DigiMOF: A Database of MOF Synthesis Information Generated via Text Mining

*Kristian Gubsch[‡], Rosalee Bence[‡], Lawson T. Glasby[‡] and Peyman Z. Moghadam**

Department of Chemical and Biological Engineering, The University of Sheffield, Sheffield S1 3JD, United Kingdom. *Email: p.moghadam@sheffield.ac.uk

ABSTRACT: The vastness of materials space, particularly that which is concerned with metal-organic frameworks (MOFs), creates the critical problem of performing efficient identification of promising materials for specific applications. Although high-throughput computational approaches, including the use of machine learning, have been useful in rapid screening and rational design of MOFs, they tend to neglect descriptors related to their synthesis. One way to improve the efficiency of MOF discovery is to data mine published MOF papers to extract the materials informatics knowledge contained within the journal articles. Here, by adapting the chemistry-aware natural language processing tool, ChemDataExtractor (CDE), we generated an open-source database of MOFs focused on their synthetic properties: the DigiMOF database. Using the CDE web scraping package alongside the Cambridge Structural Database (CSD) MOF subset, we automatically downloaded 43,281 unique MOF journal articles, extracted 15,501 unique MOF materials and text mined over 52,680 associated properties including synthesis method, solvent, organic linker, metal precursor, and topology. This centralised, structured database reveals the MOF synthetic data embedded within thousands of MOF publications. The DigiMOF database

and associated software are publicly available for other researchers to conduct further analysis of alternative MOF production pathways and create additional parsers to search for other desirable properties.

1. Introduction

Metal-organic frameworks (MOFs) are a class of crystalline materials consisting of a lattice of metal ions co-ordinately bonded by organic linkers. MOFs are well-known for their extremely high surface areas and exceptionally tuneable properties which enable their potential application in areas including gas storage [1–4], sensing [5,6], separations [7,8], drug delivery [9–11], and catalysis [12–16]. Since the first MOFs were synthesised in the 1990s, thousands of MOFs have been produced at laboratory scale. As of 2021, more than 100,000 MOF structures have been reported in the Cambridge Structural Database (CSD)[17,18]. The sheer volume of distinct real MOF materials poses significant challenges for screening and isolating the best candidates for a given application: a typical problem of finding a needle in a haystack. To some extent, this has been counteracted by the use of high-throughput computational screening and machine learning for the elucidation of structure-property relationships, in particular for gas adsorption and separation properties of MOFs [19–25]. Given that these screening methods tend to neglect synthesis data, the identification of economical and sustainable synthesis routes has remained largely a manual process, and clearly, relying on experimental trial-and-error and serendipity to develop MOFs is costly, slow, and unreliable. To address these challenges, we propose the use of high-throughput text mining to collect MOF synthesis data in a single resource and to aid the design and discovery of more practical MOFs by valorising their synthesis information.

Most chemistry literature is published as unstructured text which makes manual database creation cumbersome, time-consuming, and error prone. To address this problem, Swain and Cole developed ChemDataExtractor (CDE) to automate the extraction of chemical data from research articles and patents via text mining [26]. To date, CDE has been deployed to automatically assemble databases of magnetic materials, battery materials, UV/Vis absorption spectra, light harvesting materials, hydrogen storage applications, and nanomaterials synthesis [27]. Whilst CDE has been used to text mine both organic and inorganic chemistry literature, it has yet to be applied to MOFs, possibly due to challenges presented by the diverse nature of their building blocks and complex synthesis techniques. To the best of our knowledge, Park et al.'s text mining software was the first work which enlisted text mining to scrape MOF-related data such as pore volume and surface area.[28] More recently, Luo et al.[29] developed an automatic data mining tool using the CoRE MOF database[30], alongside web-scraping tool Puppeteer (https://pptr.dev), to text mine 6099 journal articles. These were then analysed using ChemicalTagger software[31] to extract metal source, linker(s), solvent(s), additive, synthesis time, and temperature.

The CSD MOF subset contains comprehensive structural information about MOFs, however, the data related to their synthesis is scarce and inconsistent. Here, we developed rule-based MOF compound name and property parsers within CDE to automatically generate a database of MOF synthesis data, i.e. the DigiMOF database, to facilitate digital transformation of MOFs' synthesis protocols. We envisage that DigiMOF will allow next-generation high-throughput screening and machine learning approaches to take more circumspective consideration of the synthesis information. These new features will allow MOF scientists to rapidly search for MOFs associated with specific precursors, topologies, organic linkers, and synthesis routes, offering a platform which facilitates screening and identification of sustainable and scalable materials. For each MOF compound, its corresponding DOI is also included in the database so users can access the publication where it was first reported. We highly encourage users of DigiMOF to build upon this foundational work and integrate additional MOF property extraction capabilities into the adapted CDE to expand or tailor the database according to their own research requirements.

2. Property Identification and Parsing

The principal challenge in developing text mining parsers is to identify key MOF properties for data extraction. Initially, we conducted an extensive review of existing literature to select properties that are most indicative of MOF scalability and ease of synthesis. Given the widespread interest in MOF chemistry, it is somewhat surprising that only a few MOF technoeconomic assessments (TEA), with a focus on production, have been carried out. For example, DeSantis et al. [32] demonstrated that switching from traditional solvothermal synthesis techniques to more novel, less solvent-intensive pathways such as aqueous or mechanochemical routes, could reduce MOF production costs by 34-83%. Increasing the MOF yield by a factor of 30% had a negligible impact on production cost in comparison to using a less solvent-intensive pathway. In another study, Luo et al. [33] compared traditional solvothermal synthesis with an aqueous pathway to produce UiO-66-NH_2 and found that omitting solvents from the synthesis of this MOF resulted in an 83.8% reduction in production cost. The key properties that influenced production cost were solvent, organic linker, and inorganic MOF precursors.

Following these findings, we focused on constructing parsers to extract information on four key MOF synthesis properties: solvents, inorganic and organic precursors, and synthesis methods. We also constructed a parser to extract MOF topologies as the description of topology aids mechanical stability predictions, critical for the pelletisation and industrial application of MOFs [34]. Finally, integration with the CSD Python API also allowed information such as the tested temperature, article DOI, and publication year to be merged with the parser extracted records.

3. Methods: Automatic Generation of the DigiMOF Database

The key motivation for adapting the CDE tool to text mine MOF literature was to better integrate MOF synthesis protocols, TEA considerations, and computational screening approaches into a tight feedback loop to enable more efficient MOF materials development. **Figure 1** demonstrates how the DigiMOF database and the adapted CDE parsers can be integrated into a data-driven pipeline for MOF design and discovery.

