MOFs und Maschinelles Lernen

Zitierweise: *Angew. Chem. Int. Ed.* **2022,** *61,* e202200242
Internationale Ausgabe: doi.org/10.1002/anie.202200242
Deutsche Ausgabe: doi.org/10.1002/ange.202200242

Vorhersage der MOF-Synthese durch automatisches Data-Mining und maschinelles Lernen**

Yi Luo[+], Saientan Bag[+], Orysia Zaremba, Adrian Cierpka, Jacopo Andreo, Stefan Wuttke, Pascal Friederich, und Manuel Tsotsalas**

Abstract: Trotz großer Fortschritte auf dem Gebiet der metallorganischen Gerüststrukturen (MOF) ist das volle Potential des Maschinellen Lernens (ML) für die Vorhersage von MOF-Syntheseparametern bisher noch nicht erschlossen. In diesem Beitrag wird dargestellt, wie Methoden des ML für die Rationalisierung und Beschleunigung von MOF-Entwicklungsverfahren eingesetzt werden können, indem die Synthesebedingungen der MOFs direkt anhand ihrer Kristallstruktur vorhergesagt werden. Unser Ansatz stützt sich auf: i) die Erstellung der ersten MOF-Synthese-Datenbank durch automatische Extraktion der Syntheseparameter aus der Fachliteratur, ii) das Trainieren und die Optimierung von ML-Modellen mit Daten der MOF-Datenbank und iii) die ML basierte Vorhersage der Synthesebedingungen neuer MOF-Strukturen. Schon jetzt übertreffen die Ergebnisse der Vorhersagemodelle die Vorhersagen menschlicher ExpertInnen, welche in einer Befragung ermittelt wurden. Die automatisierte Synthesevorhersage ist über ein Web-Tool unter https://mof-synthesis.aimat.science verfügbar.

Die chemische Raum der metallorganischen Gerüststrukturen (MOF) hat sich durch die Entdeckung von mehr als 100.000 MOFs rapide vergrößert[1] und enthält eine zunehmende Vielzahl von Strukturtypen, Bausteinen, Kopplungen und funktionellen Gruppen.[2] Der gesamte chemische Raum umfasst mehrere Millionen möglicher unterschiedlicher Strukturen, deren vollständige experimentelle Überprüfung aufgrund der großen Anzahl ausgeschlossen ist.[3] Simulationen und Maschinelles Lernen (ML) haben sich zu entscheidenden Werkzeugen entwickelt, mit deren Hilfe Forschende interessante Bereiche des hypothetischen chemischen Raums identifizieren können.[3a,4] Bei der Synthese neuer MOFs sind die Forschenden allerdings immer noch auf ein Ausprobieren angewiesen. Sie können sich dabei nur auf ihre Erfahrung verlassen (siehe Abbildung 1). Dieses komplizierte Verfahren ist zeit-, arbeits- und ressourcenaufwendig und stellt momentan einen Engpass bei der Entwicklung von Methoden zur Erforschung metallorganischer Gerüststrukturen dar. Es muss daher ein effizienterer Weg zur Untersuchung der optimalen MOF-Synthesebedingungen gefunden werden.

Die Entwicklung von ML–Methoden zur Vorhersage der Syntheseparameter für eine gewünschte MOF-Kristallstruktur basierend auf wissenschaftlicher Fachliteratur ist ein anspruchsvoller aber vielversprechender Ansatz, der die chemische Synthese voranbringen und beschleunigen wird. Im Laufe der letzten Jahre haben sich ML–Methoden kontinuierlich weiter entwickelt. Mit ihrer Hilfe werden komplexe Probleme gelöst, die aufgrund hochgradig nichtlinearer oder kombinatorischer Prozesse nicht mit konventionellen Verfahren lösbar sind.[5] ML-Ansätze wurden bereits erfolgreich bei der organischen und anorganischen Synthese eingesetzt.[4a,6] Bei der MOF-Synthese wurde ML kürzlich dazu verwendet, die Syntheseparameter für HKUST-1 zu optimieren und die Bedeutung der verschiedenen Parameter durch die Analyse einer Reihe teilweise fehlgeschlagener

[*] Y. Luo,[+] Dr. M. Tsotsalas
Institute of Functional Interfaces, Karlsruhe Institute of Technology
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen (Deutschland)
E-mail: manuel.tsotsalas@kit.edu

Dr. S. Bag,[+] Jun.-Prof. Dr. P. Friederich
Institute of Nanotechnology, Karlsruhe Institute of Technology
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen (Deutschland)
E-mail: pascal.friederich@kit.edu

