research is performed. It signifies a transformative shift towards managing vast data collections, digital repositories, and innovative data analytics methods. The integration of artificial intelligence and its subset machine learning, has become pivotal in addressing all these challenges. This Roadmap on Data-Centric Materials Science explores fundamental concepts and methodologies, illustrating diverse applications in electronic-structure theory, soft matter theory, microstructure research, and experimental techniques like photoemission, atom probe tomography, and electron microscopy. While the roadmap delves into specific areas within the broad interdisciplinary field of materials science, the provided examples elucidate key concepts applicable to a wider range of topics. The discussed instances offer insights into addressing the multifaceted challenges encountered in contemporary materials research.

Keywords: data, centric, materials, science, molecular simulations, roadmap

Contents

1. Introduction

Peter Benner[1], *Dierk Raabe*[2], *Jan-Michael Rost*[3], *Matthias Scheffler*[4] *and Gerhard Weikum*[5]
[1] Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany
[2] Max Planck Institute for Sustainable Materials, Düsseldorf, Germany
[3] Max Planck Institute for the Physics of Complex Systems, Dresden, Germany
[4] The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society, Berlin, Germany
[5] Max Planck Institute for Informatics, Saarbrücken, Germany

Introduction

Materials science and engineering play a pivotal role in fostering prosperity, enhancing lifestyle, and advancing the development of environmentally sustainable technologies. The field is profoundly interdisciplinary, encompassing physics, chemistry, biology, mathematics, and computer science. It addresses intriguing inquiries such as: are new semiconductors with increased efficiencies for solar modules available, and can they surpass the flexibility of materials under discussion today? Which catalyst materials would be optimal for a specific chemical reaction, e.g. splitting of water to produce hydrogen? What combination of alloying constituents imparts unique bending strength, extreme hardness, and corrosion-resistant properties of metallic alloys? Furthermore, how should a surface be coated to attain the utmost thermal protection, e.g. for improving the energy efficiency of turbines?

In recent years, materials science has entered an era marked by an unprecedented surge in data, stemming from both experiments and computations. This influx has surpassed the capacities of traditional methods to manage these data effectively. The so-called 4V challenge is clearly becoming eminent. It can be summarized as follows:

Volume: Addressing strategies to manage large datasets efficiently, exploring data storage solutions, and leveraging scalable technologies to handle voluminous data.

Variety: Discussing approaches to handle the diverse forms and meanings of data, including data normalization techniques and methods for dealing with heterogeneous datasets.

Velocity: Examining ways to cope with the rapid changes in data and the arrival of new datasets in real-time, emphasizing the importance of agile methodologies.

Veracity: Exploring methods to assess and enhance the quality and reliability of data, including data validation techniques, quality control measures, and uncertainty quantification (UQ).

Amidst these challenges, and most importantly, big data in materials science unveils extraordinary opportunities to advance scientific knowledge and to address important challenges like those noted above. To seize these opportunities, researchers must adopt fresh perspectives, innovative concepts, and novel methods. This paradigm shift, i.e. a new way of thinking, is commonly referred to as the 4th paradigm of materials research, a term made known by Jim Gray in his inspiring, final talk in 2007 [1]. In essence, 'data-centric research' and the '4th research paradigm' represent a departure from traditional research methodologies. It emphasizes the significance of correlations and statistical predictions, focusing on mean prediction values and variance (or uncertainty) as key elements in the investigative process. In this way the high intricacy of several co- and counter-acting processes is considered. It reflects that big data reveal correlations and dependencies that cannot be seen when studying small data sets, and, in difference to the past, it is accepted that a detailed causal explanation is not always possible. Causal inference, when possible, may not necessarily be expressed in terms of a simple, closed analytic equation or an insightful, simple physical model. We will get back to this point below.

Let us briefly recall the first three research paradigms. Experimental research, the initial paradigm, dates back to the Stone Age and developed first metallurgical techniques in the Copper and Bronze Ages. The control of fire marked a significant breakthrough. In the late 16th century, analytical equations became the central instrument for describing physical relationships, establishing theoretical physics as the second paradigm. The change was led by Brahe, Galileo, Kepler, and Newton. The next chapter started in the 1950s, when electronic-structure theory for solids [2, 3], the Monte Carlo method [4], and molecular dynamics [5, 6] were introduced. These developments enabled computer-based studies and analyses of thermodynamics and statistical mechanics on the one hand and of quantum mechanical properties of solids and liquids on the other hand. They define the beginning of computational materials science, what is nowadays considered the third paradigm of materials research.

Today, big data and artificial intelligence (AI) revolutionize various aspects of life, including materials science [1, 7, 8]. To navigate this 4th paradigm successfully, researchers must embrace new research concepts, and this Roadmap on Data-Centric Materials Science provides a summary of ideas for exploring the data-centric landscape of materials science and engineering. As materials science is a very broad and interdisciplinary field, only some areas of this landscape can be covered. However, we trust that the addressed examples explicate many of the basic concepts and that they can be helpful also for other topics than those addressed explicitly in the different contributions.

Science is and always has been based on data, but the terms ‘data-centric’ and the ‘4th paradigm’ of materials research signifies a transformative shift towards retrieving and managing vast data collections, digital repositories, and innovative data analytics methods. The integration of AI and its subset machine learning (ML) has become pivotal in addressing all these challenges. In the data analysis, we are looking for structures and patterns in the data. As mentioned above, materials properties and function are often not just governed by one single process but there are many. Some drive, others just facilitate, and again others hinder the materials property or function of interest. The interplay of various processes is very intricate. In analogy to genes in biology, we discuss elemental materials features (e.g. electronegativity of the atoms that build the material) that correlate with the materials property of interest. The primary features that connect with a certain materials property or function are called the relevant ‘materials genes’. Together with environmental parameters (e.g. temperature), they determine (in a statistical sense) the material’s property and function [9].

In recent years, major advances in ML and computing power, in particular the advance of hardware accelerators like graphical processing units (GPUs), have enabled deep neural networks (DNNs), with billions of trainable parameters, leading to breakthroughs in computer vision and natural language processing (NLP). A key strength of deep learning (DL) is that it addresses not only the objective for classification, regression or other tasks, but also the learning of how to represent the input data itself. Thus, there is no need for explicit feature modelling: images can be ingested as arrays of pixels, and text documents are simply sequences of tokens. High-level structures in visual or textual contents, like people interacting with objects in a scene or argumentation and sentiments in a conversation, are automatically discovered and latently captured by the DNN itself.

Obviously, this predictive methodology of DL has potential in many application areas, conceivably including materials science and particularly microscopy images. However, the success of DL builds on various assumptions, including the availability of large training data with 'independent and identically distributed' (iid) samples. These assumptions are not easily satisfied for materials data, and feature engineering and physics-based modelling is still indispensable (e.g. [10]).

At its core, ML operates as an interpolation technique, fitting and connecting the data upon which it is trained, applying regularization (or smoothening) to achieve generalization. The ML model excels in exploiting the data space covered by the training data but exhibits diminished reliability when entering uncharted data realms typically called the out-of-distribution (OOD) regime. When the training data are iid or representative of the full population, extrapolation may work. However, for materials science this requirement is hardly fulfilled, i.e. the data selection is governed by subjective and technical issues, and often it is strongly biased and unbalanced. Still, materials scientists are searching for statistically exceptional situations, and important processes are often triggered by 'rare events' that are not or not well covered by the available data set, or smoothed out by the regularization (e.g. [11]). This all implies caution when applying ML.

Similar to any scientific theory or model, an AI model possesses a range of applicability [12], often inadequately defined. Consequently, there is an argument advocating the importance of AI interpretability, as it not only sheds light on the underlying mechanism but also provides some confidence in extrapolations. The contributions by Boley *et al* (2.1), Ghiringhelli and Rossi (2.2), and Foppa and Scheffler (2.3) address these issues in more detail.

A special point in materials science is that data is typically not big. This implies that some ML methods are not suitable. In general, standard ML methods need to be used with caution and modification or new concepts have been and still need to be developed. Interestingly, Gaussian process (GP) regression and random forests (RFs) are still often and helpfully used, but several new concepts were established in recent years, e.g. crystal-graph neural network (GNN), message passing and equivariance, subgroup discovery (SGD), and SISSO (sure independence screening and sparsifying operator). In particular the latter can deal with correlations between a big (even immense) number of elemental materials features (millions or trillions) and just a few data dozens data points of the property of interest. SISSO derives an analytical equation for describing the materials property and its statistical correlation with the relevant materials genes. The approach as well as recent advancements, implementations, and challenges are described by Yao *et al* in contribution (3.1).

When data are scarce, the critical request is, that they must be highly accurate, precise, and well characterized. This is summarized by the request that experimental data must be 'clean', but it is not often achieved in materials science and rarely fulfilled in heterogeneous catalysis. The 'clean-data concept' for experimental studies is described in contribution (3.2) by Trunschke *et al.* Advancements in obtaining high-quality data from electronic-structure theory are described by Kokott *et al* in (3.3). The general challenge to find the best-suited AI method for a certain application is severe, and the reproducibility of published AI studies is often problematic. The NOMAD concept is described in contribution (3.4). A strategy to overcome the bottleneck of scarce data in DL is the augmentation of a small, accurate data set by synthetically generated data. This is discussed by Giri *et al* in contribution (3.5) and exemplified by generating synthetic Hamilton Matrices (SHMs) for DL applied to multiphotoabsorption. Spatiotemporal models like RFs and GPs have demonstrated promising outcomes in integrating data from multiple sources and guiding scientific discovery in various disciplines.

Contribution (3.6) by Xian *et al* discusses their application to materials science and hints at further directions to be explored to leverage their full potential in materials discovery. When trying to apply ML methods that have already proved successful in 'hard matter physics' to soft matter, several technical obstacles need to be overcome, including the intrinsic multiscale nature of this part of condensed matter. Bereau and Kremer argue that when this can be achieved, it would usher soft matter in a new era, where poor scale separation can be efficiently addressed, and insight will be gained for phenomena that are currently too complex for traditional methods (contribution 3.7). In contribution (3.8), Goyal *et al* show that significant computational gains can be achieved in the numerical simulation of microstructure continuum mechanics models when traditional direct numerical simulation is replaced by modern DL based methods when the AI models are informed by physical insight. Digitalizing the entire workflow in data-rich imaging techniques in material science from synthesis, sample preparation, data acquisition and post-processing in an integrated way is the topic of contribution (3.9) by Freysoldt *et al.* There, it is discussed that ML techniques can leverage the data science approach by removing the human inspection as the limiting factor to digest larger and larger amounts of data in order to discover relevant, but possibly rare patterns. Recently, large-language models (LLMs) have also entered the field of materials science. Raabe *et al* provide an overview and perspective in contribution (3.10).

Section 4 then addresses several applications of data-centric materials science, typically paired with methodological developments. Experimental methods cover photoemission, electron microscopy, and atom-probe tomography. In contribution (4.1), Purcell *et al* consider the role of AI in high-throughput materials discovery using computational workflows while Liebscher *et al* as well as Schloz *et al* discuss the roadmap to AI and ML driven data analytics in scanning transmission electron microscopy (STEM) in contributions (4.2) and (4.3), respectively. Atom probe tomography (APT) is another imaging-based technology to analyse the composition of materials at the near-atomic scale. Its enhancement using ML is the topic of contribution (4.4) by Li *et al.* In contribution (4.5), Logsdail *et al* investigate the potentials of a data-driven approach for heterogeneous catalysis. Finally, in contribution (4.6), Fratzl discusses recent advancements of x-ray scattering and diffraction for materials at the nanoscale with respect to the retrieval and analytics of large amounts of data.

Acknowledgment

Finally, we would like to express our gratitude for the various workshops, schools, contacts, and collaborations fostered by the Max Planck Society network BiGmax, which has been instrumental in shaping the subject of this roadmap over the past seven years. Our sincere appreciation goes to Rayya Douedari for her dedicated and invaluable support of this network, as well as her contributions to the meticulous preparation of this roadmap. Furthermore, we extend our thanks to Alex Wotherspoon from IOP Publishing for his skilled and amiable assistance throughout the editorial and publishing process.

2. Data and uncertainty

2.1. From prediction to action: critical role of performance estimation for ML-driven materials discovery

Mario Boley[1], *Felix Luong*[1], *Simon Teshuva*[1], *Daniel F Schmidt*[1], *Lucas Foppa*[2]
and Matthias Scheffler[2]

[1] Department of Data Science and AI, Monash University, Melbourne, Australia

[2] The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society, Berlin, Germany

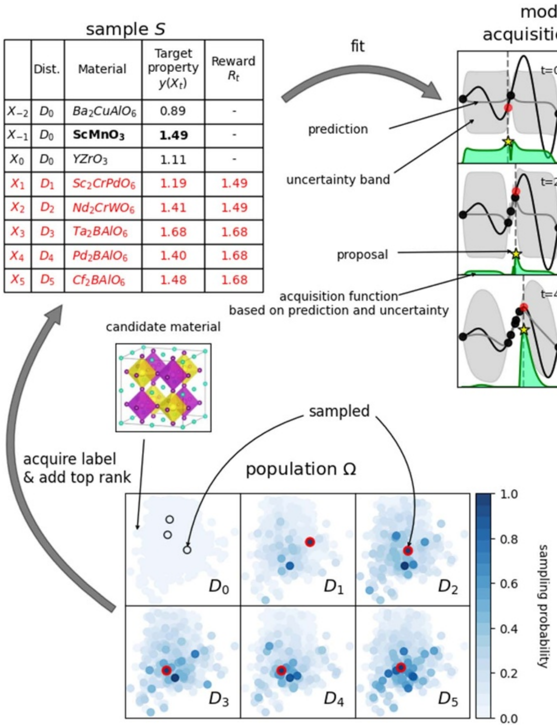
Status

In recent years, the materials science community has established a large-scale infrastructure for data sharing that promises to increase the efficiency of the ‘data-driven’ discovery of novel useful materials [13]. Growing data collections are envisioned to lead to increasingly accurate statistical models for property prediction that can significantly reduce the number of necessary experiments or first principles computations and, thus, substantially improve the cost and time for critical discoveries [14]. Indeed, the combination of public datasets and robust statistical estimation techniques like cross validation (CV) enables a collaborative improvement process (‘common task framework’ [15, 16]). As a result, there are now models that can predict certain materials properties well *on average* with respect to the same distribution as the training data. Unfortunately, the *in-distribution* expected performance, as estimated by CV, is not directly coupled with the performance for the discovery of novel materials: expected performance fails to capture the model behaviour for the very few exceptional materials that one aims to discover, and, fundamentally, in-distribution performance is irrelevant for a discovery process that is designed to generate high-performing materials more frequently than they occur in the initial training data.

Recognizing these issues, the community increasingly focusses on active learning approaches [17] like Bayesian optimization for model-driven blackbox optimization [18] (BBO). These methods manage an iterative modelling and data acquisition process and aim to optimize the cumulative ‘reward’ received for the acquired data points over time, such as the maximum property value discovered so far. This process, illustrated in figure 1, is enabled by an acquisition function that leverages the predictions of a statistical model together with its UQ to effectively manage the underlying trade-off of exploration (learning more about the candidate space) and exploitation (aim to sample high value candidates). This shift to consider actions instead of just predictions constitutes an important step towards accelerated materials discovery, but it reveals shortcomings not only in existing modelling approaches but more fundamentally in the methodological framework used to improve those models. In particular, the inapplicability of established performance estimation frameworks based on pre-generated data renders it extremely costly to conclusively compare and to systematically improve methods.

Current and future challenges

To illustrate these challenges, let us consider as example the discovery of double perovskite oxides with high *ab initio* computed bulk modulus, where we use two popular statistical models, GP regression and RF, and two BBO data acquisition strategies, *expected improvement* [19] (EI) of rewards and *pure exploitation* [20] (XT). GPs are the traditional BBO model,



	Dist.	Material	Target property $y(X_t)$	Reward R_t
X_{-2}	D_0	Ba_2CuAlO_6	0.89	-
X_{-1}	D_0	**$ScMnO_3$**	**1.49**	-
X_0	D_0	$YZrO_3$	1.11	-
X_1	D_1	Sc_2CrPdO_6	1.19	1.49
X_2	D_2	Nd_2CrWO_6	1.41	1.49
X_3	D_3	Ta_2BAlO_6	1.68	1.68
X_4	D_4	Pd_2BAlO_6	1.40	1.68
X_5	D_5	Cf_2BAlO_6	1.48	1.68

Figure 1. Schematic steps of iterative model-driven discovery process. At time t: (i) probabilistic property model is fitted to sample $\{X_{-N+1},\ldots,X_0;X_1;\ldots;X_{t-1}\}$ of materials population Ω, i.e. a conditional density function p(y|x) is learned that provides probability density of property value y for material x, which gives rise to both (mean) prediction $f(x)=\mathbb{E}_p(Y|X=x)$ and uncertainty (variance) $\sigma^2(x)=\mathbb{V}_p$ (Y|X=x) where expected value and variance are taken with respect to p; (ii) remaining population is ranked by acquisition function, e.g. 'expected improvement' of reward $a(x)=\mathbb{E}_p(R_t-R_{t-1}|X_t=x)$, which for conditionally normal property models can be computed as $a(x)=f(x)+\sigma^2(x)p(R_t|x)/(1-P(R_t|x))$ where P is the modelled cumulative distribution function; and (iii) label for top-ranked material is acquired and added to data sample generating reward, e.g. defined as $R_t=\max\{y(X_i):-N<i\leqslant t\}$ when maximizing a single property or figure of merit y, which incentivizes the discovery of materials with high y-value as early as possible in the process. While standard statistical analysis assumes the initial data points $X_{-N+1},\ldots,X_0$ to be drawn with respect to some sampling distribution D_0, this distribution does not have to be balanced or representative of the whole population. However, any concentration away from a representative, i.e. uniform, sampling distribution, poses the risk of delayed reward generation, and a misspecified acquisition function or model, in particular one with over-confident predictions, even risks to never escape local maxima represented in the initial data collection. The sampling distribution of subsequent points $D_1,D_2,\ldots D_T$ vary and depend on the combination of model p and acquisition function a. Hence, they cannot be pre-generated for new methods rendering label generation a key bottleneck in method development.

because their Bayesian approach provides a principled quantification of 'epistemic' uncertainty, i.e. uncertainty from a lack of training data related to a specific test point. However, they can struggle already with moderately high-dimensional representations such as the 24 features used in this example. In contrast, RFs are known to work robustly well with high-dimensional feature spaces [21], while their ensemble-based UQ does not represent epistemic uncertainty.

Interestingly, as shown in figure 2, CV indicates that RF has the better in-distribution predictive performance not only in terms of squared error but also in terms of log loss, which takes uncertainty into account. Nevertheless, RF is outperformed by GP in terms of the produced discovery rewards, demonstrating that standard in-distribution performance estimation techniques can suggest sub-optimal methods.

This demonstrates that already method selection is a real challenge for practical problems. However, the situation is much worse for methodological research that aims to not only determine, which of a small number of established methods works best, but to test dozens of combinations of models and acquisition functions. Absent innovation in performance estimation, comparing K methods in terms of their expected discovery reward across L repetitions of T rounds requires the acquisition of KLT labels in addition to any pre-generated initial data. This is because, even when starting from a common initial training distribution, each method produces its own sequence of proposal distributions. Since these distributions are unknown *a priori*, there is no way to pre-generate data from them, blocking the usual collaborative improvement process around an initially released dataset. Thus, the prohibitive cost of expected reward estimation currently blocks substantial progress in addressing other important challenges like unsound UQ or acquisition function optimization with infinite candidate populations, particularly when using non-invertible materials representations.

Advances in science and technology to meet challenges

Given these considerations, a central research goal should be to find reliable approaches for estimating a method's expected discovery reward based on existing data. A simple but infeasible state-of-the-art strategy is to run a method repeatedly using sub-samples of size n from the given dataset as initial data and the sub-sample complement as candidate pool, such that the ratio n/N is close to N/M where M is the overall population size. That is, one naively uses the initial dataset as proxy for the population. For at least two reasons, this simplistic approach is likely to produce misleading results (see figure 2, middle left). Firstly, the real rewards are determined by the exceptional materials in the tail of the target property distribution, which are almost certainly not well represented in the available dataset. Secondly, changing the absolute sizes of initial data and candidate population misestimates model performance and, more severely, misrepresents the real overwhelming number of uninteresting materials that an efficient search must largely avoid.

Here, we present an adjusted reward estimation approach that provides random initial and candidate sets with realistic absolute numbers of unrepresented exceptional materials as well as distinct ordinary materials to distract from them. Let $X_{(1)},\ldots,X_{(N)}$ denote the initial data elements in increasing order of their target property or figure of merit values. Based on an estimate $\hat{\alpha}$ of the unrepresented fraction of top materials $\alpha = \#\left\{X \in \Omega : y(X) > y(X_{(N)})\right\}/M$ create:

(1) **An initial dataset** by drawing a size-N bootstrap sub-sample [22], i.e. sample with replacement, from the low property value materials $X_{(1)},\ldots,X_{(N-\hat{\alpha}N)}$ and
(2) **A candidate set** consisting of an up-sampled and stochastically perturbed set $\tilde{X}_1,\ldots,\tilde{X}_{M-\hat{\alpha}M}$ from the unsampled (out-of-bag) elements of the bootstrap sample and an up-sampled and stochastically perturbed set $\tilde{X}_{M-\hat{\alpha}M+1},\ldots,\tilde{X}_M$ of the held-out top $\hat{\alpha}N$ materials.

The required α-estimate can be obtained via Monte Carlo simulations if the sampling distribution of the data or at least its level of concentration is approximately known. Alternatively,

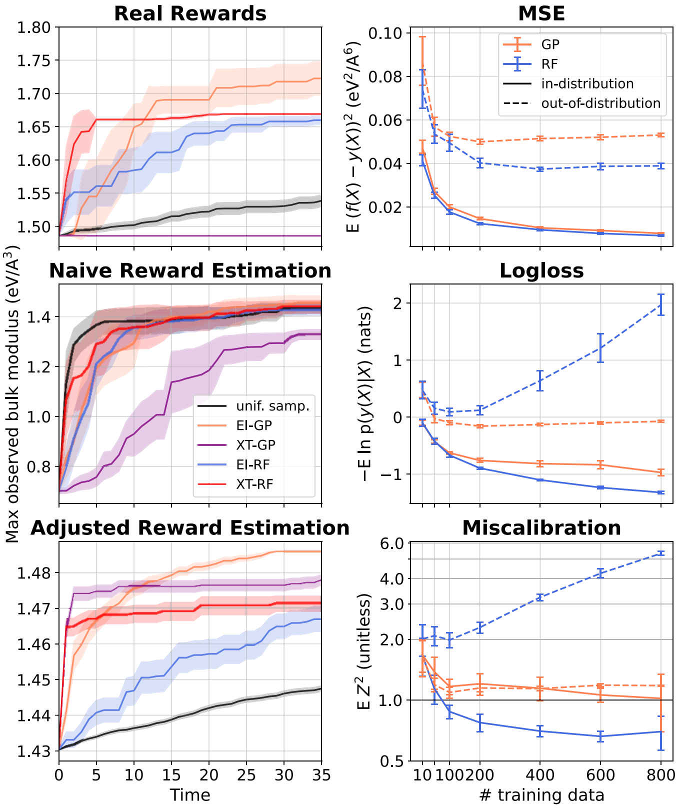


Figure 2. Performance of Gaussian process (GP) and random forest (RF) models for discovering double perovskites with high bulk modulus. Left column: Rewards generated by models with either expected improvement (EI) or pure exploitation (XT) acquisition function as well as their naïve and adjusted reward estimation using $\hat{\alpha} = 0.0294$ determined from a uniform population sample of size 100. Real rewards are mean rewards based ten repetitions (100 for uniform). Estimated rewards are the mean of 20 subsampling repetitions. All error bars correspond to 90% confidence intervals. GP with EI has the highest mean reward (1.657 eV A^{-3}) and discovers the highest bulk modulus (1.723 eV A^{-3} on average) in 35 rounds, which is qualitatively predicted by adjusted reward estimation. Right column: Model predictive performance estimates in terms of the mean squared error MSE $\mathbb{E}_D(y(X) - f(X))^2$ where $f(X)$ is the prediction for random input point X with property value $y(X)$, log loss $\mathbb{E}_D(\ln p(y(X) \mid X))$ where p is the modelled density of $y(X)$, and miscalibration indicator $\mathbb{E}_D Z^2$ with $Z = (y(X) - f(X))/\sigma(X)$ where $\sigma(X)$ is the modelled standard deviation of $y(X)$ given X and $Z^2 > 1$ and $Z^2 < 1$ correspond to over- and under-confidence, respectively. Here, all expected values refer to unknown true distributions estimated via 20 repetitions of sub-sampling with replacement from available data (i.e. bootstrap sampling). In-distribution performance is performance with respect to the initial sampling distribution D_0, out-of-distribution is with respect to the uniform mixture of the distributions D_1 to D_{30} of the data points examined by the various discovery processes. While RF provides a better mean squared error, both in- and out-of-distribution, its out-of-distribution log loss is increasing with the size of the training, indicating a failure of its uncertainty quantification.

one can obtain a relatively small uniform random sample of size U from the population minus the N previously sampled materials and, following a Bayesian estimation procedure for the success parameter of a binomial distribution, set $\hat{\alpha} = (C+1)/(U+2)$ where C is the number of elements in the uniform sample with a y-value greater than $y\left(X_{(N)}\right)$. As shown in figure 2 (bottom left), reward estimation with this approach performs much better than naïve estimation for our bulk modulus example. It accurately predicts GP with EI to produce the highest bulk modulus and highest cumulative reward out of the four candidate methods. Moreover, outside of GP with XT, which in the real experiment fails to produce any bulk modulus improvement, the adjusted reward estimation correctly predicts the relative order of all other methods. As desired, this is based entirely on the initially available data plus a small number of uniformly sampled data points without requiring the over thousand additional calculations that were needed to confirm this result.

Concluding remarks

The lack of reliable approaches to estimate expected discovery rewards from a given dataset is a serious roadblock for the development of active learning methods for materials discovery. Without such estimators, the evaluation of each candidate method requires the acquisition of a potentially large number of labels in addition to any initially available data collection, preventing the usual collaborative process that led to fast-paced improvements of predictive model performance with fixed distributions.

Naïve reward estimation from the initial data typically fails because of unsuitable data proportions and underrepresented extreme events. We presented an adjusted approach that, by correcting for these factors, successfully assesses which combination of acquisition function and statistical model works best for the exemplary task of double perovskite bulk modulus optimization. This or similar approaches could become efficiently computable proxies for real method performances and thus enable fast community-driven improvements to data-driven methods for materials discovery.

Acknowledgment

This work was supported by the Australian Research Council (DP210100045) and the ERC Advanced Grant TEC1p (European Research Council, Grant Agreement No. 740233).

2.2. Reliable quantification of uncertainties: the biggest challenge for data-centric materials modelling?

Luca M Ghiringhelli[1,2] *and Mariana Rossi*[3]

[1] Department of Materials Science and Engineering, Friedrich-Alexander Universität, Erlangen-Nürnberg, Germany

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

[3] Max Planck Institute for the Structure and Dynamics of Matter, Hamburg, Germany

Status

AI and, in particular, ML modelling is substantially increasing the reach and predictive power of material-science simulations. Such strategies are adopted for two broad classes of applications: (a) surrogate modelling of materials properties, e.g. learning energies and forces of given atomic configurations, where the Hamiltonian is known but computationally intensive to evaluate (references [23, 24] and references therein), and (b) materials genomics, i.e. the identification of the features that can explain and be used to model certain materials' property (the genes for that material and property), together with fitting of a predictive model for the given property as function of the identified genes (reference [25] and references therein).

Often, the performance of predictive models is focused on averages (e.g. the mean absolute error), and little attention is given to the distribution of errors (e.g. via the so-called violin plots) and to the inspection of the outliers, i.e. the data points that yield the largest prediction errors. Are these data points simply wrongly measured or could they herald some different physical mechanism that was not captured by the model trained to yield acceptable average errors?

Scientifically, it is equally important for a ML model to yield predicted values for new data points and, concurrently, provide reliable UQ. In other words, the model should be able to recognize if it can make a confident prediction solely from the input representation of a test data point, identifying whether it is similar to the data points used for training (interpolatory regime) or dissimilar (extrapolatory regime). The correct metric for assessing this similarity is, however, most often unknown and systematically finding it for a given ML model is one of the most difficult steps for a reliable uncertainty estimate.

Several strategies have been developed for UQ, spanning from rigorous and computationally extremely expensive Bayesian estimates to pragmatic ensemble-of-models training [26–28]. However, many such estimates have been shown to be overconfident when test data are drawn far from the sampling distribution of the training data [29–31]. This limitation represents a serious drawback for the overall reliability of ML models in atomistic simulations, where they promise to deliver first-principles quality results.

Current and future challenges

Besides the obvious intrinsic benefit of reliably quantifying the uncertainty of an ML model, these estimates are also a vital part of the so-called active-learning (AL) algorithms. AL denotes a strategy where the model constructs new (training) data points either in regions where a property of interest needs to be optimized (exploitation task) or in regions where the model uncertainty is large (exploration task), resulting in a more accurate model with a lower amount of training points. In material science, these algorithms are often desirable, because little initial information is known about a material or materials class and calculating labels (properties) is expensive.

In view of the exploitation task, it is desirable to adopt model classes that allow for a computationally inexpensive optimization (e.g. GPs). However, the biggest challenge in both surrogate modelling and materials genomics is the UQ in extrapolative regions for the exploration task. In practice, recognizing that a data point belongs to the extrapolation region is the actual conundrum. Statistics and information-theory modelling approaches rely on the fact that training data are representative of the overall population where predictions will be made. In both surrogate modelling and materials genomics applications, the unseen data may carry physical information that is not present in the model training. Electronic-structure data carries a further challenge due to its intrinsic aleatoric uncertainty stemming from numerical convergence and basis sets. It is often difficult, but necessary, to separate it from the model (epistemic) uncertainty, for defining whether training data refinement is needed or whether the model can be really improved.