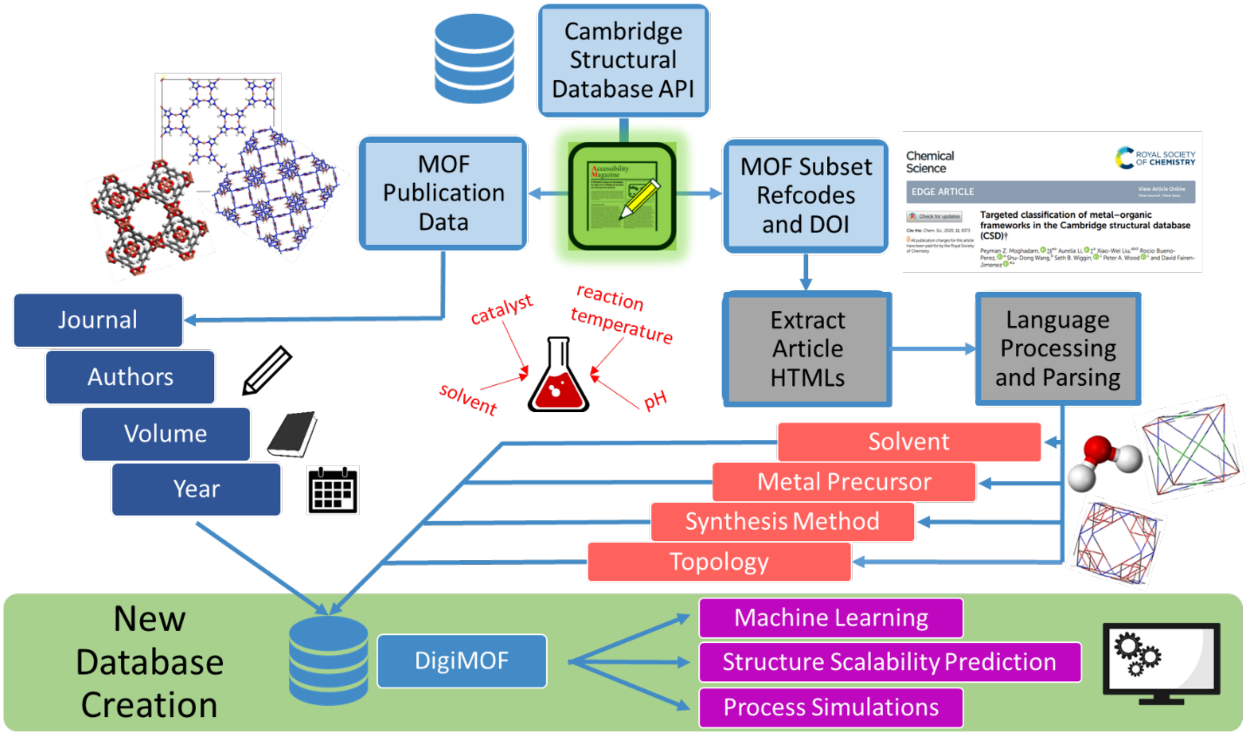
Cambridge Structural Database API
MOF Publication Data
MOF Subset Refcodes and DOI
Chemical Science
ROYAL SOCIETY OF CHEMISTRY
EDGE ARTICLE
View Article Online
Check for updates
Cite this: Chem. Sci., 2020, 11, 8373
Targeted classification of metal–organic frameworks in the Cambridge structural database (CSD)†
Peyman Z. Moghadam, Aurelia Li, Xiao-Wei Liu, Rocio Bueno-Perez, Shu-Dong Wang, Seth B. Wiggin, Peter A. Wood and David Fairen-Jimenez
Journal
Authors
Volume
Year
catalyst
reaction temperature
solvent
pH
Extract Article HTMLs
Language Processing and Parsing
Solvent
Metal Precursor
Synthesis Method
Topology
New Database Creation
DigiMOF
Machine Learning
Structure Scalability Prediction
Process Simulations

Figure 1. A flow diagram to visualise the integration of CDE into a data-driven MOF synthesis plan: from article retrieval to text mining, computational screening, and materials discovery.

We also developed a MOF-specific approach in conjunction with the CDE web scraper: DOIs associated with the CSD MOF subset were extracted using the CSD API and used to automatically download the associated articles in HTML format using the CDE web scraping script for the corresponding journal. After download, text mined MOF synthesis data was automatically extracted from each HTML file and stored in our database in a JSON format. This data can then be used for further TEA studies and integrated with other physicochemical properties obtained from either simulations or experiments to generate rich datasets for further processing.

3.1 Natural Language Processing

To identify specific MOF properties, using CDE based classes and variables, we created customised parsers which utilise Part-of-Speech (POS) taggers and chemical entity recognisers. These parsers contain specific regular expressions for the identification of MOF compound names. The natural language processing (NLP) pipeline in CDE first identifies a sentence which is then tokenised into individual words and punctuation known as tokens [26]. These tokens are marked up by POS tagging to reflect their syntactical functions, such as noun, verb, chemical mention, and adjective [26]. Entity recognition of the chemical species allows relationships to be extracted and merged with their corresponding compound by interdependency resolution [26]. Our rule-based parsers utilised Python regular expressions as well as CDE parsing elements and were tailored to extract specific properties. We generated parsing rules to identify MOF names, synthesis methods, inorganic precursors, linker names, and MOF topology abbreviations, as well as creating blacklists to exclude words which were frequently misidentified as these variables. The use of regular expressions and parsing elements shown in **Table S.1**, were crucial to improving performance.

The process of building and refining the parsers is shown in **Figure 2** following a similar process used by Huang & Cole [35]. First, basic parser functionality was achieved on individual sentences by successfully extracting the MOF compound name and corresponding property. The parsers were then tested on a series of sets containing 10 random papers and continuously refined until they achieved a precision above 80% on one test set. The last step of the process was evaluating parser performance on a final set of 50 randomly selected papers from the CSD.

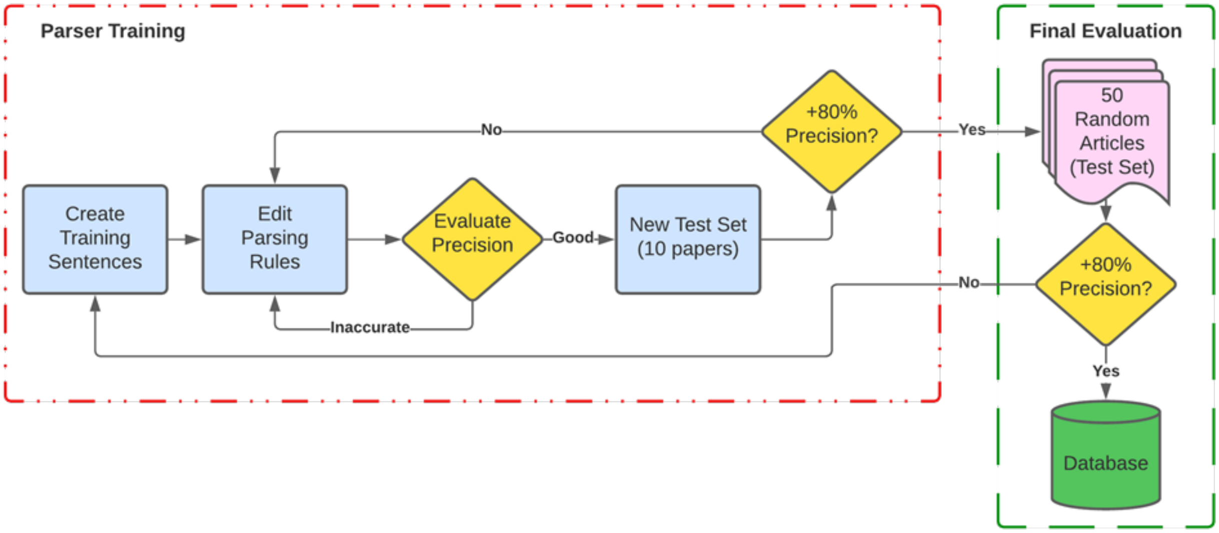

Figure 2. An iterative flow chart depicting the process through which the parsers were refined before database application.

