MOF Synthesis Prediction Enabled by Automatic Data Mining and Machine Learning

Yi Luo[1,†], Saientan Bag[2,†], Orysia Zaremba[3], Jacopo Andreo[3], Stefan Wuttke[3,4], Pascal Friederich[2,5,*], and Manuel Tsotsalas[1,6,*]

[1] Institute of Functional Interfaces, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

[2] Institute of Nanotechnology, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

[3] Basque Center for Materials, Applications & Nanostructures, Edif. Martina Casiano, Pl. 3 Parque Científico UPV/EHU Barrio Sarriena, 48940 Leioa, Bizkaia, Spain

[4] Ikerbasque, Basque Foundation for Science, Bilbao 48013, Spain

[5] Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Am Fasanengarten 5, 76131, Karlsruhe, Germany

[6] Institute of Organic Chemistry, Karlsruhe Institute of Technology, Kaiserstrasse 12, 76131 Karlsruhe, Germany

[†] These authors contributed equally

[*] Corresponding authors: pascal.friederich@kit.edu, and manuel.tsotsalas@kit.edu

Abstract

Despite rapid progress in the field of metal-organic frameworks (MOFs), the potential of using machine learning (ML) methods to predict MOF synthesis parameters is still untapped. Here, we show how ML can be used for rationalization and acceleration of the MOF discovery process by directly predicting the synthesis conditions of a MOF based on its crystal structure. Our approach is based on: (*i*) establishing the first MOF synthesis database *via* automatic extraction of synthesis parameters from the literature, (*ii*) training and optimizing ML models by employing the MOF database, and (*iii*) predicting the synthesis conditions for new MOF structures. The ML models even at an initial stage exhibit a good prediction performance, outperforming human expert predictions, obtained through a synthesis survey.

Main

Metal-organic framework (MOF) chemistry has flourished through the creation of a vast chemical space where more than 100,000 MOFs have been discovered.[1] The number is increasing rapidly with a wide and continuously expanding variety of structural types, building units, linkage chemistry, and functional groups.[2–5] In fact, the chemical space of possible MOF structures is so huge that it is impossible to fully explore it experimentally.[6–8] Simulation and machine learning (ML) have evolved as important tools for guiding researchers to computationally identify regions of interest.[6,9–11] However, in order to synthesize the novel MOF structures, the researchers still have to rely on their experience, employing a trial-and-error approach (**Fig. 1**). This is a very challenging process that is highly time-consuming, labor-intensive and requires a lot of resources. Therefore, the search for an efficient way to find the optimal MOF synthesis conditions represents the current bottleneck in speeding up MOF exploration.