An aspect that cannot be disregarded in this discussion is that the definition of the metrics for UQ is not uniform across different studies. These metrics differ on their sensitivity to outliers and performance with respect to estimating true errors [32]. Systematic testing of these metrics over a wide range of materials and properties is not yet available to the community. This hinders further progress in the field and should be urgently tackled by the community. As for any physical modelling, one does not expect a model to be predictive outside its physical scope. Yet, in the traditional development of physical theories (sometimes referred to as 'model-based', as opposed to 'data-centric', approach) *describing* the limit of validity of a theory is an essential part of it. Such limits of validity are typically expressed as inequalities as function of key parameters governing the physical property or process. We identify the data-centric identification of the limits of validity of an ML model as, arguably, the biggest challenge in AI applied to materials science.

Advances in science and technology to meet challenges

The full acceptance of ML tools within the community, for both surrogate modelling and materials genomics, may depend on two interrelated aspects: The introduction of algorithms for (a) reliable UQ, especially for data points that are outside the training distribution and (b) finding *explanations* why any given outlier is an outlier.

For the first aspect, in the realm of surrogate model potentials, Bayesian-based frameworks offer an intrinsic definition of uncertainty, which can be judiciously used [33]. For neural-network architectures, committee ensemble models can deliver some degree of uncertainty prediction. In both cases, correctly accounting for correlations in the training set data is essential for avoiding overconfident model predictions [34], but UQ can still be unreliable for out-of-sample data points. A promising alternative is the use of deep ensembles or variations thereof. Finally, because the surrogate model is trained to predict energy and forces, but these quantities are almost never the observable that is being sought in a simulation, advances in error definition and propagation through derived properties have been gaining much attention [35].

For the second aspect, a promising route is the use of SGD for the identification of the so-called domains of applicability (DAs, regions of the input space where a predictive model yields small errors) [12], which are given in form of descriptive rules, i.e. inequalities over a set of features, identified among a larger set of candidates. Although it has been shown that DAs can be found and the descriptive rules give insight on the analysed ML models, the method has not been yet further developed to systematically identify outliers and exploited to improve the underlying ML model, e.g. in an AL fashion.

Concluding remarks

The recent literature has shown that, with carefully selected training data sets and physical expertise (domain knowledge), the resulting ML predictive models allow for important discoveries in materials science. However, unleashing the full potential of data-centric approaches and fulfilling their promise to deliver results of *ab initio* quality requires that the uncertainty of the predictions be quantified. This UQ needs to be robust and reliable and the related algorithm should be relatively straightforward to implement, such that users have a transparent access to it.

Although reliability has to be prioritised, any UQ algorithm must not add a substantial computational cost to the ML model it is being applied to, since in materials modelling efficiency is often a core requirement to achieve meaningful simulations. This observation applies both to the realm of surrogate modelling where, e.g. millions of force evaluations with UQ need to be carried out, and to the realm of materials genomics where, e.g. millions of candidate systems need to be classified including this quantification. Achieving such a framework requires the community to adopt more widespread standards and work together on benchmarking efforts targeted at error prediction.

Reaching this goal would enable the systematic, fully data-centric improvement of the learned model, via the AL strategies, and the assessment of the limits of validity of the learned models.

Acknowledgment

LMG acknowledges funding from the NOMAD Center of Excellence (European Union's Horizon 2020 research and innovation program, Grant Agreement No. 951786) and the project FAIRmat (FAIR Data Infrastructure for Condensed-Matter Physics and the Chemical Physics of Solids, German Research Foundation, Project No. 460197019). We acknowledge financial support from BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials-Science. We thank Matthias Scheffler for a critical read of the earlier version of this manuscript.

2.3. Towards a multi-objective optimization of subgroups for the discovery of materials with exceptional performance

Lucas Foppa and Matthias Scheffler
The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society, Berlin, Germany

Status

AI approaches in materials science usually attempt a description of all possible scenarios with a single, global model. However, the materials that are useful for a given application, which requires a special and high performance, are often statistically exceptional. For instance, one might be interested in identifying exceedingly hard materials, or materials with band gap within a narrow range of values. Global models of materials' properties and functions are designed to perform well in average for the majority of (uninteresting) compounds. Thus, AI might well overlook the useful materials. In contrast, SGD [36, 37] identifies *local* descriptions of the materials space, accepting that a global model might be inaccurate or inappropriate to capture the useful materials subspace. Indeed, different mechanisms may govern the materials' performance across the immense materials space and SGD can focus on the mechanism(s) that result in exceptional performance.

The SGD analysis is based on a dataset $\tilde{P}$, which contains a known set of materials. $\tilde{P}$ is part of a larger space of possible materials, the full, typically infinite population P. For the materials in $\tilde{P}$, we know a target of interest Y (metric or categorical), such as a materials' property, as well as many candidate descriptive parameters φ possibly correlated with the underlying phenomena governing Y (figure 3). From this dataset, SGD generates propositions π about the descriptive parameters, e.g. inequalities constraining their values, and then identifies selectors σ, conjunctions of π, that result in SGs that maximize a quality function Q:

$$Q\left(\tilde{P},SG\right)=\left(\frac{s_{SG}}{s_{\tilde{P}}}\right)^{\gamma}*\left(u\left(SG,\tilde{P}\right)\right)^{(1-\gamma)}. \tag{1}$$

In equation (1), the ratio $s_{SG}/s_{\tilde{P}}$ is called the coverage, where s_{SG} and $s_{\tilde{P}}$ are the number of data points in the SG and in $\tilde{P}$, respectively. The utility function $u(SG,\tilde{P})$ measures how exceptional the SGs are compared to $\tilde{P}$ based on the distributions of Y values in the SG and in $\tilde{P}$. Q establishes a tradeoff between the coverage (generality) and the utility (exceptionality), which can be tuned by a tradeoff parameter γ. Typically, the identified selectors only depend on few of the initially offered candidate descriptive parameters. The identified SG selectors (or rules) describe the local behaviour in the SG and they can be exploited for the identification of new materials in P.

Current and future challenges

The potential of SGD to uncover local patterns in materials science has been demonstrated by the identification of structure-property relationships [38], and by the discovery of materials for heterogeneous catalysis [39]. Additionally, using (prediction) errors as target in SGD, we identified descriptions of the regions of the materials space in which (ML) models have low [12] or high errors [40]. Thus, the domain of applicability (DoA) of the models could be established. Despite these encouraging results, the advancement of the SGD approach in materials science requires addressing key challenges:

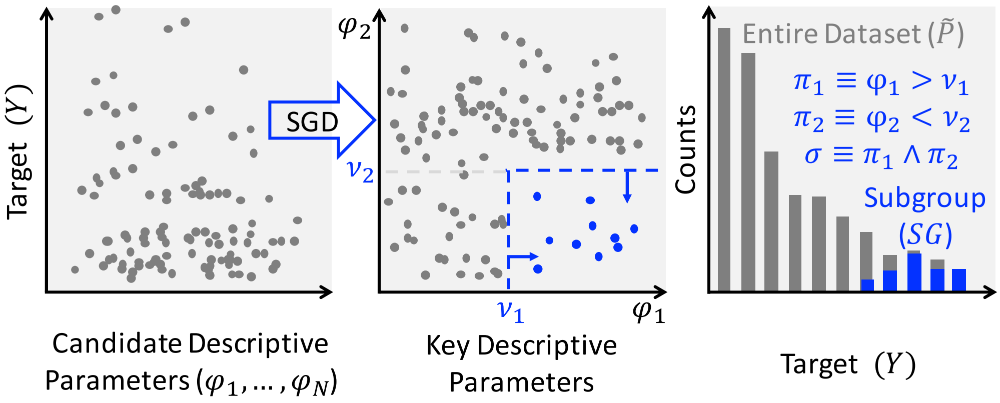

Target (Y)
Candidate Descriptive Parameters ($\varphi_1, \ldots, \varphi_N$)
SGD
φ_2
ν_2
ν_1
φ_1
Key Descriptive Parameters
Counts
Entire Dataset ($\tilde{P}$)
$\pi_1 \equiv \varphi_1 > \nu_1$
$\pi_2 \equiv \varphi_2 < \nu_2$
$\sigma \equiv \pi_1 \wedge \pi_2$
Subgroup (SG)
Target (Y)

Figure 3. Subgroup discovery (SGD) identifies descriptions of exceptional subselections of the dataset. These descriptions (rules) are selectors σ constructed as conjunctions of propositions π about the data. The symbol $\wedge$ denotes the 'AND' operator.

- The quality function introduces one generality-exceptionality tradeoff, among a multitude of possible tradeoffs that can be relevant for a given application and that can be obtained with different γ. For instance, the required hardness of a material depends on the type of device in which it will be used and the DoA of a model depends on the accuracy that is acceptable to describe a certain property or phenomenon. However, choosing the appropriate γ and assessing the similarity—or redundancy—among the multiple rules obtained with different tradeoffs are challenging tasks.
- Widely used utility functions assess the exceptionality of SGs by comparing the data distribution of the SG and that of $\tilde{P}$ via a single summary-statistics value. For example, the positive-mean-shift utility function for metric target favours the identification of SGs with high Y values only based on the means of the two distributions. Thus, it is often assumed that the distributions are well characterized by the chosen summary-statistics value and that $\tilde{P}$ is representative of the full population P. However, distributions in materials science are typically non-normal and $\tilde{P}$ might not reflect the infinitely larger, unknown P. This calls for the consideration of utility functions that circumvent the mentioned assumptions.
- The mechanisms governing materials can be highly intricate and the relevant descriptive parameters to describe a certain materials' property are often unknown. Thus, one would like to offer many possibly relevant candidate parameters and let the SGD analysis identify the key ones. However, optimizing the quality function is a combinatorial problem with respect to the number of descriptive parameters and efficient search algorithms are therefore crucial [41].

Advances in science and technology to meet challenges

In order to address some of these open questions, we approach the SGD as a multi-objective-optimization problem for the systematic identification of SG rules that correspond to a multitude of generality-exceptionality trade-offs. Coherent collections of SG rules are obtained by considering the Pareto front of optimal SGD solutions with respect to the objectives coverage and utility function, as illustrated for the example of identification of perovskites with high bulk moduli in figure 4. Once the coherent collections of SG rules are identified, the overlap

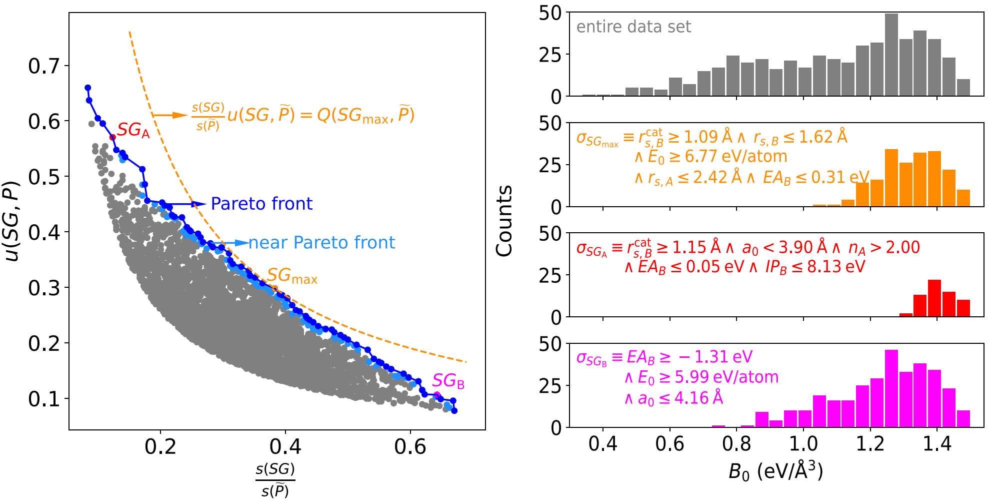


Figure 4. Left panel: a coherent collection of SG rules describing ABO_3 perovskites with high bulk modulus (B_0) is identified at the Pareto front of SGD solutions with respect to the objectives coverage and the utility function *cumulative Jensen–Shannon divergence*. Right panel: the identified rules constrain the values of the radiii of the s orbitals of isolated A, B and B^{+1} species ($r_{s,A}$, $r_{s,B}$ and $r^{\mathrm{cat}}_{s,B}$, respectively), the electron affinity and ionization potential of isolated B species (EA_B and IP_B, respectively), the expected oxidation state of A (n_A), the equilibrium lattice constant (a_0), and the cohesive energy (E_0).

between SG elements can be used to assess their similarity. A high similarity between SG rules might indicate that the rules are redundant. Thus, the similarity analysis can be used to choose the SG rules that should be considered for further investigation or exploitation.

Noteworthy, the *cumulative Jensen–Shannon divergence* (D_{JS}) [42] between the distribution of bulk moduli in the SG and in the entire dataset is used as quality function in the example of figure 4. D_{JS} assumes small values for similar distributions and increases as the distribution of target values in the SG is, e.g. shifted or narrower with respect to the distribution of the entire dataset. Crucially, D_{JS} does not assume that one single summary-statistics value represents the distributions. Divergence-based utility functions addressing, e.g. high or low target values, will thus be an important advance. We note that the utility function might also incorporate information on multiple targets or physical constraints that are specific to the scientific question being addressed [43]. However, in order to ensure that the training data is representative of the relevant materials space one would like to cover, the iterative incorporation of new data points and training of SGD rules in an AL fashion might be required.

Concluding remarks

SGD can accelerate the identification of exceptional materials that may be overlooked by global AI models because it focuses on local descriptions. However, further developments are required in order to translate the SGD concept to the typical scenario of materials science, where datasets might be unbalanced, or not be representative of the whole materials space and the most important descriptive parameters are unknown. The multi-objective perspective introduced in this contribution provides an efficient framework for dealing with the compromise between generality and exceptionality in SGD. The combination of this strategy with efficient algorithms for SG search and with a systematic incorporation of new data points to better cover the materials space will further advance the AI-driven discovery of materials.

Acknowledgment

This work was funded by the NOMAD Center of Excellence (European Union's Horizon 2020 research and innovation program, Grant Agreement No. 951786) and by the ERC Advanced Grant TEC1p (European Research Council, Grant Agreement No. 740233).

3. Methods

3.1. Building portable AI software for the exascale

Yi Yao[1], *Thomas A R Purcell*[1,2], *Sebastian Eibl*[3], *Markus Rampp*[3], *Luca M Ghiringhelli*[1,4] *and Matthias Scheffler*[1]
[1] The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society, Berlin, Germany
[2] The Department of Chemistry and Biochemistry, University of Arizona, Tucson, AZ, United States of America
[3] Max Planck Computing and Data Facility, Garching, Germany
[4] Department of Materials Science and Engineering, Friedrich-Alexander Universität, Erlangen-Nürnberg, Germany

Status

Modern high-performance computing (HPC) systems are evolving towards greater heterogeneity and diversification. The heterogeneity is due to the use of specialized processing units for specific tasks, nowadays with a strong (commercial) focus on AI-specific algorithms. This strategy, led by companies like Nvidia with their (general-purpose) GPUs and tools like CUDA, is driven by the need to enhance computational performance while containing electrical-power consumption and total cost. Present-day exascale and pre-exascale systems commonly integrate GPUs with CPUs of different architectures and vendors. Additionally, alternative accelerators like tensor processing units, neural processing units, field programmable gate arrays, and emerging technologies like neuromorphic and quantum processors add to the array of HPC options. These will further contribute to the heterogeneity and diversification of HPC but have not yet broken into scientific computing. Except for the quantum processor, the other accelerators adhere to classical architectures characterised by varying levels of parallelism.

To tap the power of accelerators, AI codes must incorporate efficient internode communication schemes (like the well-established message passing interface (MPI)) and align with programming models associated with the available accelerators. Examples include CUDA for Nvidia GPUs, ROCm for AMD GPUs, or SYCL/DPC++ for Intel GPUs. The neural network-based AI codes often rely on the availability and development of frameworks such as pyTorch and tensorflow, where the developers of these frameworks take the burden to adapt the framework to accelerators. For instance, pytorch provides versions of its framework with support for CUDA or ROCm backends. However, not all of the AI methods can be seamlessly translated into a neural-network representation and not all applications are well suited for neural networks. Consequently, significant adaptation is required, leading to limited accelerator support. For example, the widely used decision-tree-based AI library, XGBoost, offers a CUDA version, but is still lacking a ROCm equivalent.

Current and future challenges

The current challenge involves developing performance-portable and maintainable code for AI methods, in general, on HPC systems. This task will become even more challenging with the increasing heterogeneity of HPC systems. In a typical HPC system, internode communication is necessary and the MPI has proven to be a flexible and effective solution for doing this. The

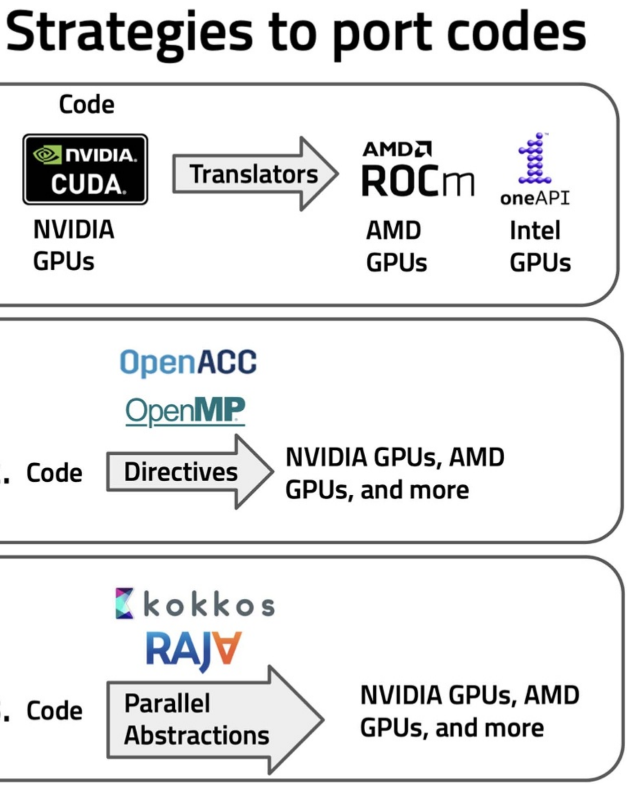


Figure 5. Strategies to port codes onto diversified and heterogeneous high-performance computers. Translators convert from one programming model to another, directives are compiler instructions to dictate how a piece of code should be compiled, and parallel abstractions define how a computation workload may be calculated in parallel. The library will then map the abstractions to GPUs.

so-called MPI + X paradigm combines MPI with intra-node parallelization models and/or accelerator offloading models (X). The choice of accelerator offloading model is largely determined by the specific accelerators in use, together with problem requirements and personal taste.

We summarise various strategies that have been developed to address this challenge in figure 5. For offloading work onto an accelerator, the most straight-forward approach would be to write the algorithms with accelerator-specific interfaces such as CUDA. While this in principle allows to tap all (performance) capabilities, these interfaces are limited to specific accelerators. Given the abundance of existing CUDA code in scientific computing, AMD and Intel have introduced tools to facilitate the translation of such code into their HIP/ROCm and SYCL/DPC++ language, which more or less resemble the semantics of CUDA. Moreover, the HIP and SYCL programming models even claim some universality by enabling code execution not only on AMD or Intel GPUs, respectively. However, the viability and broader adoption of these comparably recent approaches remains to be demonstrated.

An alternative approach involves the utilisation of architecture-independent—and typically more abstract—programming models, which come in various forms. One category employs compiler directives to manage loop parallelization and data management. Examples include OpenMP [44] and OpenACC [45]. Programming with these directives aims at a single codebase compatible with different accelerators. Directive-based approaches can also facilitate the reuse of existing CPU-based code and enable an incremental code-porting workflow by successively ‘offloading’ performance-critical parts of the code. Success-stories have been observed adapting these models [46].

Another approach are C++ portability frameworks such as Kokkos [47] and RAJA [48]. They provide high-level parallel abstractions such as the parallel implementation of the traditional ‘for’, ‘reduce’, and ‘scan’ operations, which the framework maps to specific hardware backends that use the corresponding platform-native programming models. These may, in addition, serve as forerunners for corresponding extensions to be added to the C++ standard.

Advances in science and technology to meet challenges

As an example of how the code-portability challenge can be met for an originally developed AI application which is different from DL, we outline the porting of an implementation of the sure independence screening and sparsifying operator (SISSO) [207] to GPUs using Kokkos. SISSO is a combination of symbolic regression and compressed sensing. It first generates a list of up to trillions of analytical expressions from an initial set of primary features and mathematical operators. It then uses an ℓ_0 regularised least squares regression to find the best low-dimensional linear model from the generated expressions. In preparation for (pre-)exascale computing, we converted the most computationally intensive components of SISSO, i.e. expression generation and ℓ_0 regularisation, in our initial MPI + OpenMP code to a MPI + OpenMP + Kokkos implementation, in order to demonstrate scalability and portability on exascale-ready HPC platforms.

Throughout the development, we refactored the data structures to suit the access pattern of accelerators, and carefully optimised the memory migration between host and device. This results in an approximately tenfold speedup by the GPUs of two generations of Nvidia GPUs for a test problem with $\sim$60 billion generated features and $\sim$36 billion least squares regression problems (see figure 6). The code also scales to at least 64 nodes, see figure 6. We expect that scaling to much higher node counts can be achieved with increasing the size of the training dataset. Notably, the same code also runs on AMD Instinct MI200 GPUs with a similar speedup without any code modifications, except for compilation settings. Given that the Kokkos framework supports backends for CUDA, HIP, SYCL, OpenMP, we expect our code can also be smoothly ported to other accelerators. Since Kokkos is developed and maintained with strong commitments by the US DOE laboratories, we expect it to receive continuous support and will extend to future HPC hardware.

One key question when using an abstraction framework is how close its performance comes to the ‘native’, i.e. architecture-specific programming models. In our case, we compared the performance of our batched least-squares-regression algorithm (for ℓ_0 regularization) to a native CUDA implementation co-developed with Nvidia engineers. This new CUDA code is about twice as fast as the Kokkos version. However, it is worth noting that Kokkos’ continuous development is promising. For instance, during our development, transitioning from Kokkos version 3 to version 4 resulted in a 10% speedup without requiring any code modifications from us in the application code.

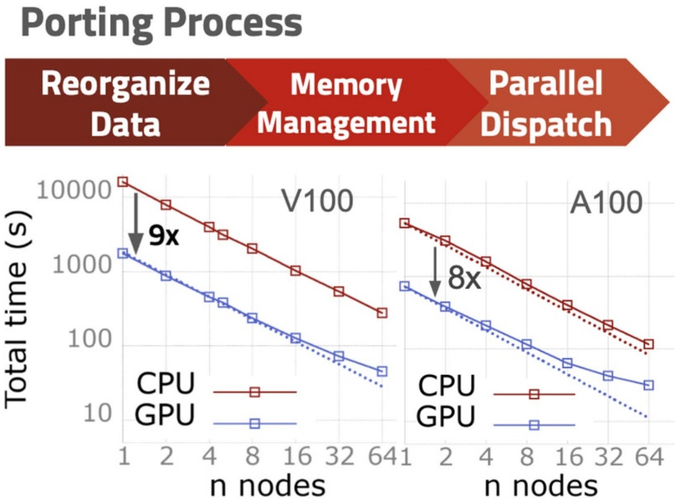


Figure 6. The process and performance of porting SISSO++ on different HPC platforms with Kokkos library. The scaling test is performed on (1), the Talos cluster at MPCDF up to 64 nodes with 40 Intel Skylake CPU cores and 2 Nvidia V100 GPUs on each node, and (2), the Raven cluster at MPCDF, on up to 64 nodes with 72 Intel Xeon Icelake CPU cores and 4 Nvidia A100 GPUs on each node.

Concluding remarks

The growing diversity and heterogeneity in (pre-)exascale HPC poses significant challenges to software developers, including performance portability and code maintainability. To tackle these issues, developers have adopted various strategies, such as code duplication (typically abstracted internally by some application-specific interfaces), (semi-)automatic code translation, directive-based portability models, and high-level abstraction frameworks. For our SISSO++ code [208], which is an AI application not readily amenable to the well-established (and portable) AI frameworks like, e.g. Tensorflow, we opted for the MPI + X paradigm which is well established in HPC, specifically using MPI + OpenMP + Kokkos. The usage of the Kokkos abstraction framework enhances both the code performance and portability, and it also helps reduce code-maintenance burdens. The Kokkos framework is also expected to pave the way for adopting the parallel abstract concepts in future C++ language standards. Due to the generality of the Kokkos framework, and already proven for SISSO++ by a seamless transition over two generations of Nvidia GPUs, we anticipate that our SISSO++ code will easily adapt also to future HPC architectures. Our porting strategy outlined here can serve as an example for other non-neural network based AI code development efforts.

Acknowledgment

We thank Patrick Atkinson and Markus Hrywniak from Nvidia for their support in optimising the CUDA-native reference implementation for our benchmark. The project received financial support from BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials Science, the NOMAD Center of Excellence (European Union's Horizon 2020 research and innovation program, Grant Agreement No. 951786) and the ERC Advanced Grant TEC1p (European Research Council, Grant Agreement No. 740233).

3.2. Clean-data concept for experimental studies

Annette Trunschke[1], *Lucas Foppa*[2] *and Matthias Scheffler*[2]
[1] Department of Inorganic Chemistry, Fritz Haber Institute of the Max Planck Society, Berlin, Germany
[2] The NOMAD Laboratory at the Fritz Haber Institute of the Fritz Haber Institute of the Max Planck Society, Berlin, Germany

Status

Materials design typically targets an application that requires the synthesis of a material which is characterised by measurable and reliable properties and functions that are maintained during its use. Inexpensive and abundant raw materials, reproducibility, and scalability are decisive factors for success. The relationships between the structure and the function of a material are usually complex and intricate and they prevent a strictly *in-silico* design for realistic conditions. Thus, the experimental input is crucial. AI methods have the potential to reduce the significant efforts related to the synthesis and characterization of materials, accelerating materials discovery. However, rigorously conducted experiments that provide consistent training data for AI are indispensable. They directly determine the reliability of generated insights.

The applications of AI in materials science are diverse [49, 50]. For example, the optimization of synthesis and functional properties in high-throughput experiments requires mathematical models which are iteratively trained in order to ensure an efficient experimental design. The elucidation of materials structures can be facilitated by AI. Besides, new materials can be predicted via the identification of correlations and patterns in experimental and computational data sets. This leads to a variety of data set structures. The interdisciplinary nature of materials science and the multitude of experimental techniques applied produce a broad spectrum of data formats, all of which can ultimately be traced back to spectroscopic, thermodynamic or kinetic relationships and are already standardised to some extent. Experimental data in materials research are usually not ‘big data’, which places additional demands on the methods of data analysis.

However, if the data becomes FAIR [51], i.e. findable, accessible, interoperable, and reusable (FAIR) and open, i.e. generally accessible after publication, machines can systematically analyse this information beyond the boundaries of a single laboratory and field of research, learn from it and develop disruptive solutions [13]. A particularly sustainable generation of insight is achieved through the use of interpretable AI algorithms that uncover descriptors, i.e. correlations between key physical parameters and the material properties and functions.

Current and future challenges

Predictions could be more reliable if the materials function of interest was determined exclusively by the bulk properties of the material. However, when the material’s function is affected, or even governed by interfacial and kinetic phenomena, such as in case of batteries, sensors, biophysical applications or catalysis, the relationships between the materials parameters and the function become extremely complex. On the one hand, this is caused by the strong influence of defects and minor impurities. On the other hand, the material properties respond to the fluctuating chemical potential of the environment in which they are used. This gives metadata such as the sequence of experimental steps and the time frame a particular importance [52].

In order to make experimental data useful for a digital analysis, the measurements have to follow so-called 'standard operating procedures' (SOPs), as is already common practice in some research areas. An important cornerstone for such workflows is the introduction of certified standards that enables the direct evaluation of measured data when they are published together with the results of the standard. Awareness of the need for rigorous work and standardization of experiments has grown considerably in academic research in recent times and it is reflected in initiatives for standardized measurement procedures and test protocols ([13] and references cited therein).

The currently most common form of publication in scientific journals does not support the direct electronic access to the data. The use of NLP tools is one approach to analyse and understand human language in published articles [53]. These computer science techniques can help to identify trends, but do not provide consistent data sets, as data in publications are not presented uniformly, for example often only in the form of graphical representations, and data as well as metadata are not necessarily provided completely.

The most effective solution to enable the use of experimental data in AI is to apply machine-readable SOPs in automated experiments. In this way, standardized and complete data and metadata sets can be generated that can be shared after publication in repositories, as is already widely done in computational materials science and synthesis of complex molecules. The latter also requires the development of ontologies. We note that digital SOPs are an important preliminary step for enabling autonomous research by robots in the future [54].

Advances in science and technology to meet challenges

The most common AI methods require large amounts of data, and only the smallest part of available data in materials science meets the quality requirements for data-efficient AI. In a use case study [55], we have shown how a 'clean' data-centric approach in interfacial catalysis enables the identification of descriptors based on a data set that can be generated in the experimental practice with reasonable effort (figure 7). Here the term 'clean data' refers to the fact that the considered materials were carefully synthesized, characterized, and tested in catalysis according to SOPs reported in an experimental handbook [52].

Large-scale applications in the field of energy storage such as water splitting and the efficient use of resources in the production of consumer goods are generally based on highly complex catalysed reactions at interfaces. The selective oxidation of the short-chain alkanes ethane, propane and *n*-butane to valuable olefins and oxygenates was chosen as an example of a reaction type that is known for its complicated reaction networks. Control over the selective formation of desired products in this network and the minimization of CO_2 formation requires sophisticated catalyst materials and adapted reaction conditions.

Experimental procedures that capture the kinetics of the formation of the active phase from the catalyst precursors have been designed and specified in a SOP [52]. A typical set of 12 chemically and structurally diverse catalyst materials was included in the study that combines rigorously conducted clean experiments in catalyst synthesis, physicochemical characterization and kinetic evaluation with interpretable AI using the SISSO symbolic regression approach [9, 207]. Previously obtained empirical findings are correctly reflected by the data analysis, which proves the value of the data set.

