Functional Material Systems Enabled by Automated Data Extraction and Machine Learning

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

Dr. Payam Kalhor
Institute of Nanotechnology, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Am Fasanengarten 5, 76131 Karlsruhe, Germany
Dr. Nicole Jung
Institute for Organic Chemistry, Karlsruhe Institute of Technology, Fritz-Haber-Weg 6, 76131 Karlsruhe, Germany
Institute of Biological and Chemical Systems - Functional Molecular Systems, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
Prof. Dr. Stefan Bräse
Institute for Organic Chemistry, Karlsruhe Institute of Technology, Fritz-Haber-Weg 6, 76131 Karlsruhe, Germany
Prof. Dr. Christof Wöll
Institute of Functional Interfaces, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
T.T.-Prof. Dr. Pascal Friederich
Institute of Nanotechnology, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Am Fasanengarten 5, 76131 Karlsruhe, Germany
Email Address: pascal.friederich@kit.edu
Dr. Manuel Tsotsalas
Institute for Organic Chemistry, Karlsruhe Institute of Technology, Fritz-Haber-Weg 6, 76131 Karlsruhe, Germany
Institute of Functional Interfaces, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
Email Address: manuel.tsotsalas@kit.edu

Keywords: *Functional Materials, Literature Data Extraction, Machine Learning, Large Language Models, Materials Design, Research Data Management*

The development of functional materials is crucial for addressing global challenges such as clean energy, water, and food supply, the discovery of new drugs and antibiotics, and climate change. However, the current process for developing and bringing new materials to market is time-consuming and requires significant financial and human resources. Functional material systems are typically composed of functional molecular building blocks, organized across multiple length scales in a hierarchical order. This large design space allows for precise tuning of properties and tailoring to specific applications, but also makes it difficult to screen for optimal structures using traditional trial and error or high-throughput experimental methods. Machine learning (ML) models have the potential to revolutionize the field of materials science by predicting synthesis and materials properties with high accuracy. However, ML models require data to be trained and validated. Methods to automatically extract data from scientific literature makes it possible to build large and diverse datasets for ML models. In this perspective article, we will discuss opportunities and challenges of data extraction and machine learning methods to accelerate the discovery of high-performing functional material systems, while ensuring that the predicted materials are stable, synthesizable, scalable, and sustainable. We discuss the potential impact of large language models (LLM) on the data extraction process and ML workflows. Additionally, we will discuss the importance of research data management tools to overcome 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, 2, 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, 5, 6, 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, 9, 10]. To support and enable research in the area of functional material systems, automated data extraction from literature, using natural language processing, combined with machine learning can be used to operate on large amounts of data representing community knowledge to complement the researchers' own knowledge and experimental results [11, 12, 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 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, e.g. 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 machine learning of the material science literature and databases [14, 15, 16]. The synthesis, characterization, processing, and application of functional material systems produce large amounts of hierarchical, or interdependent data. Making this data machine-readable and ready for machine learning and combining it with data extracted from scientific literature represents 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, 19, 20]. In this perspective, we will briefly outline the design, synthesis, and characterization of functional material systems using metal-organic frameworks as example materials. Following this outline, we will highlight selected publications on enabling functional materials systems using a combination of automated data extraction and machine learning. 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 machine learning are essential to fully explore the potential of functional material systems in addressing urgent social, economic, and environmental challenges.

2 Functional material systems

Functional material systems are typically composed of functional molecular building blocks, organized across multiple length scales in a hierarchical order. Metal-Organic Frameworks (MOFs) emerged as a particularly powerful class of functional material systems [22, 23, 24]. Their modular synthesis enables the incorporation of diverse functionalities and tuning of their structures for desired applications [25]. The chemical design space of new metal-organic frameworks (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 [26, 27]. 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 [28]. Multiple techniques were developed for the synthesis of MOFs to control their structure across multiple length scales [29]. Starting from the synthesis of the organic linkers and precursors of metal nodes to the crystal synthesis and 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 [30]. 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 [31, 32]. The MOF materials can afterward be processed, e.g. as thin films or freestanding membranes, or formulated e.g. by mixing with polymers, palleted, and processed to the required shapes for the final device [33, 34].
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 [35]. Exposure of this treasure of data and making it ready for machine learning 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 [36, 37]. To fully exploit this treasure of data, a combination of tailored research data management tools, efficient data extraction from scientific literature, and machine learning are essential.