O. Zaremba, Dr. J. Andreo, Prof. S. Wuttke
Basque Center for Materials, Applications & Nanostructures
Edif. Martina Casiano, Pl. 3 Parque Científico UPV/EHU Barrio Sarriena, 48940 Leioa, Bizkaia (Spanien)

A. Cierpka, Jun.-Prof. Dr. P. Friederich
Institute of Theoretical Informatics, Karlsruhe Institute of Technology
Am Fasanengarten 5, 76131 Karlsruhe (Deutschland)

Prof. S. Wuttke
Ikerbasque, Basque Foundation for Science
Bilbao 48013 (Spanien)

Dr. M. Tsotsalas
Institute of Organic Chemistry, Karlsruhe Institute of Technology
Kaiserstrasse 12, 76131 Karlsruhe (Deutschland)

[+] Diese Autoren haben zu gleichen Teilen zu der Arbeit beigetragen.

[**] Eine frühere Version dieses Manuskripts ist auf einem Preprint-Server hinterlegt worden (https://doi.org/10.33774/chemrxiv-2021-kgd0h).

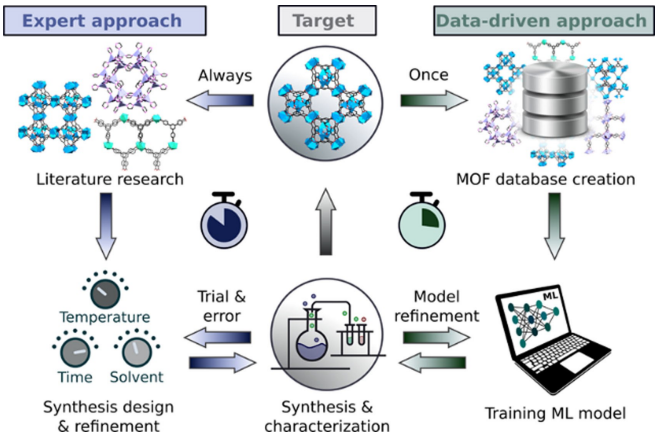

Abbildung 1. Ein neuer Ansatz in der MOF-Synthese. Der konventionelle Ansatz (linker Kreislauf) für die Synthese neuer MOFs basiert auf zeitaufwendigem Ausprobieren, d. h. dem Vergleichen einer MOF-Zielstruktur mit MOFs aus der Literatur, dem Auffinden ähnlicher Synthesebedingungen und der experimentellen Weiterentwicklung dieser Synthesebedingungen. Im datenbasierten Ansatz (rechter Kreislauf) wird ein ML-Modell mit einer Sammlung von automatisch aus der Literatur extrahierten Daten trainiert, so dass es dann in einem datenbasierten MOF-Entwicklungszyklus Synthesebedingungen vorschlagen kann. Die Aktualisierung des Modells durch neue Experimente führt zu einer kontinuierlichen Verbesserung der Vorhersagen.

Experimente zu bestimmen. Mit anderen Worten, mit ML wurde "die chemische Intuition erfasst", um die Synthese ähnlicher MOF-Systeme zu beschleunigen.[7] Die Synthese von MOFs in umgekehrter Richtung, d. h. die automatische Vorhersage geeigneter Synthesebedingungen für eine MOF-Zielstruktur (die z. B. in silico entwickelt wurde), bleibt jedoch eine ungelöste Herausforderung.

Diese Arbeit stellt einen ersten Schritt in Richtung der Vorhersage der Synthesebedingungen beliebiger MOFs dar. Im Folgenden wird der vollständige ML-Workflow für die inverse Syntheseentwicklung von MOFs dargestellt (von der Kristallstruktur bis hin zu den Synthesebedingungen): 1) automatisiertes Data-Mining in der wissenschaftlichen Fachliteratur zu MOF-Synthesebedingungen und deren Strukturinformationen, 2) Erstellung und Training der ML-Modelle und 3) Vorhersage der Synthesebedingungen für neue MOF-Strukturen sowie Vergleich mit Vorhersagen menschlicher ExpertInnen. Diese Arbeit erforscht den Übergang vom alten Trial-and-Error-Ansatz basierend auf Erfahrung und Heuristik hin zum neuen Ansatz der inversen Syntheseentwicklung von MOFs, der schlussendlich eine vollständig autonome MOF-Entwicklung in automatisierten Laboren ermöglicht.[8]