Interpretable AI goes far beyond empirical interpretations. It addresses the full complexity of the dynamically changing material and the full catalytic process by identifying non-linear property-function relationships described by mathematical equations in which the target catalytic parameters depend on several key physicochemical parameters of the material measured

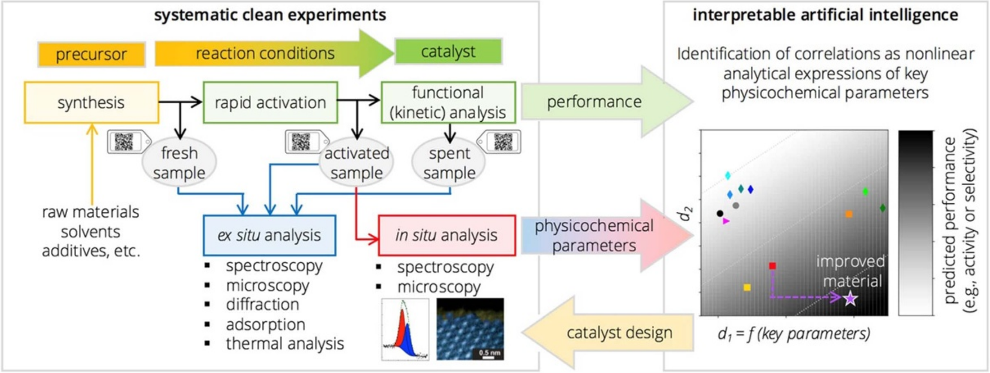

systematic clean experiments
precursor
reaction conditions
catalyst
synthesis
rapid activation
functional (kinetic) analysis
fresh sample
activated sample
spent sample
raw materials solvents additives, etc.
ex situ analysis
spectroscopy
microscopy
diffraction
adsorption
thermal analysis
in situ analysis
spectroscopy
microscopy
0.5 nm
performance
physicochemical parameters
catalyst design
interpretable artificial intelligence
Identification of correlations as nonlinear analytical expressions of key physicochemical parameters
d_2
improved material
$d_1 = f$ (key parameters)
predicted performance (e.g., activity or selectivity)

Figure 7. Clean experiments designed to capture the kinetics of the formation of the catalyst active state under reaction conditions are used to generate a consistent and detailed data set, which is then analysed via the sure-independence-screening-and-sparsifying-operator (SISSO) artificial intelligence (AI) approach in order to uncover the key physicochemical parameters describing the catalytic performance. Reproduced from [55]. CC BY 4.0.

in operando and after different stages in the life cycle of the catalyst. These key descriptive parameters, that the AI approach identifies out of many offered ones, reflect the processes triggering, favouring or hindering the catalytic performance. In analogy to genes in biology, these parameters are called ‘materials genes’ of heterogeneous catalysis, since they describe the catalyst function similarly as genes in biology and relate, for instance, to the colour of the eyes or to health issues. Thus, these materials genes capture complex relationships. They describe a correlation (with uncertainties) but they do not provide the detailed description of the underlying processes.

This data-centric approach discloses, which of the often time-consuming and expensive characterization techniques are important for the catalyst design. The chemist is also provided with practical guidelines for optimizing certain material properties to further improve the catalysts function.

Concluding remarks

Reproducibility is probably the most basic and crucial requirement of materials science. AI is an efficient tool in materials research and development, but its application requires that we change the way we work and deal with data. Complete, uniform and reliable data sets are required that comply with the FAIR principles. These can be obtained by working across laboratories according to SOPs (‘handbooks’), which also include the analysis of benchmarks. Important elements for the gradual development of autonomous materials research [56], in addition to technical progress in robotics, are the use of machine-readable handbooks, automated experiments with standardized data analysis and upload to local data infrastructures as well as the standardized publication of experimental data in overarching open repositories.

Acknowledgment

This project received financial support from BiGmax, the Max Planck Society’s Research Network on Big-Data-Driven Materials Science, the NOMAD Center of Excellence (European

Union's Horizon 2020 research and innovation program, Grant Agreement No. 951786) and the ERC Advanced Grant TEC1p (European Research Council, Grant Agreement No. 740233). The use case study 'Clean data in oxidation catalysis' was conducted in the framework of the BasCat—UniCat BASF JointLab collaboration between BASF SE, Technical University Berlin, Fritz-Haber-Institut (FHI) der Max-Planck-Gesellschaft, and the clusters of excellence 'Unified Concepts in Catalysis'/'Unifying Systems in Catalysis' (UniCat www.unicat.tu-berlin.de/ UniSysCat www.unisyscat.de). Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy -EXC2008–390540038—UniSysCat is acknowledged. Work on overarching repositories is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), in the framework of the project FAIRmat (FAIR Data Infrastructure for Condensed Matter Physics and the Chemical Physics of Solids, Project Number 460197019) FAIR Data Infrastructure for Condensed Matter Physics and the Chemical Physics of Solids, Project Number 460197019. Studies regarding the automation of experiments in catalysis research are funded by the Federal Ministry of Education and Research (BMBF) in the framework of the CatLab project, FKZ 03EW0015B.

3.3. Towards efficient and accurate input for data-driven materials science from large-scale all-electron density functional theory (DFT) simulations

Sebastian Kokott[1,2], *Andreas Marek*[3], *Florian Merz*[4], *Petr Karpov*[3], *Christian Carbogno*[1], *Mariana Rossi*[5], *Markus Rampp*[3], *Volker Blum*[6] *and Matthias Scheffler*[1]

[1] The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society, Berlin, Germany

[2] Molecular Simulations from First Principles e.V., Berlin, Germany

[3] Max Planck Computing and Data Facility, Garching, Germany

[4] Lenovo HPC Innovation Center, Stuttgart, Germany

[5] MPI for the Structure and Dynamics of Matter, Hamburg, Germany

[6] Thomas Lord Department of Mechanical Engineering and Materials Science, Duke University, Durham, United States of America

Status

The quality of input data is critical for data driven science. Detailed, high-level (i.e. quantum many-body theory based) simulations, although expensive, can provide immensely valuable data on which other methods can build, if three main issues can be addressed: First, the system size of accurate quantum-mechanical simulations is often restricted by the computational complexity of the underlying simulation algorithms. Second, the accuracy of the predicted data for new complex materials critically depends on the accuracy of the specific physical model chosen to derive quantum-mechanical simulation data, limiting subsequent data-driven models. Third, the number of atomic-scale configurations that must be covered for a statistically sound description grows dramatically with the complexity of a material, necessitating more and faster high-level calculations to provide input data for subsequent, AI-driven research. Simulations of real-world materials require addressing all three points at the same time.

Hybrid density functionals (hybrids) have emerged as a practical reference method for *ab initio* electronic-structure-based simulations because they resolve several known accuracy issues of lower levels of DFT while offering affordable computational cost on current high-performance computers. There are two main computational bottlenecks for atomistic simulations using hybrid DFT: evaluating the non-local exact exchange contribution and finding the solution of a generalized eigenvalue problem (matrix diagonalization). Here, we discuss advances and perspectives for both challenges as recently implemented in the all-electron code FHI-aims [57].

The current reach of these methods is documented by run times and scaling of hybrid DFT simulations for several challenging materials, including hybrid organic/inorganic perovskites [58] and organic crystals, with up to 30 000 atoms (50 000 electron pairs) in the simulation cell. The runtimes for the largest simulated systems are shown in figure 8. Despite such large systems sizes, the simulations can be run with moderate computational resources.

Current and future challenges

A resolution-of-identity-based real-space implementation of the exact exchange algorithm [59–61] was optimized to allow for much improved exploitation of sparsity and load balancing across ten thousands of parallel computational tasks. Results show drastically improved memory and runtime performance, scalability, and workload distribution on CPU clusters. The improvements pushed the simulation limits beyond 10 000 atoms, compared to an earlier

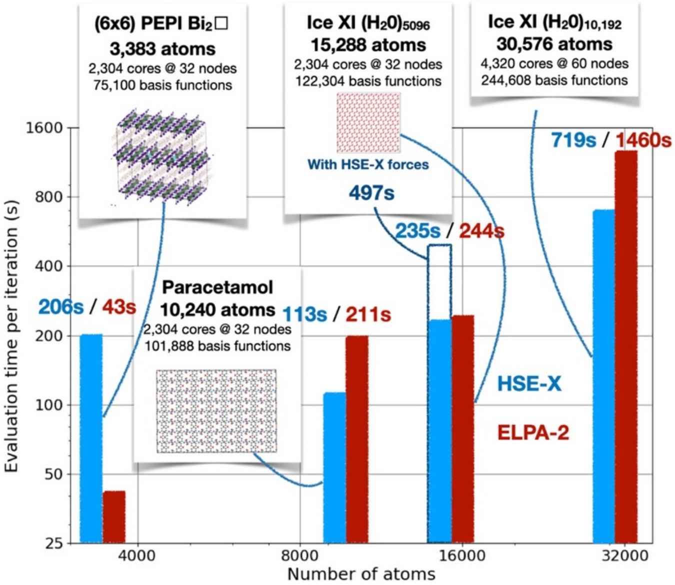


Figure 8. Average runtime to evaluate the exchange operator (blue bars) and the ELPA eigenvalue solver (red bars) per self-consistent field iteration. The HSE06 hybrid functional was used for all simulations. The following systems were simulated (from left to right): phenylethylammonium lead iodide (PEPI) with a defect complex [58], a 4 × 4 × 4 paracetamol supercell, a 15 288-atoms Ice XI supercell (including a force evaluation), and a 30 576-atoms Ice XI supercell. All calculations were carried out on the Raven HPC cluster at the MPCDF using Intel Xeon IceLake (Platinum 8360Y) nodes with 72 cores per node.

implementation that reached system sizes around 1000 heavy atoms [60]. This new code implementation can perform computation of energy, forces, and stress for periodic and non-periodic systems for several fashions of hybrid density functionals. In addition, for materials including heavy elements, perturbative spin–orbit coupling can be combined with the hybrids [62]. Due to the inherent O(N^3) scaling, the solution of the eigenvalue problem beyond 10 000 atoms becomes the bottleneck during the simulations.

The direct eigensolver library ELPA [63] has long offered unrivalled performance for parallel matrix diagonalizations. Extensive profiling, fine tuning and work on portability was carried out to adapt ELPA to the most current HPC architectures, further reducing the time for the diagonalization bottleneck for simulation sizes up to many thousands of atoms. Key to future success of ELPA is exploiting full capabilities of GPU-accelerated high-performance clusters. ELPA already has a well-established support for NVIDIA GPUs [64, 65]. Recently, we ported ELPA to AMD GPUs, enabling the solution of a problem with a matrix size with leading dimension of more than 3 million on 1024 AMD-GPU nodes of the LUMI pre-exascale system at CSC in Finland.

Although the library application programming interfaces (APIs) for AMD and NVIDIA are very similar, we find very different run-time and performance behaviour for the ELPA code. Thus, a new abstraction layer driving the GPU computations within ELPA has been implemented. Below this abstraction layer, the vendor specific implementations co-exist and can be independently developed and optimized. We believe that this very flexible approach facilitates the integration of upcoming new architectures, e.g. Intel GPUs.

Similar GPU strategies will be needed for the exact exchange algorithm, but are not yet exploited, as the porting of CPU code to GPU architecture is not at all straightforward. In the CPU implementation, the inherent sparsity of real-space approach keeps the size of matrices used for dense matrix-matrix operations moderate. Thus, with the current algorithm the full capabilities of GPUs cannot be used, and speedups would be limited by communication. An overhaul of the algorithm, and GPU-specific storage and communication patterns will be needed to make it amenable for heterogeneous, GPU-accelerated architectures.

Advances in science and technology to meet challenges

The achievements for hybrid DFT simulations demonstrated above is a big success and paves the way to efficient use of exa-scale resources in the future. Still, the accuracy of hybrids is limited by construction. The required fraction of exact exchange is an open point. A related question is the treatment of the electron correlation—hybrid density functionals addressing this point only insufficiently. Approaches using range-dependent parameters for the fraction of exact exchange or double hybrids are a way forward to improve the accuracy of the *ab initio* model. The *GW* approach and the CCSD method provide much more accurate access to electronic structure quantities per se, but the complexity of these methods will limit their application to smaller system sizes for the foreseeable future.

From a technological point of view, we think that sufficiently large memory per node and task will be needed for any enhanced electronic structure method, i.e. usually non-local operators are evaluated, which require finding a good balance between communication across nodes and tasks and storing data. Here, the tighter integration of accelerators within the HPC node, as, for example, expected for the upcoming Nvidia (Grace-Hopper) and AMD (MI300) technologies, looks very promising. There are two main hurdles for scientific software developers: library APIs for solving mathematical and physical problems are partially vendor-specific and/or not performance optimal. Addressing both points increase the reach of scientific code (and in turn reduces the need for code duplication) and will reduce the overall cost of research significantly. As a difficult task remains the optimization of communication patterns between CPUs and GPUs for specific architectures. Also new workload distribution models might be needed to better use all available resources, e.g. compute with GPUs and CPUs at the same time (right now often CPUs are idling while GPUs do the work).

Concluding remarks

The new exact exchange algorithm implemented in FHI-aims and the highly optimized ELPA library enables simulations of large system sizes at moderate runtimes. On the one hand, these implementations allow one to increase statistical sampling to address the huge configuration space that comes with the large system sizes. On the other hand, the accuracy of hybrid DFT simulations is sufficient for many applications. We believe that with the aid of future exa-scale resources in combination with sophisticated data-driven models, hybrid functionals will be established as default method for DFT simulations of materials. In general, exploiting sparsity is key to low-scaling electronic structure methods for large scale simulations. Real-space algorithms using localized wavefunctions are especially well suited. Nevertheless, the data distribution and communication pattern may need architecture-specific optimizations that complicates software design and code maintenance.

Acknowledgment

This project received financial support from BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials Science, the NOMAD Center of Excellence (European Union's Horizon 2020 research and innovation program, Grant Agreement No. 951786) and the ERC Advanced Grant TEC1p (European Research Council, Grant Agreement No. 740233).

3.4. Choosing AI analysis tools and enacting their reproducibility: the NOMAD AI toolkit

Luca M Ghiringhelli[1,2], *Luigi Sbailò*[1], *Ádám Fekete*[1], *Markus Scheidgen*[1]
and Matthias Scheffler[3]
[1] Department of Physics & CSMB, Humboldt-Universität zu Berlin, Berlin, Germany
[2] Department of Materials Science and Engineering, Friedrich-Alexander Universität, Erlangen-Nürnberg, Germany
[3] The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society, Berlin, Germany

Status

When, at the end of 2014, the NOMAD Repository & Archive [66, 67] went online, it was the first data infrastructure in computational materials science that fulfilled what was later and independently defined by the acronym FAIR. This definition and the request that scientific data should be FAIR was introduced in a very general scientific-data context by Wilkinson *et al* in 2016 [51]. As of today, the NOMAD Repository stores input and output files from more than 50 different atomistic (*ab initio* and molecular mechanics) codes and totals more than 13 million entries, uploaded by over 500 international authors from their local storage, or from other public databases. The NOMAD Archive stores the same information, but converted, normalized, and characterized by means of a metadata schema, the NOMAD Metainfo [68], which allows for the labelling of most of the raw data in a code-independent representation. One of the benefits of normalized data is that they are accessible in a format that makes them suitable for direct AI analysis.

NOMAD also offers the AI toolkit [69], a JupyterHub-based platform for running notebooks on NOMAD servers, without the need for any registration or downloaded software. The data-science community has introduced several platforms for performing AI-based analysis of scientific data, typically by providing rich libraries for AI. General-purpose frameworks such as Binder [70] and Google Colab [71], as well as materials-science dedicated frameworks such as pyIron [72], AiidaLab [73], and MatBench [74] are the most used by the community. In all these cases, a big effort is devoted to education via online and in-person tutorials. The main specificity of the NOMAD AI toolkit is its connection with the extensive NOMAD Archive. Moreover, together with the NOMAD Oasis [67], users can work with their private as well as community data within the same software platform and using the same API.

Current and future challenges

Besides providing the framework for performing custom-made AI analysis, the NOMAD AI toolkit provides a set of tutorial notebooks introducing users step by step into both the most popular and widely known AI methodologies, with showcase applications in materials science, and into the more advanced ones, i.e. methodologies that have been published in the latest years. Due to the very nature of the Jupyter technology, these tutorial notebooks are interactive, in the sense that users can modify lines of codes and check the effect of the modifications. Also, the tutorial notebooks have direct access to the whole NOMAD data, so that users can apply the learned techniques to new data, including data uploaded by them.

Importantly, the AI toolkit includes notebooks that present actual AI software as used for producing results for peer-reviewed publications. This feature suggests that scientific *reproducibility* can reach its full potential, at least for AI analysis tools. For instance, users can re-train AI models with exactly the same set of hyperparameters as used in the original publications, on exactly the same data, including the train/validation/test set splits. A piece of information that nowadays is not required in peer-reviewed publications. However, such addition would be scientifically appropriate as it would directly enable the reproducibility of reported results. The NOMAD AI toolkit enables this important step.

As already noted by the proponents of the FAIR principles for scientific software [75], providing complete information on the algorithms and software used to analyze data is all but trivial. This is particularly challenging if one wants to provide live software that can be run on demand, mainly because pieces of software, e.g. Python scripts for an AI analysis, require a virtual environment where libraries that are used for efficiently performing certain routine tasks are installed. These libraries get repeatedly updated, and unfortunately backward compatibility is not necessarily ensured. This means that the same set of commands that at release time allows to install and run a software, at a later point in time may not yield a correct installation. Besides, in the case that the software is run in a container (as for the NOMAD AI toolkit), when a new container is created the software for the container platform gets updated. In other words, special care and planning has to be devoted to maintaining the whole ecosystem of software so that exactly the same datasets yield in time exactly the same AI models and therefore exactly the same predictions over the same test data.

Advances in science and technology to meet challenges

Platforms like the NOMAD AI toolkit also foreshadow the scientific-reproducibility utopia. Much of the technology that allows for reaching these goals still needs to be developed, but some important steps have been taken already. First of all, Jupyter notebooks can be uploaded to NOMAD as easily as the data. The upload timestamps and other *provenance* metadata that allow for the unique identification of each analysis script. Furthermore, users are encouraged to provide a rich set of metadata that are made searchable and therefore will allow other users to locate the notebooks by e.g. model class for the AI analysis, or by used libraries, including their versions. In its current state, the NOMAD AI toolkit allows for the findability and interoperability of the AI-analysis software. In fact, a unique container is currently used for all the notebooks, thus allowing for a full interoperability among the different AI analysis tools. The complexity of the maintenance of such an environment rapidly increases with the number of uploaded notebooks which poses challenges to ensure that stored notebooks can run over the years and produce the same results. However, each set of obtained results, including all the intermediate results along the analysis workflow, can be stored (according to FAIR principles) and automatic tests could be run to check the conformity of the results produced by the re-trained model with the reference ones. Knowing that some piece of code is at some point in time unable to reproduce old results is the necessary condition to try and fix the code in order to conform with the reference results. This solution, which requires quite some human effort, introduces a possibly interesting generalization on the idea of reproducibility, which in some sense is a black-box requirement. I.e. in each step of the analysis, the same input needs to yield the same output, but the details inside the black box are allowed to change.

A radically alternative route is to partly renounce to a full interoperability among the notebooks and maintain several different containers within the NOMAD AI toolkit. Such an approach would allow for the creation of specific containers that are not updated, thus allowing for the software installed therein to be always executable. Although the tools used in these not-updated containers cannot always be combined with software installed into other containers, it can still be deployed on new data that have been uploaded at a later time.

Concluding remarks

The introduction and gradual implementation of the FAIR practices for scientific-data management and stewardship revealed that another crucial component of scientific research needs to adopt the FAIR concepts: the scientific software for data production and analysis. As for data, the key point is the *reproducibility* of research finding, i.e. the practical possibility to re-obtain the same results starting from the same hypotheses (the input settings) and methods.

Clearly, providing only the input data and results in a data archive, even if fully FAIR-data compliant, is not enough for reproducibility, if part of the results are obtained in an incompletely documented way and/or via some custom-tailored analysis software, which is not properly stored and versioned.

The NOMAD AI toolkit already enables re-run AI software on FAIR data for a relatively small set of Jupyter notebooks at the price of human-intensive maintenance. The grand-challenge is to develop a strategy to scale up such maintenance in a (semi-)automatic fashion, so that all AI tools from the community can be preserved according to FAIR practices, fully achieving scientific reproducibility.

Clearly, these reproducibility concepts and the use of Jupyter notebooks also imply that newcomers to AI can use the software that already exists at the NOMAD infrastructure, train themselves and adjust and advance the analysis tools towards their own but different applications.

Acknowledgment

M Scheffler thanks Fawzi Mohamed, Thomas Schulthess, Rampi Ramprasad, and Christian Carbogno for discussions on several critical development of the AI toolkit in its current state and beyond. LMG acknowledges inspiring discussions on the future directions of the AI toolkit with Thomas Hammerschmidt, Kevin Jablonka, José Márquez Prieto, and Claudia Draxl.

The project received funding from BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials Science, the NOMAD Center of Excellence (European Union's Horizon 2020 research and innovation program, Grant Agreements Nos. 676580 and 951786), the ERC Advanced Grant TEC1p (European Research Council, Grant Agreement No. 740233), and the project FAIRmat (FAIR-Data Infrastructure for Condensed-Matter Physics and the Chemical Physics of Solids, German Research Foundation, Project No. 460197019).

3.5. Synthetic Hamilton Matrices for Deep Learning

Sajal K Giri[1]*, Ulf Saalmann*[2] *and Jan M Rost*[2]
[1] Department of Chemistry, Northwestern University, Evanston, IL, United States of America
[2] Max Planck Institute for the Physics of Complex Systems, Dresden, Germany

Status

Training of deep learning (DL) models requires a large amount of data in the first place and the data set must be sufficiently diverse for the network to be transferable such that it produces unbiased predictions. At the same time, the data size needs to be balanced to compensate for the cost of their generation. Strategies to deal with scarce data problems include transfer learning (TL), self-supervised learning, generative adversarial networks (GANs), model architecture, physics-informed neural network, and deep synthetic minority oversampling techniques to name a few recent approaches, as pointed out in [76].

Here, we focus on a specific route to overcome the scarce training data bottleneck, namely the generation of *random synthetic training data* under suitable constraints determined by the physics involved [77–79]. In our approach we aim at modelling system dynamics by encoding it into a Hamilton matrix for the interaction of (bound) electrons with intense laser light. The latter can be very noisy and fluctuating from shot to shot, as produced by x-ray free electron lasers (XFELs). We vary the elements of the Hamilton matrix randomly about a matrix of an existing model system in one physical dimension (1D), creating synthetic Hamilton matrices for systems which could but do not necessarily exist in nature for which calculations can be done quickly compared to real 3D systems. From the large set of SHMs augmented with different deterministic realizations of noisy laser pulses, we compute photoelectron spectra to train a fully-connected DNN. Figure 9 shows an application of the DNN (trained with spectra from SHMs) to a real 3D system for which it predicts, without knowing the system explicitly, how the noisy spectrum would look like if a ‘clean’ (Gaussian) laser pulse would have been used. The good agreement with the ground truth demonstrates that the trained DNN can be transferred from 1D to 3D problems and gives confidence in our SHM-DL concept.

Very recently, the idea of synthetic data generation based on existing data has been taken up for composite materials [79], where a limited number of original full-field micro-mechanical simulation data are randomly rotated in physical 3D space to generate additional data to train a recurrent neural network for the non-linear elasto-plastic response of short fibre reinforced composites. Similar ideas using TL are being explored in other areas [80].

Current and future challenges

An important problem in the context of spectra generated by XFEL double pulses is the delay between the pulses which jitters in an unknown way from shot to shot. The SHM-DL approach can extract the time-delay of a double-pulse from the spectrum it has generated. Importantly, we can sort single-shot noisy spectra according to the time-delay of the double pulse with which the spectra were generated. With a second network the time-delay sorted pulses, binned over small delay intervals (1 fs) can be purified as shown in figure 10. This constitutes a substantial generalization to predict a hidden parameter (the time-delay of the pulses) [78].

The task the SHM-DL has successfully completed so far, is the mapping of spectra generated by noisy pulses to spectra generated by Gaussian (Fourier limited) pulses. Can we also predict via SHM-DL maps spectra for other clean pulse forms, e.g. for pulse forms which

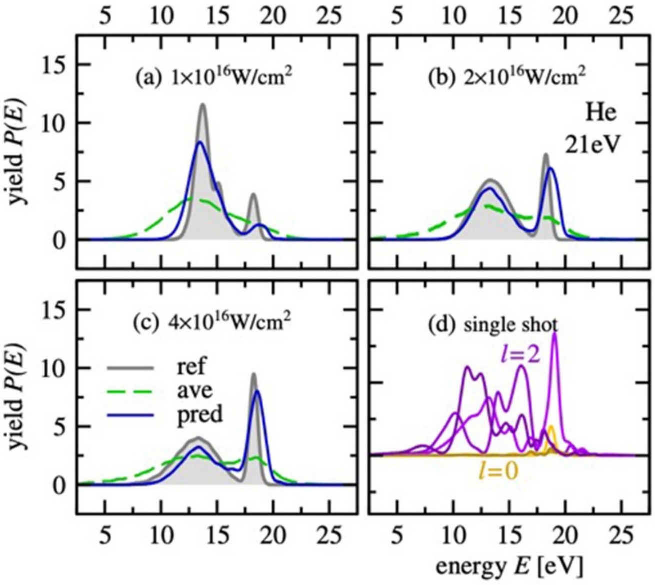


Figure 9. Photoelectron spectra for the He atom (a)–(d) from noisy laser pulses with a central photon energy of 21 eV and peak intensity I as given in the panels. In (d) the spectra from different single noisy pulses are shown with dominant contributions from angular momentum $l = 2$ (purple) and $l = 0$ (yellow) after absorption of two photons by the ionized electron at $I = 2 \times 10^{16}$ W cm^{-2}. Panels (a)–(c) show the spectra which result from averaging the single-shot spectra (green-dashed), the reference (i.e. the spectra calculated from a clean Gaussian laser pulse, grey-shaded) and the predicted spectra (blue solid) for a clean pulse by the DNN (Reproduced from [77]. CC BY 4.0).

are not even realizable experimentally? This would be very interesting for systems whose response to light cannot be computed (too complicated) but measured, e.g. with noisy pulses as described, since with SHM-DL we do not need to compute the 'true' spectrum of the desired system.

The primary vision of the SHM-DL approach is a 21st century spectroscopy. Applied to molecular ro-vibrational spectra, e.g. it could replace the traditional normal mode model for the assignment and classification of spectral, leaving it to the trained network to associate appropriate SHMs with the spectrum, thereby classifying it by means much more flexible than traditional, structurally predefined normal modes.

The long-term goal is to develop SHM-DL to become capable of identifying a single SHM (or a small group of SHMs) which describe the system so well, i.e. represent the system, such that from the reconstructed SHM(s) time-dependent system evolution in general and other observables can be computed/predicted. This would constitute a physics-rooted form of generalization which delivers at the same time physical insight as it provides an optimal parameterization of a physical system with a Hamilton matrix of chosen size. First attempts are promising that identify SHMs in relation to two- or multidimensional spectroscopy [81].

Advances in science and technology to meet challenges

Technically, even the SHM-DL approach remains a challenge regarding the computing power needed to numerically produce the spectra (training data) from the SHMs. Hence, (i) a

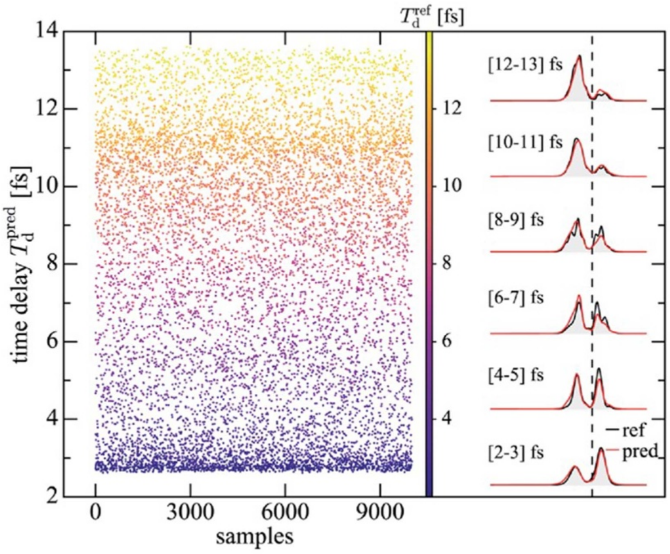


Figure 10. Reconstruction of double pulse time-delay and purification of noisy spectra for a single Hamilton matrix taken from test data. Single-shot fluctuating spectra for random time-delays are passed through the trained network to reconstruct the underlying time delays which are shown as scattered points where the colour gradient represents the reference time-delay. We consider 12 intervals of time delay in the range 2–14 fs with an interval length of 1 fs. All single-shot spectra which fall into interval of time-delay are averaged. The averaged spectra are passed through another network which maps averaged noisy spectra to purified ones. The predicted purified spectra (red) are compared to reference spectra (black). Reproduced from [78]. CC BY 4.0.

reduction of the required training data size by better knowledge of the underlying physics is desirable as well as (ii) a reduction of computational costs by ultra-efficient quantum propagation in time to obtain the spectra [82]. (iii) Furthermore, the computed spectra as training data must be balanced. For the time being this is done by simply discarding spectra from the training set which are too close to each other. However, this implies a large waste of computing time. To reduce this waste several advances are desirable: Firstly, use of an optimal metric to determine the Euclidean 'distance' between two spectra. Here, recently the Wasserstein metric has become popular [83], or approximations to it which are computationally cheaper. A more elegant, physics-oriented advance would be to find an approximate inversion of the SHM-to-spectrum map, or any other way which allows us to shift the balancing of the spectra to suitable choices of the SHMs.

Thinking ahead, the idea of SHMs could be realized not with DNNs but other DL approaches. Most promising are GANs or variants thereof, where the relevant SHM is constructed by the GAN from a random one successively with computationally costs eventually reduced compared to the present SHM-DL sampling approach. Moreover, the GAN approach would directly predict an SHM which describes the system's coupling to light.

Finally, and almost trivial since true for many DL applications: SHM-DL would benefit enormously from a possibility for inherent error quantification.

Concluding remarks

We have introduced the idea of SHMs, random representations for the dynamics of systems coupled to light, which could exist but not necessarily exist in nature. This approach enables sampling of the training space solving the dynamics with SHMs efficiently incorporating sufficiently generic features to be transferable to real systems. They serve the purpose to augment training data for DL with DNNs. We demonstrated that this SHM-DL approach works by purifying photoelectron spectra from noisy pulses and identifying pulse delays which jitter in an unknown manner, as supplied by x-ray Free-electron Lasers. The approach is physics-oriented and therefore promises physics insight beyond the prediction of spectra through the DL based identification of the relevant Hamilton matrix from a spectrum for a system, unknown to the DNN.