3 Data extraction

One of the main challenges in applying machine learning 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 machine learning models for particular tasks, e.g. the prediction of materials properties. However, the majority of potentially relevant data generated in scientific labs are 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 machine learning 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, focussing on text extraction but also discussing the extraction of information from tables, graphs, and images. Data extraction in other scientific domains, e.g. biology dates back more than 20 years [38], with seminal work in the late 90s, e.g. Andrade *et al.* [39]. One of the earliest attempts to automatically extract information from chemistry literature was OSCAR [40] and based on that the ChemicalTagger method in 2011 [41]. 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 [42, 43] extends the rule-based natural language processing approach further, among others with machine learning methods, and adds functionality for table extraction [43]. During the last years, machine learning approaches started to play an increasingly important role in literature data extraction, where e.g. article section relevance scores [44] and learned word embeddings [44, 12] were used to enhance existing information extraction methods, or conditional random field models were used. With increasing capabilities of language models such as BERT [45] and GPT [46], new possibilities for extracting information in literature are generated. Seminar examples of literature extraction methods based on large language models include MatSciBERT by Gupta *et al.* [47]; a fine-tuned BERT model for materials science by Huang *et al.* [48], BatteryBert, which among others use question-answering algorithms to translate text to structured information; and a GPT-3 based model by Dunn *et al.* [49] 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 materi-

als science literature were reported [50, 51, 52], also extracting information from sources other than scientific literature, e.g. lab notebooks to retrieve data about failed experiments which are usually not reported in scientific articles [53]. 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 [43], graphs [48], and images, in particular optical chemical structure recognition (OCSR), i.e. the extraction of chemical structures from images [54, 55, 56, 57] is beyond of 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 [58, 59, 60, 52, 61, 62] over materials stability [63] to materials properties, e.g. for magnetic and superconducting properties [64, 65, 66], semi-conductors [67], battery materials [48], thermoelectric materials [68], glasses [69] and more general knowledge graphs [12, 70]. In most cases, the databases are only a means to an end, i.e. to provide sufficient training data for machine learning models for the prediction of synthesis routes and conditions as well as materials properties of a wider range of materials.

4 Technical challenges and intrinsic limitations

Despite fast progress and promising new avenues related to the increasing use of machine learning and in particular large language models 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 is 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 e.g. abstracts, the amount of information that can be extracted is rather limited [70]. Furthermore, the use of large language models (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) [48], 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 [49]. At the same time, LLMs might potentially help analyze complex texts and sentence structures, which are not extractable using conventional approaches [71]. Beyond that, one of the main challenges in literature data extraction currently is related to the fact that large amounts of data, e.g. synthesis protocols are not tabular data but can only be represented in more complex data structures as they represent flexible, potentially multistep processes with dynamic data types and complex relations [72], which not only requires the development of extraction methods but also of flexible data blueprints for complex scientific data. One development in that direction are formal description languages for materials science and chemistry, e.g. the XDL language by Cronin and coworkers [73]. 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 [71]. Databases constructed from extracted literature data might contain noise and errors [58] 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 [48], the reported data might be highly processed and condensed (i.e. lacks possibilities for further analysis of raw data), has limits in accuracy, 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, e.g. 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, major breakthroughs can be expected in the next years regarding systematic, wide-spread efforts to transfer data and knowledge currently hidden in scientific publications into FAIR data, i.e. into accessible, findable and computer-readable databases. Main challenges on the way there include the development of more flexible yet formal and thus computer-readable descriptions of complex data structures, 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 make it possible to further develop methods for research data management and FAIR data publication [74].

5 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.
* Development and implementation of research data management tools combined with data extraction from literature and curation * RDM tools (examples we can discuss are Chemotion and NOMAD/FAIRmat) * Material genome initiative * Data formats * Data templates * Domain/level specific: synthesis of compounds, synthesis of materials, morphology control, device design * Challenges and limitations * Data cleaning and thus data reliability * Data compatibility and metadata

6 Examples where data mining and machine learning enabled the design and application of functional material systems

7 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 machine learning optimization algorithms, such as Bayesian Optimization or Genetic algorithms. Examples by B. Shields *et al.* [75] for the synthesis of organic molecules with improved yield and Moosavi *et al.* [76] for the synthesis of MOFs with improved crystallinity and BET surface area demonstrated the possibilities of using machine learning to rationally optimize the synthesis conditions for organic molecules and MOF crystals. P. Chen *et al.* [77] demonstrated the possibility to employ machine learning techniques to design MOFs with desired shapes or morphologies and L. Pilz *et al.* [78] 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 knowledge of the involved scientists to set the parameter and condition space for the optimization algorithms. By operating on large synthesis databases, M. Seeger *et al.* [79] demonstrated that retrosynthesis design is possible for small organic molecules. The work by H. Park *et al.* [72] and by Y. Luo *et al.* [58] demonstrated that automated data extraction can be combined with machine learning 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 machine learning techniques are well suited for synthesis planning, parameter prediction, and further optimiza-

tion 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.

