ADVANCED FUNCTIONAL MATERIALS
www.afm-journal.de

Functional Material Systems Enabled by Automated Data Extraction and Machine Learning

Payam Kalhor, Nicole Jung, Stefan Bräse, Christof Wöll, Manuel Tsotsalas, and Pascal Friederich**

The development of new functional materials is crucial for addressing global challenges such as clean energy or the discovery of new drugs and antibiotics. Functional material systems are typically composed of functional molecular building blocks, organized across multiple length scales in a hierarchical order. The large design space allows for precise tuning of properties to specific applications, but also makes it time-consuming and expensive to screen for optimal structures using traditional experimental methods. Machine learning (ML) models can potentially revolutionize the field of materials science by predicting chemical syntheses and materials properties with high accuracy. However, ML models require data to be trained and validated. Methods to automatically extract data from scientific literature make it possible to build large and diverse datasets for ML models. In this article, opportunities and challenges of data extraction and machine learning methods are discussed to accelerate the discovery of high-performing functional material systems, while ensuring that the predicted materials are stable, synthesizable, scalable, and sustainable. The potential impact of large language models (LLMs) on the data extraction process are discussed. Additionally, the importance of research data management tools is discussed to overcome the intrinsic limitations of data extraction approaches.

1. Introduction

A current challenge for research on functional material systems is the need to simultaneously consider multiple aspects from different disciplines to achieve optimal design.[1–3] This includes the components of materials, the structure of materials across different length scales, and every aspect of the final device and its operation conditions.[4–7] Additionally, environmental impact, circularity, and sustainability become increasingly important. All these individual aspects represent objectives for the design of functional material systems. To navigate this multidimensional design space with multiple objectives, researchers need to work across different disciplines in joint projects, considering expertise and research from these different disciplines.[8–10] To support and enable research in the area of functional material systems, automated data extraction from literature, using natural language processing, combined with ML can be used to operate on large amounts of data representing community knowledge to complement the researchers' own knowledge and experimental results.[11–13] Thus, a collaborative and interdisciplinary approach, coupled with the use of automated data extraction and machine learning, is necessary to enable the development of functional material systems that optimally meet multiple objectives. After identifying the optimal design of functional material systems, the synthesis of such complex hierarchically organized materials represents an

P. Kalhor, P. Friederich
Institute of Nanotechnology
Karlsruhe Institute of Technology
Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
E-mail: pascal.friederich@kit.edu

P. Kalhor, P. Friederich
Institute of Theoretical Informatics
Karlsruhe Institute of Technology
Am Fasanengarten 5, 76131 Karlsruhe, Germany

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adfm.202302630

DOI: 10.1002/adfm.202302630

N. Jung, S. Bräse
Institute of Organic Chemistry
Karlsruhe Institute of Technology
Fritz-Haber-Weg 6, 76131 Karlsruhe, Germany

N. Jung, S. Bräse
Institute of Biological and Chemical Systems - Functional Molecular Systems
Karlsruhe Institute of Technology
Hermann-von-Helmholtz-Platz 1, 76344Eggenstein-Leopoldshafen, Germany

C. Wöll, M. Tsotsalas
Institute of Functional Interfaces
Karlsruhe Institute of Technology
Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
E-mail: manuel.tsotsalas@kit.edu

M. Tsotsalas
Institute for Organic Chemistry
Karlsruhe Institute of Technology
Fritz-Haber-Weg 6, 76131 Karlsruhe, Germany

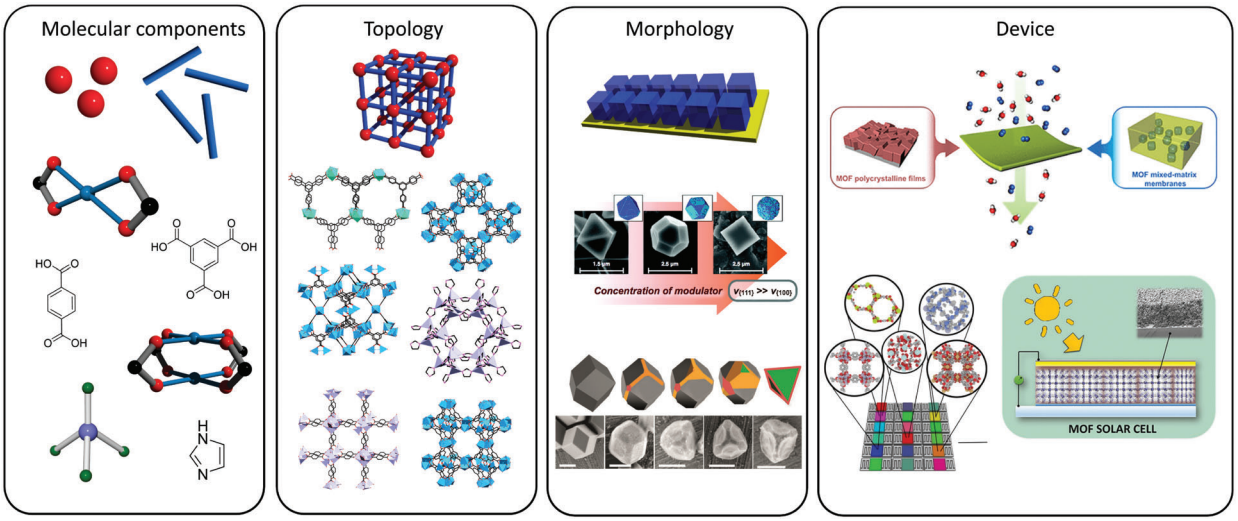

Figure 1. Hierarchical structure of functional material systems based on metal–organic frameworks (MOFs).Reproduced with permission. [4] Copyright 2022, Wiley VCH; Reproduced with permission.[39] Copyright 2020, American Chemical Society; Reproduced with permission. [40] Copyright 2021, Wiley VCH; Reproduced with permission. [41] Copyright 2018, Elsevier; Reproduced with permission. [42] Copyright 2015, Wiley VCH; Reproduced with permission. [43] Copyright 2011, American Chemical Society.

additional challenge. The synthesis of functional material systems can be subdivided into the synthesis of molecular components and the assembly of these components with specific composition and morphology in the nano- or micrometer-length scale. In the next step, the materials are processed, for example, into thin films, membranes, or certain reactor designs, in order to implement and "fit" the materials to the final device. All these steps need tailored synthesis and processing conditions to ensure their performance. Next to the design, also the synthesis, characterization, and processing of functional material systems can be supported and enabled by data extraction and ML of the material science literature and databases.[14–16] The synthesis, characterization, processing, and application of functional material systems produce large amounts of hierarchical and interdependent data. Making this data machine-readable and ready for ML and combining it with data extracted from scientific literature represent a particular challenge.[17] The use of tailored research data infrastructure is highly recommended, especially when working in large interdisciplinary consortia. Thus, the development of such research data management tools represents an essential task for the scientific community. Such tools should combine aspects from data collection, over data processing, to storing and publishing, ideally the raw and processed research data along with metadata.[18–20] In this perspective, we will briefly outline the design, synthesis, and characterization of functional material systems using metal–organic frameworks (MOFs) as example materials. Following this outline, we will highlight selected publications on enabling functional materials systems using a combination of automated data extraction and ML. We will discuss the accomplishments, prospects, challenges, and limitations of this approach. In the end, we will conclude with a discussion on research data management tools and unifying material science ontology.[21] The combination of research data management, data extraction from scientific literature, and ML are essential to fully explore the potential of functional material systems in addressing urgent social, economic, and environmental challenges.