Zur Erstellung eines Datensatzes mit MOF-Syntheseparametern und Strukturinformationen konnte auf gut gepflegte Datenbanken mit MOF-Strukturdaten zurückgegriffen werden (z. B. die Computation-Ready Experimental Metal-Organic Framework Database CoRE MOF[9] und die Cambridge Structural Database CSD[10]), in denen MOF-Strukturdaten und die dazugehörigen Publikationen mit Protokollen der erfolgreichen Synthese gespeichert sind. Die manuelle Extraktion von Syntheseverfahren aus wissenschaftlicher Fachliteratur erfordert einen immensen Zeit- und Arbeitsaufwand. Im Gegensatz dazu nutzt die automatische Datenextraktion Techniken aus dem Natural Language Processing (NLP), um Versuchsdurchführungen in Datensätze mit den erwünschten Syntheseparametern zu konvertieren und ist somit eine hocheffiziente und vielversprechende Alternative, von der zu erwarten ist, dass sie in den nächsten Jahren weiter verbessert wird.[11]

In dieser Arbeit wurde ein automatisches Verfahren zur Extraktion von Informationen über die MOF-Synthese aller öffentlich zugänglichen MOF-Strukturen in der CoRE-MOF-Datenbank entwickelt (Hintergrundinformationen Section 2.1). Es wurden sechs relevante Parameter extrahiert: Metallzentrum (bzw. -zentren), Linkermolekül(e), Lösungsmittel, Additive, Synthesezeit und Temperatur (Abbildung 2). Hierzu wurde unter Heranziehung eines Entscheidungsbaums und eines String-basierten Suchverfahrens zunächst eine Klassifizierung relevanter Literaturabschnitte vorgenommen. Im Anschluss daran wurden die Syntheseabschnitte der einzelnen MOF-Strukturen identifiziert (Hintergrundinformationen Section 2.2). Nach Bestimmung der Syntheseabschnitte wurde die Software ChemicalTagger eingesetzt. Diese erkennt und annotiert im Versuchsteil wissenschaftlicher Texte signifikante Wörter innerhalb der Sätze eines Abschnitts.[12] Um die Tagging-Genauigkeit zu erhöhen, wurden die Syntheseabschnitte aufgrund der spezifischen Beschreibungen im Bereich der MOFs leicht modifiziert (Hintergrundinformationen Section 2.3). Zur Bewertung der Genauigkeit der automatisch extrahierten Datenbank SynMOF-A wurden zusätzlich manuell korri-

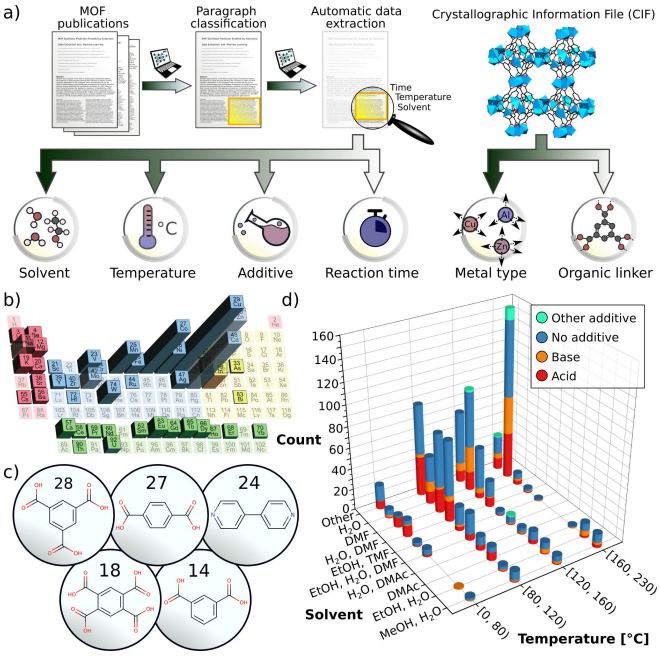

Abbildung 2. SynMOF-Datenbank. a) Data-Mining-Pipeline und Inhalt der SynMOF-Datenbank, b) Statistik der häufigsten Metallzentren und c) Struktur und Vorkommen der häufigsten Linker in der SynMOF-Datenbank, d) 3D-Grafik mit der Korrelation zwischen Lösungsmitteln, Additiven und Temperatur.

gierte Versionen erstellt: Die Datenbanken SynMOF-M und SynMOF-ME, die in Hintergrundinformationen Section 2.4 vorgestellt und diskutiert werden.

Neben der Extraktion von Syntheseinformationen aus der MOF-Fachliteratur wurden automatisch Informationen über Linker und Oxidationszustand des Metallzentrums aus den Crystallographic Information Files (CIFs) extrahiert.[13] Schließlich wurden die Synthesedaten aus den Publikationen (Metallzentrum, Linker, Lösungsmittel, Additive, Synthesezeit und Temperatur) mit den Informationen über Linker und Metallzentrum aus den CIFs in der Datenbank SynMOF zusammengeführt (Abbildung 2). Wir gehen in dieser Arbeit davon aus, dass die entstandene SynMOF-Datenbank geeignet sein wird, ML-Modelle für das Entdecken von Ähnlichkeiten in den Synthesebedingungen zu trainieren und so die Vorhersage von Syntheseprotokollen für neue MOF-Strukturen zu ermöglichen.