3.2 Technical Validation

This text mining software was evaluated for reproducibility on a randomly selected array of 'unseen' text, distinct from the training set used to refine the NLP parsers, to ensure the parser performance achieved on a limited training set can be consistently replicated for high-throughput application. The three performance metrics used in evaluation are precision, recall, and F-score which can be calculated using equations 1, 2 and 3, respectively. True positives (TP) correspond to data extracted and identified correctly. False positives (FP) correspond to data which are incorrectly identified as a match. False negatives (FN) are relevant data which should be extracted but have not been identified.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

Precision is the fraction of correctly extracted data, recall is the fraction of available data extracted, and F-score represents the harmonic mean of recall and precision. For the estimation of precision and recall, 50 MOF articles were randomly selected as the test set from a collection of over 700 articles retrieved by the web scraper from the CSD: the selected articles can be found in the Supporting Information. For each extracted record, a value of 1 was assigned if both the MOF compound name and the corresponding property (e.g. synthesis method, linker, etc.) were correctly matched, or a value of 0 if the compound name or the property were incorrectly matched. The number of total relationships was manually extracted from the same 50 journal articles and compared with the records in the auto-generated database to calculate recall and precision.

In practice, there is often a trade-off between the precision and recall of a text mining algorithm. The development and implementation of rule-based parsers prioritises high precision which reduces the overall recall as the parser is less capable of extracting values from many variations in sentence structure. More lenient parsing rules increase the overall number of records extracted and therefore improved recall, but will show reduction in specificity which reduces precision.

Generally, high precision should be given precedence over recall; low recall is acceptable provided that a large enough dataset is used to compensate for a lower proportion of the available data being extracted. Examples of the compound records from this work and previous projects utilising CDE are shown in **Table S2**. We found it extremely challenging to accommodate the considerable diversity of sentence structures observed in MOF literature without compromising the precision of the parsers. When maximising precision, extracting common and unambiguous sentences observed in MOF literature was prioritised, although it was expected that lower recall would be obtained compared to previous iterations of CDE. **Figure S1** summarises the overall performance of our parsers compared to previous CDE projects, and the MOF text mining tool from Park et al.[28] The overall precision for our parsers was 77% which we deemed satisfactory as values approaching 80% are generally considered sufficient for data-driven materials discovery via current text mining techniques[35]. A breakdown of individual parser results for synthesis route, topology, linkers, and metal precursors can be found in **Table S3**.

3.3 Parser Training

During parser training, precision was substantially improved by employing blacklists to filter out frequently observed misidentifications. The addition of common abbreviations, names, and blacklist items for metal precursors, linkers, MOFs, and topologies to the regular expressions helped to improve both precision and recall. As MOF terminology and literature is dynamic and rapidly evolving, it is crucial that continued adaptations be made to this tool to improve its performance. With this idea in mind, we have made the software open source with the aim of using open collaboration to add abbreviations or names to the blacklists and compound regular expressions, which will allow the tool to evolve and improve over time.

Figure 3 shows the process for selection of regular expressions that should be incorporated into CDE. Here, we demonstrate how regular expressions (regex) may be developed iteratively to achieve more true positives, and to eliminate false positives and negatives. **Table S4** contains examples of simplified regex used in the creation of the DigiMOF database. The actual regex which have been integrated into the MOF version of CDE are available on the associated GitHub (https://github.com/peymanzmoghadam/DigiMOF-database-master-main.git) in the chemical entity mention (CEM) and precursor parser files.

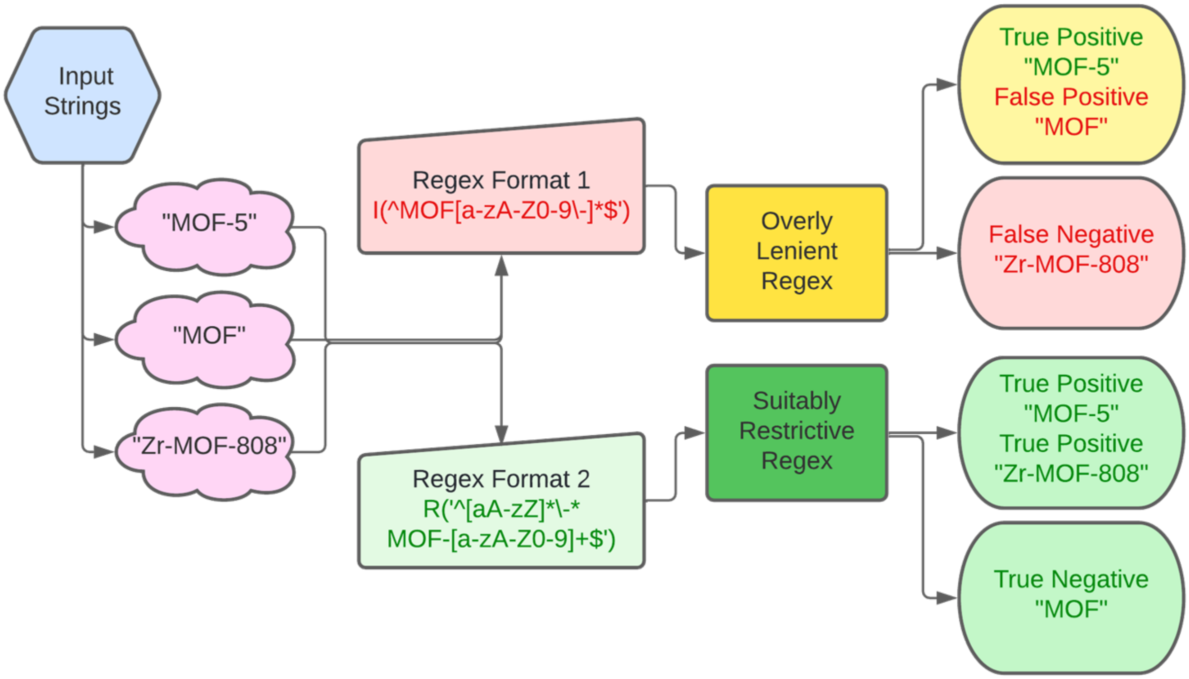

Figure 3. Flow chart displaying possible outcomes when fed an input string for high-throughput MOF name parsing.

It is often preferable to use multiple regular expressions to accommodate different formats of the same variable. Attempting to accommodate too many types of matches into a single expression can increase the number of false positives, as demonstrated by expression number 4 in **Table S.4** which is the lenient regular expression for common linker abbreviations. To accommodate a wider variety of sentence structures to help recognise MOF names, a blacklist was integrated into the regular expression rules to exclude false positives, as with expression 9 in **Table S.4**. Regular expressions within the context of blacklisting are further detailed in the supporting information in **Table S.5**.

