Unveiling the Synthesis Patterns of Nanomaterials: A Text Mining and Meta-Analysis Approach with ZIF-8 as a Case Study

Joseph R.H. Manning[*a] and Lev Sarkisov[*a]

(a) Department of Chemical Engineering, University of Manchester, UK, M13 9PL
Joseph.Manning@manchester.ac.uk
Lev.Sarkisov@manchester.ac.uk

Abstract

With the continuously growing number of scientific articles on synthesis of nanomaterials, it becomes impossible for researchers to grasp and comprehend the landscape of synthetic protocols available for a particular material. The aim of this study is to explore the feasibility of extracting the collective knowledge on synthesis of a particular material accumulated over the years from the published corpus of articles and organizing it in a systematic manner. Accordingly, we developed methods to perform detailed text mining on a single nanomaterial target for the purposes of methodology optimisation. Taking the common material ZIF-8 as a case study, we analysed 1600 synthesis protocols to identify trends in parameters, such as reagents, concentrations, and reaction time/temperature. We used this information to find the distribution of synthesis parameters and their relationships to one another, identifying the limits of common reaction parameters and revealing subtle details, such as insolubility of metal acetate reagents in alcoholic solvents, or the occurrence of amorphous oxides at low stoichiometric ratios. We then clustered similar synthesis protocols together, using their relative popularity to identify promising regions of the synthesis phase space for optimisation, reducing the need for brute force synthesis optimisation. The techniques developed here are a general tool accelerating the synthesis development of a wide range of nanomaterials by aggregating existing research trends, averting the need for laborious manual comparison of existing synthesis protocols or repetition of previously-developed techniques.

Introduction

The number of chemical syntheses reported is large and growing exponentially.[1] While naturally indicative of greater scientific progress, this leads to two significant challenges. Firstly, researchers are confronted with the growing difficulty of maintaining a comprehensive overview and understanding of the diverse landscape of synthetic routes and conditions accessible for a particular group of compounds. Secondly, although the repository of published synthesis data contains an immense wealth of information, its potential for systematic development of new synthesis protocols remains largely untapped and underutilized. In response to this, various informatics approaches have been adopted to standardise the data produced during chemical research. For example, the creation of chemical synthesis ontologies[2–4] and automated reactionware[5,6] has enabled new procedures to be directly compared against previously-published data or shared openly through chemical "programming languages".[7,8] However, the nature of reporting synthesis protocols – as unformatted prose in a written report – has remained largely unchanged.

As a result, most new publications and the entire body of prior chemical synthesis reports remains unlabelled, with the potential for far broader data mining and informatics research if these reports could be standardised. Accordingly, with the advent of text mining methods and natural language processing (NLP),[9] software has been developed to interpret chemical details from the plain text within chemistry publications[10,11] including compound structure,[12] reaction stoichiometry,[13] and performance.[14] Using these tools, large databases of organic[14,15] and inorganic[16–19] chemicals and reactions have been developed and used for novel materials discovery. For example, Cole and co-

workers created a database of organic dyes to identify ideal mixtures for broad-spectrum light absorption in dye-sensitized solar cells, regardless of the intension of the original studies.[15] Similar strategies have been used by Olivetti and co-workers to analyse how synthesis gel composition and organic structure directing agent (OSDA) can dictate crystal polymorphs for a range of zeolite syntheses.[16]

One weakness of these text mining approaches is their reliance on unambiguous identification of the chemical entities in question, using named-entity recognition (NER)[9,20] and the programmatic naming conventions defined by IUPAC[21] to succeed. In the absence of such well-accepted naming schemes – as is the case for a variety of emerging nanomaterial families like porous silicas, polymers of intrinsic microporosity, and covalent organic framework materials – large scale data mining becomes far less practical. An excellent example of this is metal-organic framework (MOF) materials - infinite condensation polymers of various organic ligands and metal ions or clusters. There are millions of possible MOFs,[22–25] and hundreds of thousands of frameworks already synthesized,[26–29] necessitating data-driven approaches to accelerate progress in the field. However, unambiguous naming conventions for MOFs have yet to be fully adopted,[30] frustrating text-mining of the primary publications themselves. Instead, informatics methods have largely been driven by the creation of a subset of the Cambridge Structural Database (CSD)[31] focused on MOF materials,[28] as these resources allow researchers to analyse the full range of experimentally known MOF structures, identifying the best experimentally-realised materials for future research and development.

To accelerate development of experimental procedures to make MOFs, however, data-mining approaches must look beyond structure into the synthesis protocols leading to different frameworks. By understanding the relationships between protocol and eventual material, new synthesis methods can be digitally generated, obviating the need for arduous trial-and-error or intuition-based approaches.[6] To this end, large-scale post-hoc analyses of experimental MOF synthesis protocols have recently been developed.[32,33] These studies apply NLP to the underlying publications in the CSD MOF subset to interpret their synthesis protocols, identifying such details as solvents used, specific reagents, solvents, and reaction parameters. As a result, broad descriptive statistics about the synthesis strategies to produce MOFs have been developed,[33] and even predictive models to suggest synthesis parameters for novel MOF materials when given a hypothetical structure.[32]

While these approaches give an excellent overview of the field of MOFs in general, they are vulnerable to bias in the papers submitting to the CSD. As the database focuses of chemical structure rather than synthesis protocols, only 1-2 synthesis examples of each framework are included. Further, the synthesis protocols are generally submitted from initial studies reporting the discovery of a material, rather than exploring the full range of potential approaches to a single target, meaning that only a very vague understanding of any individual MOF can be generated with this approach. For example, while candidate solvents and reaction parameters can be suggested, other salient parameters such as reagent ratios, product isolation methods, and alternative synthesis strategies (e.g. hydrothermal or mechanochemical versus solvent crystallisation) cannot. Deeper insight into individual MOFs and the peculiarities of their synthesis protocols can be gained through targeted meta-analysis of studies focusing on that particular material,[34] enabling regression of product properties like defect density against synthesis details. However, challenges of manually comparing synthesis protocols against one another severely limit the scale of such meta-analyses, preventing their widespread use.