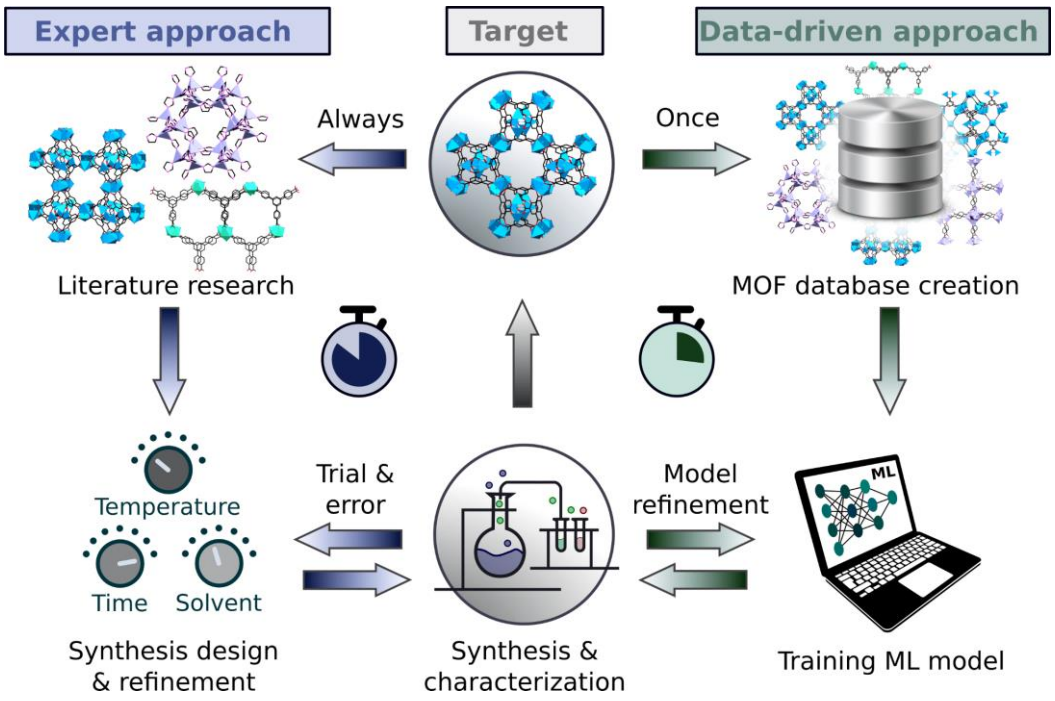

Figure 1. A new approach to MOF synthesis. *The conventional approach (left loop) of new MOF synthesis is based on a time-consuming trial-and-error approach, in which a target MOF structure is*

compared with reported MOFs from literature to find similar synthesis conditions and experimentally refine them. A data driven approach (right loop), where a ML model is trained on a library of automatically extracted literature data, to then suggest synthesis conditions in a data-driven MOF discovery cycle. Updating the ML model based on new experiments leads to continuous improvement of the predictions.

The development of ML methods to predict the synthesis parameters for a desired MOF crystal structure based on scientific literature is a challenging but promising approach that will advance and accelerate chemical synthesis. Over the last years, ML methods have rapidly evolved, solving complex problems that involve highly nonlinear or massively combinatorial processes that conventional approaches fail to answer.[12] Up till now, ML approaches have been successfully applied to address challenges in organic and inorganic synthesis.[9,13–18] In the case of MOF synthesis, only recently, ML was used to optimize synthesis parameters for HKUST-1 and to determine the importance of the different parameters by analysing a set of partially failed experiments, in other words, "capture the chemical intuition" that can help to speed up the synthesis of similar MOF systems.[19] However, the inverse synthesis design of MOFs, *i.e.* the automated prediction of suitable synthesis conditions for a targeted MOF structure (*e.g.* designed in silico) remains an unsolved challenge.

This work represents a first step towards predicting synthesis conditions for an arbitrary MOF. We show a complete ML workflow for the inverse synthesis design of MOFs (going from crystal structure to synthesis conditions), (1) starting from automated data mining from scientific literature on MOF synthesis conditions and their structural information, (2) setting up and training of ML models, and (3) prediction of synthesis conditions for new MOF structures and comparison with human experts' predictions.

Our approach marks the starting point for the transition from a trial-and-error approach that is based on experience and heuristics, towards an inverse synthesis design approach in

the MOF synthesis, ultimately enabling fully autonomous MOF discovery in automated labs.[20]

Results and discussion

Data mining from MOF literature. To create a dataset with MOF synthesis parameters and structural information, we took advantage of the fact that well-curated MOF structural databases already exist (*e.g.* the Computation-Ready Experimental Metal–Organic Framework database CoREMOF[21] and the Cambridge Structural Database CSD), in which MOF structural information and the corresponding publications with successful synthesis protocols are stored.[22] The manual extraction of synthesis procedures from scientific literature is a time-consuming task, requiring the work of many experts. Alternatively, automatic data extraction to convert experimental procedures into a set of the desired synthesis parameters by employing natural language processing (NLP) techniques is a highly efficient and promising approach that we expect to be continuously improved in the upcoming years.[23,24]

In this study, we automatically extracted information on MOF synthesis for all publicly available MOF structures in the CoRE MOF database (**SI Section 2.1**). The six relevant parameters that were extracted are metal source, linker(s), solvent(s), additive, synthesis time, and temperature (**Fig. 2**). To achieve this, we initially classified literature paragraphs, employing a decision tree with a string search method, to identify the synthesis paragraph related to each MOF structure (**SI Section 2.2**). After the synthesis paragraphs were determined, we employed the ChemicalTagger software, which focuses on the experimental part of a scientific text, recognizing significant words within the sentences, and annotating phrases inside the paragraph.[25] In an effort to increase the tagging accuracy, we slightly modified the synthesis paragraphs, accounting for MOF-domain specific descriptions (**SI Section 2.3**). To evaluate the accuracy of the automatically extracted SynMOF-A database,

we additionally generated manually corrected versions - the SynMOF-M and SynMOF-ME databases that are discussed in **SI Section 2.4**.