Die SynMOF-Datenbank enthält derzeit 983 MOF-Strukturen und liefert neben detaillierten Informationen über MOF-Synthesebedingungen auch statistische Daten über die Metallzentren und die organischen Komponenten (Abbildung 2b,c). Die Datenbank umfasst 46 verschiedene Metalle, deren Oxidationszustände meist zwischen +1 und +3 liegen. Erwartungsgemäß bestehen die meisten MOF-Strukturen aus Übergangsmetallen. Kupfer und Zink machen dabei allein schon fast 50 % aller Metalle aus. Die häufigsten Linker aus der Vielzahl organischer Moleküle sind mehrzähnige Carboxylsäuren (wie Benzol-1,3,5-Tricarboxylsäure, Benzol-1,4-Dicarboxylsäure und Benzol-1,2,4,5-Tetracarboxylsäure), gefolgt von N-haltigen Basen (wie Pyridin, Triazol und Tetrazol).

Auf der Suche nach eindeutigen Mustern wurden die häufigsten Lösungsmittel für die MOF-Synthese bezüglich verschiedener Temperaturregime und Additive analysiert (Abbildung 2d). Bei Temperaturen zwischen 80 °C und 160 °C werden am häufigsten DMF und Wasser sowie deren Mischungen mit anderen Lösungsmitteln als Lösungsmittel eingesetzt. Die Synthese bei Temperaturen über 160 °C erfolgt überwiegend in Wasser als einzigem Lösungsmittel. Die meisten MOF-Synthesereaktionen bei hohen Temperaturen (über 120 °C) finden außerdem ohne Additive statt, während bei Temperaturen unter 80 °C die Verwendung von sauren Additiven vorherrscht. Neben diesen relativ einfachen Mustern konnten wir weitere, in den Daten verborgene Korrelationen (Hintergrundinformationen Section 2.5) durch den Einsatz von ML entdecken.

Mit den Daten der SynMOF-Datenbank wurden mehrere ML-Modelle trainiert, um die Synthesebedingungen einer Reihe unterschiedlicher MOFs vorherzusagen, die nicht Teil der Trainingsdaten waren. Die Input-Darstellung der MOF-Strukturen ist für die Leistung der ML-Modelle von entscheidender Bedeutung.[14] In dieser Arbeit wurden zwei Arten der Darstellung als Input für das Training der ML-Modelle verwendet: Die erste basiert auf dem molekularen Fingerabdruck der Linker, erweitert durch Kodierungen der Metallarten und deren Oxidationszuständen (Abbildung 3a, Hintergrundinformationen Section 3.1), die andere ist die kürzlich von Kulik und KollegInnen entwickelte MOF-Repräsentation (Hintergrundinformationen Section 3.2).[15]