8 Optimization of MOF-based FMS

The design of ideal MOF structures using high throughput computational screening and machine learning is a highly active and quickly developing area of intense research, [80] enabled by well-structured databases such as the MOF Cambridge Structural Database Subset [27] and curated databases such as CoREMOF, [81] MOFX-DB, [82] ToBaCCo, [83] QMOF, [84] and others [85]. Starting from suitable databases allows the automated screening for ideal structures from a large pool of already synthesized or predicted materials [86]. 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, 87, 88]. 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 [89, 72, 90]. In addition, the communication between theoretical and experimental groups is often challenging, leading to missed opportunities to cooperate [88, 14]. Addressing these issues, pioneering work based on simulation and machine learning for the predictions of mechanical stability by Moghadam et al. [91] and synthesizability by R. Anderson et al. [92] could be realized. The alternative approach of automated data mining from scientific literature combined with machine learning proved also a valuable strategy to predict important features of MOF. Important prediction tools were developed by Batra et al. [93] for water stability and Nandy et al. [63, 94] for thermal stability and stability towards 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 machine learning based predictions can be mined from scientific literature via automated data extraction. In addition, the synthesis of the target structure can be facilitated via machine learning prediction and optimization tools.

9 Conclusions and outlook

Simulation and machine learning 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, machine learning approaches rely on sufficient data in machine readable formats. Combining machine learning 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. A machine learning 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 machine learning workflow as 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 machine learning, 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 machine learning 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 large language models (e.g. GPT-3) allows domain specialists in material science to automatically extract datasets to feed machine learning models. This workflow holds promise to accelerate the development of new functional material systems, urgently needed to tackle global challenges.

9.1 First Subsection

9.1.1 First Sub Subsection

First lowest-level subsection:

10 Conclusion

11 Experimental Section

First part of experimental section:

Second part of experimental section:

Acknowledgements
3DMMO: Todo. P.F. and P.K. acknowledge 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).

References

[1] X. Zhang, T. Zhou, K. Sundmacher, *AIChE Journal* **2022**, *68*, 9 e17788.

[2] E. Ren, P. Guilbaud, F.-X. Coudert, *Digital Discovery* **2022**, *1*, 4 355.

[3] A. S. Rosen, J. M. Notestein, R. Q. Snurr, *Current Opinion in Chemical Engineering* **2022**, *35* 100760.

[4] Y. Luo, M. Ahmad, A. Schug, M. Tsotsalas, *Advanced materials* **2019**, *31*, 26 1901744.

[5] R. Lakes, *Nature* **1993**, *361*, 6412 511.

[6] A. L. Goodwin, *Nature Communications* **2019**, *10*, 1 4461.

[7] B. Seoane, S. Castellanos, A. Dikhtiarenko, F. Kapteijn, J. Gascon, *Coordination Chemistry Reviews* **2016**, *307* 147.

[8] S. Wuttke, D. D. Medina, J. M. Rotter, S. Begum, T. Stassin, R. Ameloot, M. Oschatz, M. Tsotsalas, *Advanced Functional Materials* **2018**, *28*, 44 1801545.

[9] B. Hosseini Monjezi, K. Kutonova, M. Tsotsalas, S. Henke, A. Knebel, *Angewandte Chemie International Edition* **2021**, *60*, 28 15153.

[10] M. Taddei, C. Petit, *Molecular Systems Design & Engineering* **2021**, *6*, 11 841.

[11] R. L. Greenaway, K. E. Jelfs, *Advanced Materials* **2021**, *33*, 11 2004831.

[12] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, *Nature* **2019**, *571*, 7763 95.

[13] E. A. Olivetti, J. M. Cole, E. Kim, O. Kononova, G. Ceder, T. Y.-J. Han, A. M. Hiszpanski, *Applied Physics Reviews* **2020**, *7*, 4 041317.

[14] M. Rahimi, S. M. Moosavi, B. Smit, T. A. Hatton, *Cell reports physical science* **2021**, *2*, 4.

[15] M. Ahmad, Y. Luo, C. Wöll, M. Tsotsalas, A. Schug, *Molecules* **2020**, *25*, 21 4875.

[16] K. M. Jablonka, D. Ongari, S. M. Moosavi, B. Smit, *Chemical reviews* **2020**, *120*, 16 8066.

[17] O. Kononova, T. He, H. Huo, A. Trewartha, E. A. Olivetti, G. Ceder, *Iscience* **2021**, *24*, 3 102155.

[18] L. Himanen, A. Geurts, A. S. Foster, P. Rinke, *Advanced Science* **2019**, *6*, 21 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, et al., *Angewandte Chemie International Edition* **2020**, *59*, 50 22771.

[20] M. Scheffler, M. Aeschlimann, M. Albrecht, T. Bereau, H.-J. Bungartz, C. Felser, M. Greiner, A. Groß, C. T. Koch, K. Kremer, et al., *Nature* **2022**, *604*, 7907 635.