Acknowledgment

The project received funding from BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials Science.

3.6. Spatiotemporal models for data integration

R Patrick Xian[1], *Jason Hattrick-Simpers*[2], *Ralph Ernstorfer*[3] *and Stefan Bauer*[4]
[1] Department of Statistical Sciences, University of Toronto, Toronto, Canada
[2] Department of Materials Science and Engineering, University of Toronto, Toronto, Canada
[3] Institute for Optics and Atomic Physics, Technical University of Berlin, Berlin, Germany
[4] School of Computation, Information and Technology, Technical University of Munich & Helmholtz AI, Munich, Germany

Status

Understanding the structure-property relations of materials and optimizing chemical synthesis or device manufacturing processes requires integrating multimodal datasets from both theory and experiments [13] that often encompass spatial and temporal dependence. The explicit spatiotemporal characteristics may be exploited in model-building for data integration. Historically, spatial and spatiotemporal models were largely developed in the contexts of geoinformatics, biostatistics, and quantitative ecology, many of which remain underappreciated by the materials science community. In these models, the temporal and spatial subsystems are typically considered in 1D and 2D/3D, respectively. Spatiotemporal models describe their subsystems jointly to capture the interactions through covariance functions or dynamical processes derived from physical knowledge [84]. They are structured and interpretable and are considerably more tractable than first-principles methods.

RFs and GPs, also known as kriging, are two established categories of models designed for spatial and spatiotemporal data. RFs already have established use in the statistical modelling of microstructured materials [85], while GPs are invented in mining engineering and they are a classic example of surrogate models. We discuss here three diverse examples from materials data science that indicate their broad applicability and utility. (i) In metal additive manufacturing, Saunders *et al* [86] combined three GPs with distinct characteristics to model the pairwise relationships between materials microstructure, melt pool geometry, and mechanical property obtained from multiphysics simulation, all of which are also time-dependent. (ii) In photoemission spectroscopy, Xian *et al* [87] constructed a Markov RF with nearest-neighbour interaction and transformed the band dispersion reconstruction problem into a classification problem (see figure 11). The coordinates in their problem are the two momenta and energy of photoelectrons, the use of pre-computed energy bands from electronic structure theory provides an effective initialization. (iii) In combinatorial materials screening, Kusne *et al* [88] constructed a GP in the chemical compositional space to guide the search for the optimal stoichiometry within a family of tertiary phase-change materials. Their algorithm was integrated into a synchrotron beamline and may be run in a closed loop driven by active learning.

Current and future challenges

One defining characteristic of materials science data is its abundance of data types, from videos to images to atomic structures [13]. Comparatively, spatiotemporal models, besides the classic examples like RFs and GPs, may also take the form of point processes [89], state space models [90], and diffusion processes [91], which are as yet not used for data integration, but have their respective benefits to representing specific data types. Besides coordinates with concrete physical meaning, one could also consider direct spatial or dynamical models of the latent space in data integration, as it is often more robust to noise and dimensional scaling artefacts,

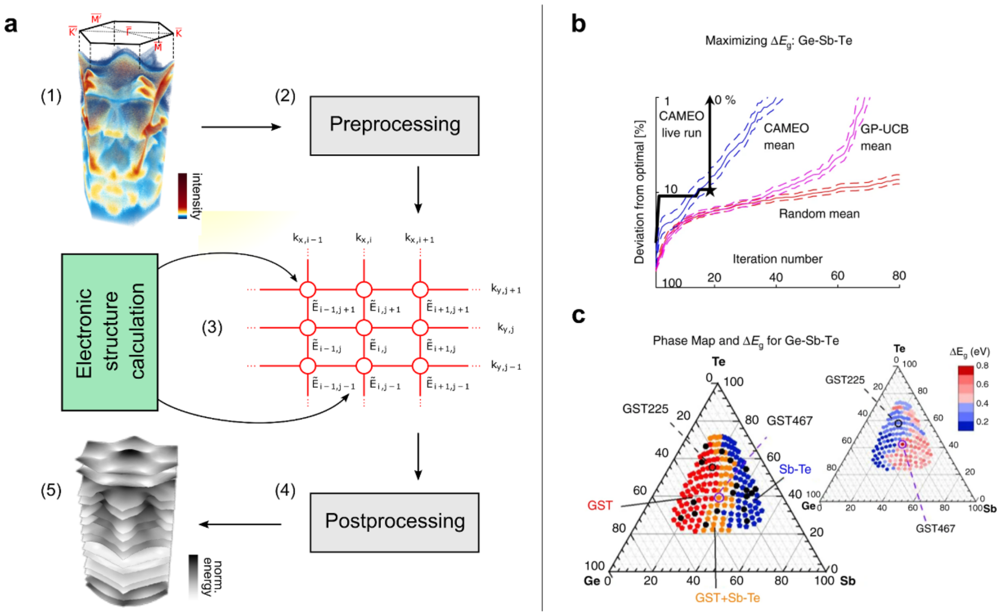


Figure 11. Illustrations of spatial models of (left) photoemission data in the energy-momentum coordinates using a Markov random field [87] and (right) combinatorial material screening data in the chemical compositional space using a Gaussian process [88]. Reproduced from [87]. CC BY 4.0. Reproduced from [88]. CC BY 4.0.

especially for multiple data modalities. This leads to the question of problem mapping from data type to model category and subsequent model specification as the primary challenges. The three examples in the previous subsection illustrate that building spatial and spatiotemporal models are not limited to the physical dimensions attached to their original meaning. The straightforward way for problem mapping is to first identify the data types related to a particular problem, then consider the native data types to the model and find the match. For example, point process models would be suitable for modelling the transport of point defects because of their sparse distribution.

Secondly, we should pay special attention to the data quality in the subsystems to be integrated, including resolution, unit size (such as pixel size for image data), missingness, structuredness, and fidelity (such as noise level). Many of these problems are not yet formally addressed, thereby motivating further research on a case-by-case basis guided by domain knowledge. For example, data resolution and fidelity affect the choice of integration ordering, i.e. from high to low or in reverse. For experimental data, the unit size is usually not the same as the resolution because of blurring introduced by the instrument response. Thirdly, we should consider the scalability of the model during development, which may be left unnoticed until later in model deployment using real-world data. For spatiotemporal models that aim to handle large datasets, scalability is often a primary determinant of model choice.

Advances in science and technology to meet challenges

The two main paradigms in materials science that benefit from advancements in spatiotemporal models are: (i) self-driving (or autonomous) laboratories [92]. They deploy robots and ML-driven sequential decision-making from streaming data to search through high-dimensional parameter spaces (such as process, composition, and property parameters) for materials optimization. A growing number of them are installed at large-scale research facilities such as x-ray or neutron sources or in regular research institutions for organic and inorganic synthesis. (ii) Combined large-scale atomistic simulation and video-mode recording of time-resolved experiments [86]. Here both the simulation and the data analysis may be powered by ML algorithms, while data integration between the two modalities through a spatiotemporal model is needed to obtain experimentally validated physical parameters. Both of these two paradigms will benefit from the following developments:

From the model development side, accurately accounting for long-range dependence (LRD) [90] in both spatial and temporal dimensions is one of the crucial yet unmet challenges. LRD manifests in the slowly decaying dependence structure, such that the Markov assumption is no longer a valid approximation. Current approaches using deep-state space models are limited to video frame classification and generation, further improvements on both spatial and temporal LRD, computational efficiency, and the accommodation of graph-structured data will be fitting for the demands in materials data integration.

From the data engineering side, the data integration process often involves the comparison of metadata from two or more sets of measurements or calculations, which require that the data formats are interoperable. Systematic documentation of metadata is crucial for successful data integration projects, which now lie at the centre stage of the FAIR principle [13]. For materials optimization platforms that depend on streaming data, the development of automated (meta)data logging systems that include anomaly and distribution shift detection is essential for the quality control of data acquisition. It will also pave the way for efficient data integration and enable online search and process optimization.

Concluding remarks

Spatiotemporal models have demonstrated promising outcomes in integrating data from multiple sources and guiding scientific discovery. The future of spatiotemporal models for materials data science should explore the interplay between the domain knowledge used in problem mapping and model specification to ensure a faithful representation of the problem context to achieve the desired interpretability and performance.

Acknowledgment

The work was partially supported by BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials Science.

3.7. Soft-matter simulations

Tristan Bereau[1] *and Kurt Kremer*[2]
[1] Institute fur Theoretical Physics, Heidelberg University, Heidelberg, Germany
[2] Max Planck Institute for Polymer Research, Mainz, Germany

Status

Soft matter is a sub-class of condensed matter that comprises systems with a characteristic energy on par with thermal energy at room temperature, $k_\mathrm{B}T_\mathrm{room}$ (about $2.5 \cdot 10^{-2}$ eV at $T = 300$ K). The low energy gives rise to significant conformational (intra-molecular) flexibility, leading to the spontaneous self-assembly of supramolecular mesoscopic structures. Relevant systems include polymers, colloids, and complex fluids, for which soft-matter physics have provided a foundational understanding [93]. Soft matter offers a slew of modern-day applications, e.g. food products, rubbers for automotives, electronics or medical applications. This makes the discipline both scientifically and technologically highly relevant.

A crucial aspect is the relevance of multiple scales: phenomena occur at various length- and time-scales, some of which decouple. This simplifies the tackling of complex systems: to build simpler models and focus solely on the relevant degrees of freedom. Scale separation takes its roots in renormalization group theory, and with significant implications in various aspects of theory (e.g. scaling concepts in polymer physics) as well as computer simulations (i.e. multiscale modelling). Figure 12 illustrates the benefits of multiscale modelling for two applications: high-throughput screening of drug-membrane permeability, and a hierarchical description of polymeric organic electronics.

Soft-matter science has gone through substantial evolution in the last half century. In polymer physics, experiments and theory have worked hand in hand early on to measure coveted critical exponents, and link to general statistical mechanics theory. Computer simulations have played an increasingly important role—they offer invaluable microscopic detail and reach out to ever-growing system sizes [96, 97]. They combine basic generic concepts with specific material properties. In the last decade, data-driven methods, and more recently ML, have become increasingly popular in soft matter. They offer an inductive approach to help bridge the scales, and more broadly solve complex structure-property relationships.

Though the penetration of ML in soft matter has been lagging against hard condensed matter, more recent developments show that the outstanding challenges faced by conformational flexibility (i.e. the role of entropy) are increasingly being addressed. In accordance with other fields of physics, chemistry, and materials science, the pursuit of stronger inductive bias (i.e. building physics in the model) systematically helps build better models in an area that is notoriously scarce in data—experimental or from computer simulations. The continued development of ML techniques for soft-matter physics, and the cross-penetration of ML with multiscale modelling, is helping push soft-matter physics toward higher-precision predictive modelling, soft-materials design and optimization, and reproducing entire experiments on the computer [98].

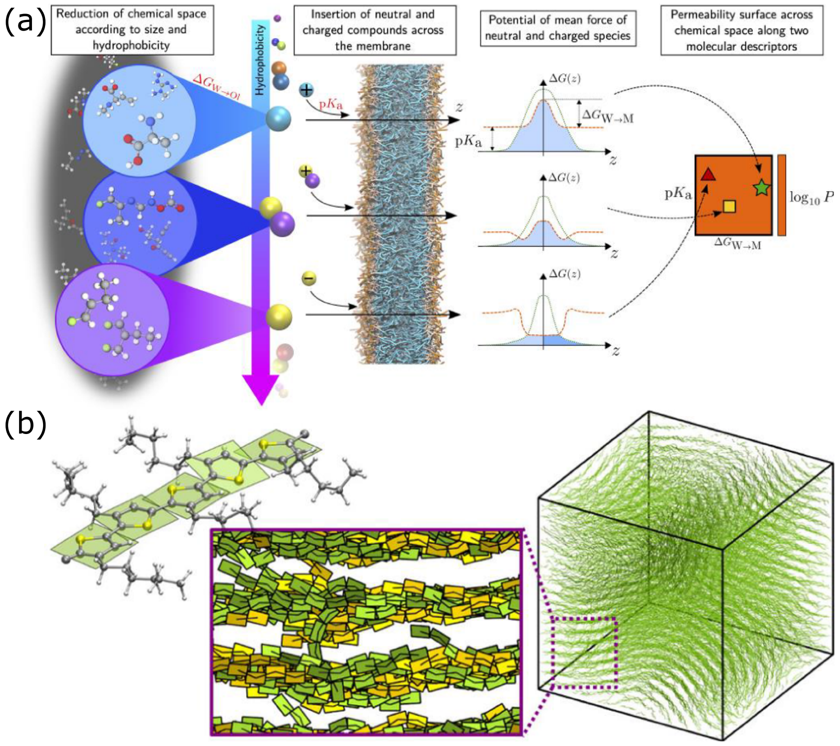

(a)
Reduction of chemical space according to size and hydrophobicity
$\Delta G_{W\rightarrow Ol}$
Hydrophobicity
Insertion of neutral and charged compounds across the membrane
pKa
z
Potential of mean force of neutral and charged species
ΔG(z)
$\Delta G_{W\rightarrow M}$
pKa
z
Permeability surface across chemical space along two molecular descriptors
pKa
$\log_{10} P$
$\Delta G_{W\rightarrow M}$
(b)

Figure 12. (a) Coarse-grained simulations of drug-membrane permeability to screen compounds at high throughput. Reprinted with permission from [94]. Copyright (2019) American Chemical Society. (b) Hierarchy of descriptions of P3HT, a prototypical polymer for organic electronics. Charges are transported primarily along the backbone of the chains, while the aliphatic side chains are needed to process the material. Reprinted with permission from [95]. Copyright (2019) American Chemical Society.

Current and future challenges

The field of big-data-driven materials science in the context of soft-matter simulations faces several outstanding challenges:

(1) The foremost challenge is tackling the ‘black box’ nature of complex ML models like DNNs. Why does a ML model make certain decisions? To this end, interpretability and explainability are paramount. Important developments have been made in the direction of symbolic regression, thereby discovering *mathematical equations* governing the complex phenomena characteristic of soft matter systems. Still, more effort is needed to gather further insight and intuition behind the underlying physics.

(2) What makes soft-matter systems fascinating is also what makes them challenging: their multiscale nature. The aggregate effect of many small parts often sums up to large-scale supramolecular behaviour—this emergent phenomenon is an outstanding challenge to effectively learn in ML models, and adequately generalize. This is the main reason why computer simulations remain essential nowadays and cannot easily be replaced by ML models alone. Looking to the future, the fusion of ML with physics-based simulation methods (e.g. molecular dynamics or Monte Carlo methods) is expected to have a strong impact.

The recent combination of LLMs with multi-agent collaborations strikes us as a relevant component toward modelling complex soft-matter systems [99].

(3) Furthermore, navigating non-equilibrium dynamics stands as a colossal challenge. Almost all soft matter systems—including all of life—exist far from equilibrium. Worse, even systems that appear in equilibrium typically depend on non-equilibrium effects via their *processing*: the mere preparation (e.g. synthesis and subsequent treatment) impacts the final product [100]. The absence of a well-established theory for non-equilibrium statistical mechanics can be an opportunity for inductive methods.

Though soft-matter science is already traditionally an interdisciplinary field, bringing together physics, chemistry, biology, and materials science, the advent of data-driven methods and ML further reaches out into computer science. The training of scientists that can efficiently work and communicate between these different fields is more important than ever.

Advances in science and technology to meet challenges

Compared to hard condensed matter, soft matter lags behind in terms of ML integration, in large part due to the need to address the associated conformational flexibility. One outstanding challenge lies at the level of system representation, i.e. how to encode the fluctuating system configuration for input to an ML model. Atomic representations developed for electronic properties have focused on single configurations (e.g. [101]). Here instead, observables are averaged over a typically very broad Boltzmann distribution of configurations. Much less work has been proposed in the context of ensemble-averaged ML representations, though ideas have been proposed [102, 103].

Capturing multiscale phenomena lies at the heart of soft-matter physics—from microscopic molecular architecture to mesoscopic structure, to macroscopic behaviour. Limitations in the generalizability of ML models strongly limits the current prospects of replacing physics-based models. It is not so clear how extensive the training of an ML model ought to be to reproduce emergent phenomena, such as the self-assembly of soap bubbles from amphiphilic molecules. Coarse-grained modelling has been at the forefront of soft-matter simulations—it exploits scale separation to focus on the most relevant degrees of freedom. Advances in combining coarse-grained modelling with ML is key to further develop data-driven soft-matter simulations. Much work is currently focused on ML-based coarse-grained potentials [104, 105], where striking an adequate balance between accuracy and computational speed is of critical importance. Longer term, it is not clear to what extent ML models might be able to generalize enough to replace the integration of classical equations of motion.

Concluding remarks

It is difficult to overstate the significant impact of first theory, and later computer simulations, on our understanding of soft matter. Bringing soft matter to the fourth paradigm of science (i.e. data-driven methods) will require the tackling of several outstanding challenges. The ongoing developments of ML will hopefully continue to naturally evolve from hard condensed matter to soft matter, thereby addressing the needs to model configurational entropy. We foresee that these technical hurdles may help usher soft matter in a new era, where poor scale separation can be efficiently addressed, and insight can be gained for phenomena that are too complex for traditional methods.

Acknowledgment

TB and KK thank BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials-Science, for support and stimulating interdisciplinary interactions with members of the consortium.

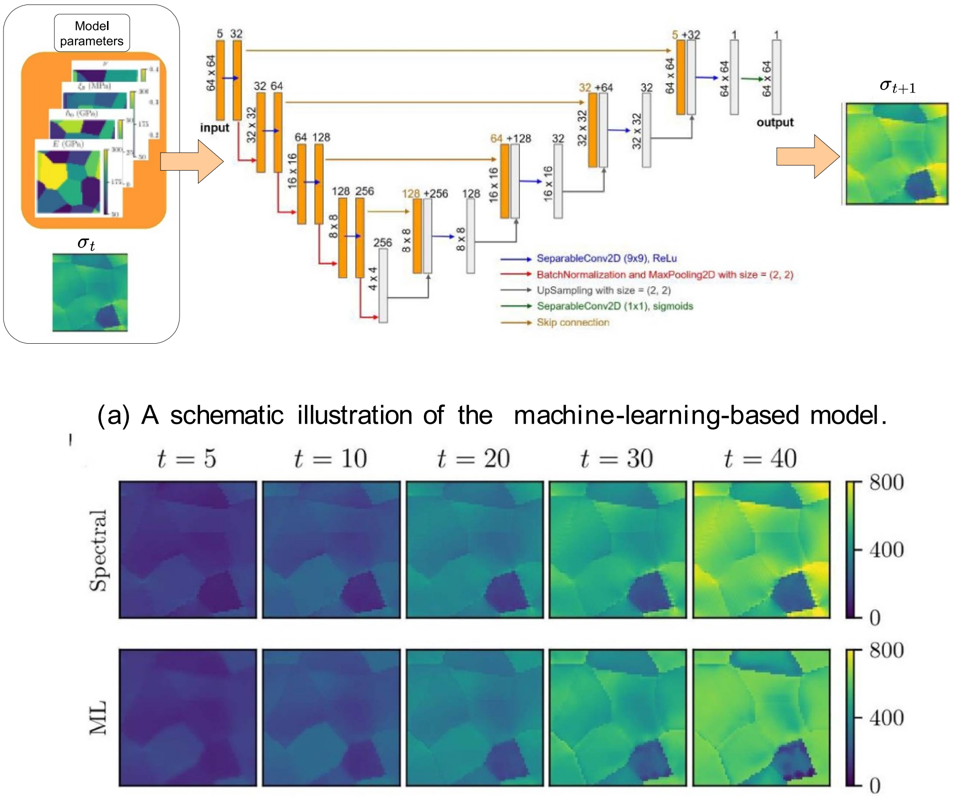
3.8. Physics-enhanced ML based surrogate modelling for continuum mechanics

Pawan Goyal[1], *Mohammad S Khorrami*[2], *Jaber R Mianroodi*[2], *Peter Benner*[1]
and Dierk Raabe[2]
[1] Max-Planck-Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany
[2] Max Planck Institute for Sustainable Materials, Düsseldorf, Germany

Status

Mathematical modelling plays a pivotal role in the study of continuum mechanics and material design, offering profound insights into material behaviours and microstructures, which in turn, support and guide material optimization and design. Typically, this modelling process involves formulating partial differential equations (PDEs) based on fundamental physical principles such as mass and energy conservation as well as force equilibrium. Non-equilibrium aspects such as the role of microstructure and the underlying carriers of inelastic deformation such as dislocations and mechanical twins are also represented by PDEs, e.g. in the form of mean field defect density equations that couple to stress and strain [106]. These PDEs, for given initial conditions and boundary values, are subsequently solved using numerical methods, with finite element and spectral methods being popular choices. Unfortunately, these traditional numerical techniques are computational very costly, a challenge that becomes particularly pronounced when dealing with design studies that require a multitude of simulations under varying configurations. To address this computational burden and streamline the design cycle, there is a pressing need to develop surrogate models that can replace the traditional simulations, often reliant on the methods mentioned above, like finite elements, spectral methods, or finite volume techniques. These surrogate models are particularly valuable during the design phase, offering a more computationally tractable solution.

The use of artificial neural networks (ANNs) in surrogate modelling, driven by advances in ML and DL, has therefore become a field of growing interest [107]. While neural networks-based material modelling can be traced back to [108], it is in the last decade, with the rapid progress in DL and the availability of powerful hardware, that the development of surrogate models using ANNs has surged and continues to expand. Data utilized for constructing these surrogate models can comprise a combination of experimental data, empirical knowledge, and synthetic data generated through numerical solvers. Within the realm of continuum mechanics, numerous methodologies have emerged for building surrogate models using ANNs. For example, in [109] a neural network architecture, namely, conditional GANs, has been employed to predict stress and stress fields for a given microstructure geometry [110]; employed a convolutional neural network (CNN) to estimate von Mises stress for microstructures consisting of isotropic elastic and elasto-plastic grains within microstructures, with extensions to heterogeneous periodic microstructures [111] as depicted in figure 13. Furthermore, [112] explores the application of the Fourier neural operator (FNO) for the surrogate modelling of stress and strain in heterogeneous composites. Additionally, in recent times, there have been attempts to generate physics-based solutions by using large language models (LLMs) based on the requirements mentioned by a user, see [113]; however, it targets at generating scripts (e.g. Python) to generate finite-element simulation code. However, these simulation codes are still of high fidelity, hence, are computationally still expensive.



(a) A schematic illustration of the machine-learning-based model.

(b) A comparison of prediction quality.

Figure 13. The schematic illustration of the machine-learning-based surrogate model for predicting the history-dependent local von Mises stress in a solid aggregate that comprises sets of crystalline grains which are characterized here by different elastic-plastic stiffness. The material parameters considered (varied from grain-to-grain) in the simulations include E: Young's modulus, μ: Poisson ratio, ζ_0: initial inelastic flow resistance (viz. stress value, where plastic deformation starts), h_0: initial isotropic hardening, and σ: von Mises equivalent stress. We construct a surrogate model of stress fields for visco-plastic polycrystalline materials using the U-Net architecture as shown in (a), which predicts von Mises stress field 500 times faster than conventional spectral solvers, see (b). Reproduced from [110]. CC BY 4.0. Reproduced from [111]. CC BY 4.0.

Current and future challenges in using big data methods for continuum mechanics

Often, surrogate modelling is conducted purely based on large amounts of data, mostly by training ANNs with them. However, within the context of continuum (micro-)mechanics, there exists a wealth of established physical and empirical knowledge [114] that ANN-based surrogate methods have yet to fully incorporate. In the following, we discuss the notable challenges in bridging this gap for surrogate modelling in continuum mechanics.

(a) *Physics-enhanced surrogate modelling*: Incorporating physics-based knowledge into surrogate modelling is an active research field. For instance, in [115] and [116], physics-based knowledge, including the underlying PDEs and empirical knowledge, has been leveraged to introduce biases into ANNs, resulting in outputs that approximate the underlying physics, such as enforcing divergence-free conditions as well as mass and energy conservation. However, it is essential to note that these approaches primarily aim to satisfy the physical laws in a weak sense. Hence, the output from the trained ANNs may not be fully physically meaningful, particularly at a local scale. Therefore, we need to explore the design of neural network architectures that are capable to inherently produce an output that satisfies physics in a strong sense, with a particular focus on critical properties like divergence-free behaviour, as well as mass and energy conservation both, on a global and local scale.

(b) *Stable dynamic prediction*: Surrogate modelling has been used for predicting time-dependent stress and stress fields of heterogeneous solids subject to homogenous steady-state external loading conditions. Within this framework, these surrogate models can be regarded as dynamical systems. Given that surrogate models typically emulate stable physical behaviour thereby mimicking the basic rules of continuum mechanics, it is essential for them to possess inherent stability, i.e. mimicking also convergence. This stability ensures that predictions remain consistently stable and bounded. Consequently, it is imperative that ANN-based surrogate models are designed to have these stability properties inherently embedded.

(c) *Learning low-dimensional latent space representation*: Often, the field of interest in continuum mechanics is two or three-dimensional real space, ideally also informed by the solid's crystal and phase state (including also non-equilibrium features such as crystal defects and related inner structural descriptors, often also referred to as microstructure), adding further dimensions and anisotropy features to the problem to be solved. Consequently, the data obtained for these scenarios are high-dimensional, especially while dealing with high-resolution spatial fields. However, it is a common observation that such high-dimensional solutions can often be accurately represented in a low-dimensional latent space. The creation of this low-dimensional space is further guided by constraints designed to simplify the dynamics and engineering design processes. For instance, it is possible to construct a latent space in such a way that the system dynamics evolve in a nearly linear fashion, aligning with principles like Koopman theory and dimensionality reduction techniques.

(d) *Predicting multi-functionality of materials*: An aspect going beyond continuum mechanics is the quest for an efficient multi-dimensional descriptor representation of materials and the underlying predictive ANN and active learning models when it comes to the multi functionality of materials. This notion refers to a material design challenge, were not just the mechanical response but also its often non-linear interplay with functional properties such as magnetism, electrochemistry or electrical features is targeted. Example scenarios would be mechanically strong invar alloys and magnets [11], elastically compliant materials with high biocompatibility [117], or high-performance materials that are free of any critical and expensive elements [118] to name but a few examples. For such purposes the original low-dimensional descriptions are often insufficient and ANN models as well as latent space representations will increasingly have to embrace these multi-functionalities, because this is particularly a field where conventional simulation spaces become too large to be tractable by classical physics-based theory alone.

Advances to meet these challenges

In our pursuit of designing neural network architectures that inherently adhere to physical properties (e.g. divergence-free, energy preserving), we seek to utilize fundamental mathematical vector calculus. For illustration, in order to design ANNs to produce divergence-free quantities, we seek to obtain intermediate quantities so that divergence-free quantities are obtained by taking the curl of those intermediate quantities. Such techniques find widespread use in solving PDEs (e.g. Maxwell equation) with divergence-based constraints. Additionally, for achieving stable time evolutions through neural networks, we extend concepts proposed in [119] to encompass high-dimensional spatial and temporal data. Furthermore, our empirical studies indicate that CNNs that explore local features underperform compared to FNO, which explore global features present in the data. Therefore, our exploration centres on incorporating these physical properties within the context of FNO. We further need to explore how these trained networks are used for engineering studies such as predicting optimal material property configurations, drawing inspiration from [120]. What is more, we seek to discover suitable low-dimensional latent representation through autoencoders with the intent to simplify the task of predictions and engineering studies. Algorithmic developments in this direction have been pursued in [121], which requires further investigation in the context of continuum mechanics.

Concluding remarks

We conclude by emphasising that it is imperative to develop new ML and DL methodologies for tackling problems pertaining to the continuum mechanics of heterogeneous and anisotropic solids that adhere to the strong forms of essential physical principles both on a global and on a local scale. Doing so offers several advantages: firstly, it enhances the interpretability and generalizability of ML-based surrogate models. Secondly, it reduces the amount of required training data. Thirdly, it can enhance solver performance by up to several thousand times compared to conventional solution methods such as FEM or spectral methods. As an initial endeavour in this direction, we have demonstrated how to construct ML surrogate models that inherently produce divergence-free stress fields, thereby satisfying mechanical equilibrium conditions. Learning suitable low-dimensional latent representations not only reduces online inference time but also facilitates engineering studies with minimal computational resources. Additionally, acquiring training data for engineering applications is both economically expensive and time-consuming. Therefore, it is crucial to devise strategies for cleverly gathering training data, ensuring that the limited data covers a wide range of parameter space.

Acknowledgment

The authors acknowledge funding by the Max Planck Research Network on big-data-driven materials science BiGmax.

3.9. Digitalization of advanced experimental techniques for microstructures

Christoph Freysoldt[1], *Baptiste Gault*[1,2], *Christian H Liebscher*[1], *Pawan Goyal*[3] *and Jörg Neugebauer*[1]

[1] Max Planck Institute for Sustainable Materials, Düsseldorf, Germany

[2] Department of Materials, Imperial College, London, United Kingdom

[3] Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany

Status

Most modern engineering materials exhibit a complex microstructure that underpins the properties of the material in beneficial—or sometimes detrimental—ways. This applies to structural alloys, to ceramic materials like concrete or protective coatings, as well as to functional materials for energy storage, electronics, heterogeneous catalysis, etc. Steels, for example, consist of several meta-stable phases formed during casting, thermo-mechanical processing, or in operation. To image the interplay of grain morphology and texture, chemical composition, crystallographic relationships, and local properties of distinct regions, various complementary 'imaging' experiments are available, such as electron microscopy, APT, beam diffraction (electron, x-ray, synchrotron), or spatially resolved spectroscopy. Thanks to progress in experimentation, data storage capacities, and digital data processing, these techniques yield an ever-increasing data pool. A single experiment can provide GBs or even TBs of data, which is further multiplied by high-throughput experimentation (HTE) or *in situ* monitoring of transformations that add a time dimension. This big data is both a challenge and a great opportunity for data-driven research.

So far, it is mostly up to human experts to identify the microstructural features of interest within the experimental data. Often, it is not clear *a priori* what features relate to performance in the applied context. Once identified, one would like to quantify their number density, size distribution, chemical characteristics, and functional properties, in order to extract quantitative processing-microstructure-property relationships that facilitate material design. To automate this process, pattern recognition algorithms are actively being developed [122–124], often specifically targeted at a particular experiment for a particular type of material. Upon success, they provide a secondary characterization of the material in a reproducible and scalable way. This becomes particularly attractive when combined with HTE to systematically explore a material space.