4. Results and Discussion

We note that in order for a MOF compound name and corresponding property relationship to be entered into the DigiMOF database, both the MOF compound name and property had to be recognized by the parsers. Overall, 15,501 MOF compound name and property relationships with over 52,680 associated properties were extracted from the CSD MOF subset which contains 43,281 unique MOF publications and over 100,000 MOFs. **Table 1** displays the total number of each type of synthesis property associated with these MOFs, in addition to the total number of unique properties of each type. The full list of MOF names and their relevant properties can be found in the Supporting Information.

Table 1. Total number of extracted properties and number of unique properties for each MOF property in the DigiMOF database.

Property	Total Extracted	Total Unique Properties Extracted
MOF compound names	15,501	-
Synthesis Route	9,705	8
Solvents	1,211	81
Topologies	6,680	154
Linkers	24,116	10,690
Metals including ions	10,968	1,803
Metals excluding ions and element names	5,163	1,476

Our database contains a MOF compound name and corresponding topology, organic linker, metal precursor, synthesis method, or solvent for approximately 15% of structures within the CSD MOF subset. One important factor to consider is that not every publication discusses all of these properties. If a compound is labelled as “1” or “2” without a specifier such as “compound”, “complex”, or “MOF” then the parsers will not associate the label with anything, and so cannot extract a property relationship. We must also note that full access to every article within the CSD was not possible, either due to the location in which the article was published or that the corresponding papers were written in languages other than English. An extended discussion on how the parsers function is located in the Database Overview and Performance section of the Supporting Information**.** In the following sections, we summarise our key findings after text mining the CSD MOF subset.

5. Data Analysis

Figure 4 shows the most frequently extracted MOF synthesis data from the CSD MOF subset. In the subsequent sections we discuss each MOF property in more detail.

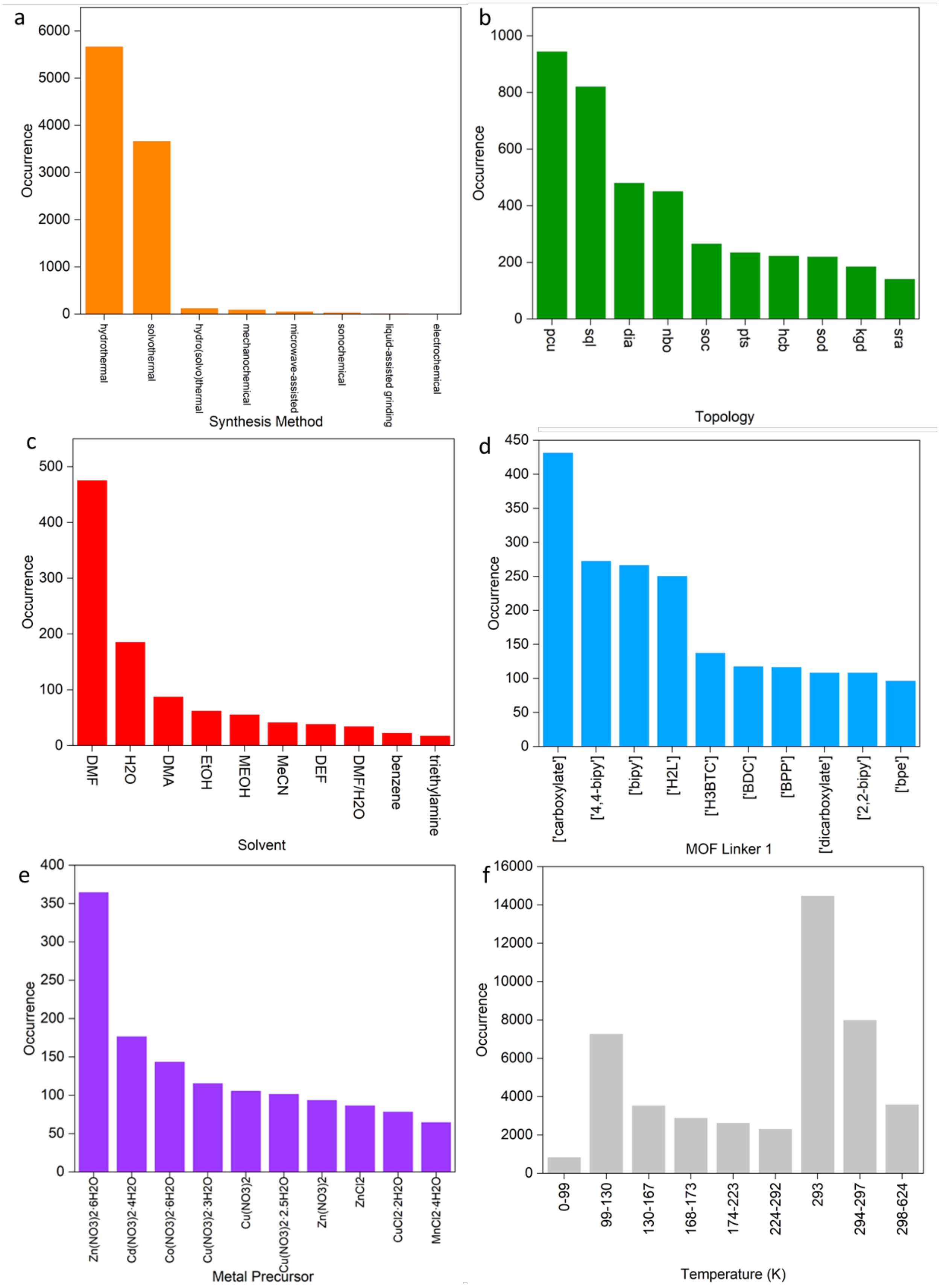
a
Occurrence
6000
5000
4000
3000
2000
1000
0
hydrothermal
solvothermal
hydro(solvo)thermal
mechanochemical
microwave-assisted
sonochemical
liquid-assisted grinding
electrochemical
Synthesis Method
b
Occurrence
1000
800
600
400
200
0
pcu
sql
dia
nbo
soc
pts
hcb
sod
kgd
sra
Topology
c
Occurrence
500
400
300
200
100
0
DMF
H2O
DMA
EtOH
MEOH
MeCN
DEF
DMF/H2O
benzene
triethylamine
Solvent
d
Occurrence
450
400
350
300
250
200
150
100
50
0
['carboxylate']
['4,4-bipy']
['bipy']
['H2L']
['H3BTC']
['BDC']
['BPP']
['dicarboxylate']
['2,2-bipy']
['bpe']
MOF Linker 1
e
Occurrence
400
350
300
250
200
150
100
50
0
Zn(NO3)2·6H2O
Cd(NO3)2·4H2O
Co(NO3)2·6H2O
Cu(NO3)2·3H2O
Cu(NO3)2
Cu(NO3)2·2.5H2O
Zn(NO3)2
ZnCl2
CuCl2·2H2O
MnCl2·4H2O
Metal Precursor
f
Occurrence
16000
14000
12000
10000
8000
6000
4000
2000
0
0-99
99-130
130-167
168-173
174-223
224-292
293
294-297
298-624
Temperature (K)

Figure 4. Histograms showing the most common MOF properties extracted in the DigiMOF database. **a.** synthesis methods, **b.** topologies, **c.** solvents, **d.** organic linkers, **e.** metal precursors, and **f.** temperature.