[21] L. M. Ghiringhelli, C. Baldauf, T. Bereau, S. Brockhauser, C. Carbogno, J. Chamanara, S. Cozzini, S. Curtarolo, C. Draxl, S. Dwaraknath, et al., *arXiv preprint 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, et al., *Angewandte Chemie International Edition* **2021**, *60*, 45 23946.

[23] O. M. Yaghi, M. O'Keeffe, N. W. Ockwig, H. K. Chae, M. Eddaoudi, J. Kim, *Nature* **2003**, *423*, 6941 705.

[24] S. Kitagawa, R. Kitaura, S.-i. Noro, *Angewandte Chemie International Edition* **2004**, *43*, 18 2334.

[25] R. L. Siegelman, E. J. Kim, J. R. Long, *Nature materials* **2021**, *20*, 8 1060.

[26] Y. G. Chung, E. Haldoupis, B. J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling, J. S. Camp, et al., *Journal of Chemical & Engineering Data* **2019**, *64*, 12 5985.

[27] P. Z. Moghadam, A. Li, S. B. Wiggin, A. Tao, A. G. Maloney, P. A. Wood, S. C. Ward, D. Fairen-Jimenez, *Chemistry of Materials* **2017**, *29*, 7 2618.

[28] S. M. Moosavi, K. M. Jablonka, B. Smit, *Journal of the American Chemical Society* **2020**, *142*, 48 20273.

[29] S. Furukawa, J. Reboul, S. Diring, K. Sumida, S. Kitagawa, *Chemical Society Reviews* **2014**, *43*, 16 5700.

[30] N. Stock, S. Biswas, *Chemical reviews* **2012**, *112*, 2 933.

[31] S. Dissegna, K. Epp, W. R. Heinz, G. Kieslich, R. A. Fischer, *Advanced Materials* **2018**, *30*, 37 1704501.

[32] H. Gliemann, C. Wöll, *Materials today* **2012**, *15*, 3 110.

[33] J. Dechnik, J. Gascon, C. J. Doonan, C. Janiak, C. J. Sumby, *Angewandte Chemie International Edition* **2017**, *56*, 32 9292.

[34] M. Tsotsalas, A. Umemura, F. Kim, Y. Sakata, J. Reboul, S. Kitagawa, S. Furukawa, *Journal of Materials Chemistry* **2012**, *22*, 20 10159.

[35] X. Yin, C. E. Gounaris, *Computers & Chemical Engineering* **2022**, *167* 108022.

[36] H. Lyu, Z. Ji, S. Wuttke, O. M. Yaghi, *Chem* **2020**, *6*, 9 2219.

[37] S. Chong, S. Lee, B. Kim, J. Kim, *Coordination Chemistry Reviews* **2020**, *423* 213487.

[38] A. M. Cohen, W. R. Hersh, *Briefings in bioinformatics* **2005**, *6*, 1 57.

[39] M. A. Andrade, A. Valencia, In *Ismb*, volume 5. **1997** 25–32.

[40] P. Corbett, P. Murray-Rust, In *Computational Life Sciences II: Second International Symposium, CompLife 2006, Cambridge, UK, September 27-29, 2006. Proceedings 2*. Springer, **2006** 107–118.

[41] L. Hawizy, D. M. Jessop, N. Adams, P. Murray-Rust, *Journal of cheminformatics* **2011**, *3* 1.

[42] M. C. Swain, J. M. Cole, *Journal of chemical information and modeling* **2016**, *56*, 10 1894.

[43] J. Mavracic, C. J. Court, T. Isazawa, S. R. Elliott, J. M. Cole, *Journal of Chemical Information and Modeling* **2021**, *61*, 9 4280.

[44] E. Kim, K. Huang, A. Tomala, S. Matthews, E. Strubell, A. Saunders, A. McCallum, E. Olivetti, *Scientific data* **2017**, *4*, 1 1.

[45] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, *arXiv preprint arXiv:1810.04805* **2018**.

[46] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., *OpenAI blog* **2019**, *1*, 8 9.

[47] T. Gupta, M. Zaki, N. A. Krishnan, *npj Computational Materials* **2022**, *8*, 1 102.

[48] S. Huang, J. M. Cole, *Chemical Science* **2022**, *13*, 39 11487.

[49] A. Dunn, J. Dagdelen, N. Walker, S. Lee, A. S. Rosen, G. Ceder, K. Persson, A. Jain, *arXiv preprint arXiv:2212.05238* **2022**.

[50] L. Ghadbeigi, J. K. Harada, B. R. Lettiere, T. D. Sparks, *Energy & Environmental Science* **2015**, *8*, 6 1640.