Alongside retrieving synthesis information from the MOF literature, we used the crystallographic information files (CIFs) from MOF databases to automatically extract the structural information of the linker and the oxidation state of the metal center.[26] Ultimately, we combined the extracted synthesis details (*i.e.* metal source, linkers, temperature, synthesis time, solvents and additives) from the publications and information of the linker and the metal source from the CIF into the SynMOF database (**Fig. 2**). Our central assumption in this work is that the established SynMOF database can be used for the training of ML models to facilitate the discovery of similarity patterns in the synthesis conditions to reach the final goal of predicting synthesis protocols for new MOF structures.

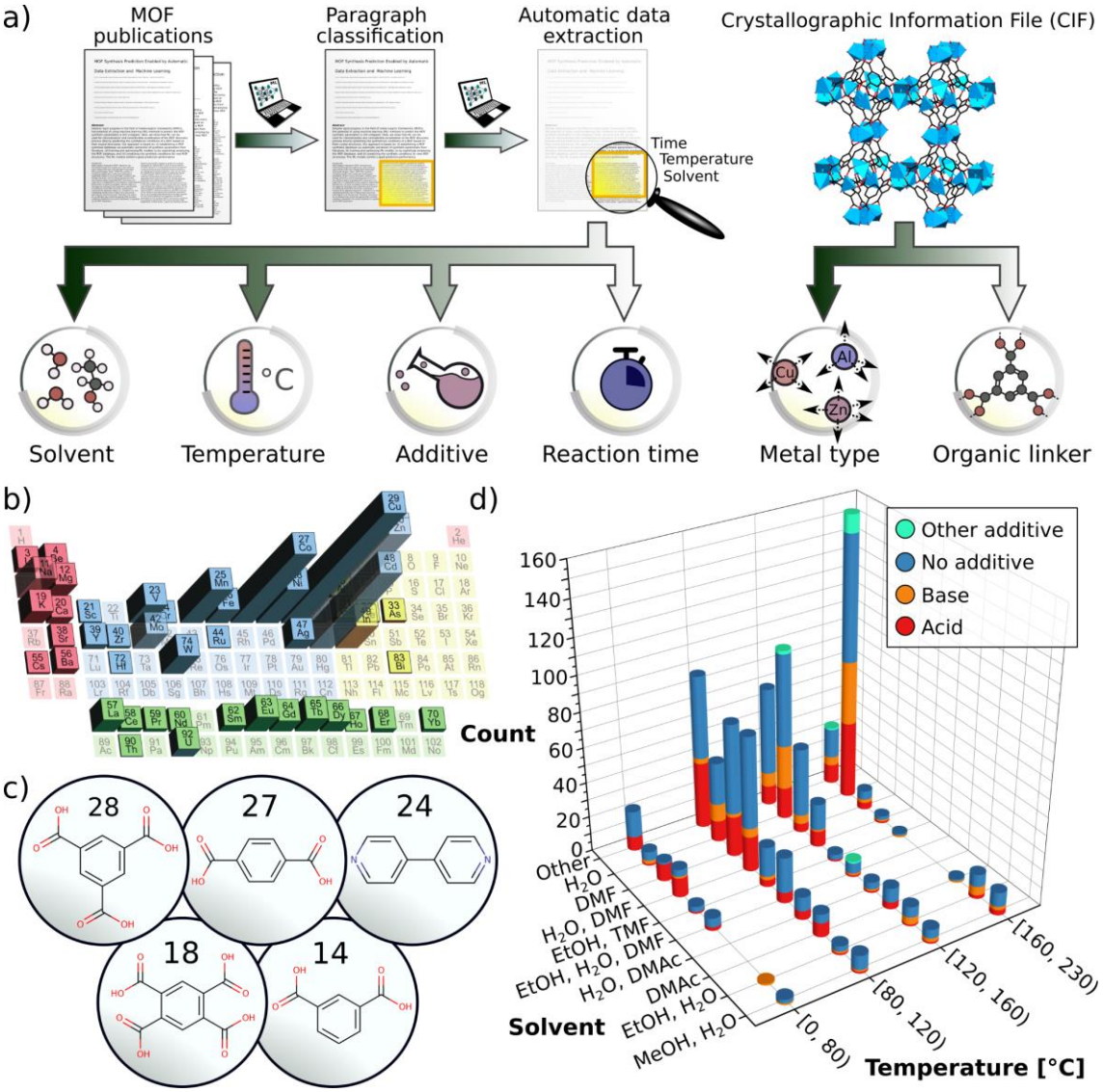

Figure 2. SynMOF database. *a) Data mining pipeline and content of the SynMOF database; b) the statistics on the most common metal source and c) structures and occurrences of the most common linkers in the SynMOF database; d) 3D graph exhibiting correlation between solvent type, additive and temperature.*

Apart from the detailed information on MOF synthesis conditions, our SynMOF database, currently consisting of 983 MOF structures, provides the statistical data on the metal source and organic components (**Figs. 2b and 2c**). It contains 46 different metals with most common oxidation states ranging from +1 to +3. As expected, most MOF structures are composed of transition metals with copper and zinc comprising almost 50% of all metal

types. Among the diverse organic molecules, the most commonly employed linkers for MOF synthesis are multidentate carboxylic acids (*i.e.* benzene-1,3,5-tricarboxylic acid, benzene-1,4-dicarboxylic acid, and benzene-1,2,4,5-tetracarboxylic acid) followed by N-containing bases (*i.e.* pyridine, triazole, and tetrazole).

In search of obvious patterns, we analysed the most common solvents used during MOF synthesis with respect to different temperature regimes and additives (**Fig. 2d**). At temperatures ranging from 80 °C to 160 °C, DMF and water, as well as their mixtures with other solvents are the most commonly used solvents. Synthesis at temperatures above 160 °C is predominantly carried out in water as a single solvent. Besides, the majority of MOF synthesis reactions at high temperatures (above 120 °C) are performed without additives, while at temperatures below 80 °C, the addition of acidic additives dominates. Beyond such relatively simple patterns, we expected more correlations to be hidden in the data (**SI Section 2.5**), which we exploit using ML approaches.