To address these issues, in this article we pose the following questions: can we leverage previously-developed chemistry text mining tools to analyse the synthesis protocols for a single target

nanomaterial? If so, can we develop methods to process the extracted information on a uniform basis, enabling like-for-like comparison regardless of original format? Finally, can we harness this information to accelerate synthesis refinement of the material e.g. by generating proposed synthesis conditions correlated to high material quality and yield?

As a case study, we consider ZIF-8, a commonly synthesized MOF material which has been extensively studied within the literature. ZIF-8 is constructed from a combination of zinc ions and 2-methylimidazole in the sodalite topology, held together with metal-amine bonds rather than the more common metal-carboxylate bonds, thus rendering the material both hydrophobic and water-stable.[35,36] Accordingly, ZIF-8 has garnered significant interest in the literature for applications including gas storage and separation, adsorptive refrigeration,[37] biomolecule encapsulation,[38] catalysis,[39] and sensing.[40] Further, ZIF-8 can be synthesized from a number of strategies – for example using protic or aprotic solvents,[41] a range of temperatures,[42] reagent concentrations,[43] modulators and crystal growth modifiers,[44] and acid/base conditions.[45] In sum, over 7500 papers have been published regarding ZIF-8 to date. Given the breadth of synthesis protocols established for ZIF-8, it practically impossible to manually compare all possible synthesis methodologies to one another. Applying text mining methods to automatically and quantitively analyse ZIF-8 synthesis protocols would enable larger-scale analysis and the identification of promising synthesis strategies.

In this study we developed methods to extract and aggregate synthesis protocols in a uniform format. We studied 1600 synthesis protocols of ZIF-8 and related materials from 3197 original articles, performing an automated meta-analysis of the synthesis methods contained. We analysed the chemical identities used alongside quantities and reaction conditions to provide a systematic design space for ZIF-8, identifying key trends in the approaches used. Finally, we group similar synthesis protocols together with unsupervised clustering techniques, identifying hidden patterns in the data.

Methods development

The workflow of extracting and analysing synthesis protocols was split into four overarching steps: text collection, where a corpus of research papers is identified and downloaded; paragraph identification, where raw synthesis protocols are identified within the prose; grammar parsing, where the natural language is converted into hierarchical data for later interpretation; and synthesis protocol extraction, where the extracted data is standardised to produce a structured "recipe" for each synthesis protocol. Key steps in the workflow are depicted in Figure 1. The first three steps have been widely described elsewhere, and only a brief description is provided in this section (with associated code provided by the authors on GitHub at https://github.com/SarkisovTeam/SynOracle-preprocessing). The final stage of the workflow was developed in this study using python 3.9,[46] and is made freely available by the authors on GitHub at https://github.com/SarkisovTeam/SyntheticOracle.

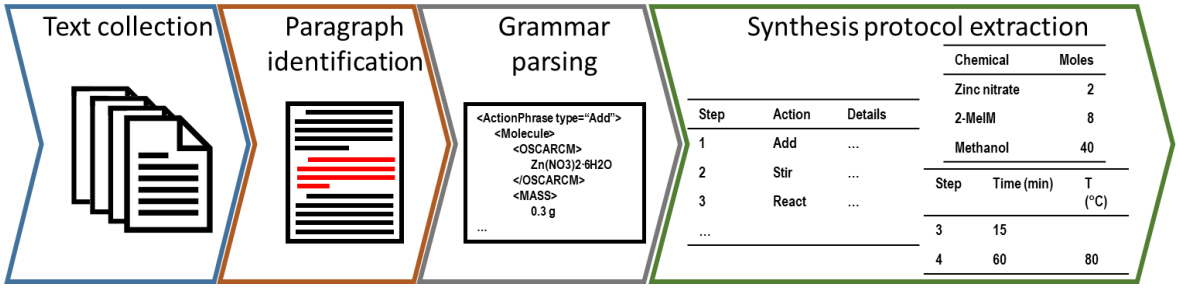

Figure 1 – Scheme of the data processing pipeline used in this study.

Text collection, paragraph identification, and grammar parsing

To produce a corpus of ZIF-8 synthesis protocols, we initially followed established methods to download collections of papers and identify synthesis protocols within them.[32,33] Synthesis papers were identified by searching the SCOPUS database using Elsevier's elsapy software (https://github.com/ElsevierDev/elsapy). Papers were identified using the search term "ZIF OR zeolitic imidazol* AND synthesis," returning 4198 results. These were then categorised by publisher, from which the three largest groups were targeted for downloading (ACS, RSC, and Elsevier), reducing the total corpus to 3179 papers. XML or HTML versions of each paper were then downloaded according to their publisher's specifications – using elsapy in the case of Elsevier, web scraping in the case of the RSC, and through the text and data mining service at the ACS.

Once downloaded, synthesis paragraphs were identified using ChemDataExtractor2.1[10] according to previously developed protocols for identifying MOF synthesis methods.[32,33] In this procedure, chemical named entity recognition was performed using BERT[47] to identify potential reagents, and part-of-speech (POS) tagging was carried out on the remaining tokens to interpret sentence grammar. Chemical quantities were identified from the POS tags as CD-NN bigrams (phrases consisting of a cardinal number followed by a noun), and regex matching of the noun against a library of SI units. Synthesis paragraphs were identified as containing three or more chemical named entities and three or more chemical quantities, after which each paragraph was extracted as plain text for manual confirmation and later analysis.