Merging such derived data, possibly even from different experiments, with traditional materials' characterization across multiple samples while tracking their synthesis and processing history alongside necessitates a careful data management. Electronic lab books [125], integrated work-flow environments [72], structured material databases [126], and flexible data sharing platforms [67] cover some, but not all aspects. The barriers between them effectively limit data-driven material's design.

Current and future challenges

Suitable algorithms for pattern recognition are available from other fields, but must be adapted to a specific research question, see figure 14. Exploiting domain knowledge to define suitable descriptors and selecting robust algorithms [122, 123] will remain a scientific challenge in the coming years due to the vast variety of relevant phenomena and patterns. The

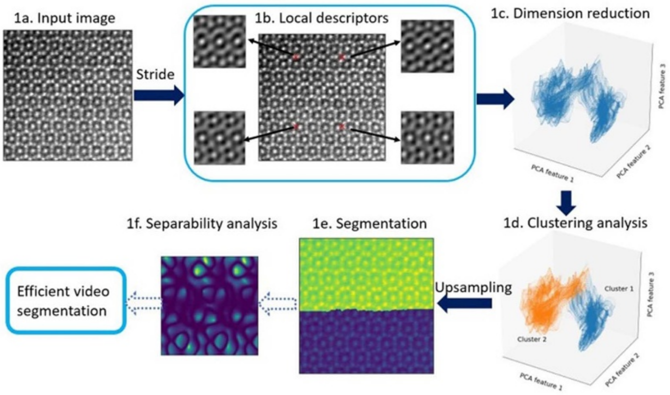


Figure 14. Crystallographic segmentation of atomically resolved STEM-HAADF frames via symmetry descriptors, clustering, and distinguishing feature selection for enhanced performance. Reproduced from [122]. CC BY 4.0.

actual integration of automatic microstructure evaluation in research practice is still at infancy. Progress is presently hindered by:

(1) A lack of established data and file formats. Experimental raw data is typically acquired in instrument-specific file formats. Extracting all potentially relevant data for ML workflows is often not possible or impaired. Community efforts have been undertaken to establish open data formats [127, 128]. An alternative effort aims at read-function libraries that support multiple formats [129]. For storing analysis output, ideally in conjunction with input data, no standard exists. Similarly, exchanging data between different data management systems is severely hindered by inherent heterogeneity in data structures and metadata, in the naming and unit convention of data fields, and by assuming implicit context (e.g. providing an instrument's name rather than its measurement parameters).
(2) A lack of flexible workflows or tool chains. Material science research routinely combines different characterization methods, but rarely so in a digitally integrated way. Researchers have their individual ways to document a material's synthesis and processing history, how each experiment's specimen was prepared, and how data was post-processed. Common approaches (via file name, free-form notes, folders, …) are ill-suited for automatic processing. Electronic lab-book systems help to manage those data [125], but typically reach their limits in collaborations across labs.
(3) A cultural gap between experimentalists accustomed to graphical user interfaces (GUIs) and programming-oriented data scientists. Present-day analysis strongly relies on humans to inspect the data. Instrument manufacturers therefore provide monolithic visualization tools with a GUI, that read the instrument's raw files, provide a fixed set of processing schemes and export results in established general-purpose image (jpg, tiff) or data formats (csv, hdf5) that drop context. In contrast, the wider ML field thrives on plugging together open-source libraries and code snippets on demand, that require significant coding skills.

Advances in science and technology to meet challenges

To reconcile the cultural gap, today's interactive data visualization and future advanced data processing must be interlinked. GUI-based visualization tools could open up by establishing plug-in mechanisms to exchange data and visualization items with external modules. An

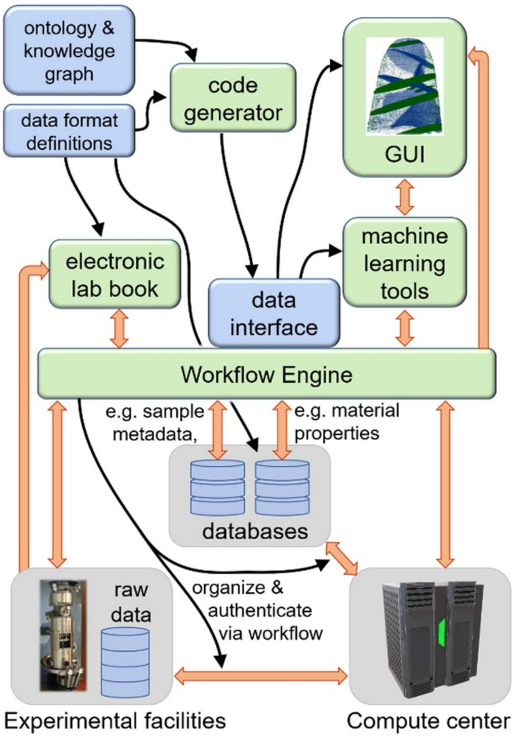


Figure 15. Sketch of a possible digital infrastructure for handling data from experimental imaging techniques in an integrated workflow. The experimental facility image has been extracted from [Electron microscope, RCA—Museum of Science and Industry (Chicago)—DSC06730.JPG] as obtained from wikimedia under a CC0-1.0 (public domain) license. Compute cluster image has been reproduced from Microsoft PowerPoint.

alternative route, that circumvents the GUI integration challenges, is to follow the successful model in *computational* material science [67, 72]: focus on input/output data format normalization, and employ separate tools that work with these formats for analysis and visualization, all coupled together by a managing framework, see figure 15. Further efforts to standardize input, but more importantly for recurring output such as classification signatures, segmentation maps, interface location, geometric shape information, etc are urgently needed.

In this context, exploiting automatic code generation from machine-readable data format definitions—in conjunction with ontologies and knowledge graphs—could be a game-changer to speed up the development, as they reduce the human effort in defining standards *and* implementing corresponding code for possibly different programming languages. Similarly, the trend towards higher abstraction in ML software should be exploited to generate processing metadata. When the transformation chain is built at run-time via high-level objects (which later generate the actual code for the hardware at hand on the fly), the high-level representation should automatically annotate the data output with the details of the processing chain.

At a higher level, workflow and data management tools must be adapted to deal with the specific challenges of experimental data. As experimental data sets can become very large, moving or copying around entire data sets is prohibitive. In most cases, raw data will be stored close to where it was generated. Computational resources for advanced ML might be located elsewhere, and only need part of the data, or specifically pre-processed data that reduces transfer size via dimensionality reduction or compressed sensing. Thus, workflows that deal with both distributed data and distributed computation will be needed, while maintaining consistency in metadata and ensuring that data access across computer systems is reliably authenticated to avoid premature publication or leakage of confidential data.

Concluding remarks

The success of data-rich imaging techniques in material science lies in the promise that materials' properties are linked to recurring patterns that can be discovered by inspecting a few representative examples. ML techniques can leverage this approach by removing the human inspection as the limiting factor to digest larger and larger amounts of data in order to discover relevant, but possibly rare patterns. At the same time, they offer the unique chance to characterize the underlying distributions in a statistically significant manner as more data becomes available, thus generating secondary high-level characterizing data that might serve as valuable descriptors for associated properties. Digitalizing the entire workflow from synthesis, sample preparation, data acquisition and post-processing in an integrated way as sketched in figure 15 is critical to achieve these goals.

Acknowledgment

This project received financial support from BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials Science.

3.10. Opportunities and pitfalls of using LLMs in materials science and engineering

Dierk Raabe[1], *Zongrui Pei*[2], *Junqi Yin*[3], *James Saal*[4], *Kasturi Narasimha Sasidhar*[5] *and Jörg Neugebauer*[1]

[1] Max Planck Institute for Sustainable Materials, Düsseldorf, Germany
[2] New York University, New York, NY 10012, United States of America
[3] Oak Ridge National Laboratory, Oak Ridge, TN 37831, United States of America
[4] Citrine Informatics, Inc., Redwood City, CA 94063, United States of America
[5] University of Wisconsin, Madison 53706, United States of America

Status

LLMs could grow into a transformative research tool in materials science, bridging the gap between text-based data and actionable insights. Opportunities lie in accelerated materials synthesis, discovery, processing and property design. LLMs can be used as indirect or direct tools. By indirect, we refer to situations where LLMs help in extracting data and building databases from scattered sources. This does not involve the actual process of materials discovery but it is a precursor step. Such databases can then be used by other ML methods. With their direct role in materials discovery we mean that LLMs can even extract causal relationships from collected data, serve to build domain-specific knowledge graphs, render hypotheses and guide progress-critical experiments, data collection and simulation [130–132]. The latter aspect is essential because LLMs do not obey any built-in causal rules. Instead, they connect language tokens in a probabilistic way, without considering logic, self-consistency or conservation laws. This means that they can violate elementary scientific rules. They mimic scientific context by using probability measures that rest on majority but not on proof or logic. This explains why there are opportunities but also pitfalls. The latter can be mitigated by combining LLMs with other methods such as classical theory, thermodynamics, kinetics, materials property data bases, explainable AI, active learning etc. LLMs are also capable of generating hypotheses and they can be used to build domain-specific knowledge graphs which in turn can enhance predictive models [133, 134].

Materials science stands at the confluence of several disciplines. Research topics range from latest quantum mechanical insights into the behaviour of electrons in complex systems to large-scale processing of billions of tons of material (concrete, steel) and materials exposed to harsh environmental conditions (catalysts, corroding products). Developing data-centric methods to leverage disruptive progress in this field must, therefore, reflect and embrace this heterogeneity in the underlying data from which knowledge can be extracted, combined and used.

In the portfolio of model-based AI methods, LLMs seem to offer new opportunities to discover materials and processes that may otherwise remain hidden in the complexity and scattered information that already exists [135]. One avenue to use LLMs is accelerated materials discovery [10, 136, 137]. This is due to the fact that language-based token systems that connect words based on probability are particularly strong in extracting and combining knowledge that already exists in text form. Therefore, while LLMs may not be necessarily suited for disruptive conceptual discoveries from text connections, they can accelerate design based on existing concepts [138]. Although this is a rather conservative approach, it is already a big step forward, because the traditional trial-and-error approach of material discovery is time-consuming and resource-intensive. Also, LLMs can analyse vast datasets, extracting patterns and correlations that would elude human researchers. For instance, LLMs can process published literature, patents, and experimental data to suggest combinations of novel material compositions and even possible properties, as will be shown below in more detail. By integrating databases like the Materials Project or the Cambridge Structural Database, LLMs can offer quantitative predictions about material structures, compositions, and potential applications, significantly reducing the time from conception to application.

However, it should also be noted that Krenn and Zeilinger [139] recently suggested a more disruptive approach to use LLMs. They introduced SemNet, a dynamic knowledge organization method in the form of a continuously evolving network, constructed from 750 000 scientific papers dating back to 1919. Each node in SemNet represents a physical concept, and a link is established between two nodes when the concepts are jointly explored in articles. SemNet has proven its utility by enabling the authors to pinpoint influential research topics from the past. The authors trained SemNet to forecast trends in quantum physics, and these predictions have been validated using historical data.

A few examples of using LLMs in materials science have been recently presented. Jablonka *et al* [131] conducted a hackathon using LLMs such as the generative pre-trained transformer 4 model (GPT-4) for chemistry and materials science. The participants leveraged LLMs for a variety of purposes, such as predicting properties of molecules and materials, creating new tool interfaces, extracting knowledge from unstructured data, and developing educational applications. Being more specific, An *et al* [140] argued that the construction of knowledge graphs for domain-specific applications like metal–organic frameworks (MOFs) can be resource-intensive. LLMs, particularly domain-specific pre-trained models, have been successfully employed to create such graphs. For example, a study explored the use of state-of-the-art pre-trained general-purpose and domain-specific language models to extract knowledge triples for MOFs [140]. The authors constructed a knowledge graph benchmark with 7 relations for 1248 published MOF synonyms. Experimental probing revealed that such domain-specific pre-trained language models (PLMs) outperformed general-purpose PLMs for predicting MOF related triples. The authors also conceded from their overall benchmarking results that the use of PLMs alone to create domain-specific knowledge graphs is still far from being practical and requires the development of better-informed PLMs for specific materials design tasks. The group of Olivetti used LLMs to generate knowledge graphs (MatKG2) for the entire domain of materials science, taking ontological information into account as opposed to using statistical co-occurrence alone [141]. Zhao *et al* [142] used fine-tuned bidirectional encoder representations from a transformer (BERT) model and tested it with respect to data extraction from published corpora. They reported that the model achieved an impressive F-score of 85% for the task of materials named entity recognition. The F-score is a metric used to evaluate the accuracy of a model in binary classification tasks. Sasidhar *et al* [143] integrated NLP and DL for the design of corrosion-resistant alloys [120]. They also highlighted the general challenges in utilizing textual data in ML models for material datasets and proposed an automated approach to transform language data into a format suitable for subsequent DNN processing. This method significantly improved the accuracy of pitting corrosion potential predictions for alloys, providing insights into the critical descriptors for an alloy's resistance to such type of environmental decay like configurational entropy and atomic packing efficiency. Pei *et al* [138] proposed a concept of 'context similarity' to select chemical elements with high mutual solubility for discovering high-entropy alloys. They trained a word-embedding language model with the abstracts of 6.4 million papers to calculate the 'context similarity'. With this approach they designed a workflow to design lightweight high-entropy alloys, which suggested even 6- and 7-component lightweight high-entropy alloys by finding nearly 500 promising metallic mixtures out of 2.6 million candidates.

Gupta *et al* [137] developed MatSciBERT, a materials domain-specific language model for text mining and information extraction. They argued that conventional language processing alone, such as encoded in the form of BERT models, may not yield optimal results when applied to materials science due to their lack of training in materials-specific notations and terminology. To address this challenge, the authors introduced a specific materials-aware language model they refer to as MatSciBERT. This model was trained on an extensive corpus of peer-reviewed materials science publications. The authors claimed that their model surpasses SciBERT, a LLM trained on a broader and less materials-specific scientific corpus, in three critical tasks, named entity recognition, relation classification, and abstract classification. The developers made trained weights of MatSciBERT publicly accessible, enabling accelerated materials discovery and information extraction from materials science texts. A recent study introduced a larger GPT version, named MatGPT [144], based on a larger scientific corpus than MatSciBERT. In their study the group claim that the MatGPT model embeddings outperform MatSciBERT and achieve an improved band gap prediction based on the Materials Project combined with GNNs.

Current and future challenges: LLMs and knowledge graphs for materials discovery

The current flagship in the world of LLMs is the GPT-4 from OpenAI. It is based on 8 separate models, each containing dozens of network layers and 220 billion parameters, which are supposedly linked together using the mixture of experts architecture. GPT-4 is built on a transformer architecture, combining self-attention and feed-forward neural networks to process input tokens. Each token represents a text string containing a word or phrase. Therefore, the token limit represents the amount of text that an LLM can consider at a given time as input. Early LLM releases had very low token limits since LLM calculation time is strongly dependent on token length. Initial releases of GPT3 had a token limit of 2048 tokens, but the recent release of GPT-4 has a token limit of 128 000 tokens. To put this into context, the average length of a PubMed abstract is 114 tokens (sd 48.83) and an article is 2378 tokens (sd 1604.79). So while increasing complexity of the LLMs has enabled using entire papers (or even groups of papers) as input, there is still the computational cost of running GPT-4 calculations to consider. As an example, when asking a question of medium complexity via a string of fewer than 10 tokens, such as ‘Composition and property ranges of material XY’, then the rough total cost estimate to answer this question for GPT-4 is about 7–10 Euros. Getting the same answer from a classical knowledge graph would incur only about one-hundredth of this cost and also take less time, provided the information is in the corpora and mapped in a graph accessible by search engines.

Using knowledge graphs also removes the hallucination effect, an error made by LLMs when rendering combinations that appear plausible to the model’s probability measures but false when tested against high-fidelity information or logic. It appears due to multiple factors, such as when LLMs are trained on contradictory datasets, overfitting, etc. An urgent and vital topic in LLMs is, therefore, quantifying the level of the hallucination effects and developing a systematic method to recognise and mitigate them. On the other hand, LLMs have the advantage that they can process and structure the context from scientific literature, patents, and database entries. When combined with knowledge graphs that can help to check and scientifically organize this information, it provides a rich database of materials science knowledge which can be readily queried. This integration allows for the rapid assimilation of existing knowledge and the identification of knowledge gaps.

Vice versa, LLMs, with their ability to process and generate large volumes of text, can also serve to construct domain-specific knowledge graphs, optimize algorithms for faster discovery, and enable more efficient design and exploration of materials. The synergy between LLMs and knowledge graphs could hence be a useful next step in materials discovery, offering a paradigm shift from traditional, iterative experimental methods to a more quality-controlled data-driven model. This combination would allow better alignment of reliable high-quality data exploitation (through knowledge graphs) and semantic contextualization (through LLMs).

LLMs can also analyse patterns and relationships within a knowledge graph to generate hypotheses-based suggestions for suitable search spaces pertaining to potentially novel materials and properties. For instance, by understanding the relationship between crystal structure and electronic properties, LLMs coupled to knowledge graphs could likely be used to suggest new compositions or corresponding search spaces for magnets, battery materials or solar cell absorbers.

Advances to meet challenges associated with the use of LLMs in materials science

While the opportunities are vast, applying LLMs in materials science also has challenges. Data quality and availability are critical as models are only as good as the data they train on. Ensuring data integrity and representativeness is paramount. Furthermore, the interpretability of LLM outputs is crucial for gaining trust in their predictions. Developing models that can provide not just predictions but also insights into the underlying mechanisms is an essential goal. Another point is the Chain of Thought Prompting, an approach to enhance LLMs' comprehension of causal relationships and reduce hallucination. It involves forcing the models to verbalize different steps of reasoning they have gone through in reaching conclusions. This makes the process more transparent. Such ideas have not been implemented in materials science but in other areas such as medical science [130].

The quality of the information that can be extracted from LLMs depends on the quality and timeliness of the input text. For material science that can be only achieved if the latest literature that has been going through proper peer review processes is being used. However, only one-third of the current scientific corpora is open access. Therefore, some of the corpora currently used for training LLMs is in part of questionable quality. Also, current LLMs might simply miss the latest literature. This means that the model weights are not fitted to the latest state of the art. These two aspects show that fine-tuning prior to the use of such LLMs is recommendable. On the other hand, recent literature sometimes also overlooks knowledge that already exists long in the literature so that some findings reported in papers are more like re-discoveries, a problem that can be likely mitigated when LLMs are used. In this context, is it worth emphasizing that APIs are increasingly being offered by a few companies to allow accessing millions of publications along with metadata. Another issue is that extracting text from PDF files, the standard format of the literature, results in poorly formatted corpora with numerous errors (e.g. missing text, insertion of text from other items such as tables in the middle of sentences, headers and page numbers, etc).

An unresolved open front of LLMs is the potential violation of existing copyright when tapping into web-based resources, which becomes an obvious issue with the use of journals, textbooks, and other scientific literature in training. Another concern is if further tuning of LLMs leads to slow asymptotic knowledge increase because high-quality peer-reviewed content on certain topics is not growing at a sufficiently high rate and is often not freely accessible for training. In other words, it is not likely that LLMs can gain and organize knowledge quicker than the generic basic research used to train them. To meet both challenges, the rapidly growing fraction of open-access literature and the use of pre-publication and self-archiving services is of great value, likely leading to higher quality improvement and less hallucination of LLMs. Some of these aspects also connect to general limit considerations regarding model capacity and scaling laws, which were recently shown to depend essentially on the number of model parameters, the size of the dataset and the amount of computation power used for training. Performance was shown to depend less on other architectural hyperparameters such as depth and width. However, irrespective of these theoretical considerations, it appears that the scientific community has not yet fully discovered or exploited the capacity limits of the GPT model in current applications. This means that for the same data size, the GPT model improved further as the number of parameters was further increased.

Concluding remarks

LLMs offer great potential in the complex interplay between advanced computational methods and the nuanced, often experimentally and empirically grounded field of materials science. Opportunities lie in accelerated material discovery; enhancement and improved pattern and result analysis of data obtained from existing computational tools such as atomistic simulations; better knowledge synthesis and data management from research articles, reports, and property studies; support in hypothesis development and outlier analysis; and advanced decision-making support in materials selection and design, including aspects such as costs, sustainability and regulatory constraints. Pitfalls exist regarding the quality, availability, bias and legal status of the training data; lack of built-in logic or conservation laws; lack of the reflection of microstructure, synthesis, sustainability and processing complexity; and the danger of over-reliance and even complacency regarding LLM predictions, i.e. the decay of individuals' own critical thinking, rigorous validation or falsification and the thrive towards deep understanding of the underlying causality behind phenomena, which are all essential key factors that have made the scientific method the most successful and reliable approach in human history.

This contemplation about a few generic pro and con aspects shows that while LLMs offer transformative potential in materials science, their successful integration into the field necessitates careful consideration of the quality and completeness of the data they are trained on, a thorough understanding of the underlying physical and chemical principles, and a balanced approach to leveraging their computational power with critical human expertise.

Acknowledgment

The authors acknowledge funding by the Max Planck Research Network on big-data-driven materials science BiGmax.

4. Material discovery and applications

4.1. High-throughput materials discovery with AI-guided workflows

Thomas A R Purcell[1,2], *Luca M Ghiringhelli*[1,3], *Christian Carbogno*[1] *and Matthias Scheffler*[1]
[1] The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society, Berlin, Germany
[2] The Department of Chemistry and Biochemistry, University of Arizona, Tucson, AZ, United States of America
[3] Department of Materials Science and Engineering, Friedrich-Alexander Universität, Erlangen-Nürnberg, Germany

Status

Computational, high-throughput materials discovery is seen as a promising route to advance a myriad of technologies including batteries [145], renewable energy [146], and pharmaceuticals [147]. With the increasing amount of computer power over the past several decades, several properties were calculated on millions of materials, with the aid of high-throughput workflow. These efforts led to the recent discovery of 2.2 materials below the convex hull by Merchant *et al* [148]. Such workflows allow a user to define a set of calculation parameters and run those calculations for a large set of materials. The results then populated several large databases, e.g. Materials Project, AFLOW, Open Quantum Materials Database, NOMAD, etc [149]. However, as the materials space is practically infinite, such studies can only address a marginal part of it, even for relatively simple properties.

The computational funnel model [150] extends high-throughput studies to complex materials properties by screening out materials after each step according to selection criteria based on the expected result. In theory, this means that the costliest calculations or experiments are done only for the most promising candidates. Naturally, this process is the more successful, the faster and the more reliable undesired materials can be disregarded.

Active learning provides one way of combining AI models and high-throughput workflows [151]. The goal of this class of algorithms is to balance exploitation and exploration to select new data points that can optimize a property or better train an AI model for a given material property, while simultaneously finding a global optimum for it in a data-efficient manner. By using an acquisition function that finds this balance, these frameworks can improve the efficiency of the studies by selecting which materials enter the funnel in a non-subjective manner (see section 2.1). In essence, both the active learning framework and the selection funnels attempt to achieve the same goal synergistically: active learning suggests which materials enter the workflows and the high-throughput funnel removes the unpromising candidates after each step of the workflow.

In fact, active learning already has been used in several different applications using DFT and materials discovery frameworks. The need for these frameworks is highlighted in a recent publication from Li *et al* who demonstrated that 95% of existing data is redundant and that uncertainty-based active learning frameworks can create smaller, but as effective datasets for ML [152]. Hengrui and coworkers recently developed an entropy-targeted active learning to explore parts of materials space that is under-explored, complimenting these results [153]. Finally active learning codes can also be applied to other problems, such as searching configurational space of molecules on surfaces [154].

Current and future challenges

The main challenge in fully realising the potential of AI-guided workflows is integrating active learning schemes, as well as the AI models and the suited acquisition function they are based on, into advanced selection funnels. The criteria used for each step of the funnels are either based on an expected error bound of a lower accuracy calculation or a physics-informed heuristic, e.g. a material having too large of an electronic band gap or being too dense. The end goal of the screening criteria is to reduce the overall cost and time of a study, while still exploring the relevant parts of materials space.

While useful, the current screening criteria are a potential obstacle when combining active learning with high-throughput workflows. Because they are not necessarily derived from the data that underlies the AI-model, an overzealous screening procedure can exclude materials that would drastically improve model performance and possibly correct an initial bias. Importantly, the heuristics used to screen out materials may not directly relate to the target property, but be controlled by an unknown third process, leading to an incorrect physical interpretation. Furthermore, adding selection funnels to active learning frameworks could perpetuate the initial bias of the models as the dataset will be directed towards the existing conditions. One potential solution to this problem is through using multi-objective learning to simultaneously optimize both the screening criteria and the target property. However, a less complex solution would be preferable.

The final challenge with creating these workflows is to incorporate them into existing materials discovery frameworks. Currently, the tools used for materials discovery such as AFLOW [155], atomate [156], and Aiida [157]. Without native integration, multiple, potentially incompatible solutions must be created leading to a less transparent ecosystem. An additional benefit of fully integrating these methods is an improved selection procedure. The use of cost-aware and efficient acquisition functions is becoming increasingly popular [158], and including the AI model training and selection steps inside the workflow libraries themselves will improve the estimated costs for these acquisition functions and multi-fidelity approaches.

Advances in science and technology to meet challenges

An expanded use of explainable AI methods presents a clear path to achieve the necessary combination of methods presented above. By learning the conditions for screening out materials from the AI models themselves, explainable AI attempts to expands the predictive power of ML models, and give insights into the relationship between the input physico-chemical materials features and target properties. These methods can relate to either the regression method used, e.g. linear or symbolic regression, or post-processing techniques that uncover the relationships. By better understanding the connections between the input features and a target property, one can then replace the physics-derived heuristics with ones from the model itself.

We recently demonstrated the capabilities of this approach, by creating an AI-guided workflow for finding thermal insulators [159]. For this project we modelled the thermal conductivity, κ_L, of a material based on its structural, harmonic, and the anharmonic properties (see [159] for a complete list). We then applied feature importance metrics, and found only three inputs were important. From here we were able to map the expected value of κ_L against each of these inputs to find the screening procedure highlighted in figure 16(b). With this workflow we were able to efficiently find 16 predicted ultra-thermal insulators with a κ_L less than 1 W mK^{-1} out of an initial set of 732 materials [159].

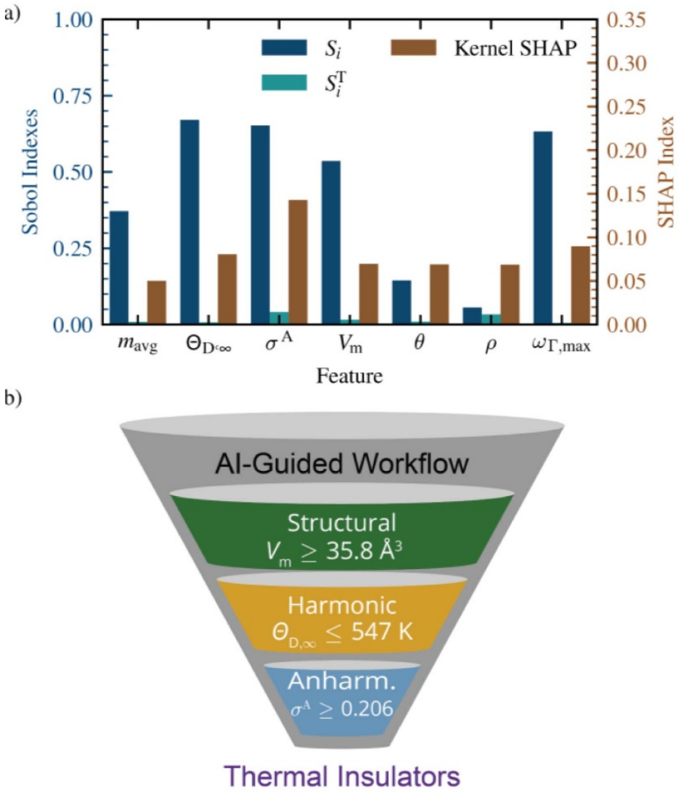


Figure 16. Example of the proposed workflows. (a) The SHAP (brown bars) and Sobol indexes (blue bars) for the AI model found in [159] for a model of the thermal conductivity of a material using its structural (average mass, m_{avg}; density, ρ; molar volume, V_{m}; and reduced mass, μ), harmonic properties (Debye temperature, $\Theta_{\mathrm{D},\infty}$, and the maximum Γ-point frequency, $\omega_{\Gamma,\mathrm{max}}$), and the anharmonicity factor, σ^{A}. (b) The workflow obtained from the expected thermal conductivity for a given input features of the most important inputs. Reprinted from [159].

To fully address the challenges associated with creating sustainable, AI-guided workflows, active learning techniques must be integrated into them. While the selection funnels can find a list of hundreds of possible candidate materials, it cannot identify which predictions are the most important to validate next. However, introducing an acquisition function the workflows can then maximize the quality of information gained per calculation or experiment. In turn this will allow us to speed up the discovery of good materials for vital applications. More importantly, by redoing the feature importance study after each iteration we can further refine the screening criteria and continue calculations that were initially discarded because they broke one of the old metrics.

Concluding remarks

AI-guided workflows have the potential to revolutionise materials discovery frameworks by focusing calculations or experiments on the most promising materials, and potentially remove the initial bias of data selection. By using an appropriate acquisition function to determine which experiments or computations to run next, we can automate these calculations. In turn the focus of the researchers working on these problems can instead be on further developing new methods and not managing a large set of calculations. Furthermore, explainable-AI methods will help elucidate why the models are deciding which candidates to calculate next. With this insight, part of the physical mechanisms driving, facilitating, or hindering the different processes may also be understood. Finally, as the frameworks become better focused the overall efficiency of these efforts will be significantly enhanced.

Acknowledgment

This project was supported by the NOMAD Center of Excellence (European Union's Horizon 2020 research and innovation program, Grant Agreement No. 951786) and the ERC Advanced Grant TEC1p (European Research Council, Grant Agreement No. 740233).