[51] S. R. Young, A. Maksov, M. Ziatdinov, Y. Cao, M. Burch, J. Balachandran, L. Li, S. Somnath, R. M. Patton, S. V. Kalinin, et al., *Journal of Applied Physics* **2018**, *123*, 11 115303.

[52] F. Baum, T. Pretto, A. Koche, M. J. L. Santos, *The Journal of Physical Chemistry C* **2020**, *124*, 44 24298.

[53] 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*, 7601 73.

[54] J. R. McDaniel, J. R. Balmuth, *Journal of chemical information and computer sciences* **1992**, *32*, 4 373.

[55] M. Oldenhof, A. Arany, Y. Moreau, J. Simm, *Journal of chemical information and modeling* **2020**, *60*, 10 4506.

[56] K. Rajan, A. Zielesny, C. Steinbeck, *Journal of Cheminformatics* **2020**, *12*, 1 1.

[57] S. Yoo, O. Kwon, H. Lee, In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, **2022** 3393–3397.

[58] Y. Luo, S. Bag, O. Zaremba, A. Cierpka, J. Andreo, S. Wuttke, P. Friederich, M. Tsotsalas, *Angewandte Chemie International Edition* **2022**, *61*, 19 e202200242.

[59] A. Davariashtiyani, Z. Kadkhodaie, S. Kadkhodaei, *Communications Materials* **2021**, *2*, 1 115.

[60] Z. Jensen, E. Kim, S. Kwon, T. Z. Gani, Y. Román-Leshkov, M. Moliner, A. Corma, E. Olivetti, *ACS central science* **2019**, *5*, 5 892.

[61] C. Karpovich, E. Pan, Z. Jensen, E. Olivetti, *Chemistry of Materials* **2023**.

[62] O. Kononova, H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan, G. Ceder, *Scientific data* **2019**, *6*, 1 203.

[63] A. Nandy, G. Terrones, N. Arunachalam, C. Duan, D. W. Kastner, H. J. Kulik, *Scientific Data* **2022**, *9*, 1 74.

[64] C. J. Court, J. M. Cole, *npj Computational Materials* **2020**, *6*, 1 18.

[65] C. J. Court, A. Jain, J. M. Cole, *Chemistry of Materials* **2021**, *33*, 18 7217.

[66] C. J. Court, J. M. Cole, *Scientific data* **2018**, *5*, 1 1.

[67] Q. Dong, J. M. Cole, *Scientific Data* **2022**, *9*, 1 193.

[68] O. Sierepeklis, J. M. Cole, *Scientific Data* **2022**, *9*, 1 648.

[69] M. Zaki, N. A. Krishnan, et al., *Chemical Engineering and Processing-Process Intensification* **2022**, *180* 108607.

[70] Z. Nie, S. Zheng, Y. Liu, Z. Chen, S. Li, K. Lei, F. Pan, *Advanced Functional Materials* **2022**, *32*, 26 2201437.

[71] T. He, W. Sun, H. Huo, O. Kononova, Z. Rong, V. Tshitoyan, T. Botari, G. Ceder, *Chemistry of Materials* **2020**, *32*, 18 7861.

[72] H. Park, Y. Kang, W. Choe, J. Kim, *Journal of Chemical Information and Modeling* **2022**, *62*, 5 1190.

[73] S. H. M. Mehr, M. Craven, A. I. Leonov, G. Keenan, L. Cronin, *Science* **2020**, *370*, 6512 101.

[74] J. D. Evans, V. Bon, I. Senkovska, S. Kaskel, *Langmuir* **2021**, *37*, 14 4222.

[75] 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*, 7844 89.

[76] S. M. Moosavi, A. Chidambaram, L. Talirz, M. Haranczyk, K. C. Stylianou, B. Smit, *Nature communications* **2019**, *10*, 1 539.

[77] P. Chen, Z. Tang, Z. Zeng, X. Hu, L. Xiao, Y. Liu, X. Qian, C. Deng, R. Huang, J. Zhang, et al., *Matter* **2020**, *2*, 6 1651.

[78] L. Pilz, C. Natzeck, J. Wohlgemuth, N. Scheuermann, P. G. Weidler, I. Wagner, C. Wöll, M. Tsotsalas, *Advanced Materials Interfaces* **2022**, 2201771.

[79] M. H. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 7698 604.

[80] Y. J. Colón, R. Q. Snurr, *Chemical Society Reviews* **2014**, *43*, 16 5735.

[81] 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, *Chemistry of Materials* **2014**, *26*, 21 6185.

[82] 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, et al., *Journal of Chemical & Engineering Data* **2023**.

[83] Y. J. Colón, D. A. Gomez-Gualdron, R. Q. Snurr, *Crystal Growth & Design* **2017**, *17*, 11 5801.