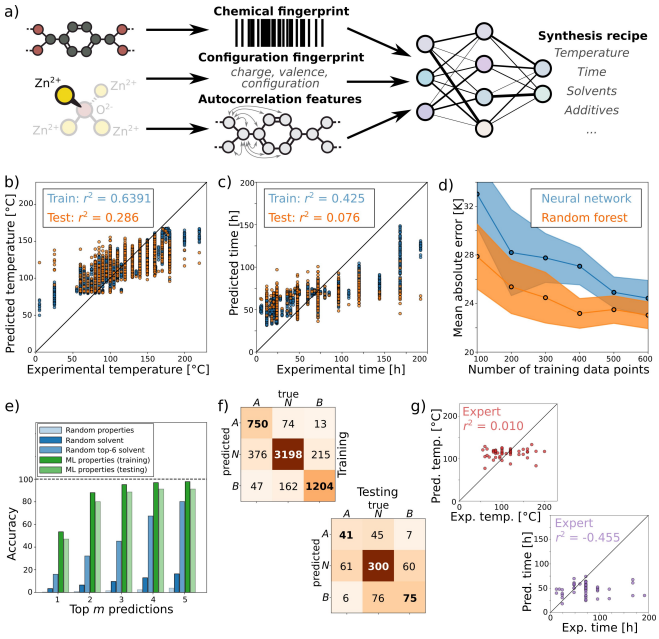

Abbildung 3. Modelle trainiert durch Maschinelles Lernen mit der SynMOF-A-Datenbank. a) ML-Workflow mit Fingerabdruck-Darstellung der Linker und Darstellungen des Metalltyps und des Oxidationszustands, b) und c) Vergleich der ML–Vorhersagen für Zeit und Temperatur bei Trainingsdatensätzen und Testdatensätzen der ursprünglich aus der Fachliteratur extrahierten Daten, d) Lernkurve bei den Temperaturvorhersagen, d. h. der mittlere absolute Fehler als Funktion der Größe des Trainingsdatensatzes, für Modelle mit neuronalen Netzen und mit Random-Forest-Regression, e) Genauigkeit der ML–Vorhersagen für Lösungsmittel bei einer Teilgruppe von MOFs mit einem einzigen Lösungsmittel, verglichen mit Zufallsvorhersagemethoden, f) Leistungsdaten zur Klassifikation von Additiven bei Trainings- und Testdatensätzen, dabei stehen A, B und N für sauer, basisch und keine Additive und g) Durchschnittswert von 11 Vorhersagen für Zeit und Temperatur von menschlichen Experten für 50 MOFs zur Bewertung der Komplexität der Problemstellung.

Es ist zu beachten, dass das Gebiet der MOF-Entwicklung einem stetigen Wachstum unterliegt. Mit der Zeit wird daher also eine immer größere Anzahl neuer Strukturen und Syntheseparameter zum Trainieren und Optimieren von ML-Modellen und damit zum Erzielen bestmöglicher Leistungen verfügbar sein. Neue ML Modelle, wie z. B. Graph neuronale Netze, werden in Zukunft voraussichtlich präziser sein als Modelle, die auf von Hand erstellten Darstellungen von Merkmalen angewiesen sind.[16]

Die Vorhersage von Synthesezeit und Temperatur erfolgte in dieser Studie mittels Regressionsmodellen wie Random Forests und Neuronalen Netze (Hintergrundinformationen Section 3.3, 3.4, 3.5). Um diskrete Syntheseparameter wie Lösungsmittel und Additive vorherzusagen, könnten grundsätzlich auch Klassifikationsmodelle eingesetzt werden. Allerdings hat sich dies aus mehreren Gründen als unzweckmäßig erwiesen. In der Literatur wird eine Vielzahl möglicher Lösungsmittel und Additive angegeben, aus denen sich eine große Zahl an Kategorien und somit äußert unausgewogene Datensätze ergeben. Darüber hinaus können sich die Eigenschaften einiger Lösungsmittel so sehr

ähneln, dass sie bei der Synthese austauschbar sind. Dies führt zu mehrdeutigen Ergebnissen. Zusätzlich sind in der Praxis für eine erfolgreiche MOF-Synthese teils auch Kombinationen mehrerer Lösungsmittel erforderlich. Daher wurde für diese Arbeit ein ML-Modell entwickelt, das die Eigenschaften der Lösungsmittel, wie zum Beispiel den Verteilungskoeffizienten und den Siedepunkt (Hintergrundinformationen Section 3.6), vorhersagt, nicht aber das Lösungsmittel selbst. Eine Nächste-Nachbarn-Suche im Bereich der Lösungsmittel-Eigenschaften ergibt eine Liste potentieller Lösungsmittel, deren Eigenschaften denen nahekommen, die mit Hilfe des ML-Modells vorhergesagt wurden. So können neue Lösungsmittel und auch Lösungsmittel, die nur ein einziges Mal in der Literatur vorkommen, beim Trainieren des Modells berücksichtigt werden. Für die Additive wurde der pH-Wert als wichtigster Parameter zur Unterscheidung ermittelt. Daher wurden die Datensätze in drei Gruppen eingeteilt (sauer, basisch, keine Additive) und es wurde für die Vorhersage der Additive ein Klassifikationsmodell verwendet.

Die Ergebnisse der trainierten ML-Modelle werden in Abbildung 3b–f dargestellt. Reproduzierbar positive Korrelationskoeffizienten r^2 bei neuen Testdatensätzen zeigen, dass die ML-Modelle in der Lage sind, vorhersehbare und aussagekräftige Beziehungen zwischen der MOF-Zielstruktur und den nötigen Synthesebedingungen herzustellen, insbesondere hinsichtlich Temperatur und Zeit (Abbildung 3b,c). Mit dem gegenwärtigem Stand der aus der Fachliteratur extrahierten Datenmenge zeigt sich, dass die Random-Forest-Modelle bei allen vorhergesagten Parametern die besten Ergebnisse liefern. Allerdings lernen neuronale Netze mit wachsenden Datensatzgrößen schneller, bessere Vorhersagen zu machen (siehe Lernkurven in Abbildung 3d) und sind der Lage, Korrelationen zwischen verschiedenen Syntheseparametern (z. B. Lösungsmittel und Temperatur) auszunutzen, anstatt sie unabhängig voneinander vorherzusagen. Demnach ist zu erwarten, dass komplexere Modelle die Ergebnisse der Random-Forest-Modelle in absehbarer Zukunft übertreffen werden.