2. Data Extraction for Functional Material Systems

2.1. Functional Material Systems

Functional material systems are typically composed of functional molecular building blocks, organized across multiple length scales in a hierarchical order (illustrated in **Figure 1**). MOFs emerged as a particularly powerful class of functional material systems.[22–25] Their modular synthesis enables the incorporation of diverse functionalities and tuning of their structures for desired applications.[26] The chemical design space of new MOFs is virtually unlimited, due to the numerous possibilities of combining metal nodes and organic linkers. Currently, about 100 000 MOFs have been synthesized and over 500 000 predicted.[27,28] However, the wide design space also makes it impossible to screen for optimal structures via brute force trial and error or traditional high-throughput experimental screening approaches.[29] Multiple techniques were developed for the synthesis of MOFs to control their structure across multiple length scales.[30] This starts from the synthesis of the organic linkers and precursors of metal nodes, all the way to crystal synthesis and further processing in the desired shape and formulation. The synthesis of MOFs started with solvothermal synthesis via multiple heating methods. Over time, new techniques were added, such as mechanochemical, vapor phase synthesis, and sacrificial or epitaxial growth.[31] The choice of synthesis conditions and the synthesis method dictates the final MOF crystal quality, defect density, crystal size, and morphology and enables interfacial growth.[32,33] The MOF materials can afterward be processed, for example, as thin films or freestanding membranes, or formulated, for example, by mixing with polymers, palleted, and processed to the required shapes for the final device.[34,35] The enormous amount of research related to functional material systems based on MOFs, starting from the synthesis of the molecular components, their assembly into MOF crystals with different topologies and morphologies, and their integration

and testing in the final device represents a hidden treasure.[36] Exposure of this treasure of data and making it ready for ML applications could lead to the development of tools that guide researchers and accelerate their efforts in the preparation of MOF-based devices that can address global challenges.[37,38] To fully exploit this treasure of data, a combination of tailored research data management tools, efficient data extraction from scientific literature, and ML are essential.

2.2. Data Extraction

One of the main challenges in applying ML to problems of high scientific relevance is the lack of openly accessible, structured, and machine-readable data. Existing databases, typically maintained and extended by particular scientific communities (e.g., protein structure database, certain MOF databases, crystal structure databases, etc.), can be used to train ML models for particular tasks, for example, the prediction of materials properties. However, the majority of potentially relevant data generated in scientific labs is not published at all, and from the fraction that is published, the majority is published in the form of graphs, tables, and non-structured text. Therefore, the extraction of data from scientific literature opens a vast amount of yet untapped possibilities to train ML models and use them to predict materials properties, extract and learn relevant relationships in the data, and eventually discover or design new materials. In the following, we will describe approaches to extract structured data from publications, focusing on text extraction but also discussing the extraction of information from tables, graphs, and images. Data extraction in other scientific domains, for example, biology dates back more than 20 years,[39] with seminal work in the late 90s, for example, Andrade et al.[40] One of the earliest attempts to automatically extract information from chemistry literature was OSCAR[41] and based on that the ChemicalTagger method in 2011.[42] ChemicalTagger is a rule-based multistep method based on tokenization (preprocessing of raw text), tagging (using OSCAR and regular expressions), phrase parsing (assignment of syntactical structure to text), and finally action phrase identification (extraction of chemical information) based on parse trees. The ChemDataExtractor Toolkit developed by Cole and coworkers starting in 2016[43,44] extends the rule-based natural language processing approach further, among others with ML methods, and adds functionality for table extraction[44] During the last years, ML approaches started to play an increasingly important role in literature data extraction, where, for example, article section relevance scores[45] and learned word embeddings[12,45] were used to enhance existing information extraction methods, or conditional random field models were used. With increasing capabilities of language models such as BERT[46] and GPT,[47] new possibilities for extracting information from literature are generated. Seminal examples of literature extraction methods based on LLMs include MatSciBERT by Gupta et al.;[48] a fine-tuned BERT model for materials science by Huang et al.,[49] BatteryBert, which among others use question-answering algorithms to translate text to structured information; and a GPT-3 based model by Dunn et al.[50] which uses fine tuning to directly translate scientific text to structured tabular data in JSON format. Also, semi-manual and crowd-sourcing-based approaches to extract information from chemistry and materials science literature were reported,[51–53] also extracting information from sources other than scientific literature, for example, lab notebooks to retrieve data about failed experiments which are usually not reported in scientific articles.[54] The automated extraction of data from tables, graphs, and images in many cases poses even larger challenges than the extraction of data from text. However, a detailed discussion of methods to extract data from tables,[44] graphs,[49] and images, in particular optical chemical structure recognition (OCSR), that is, the extraction of chemical structures from images[55–58] is beyond the scope of this article. Using various ways of literature data extraction, a large number of databases was generated and published, spanning from synthesis conditions[53,59–63] over materials stability[64] to materials properties, for example, for magnetic and superconducting properties,[65–67] semi-conductors,[68] battery materials,[49] thermoelectric materials,[69] glasses[70] and more general knowledge graphs.[12,71] In most cases, the databases are only a means to an end, that is, to provide sufficient training data for ML models for the prediction of synthesis routes and conditions as well as materials properties of a wider range of materials.

2.2.1. Technical Challenges and Intrinsic Limitations

Despite fast progress and promising new avenues related to the increasing use of ML and in particular LLMs in literature data extraction, there are still a range of important limitations and challenges. These can be grouped in technical challenges, which can in principle be solved by improving the data extraction methods, and intrinsic challenges, which concern inherent problems of unstructured literature as well as the quality and reliability of data that can be extracted from that. Technical challenges include current limitations of LLMs such as GPT-3 and similar models, which are either only obtainable via OpenAI's commercial APIs, or require state-of-the-art GPUs with large amounts of memory for prediction and retraining, both of which are only affordable for a small group of researchers worldwide. Another limitation is the availability and free accessibility of research papers, which makes automated access difficult and again excludes a large number of researchers who do not have access to all journals and publishers. Furthermore, if access is limited to, for example, abstracts, the amount of information that can be extracted is rather limited.[71] Furthermore, the use of LLMs (compared to algorithmic, rule-based models) comes at the cost of potentially higher processing times due to the size and computational cost of the models (even after retraining),[49] as well as a non-negligible amount of uncertainty regarding the question whether the output of LLMs is fully trustworthy, or if they can potentially output wrong information and give wrong answers or generate data, which is not contained in the input text.[50] At the same time, LLMs might potentially help to analyze complex texts and sentence structures, which are not extractable using conventional approaches.[72] Beyond that, one of the main challenges in literature data extraction currently is related to the fact that large amounts of data, for example, synthesis protocols, are not tabular data but can only be represented in more complex data structures. Examples of that are flexible, potentially multistep processes with dynamic data types and complex relations,[59,73] which not only