5.1 Synthesis Methods: When analysing the data for synthesis methods, we first investigated how synthesis methods have changed over time. A total of 9,705 synthesis route records were extracted from 43,281 papers. **Figure 5** shows the cumulative sum of records extracted for various types of synthesis route from 1995 to 2020. Solvothermal synthesis in the context of MOFs generally refers to the use of one or more organic solvents such as DMF and methanol at high temperatures. Hydro(solvo)thermal synthesis generally refers to reactions where water is employed as part of a solvent mixture. Hydrothermal synthesis refers to reactions where water is the primary solvent and is itself a type of solvothermal synthesis. A significant result was the extraction of more hydrothermal (5,677) synthesis methods than solvothermal (3,672). This is surprising as the most common lab-scale MOF synthesis routes are solvothermal, however, many papers do not explicitly name this as their synthesis route but instead imply it by mentioning the use of solvents and high temperatures in the experimental methods section. These implicit synthesis routes could be easily deduced by a reader but are challenging to extract using rule-based NLP algorithms which are looking for a specifier word such as "solvothermal". **Figure 4a.** also shows that hydrothermal synthesis was the most common alternative/low-solvent synthesis route extracted by the parsers.

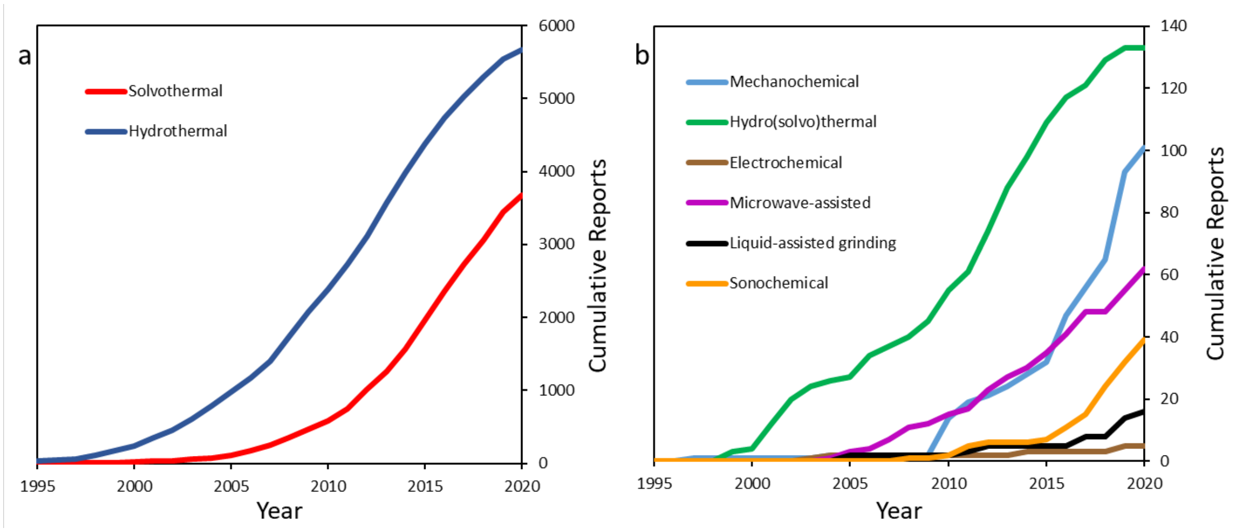

Figure 5. a. The cumulative sum of the two main MOF synthesis methods from 1995 to 2020. **b.** Cumulative sum of alternative and emerging synthesis methods showing periods where these techniques were first introduced for MOF synthesis.

We also note that the majority of the synthesis route records are from articles published in the last 10 years, this reflects the rapidly increasing interest and investment in MOF compounds, and in alternatives to the solvothermal synthesis method. In fact, 6,033 (62.2%) of the total synthesis route records may be classified as alternatives to solvothermal synthesis which reflects greater interest in developing alternative synthesis routes, particularly when considering that high solvent-use is inhibiting MOF scalability. Rapid increases can be observed for more novel synthesis routes, with an overwhelming majority of solvent-free synthesis papers published after 2010 (76% microwave-assisted, 95% sonochemical, 86% mechanochemical, and 88% liquid-assisted grinding). There is also likely to be some crossover between these methods as liquid-assisted grinding and sonochemical methods are themselves subsets of mechanochemical methods, and may be used in various combinations for MOF synthesis. This trend of utilising greener synthesis methods is also reflected in innovative MOF commercialisation efforts such as the ton-scale water-based processes that BASF has developed[36] and the mechanochemical process from MOF Technologies[37].

The DigiMOF database allows users to search for potentially scalable MOFs via synthesis method, and discover MOFs that can be more easily synthesised and tested with the equipment and resources available to them. In the future, an alternative web search query method of database assembly could be used in place of the CSD reference code method to assemble a corpus using queries such as 'solvent-free MOF synthesis' or 'mechanochemical MOF synthesis', expanding the database to include more MOFs that can be produced using alternative synthesis methods, and novel synthesis techniques for MOFs already logged in the database with more conventional synthesis routes. The synthesis method parser should be continually updated to allow it to parse novel synthesis methods and procedures as and when they become more prominent in MOF literature, and may be extended to parse for post-synthetic methods such as linker substitution.

5.2 Topology: Topological characterisation of MOFs is important as it can constrain key structural properties such as the pore shape, size and chemistry, and it is directly related to mechanical properties[34]. **Figure 4b**. shows the distribution of topologies identified in the CSD MOF subset: we extracted 112 unique topologies, across a total of 6680 results. The most frequently occurring topology was **pcu** with 946 hits, followed by **sql** and **dia** with 822 and 482 counts, respectively. In some publications, the parsers picked up variations of certain topologies e.g. **sql**, **44-sql**, **(4,4)-sql**, and **(44)-sql** as separate entries. From the top ten topologies shown in **Figure 4b**., **sql**, **hcb**, and **kgd** are 2-periodic and the remaining seven exhibit 3-periodic frameworks. The supporting information provides a full list of MOF names and topologies identified. We anticipate that this topological characterisation of the MOFs will guide future efforts to identify mechanically stable MOFs.

5.3 Solvent: As shown in Figure 4c, DMF is the most frequently extracted solvent by a considerable margin, representing 469 of the 1,211 extracted solvents. Water is the second most frequently extracted solvent with 186 counts for which 127 were paired with hydrothermal synthesis routes. The remainder of the water solvent records were merged with solvothermal or hydro(solvo)thermal synthesis routes, which could reflect the common use of solvent mixtures containing multiple reagents such as DMF, water, and ethanol. The parser does not have the capability to extract lists or mixtures of solvents unless they appear consecutively in a string without whitespace e.g. 'DMF/H_2O'.

The presence of organic solvents such as DMF, DMA, ethanol, and acetonitrile demonstrates that despite increased research into alternative synthetic pathways, many existing synthetic procedures are still reliant on organic solvents and failure to eliminate large volumes of such solvents in MOF synthesis is one of the largest barriers to MOF commercialisation. It should be noted that whilst the CSD includes solvent information, most of these records are missing from the database. These parsers offer the ability to search for MOF synthesis routes associated with a given solvent, thereby allowing researchers to limit screening to hydrothermal synthesis or to solvothermal synthesis techniques with cheaper, less toxic, or more readily recoverable solvents.