Once confirmed that each extracted paragraph contained a synthesis procedure, hierarchical grammar parsing was performed in the ChemicalTagger software[11] to associate chemical named entities with quantities and specific synthesis actions (termed *ActionPhrases*). These were stored as nested tags within an XML document.

Synthesis protocol extraction

To interpret and compare synthesis protocols against one another, data about synthesis steps, conditions, and chemicals involved had to be converted from nested XML data into useful information using the software developed in this study. To perform this, XML data extracted from ChemicalTagger was recursively parsed into strings within a pandas[48,49] DataFrame object such that each row consisted of a single *ActionPhrase*, its associated time and temperature, and details of any chemical entity involved.

Chemical identities were first confirmed by cross-referencing identified chemical names against the PubChem database[50] using the pubchempy python library (https://pubchempy.readthedocs.io/en/latest/index.html). From this, a unique identifier for each individual chemical was generated, enabling extraction of key information about each chemical and summation of identical chemicals together. To prevent semantically identical reagents from being considered separately (e.g. zinc nitrate and their hydrates), PubChem identifiers were supplemented with structural information gathered from the cheminformatics tool RDkit.[51] Specifically, chemicals whose formulae contained the elements zinc or cobalt, as well as the nitrate, acetate, sulfate, and imidazole substructures were separately identified.

Then, numerical quantities associated with each chemical were calculated. To do this, chemical quantities were categorised by type from the structured XML output of ChemicalTagger (e.g. by volume, moles, mass etc.), and parsed into physically meaningful units with the pint python library (https://pint.readthedocs.io/en/0.20.1/index.html). To prevent double-counting in situations where two units were mentioned, e.g. by the common phrase "5 g of [reagent] (0.8 mmol)," only a single

unit type was considered for each chemical entity according to the priority list (moles > mass > volume). These units were then converted into moles using the molecular mass identified from the PubChem identity. In the case of converting volume to moles, densities were estimated from the ChEDL database of critical point properties[52] using the COSTALD method.[53] Once chemical identities and quantities had been fully converted, these were aggregated into a single bill of materials for each synthesis (visualised in Table 1). Conditions (i.e. time and temperature values) were similarly parsed from strings into meaningful units using the pint python library, and stored as minutes and degrees Kelvin, respectively.

Table 1 – example of a synthesis protocol bill of materials taken from reference [54]

PubChem Identifier	Chemical name	Original quantities	Amount (millimoles)
12749	2-methylimidazole	0.24 g, 3.4 mmol	3.4
15865313	Zn(NO3)2.6H2O	0.956 g, 3.2 mmol	3.2
6212	Chloroform	40 mL	500
6228	DMF	70 mL	1210

Finally, to reduce semantically meaningless differences between different synthesis sequences, synthesis actions were grouped using a similar technique to the recently developed ULSA for inorganic nanomaterials syntheses.[55] Synthesis actions were categorised as either being related to "addition," "extraction," "reaction," or "other" (Table 2) and collocated steps of the same kind were grouped together. A fifth category, "start," was used to signify opening statements of synthesis protocols (e.g. "ZIF-8 was produced by our previously published method"), which would otherwise be miscategorised as an "extraction" or "other" action. "Start" actions were then excluded from further analysis.

Table 2 – Relationship between ChemicalTagger-identified ActionPhrase types and aggregated action types used here.

Action type	ActionPhrase
"addition"	Add, Dissolve, Stir
"reaction"	ApparatusAction, Synthesize, Wait
"extraction"	Degass, Dry, Extract, Filter, Partition, Precipitate, Purify, Quench, Recover, Remove, Yield
"other"	Concentrate, Cool, Heat

Grouping similar synthesis protocols together

To group synthesis protocols together, we related individual syntheses to one another by the identity of the reagents used only. To calculate the mathematical relationship between different synthesis protocols the list of chemicals was first vectorised, creating a numerical representation of the chemical combination used in each synthesis. Briefly, an $M \times N$ matrix was created, where *M* is the number of synthesis protocols, and *N* is the number of unique chemicals present across all of synthesis protocols studied. To reduce noise in the data, only synthesis protocols containing 2-

methylimidazole were considered, and metal sources were grouped by chemical substructures as described previously. In total, 139 unique chemicals were identified across 1134 synthesis protocols.

For each synthesis protocol, a vector was generated using the term frequency–inverse document frequency algorithm (TF-IDF), a commonly used text mining method to estimate the importance of words in a group of documents.[56] The TF-IDF algorithm weights the frequency of a word used in each document against its frequency across the group of documents – words present in many documents are given a low weight, while words occurring in only rarely are given a high weight. This is shown in Equation 1, which calculates the weight of word *t* in the individual document *d* as part of the group of documents *D*, where *f* is the frequency the word occurs. As in this study the "words" are chemical names, common chemicals like methanol are afforded a low weight, while rarer chemicals like CTAB are afforded a relatively higher weight.

$$tfidf(t,d,D) = f_{t,d} \cdot \log_{10}\left(\frac{1+n}{1+f_{t,D}}\right)$$

Once the chemical identities had been vectorised, similarity was calculated by the DBSCAN clustering method.[57] DBSCAN calculates the local density of data points in Euclidean space (synthesis protocols in the case of this study), defined as the number of neighbours closer than a threshold distance from each data point. Clusters are identified as disconnected regions containing a high density of data points, while isolated data points with no connection to a larger cluster as identified as noise.

To visualise the results of the clustering analysis, the high dimensional data were projected into two dimensions using the t-distributed stochastic neighbour embedding (t-SNE) method.[58] To do this the algorithm calculates the distances between each datapoint in high dimensional space, and estimates low-dimensional coordinates for each datapoints which preserves the distance between each point and its neighbours.