Machine learning training, prediction and evaluation. Employing the data stored in the SynMOF database, we trained multiple ML models to predict synthesis conditions of a diverse set of MOFs unseen during training. The input representation of the MOF structures is of crucial importance for the ML models performance.[27] In this study, we used two types of representations as an input for the ML models training: One based on molecular fingerprints of the linkers, extended with encodings of the metal type and its oxidation state (**Fig. 3a, SI Section 3.1**), and the recently developed MOF representation by Kulik and co-workers (**SI Section 3.2**).[28] It is to be noted that the MOF field is still expanding, and an increasing amount of new structures and corresponding synthesis parameters will be available over time that can be used for training and refinement of ML models to achieve the highest possible performance. In this case, representation learning methods such as graph neural networks will then likely become more accurate than models relying on hand-crafted feature representations.[29–31]

The prediction of synthesis time and temperature was achieved *via* regression models, such as random forests or neural networks (**SI Sections 3.3, 3.4, 3.5**). To predict discrete synthesis parameters, such as solvent and additives, classification models could be, in principle, used. However, for multiple reasons this turns out to be impractical: There is a wide variety of possible solvents and additives reported in literature, leading to a large number of categories, and, in turn, strongly imbalanced datasets. Furthermore, the properties of solvents can be very similar, making them interchangeable in synthesis, which leads to ambiguous solutions. In practice, also combinations of various solvents are required for successful MOF synthesis. Therefore, we developed a ML model which predicts solvent properties, such as partition coefficients, boiling point (**SI Section 3.6**), rather than the specific solvent. A nearest neighbor search in solvent property space yields lists of possible solvents that have properties similar to those predicted by the ML model. In this way, new solvents can be incorporated easily, and even solvents occurring only once in literature can be used to train the model. In the case of additives, we found that the main parameter that distinguishes different additives is their acidity/basicity strength. Thus, we split the dataset into three groups (acidic, basic or no additive) and used a classification model for additive prediction.