5.4 Organic Linkers: Histograms in **Figure 4d.** shows that carboxylate-type linkers were the most frequently extracted type of organic linker, with over 432 associated records. Specific carboxylate linkers e.g. benzene dicarboxylate acid (BDC) were not extracted more frequently because these linkers are more generically referred to as carboxylate or dicarboxylate without specification of the exact structure. Other challenges with NLP parsing of MOF linkers in the literature were inconsistencies in linker abbreviations and naming conventions. For example, 'bpy' and 'bipy' are used to denote specific bipyridine type linkers such as 2,2-bipyridine and 4,4-bipyridine [38,39].

Whilst researchers may be referring to specific linkers when using these abbreviations, these labels are not consistently used to refer to any one distinct structure. Records for 'bpy' and 'bipy' were merged as 'bipy' to denote generic bipyridine type linkers. Following data cleaning where instances of '4,4-bipyridine', '4,4-bipy' and '4,4-bpy' were merged as '4,4-bipy', 273 records were associated with '4,4-bipy', and 267 with 'bipy' representing the 2nd and 3rd most extracted linkers, respectively. Similar cleaning was conducted for 2,2-byripidine linkers with 109 records. Carboxylate (H3BTC, BDC, carboxylate, dicarboxylate) and pyridyl type linkers (4,4-bipy, 2,2-bipy, bipy, bpe and bpp) were the most dominant linker types extracted by the parsers. Other notable linkers included imidazole type bridging ligands such as 'bimb' (phenylenebis(methylene)bis(1H-imidazole)). 'H2L' was the 4th most extracted linker with 251 associated records. This refers not to a specific chemical structure, it is instead a generic label used within MOF literature to refer to a number or organic linkers[40]. This means that the linker chemical formulae may be explicitly named in one part of the text and then simply be referred to as L, posing considerable challenges for NLP parsing. In some instances, researchers do not elaborate on the chemical formula of the linker within any part of the text and use generic L-type notation or refer to the general structure (e.g. carboxylate). The usage of generic labels and general compound class names may reflect increased trends towards more complex and functionalised linkers in MOF synthesis, which may make consistent identification and naming of these structures more challenging[41].

5.5 Metal Precursor: The choice of metal precursors is also important for MOF synthesis; certain metal clusters such as metal oxides can provide cost-effective and flexible MOF production routes as well as control over structural topology and shape. Our parser extracted many metal precursors in the form of a metal element, ion name, or symbol: this is shown in **Figure 4e.** Zinc-based precursors were most frequently extracted, with ‘$Zn(NO_3)_2$·6H2O’ representing 365 of the merged records. Zinc salts represented three of the most extracted metal precursors accounting for 36% of the 1481 records. This is unsurprising given the prevalence and popularity of zinc-based MOFs; however, an absence of zirconium salts from the top 10 metal precursors is unexpected. One reason for the lack of zirconium salts is that papers discuss zirconium precursors as “Zr”, as can be seen by 212 hits in the database for “Zr”, shown in **Figure S3**. Additionally, compared to zinc and copper based MOFs, Zr-based MOFs were not widely produced until after 2012[42]. The second most frequently extracted metal salt was ‘$Cd(NO_3)_2 \cdot 4H_2O$’ with 177 merged records followed by nitrate salts of Zn, Co, and Cu. The ability to cross-reference MOF structures with their metal precursors from proven synthesis procedures will allow MOF scientists to rapidly screen structures for criteria such as metal nodes or precursors associated with desirable properties, greater material abundances and lower costs. Searching by metal precursor will also provide valuable insight into MOF building blocks in cases where records include MOF names which are not directly based on the MOF structure or formula.

5.6 Temperature: The CSD database contains temperature entries for almost all deposited structures, when DOI records were extracted from the CSD Python API, it was also possible to extract corresponding temperature records. The results of these extractions can be seen in **Figure 4f** — it is important to note that these values are not the synthesis temperatures of the materials but the variable-temperature crystallographic studies. Typically, the structures are tested and reported at or around room temperature explaining the spike in records at 293 K. It is also common for a cryostream or other device to be used to cool a sample for low-temperature crystallographic testing.

6. Conclusions and Future Directions

To the best of our knowledge, DigiMOF database is the first automatically generated database of MOF synthesis properties which uses ChemDataExtractor to text mine MOF literature. After an iterative training process, the parsers yielded an overall precision of 77%. DigiMOF will allow researchers to search for key MOF properties related to their large-scale production e.g. synthesis routes and solvents used, organic linkers, metal precursors, and structure topology. We envisage DigiMOF as an invaluable tool to both MOF scientists conducting high-throughput computational screening and experimentalists evaluating MOF properties empirically. The software and the parsers developed here are open-source, which allows researchers to update regular expressions as new compounds emerge to ensure the algorithms are able to identify new MOF-property relationships. This means that with minimal additional effort, researchers can employ the modified CDE scripts to generate their own database; with more focused search queries to study alternative MOF production pathways with very basic alterations to the parsers. The ability to cross-reference and merge data using DOIs allows researchers to readily merge or expand this database to include other properties which pique their interest.

DigiMOF is mainly focused on the production of MOF compounds; future databases should aim to compile properties that are important to both the production and application of MOFs. Additional parsers can be developed to extract properties related to scalability and synthesis, such as the reaction temperature, space-time yield, heat of adsorption, reaction time, and regeneration time. Additionally, structural information related to MOF families, functional groups, and SBUs could be integrated into the database to observe any potential structure-scalability relationships. For example, the incorporation of parsers to extract secondary building units in addition to metal salt precursors could further elucidate the structure and nature of materials in the merged records.

We recommend that future MOF synthesis publications contain specifically formatted tables of key information, presented in a way that is friendly to text mining algorithms to enable the scraping of data using a high throughput screening approach. Not only would this improve the precision and recall of structure property parsing, it has the potential to enable an extremely accurate and reliable database of synthesis data to be created in the public domain.

The aim of this work has been to lay a foundation for enabling digital manufacturing of MOFs and facilitate identification of commercially-viable MOF production pathways. With over 15,000 unique MOF records, this data can be used to further assess the viability of alternative MOF synthesis routes and to drive further techno-economic assessment, life-cycle assessment, and experimental validation work. DigiMOF could therefore help to reduce the overdependence within the MOF community on unsustainable synthesis routes which currently precludes the application of these structures in decarbonisation technologies that motivate many contemporary MOF research proposals. DigiMOF should encourage and augment MOF scientists' expertise, allowing them to design more efficient MOF discovery pathways and advance the synthesis of these fascinating materials.

Supporting Information.

The following files are available free of charge.

Supporting Information (PDF)

CSD and ACS merged data (CSV)

5D visualization platform on Wiz

All of the graphs presented in this paper can be reproduced online at https://wiz.shef.ac.uk/. Furthermore, visitors to the site can explore the entire data set interactively, with all the variables plotted on each of the five axes according to the user's interest. MOFs can be searched for and filtered either by name or by property, or by selecting them from the graph[43].

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. ‡These authors contributed equally.