Results and discussion

Validation against manually-extracted information

To perform a quantitative meta-analysis of ZIF-8 synthesis, we first demonstrate the validity of the information extracted by comparing the performance of our text mining approach against a manually identified "ground truth" from a small number of papers sourced from the NIST database of emerging adsorbent materials. Using this database served two purposes: it was sufficiently small to provide a tractable number of articles for high-fidelity analysis, and each synthesis report was confirmed to contain ZIF-8 by the isotherm data provided. Overall, 44 publications describing ZIF-8 synthesis were identified, of which full information could be extracted for 42. The manuscripts were downloaded from their publisher, synthesis paragraphs manually identified, and synthesis information extracted both manually and using our software. In all cases, data reported within the paper and manually collated were considered as the ground truth.

From these paragraphs, three key parameters were extracted: a sequence of synthesis actions taken, a table of constituent chemicals, and the reaction conditions (i.e. temperatures and quoted times). For each parameter, the F1-score was calculated providing a numeric score for each text mining task compared against the manually-extracted ground truth. Extracted chemical identities were cross-referenced against the PubChem database of compounds to act as both a unique identifier and source of key information about each species. Finally, physical quantities – the values of time, temperature, and chemical quantity – were converted from plain text to numerical units

using the pint python library and compared against their manually extracted counterparts. These data are summarised in Table 3.

Table 3 – Parsing fidelity metrics as a percentage for manually-labelled quantities in the NIST ISODB corpus of ZIF-8 synthesis procedures.

Metric	*Precision*	*Recall*	*F1-score*	*Matching quantities*
Synthesis actions	79	53	63	-
Aggregated actions	94	94	93	-
Reagent identification	60.2	77.4	66.7	81.7
Temperature parsing	74.4	72.5	73.4	68.3
Time parsing	73.8	91.2	81.6	73.8

Individual synthesis actions were relatively poorly identified with text mining, with a F1-score of ca. 60%. This was primarily due to the low recall (i.e. true positive) rate of action parsing. Inspection of the synthesis paragraphs themselves showed that actions that were implicitly repeated, for example in the phrase "washed with water and methanol subsequently for 3 times",[59] were not captured by ChemicalTagger thereby leading to lower scores. Conversely, when synthesis actions were converted to their conceptual types and aggregated, the F1-score increased significantly to over 90% indicating that all synthesis stages were identified even if the specific *ActionPhrases* themselves were not. Therefore, we conclude that the text mining captures the essence of the synthesis protocol, but is unable to fully summarise the semantics of synthesis due to "linguistic noise" i.e. variability between different authors writing styles.

In terms of synthesis parameters, F1-scores and quantity matching were between 60-80% in all cases. These range of scores are slightly lower than previous text-mining efforts, which generally score between 60-98%.[1,60] We ascribe this relatively low score to more stringent criteria used in this study: as we define true positive to be the successful identification of a PubChem database entry, precision is lowered when cross-referencing fails. This is further exacerbated by the presence of typographical errors and colloquial chemical names which are not recognised by an automated PubChem database search (e.g. 2-methyli**n**idazole or 2-MeIM, rather than 2-methylimidazole). Failure to successfully convert numerical quantities similarly reduced the F1-score during time and temperature parsing.

In sum, while individual synthesis features could be reliably extracted using the methods developed here, it is currently impossible to reliably reproduce the entirety of any specific synthesis protocol. To achieve such high-fidelity reproduction, methods would have to be developed to estimate the completeness of a synthesis protocol, requiring a much larger set of manually-labelled synthesis sequences, similar to that developed by Wang *et al.* for individual synthesis actions.[55] Efforts to create such a dataset are ongoing in our research group. Instead, further analysis in this study is performed by compiling a group of similar synthesis protocols to extract a representative aggregate of synthesis details, hence enabling quantitative meta-analysis.

Interpreting ZIF-8 synthesis strategies

Given the effectiveness of our text mining methods to extract synthesis information from text, we progressed to a larger dataset of 3179 experimental synthesis reports of ZIF-8. From this dataset we processed 1600 synthesis protocols, enabling strong statistical analysis of the synthesis options which have been explored.

We first analysed the reagent compounds used during synthesis, which should consist of 2-methylimidazole and Zn salts only. As can be seen in Figure 2, this is not the case: while methylimidazole was by far the most common linker molecule mentioned (Figure 2A), 34% of the synthesis protocols mentioned cobalt salts. In fact, 32% of the synthesis protocols omitted zinc entirely, indicating that these were synthesis protocols of ZIF-67 instead – the cobalt equivalent of ZIF-8. The remaining cobalt-mentioning synthesis protocols also contained zinc, indicating that they may be mixed-metal systems. This ambiguity highlights some of the key nomenclature issues with MOF materials – ZIF-8 and -67 are practically the same material in terms of synthesis protocol but this proximity is not reflected in the common name. The use of tools such as MOFid[30] can avoid this linguistic ambiguity, even accurately describing the continuous transition between the two frameworks.

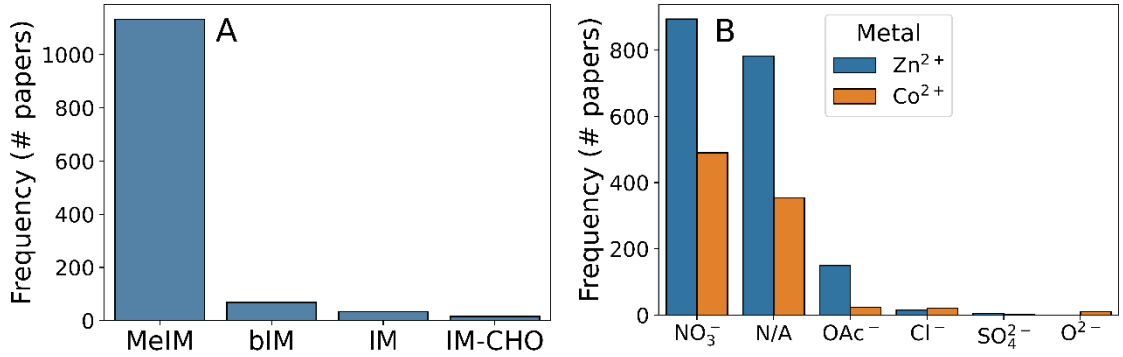

Figure 2 – Histograms of reagent compound frequency in ZIF-8 syntheses, broken down by (A) linker choice and (B) metal choice. Abbreviated chemical names refer to: MeIM – 2-methyilimidazole; bIM – 2-benzylimidazole; IM – imidazole; IM-CHO – imidazole-2-carbaldehyde