The results of our trained ML models are shown in **Fig. 3b-f**. Reproducibly positive correlation coefficients r^2 on unseen test datasets show that the ML models are capable of identifying meaningful and predictive relations between the target MOF structure and the required synthesis conditions, in particular temperature and time (**Figs. 3b, 3c**). Given the amount of data that we have currently extracted from literature, we find that the random forest models have the highest performance across all predicted parameters. However, neural networks learn to make better predictions with growing dataset sizes faster (see learning curves in **Fig. 3d**) and even exploit correlations between different synthesis parameters (*e.g.,* solvent and temperature) rather than predicting them separately. Hence, we expect that more complex models will outperform random forests in the near future.

To evaluate ML-based solvent prediction, we focused on a subset of MOFs which are synthesized using only one solvent. We compared the accuracy of the top 6 ML predictions with multiple random baseline methods (**Fig. 3e**), including selection of a random solvent out of all solvents as well as out of the six most frequent solvents that are used in 96% of the single-solvent SynMOF database. We found that the ML model outperforms the random selection, in particular for the top 1 - 3 solvent predictions, where the ML model reaches an accuracy of > 90%. In the case of additive predictions (**Fig. 3f**), the task of the ML model is to classify required additives as acidic, basic, and no additive. While performing well on the training set, the generalization to unseen test data suffers from an imbalanced dataset (most database entries do not use an additive). We use balance correcting weights of the training data points, leading to predictions which distinguish very well between synthesis procedures involving basic and acidic additives. However, the differentiation between acidic and no additive or basic and no additive is less pronounced. One of the reasons might be related to the hidden variables such as type and function of additives: Some of them (inorganic acids and bases) have only the role of pH regulation, while others (organic acids and bases) are also involved in modulation of the MOF growth. Besides, concentration and strength of additives are additional important parameters, influencing the role of additive. A larger amount of training data in the future will enable refinement of the additive representation and improvement of our ML model, thus opening new prospects in synthesis condition prediction.

To put the ML performance into perspective, we performed tests with 11 human MOF synthesis experts. We developed an online quiz based on 50 MOFs randomly selected from the SynMOF database which will be publicly available. The participants were given the 3D structures of MOFs, chemical structures of the linkers and information on the metal ion, and asked to estimate synthesis conditions such as temperature, time, solvents and additives without any help from literature or other external sources (**SI Section 3.7**). After each MOF synthesis prediction, we also asked the participants to estimate how certain they are in the

answer. The correlation coefficients r^2 between the experts' temperature and time predictions and the reported synthesis conditions are close to zero, even after averaging over 11 estimates by different researchers (**Fig. 3g**) and after sorting only by predictions with high certainty. This rather surprising result shows that even small correlations learned and exploited by the ML model will help to estimate better synthesis conditions. In summary, we showed that the ML models are able to learn generalized patterns and correlations in the SynMOF database, which exceed the experts' general intuition, and thus, could be used to identify good first guesses for experimental synthesis attempts of new MOFs.