Acknowledgements

P. Z. M. acknowledges funding from the Engineering and Physical Science Research Council (EPSRC) and the Cambridge Crystallographic Data Centre (CCDC) for the provision of PhD studentship funding to L. T. G.

References

(1) Li, B.; Wen, H.-M.; Zhou, W.; Chen, B. Porous Metal–Organic Frameworks for Gas Storage and Separation: What, How, and Why? *J. Phys. Chem. Lett.* **2014**, *5* (20), 3468–3479. https://doi.org/10.1021/jz501586e.

(2) Farha, O. K.; Yazaydın, A. Ö.; Eryazici, I.; Malliakas, C. D.; Hauser, B. G.; Kanatzidis, M. G.; Nguyen, S. T.; Snurr, R. Q.; Hupp, J. T. De Novo Synthesis of a Metal-Organic Framework Material Featuring Ultrahigh Surface Area and Gas Storage Capacities. *Nat Chem* **2010**, *2* (11), 944–948. https://doi.org/10.1038/nchem.834.

(3) Ma, S.; Zhou, H.-C. Gas Storage in Porous Metal–Organic Frameworks for Clean Energy Applications. *Chem. Commun.* **2010**, *46* (1), 44–53. https://doi.org/10.1039/B916295J.

(4) Mason, J. A.; Veenstra, M.; Long, J. R. Evaluating Metal–Organic Frameworks for Natural Gas Storage. *Chem. Sci.* **2013**, *5* (1), 32–51. https://doi.org/10.1039/C3SC52633J.

(5) Miller, S. E.; Teplensky, M. H.; Moghadam, P. Z.; Fairen-Jimenez, D. Metal-Organic Frameworks as Biosensors for Luminescence-Based Detection and Imaging. *Interface Focus* **2016**, *6* (4), 20160027. https://doi.org/10.1098/rsfs.2016.0027.

(6) Kreno, L. E.; Leong, K.; Farha, O. K.; Allendorf, M.; Van Duyne, R. P.; Hupp, J. T. Metal-Organic Framework Materials as Chemical Sensors. *Chem Rev* **2012**, *112* (2), 1105–1125. https://doi.org/10.1021/cr200324t.

(7) Li, J.-R.; Sculley, J.; Zhou, H.-C. Metal–Organic Frameworks for Separations. *Chem. Rev.* **2012**, *112* (2), 869–932. https://doi.org/10.1021/cr200190s.

(8) Hiraide, S.; Sakanaka, Y.; Kajiro, H.; Kawaguchi, S.; Miyahara, M. T.; Tanaka, H. High-Throughput Gas Separation by Flexible Metal–Organic Frameworks with Fast Gating and Thermal Management Capabilities. *Nat Commun* **2020**, *11* (1), 3867. https://doi.org/10.1038/s41467-020-17625-3.

(9) Teplensky, M. H.; Fantham, M.; Li, P.; Wang, T. C.; Mehta, J. P.; Young, L. J.; Moghadam, P. Z.; Hupp, J. T.; Farha, O. K.; Kaminski, C. F.; Fairen-Jimenez, D. Temperature Treatment of Highly Porous Zirconium-Containing Metal–Organic Frameworks Extends Drug Delivery Release. *J. Am. Chem. Soc.* **2017**, *139* (22), 7522–7532. https://doi.org/10.1021/jacs.7b01451.

(10) Abánades Lázaro, I.; Haddad, S.; Sacca, S.; Orellana-Tavra, C.; Fairen-Jimenez, D.; Forgan, R. S. Selective Surface PEGylation of UiO-66 Nanoparticles for Enhanced Stability, Cell Uptake, and PH-Responsive Drug Delivery. *Chem* **2017**, *2* (4), 561–578. https://doi.org/10.1016/j.chempr.2017.02.005.

(11) Lawson, H. D.; Walton, S. P.; Chan, C. Metal–Organic Frameworks for Drug Delivery: A Design Perspective. *ACS Appl. Mater. Interfaces* **2021**, *13* (6), 7004–7020. https://doi.org/10.1021/acsami.1c01089.

(12) Yoon, M.; Srirambalaji, R.; Kim, K. Homochiral Metal–Organic Frameworks for Asymmetric Heterogeneous Catalysis. *Chem. Rev.* **2012**, *112* (2), 1196–1231. https://doi.org/10.1021/cr2003147.

(13) Corma, A.; García, H.; Llabrés i Xamena, F. X. Engineering Metal Organic Frameworks for Heterogeneous Catalysis. *Chem Rev* **2010**, *110* (8), 4606–4655. https://doi.org/10.1021/cr9003924.

(14) Ma, L.; Abney, C.; Lin, W. Enantioselective Catalysis with Homochiral Metal–Organic Frameworks. *Chem. Soc. Rev.* **2009**, *38* (5), 1248–1256. https://doi.org/10.1039/B807083K.

(15) Pascanu, V.; González Miera, G.; Inge, A. K.; Martín-Matute, B. Metal–Organic Frameworks as Catalysts for Organic Synthesis: A Critical Perspective. *J. Am. Chem. Soc.* **2019**, *141* (18), 7223–7234. https://doi.org/10.1021/jacs.9b00733.

(16) Shen, Y.; Pan, T.; Wang, L.; Ren, Z.; Zhang, W.; Huo, F. Programmable Logic in Metal–Organic Frameworks for Catalysis. *Advanced Materials* **2021**, *33* (46), 2007442. https://doi.org/10.1002/adma.202007442.

(17) Z. Moghadam, P.; Li, A.; Liu, X.-W.; Bueno-Perez, R.; Wang, S.-D.; B. Wiggin, S.; A. Wood, P.; Fairen-Jimenez, D. Targeted Classification of Metal–Organic Frameworks in the Cambridge Structural Database (CSD). *Chemical Science* **2020**, *11* (32), 8373–8387. https://doi.org/10.1039/D0SC01297A.

(18) Butt, T. E.; Javadi, A. A.; Nunns, M. A.; Beal, C. D. Development of a Conceptual Framework of Holistic Risk Assessment — Landfill as a Particular Type of Contaminated Land. *Science of The Total Environment* **2016**, *569–570*, 815–829. https://doi.org/10.1016/j.scitotenv.2016.04.152.

(19) Moghadam, P. Z.; Islamoglu, T.; Goswami, S.; Exley, J.; Fantham, M.; Kaminski, C. F.; Snurr, R. Q.; Farha, O. K.; Fairen-Jimenez, D. Computer-Aided Discovery of a Metal–Organic Framework with Superior Oxygen Uptake. *Nat Commun* **2018**, *9* (1), 1378. https://doi.org/10.1038/s41467-018-03892-8.

(20) Moghadam, P. Z.; Fairen-Jimenez, D.; Snurr, R. Q. Efficient Identification of Hydrophobic MOFs: Application in the Capture of Toxic Industrial Chemicals. *J. Mater. Chem. A* **2015**, *4* (2), 529–536. https://doi.org/10.1039/C5TA06472D.

(21) Wilmer, C. E.; Farha, O. K.; Bae, Y.-S.; Hupp, J. T.; Snurr, R. Q. Structure–Property Relationships of Porous Materials for Carbon Dioxide Separation and Capture. *Energy Environ. Sci.* **2012**, *5* (12), 9849–9856. https://doi.org/10.1039/C2EE23201D.