[84] A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein, R. Q. Snurr, *Matter* **2021**, *4*, 5 1578.

[85] S. M. Moosavi, A. Nandy, K. M. Jablonka, D. Ongari, J. P. Janet, P. G. Boyd, Y. Lee, B. Smit, H. J. Kulik, *Nature communications* **2020**, *11*, 1 1.

Table 1: Table 1 caption

Description 1	Description 2	Description 3
Row 1, Col 1	Row 1, Col 2	Row 1, Col 3
Row 2, Col 1	Row 2, Col 2	Row 2, Col 3

[86] H. Daglar, H. C. Gulbalkan, G. Avci, G. O. Aksu, O. F. Altundal, C. Altintas, I. Erucar, S. Keskin, *Angewandte Chemie International Edition* **2021**, *60*, 14 7828.

[87] 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, *Chemistry of Materials* **2014**, *26*, 19 5632.

[88] A. Li, R. Bueno-Perez, D. Madden, D. Fairen-Jimenez, *Chemical Science* **2022**, *13*, 27 7990.

[89] F. T. Szczypiński, S. Bennett, K. E. Jelfs, *Chemical Science* **2021**, *12*, 3 830.

[90] S. Bennett, F. T. Szczypinski, L. Turcani, M. E. Briggs, R. L. Greenaway, K. E. Jelfs, *Journal of Chemical Information and Modeling* **2021**, *61*, 9 4342.

[91] 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, et al., *Matter* **2019**, *1*, 1 219.

[92] R. Anderson, D. A. Gómez-Gualdrón, *Chemistry of Materials* **2020**, *32*, 19 8106.

[93] R. Batra, C. Chen, T. G. Evans, K. S. Walton, R. Ramprasad, *Nature Machine Intelligence* **2020**, *2*, 11 704.

[94] A. Nandy, C. Duan, H. J. Kulik, *Journal of the American Chemical Society* **2021**, *143*, 42 17535.