To further analyse the reagents used we grouped the metal salts used by anion type (Figure 2B), assuming that there was no consequence of using anhydrous versus hydrated salts. Nitrate was the most commonly used counterion, being present in 75% of syntheses. Ambiguous mentions of zinc and cobalt compounds were present in 17.2% of the 1600 protocols, encompassing minor zinc salts (e.g. $Zn(OH)_2$ in the case of reference 61), indirect reference to zinc precursors in synthesis (e.g. "The sample obtained with Zn"[62]), or mis-identified zinc compounds due to word tokenisation errors (e.g. "Firstly , 645 mg (2.469 mmol) of Zn $(NO3)_2$.$4H_2O$ was dissolved"[63], where the space character between "Zn" and its counterions causes incorrect chemical parsing). Aside from nitrates and ambiguous mentions, the only other commonly-mentioned metal salt was zinc acetate (present in 11.5% of synthesis protocols). The presence of chloride, acetate, and oxide precursors indicate that the synthesis is compatible to a range of electrolyte environments, agreeing with experimental reports which have shown that counterion choice significantly alters crystal nucleation and growth rates.[64,65] Despite the utility of these other salts, the overwhelming popularity of nitrate counterions found during our analysis indicates that other factors e.g. cost may have been prohibitive to their widespread adoption.

In addition to reagent identity, our text mining method provides information about the quantity of each reagent used, enabling analysis of synthesis protocol scale and reaction stoichiometry (Figure 3). The scale of ZIF-8 synthesis follows approximately a log-normal distribution, with 95% of

synthesis using 0.18-46 millimoles of metal ions and 0.73-330 millimoles of 2-methylimidazole (Figure 3A and B, respectively), demonstrating the flexibility of ZIF-8 synthesis with respect to scale. In terms of reaction stoichiometry, most synthesis protocols use an excess of linkers compared to the stoichiometric ratio of 2:1 (Figure 3C). This excess has been shown to control particle sizes by slowing the rate of crystal growth,[38,66–68] although few synthesis protocols use a higher ratio than 8:1. Interestingly, despite clear evidence that excess concentration of metal ions forms undesired by-products such as $Zn(OH)(NO_3)(H_2O)$,[43,68–70] 6% of the synthesis protocols analysed used a molar ratio of 1:1 or lower.

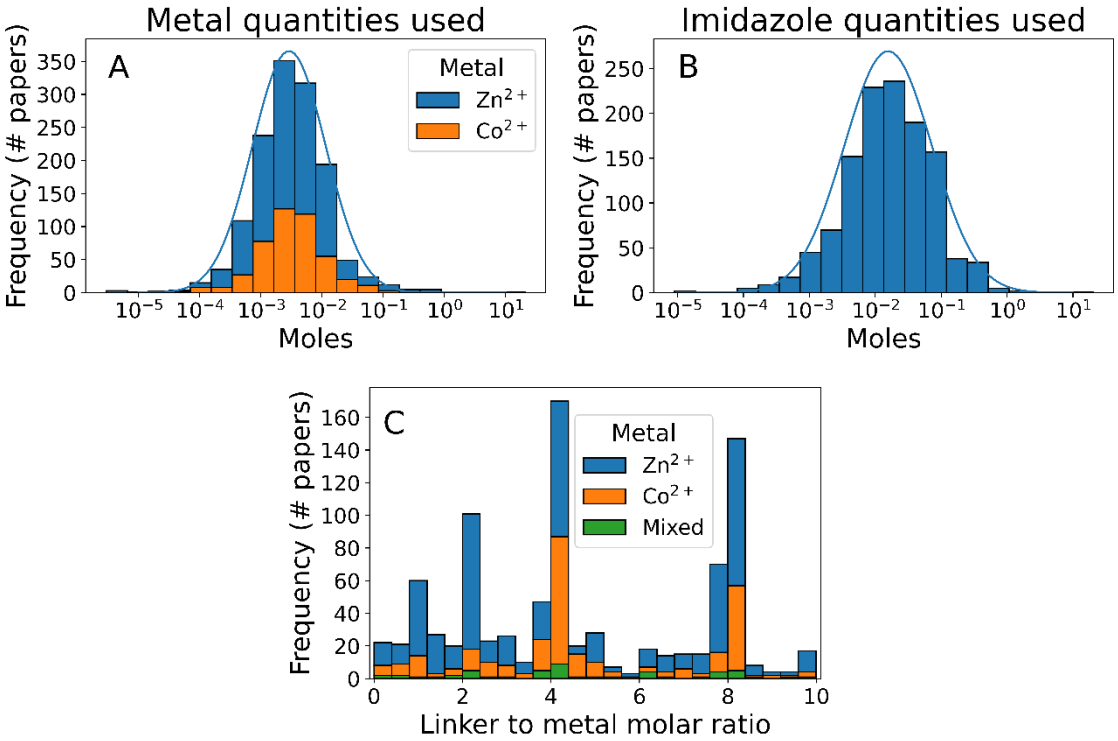

Figure 3 - Histograms of reagent quantities used. (A) metals, (B) linkers, and (C) metal/linker ratios broken down by synthesis metal. Where multiple variables are plotted in A and C, data bars are stacked on top of one another.