Um die ML-basierte Vorhersage der Lösungsmittel zu bewerten, wurde eine Teilgruppe von mit nur einem Lösungsmittel synthetisierten MOFs genauer betrachtet. Ein Vergleich der Genauigkeit der sechs besten ML–Vorhersagen mit mehreren Zufallsmethoden als Baseline (Abbildung 3e), darunter die Auswahl eines zufälligen Lösungsmittels aus allen verfügbaren sowie den sechs häufigsten Lösungsmitteln, die in 96 % Prozent der Fälle mit nur einem Lösungsmittel in der SynMOF-Datenbank zum Einsatz kommen, hat ergeben, dass das ML-Modell die Zufallsauswahl übertrifft. Dies ist insbesondere der Fall für die besten 3 Lösungsmittelvorhersagen, bei denen das ML-Modell eine Genauigkeit von $>90\,\%$ erreicht. Bei der Vorhersage der Additive (Abbildung 3f) ist es die Aufgabe des ML-Modells, die benötigten Additive als sauer, basisch oder als keine Additive enthaltend zu klassifizieren. Während das Modell bei den Trainingsdaten gute Ergebnisse liefert, leidet die Ergebnisqualität bei neuen Testdaten darunter, dass die Datensätze sehr unausgewogen sind (die meisten Einträge in der Datenbank berücksichtigen keine Additive). Die Verwendung von Ausgleichsgewichtungen der Trainingsdatenpunkte ermöglicht Vorhersagen, die gut zwischen Synthesevorgängen mit sauren und basischen Additiven unterscheiden. Die Unterscheidung zwischen sauren Additiven und dem Fehlen von Additiven oder basischen Additiven und dem Fehlen von Additiven ist dagegen weniger stark ausgeprägt. Ein Grund dafür könnte in den verborgenen Variablen, wie der Art und Funktion der Additive, zu finden sein. Manche Additive (anorganische Säuren und Basen) regulieren nur den pH-Wert, während andere (organische Säuren und Basen) auch an der Modulation des MOF-Wachstums beteiligt sind. Daneben sind auch die Konzentration und Stärke der Additive wichtige Parameter für die Rolle der Additive bei der Synthese. Weitere Trainingsdaten werden in der Zukunft die Darstellung der Additive verfeinern und eine Verbesserung des ML-Modells ermöglichen. Hierdurch werden sich neue Perspektiven für die Vorhersage von Synthesebedingungen ergeben.

Zu beachten ist, dass die Vorhersage von MOF-Synthesebedingungen eine uneindeutige Aufgabe darstellt: anstatt einer einzigen korrekten Antwort existiert ein ganzes Spektrum verschiedener Bedingungen, die zusammen eine erfolgreiche Synthese ermöglichen. Die in der Fachliteratur veröffentlichten Daten sind sehr heterogen, da nur einige der Reaktionen für die Ausbeute oder andere Zwecke ausgelegt optimiert wurden. Je nach MOF kann das Fenster der nahezu optimalen Bedingungen außerdem größer oder kleiner sein. Im Gegensatz zu anderen Anwendungen für maschinelles Lernen ist es daher unwahrscheinlich, dass ein Modell, selbst ein perfektes, einen r^2-Wert von 1 erreicht.

Zur Einschätzung der Leistung des ML-Modells wurden 11 menschliche ExpertInnen auf dem Gebiet der MOF-Synthese befragt. Dazu wurde ein öffentlich zugänglicher Online-Test mit 50 zufällig aus der SynMOF-Datenbank ausgewählten MOFs entwickelt. Die Teilnehmer erhielten die 3D-Strukturen der MOFs, die chemische Struktur der Linker sowie Informationen über die Metallionen und wurden gebeten, die Synthesebedingungen wie Temperatur, Zeit, Lösungsmittel und Additive abzuschätzen, ohne auf Literatur oder andere externe Quellen zurückzugreifen (Hintergrundinformationen Section 3.7). Im Anschluss an jede Vorhersage waren die TeilnehmerInnen zudem dazu aufgefordert anzugeben, wie sicher sie sich bezüglich ihrer jeweiligen Einschätzungen waren. Die Korrelationskoeffizienten r^2 zwischen den Zeit- und Temperaturangaben der ExpertInnen und den angegebenen Bedingungen liegen nahe null, selbst für einen Durchschnittswert von 11 Schätzungen verschiedener Forschender (Abbildung 3g) und auch dann, wenn nur Schätzungen mit höherer Sicherheit berücksichtigt werden. Dieses recht überraschende Ergebnis zeigt, dass auch geringe vom ML-Modell gelernte und genutzte Korrelationen bei der Vorhersage von Synthesebedingungen hilfreich sind.