References

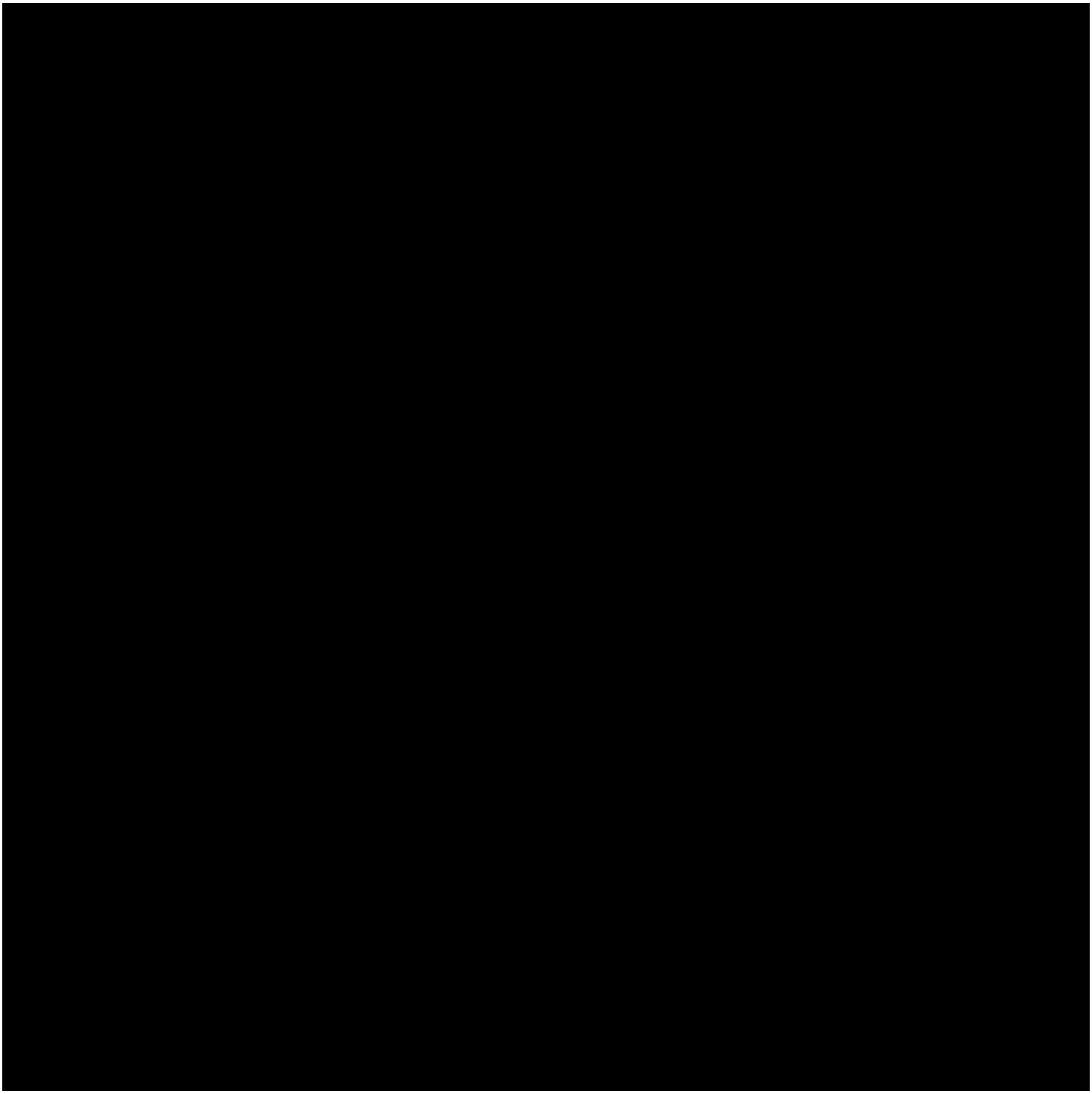

Figure 1: Figure 1 caption goes here. Reproduced with permission.[Ref.] Copyright Year, Publisher.

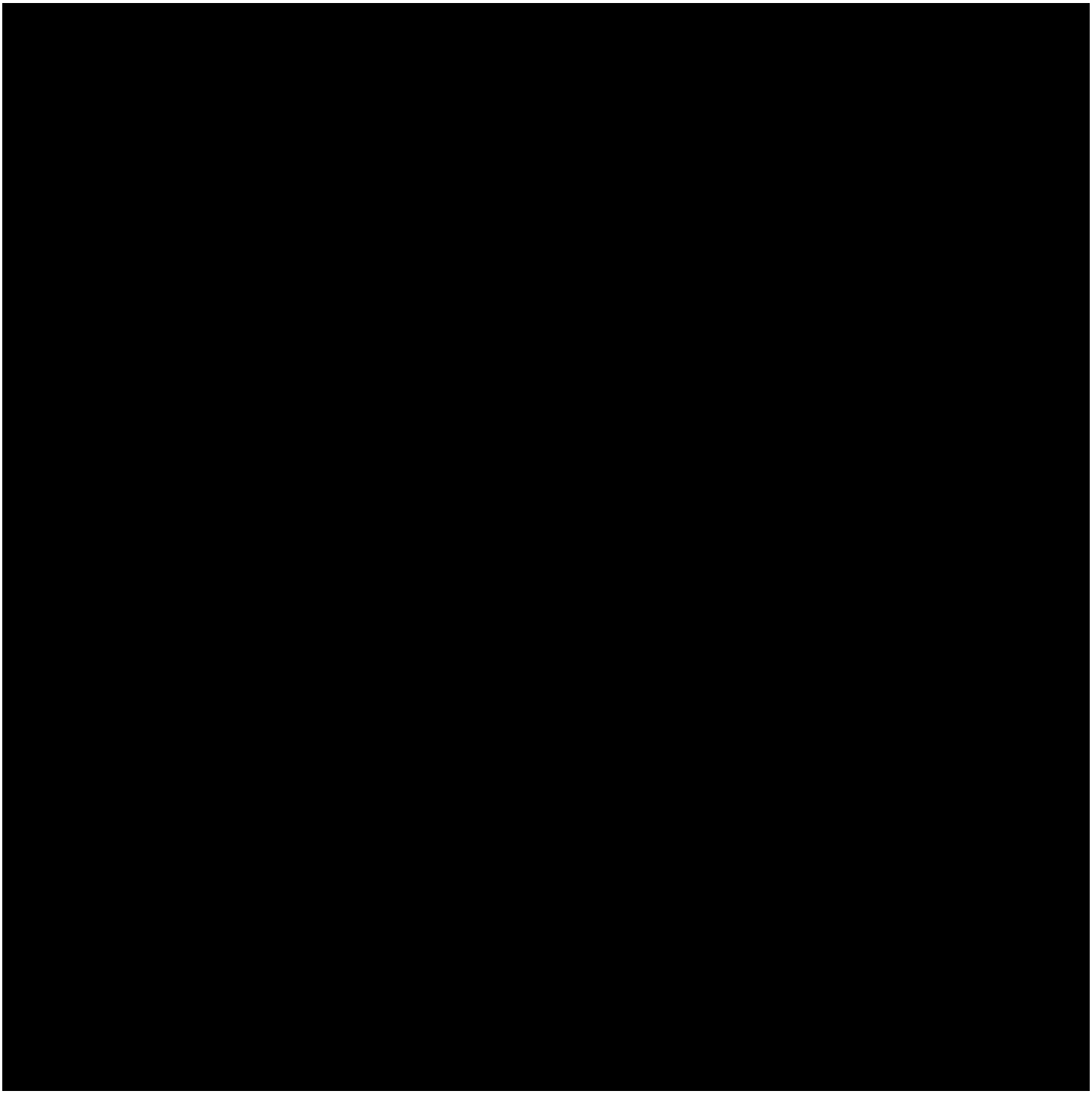

Figure 2: Figure 2 caption goes here. Reproduced with permission.[Ref.] Copyright Year, Publisher.