requires the development of more sophisticated extraction methods but furthermore need flexible data blueprints for complex scientific data. One development in that direction is formal description languages for materials science and chemistry, for example, the XDL language by Cronin and coworkers.[74] Intrinsic limitations mostly refer to the completeness, reliability, unambiguousness, and precision of data reported in scientific literature. Materials entity names might not always be unique and pose fundamental challenges to extraction algorithms.[72] Databases constructed from extracted literature data might contain noise and errors[59] due to differences in experimental setups, experimental measurement conditions, reporting accuracies, and missing metadata. Furthermore, even if data extraction from graphs and figures becomes possible and reliable,[49] the reported data might be highly processed and condensed (i.e., lacks possibilities for further analysis of raw data), has limits in accuracy and completeness, and might in many cases be ambiguous. Those intrinsic challenges are inherent to all approaches that aim to extract and collect data from published literature, independent of the reliability of the extraction methods used. Such intrinsic limitations can only be overcome if access to high-quality data and metadata is given directly by the research groups that produce the data, for example, through publication in repositories and databases, rather than through the "information bottleneck" of scientific literature. Given the rapid recent progress in the development of data extraction methods and more generally natural language processing tools, LLMs will likely become one of the most widely used tools to extract (also complex and heterogeneous) data from the literature, as LLMs are capable of systematically analyzing natural language and also generating formal languages, for example, tabular formats or structured data templates. Retraining LLMs on small datasets can help to improve their accuracy for given tasks, which will also become more affordable due to the development of smaller and more efficient LLMs. Major breakthroughs can be expected in the next years regarding systematic, wide-spread efforts to disclose data and knowledge currently hidden in scientific publications. The sustainable provision of—if possible—FAIR data, that is, findable, accessible, interoperable, and re-usable data in suitable databases and repositories could maximize the benefit for the whole scientific community. Main challenges on the way there include the development of more flexible yet formal and thus computer-readable descriptions of complex data structures, the standardization of data and metadata, as well as the further development of data extraction methods to reduce the amount of data missed during extraction as well as to reduce the error rate. However, intrinsic limitations of data extractable from scientific literature indicate that loss of data is unavoidable if the publication of data will not change in the future, implying that further development and use of methods for research data management and FAIR data publication[75] is of the highest importance, to ensure best possible outcomes in data science and ML approaches applied to questions in materials science, chemistry, and beyond.[76]

2.3. Research Data Management to Publishing Data in a FAIR Way

So far, we discussed approaches to extract published data from text, tables, and graphs of research papers and other scientific texts, along with associated limitations and perspectives. However, even if data extraction methods can be perfected, one of the main challenges cannot be solved with this approach, which is the fact that a lot of valuable data is not published at all, as it was considered not successful, not publication-relevant, or not published for other reasons. Nonetheless, this data can be highly relevant and thus valuable in other contexts, indicating the relevance of approaches to decrease the difficulty and thus the barrier to publishing the majority of generated data in a FAIR way, to make it accessible and also findable for other researchers.

It is well-accepted that the systematic collection of research data in digital form and its disclosure is highly important for the transparency and reproducibility of scientific work. If research data management (RDM) can be tied to the FAIR data principles, RDM processes have enormous potential to systematically provide any data the research community needs for a variety of projects. In past decades, the use of efficient tools for digital RDM was difficult to achieve for the wider materials science community due to the lack of the necessary software tools, storage resources, and policies. Meanwhile, great progress has been made in all three areas, especially in recent years, thus at least partial digitization of research processes and modern methods of RDM are technically achievable goals for scientists of many disciplines now. For best-practice guides, we refer to Talley et al.[77] and Herres-Pawlis et al.[78] Nevertheless, for a broad adaptation of RDM processes by scientists, a cultural change—that is, a change of the mindset with respect to the importance of research data and its appropriate storage—is needed.

While this cultural change is progressing only very slowly worldwide, scientists in Germany are facing an important turning point: After several years of preparation, the requirements for the practice of data provision have been changed by the German Research Foundation (DFG), one of the most important funding agencies in Germany, and an extension of the obligation to disclose research data will come into force in 2023.[79,80] Additionally, the importance of a FAIR provision of research data has recently been reaffirmed and strengthened by establishing the National Research Data Infrastructure (NFDI),[81] an infrastructure to store and preserve FAIR research data in Germany. As a result of these and many other changes in the scientific system, researchers are making more and more efforts to adopt existing RDM offerings which can make valuable contributions to the provision of high-quality, standardized, machine-readable data in the long run. In this regard, three essential steps can be described:

2.3.1. Methods and Software for Digitalization Strategies and Data Availability

Electronic laboratory journals (ELNs) or Laboratory Information and Management Systems (LIMS) have been used for decades in the industry as valuable RDM tools for the digitization of research processes. With the availability of powerful open-source software as an alternative to commercial systems, many academic institutions can use these RDM tools now. Thus, research data can be digitally stored and tagged with the relevant metadata as soon as they are created. Open source ELNs such as Chemotion ELN,[82] eLAB,[83] NOMAD ELN,[20] Kadi4Mat,[84] and many others, bring direct advantages, especially with regard to potential subsequent

use of the data: Since they can be extended by own developments and thus, if necessary, also reflect changing requirements, the necessary data and metadata schemas can be made available to the scientists on a permanent basis. Open source ELN software allows the scientists themselves to specify the type and level of detail of the stored information. Automatic test protocols and algorithms can be integrated and used to achieve high data quality and, if necessary, to offer correction suggestions to the scientists. In this way, open source systems in particular form an important basis for the self-determined acquisition of research data and content. The central collection of data enables the collection of all relevant data via one UI. If data from measurement devices are consistently integrated into the ELN/LIMS process, the experiments can be linked to the measurement data without data loss or errors. Especially with regard to the importance of the completeness of data and its quality for ML, this step is a milestone to improve the data situation for various reuse purposes. ELNs are on the one hand the means for complete documentation of research processes for the individual scientist and on the other hand a powerful tool for building community-driven databases that can be searched and reused.

2.3.2. Standardization of Discipline-Specific Data, Processes, and Metadata

In addition to the systematic digital recording and linking of research processes and data, the standardization of data and metadata is particularly important in order to ensure their efficient subsequent use by others. The goal of standardization is to ensure the completeness of reported data and metadata, that is, all relevant variables and parameters should be included in a data standard to ensure qualitative reproducibility, which also includes external conditions which are known to be crucial for the respective experiment. Furthermore, the data accuracy should be high enough to ensure also quantitative reproducibility. Metadata schemas and ontologies are helpful for the standardization of data and processes, as well as to link data published by different researchers. The use of metadata schemas and ontologies becomes accessible to a broad range of scientists through their integration into ELNs. Currently, tools for standardizing data and metadata are also being developed in many initiatives. Examples are the long-existing working groups within IUPAC,[85] NIST,[86] and many others, but also newer interest groups of RDA[87] and subgroups of the NFDI consortia.[88–91] With the mostly direct involvement of scientists, freely available descriptions and software can thus be obtained to enable uniform storage of information. As an example within the Excellence Cluster 3DMM2O, data converters are developed to obtain open, standardized data from non-standard, proprietary file formats that are available for comparative display, analysis, and interpretation.[92] This allows to extract also metadata and to merge it with standardized metadata schemas in a way that enables reuse without the need to develop custom scripts. When these digital tools and standardization elements are embedded in ELNs, the standardized data and metadata can then be made directly usable with appropriate interfaces and form an ever-growing resource for ML by machine-readable data.