After considering reagents, the next most import aspect of a synthesis protocol lies in the choice of solvent environment for the reaction. Solvent choice has ramifications on the reaction mixture dielectric constant, in turn dictating factors such as reagent solubility and reaction kinetics. Further, the choice between protic and aprotic solvents, can accelerate reaction mechanisms relying on proton transfer, such as the linker deprotonation present during ZIF-8 synthesis.[66] Finally, overall reaction concentration is critical for determining whether the reaction mixture will act as an ideal solution, and in terms of the relative mass efficiency of the synthesis, both of which have consequences in terms of synthesis protocol viability in terms of scaleup to process-level manufacture.

The vast majority of synthesis protocols studied here contain one of methanol, ethanol, water, and DMF. Methanol was by far the most frequently mentioned solvent, present in 66% of synthesis protocols (Figure 4A), followed by water (40% of synthesis protocols), ethanol (27%), and finally DMF (12%). Less frequently used solvents included chloroform (1.4%), toluene (1.0%), and ethylene glycol (0.88%). To analyse the usage of each solvent present, we separated them by “synthesis” and “workup” procedure steps, as well as incorporating binary solvent mixtures (Figure 4B). This analysis revealed that, while ethanol was the third most prevalent solvent overall, it was the second most

common solvent used for washing and purification (and the fifth most common reaction solvent). Mixed solvent systems, primarily methanol-water, were present in 8% of syntheses presumably to tune the reaction dielectric and proton transfer catalysis rate.[71]

The distribution of solvent quantities used within the syntheses studied (Figure 4C) showed that each solvent followed approximately lognormal distributions. Both DMF and ethanol were used in smaller quantities than methanol or water (means of 0.4, 0.6, 1.4, and 1.6 moles per synthesis, respectively), indicating that the latter two solvents were more appropriate for scaling up the synthesis. Finally, we analysed the total solids concentration of synthesis protocols by dividing total reagent amounts by the solvent amounts used (Figure 4D). As with individual reagent concentrations, the total solids concentration followed an approximately log-normal distribution between 0.1-10 %mol. Separately, 7.7% of synthesis protocols had a solids loading of approximately 100 %mol - signifying mechanochemical synthesis protocols. Although mechanochemistry is a promising synthesis route due to its high yields[72] and low environmental impact[73] compared to conventional solvent synthesis methods, the relatively low popularity may be explained due to practical difficulties of mechanochemical synthesis e.g. prevention of hot-spot formation in the reaction vessel.[74]

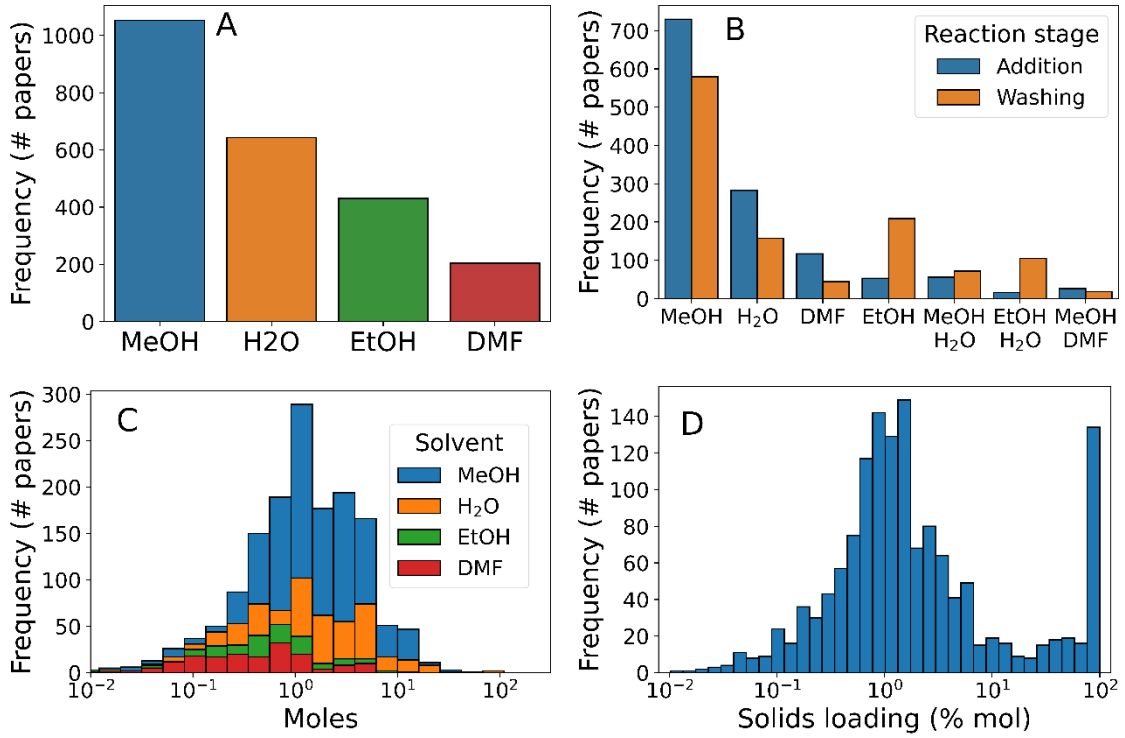

Figure 4 - histograms of solvent usage in ZIF-8 synthesis. (A) frequency of solvent mentions in all synthesis procedures, (B) frequency of solvent usage broken down by stage of the procedure, (C) quantity of solvent used, broken down by solvent type, and (D) total solids loading. Where multiple variables are plotted in C, data bars are stacked on top of one another.