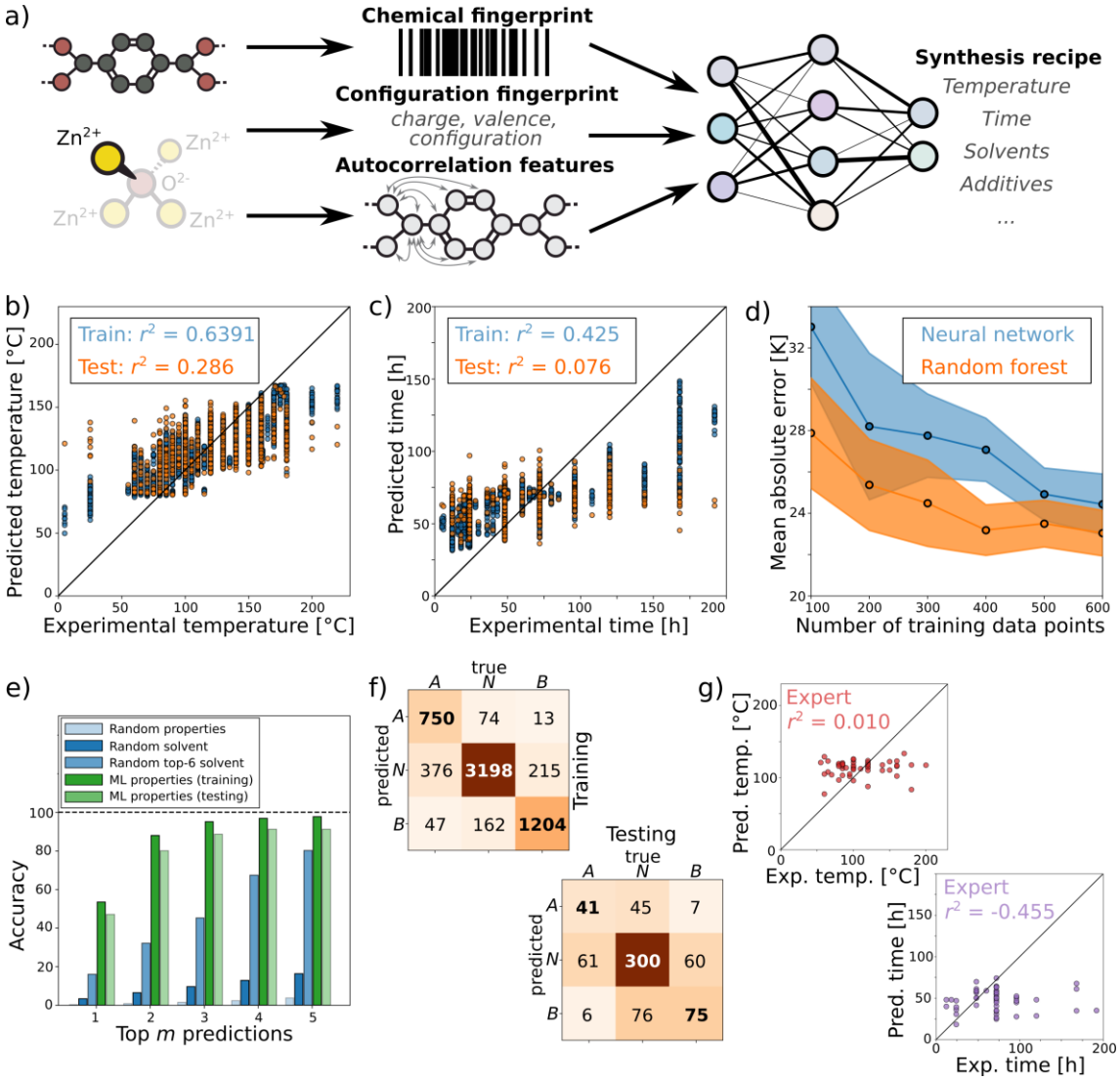

Figure 3. Machine learning models trained on the SynMOF-A database. *a) ML workflow, including fingerprint representation of the linkers and the feature representation of the metal type and oxidation state; b) and c) comparison of ML predictions of temperature and time for training and test sets with the initial data extracted from literature; d) learning curve of temperature predictions, i.e. mean absolute error as a function of the training set size, for neural network and random forest regression models; e) ML solvent prediction accuracy for a subset of single-solvent MOFs, compared to different methods of random predictions; f) training and test set performance of additive classification where A, B and N correspond to acid, base, and no additive respectively and g) average of eleven human expert predictions of temperature and time for 50 MOFs to evaluate the complexity of the problem.*