2.3.3. Data Publication in Openly Accessible Repositories

If all locally available resources such as ELNs are brought together, there is huge potential for making data available across the entire materials science and chemistry community. This is possible, for example, through the use of research data repositories.[19,93,94] Research data repositories, especially if they provide a subject focus with appropriate support for relevant data and metadata standards, can serve as a central resource for decentralized provided data. In the case of curated repositories, the data can be further enhanced by author-independent, partly automated checks for consistent data quality.[95] Repositories offer many more options to host data in the long run: In addition to the most prominent functions to date for storing and providing research data contributed by the authors themselves, repositories can also be used to provide data extracted from the literature. This can provide a combination of research data repository and database, which may be able to provide a much larger number of data sets than would be possible through direct active contribution by the community based on actual papers. An example of this can be given with the extraction of chemical reactions from several supplementary information files, which has been used in the past to enrich the database of the Chemotion repository.[96] Methods of data extraction can, of course, be used to enrich data available in internal environments such as ELNs—but then the benefit to the community is limited. Being openly accessible, repositories could become a key infrastructure for materials science and the development of new AI methods in the future: Repositories could become the primary resource for obtaining data for ML and many other methods in the future. They could also be the perfect environment for many models obtained through AI to be tested and, if necessary, put to long-term use. AI models, for example, that enable data simulation, data analysis, or curation could enrich repositories with important functions that can be directly harnessed by scientists.[97] Thus, in the long term, repositories could provide the solution to current problems by making data available: Models trained on repository data could contribute to the curation of new data in the future and thus successively increase data quality (for previously non-curated repositories) or decrease processing time and time investment (for curated repositories). Furthermore, when it comes to computational studies, including the development and application of ML methods, not only data but also code should be published to improve reproducibility and accelerate development cycles within the scientific community. Repositories for sharing of open-access code, for example, GitHub are widely used. Best practice guides can be found in Coudert,[98] Wang et al.[99] and Artrith et al.[76]

3. Examples Where Data Mining and Machine Learning Enabled the Design and Application of Functional Material Systems

In this section, we will briefly describe the reuse of structured data with ML models and the benefit for the synthesis and optimization of functional materials systems. The selected examples focus on literature data extraction and ML related to MOF research, but the application of both is of course not limited to the scientific challenges of MOFs and could be applied to many other topics.

The most general use-case of data about materials, molecules, and their properties is computational (inverse) design. To introduce new materials, the conventional trial-and-error approach which involves a long stepwise procedure from molecule design down to experimental assessment has been recently proposed to be replaced by the fully data-driven inverse design methodology to directly design the target molecules.[100,101] Inverse design of materials focuses on identifying the desired properties of materials first and then determining the optimal structure and composition to achieve those properties. The traditional forward design process involves synthesizing and testing a large number of materials in order to find one with the desired characteristics. Inverse design relies on computational methods, in particular ML and thus large amounts of data, to explore vast chemical and structural design spaces more efficiently.[102] As the early applications of inverse design in materials science, Zunger et al. used a genetic algorithm to design solid-state materials with desired electronic properties.[103,104] More recent applications of inverse design were focused on the generative design of polymer dielectrics,[105] MOF membranes,[106] nanomaterials,[107] multilayer metasurfaces,[108] metamaterials,[109] and high entropy alloys.[110]

One of the most important aims of materials design is the improvement of sustainability of technologies, implied by, for example, the sustainable development goals by the UN, aiming to provide a more sustainable future for human society. A paradigm shift in how materials and chemicals discovery is approached, that is, a shift from conventional experimental exploration to computer-aided design and AI-facilitated experimentation, can help to reach sustainability goals. Using ML methods to learn from existing data as much as possible, in order to avoid redundant computations and experiments which consume energy and resources is critical. Without efficiently collecting and reusing previously published data and without reporting newly generated data in FAIR ways, this potential cannot be fully used. At the same time, ML methods can be used to develop materials for sustainable technology. For example, Hardian et al. described how ML methods can be used to produce MOFs in an environmentally friendly way.[111] Kumar et al. investigated several green solvents for the sustainable synthesis of covalent organic frameworks.[112] ML techniques can also be used in most steps of a conventional environmental risk assessment of, for example, smart nanomaterials to ensure sustainability.[113] Moreover, to develop sustainable and eco-friendly alkali-activated material (AAM) or geopolymers, Shah et al. employed ML methods to facilitate and accelerate the development of a one-part AAM binder with the desired properties.[114] Electrocatalysis has received enormous attention as a clean and sustainable technology. In this regard, Chen et al. reviewed the application of ML in electrocatalyst design as a circumvention of the traditional trial-and-error preparation method.[115]

3.1. Synthesis of FMS

The synthesis of MOF-based functional material systems involves multiple steps, starting from the molecular precursors, over the topology and morphology until the final device integration. Pioneering articles showed the possibilities to support researchers in finding suitable conditions using ML optimization algorithms, such as Bayesian optimization or genetic algorithms. Examples by Shields et al.[116] for the synthesis of organic molecules with improved yield and Moosavi et al.[117] for the synthesis of MOFs with improved crystallinity and BET surface area demonstrated the possibilities of using ML to rationally optimize the synthesis conditions for organic molecules and MOF crystals. Chen et al.[118] demonstrated the possibility to employ ML techniques to design MOFs with desired shapes or morphologies and Pilz et al.[119] demonstrated the possibility to optimize crystallinity preferential orientation of interfacially grown SURMOF thin films. However, these approaches rely on the generation of synthesis data on which the algorithms can operate and additionally require the knowledge of the involved scientists to set the parameter and condition space for the optimization algorithms. By operating on large synthesis databases, Segler et al.[120] demonstrated that retrosynthesis design is possible for small organic molecules. The work by Park et al.[73] and Luo et al.[59] demonstrated that automated data extraction can be combined with ML models to predict the synthesis conditions of new MOFs and gain insights into the synthesis process. Taken together, these selected examples demonstrate that automated data extraction and ML techniques are well suited for synthesis planning, parameter prediction, and further optimization of MOF-based functional material systems, starting from the molecular components up to the final MOF structures with desired topology, morphology, and crystal orientation. The combination of such tools promises to accelerate the discovery of new MOFs, especially if additional data become available via extraction from scientific literature or collected in tailored electronic lab notebooks and deposited in openly accessible repositories.

3.2. Optimization of MOF-Based FMS

The design of ideal MOF structures using high throughput computational screening and ML is a highly active and quickly developing area of intense research,[121] enabled by well-structured databases such as the MOF Cambridge structural database subset[28] and curated databases such as CoREMOF,[122] MOFX-DB,[123] ToBaCCo,[124] QMOF,[125] and others.[126] Starting from suitable databases allows the automated screening for ideal structures from a large pool of already synthesized or predicted materials.[127] Despite numerous publications on the design of MOFs via high-throughput computational screening and inverse design, there are only very few experimentally realized target structures.[11,128,129] The reasons why many interesting structures have not been realized experimentally are on the one hand their difficult or very expensive synthesis and on the other hand, their poor stability.[73,130,131] In addition, the communication between theoretical and experimental groups is often challenging, leading to missed opportunities to cooperate.[14,129] Addressing these issues, pioneering work based on simulation and ML for the predictions of mechanical stability by Moghadam et al.[132] and synthesizability by Anderson et al.[133] could be realized. The alternative approach of automated data mining from scientific literature combined with ML proved also a valuable strategy to predict important features of MOF. Important prediction tools were developed by Batra et al.[134] for water stability and Nandy et al.[64,135]