Zusammenfassend konnte gezeigt werden, dass die ML-Modelle in der Lage sind, generalisierende Muster und Korrelationen aus der SynMOF-Datenbank zu lernen, mit denen die allgemeine Intuition eines Experten übertroffen wird. Somit lässt sich mit Hilfe der Modelle eine gute erste

Schätzung für die experimentellen Syntheseversuche neuer MOFs erhalten.

Es wurde eine unter https://mof-synthesis.aimat.science/ verfügbare Webseite entwickelt, auf der MOF-Synthesebedingungen durch die Modelle aus der vorliegenden Arbeit vorhergesagt werden. NutzerInnen können auf dieser Webseite eigene MOF-CIFs hochladen. Das Web-Tool macht daraufhin eine Vorhersage einschließlich der Synthesezeit, Temperatur, Lösungsmittel und Additive (sauer, basisch oder keine Additive).

Abschließend lässt sich zusammenfassen, dass der Mangel an maschinenlesbaren, gepflegten MOF-Synthesedaten die Entwicklung eines digitalen ML-Tools für die Vorhersage von MOF-Synthesebedingungen bis jetzt verhindert hat. In dieser Arbeit wurde nun durch automatische Datenextraktion mit NLP-Methoden eine SynMOF-Datenbank erstellt, die Synthesebedingungen und Strukturinformationen von mehr als 900 MOFs enthält. Mit diesen Daten wurden ML-Modelle zur Identifizierung von Mustern in der MOF-Synthese trainiert. Es ist zu erwarten, dass die entstandene Datenbank die NLP-Forschung in MOF-Fachkreisen intensivieren und dass die entwickelte Plattform für die ML–Vorhersage der Synthese der neue Standard in der datenbasierten MOF-Entwicklung werden wird. Obwohl die Arbeiten erst am Beginn stehen, übertreffen die Ergebnisse der Vorhersagemodelle die Vorhersagen von MOF-Experten. Dies verdeutlicht einerseits die Komplexität des Syntheseverfahrens und andererseits den dringenden Bedarf an digitalen Tools zur Vorhersage. Die automatisierte Synthesevorhersage, die durch diese Arbeit auf Abruf verfügbar wird, wird zu einer deutlichen Beschleunigung bei der Entwicklung neuer MOFs führen und nicht nur für die MOF-Forschungsgemeinschaft ein wertvolles Tool darstellen.

Danksagung

Wir bedanken uns bei 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 Université), Dr. Akira Hinokimoto, Dr. Eli Sanchez Gonzalez, Dr. Javier Troyano (Universität Kyoto) und Dr. Romy Ettlinger (Universität Augsburg) für ihre Teilnahme als Experten für MOF-Synthesebedingungen. Weiterhin danken wir Matthias Schniewind für seine Hilfe bei der Entwicklung der Internet-Plattform. Y.L. dankt Dr. Cam An Nguyen Thanh für die Unterstützung bei der Java-Programmierung. S.W. bedankt sich für die erhaltene Finanzierung durch das Baskische Ministerium für Industrie im Rahmen der Programme ELKARTEK und HAZITEK. M.T. bedankt sich für die Finanzierung durch den Impuls- und Vernetzungsfonds der Helmholtz-Gemeinschaft (Grant VH-NG-1147) und durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen des Projektes FAIRmat – FAIR Data Infrastructure for Condensed-Matter Physics and the Chemical Physics of Solids. Y.L. dankt dem China Scholarship Council für die finanzielle Unterstützung (No. 201706270179). Die AutorInnen bedanken sich für die Unterstützung durch das Land Baden-Württemberg im Rahmen von bwHPC. Open Access Veröffentlichung ermöglicht und organisiert durch Projekt DEAL.

Interessenkonflikt

Die Autoren erklären, dass keine Interessenkonflikte vorliegen.

Erklärung zur Datenverfügbarkeit

Die Datenbanken SynMOF-A, SynMOF-M und SynMOF-ME, die Codes für die Extraktion der Syntheseparameter, für das ML-Training und die Vorhersage sowie die Expertenumfrage sind kostenlos erhältlich unter https://github.com/Tsotsalas-Group/MOF_Literature_Extraction und https://github.com/aimat-lab/MOF_Synthesis_Prediction.