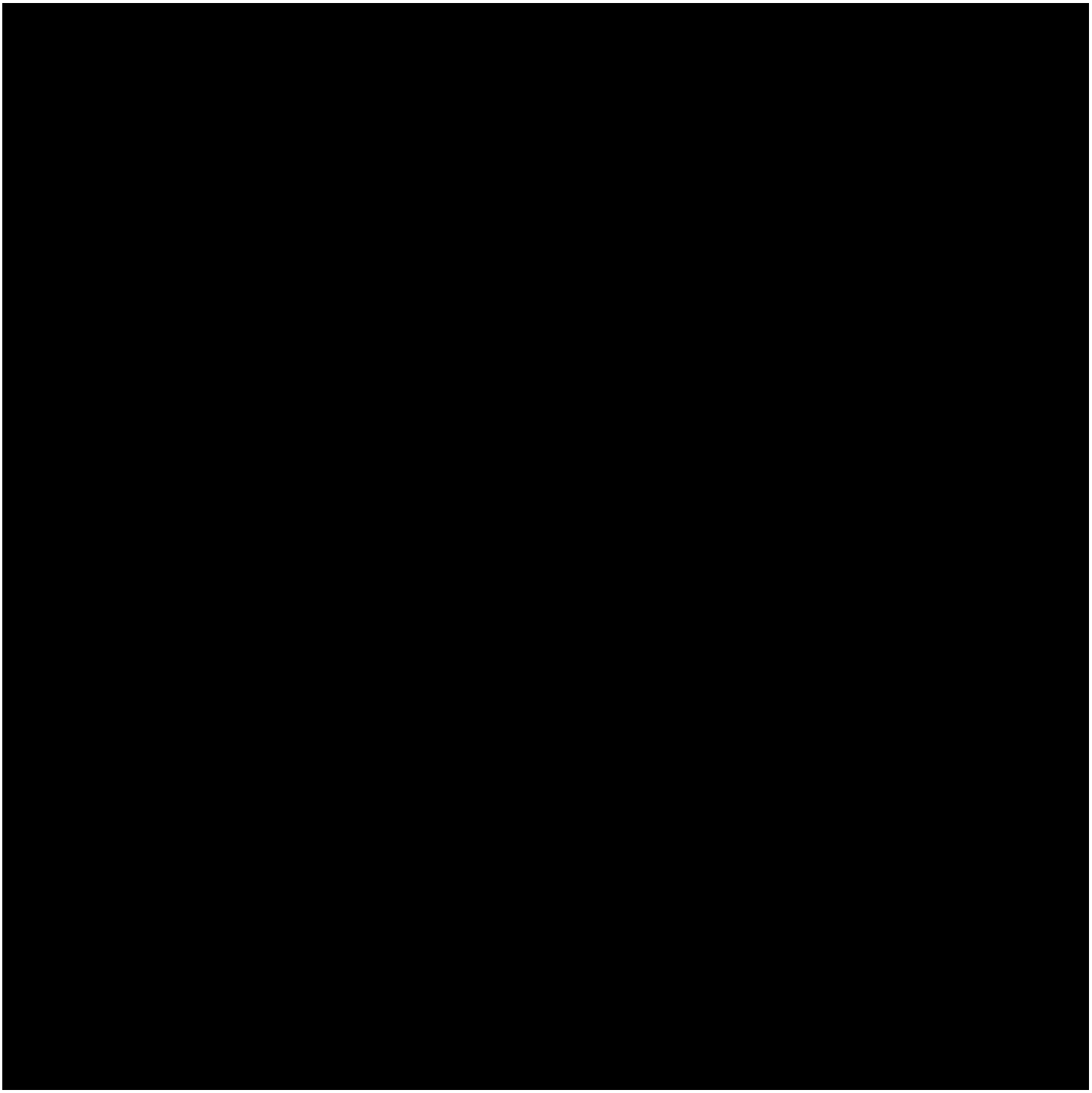

Figure 3: Figure 3 caption goes here. Reproduced with permission.[Ref.] Copyright Year, Publisher.

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. In 2019, he completed his habilitation at KIT. 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.

After his Ph.D. in physics under the supervision of Wolfgang Wenzel, **Pascal Friederich** 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 Pascal Friederich 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.

Table of Contents

ToC Entry