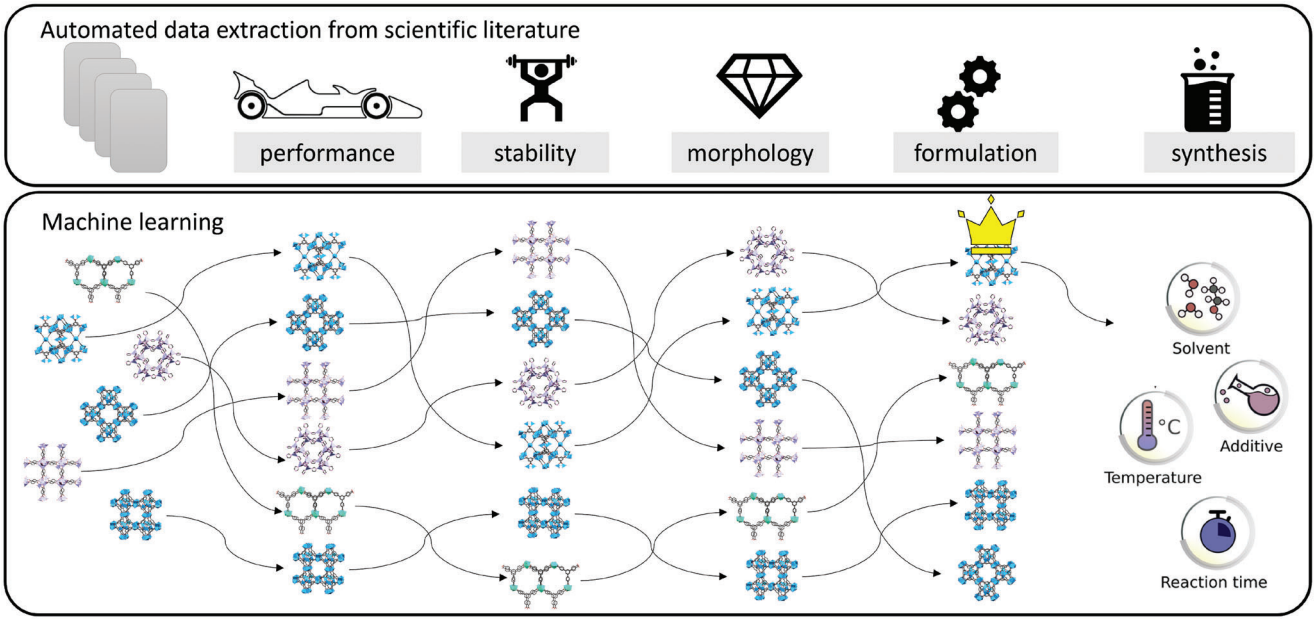

Figure 2. Automated data extraction and machine learning enable researchers to select and synthesize functional material systems tailored for their desired applications.

for thermal stability and stability toward solvent removal. Exploiting the large community knowledge hidden within the scientific literature will further refine these tools and enable the prediction of tailored MOF-based functional material systems for desired applications, that simultaneously fulfill multiple objectives imposed by the processing and operation conditions. **Figure 2** describes the identification of functional material systems for a target application, biased by multiple objectives. The relevant data for such ML-based predictions can be mined from scientific literature via automated data extraction. In addition, the synthesis of the target structure can be facilitated via ML prediction and optimization tools.

4. Conclusions and Outlook

Simulation and machine learning (ML) have evolved as important tools for guiding researchers and for identifying materials of interest. By replacing the traditional heuristic approach, associated with labor and time-intensive trial and error experiments, the computational discovery or inverse design promises to speed up the development of new materials. However, ML approaches rely on sufficient data in machine-readable formats. Combining ML with automated data extraction from scientific literature, using natural language processing, allows not only to gain insights into the ideal design of functional material systems for a desired application but also allows to collect information on important features such as thermal or mechanical stability. An ML workflow can be implemented to utilize the extracted data and identify the ideal design, starting from the composition over the structure across several length scales to the final device. The additional features, such as stability, cost, or abundance of the components can be implemented in the ML workflow as a bias to identify the ideal material under the operating conditions of the desired application. In addition, the use of automatically extracted data on synthesis conditions, in combination with ML, can guide researchers to realize the target materials experimentally. Efficiently operating with such complex interconnected and hierarchical data, involved in functional materials systems, requires the use of advanced research data management tools. In addition, electronic lab notebooks can facilitate the implementation of feedback loops and the complementary use of new experimental data. Although at an early stage, the combination of automated data extraction and ML already showed promising results for the prediction of important properties and synthesis conditions as well as for high throughput computational screening and inverse design of functional material systems. The development of advanced tools such as LLMs (e.g., GPT-3) allows domain specialists in material science to automatically extract datasets to feed ML models. This workflow holds promise to accelerate the development of new functional material systems, urgently needed to tackle global challenges.

Acknowledgements

C.W. and S.B. acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy via the Excellence Cluster "3D Matter Made to Order" (3DMM2O, EXC-2082/1-390761711, Thrust A1 and A3). P.F. acknowledges support by the Federal Ministry of Education and Research (BMBF) under Grant No. 01DM21001B (German-Canadian Materials Acceleration Center). M.T. acknowledges support by the DACStorE project, funded by the Initiative and Networking Fund of the Helmholtz Association (grant agreement number KA2-HSC-12). Initial versions of parts of the manuscript were generated using Chat-GPT based on keypoint lists but fully rewritten by the authors.

Open access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

P.F., M.T., N.J., S.B. and C.W. contributed to the conceptualization of the article. P.K., P.F., N.J. and M.T. contributed to the writing of the original draft. All authors contributed to review and editing of the final draft. M.K. and P.F. contributed to the visualizations.

Keywords

FAIR data, functional materials, large language models, literature data extraction, machine learning, research data management

Received: March 7, 2023
Revised: April 21, 2023
Published online:

[1] X. Zhang, T. Zhou, K. Sundmacher, *AIChE J.* **2022**, *68*, e17788.
[2] E. Ren, P. Guilbaud, F.-X. Coudert, *Digital Discov.* **2022**, *1*, 355.
[3] A. S. Rosen, J. M. Notestein, R. Q. Snurr, *Curr. Opin. Chem. Eng.* **2022**, *35*, 100760.
[4] Y. Luo, M. Ahmad, A. Schug, M. Tsotsalas, *Adv. Mater.* **2019**, *31*, 1901744.
[5] R. Lakes, *Nature* **1993**, *361*, 511.
[6] A. L. Goodwin, *Nat. Commun.* **2019**, *10*, 4461.
[7] B. Seoane, S. Castellanos, A. Dikhtiarenko, F. Kapteijn, J. Gascon, *Coord. Chem. Rev.* **2016**, *307*, 147.
[8] S. Wuttke, D. D. Medina, J. M. Rotter, S. Begum, T. Stassin, R. Ameloot, M. Oschatz, M. Tsotsalas, *Adv. Funct. Mater.* **2018**, *28*, 1801545.
[9] B. Hosseini Monjezi, K. Kutonova, M. Tsotsalas, S. Henke, A. Knebel, *Angew. Chem., Int. Ed.* **2021**, *60*, 15153.
[10] M. Taddei, C. Petit, *Mol. Syst. Des. Eng.* **2021**, *6*, 841.
[11] R. L. Greenaway, K. E. Jelfs, *Adv. Mater.* **2021**, *33*, 2004831.
[12] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, *Nature* **2019**, *571*, 95.
[13] E. A. Olivetti, J. M. Cole, E. Kim, O. Kononova, G. Ceder, T. Y.-J. Han, A. M. Hiszpanski, *Appl. Phys. Rev.* **2020**, *7*, 041317.
[14] M. Rahimi, S. M. Moosavi, B. Smit, T. A. Hatton, *Cell Rep. Phys. Sci.* **2021**, *2*, 4.
[15] M. Ahmad, Y. Luo, C. Wöll, M. Tsotsalas, A. Schug, *Molecules* **2020**, *25*, 4875.
[16] K. M. Jablonka, D. Ongari, S. M. Moosavi, B. Smit, *Chem. Rev.* **2020**, *120*, 8066.
[17] O. Kononova, T. He, H. Huo, A. Trewartha, E. A. Olivetti, G. Ceder, *Iscience* **2021**, *24*, 102155.
[18] L. Himanen, A. Geurts, A. S. Foster, P. Rinke, *Adv. Sci.* **2019**, *6*, 1900808.
[19] P. Tremouilhac, C.-L. Lin, P.-C. Huang, Y.-C. Huang, A. Nguyen, N. Jung, F. Bach, R. Ulrich, B. Neumair, A. Streit, S. Bräse, *Angew. Chem., Int. Ed.* **2020**, *59*, 22771.
[20] M. Scheffler, M. Aeschlimann, M. Albrecht, T. Bereau, H.-J. Bungartz, C. Felser, M. Greiner, A. Groß, C. T. Koch, K. Kremer, W. E. Nagel, M. Scheidgen, C. Wöll, C. Draxl, *Nature* **2022**, *604*, 635.
[21] L. M. Ghiringhelli, C. Baldauf, T. Bereau, S. Brockhauser, C. Carbogno, J. Chamanara, S. Cozzini, S. Curtarolo, C. Draxl, S. Dwaraknath, A. Fekete, J. Kermode, C. T. Koch, M. Kühbach, A. N. Ladines, P. Lambrix, M.-O. Lenz-Himmer, S. Levchenko, M. Oliveira, A. Michalchuk, R. Miller, B. Onat, P. Pavone, G. Pizzi, B. Regler, G.-M. Rignanese, J. Schaarschmidt, M. Scheidgen, A. Schneidewind, T. Sheveleva, et al., *arXiv:2205.14774* **2022**.
[22] R. Freund, S. Canossa, S. M. Cohen, W. Yan, H. Deng, V. Guillerm, M. Eddaoudi, D. G. Madden, D. Fairen-Jimenez, H. Lyu, L. K. Macreadie, Z. Ji, Y. Zhang, B. Wang, F. Haase, C. Wöll, O. Zaremba, J. Andreo, S. Wuttke, C. S. Diercks, *Angew. Chem., Int. Ed.* **2021**, *60*, 23946.
[23] O. M. Yaghi, M. O'Keeffe, N. W. Ockwig, H. K. Chae, M. Eddaoudi, J. Kim, *Nature* **2003**, *423*, 705.
[24] S. Kitagawa, R. Kitaura, S.-i. Noro, *Angew. Chem., Int. Ed.* **2004**, *43*, 2334.
[25] D.-H. Chen, H. Gliemann, C. Wöll, *Chem. Phys. Rev.* **2023**, *4*, 011305.
[26] R. L. Siegelman, E. J. Kim, J. R. Long, *Nat. Mater.* **2021**, *20*, 1060.
[27] Y. G. Chung, E. Haldoupis, B. J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling, J. S. Camp, B. Slater, J. I. Siepmann, D. S. Sholl, R. Q. Snurr, *J. Chem. Eng. Data* **2019**, *64*, 5985.
[28] P. Z. Moghadam, A. Li, S. B. Wiggin, A. Tao, A. G. Maloney, P. A. Wood, S. C. Ward, D. Fairen-Jimenez, *Chem. Mater.* **2017**, *29*, 2618.
[29] S. M. Moosavi, K. M. Jablonka, B. Smit, *J. Am. Chem. Soc.* **2020**, *142*, 20273.
[30] S. Furukawa, J. Reboul, S. Diring, K. Sumida, S. Kitagawa, *Chem. Soc. Rev.* **2014**, *43*, 5700.
[31] N. Stock, S. Biswas, *Chem. Rev.* **2012**, *112*, 933.
[32] S. Dissegna, K. Epp, W. R. Heinz, G. Kieslich, R. A. Fischer, *Adv. Mater.* **2018**, *30*, 1704501.
[33] H. Gliemann, C. Wöll, *Mater. Today* **2012**, *15*, 110.
[34] J. Dechnik, J. Gascon, C. J. Doonan, C. Janiak, C. J. Sumby, *Angew. Chem., Int. Ed.* **2017**, *56*, 9292.
[35] M. Tsotsalas, A. Umemura, F. Kim, Y. Sakata, J. Reboul, S. Kitagawa, S. Furukawa, *J. Mater. Chem.* **2012**, *22*, 10159.
[36] X. Yin, C. E. Gounaris, *Comput. Chem. Eng.* **2022**, *167*, 108022.
[37] H. Lyu, Z. Ji, S. Wuttke, O. M. Yaghi, *Chem* **2020**, *6*, 2219.
[38] S. Chong, S. Lee, B. Kim, J. Kim, *Coord. Chem. Rev.* **2020**, *423*, 213487.
[39] A. M. Cohen, W. R. Hersh, *Briefings Bioinf.* **2005**, *6*, 57.
[40] M. A. Andrade, A. Valencia, *In Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1997**, *5*, 25.
[41] P. Corbett, P. Murray-Rust, in *Proceedings of theComputational Life Sciences II: Second International Symposium, CompLife*, Springer, Berlin, Heidelberg **2006**, pp. 107–118.
[42] L. Hawizy, D. M. Jessop, N. Adams, P. Murray-Rust, *J. Chem.* **2011**, *3*, 17.
[43] M. C. Swain, J. M. Cole, *J. Chem. Inf. Model.* **2016**, *56*, 1894.
[44] J. Mavracic, C. J. Court, T. Isazawa, S. R. Elliott, J. M. Cole, *J. Chem. Inf. Model.* **2021**, *61*, 4280.
[45] E. Kim, K. Huang, A. Tomala, S. Matthews, E. Strubell, A. Saunders, A. McCallum, E. Olivetti, *Sci. Data* **2017**, *4*, 170127.
[46] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, *arXiv:1810.04805* **2018**.
[47] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *OpenAI blog* **2019**, *1*, 9.
[48] T. Gupta, M. Zaki, N. A. Krishnan, *npj Comput. Mater.* **2022**, *8*, 102.
[49] S. Huang, J. M. Cole, *Chem. Sci.* **2022**, *13*, 11487.
[50] A. Dunn, J. Dagdelen, N. Walker, S. Lee, A. S. Rosen, G. Ceder, K. Persson, A. Jain, *arXiv:2212.05238* **2022**.
[51] L. Ghadbeigi, J. K. Harada, B. R. Lettiere, T. D. Sparks, *Energy Environ. Sci.* **2015**, *8*, 1640.
[52] S. R. Young, A. Maksov, M. Ziatdinov, Y. Cao, M. Burch, J. Balachandran, L. Li, S. Somnath, R. M. Patton, S. V. Kalinin, R. K. Vasudevan, *J. Appl. Phys.* **2018**, *123*, 115303.
[53] F. Baum, T. Pretto, A. Köche, M. J. L. Santos, *J. Phys. Chem. C* **2020**, *124*, 24298.
[54] P. Raccuglia, K. C. Elbert, P. D. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, *Nature* **2016**, *533*, 73.