In addition to reagents and solvents, ancillary chemicals such as surfactants, pH modifiers, and modulators are often key to ensure the success of MOF syntheses as well as dictating secondary particle characteristics such as size and crystal form. Three chemical types were prevalent within the synthesis protocols studied: acids, bases, and surfactant compounds. Unlike solvents and reagents, no individual ancillary chemical was identified in more than 3.5% of synthesis protocols (Figure 5). However, bases were present in 18% of all the synthesis protocols analysed, carrying out the important role of deprotonating the linker molecule in the reaction mixture. From the variety of

distinct molecules used for this role, it appears that no molecular recognition occurs, simply pH control. Despite the requirement for methylimidazole deprotonation for the reaction to progress, acids were detected in 6.3% of syntheses, however from inspection of the individual synthesis protocols acids only appeared during post-synthetic modification of the ZIF-8 materials e.g. after carbonisation[75] or impregnation into silicas.[76] Finally, surfactants like cetyltrimethylammonium bromide (CTAB) or sodium dodecylsulfate (SDS) were present in 4.6% of synthesis protocols, being used to slow the growth of individual ZIF-8 crystals and therefore control the particle shape.[59,77]

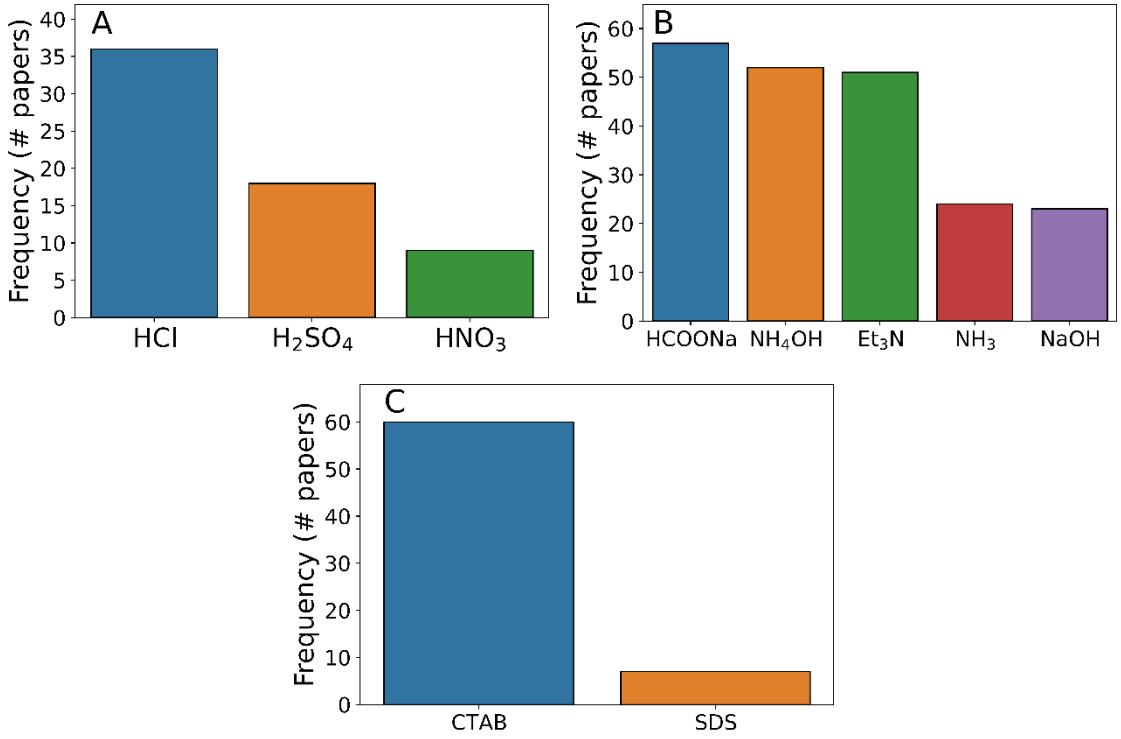

Figure 5 -Histograms of ancillary chemical prevalence in ZIF-8 synthesis. (A) acids, (B) bases, and (C) surfactants.

While it is possible to identify broad differences in synthesis strategy from feedstock compounds alone, it is impossible to understand why one chemical is chosen over another without further detail about the synthesis protocol being described. For example, the modulator sodium formate has been shown to perform different roles in room-temperature syntheses compared to hydrothermal alternatives.[44,78] In the first instance, we also consider the conditions (i.e. time and temperature) during the process. These are shown in Figure 6, demonstrating that the majority of protocols have synthesis times under six hours. Even after disregarding protocols with a reported synthesis time of 0 minutes as being spurious, it is clear that synthesis can be completed very quickly. In terms of synthesis temperature, the majority of the extracted temperatures were found to be room temperature indicating that thermal driving forces were not necessary for the formation of ZIF-8. This is further corroborated by the relative lack of procedures mentioning heated reaction conditions compared to heated drying conditions (Figure 6B).

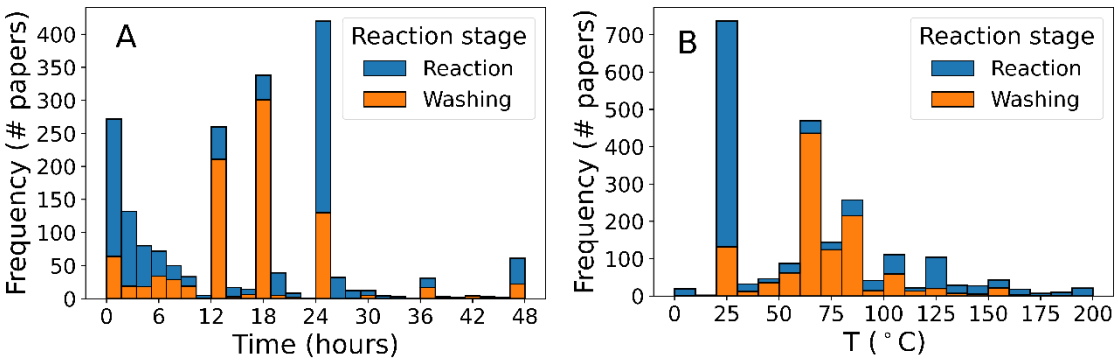

Figure 6 – Histograms of conditions during ZIF-8 synthesis processes. (A) total time elapsed and (B) temperatures used during synthesis. Annotational on (B) indicate the boiling points of the four most common solvents identified. Data are broken down by reaction step type as defined in Table 2. Where multiple variables are plotted, data bars are stacked on top of one another.