Conclusions

The lack of machine readable and curated MOF-synthesis data up till now hindered the development of digital ML tools for predicting MOF synthesis conditions. Here, we established a SynMOF database by automatic data extraction via NLP methods that provides synthesis conditions and structural information for more than 900 MOFs, and trained ML models based on these data to identify patterns in MOF synthesis. We expect that the created SynMOF database will boost the NLP research within the MOF community, while our ML synthesis prediction platform will be the new gold standard for data-driven MOF discovery. Even at an initial stage, our ML models outperformed MOF experts' synthesis prediction, underlying both the complexity behind the synthesis process and a pressing need in developing digital predictive tools. Our automated on-demand synthesis prediction will considerably accelerate the discovery of new MOFs and serve a valuable tool for the MOF community and beyond.

Methods

Data Mining. We extracted the synthesis conditions from MOF publications using different NLP techniques. To select synthesis paragraphs, we developed a decision tree algorithm based on a keyword list selected from 100 MOF synthesis papers. To analyse the synthesis paragraph and identify information about chemical entities, experimental steps, and corresponding conditions associated with those steps, we applied the ChemicalTagger software. When precursors, solvents and additives, as well as solvothermal synthesis conditions were extracted, we compared the metal element from the automatically formed synthesis protocol to the CoRE MOF database to eliminate mismatched conditions. The results of this fully automated data extraction are collected in the SynMOF-A database.

Machine Learning. We developed a code to extract the MOF linker from the CIF. The RDKit library was further used to evaluate the molecular fingerprint of the extracted linker. The

MOF metal nodes were represented by their full electronic configuration. The molecular fingerprint of the linker and the full electronic configuration of the metal node, accounting for its oxidation state, were combined to form the input of the ML model. This input representation was compared to the MOF representation developed by Kulik and co-workers, relying on autocorrelation features of the metal cores and the linkers. The output of the ML model was the MOF synthesis conditions, namely temperature, synthesis time, solvent properties and additive type. Depending on the specific synthesis conditions, we evaluated several regression models, in particular random forest regression and neural networks. The scikit-learn library in Python was used for the implementation of the ML models. 70% of the full dataset was used to train the ML model, while the remaining data was used to test the model. In the case of solvent property prediction, we limited the data to MOFs with single-solvent synthesis. To quantify the accuracy of the trained ML model, we calculated the mean absolute error and the correlation coefficient r^2 of the training and test dataset for the regression tasks. The accuracy of the ML model for the classification tasks were quantified by calculating the normalized confusion matrix.

Acknowledgements

We thank Dr. Christian Diercks (Scripps Research Institute), Dr. Julien Reboul (Sorbonne Université), Dr. Roberto Fernández de Luis (BCMaterials), Dr. João Marreiros (KU Leuven), Dr. Stéphane Diring (Nantes University), Dr. Akira Hinokimoto, Dr. Eli Sanchez Gonzalez, Dr. Javier Troyano (Kyoto University) and Dr. Romy Ettlinger (University of Augsburg) for participation as experts in the prediction of MOF synthesis conditions. Y.L. thanks Dr. Cam An Nguyen Thanh for java programming support. S.W. acknowledges funding from the Basque Government Industry Department under the ELKARTEK and HAZITEK programs. M.T. acknowledges funding from the Helmholtz Association's Initiative and Networking

Fund (Grant VH-NG-1147). Y.L. acknowledges funding from the China Scholarship Council for the financial support (No. 201706270179).

Author contributions

M.T. and P.F. designed the study, Y.L. developed the literature extraction method, S.B. trained the machine learning models, S.W., J.A. and O.Z. worked on the expert survey. All authors contributed in discussing the results and writing the manuscript.

Competing interests

The authors declare no competing interest.

Supporting Information

Supporting Information is available free of charge online.

Data and code availability

The databases SynMOF-A, SynMOF-M and SynMOF-ME, the codes for the synthesis parameter extraction, for ML training and prediction, and the expert survey are available free of charge on https://github.com/Tsotsalas-Group/MOF_Literature_Extraction and https://github.com/aimat-lab/MOF_Synthesis_Prediction.

References

1. Freund, R. *et al.* The Current Status of MOF and COF Applications. *Angew. Chem. Int. Ed.*

60, 2–29 (2021).