[55] J. R. McDaniel, J. R. Balmuth, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 373.
[56] M. Oldenhof, A. Arany, Y. Moreau, J. Simm, *J. Chem. Inf. Model.* **2020**, *60*, 4506.
[57] K. Rajan, A. Zielesny, C. Steinbeck, *J. Cheminf.* **2020**, *12*, 65.
[58] S. Yoo, O. Kwon, H. Lee, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Piscataway, NJ **2022**, pp. 3393–3397.
[59] Y. Luo, S. Bag, O. Zaremba, A. Cierpka, J. Andreo, S. Wuttke, P. Friederich, M. Tsotsalas, *Angew. Chem., Int. Ed.* **2022**, *61*, e202200242.
[60] A. Davariashtiyani, Z. Kadkhodaie, S. Kadkhodaei, *Commun. Mater.* **2021**, *2*, 115.
[61] Z. Jensen, E. Kim, S. Kwon, T. Z. Gani, Y. Román-Leshkov, M. Moliner, A. Corma, E. Olivetti, *ACS Cent. S.* **2019**, *5*, 892.
[62] C. Karpovich, E. Pan, Z. Jensen, E. Olivetti, *Chem. Mater.* **2023**.
[63] O. Kononova, H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan, G. Ceder, *Sci. Data* **2019**, *6*, 203.
[64] A. Nandy, G. Terrones, N. Arunachalam, C. Duan, D. W. Kastner, H. J. Kulik, *Sci. Data* **2022**, *9*, 74.
[65] C. J. Court, J. M. Cole, *npj Comput. Mater.* **2020**, *6*, 18.
[66] C. J. Court, A. Jain, J. M. Cole, *Chem. Mater.* **2021**, *33*, 7217.
[67] C. J. Court, J. M. Cole, *Sci. Data* **2018**, *5*, 18.
[68] Q. Dong, J. M. Cole, *Sci. Data* **2022**, *9*, 193.
[69] O. Sierepeklis, J. M. Cole, *Sci. Data* **2022**, *9*, 648.
[70] M. Zaki, Jayadeva, N. A. Krishnan, *Chem. Eng. Process.* **2022**, *180*, 108607.
[71] Z. Nie, S. Zheng, Y. Liu, Z. Chen, S. Li, K. Lei, F. Pan, *Adv. Funct. Mater.* **2022**, *32*, 262201437.
[72] T. He, W. Sun, H. Huo, O. Kononova, Z. Rong, V. Tshitoyan, T. Botari, G. Ceder, *Chem. Mater.* **2020**, *32*, 7861.
[73] H. Park, Y. Kang, W. Choe, J. Kim, *J. Chem. Inf. Model.* **2022**, *62*, 1190.
[74] S. H. M. Mehr, M. Craven, A. I. Leonov, G. Keenan, L. Cronin, *Science* **2020**, *370*, 101.
[75] J. D. Evans, V. Bon, I. Senkovska, S. Kaskel, *Langmuir* **2021**, *37*, 4222.
[76] N. Artrith, K. T. Butler, F.-X. Coudert, S. Han, O. Isayev, A. Jain, A. Walsh, *Nat. Chem.* **2021**, *13*, 505.
[77] K. R. Talley, R. White, N. Wunder, M. Eash, M. Schwarting, D. Evenson, J. D. Perkins, W. Tumas, K. Munch, C. Phillips, A. Zakutayev, *Patterns* **2021**, *2*, 100373.
[78] S. Herres-Pawlis, F. Bach, I. J. Bruno, S. J. Chalk, N. Jung, J. C. Liermann, L. R. McEwen, S. Neumann, C. Steinbeck, M. Razum, O. Koepler, *Angew. Chem., Int. Ed.* **2022**, *61*, e202203038.
[79] Good research practice, https://www.dfg.de/en/research_funding/principles_dfg_funding/good_scientific_practice (accessed: May 2023).
[80] Deutsche Forschungsgemeinschaft, *Code of Conduct* **2019** https://zenodo.org/record/6472827#.ZGeOedJBwUE (accessed: May 2023).
[81] N. Hartl, E. Wössner, Y. Sure-Vetter, *Inf. Spektrum* **2021**, *44*, 370.
[82] P. Tremouilhac, A. Nguyen, Y.-C. Huang, S. Kotov, D. S. Lütjohann, F. Hübsch, N. Jung, S. Bräse, *J. Cheminf.* **2017**, *9*, 54.
[83] N. CARP, A. Minges, M. Piel, *J. Open Source Software* **2017**, *2*, 146.
[84] N. Brandt, E. Schoof, P. Zschumme, M. Selzer, *A Research Data Infrastructure for Materials Science*, Heidelberg, University, Heidelberg **2021**.
[85] International union of pure & applied chemistry, https://iupac.org/who-we-are (accessed: May 2023).
[86] National institute of standards and technology, https://www.nist.gov (accessed: May 2023).
[87] RDA/CODATA materials data, infrastructure & interoperability IG, https://www.rd-alliance.org/groups/rdacodata-materials-data-infrastructure-interoperability-ig.html (accessed: May 2023).
[88] C. Eberl, M. Niebel, E. Bitzek, T. Dahmen, F. Fritzen, P. Gumbsch, T. Hickel, S. Klein, F. Mücklich, M. S. Müller, et al., **2021** https://doi.org/10.5281/zenodo.5082836.
[89] C. Steinbeck, O. Koepler, S. Herres-Pawlis, F. Bach, N. Jung, M. Razum, J. C. Liermann, S. Neumann, *Chem. Int.* **2023**, *45*, 8.
[90] C. Steinbeck, O. Koepler, F. Bach, S. Herres-Pawlis, N. Jung, J. Liermann, S. Neumann, M. Razum, C. Baldauf, F. Biedermann, T. W. Bocklitz, F. Boehm, F. Broda, P. Czodrowski, T. Engel, M. G. Hicks, S. M. Kast, C. Kettner, W. Koch, G. Lanza, A. Link, R. A. Mata, W. E. Nagel, A. Porzel, N. Schlörer, T. Schulze, H.-G. Weinig, W. Wenzel, L. A. Wessjohann, S. Wulle, *Res. Ideas Outcomes* **2020**, *6*, e55852.
[91] H. Junkes, P. Oppermann, R. Schlögl, A. Trunschke, M. Krieger, H. Weber, in *18th Int. Conf. on Accelerator and Large Experimental Physics Control Systems*, JACoW Publishing, **2021**, pp. 558–563.
[92] J. Klar, M. Starman, P. C. Huang, Complat - compound platform @ karlsruhe institute of technology (kit) **2023**, https://github.com/ComPlat/chemotion-converter-app (accessed: May 2023).
[93] C. Henken, M. Schmidt, Chemotion eln and repository, https://www.youtube.com/watch?v=tZHaP6DW-Dw (accessed: May 2023).
[94] C. Draxl, M. Scheffler, *MRS Bull.* **2018**, *43*, 676.
[95] P. Tremouilhac, P.-C. Huang, C.-L. Lin, Y.-C. Huang, A. Nguyen, N. Jung, F. Bach, S. Bräse, *Chem. Methods* **2021**, *1*, 8.
[96] A. Nguyen, Y.-C. Huang, P. Tremouilhac, N. Jung, S. Bräse, *J. Cheminf.* **2019**, *11*, 77.
[97] L. Sbailò, Á. Fekete, L. M. Ghiringhelli, M. Scheffler, *npj Comput. Mater.* **2022**, *8*, 250.
[98] F.-X. Coudert, *Chem. Mater.* **2017**, *29*, 2615.
[99] A. Y.-T. Wang, R. J. Murdock, S. K. Kauwe, A. O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson, T. D. Sparks, *Chem. Mater.* **2020**, *32*, 4954.
[100] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 268.
[101] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360.
[102] K. Kim, S. Kang, J. Yoo, Y. Kwon, Y. Nam, D. Lee, I. Kim, Y.-S. Choi, Y. Jung, S. Kim, W.-J. Son, J. Son, H. S. Lee, S. Kim, J. Shin, S. Hwang, *npj Comput. Mater.* **2018**, *4*, 67.
[103] S. Dudiy, A. Zunger, *Phys. Rev. Lett.* **2006**, *97*, 046401.
[104] P. Piquini, P. A. Graf, A. Zunger, *Phys. Rev. Lett.* **2008**, *100*, 186403.
[105] R. Gurnani, D. Kamal, H. Tran, H. Sahu, K. Scharm, U. Ashraf, R. Ramprasad, *Chem. Mater.* **2021**, *33*, 7008.
[106] M. Zhou, A. Vassallo, J. Wu, *J. Membr. Sci.* **2020**, *598*, 117675.
[107] S. Li, A. S. Barnard, *Adv. Theory Simul.* **2022**, *5*, 2100414.
[108] P. Naseri, S. V. Hum, *IEEE Trans. Antennas Propag.* **2021**, *69*, 5725.
[109] Y. Wang, Q. Zeng, J. Wang, Y. Li, D. Fang, *Comput. Methods Appl. Mech. Eng.* **2022**, *401*, 115571.
[110] Y. Zeng, M. Man, C. K. Ng, D. Wuu, J. J. Lee, F. Wei, P. Wang, K. Bai, D. C. Cheh Tan, Y.-W. Zhang, *APL Mater.* **2022**, *10*, 101104.
[111] R. Hardian, Z. Liang, X. Zhang, G. Szekely, *Green Chem.* **2020**, *22*, 7521.
[112] S. Kumar, G. Ignacz, G. Szekely, *Green Chem.* **2021**, *23*, 8932.
[113] J. J. Scott-Fordsmand, M. J. Amorim, *Sci. Total Environ.* **2023**, *859*, 160303.
[114] S. F. A. Shah, B. Chen, M. Zahid, M. R. Ahmad, *Constr. Build. Mater.* **2022**, *360*, 129534.
[115] L. Chen, X. Zhang, A. Chen, S. Yao, X. Hu, Z. Zhou, *Chin. J. Catal.* **2022**, *43*, 11.
[116] B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams, A. G. Doyle, *Nature* **2021**, *590*, 89.
[117] S. M. Moosavi, A. Chidambaram, L. Talirz, M. Haranczyk, K. C. Stylianou, B. Smit, *Nat. Commun.* **2019**, *10*, 539.