Stichwörter: Data-Mining · Maschinelles Lernen · Metallorganische Gerüststrukturen · Mikroporöse Materialien · Synthesevorhersage

[1] R. Freund, O. Zaremba, G. Arnauts, R. Ameloot, G. Skorupskii, M. Dincă, A. Bavykina, J. Gascon, A. Ejsmont, J. Goscianska, M. Kalmutzki, U. Lächelt, E. Ploetz, C. S. Diercks, S. Wuttke, *Angew. Chem. Int. Ed.* **2021**, *60*, 23975–24001; *Angew. Chem.* **2021**, *133*, 24174–24202.

[2] a) G. Férey, *Chem. Soc. Rev.* **2008**, *37*, 191–214; b) H. Furukawa, K. E. Cordova, M. O'Keeffe, O. M. Yaghi, *Science* **2013**, *341*, 1230444; c) S. Kitagawa, R. Kitaura, S.-I. Noro, *Angew. Chem. Int. Ed.* **2004**, *43*, 2334–2375; *Angew. Chem.* **2004**, *116*, 2388–2430; d) C. Gropp, S. Canossa, S. Wuttke, F. Gándara, Q. Li, L. Gagliardi, O. M. Yaghi, *ACS Cent. Sci.* **2020**, *6*, 1255–1273.

[3] a) K. M. Jablonka, D. Ongari, S. M. Moosavi, B. Smit, *Chem. Rev.* **2020**, *120*, 8066–8129; b) H. Lyu, Z. Ji, S. Wuttke, O. M. Yaghi, *Chem* **2020**, *6*, 2219–2241; c) Y. Luo, M. Ahmad, A. Schug, M. Tsotsalas, *Adv. Mater.* **2019**, *31*, 1901744.

[4] a) V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, *Nature* **2019**, *571*, 95–98; b) P. S. Gromski, A. B. Henson, J. M. Granda, L. Cronin, *Nat. Chem. Rev.* **2019**, *3*, 119–128; c) M. Ahmad, Y. Luo, C. Wöll, M. Tsotsalas, A. Schug, *Molecules* **2020**, *25*, 4875.

[5] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, *npj Comput. Mater.* **2017**, *3*, 54.

[6] a) Z. Jensen, E. Kim, S. Kwon, T. Z. H. Gani, Y. Román-Leshkov, M. Moliner, A. Corma, E. Olivetti, *ACS Cent. Sci.* **2019**, *5*, 892–899; b) K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547–555; c) P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, *Nature* **2016**, *533*, 73–76; d) Z. Jensen, S. Kwon, D. Schwalbe-Koda, C. Paris, R. Gómez-Bombarelli, Y. Román-Leshkov, A. Corma, M. Moliner, E. A. Olivetti, *ACS Cent. Sci.* **2021**, *7*, 858–867; e) P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, A. A. Lee, *ACS Cent. Sci.* **2019**, *5*, 1572–1583; f) J. P. Reid, M. S. Sigman, *Nature* **2019**, *571*, 343–348.

[7] S. M. Moosavi, A. Chidambaram, L. Talirz, M. Haranczyk, K. C. Stylianou, B. Smit, *Nat. Commun.* **2019**, *10*, 539.

[8] D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson, A. Aspuru-Guzik, *Nat. Rev. Mater.* **2018**, *3*, 5–20.
[9] 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*, 5985–5998.
[10] C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, *Acta Crystallogr. Sect. B* **2016**, *72*, 171–179.
[11] a) 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; b) E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, E. Olivetti, *Chem. Mater.* **2017**, *29*, 9436–9444.
[12] L. Hawizy, D. M. Jessop, N. Adams, P. Murray-Rust, *J. Cheminf.* **2011**, *3*, 17.
[13] K. M. Jablonka, D. Ongari, S. M. Moosavi, B. Smit, *Nat. Chem.* **2021**, *13*, 771–777.
[14] O. A. von Lilienfeld, K. Burke, *Nat. Commun.* **2020**, *11*, 4895.
[15] 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.
[16] a) J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, *Proceedings of the 34th International Conference on Machine Learning, Bd. 70*, **2017**, S. 1263–1272; b) K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, *J. Chem. Phys.* **2018**, *148*, 241722; c) P. Friederich, F. Häse, J. Proppe, A. Aspuru-Guzik, *Nat. Mater.* **2021**, *20*, 750–761.

Manuskript erhalten: 6. Januar 2022
Akzeptierte Fassung online: 1. Februar 2022
Endgültige Fassung online: 10. März 2022