2. Férey, G. Hybrid porous solids: past, present, future. *Chem. Soc. Rev.* **37**, 191–214 (2007).
3. Furukawa, H., Cordova, K. E., O'Keeffe, M. & Yaghi, O. M. The Chemistry and Applications of Metal-Organic Frameworks. *Science* **341**, (2013).
4. Kitagawa, S., Kitaura, R. & Noro, S. Functional Porous Coordination Polymers. *Angew. Chem. Int. Ed.* **43**, 2334–2375 (2004).
5. Gropp, C. *et al.* Standard Practices of Reticular Chemistry. *ACS Cent. Sci.* **6**, 1255–1273 (2020).
6. Jablonka, K. M., Ongari, D., Moosavi, S. M. & Smit, B. Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chem. Rev.* **120**, 8066–8129 (2020).
7. Lyu, H., Ji, Z., Wuttke, S. & Yaghi, O. M. Digital Reticular Chemistry. *Chem* **6**, 2219–2241 (2020).
8. Luo, Y., Ahmad, M., Schug, A. & Tsotsalas, M. Rising Up: Hierarchical Metal–Organic Frameworks in Experiments and Simulations. *Adv. Mater.* **31**, 1901744 (2019).
9. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
10. Gromski, P. S., Henson, A. B., Granda, J. M. & Cronin, L. How to explore chemical space using algorithms and automation. *Nat. Rev. Chem.* **3**, 119–128 (2019).
11. Ahmad, M., Luo, Y., Wöll, C., Tsotsalas, M. & Schug, A. Design of Metal-Organic Framework Templated Materials Using High-Throughput Computational Screening. *Molecules* **25**, 4875 (2020).
12. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *Npj Comput. Mater.* **3**, 54 (2017).
13. Jensen, Z. *et al.* A Machine Learning Approach to Zeolite Synthesis Enabled by Automatic Literature Data Extraction. *ACS Cent. Sci.* **5**, 892–899 (2019).
14. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for

molecular and materials science. *Nature* **559**, 547–555 (2018).

15. Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
16. Jensen, Z. *et al.* Discovering Relationships between OSDAs and Zeolites through Data Mining and Generative Neural Networks. *ACS Cent. Sci.* **7**, 858–867 (2021).
17. Schwaller, P. *et al.* Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
18. Reid, J. P. & Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **571**, 343–348 (2019).
19. Moosavi, S. M. *et al.* Capturing chemical intuition in synthesis of metal-organic frameworks. *Nat. Commun.* **10**, 539 (2019).
20. Tabor, D. P. *et al.* Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **3**, 5–20 (2018).
21. Chung, Y. G. *et al.* Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data* **64**, 5985–5998 (2019).
22. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **72**, 171–179 (2016).
23. Olivetti, E. A. *et al.* Data-driven materials research enabled by natural language processing and information extraction. *Appl. Phys. Rev.* **7**, 041317 (2020).
24. Kim, E. *et al.* Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chem. Mater.* **29**, 9436–9444 (2017).
25. Hawizy, L., Jessop, D. M., Adams, N. & Murray-Rust, P. ChemicalTagger: A tool for semantic text-mining in chemistry. *J. Cheminf.* **3**, 17 (2011).
26. Jablonka, K. M., Ongari, D., Moosavi, S. M. & Smit, B. Using collective knowledge to assign oxidation states of metal cations in metal–organic frameworks. *Nat. Chem.* 1–7 (2021)
27. von Lilienfeld, O. A. & Burke, K. Retrospective on a decade of machine learning for

chemical discovery. *Nat. Commun.* **11**, 4895 (2020).

28. Moosavi, S. M. *et al.* Understanding the diversity of the metal-organic framework ecosystem. *Nat. Commun.* **11**, 4068 (2020).
29. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural Message Passing for Quantum Chemistry. *ArXiv170401212 Cs* (2017).
30. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. SchNet – A deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
31. Friederich, P., Häse, F., Proppe, J. & Aspuru-Guzik, A. Machine-learned potentials for next-generation matter simulations. *Nat. Mater.* **20**, 750–761 (2021).