[118] P. Chen, Z. Tang, Z. Zeng, X. Hu, L. Xiao, Y. Liu, X. Qian, C. Deng, R. Huang, J. Zhang, Y. Bi, R. Lin, Y. Zhou, H. Liao, D. Zhou, C. Wang, W. Lin, *Matter* **2020**, *2*, 1651.
[119] L. Pilz, C. Natzeck, J. Wohlgemuth, N. Scheuermann, P. G. Weidler, I. Wagner, C. Wöll, M. Tsotsalas, *Adv. Mater. Interfaces* **2023**, *10*, 2201771.
[120] M. H. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604.
[121] Y. J. Colón, R. Q. Snurr, *Chem. Soc. Rev.* **2014**, *43*, 5735.
[122] Y. G. Chung, J. Camp, M. Haranczyk, B. J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O. K. Farha, D. S. Sholl, R. Q. Snurr, *Chem. Mater.* **2014**, *26*, 6185.
[123] N. S. Bobbitt, K. Shi, B. J. Bucior, H. Chen, N. Tracy-Amoroso, Z. Li, Y. Sun, J. H. Merlin, J. I. Siepmann, D. W. Siderius, R. Q. Snurr, *J. Chem. Eng. Data* **2023**, *68*, 483.
[124] Y. J. Colón, D. A. Gomez-Gualdron, R. Q. Snurr, *Cryst. Growth Des.* **2017**, *17*, 5801.
[125] A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein, R. Q. Snurr, *Matter* **2021**, *4*, 1578.
[126] S. M. Moosavi, A. Nandy, K. M. Jablonka, D. Ongari, J. P. Janet, P. G. Boyd, Y. Lee, B. Smit, H. J. Kulik, *Nat. Commun.* **2020**, *11*, 4068.
[127] H. Daglar, H. C. Gulbalkan, G. Avci, G. O. Aksu, O. F. Altundal, C. Altintas, I. Erucar, S. Keskin, *Angew. Chem., Int. Ed.* **2021**, *60*, 7828.
[128] D. A. Gomez-Gualdron, O. V. Gutov, V. Krungleviciute, B. Borah, J. E. Mondloch, J. T. Hupp, T. Yildirim, O. K. Farha, R. Q. Snurr, *Chem. Mater.* **2014**, *26*, 5632.
[129] A. Li, R. Bueno-Perez, D. Madden, D. Fairen-Jimenez, *Chem. Sci.* **2022**, *13*, 7990.
[130] F. T. Szczypiński, S. Bennett, K. E. Jelfs, *Chem. Sci.* **2021**, *12*, 830.
[131] S. Bennett, F. T. Szczypinski, L. Turcani, M. E. Briggs, R. L. Greenaway, K. E. Jelfs, *J. Chem. Inf. Model.* **2021**, *61*, 4342.
[132] P. Z. Moghadam, S. M. Rogge, A. Li, C.-M. Chow, J. Wieme, N. Moharrami, M. Aragones-Anglada, G. Conduit, D. A. Gomez-Gualdron, V. Van Speybroeck, D. Fairen-Jimenez, *Matter* **2019**, *1*, 219.
[133] R. Anderson, D. A. Gómez-Gualdrón, *Chem. Mater.* **2020**, *32*, 8106.
[134] R. Batra, C. Chen, T. G. Evans, K. S. Walton, R. Ramprasad, *Nat. Mach. Intell.* **2020**, *2*, 704.
[135] A. Nandy, C. Duan, H. J. Kulik, *J. Am. Chem. Soc.* **2021**, *143*, 17535.

Payam Kalhor received his Ph.D. in Chemistry under the supervision of Zhi-Wu Yu at Tsinghua University where he studied deep eutectic solvents using (excess) spectroscopy and computational chemistry. Then, he was awarded a Postdoctoral Fellowship at Peking University to continue his research on green solvents using (path integral) molecular simulation at Jian Liu's group. He is now a postdoctoral researcher at the Karlsruhe Institute of Technology, working with Pascal Friederich to develop and apply (multi-task) machine learning models for prediction purposes.

Manuel Tsotsalas is an experimental chemist who completed his Ph.D. under the supervision of Luisa DeCola at the University of Münster. He then moved to Kyoto University for a postdoctoral stay with Susumu Kitagawa before joining the Karlsruhe Institute of Technology (KIT) as a Helmholtz young investigator group leader. His research interests center on the interfacial synthesis and hierarchical structuring of adaptive materials, as well as their applications in life science and sustainability. He is currently a visiting scientist at Northwestern University, where he works in the group of Randall Snurr on combining experimental and computational workflows for accelerated material discovery.

Pascal Friederich, after his Ph.D. in physics under the supervision of Wolfgang Wenzel, received a Marie-Sklodowska-Curie Postdoctoral Fellowship at Harvard University and the University of Toronto where he worked with Alán Aspuru-Guzik on machine learning methods for chemistry. In 2020 he was appointed assistant professor at the Informatics Department of the Karlsruhe Institute of Technology, leading the AI for Materials Science (AiMat) research group. His research focuses on developing and applying machine learning methods for property prediction, simulation, understanding, and design of molecules and materials. In 2022, Pascal Friederich received the Heinz-Maier-Leibnitz Prize from the German Research Foundation.