(22) Rampal, N.; Ajenifuja, A.; Tao, A.; Balzer, C.; S. Cummings, M.; Evans, A.; Bueno-Perez, R.; J. Law, D.; W. Bolton, L.; Petit, C.; Siperstein, F.; P. Attfield, M.; Jobson, M.; Z. Moghadam, P.; Fairen-

Jimenez, D. The Development of a Comprehensive Toolbox Based on Multi-Level, High-Throughput Screening of MOFs for CO/N 2 Separations. *Chemical Science* **2021**, *12* (36), 12068–12081. https://doi.org/10.1039/D1SC01588E.

(23) Rogacka, J.; Seremak, A.; Luna-Triguero, A.; Formalik, F.; Matito-Martos, I.; Firlej, L.; Calero, S.; Kuchta, B. High-Throughput Screening of Metal – Organic Frameworks for CO2 and CH4 Separation in the Presence of Water. *Chemical Engineering Journal* **2021**, *403*, 126392. https://doi.org/10.1016/j.cej.2020.126392.

(24) Avci, G.; Velioglu, S.; Keskin, S. High-Throughput Screening of MOF Adsorbents and Membranes for H2 Purification and CO2 Capture. *ACS Appl. Mater. Interfaces* **2018**, *10* (39), 33693–33706. https://doi.org/10.1021/acsami.8b12746.

(25) Halder, P.; Singh, J. K. High-Throughput Screening of Metal–Organic Frameworks for Ethane–Ethylene Separation Using the Machine Learning Technique. *Energy Fuels* **2020**, *34* (11), 14591–14597. https://doi.org/10.1021/acs.energyfuels.0c03063.

(26) Swain, M. C.; Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **2016**, *56* (10), 1894–1904. https://doi.org/10.1021/acs.jcim.6b00207.

(27) Cole, J. M. How the Shape of Chemical Data Can Enable Data-Driven Materials Discovery. *Trends in Chemistry* **2021**, *3* (2), 111–119. https://doi.org/10.1016/j.trechm.2020.12.003.

(28) Park, S.; Kim, B.; Choi, S.; Boyd, P. G.; Smit, B.; Kim, J. Text Mining Metal–Organic Framework Papers. *J. Chem. Inf. Model.* **2018**, *58* (2), 244–251. https://doi.org/10.1021/acs.jcim.7b00608.

(29) Luo, Y.; Bag, S.; Zaremba, O.; Cierpka, A.; Andreo, J.; Wuttke, S.; Friederich, P.; Tsotsalas, M. MOF Synthesis Prediction Enabled by Automatic Data Mining and Machine Learning. *Angewandte Chemie International Edition n/a* (n/a). https://doi.org/10.1002/anie.202200242.

(30) Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S.; Slater, B.; Siepmann, J. I.; Sholl, D. S.; Snurr, R. Q. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data* **2019**, *64* (12), 5985–5998. https://doi.org/10.1021/acs.jced.9b00835.

(31) Hawizy, L.; Jessop, D. M.; Adams, N.; Murray-Rust, P. ChemicalTagger: A Tool for Semantic Text-Mining in Chemistry. *Journal of Cheminformatics* **2011**, *3* (1), 17. https://doi.org/10.1186/1758-2946-3-17.

(32) DeSantis, D.; Mason, J. A.; James, B. D.; Houchins, C.; Long, J. R.; Veenstra, M. Techno-Economic Analysis of Metal–Organic Frameworks for Hydrogen and Natural Gas Storage. *Energy Fuels* **2017**, *31* (2), 2024–2032. https://doi.org/10.1021/acs.energyfuels.6b02510.

(33) Luo, H.; Cheng, F.; Huelsenbeck, L.; Smith, N. Comparison between Conventional Solvothermal and Aqueous Solution-Based Production of UiO-66-NH2: Life Cycle Assessment, Techno-Economic Assessment, and Implications for CO2 Capture and Storage. *Journal of Environmental Chemical Engineering* **2021**, *9* (2), 105159. https://doi.org/10.1016/j.jece.2021.105159.

(34) Moghadam, P. Z.; Rogge, S. M. J.; Li, A.; Chow, C.-M.; Wieme, J.; Moharrami, N.; Aragones-Anglada, M.; Conduit, G.; Gomez-Gualdron, D. A.; Van Speybroeck, V.; Fairen-Jimenez, D. Structure-Mechanical Stability Relations of Metal-Organic Frameworks via Machine Learning. *Matter* **2019**, *1* (1), 219–234. https://doi.org/10.1016/j.matt.2019.03.002.

(35) Huang, S.; Cole, J. M. A Database of Battery Materials Auto-Generated Using ChemDataExtractor. *Sci Data* **2020**, *7* (1), 260. https://doi.org/10.1038/s41597-020-00602-2.

(36) Czaja, A. U.; Trukhan, N.; Müller, U. Industrial Applications of Metal–Organic Frameworks. *Chem. Soc. Rev.* **2009**, *38* (5), 1284–1293. https://doi.org/10.1039/B804680H.

(37) James, S. L.; Lazuen-Garay, A.; Pichon, A. Use of Grinding in Chemical Synthesis. WO2007023295A3, May 18, 2007.

(38) Fei, H.; Cohen, S. M. A Robust, Catalytic Metal–Organic Framework with Open 2,2′-Bipyridine Sites. *Chem. Commun.* **2014**, *50* (37), 4810–4812. https://doi.org/10.1039/C4CC01607F.

(39) Lu, J. Y.; Cabrera, B. R.; Wang, R.-J.; Li, J. Cu-X-Bpy (X = Cl, Br; Bpy = 4,4‘-Bipyridine) Coordination Polymers: The Stoichiometric Control and Structural Relations of [Cu2X2(Bpy)] and [CuBr(Bpy)]. *Inorg. Chem.* **1999**, *38* (20), 4608–4611. https://doi.org/10.1021/ic990536p.

(40) Tansell, A. J.; Jones, C. L.; Easun, T. L. MOF the Beaten Track: Unusual Structures and Uncommon Applications of Metal–Organic Frameworks. *Chemistry Central Journal* **2017**, *11* (1), 100. https://doi.org/10.1186/s13065-017-0330-0.

(41) Schukraft, G. E. M.; Ayala, S.; Dick, B. L.; Cohen, S. M. Isoreticular Expansion of PolyMOFs Achieves High Surface Area Materials. *Chem. Commun.* **2017**, *53* (77), 10684–10687. https://doi.org/10.1039/C7CC04222A.

(42) Bai, Y.; Dou, Y.; Xie, L.-H.; Rutledge, W.; Li, J.-R.; Zhou, H.-C. Zr-Based Metal–Organic Frameworks: Design, Synthesis, Structure, and Applications. *Chem. Soc. Rev.* **2016**, *45* (8), 2327–2367. https://doi.org/10.1039/C5CS00837A.

(43) Balzer, C.; Oktavian, R.; Zandi, M.; Fairen-Jimenez, D.; Moghadam, P. Z. Wiz: A Web-Based Tool for Interactive Visualization of Big Data. *Patterns* **2020**, *1* (8), 100107. https://doi.org/10.1016/j.patter.2020.100107.