4.2. Roadmap for big data and AI driven data analytics in scanning/transmission electron microscopy (S/TEM)

C H Liebscher[1,2], *G Dehm*[1], *C Freysoldt*[1], *A Leitherer*[3,4] *and L M Ghiringhelli*[3,5]
[1] Max Planck Institute for Sustainable Materials, Düsseldorf, Germany
[2] Present address: RC FEMS & Faculty of Physics and Astronomy, Ruhr University Bochum, Bochum, Germany
[3] The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society, Berlin, Germany
[4] Present address: ICFO-Institut de Ciencies Fotoniques, The Barcelona Institute of Science and Technology, Castelldefels (Barcelona), Spain
[5] Department of Materials Science and Engineering, Friedrich-Alexander Universität, Erlangen-Nürnberg, Germany

Status

Recent developments in aberration-corrected electron optics, spectrometer and detector technologies enable to capture multimodal signals within a single experiment in a S/TEM down to the atomic level. These advancements have greatly expanded our understanding of the atomic constitution of materials, which is largely driven by the rich and multimodal data streams. Spectroscopic techniques such as energy dispersive x-ray or electron energy-loss spectroscopy (EELS) can nowadays probe the local composition and electronic structure of complex materials at the atomic level. New scanning diffraction methods, termed 4D-STEM, capture 2D electron diffraction patterns in each probe position of the 2D raster grid and have facilitated to image light elements at atomic resolution, determine local structures and strain with subnanometre precision [160]. Further, the spectroscopic and 4D-STEM techniques can be combined with tomographic approaches to obtain the 3D nature of materials. Advances in *in situ* probing capabilities and fast electron detectors make it possible to directly observe the dynamic evolution of materials under different external stimuli with high spatial and temporal resolution. The common theme of these techniques is that nowadays the experimental data is often represented as a three- or higher-dimensional data set as shown in figure 17 (left) [161].

The ever-growing data complexity, size, and speed at which it is created in experiment renders human-based analysis not only impractical, but also largely limits the discovery of latent features, which often equip a material with a certain functionality [161]. This has stimulated the development of automated computer-based and ML analysis algorithms to harvest the rich information contained in the data and to turn the data into interpretable physical quantities [160]. For example, principle component analysis and clustering were employed to automatically separate different phases in a bismuth ferrite sample at atomic resolution obtained from a multi-gigabyte 4D-STEM data set [162]. The development of open-source-based data-analysis tools has been paramount for treating and interpreting multidimensional and large-scale data sets from different microscope manufacturers in an efficient manner and provide flexible platforms towards on-the-fly data analysis even of big data sets [163].

Current and future challenges

Incremental data acquisition and analysis are still common even in modern microscopy laboratories. The experiment is sequentially followed by the interpretation of the collected signals and eventually the loop repeats with refined measurements until sufficient insights into physical

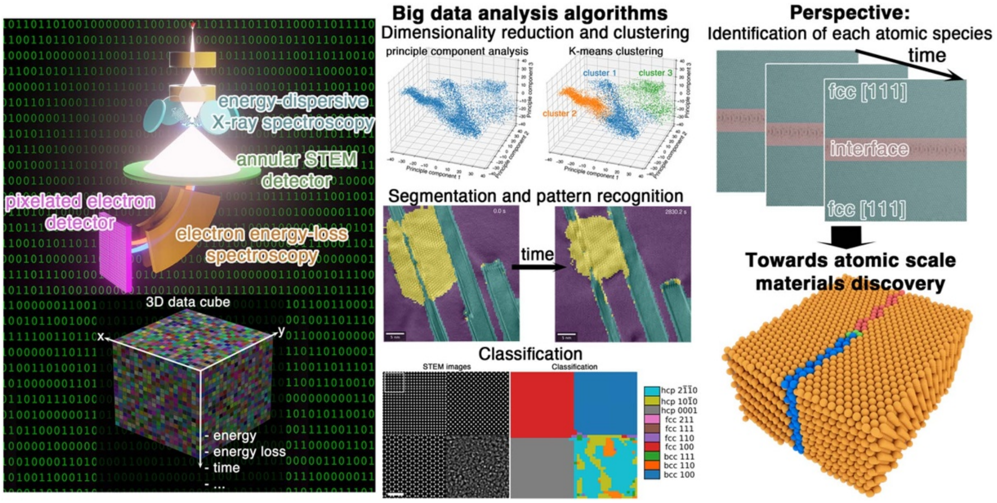

energy-dispersive X-ray spectroscopy
annular STEM detector
pixelated electron detector
electron energy-loss spectroscopy
3D data cube
x
y
- energy
- energy loss
- time
- ...
Big data analysis algorithms
Dimensionality reduction and clustering
principle component analysis
K-means clustering
cluster 1
cluster 2
cluster 3
Segmentation and pattern recognition
time
Classification
STEM images
Classification
hcp $2\bar{1}\bar{1}0$
hcp $10\bar{1}0$
hcp 0001
fcc 211
fcc 111
fcc 110
fcc 100
bcc 111
bcc 110
bcc 100
Perspective:
Identification of each atomic species
time
fcc [111]
interface
fcc [111]
Towards atomic scale materials discovery

Figure 17. Multimodal data streams and related high-dimensional data representation (left). Data-analysis algorithms for dimension reduction of large-scale data, automatic pattern recognition/segmentation and quantitative classification of the data (middle). Perspective to harvest the variety of signals and content contained in big data to uniquely identify the 3D physical structure of a sample with atomic precision, its evolution with high time resolution to discover new material phenomena on the atomic and electronic scale. Reproduced from [164]. CC BY 4.0.

material quantities are gained. There are several challenges associated with this incremental approach in the era of big microscopy data:

(1) handling, storage, and labelling of the data to enable reproducible data analysis
(2) human-based data analysis often largely exceeds experimental time frames
(3) limited interdigitation of data acquisition and analysis
(4) lack of automated or autonomous data analysis tools.

These technical restrictions often directly compromise material characterization and with this new material discoveries. One of the greatest challenges is the interdigitation of data-stream generation in a microscopy experiment and its direct analysis to provide live feedback to the researcher. Different approaches can be envisioned here where parallelized high-performance computation (HPC) utilizing modern GPU capabilities is directly performed at the microscope computer [165] or edge computing in a distributed system, where the HPC tasks are performed either on cloud servers or at HPC centres [166].

The broad variety of data streams utilized to probe materials ranging from simple 2D images to 3D or higher dimensional hyperspectral data sets, to time series probing material evolution or 3D tomographic reconstructions require the development of versatile and autonomous data analysis algorithms. Typically, advanced algorithms to reduce the dimensionality of hyperspectral data, segment or recognize patterns in images, and classify features in multidimensional data sets are employed as separate or sequential instances as shown in figure 17 (middle) [160, 161]. It has been shown that unsupervised ML is capable to automatically segment different crystalline regions in atomic resolution images and video sequences solely based on crystal structure symmetry without requiring prior knowledge on the underlying structure [122].

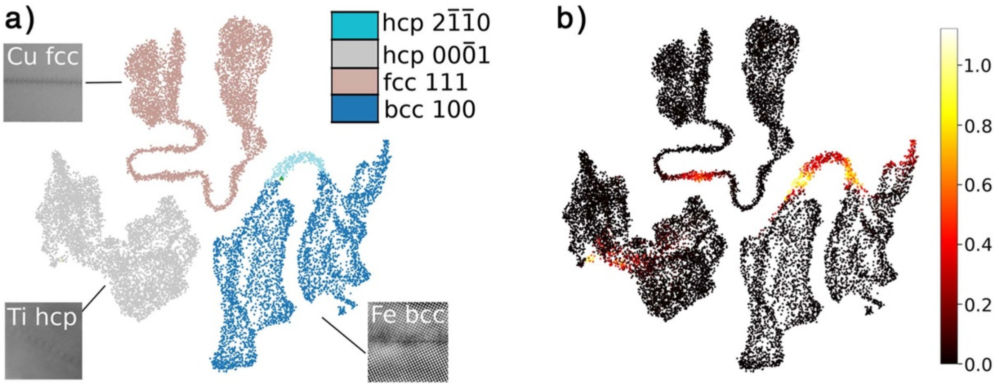


Figure 18. Dimension reduction of neural-network representations of a classification model trained on simulated atomic resolution STEM images of pristine crystal structures. Each point in the scatter plots corresponds to a local image patch of an experimental image. The colour scale corresponds to two items of information that the model provides: (a) Classification assignments of experimental STEM images of interfaces, here grain boundaries, in face-centred cubic (fcc) Cu, body-centred cubic (bcc) Fe and hexagonal close packed (hcp) Ti. (b) Mutual information quantifying the uncertainty of the prediction of the deep learning model. Bulk regions appear as clusters of low model uncertainty while interfaces correspond to diluted regions with high model uncertainty. Reproduced from [164]. CC BY 4.0.

Using a trained Bayesian DNN, it is even possible to classify crystal structures in atomically resolved images and identify defective regions or interfaces by considering the uncertainty in the prediction [164]. In a future direction, one would envision that novel big-data and ML algorithms will be integrated in hybrid algorithm architectures that perform automatic or even autonomous tasks.

Advances in science and technology to meet challenges

Advances in computing architectures for microscope laboratories are one side of the coin, but integrated or hybrid ML based algorithms need to be deployed alongside to enable automatic analysis of large-scale data. Recent developments in ML and in particular DL approaches in electron microscopy hold great promise for laying the foundation for autonomous data-analysis and electron-microscope operation [160, 167]. Ultimately, the aim is to enable the discovery of new material phenomena and to probe the physical properties of materials and their evolution with atomic precision. Since the physical nature of electron wave propagation and interaction in a crystalline material is well understood, ground truth training data for a DL model can be efficiently generated [168]. However, a large DL model would need to contain information not only of all known crystal structures and phases, but more importantly of different point, line, or planar defect types. Recognizing defects from supervised learning, however, is nearly impossible to achieve at the day of writing, since the atomic configurations existing in nature are not necessarily known or understood. Instead, a CNN can be trained on simulated images of pristine crystal structures, while still localizing and obtaining information on material imperfections [164].

Figure 18 shows the neural network representations obtained after dimension reduction of the fully connected layer before the classification and the corresponding uncertainty of the prediction. Although the model was trained on pristine crystal structures, it is capable to distinguish the different types of interfaces (here: grain boundaries) and the model uncertainty provides an indirect way to locate material imperfections. Furthermore, the unsupervised-learning analysis of the structure of the latent space, might be able to identify OOD interface structures that are similar to one another while significantly different from the training data. Until now, the model cannot relate these interface structures to known or unknown building blocks of the interface. Approaches combining supervised, unsupervised and active learning are needed to further explore regions in data sets with high uncertainty, which may represent an unknown interface structure or surface configuration. Furthermore, the classification tasks have to be extended to also consider local composition and electronic structure to fully exploit the data and yield a holistic picture of the physical nature of a material on the atomic level. Future models should enable live feedback at high time resolution to facilitate autonomous steering of the experiment and consider active re-training to include disturbed or unknown atomic structures.

Concluding remarks

Big data in electron microscopy is already a reality and will play an increasing role in the future not only for the sake of data acquisition, but to holistically characterize every single atom in a material paving the way for atomic scale materials discovery. Spectroscopic and scanning diffraction data sets (e.g. 4D-STEM) contain information on the elemental nature, the electronic and 3D structure of a material and hence this information needs to be fully harvested. Technological advancements in computing infrastructure have to be developed in parallel with hybrid ML algorithms in electron microscopy laboratories to move away from incremental experimentation. Combinations of unsupervised and supervised learning approaches have the potential to automatically identify and label different crystal structures and atomic species in complex data sets and will eventually uncover latent patterns in an automatic fashion. This will guide scientists to interesting regions in a sample and accelerates the deployment of physical material models. Ultimately, novel hardware developments will be needed that can make independent decisions on the next measurement steps to reduce the generated amount of data, while still providing essential information on the underlying physical nature of the material.

Acknowledgment

The authors acknowledge founding by the Max Planck Research Network on big-data-driven materials science BiGmax.

4.3. Applications of ML in the acquisition of and knowledge extraction from experimental data with a focus on electron microscopy

Marcel Schloz[1]*, Anton Gladyshev*[1]*, Meng Zhao*[1]*, Thomas Kosch*[2] *and Christoph T Koch*[1]
[1] Department of Physics & CSMB, Humboldt-Universität zu Berlin, Berlin, Germany
[2] Department of Computer Science, Humboldt-Universität zu Berlin, Berlin, Germany

Status

The success of inherently data-based ML in materials science can also be observed in its subfield of structure research through (S)TEM [160, 169]. Here, ML has become a game changer for post-acquisition data analysis, such as image reconstruction [170], improvement of data by denoising and resolution enhancement [160, 169, 171] and structure recognition [160, 168, 169, 172, 173]. One of the bottle-necks for the efficiency with which electron microscopes can generate materials knowledge is also the investment of time and human, highly microscope-specific expertise required to align the instrument for optimal performance, especially, when the materials question to be solved requires switching between different modes of operation. While some data-driven ML models have already demonstrated to be capable of measuring aberrations very quickly [174], they are not yet capable of handling the complexity of a modern microscope which, for some instruments, requires managing more than 500 current supplies. A few groups are also applying ML methods towards real-time data analysis and automating experiments as illustrated in figure 19 for the case of STEM [175]. In contrast to the field of cryo-electron microscopy, where fully automated experiments can run for multiple days by repeating the same image acquisition process for automatically exchanged samples at many pre-defined sample positions, the complexity of adaptive ML-driven experiments in materials science (S)TEM experiments is much higher, given the inhomogeneity of most samples, the wide variety of signals to choose from and switch between, and the sequential process with which the data is acquired. Conventional ML methods used on already acquired data sets can simply not be applied one-to-one. New ML approaches for real-time applications in electron microscopy are still rare and most notably their on-the-fly implementation on the microscope has so far not been realized [175].

Current and future challenges

In addition to the requirement for very fast data processing and fast access to electron optical components of the microscope, method developments will also need to consider the following two crucial components: The first key component is the fast handling of huge microscopy data. Electron microscopes can nowadays acquire several GBs of data within seconds which, means that ML methods for real-time applications should be capable of processing huge data sets within a fraction of a second. Obviously, a tight integration between hardware and software will play a crucial part in the solution to this problem. Edge computing and camera integrated compression techniques [176] are here just two examples to be mentioned. Another important component for the development of new real-time ML methods is a high level of adaptability. The environment in the microscope constantly changes between, but sometimes even during experimental sessions. ML methods need to deal in real-time with data that has been acquired under these circumstances without a significant loss in performance. Furthermore, methods that aim for an automation of the experiment are required to easily adapt to different experimental goals.

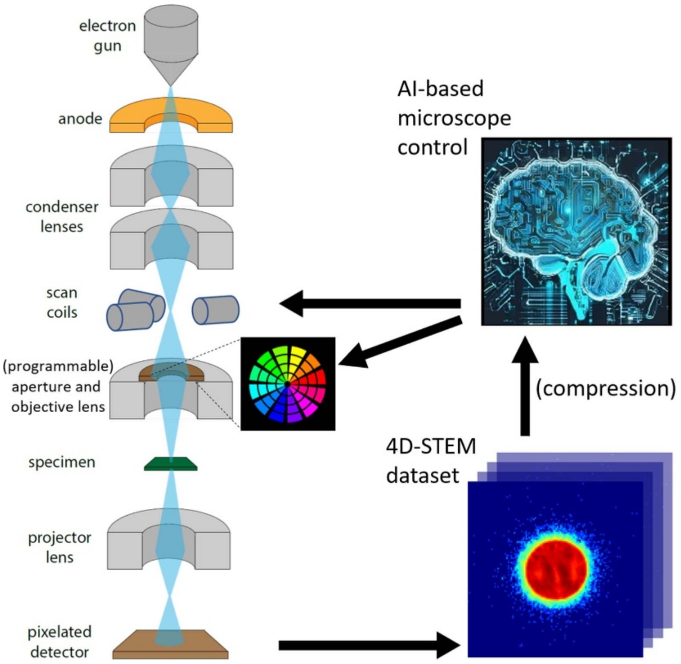


Figure 19. Schematic of an AI-controlled scanning transmission electron microscope using a pixelated detector to acquire 4D-STEM datasets. Electrons are emitted from the electron gun and guided through a system of electromagnetic lenses and are deflected by scan coils before they interact with the specimen. The electron beam is then guided to the detector, which records a diffraction pattern of approx. 1 MB size for thousands of scan position, thus, resulting in a 4D-STEM dataset of typically 10 s of GB in uncompressed size. ML methods analyse the raw or compressed data in real-time and control the hardware components of the microscope to optimize the experiment. The controlled components shown here are the scan coils and a programmable phase plate (inspired by the commercially available design by adaptem.eu), but it can also be lens currents, aberration-corrector settings, etc.

The high complexity and cost of ownership of state-of-the-art electron microscopes allows only a few labs staffed with expert operators who have undergone extensive microscope-specific training to run them. Maximizing these instrument's scientific output per time as well as democratizing access to them calls for improving their user interface in analogy to how modern chatbots have recently started to enable anybody to write complex computer programs.

Advances in science and technology to meet challenges

Advances in the method development that combines DL and reinforcement learning (RL) show promise that dynamic decision-making problems can be solved with a strong performance by a machine alone. Operating an electron microscope in an automated fashion could therefore benefit from this development. A first step towards this direction has been proposed in [175], where the combination of DL and RL offers the possibility to perform low-dose experiments for electron ptychography through adaptive scanning. A schematic of the adaptive scanning workflow is shown in figure 20. The advantage of this approach is that it is highly adaptable to a wide range of scanning microscopy techniques through the modification of a reward function

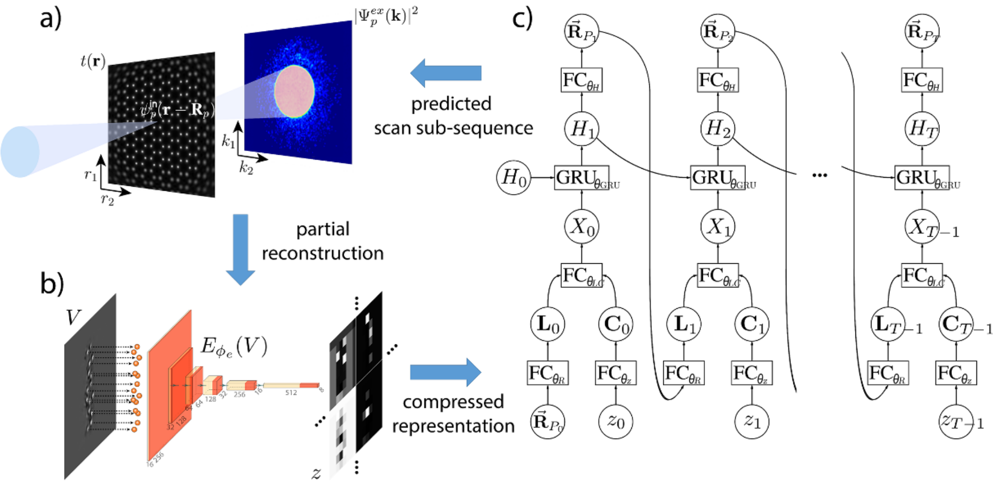


Figure 20. Schematic of a ML-driven adaptive scanning workflow for the purpose of optimizing the scanning in a 4D-STEM experiment in real-time. The employed ML methods consist of a convolutional neural network for the atomic structure extraction and a recurrent neural network for the sequential prediction of scan positions. Training of the networks is performed through RL. Reproduced from [175], with permission from Springer Nature.

that expresses the research goal. Hence, various imaging and spectroscopy techniques, such as STEM EELS, that have already been shown to benefit from an optimized scanning scheme, could be further advanced through a successful automation of the experiment. But also many other parameters of the experiment, such as adjustable aberrations, lens currents, or the phase shifts in programmable phase plates (see figure 19) can be optimized to improve the efficiency of the experiment with which a given research question is being addressed. Recent developments in software for processing natural language are likely to result in the highly technical user interface of electron microscopes being extended by chatbots and comparable features.

In order to deal with 10 s of GB of data per scan, it has been shown [176] that compression based on data-dependent linear transformations yields superior results when compared to conventional techniques like binning, or singular value decomposition, both in terms of compression ratio and quality of the information extractable from the data-set. The integration of ANN-based feature recognition techniques has the potential to further enhance compression performance. Before collecting the main data-set, a network can be pre-trained in a similar way to adaptive scanning [175], but with the aim of capturing the diffraction patterns as best as possible with as few values as possible.

Concluding remarks

In summary, data-based ML methods have already shown to be very powerful for post-processing tasks of electron microscopy data, but given the high complexity of these microscopes, their application in useful real-time data analysis and experiment automation methods still lags behind. Some initial developments of workflows that leverage ML methods to perform and optimize specific tasks of an electron microscope show promise for transitioning this fully human-controlled instrument to a (partially) autonomously operating machine being capable of carrying out precision measurements in a fully documented and fully reproducible manner. We expect that this development will largely increase the research output obtained from this type of instrumentation.

Acknowledgment

M S and C T K acknowledge financial support from the DFG—Project-ID 414984028—SFB 1404.

4.4. ML for analysing APT data

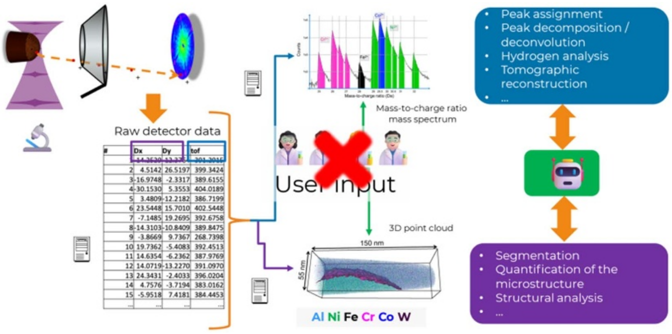
Yue Li[1]*, Ye Wei*[2]*, Alaukik Saxena*[1]*, Christoph Freysoldt*[1] *and Baptiste Gault*[1,3]
[1] Max Planck Institute for Sustainable Materials, Düsseldorf, Germany
[2] Ecole Polytechnique Fédérale de Lausanne, School of engineering, Lausanne, Switzerland
[3] Department of Materials, Imperial College, London, United Kingdom

Status

APT is a burgeoning characterization technique that provides compositional mapping of materials in three-dimensions at the near-atomic scale [177]. The data obtained by APT takes the form of a mass spectrum, from which the composition of the analysed material can be extracted, and a point cloud that reflects the distribution of all the elements within the region-of-interest of the material being studied. Material-relevant data must be extracted from this point cloud through the use of data processing or mining techniques. These go from simply the local composition of a phase or a microstructural object, sometimes extracted via cluster-finding or nearest-neighbour algorithms today classified as ML but used in the APT community for many decades [178]. Phase morphology or even partial structural information can be obtained but the information can be limited or distorted because of trajectory aberrations that are caused by heterogenities in the specimen's end shape down to the near-atomic scale. Today, data reconstruction and processing is most often done in commercially-available software, which does not allow for exploiting the cutting-edge methods arising from big data and ML, and also remains very much user-dependent [179, 209]. The enormous potential to mine atom probe data is clear, but this requires complete FAIR-compliant analysis workflows that make use of ML to facilitate more reliable and reproducible data processing and extraction, to really go beyond what human users can achieve. This section reviews challenges of APT data analysis (partially) solved by the application of ML and points out the remaining crucial locks to be addressed in the future.

Current and future challenges

A critical challenge is that present-day APT data processing tools and workflows are inherited from 'traditional' interactive data analysis based on user-interactions, through a fixed set of data analysis techniques and visualization that leaves little flexibility to explore novel and processes, as summarised in figure 21. User-input includes assignment of peaks to particular atomic or molecular species to manually retrieved structural information and microstructure segmentation and quantification. ML has the potential to automate many of these analysis steps, with models that are based on physical input and constraints. Some progress has been made across the community with dedicated ML algorithms to mine compositional [180] and structural information [181]. For instance, for mass peak assignment, we introduced an approach that uses known isotopic abundances to identify patterns in mass spectra, outperforming human users without loss of accuracy [182]. Following reconstruction of the 3D point clouds, automated identification and quantification of grain boundaries were proposed, and for more general microstructure segmentation, Saxena *et al* [123] introduced an approach that uses clustering in the compositional space, demonstrating unique capabilities for segmentation of the various phases, along with the quantification of their composition and morphologies. These would normally have been extracted through manually positioned regions of interest, which is time-consuming and error-prone. Structural imaging by APT is hindered by the anisotropic



#	Dx	Dy	tof
1	-14.2520	-12.37[illegible]	391.2015
2	4.5142	26.5197	399.3424
3	-16.9748	-2.3317	389.6155
4	-30.1530	5.3553	404.0189
5	3.4809	-12.2182	386.7199
6	23.5448	15.7010	402.5448
7	-7.1485	19.2695	392.6758
8	-14.3103	-10.8409	389.8475
9	-3.8669	9.7367	268.7398
10	19.7362	-5.4083	392.4513
11	14.6354	-6.2362	387.9769
12	14.0719	-13.2270	391.0970
13	24.3431	-2.4033	396.0204
14	4.7576	-3.7194	383.0162
15	-5.9518	7.4181	384.4453
...	...	...	...

Figure 21. Summary of typical APT data analysis workflow, from the processing of the experimental data to form and analyse the mass spectrum to the reconstruction of the 3D point cloud, all of these steps typically require user input, highlighting potential for ML-learning and possible complete data processing workflows.

spatial resolution and the limited detection efficiency [178]. Recent efforts have managed to overcome these for the ever-challenging analysis of chemical short-range order (CSRO) by using CNNs, using the workflow in figure 22 [183]. A key challenge for the future is to move away from the developing individual tools to tackle isolated problems to think about complete data analysis workflows, from patches to a logical patchwork that will also facilitate adoption. A way to solve this would be to open the programs themselves via APIs at all levels, or at least facilitate data exchange through open data formats accessible to external processing by independent tools.

Advances in science and technology to meet challenges

For APT, post-processing is mostly executed with proprietary software tools, which can often be opaque in their execution and have often limited performance and preclude the facile deployment of novel data processing methods. There is a need to agree on a more opened data format and metadata conventions as a critical prerequisite. Development of ML optimized hardware and software remains plagued by the use of proprietary, specific data format, which limits usage across software, techniques and communities. And this should include the raw data, not only what has already been processed. As such, the community will have push to provide a complete set of tools equivalent to the currently available integrated beginning-to-end workflows, i.e. from an experiment to a publishable image, yet these will have to be open and extendable to include ML steps and fully documented to also include traceable information regarding the sample and the specimen with appropriate metadata. A prerequisite is also the use of open and documented data formats. As a preliminary effort in this directions, let us mention here Paraprobe [184], that is fully open-source and provides clear documentation of each analysis step for post-processing APT datasets that offers orders of magnitude performance gain, automation, and reproducibility. For now, these open tools are seldom used, and the community seems to wait user-friendly platforms, which so far do not exist. This hinders complete FAIR-workflows that are so far lacking, which precludes direct correlations with other computational or experimental techniques, but also wider meta-analyses as introduced by Meier *et al* [180]. Finally, there is a need for a repository of benchmark datasets that would allow to evaluate the performance of new developments in a transparent way across the community.

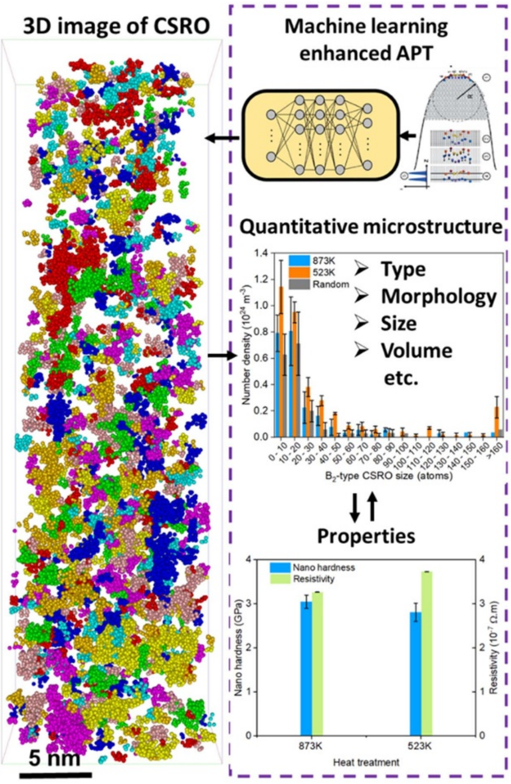


Figure 22. Machine-learning enhanced APT to break the inherent resolution limitation of atom probe tomography, and precisely image multiple arrangements of atoms associated with CSRO, in 3D. Reproduced from [183]. CC BY 4.0.

Concluding remarks

Although the above-mentioned efforts have demonstrated the potential for state-of-the-art ML to meet existing challenges for APT data processing, some major aspects remain to be tackled to fulfil the full potential. ML has the potential to help address many of APT shortcomings, and, for instance, resolve some aberrations that plague the accuracy of the measurements by better interfacing with modelling efforts in APT. This is necessary to reach true atomic-resolution that will help extract more precise local atomic arrangements. Optimisation of the data acquisition, and establishing a dialog between the instrument and the data processing are also areas that will need exploring—in this regards, APT is far behind other high end microscopy techniques. Overall, we are only at the beginning of the use of ML for APT, but the preliminary work that has been done across the community lays solid ground to build better, more encompassing and efficient tools in the future.

Acknowledgment

Y Li acknowledges the research fellowship provided by the Alexander von Humboldt Foundation. A S appreciates funding by Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE). BG acknowledges support from the Deutsche Forschungsgemeinschaft

(DFG) for funding from the Leibniz Prize 2020 (GA 2450/2-1) and for funding of the TRR 270 HoMMage (INST 163/578-1). B G, Y W, Y L and C F are grateful for financial support from BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials Science. The colleagues from the Max Planck Computing and Data Facility, Garching, Germany are warmly acknowledged for their help and support.

4.5. Data-driven approaches for heterogeneous catalysis

Andrew J Logsdail[1,2], *C Richard A Catlow*[1,2,3,4], *Lara Kabalan*[1,2,5], *Igor Kowalec*[1,2] *and Zhongwei Lu*[1,2]

[1] Max Planck Centre on the Fundamentals of Heterogeneous Catalysis (FUNCAT), School of Chemistry, Cardiff University, Cardiff, United Kingdom

[2] Cardiff Catalysis Institute, School of Chemistry, Cardiff University, Cardiff, United Kingdom

[3] UK Catalysis Hub, Research Complex at Harwell, RAL, Oxford, United Kingdom

[4] Department of Chemistry, University College London, London, United Kingdom

[5] STFC Hartree Centre, Daresbury Laboratory, Daresbury, Warrington, United Kingdom

Status

Heterogeneous catalysis is vital to sustain humanity and to address important societal challenges such as achieving net zero. Heterogeneous catalysis is also chemically complex, and the realisation of new catalysts is challenging. Catalysts themselves can contain multiple active elements; for example, the catalyst used for the Haber–Bosch process, which is integral to feeding 50% of the global population, typically contains iron, aluminium, calcium, potassium, and oxygen, with activity subtly dependent on composition [185, 186]. The composition of catalytic materials can be explored successfully via data-driven approaches, yet catalytic reactions occur at the surfaces and interfaces of these materials, and therefore the material properties must be investigated also as a function of the reactive surfaces and interacting medium [187]; furthermore, a rational design process must also consider the intricacy of reaction mechanisms to ensure appropriate reactivity and product selectivity, which includes sensitivity to temperature and pressure, to result in truly industrially relevant catalysts. The complexity of such catalytic systems quickly becomes intractable to fully explore with current experimental or computational efforts.

Historically, catalysts have been identified and their application optimised via empirical investigations, using previous success to guide future decision-making. Such 'top-down' experimentation has recently seen the integration of HTE into workflows, accelerating catalyst discovery through parallelisation of testing; in the more advanced cases, the HTE is coupled with data-driven analysis of reactivity/selectivity and automation to self-consistently optimise the efficacy of the catalytic system towards a target property, working within a defined parameter space [188]. The current HTE approaches do not typically include advanced *in situ* or *operando* characterisation, but these methods are increasingly available separately and benefit from similar emergent capabilities in automated data-driven analysis.

Alongside experiment, the advancement of computational capabilities allows the 'bottom-up' interrogation of elemental and structural knowledge from across the periodic table, presenting significant opportunities for accelerated data-driven discovery. Promising materials can be considered further using parameterised models to explore surface structures and composition as a function of operating conditions [187, 189], and reaction mechanisms derived using automated construction of chemical reaction networks, providing vast quantities of data relating to a reaction landscape [190] from which rates and product distributions are accessible via kinetic modelling. With the knowledge calculated within this sampling space, the efficacy of the catalytic system can be linked against key 'descriptors' of the catalyst and its operating conditions, providing powerful shortcuts when navigating across the reaction landscape to find better catalysts via e.g., active learning protocols. In the most state-of-the-art approaches, descriptors are derived as compound functions of both experimental and computational information, via multi-fidelity data models [39].

Current and future challenges

Data-driven models are dependent on large, accurate, and complete datasets, yet such experimental data remains challenging to locate, access, use, or reproduce. Historically, the reward structure of the research community has been towards positive results, which means that negative results are not shared [191]. Incomplete datasets lead to sampling bias and inaccuracy of data models; furthermore, hidden data can also present a challenge for reproducibility, whereby not all the experiment parameters are reported for future investigators. Data quantity and quality are also important aspects, yet most experimental investigations typical focus in a small chemical space, which lead to small datasets. Indeed, data completeness can again become challenging when only the 'best' catalysts are considered for higher-level characterisation methods, such as *in situ* electron microscopy; and simultaneously, data quality is compromised, as differing standards of analysis are introduced, and outcomes reported in contrasting formats [52]. The combination of identifiable data sources is also a current challenge, as the quality and quantity of information can vary in relation to synthetic methods, catalytic testing, and characterisation; and these data may be embedded in images, making collecting accurate data a challenge.

Similar challenges relating to data completeness and accuracy exist in the computational catalysis domain. Here, greater efforts have been made to creating standardised, complete, and publicly available datasets [192, 193], yet the realisations often remain limited to subsets of catalysts/reactants/products (e.g. oxygen evolution reaction electrocatalysts [193]) and a current challenge is to expand knowledge space. More pertinent is the need for accurate computational data that can be confidently correlated with experiment. Considering machine-learning forcefields (MLFFs), which are a notable success from the application of data-driven approaches in materials modelling, a current challenge is to build these approaches to reproduce *experiment*, and not just higher-level computational models. Further extension of the MLFFs should then include multiple compositional and environmental aspects of a fully operational heterogeneous catalyst; and for this more efficient modelling paradigms are needed to create bigger datasets. Future challenges then arise with the integration of computational and experimental datasets, whereby parameters and observables from each respective domain must be collated and compared on an equal footing to provide value to the researchers of the future.

Advances in science and technology to meet challenges

There are multiple technological advances identifiable to meet the challenges and fully achieve the potential of data-driven approaches. Within the laboratory, greater accessibility of automated high-throughput facilities, capable of synthesising, testing, and characterising catalysts, will be powerful in facilitating on-the-fly data-driven catalyst discovery, and must be coupled with public accessibility in centralised repositories to achieve larger, consistent, and more complete datasets. For modelling, improved software models are still needed to simulate a more accurate description of complete catalytic conditions, including the effects of temperature, pressure, and solvents, to provide accurate surrogate models of energy landscapes that can be explored rapidly, with automated discovery again presenting an opportunity. And at the interface of computation and experiment, greater integration of catalytic datasets to provide holistic coverage is necessary to account for deficits in knowledge from either the experimental or computational domains alone; indeed, one needs to harness the individual strengths of 'top-down' and 'bottom-up' perspectives to derive complementary data, rather than distinct.

These scientific and technological advances are coupled also with a need for greater discussion between members of the catalytic community, and advocacy of standardisation. Whilst the principles of FAIR data have developed strong roots in the computational modelling domain, the distribution or centralisation of experimental data remains limited, and focused on positive results. The value of *all* data should be championed, and the importance of metadata to aid users in understanding value and limitations of a given dataset; deposition of results in an accessible resource should be encouraged, especially for experiment, where uptake is more urgently needed. The communication between researchers should include experimental and computational communities, and span academia, industry, and third-party organisations, at all levels of scientific investigation, in order to deliver better understanding of data needs and standardisation of data-collection procedures. The work here is implicitly multidisciplinary, and so the interaction of chemists, materials scientists, physicists, computer scientists, data scientists and other domain experts should be encouraged to maximise the opportunity for multi-fidelity models that address shortcomings arising in individual research domains. Finally, there is the need to train and distribute knowledge among researchers of the value of their data; we should be educating in a cross-disciplinary manner about the importance of detailed digital data collection, in both experiment and computation. Such action will lead to engagement and investment towards necessary tools to accelerate the big-data driven discovery in heterogeneous catalysis; such software capabilities already exist, driven by the explosion in interest towards data-driven discovery, but the potential is yet to be realised.

Concluding remarks

The status for data-driven approaches in heterogeneous catalysis is promising, with strong application in computational fields and increasing demonstrations of potential in experimental laboratories. However, challenges remain with respect to ensuring the quality and completeness of individual datasets, as well as improving accessibility and standardisation. Opportunities have been highlighted that include increased automation within research environments, improved cross-discipline communication, and efforts among users to reach distribution standards that will benefit emergent as well as established researchers. Catalysis is an extremely challenging but valuable field, with impact on all of humanity. Adoption of the outlined approaches can facilitate the update of emergent data-driven methods for a transition to cleaner, more active heterogeneous catalysts that benefit the global population. There are many examples of good practice, but efforts are still needed if we are to maximise the potential value for all.

Acknowledgment

The authors are grateful for funding by the EPSRC Centre-to Centre Project (Grant Reference. EP/S030468/1). AJL acknowledges funding by the UKRI Future Leaders Fellowship program (MR/T018372/1). ZL acknowledges funding by the China Scholarship Council.

4.6. Synchrotron small-angle x-ray scattering (SAXS)—perspectives of ML

Peter Fratzl
Max Planck Institute of Colloids and Interfaces, Potsdam, Germany

Status

X-ray scattering and diffraction pertain to a major set of techniques to characterize the structure of materials at the nanoscale. SAXS, in particular, has been developed in the 1950s to resolve structures in the size range 1–100 nm [194]. Despite the development of electron microscopes some years later, it remained an important technique, mostly because x-rays are less strongly absorbed than electrons, which allows for in-operando experiments, studying the effect of physical stimuli, such as temperature, pH or humidity on material structure. A strong boost in the use of small-angle scattering came with the availability of synchrotron radiation that improved the time resolution of in-operando experiments, but also opened to possibility to transform SAXS into a multiscale imaging tool. In this approach, the general idea is that nanoscale information is extracted from analysing the scattering patterns, while mapping of the specimens provides the information at the microscale (see figure 23). The first attempts with SAXS-based imaging go back to the 1990s [195]. This evolved until the development of SAXS tomography which yields six-dimensional data: three dimensions in real space through scanning and rotating the specimen (typically with micrometre resolution), as well as three additional dimensions from the scattering patterns within each voxel (containing nanoscale information) [196, 197].

The enormous advance in the brilliance of x-ray beams, as well as in x-ray optics enables not only the collection of multidimensional SAXS-tomography data but also the measurement of massive numbers of specimens even within short times.

Current and future challenges

These advances upstream of the specimen in the experiment, however, lead to new challenges downstream of the specimen, linked to the treatment and the evaluation of massive amounts of data. A schematic of the workflow in a SAXS measurement is shown in figure 24. The traditional way of conducting such an experiment would be the path symbolized by (A) and (B) in this figure. (A) represents specimen preparation and the experiment planning and (B) the data collection. These data would then be brought back from the synchrotron experiment for treatment and analysis. However, with the increased speed of data collection, a general challenge in this approach resides in the fact that the experimentalist is essentially blind without some capabilities of data diagnostics. This requires elementary pre-analysis of the data to see whether a modification of the beamline setup could improve the experiment. Recognizing this, software packages involving fast data diagnostics were developed, an example being DPDAK, an open code software introduced at the BESSY and the DESY synchrotrons (in Berlin and Hamburg, respectively) [199].

With the amount of data collected in each beamtime session increasing continuously over the years, a number of additional challenges appear from the fact that manual data treatment becomes impossible. This applies to the cleaning of data (such as denoising, background subtraction, image reconstruction, normalization, etc) and even more to the data analysis, which in SAXS often involves data fitting. These steps are indicated by the arrows (D) and (E) in figure 24.

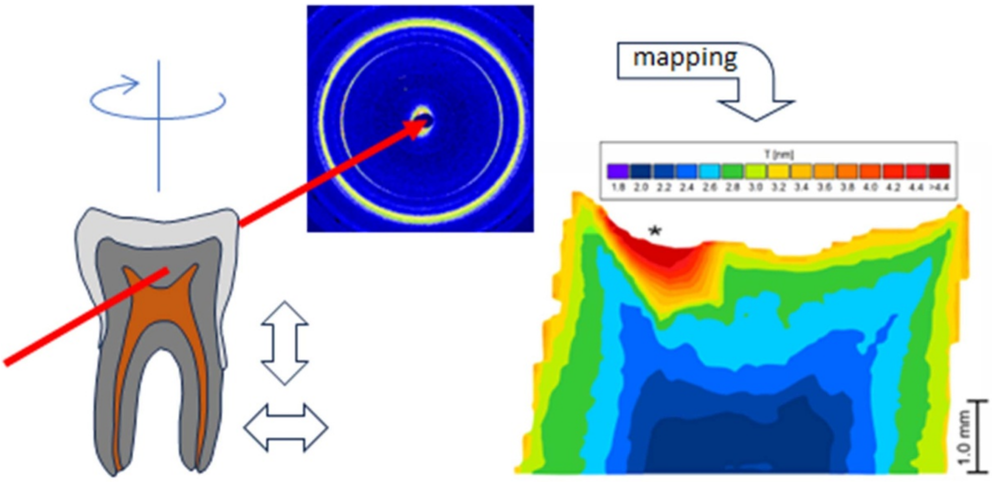


Figure 23. Principle of scanning-SAXS imaging. The specimen (for example a tooth section) is scanned across the x-ray beam with a diameter between tens of nanometres and several micrometres. Parameters extracted from the scattering patterns can then be mapped with a resolution corresponding to the x-ray beam diameter. In the figure, this is the thickness of mineral particles in dentin (the star indicates an area with a caries lesion). Reprinted from [198], Copyright (2010), with permission from Elsevier.

Advances in science and technology to meet challenges

Especially in SAXS tomography experiments, radiation damage should not be underestimated, since every specimen position will be hit several times by an intensive x-ray beam due to the required rotation of the specimen around multiple axes [197]. A typical strategy is then to reduce the irradiation time, which inevitably increases the noise in the data. To avoid problems with this noise in the 6D data reconstruction after the measurements, Zhou and coworkers propose a ML algorithm for the denoising of scattering data [200]. This approach facilitates step (D) in the diagram of figure 24.

The reconstruction of SAXS tomography data is equally challenging due to their high dimensionality. A possible traditional approach consists in calculating invariants of the SAXS data before reconstruction, which replaces the three-dimensional SAXS data by scalars that can be reconstructed much more efficiently [201]. SAXS invariants are useful, since they contain information about volume and surface of nano-size objects in the specimen [194] and allow, for example, the calculation of particle sizes in bone or dentin [195, 198, 201]. In the last few years, ML approaches are being developed for tomographic data reconstruction. Omori and coworkers review these developments for tomography using SAXS but also x-ray diffraction and other modalities [202]. While these advances relate to step (D) in figure 24, the review also addresses ML approaches for segmentation and analysis of the reconstructed data [202] (step (E) in figure 24).

Once data are reconstructed, every voxel in SAXS tomography data contains a scattering pattern to be analysed. This means a massive effort for data analysis (step (E) in figure 24) after reconstruction. Similar numbers of SAXS patterns need to be analysed in other situations, for example when material structures are studied as function of physical parameters (temperature, pressure, pH, humidity, etc) in multiple measurements. A recent review by Anker and coworkers addresses ML approaches to analyse a range of synchrotron-based experiment data,

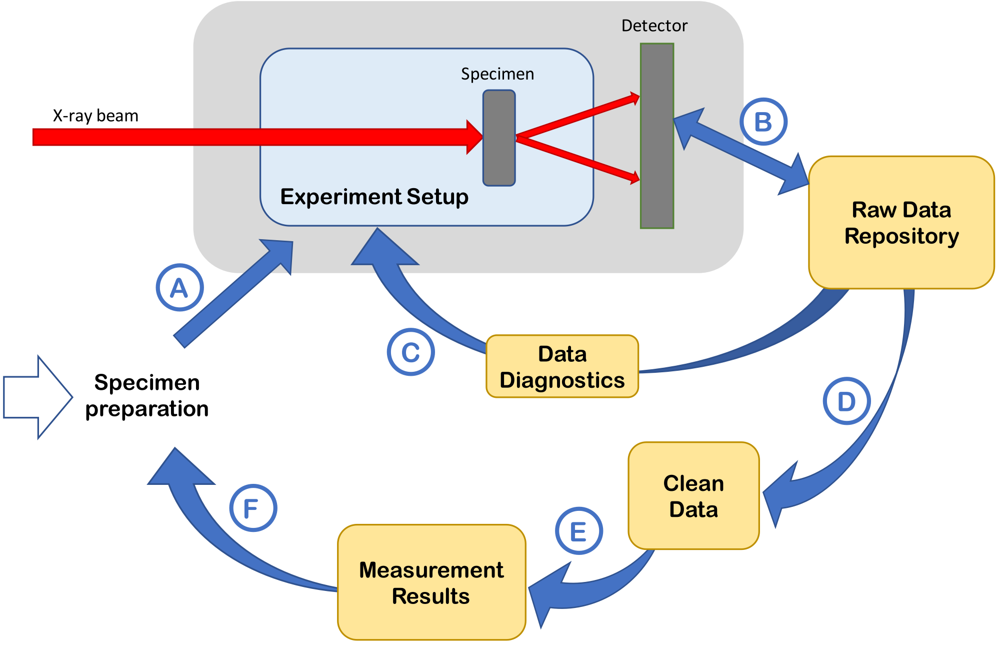


Figure 24. Schematic workflow of a small-angle x-ray scattering experiment. The traditional approach would be characterized by the arrows (A) (experiment planning to define the experiment setup) and (B) (data collection). With increasing data rates, several feedback loops involving machine learning are beginning to improve the quality and speed of the experiment: (C) is a readjustment of the experiment setup based on rapid data diagnostics, (D) is data reduction and denoising, (E) is data analysis and (F) automatic material synthesis based on the measurement results.

including SAXS but also powder diffraction, pair distribution function, inelastic neutron scattering and x-ray absorption spectroscopy data. While the traditional approach would be to fit a physical model to the data, supervised ML can be used to train a model for the prediction of structure based on data, but also to predict the scattering data based on a known structure and also to predict parameters based on some physical understanding of the system [203]. In another recent work [204], a ML-based analysis of SAXS data is proposed, which is based on Gaussian RFs that avoids the common model fitting of the data.

The approaches discussed until now are improving workflows in nearly all steps of SAXS experimentation (step (C)–(E) in figure 24). A last step (F) potentially closes the loop towards a fully automatized experimentation. This challenge is currently being taken up under the label of Autonomous Experimentation. Beaucage and Martin report on the development of an open liquid handling platform for autonomous formulation and x-ray scattering [205]. Yager and coworkers review this new paradigm and show how autonomous x-ray scattering can enhance efficiency and help discover new materials [206].

Concluding remarks

SAXS is an old method that is currently seeing an enormous increase in activity due to highly brilliant x-ray sources, more performant x-ray optics and—most recently—rapid progress in the treatment and the analysis of large amounts of data. As discussed above, several approaches have been developed addressing some of the steps in the workflow sketched in figure 23 through ML, but there are many more opportunities for applying such methods. Faster and, therefore, more effective tools for online data diagnostics based on ML during the experiment could bring a major improvement. Indeed, this has the potential to significantly reduce measurement times and radiation damage on sensitive specimens by allowing dynamic experiment planning. Moreover, better automatic tools for data cleaning, noise reduction, as well as to correct for background and instrumental resolution will be essential for high-throughput experiments or tomographic measurements. Finally, there are further needs for combining physical models with ML methods in data fitting, a development which may require a wide-spread effort in training relevant models for a variety of material classes. In conclusion, ML approaches have an important role to play in many areas of synchrotron x-ray scattering and the developments in this direction have only just begun.

Acknowledgment

The author thanks BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials-Science, for support and stimulating interdisciplinary interactions with members of the consortium.

Data availability statement

No new data were created or analysed in this study.

ORCID iDs

Peter Benner https://orcid.org/0000-0003-3362-4103
Tristan Bereau https://orcid.org/0000-0001-9945-1271
Volker Blum https://orcid.org/0000-0001-8660-7230
Mario Boley https://orcid.org/0000-0002-0704-4968
Christian Carbogno https://orcid.org/0000-0003-0635-8364
C Richard A Catlow https://orcid.org/0000-0002-1341-1541
Sebastian Eibl https://orcid.org/0000-0002-1069-2720
Lucas Foppa https://orcid.org/0000-0003-3002-062X
Christoph Freysoldt https://orcid.org/0000-0002-7896-3478
Baptiste Gault https://orcid.org/0000-0002-4934-0458
Pawan Goyal https://orcid.org/0000-0003-3072-7780
Lara Kabalan https://orcid.org/0000-0001-5715-4332
Petr Karpov https://orcid.org/0000-0003-1388-9841
Christoph T. Koch https://orcid.org/0000-0002-3984-1523
Sebastian Kokott https://orcid.org/0000-0003-1066-6909
Igor Kowalec https://orcid.org/0000-0002-9470-1275
Kurt Kremer https://orcid.org/0000-0003-1842-9369
Andreas Leitherer https://orcid.org/0000-0001-7747-4122
Yue Li https://orcid.org/0000-0003-3377-6676
Christian H Liebscher https://orcid.org/0000-0001-8620-4597
Andrew J Logsdail https://orcid.org/0000-0002-2277-415X

Felix Luong https://orcid.org/0000-0001-7821-295X
Andreas Marek https://orcid.org/0000-0001-5403-7528
Jaber R Mianroodi https://orcid.org/0000-0003-4778-3260
Jörg Neugebauer https://orcid.org/0000-0002-7903-2472
Zongrui Pei https://orcid.org/0000-0003-0748-4629
Thomas A R Purcell https://orcid.org/0000-0003-4564-7206
Dierk Raabe https://orcid.org/0000-0003-0194-6124
Markus Rampp https://orcid.org/0000-0001-8177-8698
Mariana Rossi https://orcid.org/0000-0002-3552-0677
Jan-Michael Rost https://orcid.org/0000-0002-8306-1743
Ulf Saalmann https://orcid.org/0000-0003-3208-8273
Marcel Schloz https://orcid.org/0000-0001-6295-1715
Annette Trunschke https://orcid.org/0000-0003-2869-0181
Ye Wei https://orcid.org/0000-0003-1965-2298
R Patrick Xian https://orcid.org/0000-0001-9895-6956
Matthias Scheffler https://orcid.org/0000-0002-1280-9873

References

[1] The concept of a fourth paradigm was probably first discussed by J Gray at a workshop on January 11, 2007 before he went missing at the Pacific on January 28, 2007 Hey T, Tansley S and Tolle K (eds) 2009 *The Fourth Paradigm, Data Intensive Discovery* (Microsoft Research)

[2] Slater J C 1937 Wave functions in a periodic potential *Phys. Rev.* **51** 846
Slater J C 1953 An augmented plane wave method for the periodic potential problem *Phys. Rev.* **92** 603
Slater J C 1965 *Quantum Theory of Molecules and Solids, Symmetry and Energy Bands in Crystals* vol 2 (McGraw-Hill)
Slater J C 1967 *Quantum Theory of Molecules and Solids, Insulators, Semiconductors and Metals* vol 3 (McGraw-Hill)
Slater J C and Johnson K H 1972 Self-consistent-field Xα cluster method for polyatomic molecules and solids *Phys. Rev.* B **5** 844

[3] Hohenberg P and Kohn W 1964 Inhomogeneous electron gas *Phys. Rev.* **136** B864

[4] Metropolis N, Rosenbluth A W, Rosenbluth M N and Teller E 1953 Equation of state calculations by fast computing machines *J. Chem. Phys.* **21** 1087

[5] Alder B J and Wainwright T E 1958 Molecular dynamics by electronic computers *Int. Symp. on Transport Processes in Statistical Mechanics* ed I Prigogine (Wiley) pp 97–131
Alder B J and Wainwright T E 1962 Phase transition in elastic disks *Phys. Rev.* **127** 359–361
Alder B J and Wainwright T E 1970 Decay of velocity autocorrelation function *Phys. Rev.* A **1** 18–2

[6] Rahman A 1964 Correlations in the motion of atoms in liquid argon *Phys. Rev.* **136** A405–11

[7] Agrawal A and Choudhary A 2016 Perspective: materials informatics and big data: realization of the “fourth paradigm” of science in materials science *APL Mater.* **4** 053208

[8] Draxl C and Scheffler M 2020 Big data-driven materials science and its FAIR data infrastructure *Handbook of Materials Modeling* ed W Andreoni and S Yip (Springer)

[9] Foppa L *et al* 2021 Materials genes of heterogeneous catalysis from clean experiments and artificial intelligence *MRS Bull.* **46** 1016–26

[10] Raabe D, Mianroodi J R and Neugebauer J 2023 Accelerating the design of compositionally complex materials via physics-informed artificial intelligence *Nat. Comput. Sci.* **3** 198–209

[11] Rao Z *et al* 2022 Machine learning–enabled high-entropy alloy discovery *Science* **378** 78–85

[12] Sutton C, Boley M, Ghiringhelli L M, Rupp M, Vreeken J and Scheffler M 2020 Identifying domains of applicability of machine learning models for materials science *Nat. Commun.* **11** 4428

[13] Scheffler M *et al* 2022 FAIR data enabling new horizons for materials research *Nature* **604** 635–42

[14] Schmidt J, Marques M R, Botti S and Marques M A 2019 Recent advances and applications of machine learning in solid-state materials science *npj Comput. Mater.* **5** 83

[15] Donoho D 2017 50 years of data science *J. Comput. Graph. Stat.* **26** 745–66

[16] Sutton C, Ghiringhelli L M, Yamamoto T, Lysogorskiy Y, Blumenthal L, Hammerschmidt T, Golebiowski J R, Liu X, Ziletti A and Scheffler M 2019 Crowd-sourcing materials-science challenges with the NOMAD 2018 Kaggle competition *npj Comput. Mater.* **5** 111

[17] Lookman T, Balachandran P V, Xue D and Yuan R 2019 Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design *npj Comput. Mater.* **5** 21

[18] Shahriari B, Swersky K, Wang Z, Adams R P and de Freitas N 2016 Taking the human out of the loop: a review of Bayesian optimization *Proc. IEEE* **104** 148–75

[19] Zhan D and Xing H 2020 Expected improvement for expensive optimization: a review *J. Glob. Optim.* **78** 507–44

[20] De Ath G, Everson R M, Rahat A A and Fieldsend J E 2021 Greed is good: exploration and exploitation trade-offs in Bayesian optimisation *ACM Trans. Evol. Learn. Optim.* **1** 1–27

[21] Biau G and Scornet E 2016 A random forest guided tour *Test* **25** 197–227

[22] Efron B 1979 Bootstrap methods: another look at the jackknife *Ann. Stat.* **7** 1–26

[23] Behler J 2021 Four generations of high-dimensional neural network potentials *Chem. Rev.* **121** 10037–72

[24] Deringer V L, Bartók A P, Bernstein N, Wilkins D M, Ceriotti M and Csányi G 2021 Gaussian process regression for materials and molecules *Chem. Rev.* **121** 10073–141

[25] de Pablo J J *et al* 2019 New frontiers for the materials genome initiative *npj Comput. Mater.* **5** 41

[26] Bartók A P, Payne M C, Kondor R and Csányi G 2010 Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons *Phys. Rev. Lett.* **104** 136403

[27] Musil F, Willatt M J, Langovoy M A and Ceriotti M 2019 Fast and accurate uncertainty estimation in chemical machine learning *J. Chem. Theory. Comput.* **15** 906–15

[28] Jeong W, Yoo D, Lee K, Jung J and Han S 2020 Efficient atomic-resolution uncertainty estimation for neural network potentials using a replica ensemble *J. Chem. Phys. Lett.* **11** 6090–6

[29] Hirschfeld L, Swanson K, Yang K, Barzilay R and Coley C 2020 Uncertainty quantification using neural networks for molecular property prediction *J. Chem. Inf. Modeling* **60** 3770–80

[30] Kahle L and Zipoli F 2022 Quality of uncertainty estimates from neural network potential ensembles *Phys. Rev.* E **105** 015311

[31] Tan A R, Urata S, Goldman S, Dietschreit J C and Gómez-Bombarelli R 2023 Single-model uncertainty quantification in neural network potentials does not consistently outperform model ensembles (arXiv:2305.01754)

[32] Scalia G, Grambow C A, Pernici B, Li Y and Green W H 2020 Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction *J. Chem. Inf. Modeling* **60** 2697–717

[33] Jinnouchi R, Karsai F and Kresse G 2019 On-the-fly machine learning force field generation: application to melting points *Phys. Rev.* B **100** 014105

[34] Palmer G, Du S, Politowicz A, Emory J P, Yang X, Gautam A, Gupta G, Li Z, Jacobs R and Morgan D 2022 Calibration after bootstrap for accurate uncertainty quantification in regression models *npj Comput. Mater.* **8** 115

[35] Raimbault N, Grisafi A, Ceriotti M and Rossi M 2019 Using Gaussian process regression to simulate the vibrational Raman spectra of molecular crystals *New J. Phys.* **21** 105001

[36] Wrobel S 1997 An algorithm for multi-relational discovery of subgroups *European Conf. on Principles of Data Mining and Knowledge Discovery* (*Trondheim, Norway*) pp 78–87

[37] Friedman J H and Fisher N I 1999 Bump hunting in high-dimensional data *Stat. Comput.* **9** 123–44

[38] Goldsmith B R, Boley M, Vreeken J, Scheffler M and Ghiringhelli L M 2017 Uncovering structure-property relationships of materials by subgroup discovery *New J. Phys.* **19** 013031

[39] Foppa L, Sutton C, Ghiringhelli L M, De S, Löser P, Schunk S A, Schäfer A and Scheffler M 2022 Learning design rules for selective oxidation catalysts from high-throughput experimentation and artificial intelligence *ACS Catal.* **12** 2223–32

[40] Foppa L and Ghiringhelli L M 2022 Identifying outstanding transition-metal-alloy heterogeneous catalysts for the oxygen reduction and evolution reactions via subgroup discovery *Top. Catal.* **65** 196–206

[41] Grosskreutz H, Rüping S and Wrobel S 2008 Tight optimistic estimates for fast subgroup discovery *Machine Learning and Knowledge Discovery in Databases* (*Antwerp, Belgium*) pp 440–56

[42] Nguyen H V and Vreeken J 2015 Non-parametric Jensen-Shannon divergence *Machine Learning and Knowledge Discovery in Databases: European Conf.* (*Porto, Portugal*) pp 173–89

[43] Mazheika A, Wang Y-G, Valero R, Viñes F, Illas F, Ghiringhelli L M, Levchenko S V and Scheffler M 2022 Artificial-intelligence-driven discovery of catalyst genes with application to CO_2 activation on semiconductor oxides *Nat. Commun.* **13** 419

[44] Lee S, Min S-J and Eigenmann R 2009 OpenMP to GPGPU *ACM Sigplan Notices* **44** 101–10

[45] The OpenACC application programming interface version 3.3 (available at: https://www.openacc.org/sites/default/files/inline-images/Specification/OpenACC-3.3-final.pdf) (Accessed November 2022)

[46] Maintz S and Wetzstein M 2018 Strategies to accelerate VASP with GPUs using OpenACC *Proc. Cray User Group* (available at: https://cug.org/proceedings/cug2018_proceedings/includes/files/pap153s2-file1.pdf)

[47] Edwards H C, Trott C R and Sunderland D 2014 Kokkos: enabling manycore performance portability through polymorphic memory access patterns *J. Parallel Distrib. Comput.* **74** 3202–16

[48] Beckingsale D, Burmark J, Hornung R, Jones H, Killian W, Kunen A J, Pearce O, Robinson P, Ryujin B S and Scogland T R 2019 RAJA: portable performance for large-scale scientific applications *2019 IEEE/ACM Int. Workshop on Performance, Portability and Productivity in HPC (P3HPC)* (U.S. Department of Energy Office of Scientific and Technical Information) (https://doi.org/10.1109/p3hpc49587.2019.00012)

[49] Peng J *et al* 2022 Human- and machine-centred designs of molecules and materials for sustainability and decarbonization *Nat. Rev. Mater.* **7** 991–1009

[50] Pilania G 2021 Machine learning in materials science: from explainable predictions to autonomous design *Comput. Mater. Sci.* **193** 13

[51] Wilkinson M D *et al* 2016 The FAIR guiding principles for scientific data management and stewardship *Sci. Data* **3** 160018

[52] Trunschke A *et al* 2020 Towards experimental handbooks in catalysis *Top. Catal.* **63** 1683–99

[53] Smith A, Bhat V, Ai Q and Risko C 2022 Challenges in information-mining the materials literature: a case study and perspective *Chem. Mater.* **34** 4821–7

[54] Marshall C P, Schumann J and Trunschke A 2023 Achieving digital catalysis: strategies for data acquisition, storage and use *Angew. Chem., Int. Ed.* **62** e202302971

[55] Foppa L *et al* 2023 Data-centric heterogeneous catalysis: identifying rules and materials genes of alkane selective oxidation? *J. Am. Chem. Soc.* **145** 3427–42

[56] Trunschke A 2022 Prospects and challenges for autonomous catalyst discovery viewed from an experimental perspective *Catal. Sci. Technol.* **12** 3650–69

[57] Blum V, Gehrke R, Hanke F, Havu P, Havu V, Ren X, Reuter K and Scheffler M 2009 Ab initio molecular simulations with numeric atom-centered orbitals *Comput. Phys. Commun.* **180** 2175–96

[58] Lu H *et al* 2023 Electronic impurity doping of a 2D hybrid lead iodide perovskite by Bi and Sn *PRX Energy* **2** 023010

[59] Ihrig A C, Wieferink J, Zhang I Y, Ropo M, Ren X, Rinke P, Scheffler M and Blum V 2015 Accurate localized resolution of identity approach for linear-scaling hybrid density functionals and for many-body perturbation theory *New J. Phys.* **17** 093020

[60] Levchenko S V, Ren X, Wieferink J, Johanni R, Rinke P, Blum V and Scheffler M 2015 Hybrid functionals for large periodic systems in an all-electron, numeric atom-centered basis framework *Comput. Phys. Commun.* **192** 60–69

[61] Knuth F, Carbogno C, Atalla V, Blum V and Scheffler M 2015 All-electron formalism for total energy strain derivatives and stress tensor components for numeric atom-centered orbitals *Comput. Phys. Commun.* **190** 33–50

[62] Huhn W P and Blum V 2017 One-hundred-three compound band-structure benchmark of post-self-consistent spin-orbit coupling treatments in density functional theory *Phys. Rev. Mater.* **1** 033803

[63] Marek A, Blum V, Johanni R, Havu V, Lang B, Auckenthaler T, Heinecke A, Bungartz H J and Lederer H 2014 The ELPA library: scalable parallel eigenvalue solutions for electronic structure theory and computational science *J. Phys.: Condens. Matter* **26** 213201

[64] Kůs P, Marek A, Köcher S S, Kowalski -H-H, Carbogno C, Scheurer C, Reuter K, Scheffler M and Lederer H 2019 Optimizations of the eigensolvers in the ELPA library *Parallel Comput.* **85** 167–77

[65] Yu V W-Z, Moussa J, Kus P, Marek A, Messmer P, Yoon M, Lederer H and Blum V 2021 GPU-acceleration of the ELPA2 distributed eigensolver for dense symmetric and Hermitian eigenproblems *Comput. Phys. Commun.* **262** 107808

[66] Draxl C and Scheffler M 2018 NOMAD: the FAIR concept for big data-driven materials science *MRS Bull.* **43** 676–82

[67] Scheidgen M *et al* 2023 NOMAD: a distributed web-based platform for managing materials science research data *J. Open Source Softw.* **8** 5388

[68] Ghiringhelli L M *et al* 2023 Shared metadata for data-centric materials science *Sci. Data* **10** 626

[69] Sbailò L, Fekete Á, Ghiringhelli L M and Scheffler M 2022 The NOMAD artificial-intelligence toolkit: turning materials-science data into knowledge and understanding *npj Comput. Mater.* **8** 250

[70] Ragan-Kelley B *et al* 2018 Binder 2.0-reproducible, interactive, sharable environments for science at scale *Proc. 17th Python in Science Conf.* ed F Akici, D Lippa, D Niederhut and M Pacer pp 113–20

[71] Google Research, Google Colaboratory 2018 (available at: https://colab.research.google.com/) (Accessed 19 December 2023)

[72] Janssen J, Surendralal S, Lysogorskiy Y, Todorova M, Hickel T, Drautz R and Neugebauer J 2019 pyiron: an integrated development environment for computational materials science *Comput. Mater. Sci.* **163** 24–36

[73] Yakutovich A V *et al* 2021 AiiDAlab–an ecosystem for developing, executing, and sharing scientific workflows *Comput. Mater. Sci.* **188** 110165

[74] Dunn A, Wang Q, Ganose A, Dopp D and Jain A 2020 Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm *npj Comput. Mater.* **6** 1–10

[75] Barker M *et al* 2022 Introducing the FAIR principles for research software *Sci. Data* **9** 622

[76] Alzubaidi L *et al* 2023 A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications *J. Big Data* **10** 46

[77] Giri S K, Saalmann U and Rost J M 2020 Purifying electron spectra from noisy pulses with machine learning using synthetic Hamilton matrices *Phys. Rev. Lett.* **124** 113201

[78] Giri S K, Alonso L, Saalmann U and Rost J M 2021 Perspectives for analyzing non-linear photo-ionization spectra with deep neural networks trained with synthetic Hamilton matrices *Farad. Discuss.* **228** 502

[79] Cheung H L, Uvdal P and Mirkhalaf M 2023 Augmentation of scarce data—a new approach for deep-learning modeling of composites (arXiv:2311.14557)

[80] Ghane E, Fagerström M and Mirkhalaf M 2023 Recurrent neural networks and transfer learning for elasto-plasticity in woven composites (arXiv:2311.13434v2)

[81] Giri S K, Saalmann U and Rost J M 2024 in preparation

[82] Selstø S 2022 Absorbers as detectors for unbound quantum systems *Phys. Rev.* A **106** 042213

[83] Leo J, Ge E and Li S 2023 Wasserstein distance in deep learning *SSRN Electron. J.* (https://doi.org/10.2139/ssrn.4368733)

[84] Cressie N and Wikle C K 2011 *Statistics for Spatio-Temporal Data* 1st edn (Wiley)

[85] Ostoja-Starzewski M 2007 *Microstructural Randomness and Scaling in Mechanics of Materials* (Chapman and Hall/CRC) (https://doi.org/10.1201/9781420010275)

[86] Saunders R N, Teferra K, Elwany A, Michopoulos J G and Lagoudas D 2023 Metal AM process-structure-property relational linkages using Gaussian process surrogates *Addit. Manuf.* **62** 103398

[87] Xian R P *et al* 2023 A machine learning route between band mapping and band structure *Nat. Comput. Sci.* **3** 101–14

[88] Kusne A G *et al* 2020 On-the-fly closed-loop materials discovery via Bayesian active learning *Nat. Commun.* **11** 5966

[89] Chen R T Q, Amos B and Nickel M 2020 Neural spatio-temporal point processes *Int. Conf. on Learning Representations* (available at: https://openreview.net/forum?id=XQQA6-So14)

[90] Smith J T H, De Mello S, Kautz J, Linderman S W and Byeon W 2023 Convolutional state space models for long-range spatiotemporal modeling (arXiv:2310.19694)

[91] Chang Z, Koulieris G A and Shum H P H 2023 On the design fundamentals of diffusion models: a survey (arXiv:2306.04542)

[92] Abolhasani M and Kumacheva E 2023 The rise of self-driving labs in chemical and materials sciences *Nat. Synth.* **2** 6

[93] Doi M 2015 *Soft Matter Physics* Reprinted with Correction (Oxford University Press)
[94] Menichetti R, Kanekal K H and Bereau T 2019 Drug–membrane permeability across chemical space *ACS Cent. Sci.* **5** 290–8
[95] Greco C, Melnyk A, Kremer K, Andrienko D and Daoulas K C 2019 Generic model for lamellar self-assembly in conjugated polymers: linking mesoscopic morphology and charge transport in P3HT *Macromolecules* **52** 968–81
[96] Potestio R, Peter C and Kremer K 2014 Computer simulations of soft matter: linking the scales *Entropy* **16** 4199–245
[97] Schmid F 2023 Understanding and modeling polymers: the challenge of multiple scales *ACS Polym. Au* **3** 28–58
[98] Jackson N E, Webb M A and De Pablo J J 2019 Recent advances in machine learning towards multiscale soft materials design *Curr. Opin. Chem. Eng.* **23** 106–14
[99] Ni B and Buehler M J 2024 MechAgents: large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge *Extreme Mech. Lett.* **67** 102131
[100] Suryanarayana C 1999 *Non-Equilibrium Processing of Materials* (Elsevier)
[101] Musil F, Grisafi A, Bartók A P, Ortner C, Csányi G and Ceriotti M 2021 Physics-inspired structural representations for molecules and materials *Chem. Rev.* **121** 9759–815
[102] Weinreich J, Lemm D, Von Rudorff G F and Von Lilienfeld O A 2022 Ab initio machine learning of phase space averages *J. Chem. Phys.* **157** 024303
[103] Mohr B, Van Der Mast D and Bereau T 2023 Condensed-phase molecular representation to link structure and thermodynamics in molecular dynamics *J. Chem. Theory. Comput.* **19** 4770–9
[104] Wang J, Olsson S, Wehmeyer C, Pérez A, Charron N E, de Fabritiis G, Noé F and Clementi C 2019 Machine learning of coarse-grained molecular dynamics force fields *ACS Cent. Sci.* **5** 755–67
[105] Durumeric A E P and Voth G A 2019 Adversarial-residual-coarse-graining: applying machine learning theory to systematic molecular coarse-graining *J. Chem. Phys.* **151** 124110
[106] Roters F *et al* 2019 DAMASK—the Düsseldorf Advanced Material Simulation Kit for modeling multi-physics crystal plasticity, thermal, and damage phenomena from the single crystal up to the component scale *Comput. Mater. Sci.* **158** 420–78
[107] Raabe D, Mianroodi J R and Neugebauer J 2023 Computational design of compositionally complex materials *Nat. Comput. Sci.* **3** 198–209
[108] Wu X 1991 Neural network-based material modeling *PhD Thesis* Department of Civil and Environmental Engineering at the University of Illinois Urbana–Champaign, Urbana, Illinois
[109] Yang Z, Yu C-H and Buehler M J 2021 Deep learning model to predict complex stress and strain fields in hierarchical composites *Sci. Adv.* **7** eabd7416
[110] Mianroodi J R, Siboni N H and Raabe D 2021 Teaching solid mechanics to artificial intelligence—a fast solver for heterogeneous materials *npj Comput. Mater.* **7** 99
[111] Khorrami M S, Mianroodi J R, Siboni N H, Goyal P, Svendsen B, Benner P and Raabe D 2023 An artificial neural network for surrogate modeling of stress fields in viscoplastic polycrystalline materials *npj Comput. Mater.* **9** 37
[112] Rashid M M, Pittie T, Chakraborty S and Krishnan N M A 2022 Learning the stress-strain fields in digital composites using Fourier neural operator *iScience* **25** 105452
[113] Ni B and Buehler M J 2024 MechAgents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge *Extreme Mech. Lett.* **67** 102131
[114] Roters F, Eisenlohr P, Hantcherli L, Tjahjanto D D, Bieler T R and Raabe D 2010 Overview of constitutive laws, kinematics, homogenization and multiscale methods in crystal plasticity finite-element modeling: theory, experiments, applications *Acta Mater.* **58** 1152–211
[115] Wang S, Wang H and Perdikaris P 2021 Learning the solution operator of parametric partial differential equations with physics-informed DeepONets *Sci. Adv.* **7** eabi8605
[116] Li Z, Zheng H, Kovachki N, Jin D, Chen H, Liu B, Azizzadenesheli K 2021 Physics-informed neural operator for learning partial differential equations (arXiv:2111.03794)
[117] Raabe D, Sander B, Friák M, Ma D and Neugebauer J 2007 Theory-guided bottom-up design of β-titanium alloys as biomaterials based on first principles calculations: theory and experiments *Acta Mater.* **55** 4475–87
[118] Sandlöbes S, Friák M, Korte-Kerzel S, Pei Z, Neugebauer J and Raabe D 2017 A rare-earth free magnesium alloy with improved intrinsic ductility *Sci. Rep.* **7** 1–8

[119] Goyal P, Duff I P and Benner P 2023 Guaranteed stable quadratic models and their applications in SINDy and operator inference (arXiv:2308.13819)

[120] Sasidhar K N, Siboni N H, Mianroodi J R, Rohwerder M, Neugebauer J and Raabe D 2022 Deep learning framework for uncovering compositional and environmental contributions to pitting resistance in passivating alloys *npj Mater. Degrad.* **6** 71

[121] Lusch B, Kutz J N and Brunton S L 2018 Deep learning for universal linear embeddings of nonlinear dynamics *Nat. Commun.* **9** 4950

[122] Wang N, C.Freysoldt C, Zhang S, Liebscher C H and Neugebauer J 2021 Segmentation of static and dynamic atomic-resolution microscopy data sets with unsupervised machine learning using local symmetry descriptors *Microsc. Microanal.* **27** 1454–64

[123] Saxena A, Polin N, Kusampudi N, Katnagallu S, Molina-Luna L, Gutfleisch O, Berkels B, Gault B, Neugebauer J and Freysoldt C 2023 A machine learning framework for quantifying chemical segregation and microstructural features in atom probe tomography data *Microsc. Microanal.* **29** 1658–70

[124] Kalinin S V, Dyck O, Jesse S and Ziatdinov M 2021 Exploring order parameters and dynamic processes in disordered systems via variational autoencoders *Sci. Adv.* **7** eabd5084

[125] eLabFTW—a free and open source electronic lab notebook (available at: https://www.elabftw.net/) (Accessed 24 October 2023)

[126] Jain A *et al* 2013 Commentary: the materials project: a materials genome approach to accelerating materials innovation *APL Mater.* **1** 011002-1–11

[127] APT-HDF5 file specification (available at: http://fieldemission.org/files/APT-HDF5-2020-10.pdf) (Accessed 24 October 2023)

[128] Electron microscopy datasets (available at: https://emdatasets.com/format/) (Accessed 24 October 2023)

[129] Hyperspy user guide io module (available at: https://hyperspy.org/hyperspy-doc/current/user_guide/io.html) (Accessed 24 October 2023)

[130] Ott S, Hebenstreit K, Liévin V, Hother C E, Moradi M, Mayrhauser M, Praas R, Winther O and Samwald M 2023 ThoughtSource: a central hub for large language model reasoning data *Sci. Data* **10** 1–12

[131] Jablonka K M *et al* 2023 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon *Digit. Discov.* 2 1233–50

[132] Park Y J, Kaplan D, Ren Z, Hsu C W, Li C, Xu H, Li S and Li J 2023 Can ChatGPT be used to generate scientific hypotheses? *J. Mater.* **10** 1–37

[133] Szymanski N J *et al* 2023 An autonomous laboratory for the accelerated synthesis of novel materials *Nature* **624** 86–91

[134] Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, Persson K A, Ceder G and Jain A 2019 Unsupervised word embeddings capture latent knowledge from materials science literature *Nature* **571** 95–98

[135] Zheng Z, Zhang O, Borgs C, Chayes J T and Yaghi O M 2023 ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis *J. Am. Chem. Soc.* **145** 18048–62

[136] Kim E, Huang K, Saunders A, McCallum A, Ceder G and Olivetti E 2017 Materials synthesis insights from scientific literature via text extraction and machine learning *Chem. Mater.* **29** 9436–44

[137] Gupta T, Zaki M, Krishnan N M A and Mausam M 2022 MatSciBERT: a materials domain language model for text mining and information extraction *npj Comput. Mater.* **8** 1–11

[138] Pei Z, Yin J, Liaw P K and Raabe D 2023 Toward the design of ultrahigh-entropy alloys via mining six million texts *Nat. Commun.* **14** 54

[139] Krenn M and Zeilinger A 2020 Predicting research trends with semantic and neural networks with an application in quantum physics *Proc. Natl Acad. Sci. USA* **117** 1910–6

[140] An Y *et al* 2022 Exploring pre-trained language models to build knowledge graph for metal-organic frameworks (MOFs) *2022 IEEE Int. Conf. on Big Data (Big Data)* (*Osaka, Japan*) pp 3651–8

[141] Devi M A, Prakash C P S, Chinnannavar R P, Joshi V P, Palada R S and Dixit R 2020 An informatic approach to predict the mechanical properties of aluminum alloys using machine learning techniques *2020 Int. Conf. on Smart Electronics and Communication (ICOSEC)* (*Trichy, India*) pp 536–41

[142] Zhao X, Greenberg J, An Y and Hu X T 2021 Fine-tuning BERT model for materials named entity recognition *2021 IEEE Int. Conf. on Big Data (Big Data)* (*Orlando, FL, USA*) pp 3717–20

[143] Sasidhar K N, Siboni N H, Mianroodi J R, Rohwerder M, Neugebauer J and Raabe D 2023 Enhancing corrosion-resistant alloy design through natural language processing and deep learning *Sci. Adv.* **9** 7992

[144] Yin J, Bose A, Cong G, Lyngaas I and Anthony Q 2024 Comparative study of large language model architectures on frontier (arXiv:2402.00691)

[145] Kirklin S, Meredig B and Wolverton C 2013 High-throughput computational screening of new Li-ion battery anode materials *Adv. Energy Mater.* **3** 252–62

[146] Rodríguez-Martínez X, Pascual-San-José E and Campoy-Quiles M 2021 Accelerating organic solar cell material's discovery: high-throughput screening and big data *Energy Environ. Sci.* **14** 3301–22

[147] Bajorath J 2002 Integration of virtual and high-throughput screening *Nat. Rev. Drug Discov.* **1** 882–94

[148] Merchant A, Batzner S, Schoenholz S S, Aykol M, Cheon G and Cubuk E D 2023 Scaling deep learning for materials discovery *Nature* **624** 80–85

[149] Andersen C W *et al* 2021 OPTIMADE, an API for exchanging materials data *Sci. Data* **8** 217

[150] Pyzer-Knapp E O, Suh C, Gómez-Bombarelli R, Aguilera-Iparraguirre J and Aspuru-Guzik A 2015 What is high-throughput virtual screening? A perspective from organic materials discovery *Annu. Rev. Mater. Res.* **45** 195–216

[151] Settles B 2011 From theories to queries: active learning in practice *Active Learning and Experimental Design Workshop in Conjunction with AISTATS 2010* vol 16 (Proceedings of Machine Learning Research) p 1

[152] Li K, Persaud D, Choudhary K, DeCost B, Greenwood M and Hattrick-Simpers J 2023 Exploiting redundancy in large materials datasets for efficient machine learning with less data *Nat. Commun.* **14** 7283

[153] Zhang H, Chen W W, Rondinelli J M and Chen W 2023 ET-AL: entropy-targeted active learning for bias mitigation in materials data *Appl. Phys. Rev.* **10** 021403

[154] Todorović M, Gutmann M U, Corander J and Rinke P 2019 Bayesian inference of atomistic structure in functional materials *npj Comput. Mater.* **5** 35

[155] Curtarolo S *et al* 2012 AFLOW: an automatic framework for high-throughput materials discovery *Comput. Mater. Sci.* **58** 218–26

[156] Mathew K *et al* 2017 Atomate: a high-level interface to generate, execute, and analyze computational materials science workflows *Comput. Mater. Sci.* **139** 140–52

[157] Pizzi G, Cepellotti A, Sabatini R, Marzari N and Kozinsky B 2016 AiiDA: automated interactive infrastructure and database for computational science *Comput. Mater. Sci.* **111** 218–30

[158] Foumani Z Z, Shishehbor M, Yousefpour A and Bostanabad R 2023 Multi-fidelity cost-aware Bayesian optimization *Comput. Methods Appl. Mech. Eng.* **407** 115937

[159] Purcell T A R, Scheffler M, Ghiringhelli L M and Carbogno C 2023 Accelerating materials-space exploration for thermal insulators by mapping materials properties via artificial intelligence *npj Comput. Mater.* **9** 112

[160] Kalinin S V *et al* 2022 Machine learning in scanning transmission electron microscopy *Nat. Rev. Methods Primers* **2** 11

[161] Spurgeon S R *et al* 2021 Towards data-driven next-generation transmission electron microscopy *Nat. Mater.* **20** 274–9

[162] Jesse S, Chi M, Belianinov A, Beekman C, Kalinin S V, Borisevich A Y and Lupini A R 2016 Big data analytics for scanning transmission electron microscopy ptychography *Sci. Rep.* **6** 1–8

[163] Cautaerts N, Crout P, Ånes H W, Prestat E, Jeong J, Dehm G and Liebscher C H 2022 Free, flexible and fast: orientation mapping using the multi-core and GPU-accelerated template matching capabilities in the Python-based open source 4D-STEM analysis toolbox Pyxem *Ultramicroscopy* **237** 113517

[164] Leitherer A, Yeo B C, Liebscher C H and Ghiringhelli L M 2023 Automatic identification of crystal structures and interfaces via artificial-intelligence-based electron microscopy *npj Comput. Mater.* **9** 179

[165] Yin W *et al* 2020 A petascale automated imaging pipeline for mapping neuronal circuits with high-throughput transmission electron microscopy *Nat. Commun.* **11** 1–12

[166] Mukherjee D *et al* 2022 A roadmap for edge computing enabled automated multidimensional transmission electron microscopy *Micros. Today* **30** 10–19

[167] Treder K P, Huang C, Kim J S and Kirkland A I 2022 Applications of deep learning in electron microscopy *Microscopy* **71** i100–15

[168] Madsen J, Liu P, Kling J, Wagner J B, Hansen T W, Winther O and Schiøtz J 2018 A deep learning approach to identify local structures in atomic-resolution transmission electron microscopy images *Adv. Theory Simul.* **1** 1800037

[169] Botifoll M, Pinto-Huguet I and Arbiol J 2022 Machine learning in electron microscopy for advanced nanocharacterization: current developments, available tools and future outlook *Nanoscale Horiz.* **7** 1427–77

[170] Friedrich T, Yu C-P, Verbeeck J and Van Aert S 2023 Phase object reconstruction for 4D-STEM using deep learning *Microsc. Microanal.* **29** 395–407

[171] Wang F, Eljarrat A, Müller J, Henninen T R, Erni R and Koch C T 2020 Multi-resolution convolutional neural networks for inverse problems *Sci. Rep.* **10** 5730

[172] Ziatdinov M, Dyck O, Maksov A, Li X, Sang X, Xiao K, Unocic R R, Vasudevan R, Jesse S and Kalinin S V 2017 Deep learning of atomically resolved scanning transmission electron microscopy images: chemical identification and tracking local transformations *ACS Nano* **11** 12742–52

[173] Munshi J, Rakowski A, Savitzky B H, Zeltmann S E, Ciston J, Henderson M, Cholia S, Minor A M, Chan M K Y and Ophus C 2022 Disentangling multiple scattering with deep learning: application to strain mapping from electron diffraction patterns *npj Comput. Mater.* **8** 254

[174] Bertoni G, Rotunno E, Marsmans D, Tiemeijer P, Tavabi A H, Dunin-Borkowski R E and Grillo V 2023 Near-real-time diagnosis of electron optical phase aberrations in scanning transmission electron microscopy using an artificial neural network *Ultramicroscopy* **245** 113663

[175] Schloz M, Müller J, Pekin T C, Van den Broek W, Madsen J, Susi T and Koch C T 2023 Deep reinforcement learning for data-driven adaptive scanning in ptychography *Sci. Rep.* **13** 8732

[176] Gladyshev A, Schloz M, Pekin T C and Koch C T 2022 Comparison of compression methods for ptychographic reconstructions through decomposition of the diffraction patterns in orthonormal bases *Microsc. Microanal.* **28** 394–7

[177] Gault B, Chiaramonti A, Cojocaru-Mirédin O, Stender P, Dubosq R, Freysoldt C, Makineni S K, Li T, Moody M and Cairney J M 2021 Atom probe tomography *Nat. Rev. Method Primers* **1** 51

[178] Marquis E A and Hyde J M 2010 Applications of atom-probe tomography to the characterisation of solute behaviours *Mater. Sci. Eng.* R **69** 37–62

[179] Haley D, London A J and Moody M P 2020 Processing APT spectral backgrounds for improved quantification *Microsc. Microanal.* **26** 964–77

[180] Meier M S, Bagot P A J, Moody M P and Haley D 2023 Large-scale atom probe tomography data mining: methods and application to inform hydrogen behavior *Microsc. Microanal.* **29** 879–89

[181] Li Y, Zhou X, Colnaghi T, Wei Y, Marek A, Li H, Bauer S, Rampp M and Stephenson L T 2021 Convolutional neural network-assisted recognition of nanoscale L1_2 ordered structures in face-centred cubic alloys *npj Comput. Mater.* **7** 1–9

[182] Wei Y *et al* 2021 Machine-learning-enhanced time-of-flight mass spectrometry analysis *Patterns* **2** 100192

[183] Li Y *et al* 2023 Quantitative three-dimensional imaging of chemical short-range order via machine learning enhanced atom probe tomography *Nat. Commun.* **14** 7410

[184] Kühbach M, Bajaj P, Zhao H, Çelik M H M H, Jägle E A and Gault B 2021 On strong-scaling and open-source tools for analyzing atom probe tomography data *npj Comput. Mater.* **7** 1–10

[185] Humphreys J, Lan R and Tao S 2020 Development and recent progress on ammonia synthesis catalysts for Haber–Bosch process *Adv. Energy Sustain. Res.* **2** 2000043

[186] Foster S L, Bakovic S I P, Duda R D, Maheshwari S, Milton R D, Minteer S D, Janik M J, Renner J N and Greenlee L F 2018 Catalysts for nitrogen reduction to ammonia *Nat. Catal.* **1** 490–500

[187] Li H, Jiao Y, Davey K and Qiao S 2023 Data-driven machine learning for understanding surface structures of heterogeneous catalysts *Angew. Chem.* **135** e202216383

[188] Burger B *et al* 2020 A mobile robotic chemist *Nature* **583** 237–41

[189] Mou T, Pillai H S, Wang S, Wan M, Han X, Schweitzer N M, Che F and Xin H 2023 Bridging the complexity gap in computational heterogeneous catalysis with machine learning *Nat. Catal.* **6** 122–36

[190] Margraf J T, Jung H-W, Scheurer C and Reuter K 2023 Exploring catalytic reaction networks with machine learning *Nat. Catal.* **6** 112–21

[191] Taniike T and Takahashi K 2023 The value of negative results in data-driven catalysis research *Nat. Catal.* **6** 108–11

[192] Chanussot L *et al* 2021 Open catalyst 2020 (OC20) dataset and community challenges *ACS Catal.* **11** 6059–72

[193] Tran R *et al* 2023 The open catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts *ACS Catal.* **13** 3066–84

[194] Guinier A and Fournet G 1955 *Small-Angle Scattering of X-Rays* (Wiley)

[195] Fratzl P, Jakob H F, Rinnerthaler S, Roschger P and Klaushofer K 1997 Position-resolved small-angle x-ray scattering of complex biological materials *J. Appl. Crystallogr.* **30** 765–9

[196] Liebi M, Georgiadis M, Menzel A, Schneider P, Kohlbrecher J, Bunk O and Guizar-Sicairos M 2015 Nanostructure surveys of macroscopic specimens by small-angle scattering tensor tomography *Nature* **527** 349

[197] Schaff F, Bech M, Zaslansky P, Jud C, Liebi M, Guizar-Sicairos M and Pfeiffer F 2015 Six-dimensional real and reciprocal space small-angle x-ray scattering tomography *Nature* **527** 353–8

[198] Märten A, Fratzl P, Paris O and Zaslansky P 2010 On the mineral in collagen of human crown dentine *Biomaterials* **31** 5479–90

[199] Benecke G *et al* 2014 A customizable software for fast reduction and analysis of large x-ray scattering data sets: applications of the new DPDAK package to small-angle x-ray scattering and grazing-incidence small-angle x-ray scattering *J. Appl. Crystallogr.* **47** 1797–803

[200] Zhou Z *et al* 2023 A machine learning model for textured x-ray scattering and diffraction image denoising *npj Comput. Mater.* **9** 58

[201] De Falco P *et al* 2021 Tomographic x-ray scattering based on invariant reconstruction: analysis of the 3D nanostructure of bovine bone *J. Appl. Crystallogr.* **54** 486–97

[202] Omori N E, Bobitan A D, Vamvakeros A, Beale A M and Jacques S D M 2023 Recent developments in x-ray diffraction/scattering computed tomography for materials science *Phil. Trans. R. Soc.* A **381** 20220350

[203] Anker A S, Butler K T, Selvan R and Jensen K M O 2023 Machine learning for analysis of experimental scattering and spectroscopy data in materials chemistry *Chem. Sci.* **14** 14003–19

[204] Röding M, Tomaszewski P, Yu S, Borg M and Rönnols J 2022 Machine learning-accelerated small-angle x-ray scattering analysis of disordered two- and three-phase materials *Front. Mater.* **9** 956839

[205] Beaucage P A and Martin T B 2023 The autonomous formulation laboratory: an open liquid handling platform for formulation discovery using x-ray and neutron scattering *Chem. Mater.* **35** 846–52

[206] Yager K G, Majewski P W, Noack M M and Fukuto M 2023 Autonomous x-ray scattering *Nanotechnology* **34** 322001

[207] Ouyang R, Curtarolo S, Ahmetcik E, Scheffler M and Ghiringhelli L M 2018 SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates *Phys. Rev. Mater.* **2** 083802

[208] Purcell T A R, Scheffler M, Carbogno C and Ghiringhelli L M 2022 SISSO++: a C++ implementation of the sure-independence screening and sparsifying operator approach *J. Open Source Softw.* **7** 3960

[209] Cairney J M, Rajan K, Haley D, Gault B, Bagot P A J, Choi P-P, Felfer P J, Ringer S P, Marceau R K W and Moody M P 2015 Mining information from atom probe data *Ultramicroscopy* **159** 324–37