Overall, the tools developed in this study provide wide-ranging descriptive statistics of various ZIF-8 synthesis routes. The data generated are an excellent addition to existing literature review methods, facilitating the interpretation of different synthesis aspects e.g. reagent choices, stoichiometric ratios and reaction conditions. From these data we are able to identify gaps in the existing literature or synthesis conditions most likely to succeed, as well as providing useful input data for later technoeconomic analysis.

Harnessing synthesis information for accelerated methodology development

While the analysis performed is useful as a means of understanding the ZIF-8 reaction system, a key aim of this study was to systematize the collective synthesis knowledge for the material, thereby connecting synthesis protocols to some key performance indicators either of the synthesis (e.g. yield) or material (e.g. crystal form, surface area). One crucial barrier to this goal was the correlation of material performance data with synthesis protocol information: research papers are inconsistent in reporting of material properties (primarily as different quality metrics are used depending on the motivation of the original research), and the sample naming conventions used within research articles prevent unambiguous linking between the described protocols and materials produced. For example, while a synthesis paragraph might detail the synthesis of "*nano-sized ZIF-8*," later mentions in the text may be labelled differently e.g. "*ZIF-8*$_{nano}$,"[79] confounding attempts for automated identification of reaction products using regular expressions.[33] While this issue will undoubtably be resolved by the adoption of transformer-based language models such as BERT[80] and GPT-4,[81] such models became available only recently and the scientific community, including our group, is in the process of probing their extension to scientific data mining. In fact, the current study highlighted a number of issues with the current structure and completeness of reported synthetic protocols, understanding of which will be very helpful in engineering and fine-tuning GPT-based models.

As a result, the analysis performed in this study can only provide insight into how the MOF material is made rather than linking different synthesis features to specific outcomes like yield or quality. In the absence of such synthesis outcome information, we instead focus on how best to prepare the information gathered in this study for the generation of predictive models for ZIF-8 materials quality. A key challenge when attempting to optimise synthesis protocols either through systematic experimentation[5] or by training machine learning models[6] is the high dimensionality of the information contained in each synthesis. For example, 8 unique reagent chemicals were discussed in the previous section – 3 metal sources, 1 linker, and 4 solvents – meaning that to fully explore the 8-dimensional chemical space alone, N^8 experiments would be required (where N is the number of

quantity values tested for each variable). Even when limits on the complexity of each individual reaction are used – i.e. to contain a maximum of two metal salts and solvents – the dimensionality is only reduced to $12(N^5)$ experiments. While theoretically this dimensionality would scale with the number of synthesis steps used, we were unable to identify meaningfully distinct groups of synthesis actions (data not shown here, for brevity) and hence did not consider the sequence as impacting the synthesis outcome.

To enable faster and more efficient searching of the synthesis phase space, we used clustering to identify lower-dimensional sub-regions of the synthesis phase space which have been widely researched in experimental papers – essentially using a chemical combination's popularity as a proxy for its importance. The chemical identities used were encoded using TF-IDF vectorisation, then similar synthesis protocols were grouped by their density in the encoded space. The outcome of this clustering analysis is visualised using a 2-d projection in Figure 7 and summarised in Table 4, where the distance between points is indicative of each protocol's similarity to its neighbours. Eight clusters of reagent combinations were identified each containing 2-4 chemicals of a total of 6 reagents. We posit that these clusters represent well defined strategies to synthesize ZIF-8, which can be explored separately, therefore reducing the total amount of information required to explore these regions of the synthesis space.

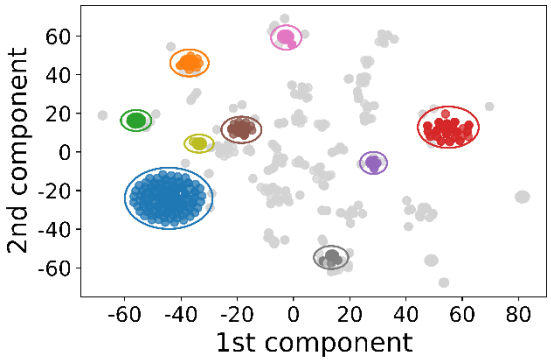

Figure 7 – 2-dimensional representation of the chemical combination space for ZIF-8 synthesis, generated using the t-SNE algorithm. Major synthesis pathways are identified using the DBSCAN clustering method and colour coded, while noise data is shown in light grey. Clusters are circled and described in Table 4.

Table 4 - Cluster labels and common features from Figure 7. N.B. all synthesis protocols included 2-methylimidazole, which was omitted for brevity.

Cluster number (colour)	*Common chemicals*	*Protocols in cluster*
1 (blue)	zinc, nitrate, methanol	225
2 (red)	cobalt, nitrate, methanol	147
3 (brown)	zinc, nitrate, water	50
4 (orange)	Zinc, nitrate	39
5 (green)	Zinc, cobalt, nitrate, methanol	31
6 (pink)	cobalt, nitrate, water	25
7 (purple)	Zinc, acetate, water	22
8 (olive)	Zinc, nitrate, methanol, water	20
9 (grey)	Zinc, nitrate, DMF	17

The well-defined synthesis strategies clustered in Figure 7 are notably different from the analysis performed in the previous section. In the first instance, ethanol was fully absent signifying its insignificance as a reaction solvent and matching the earlier analyses. Separately, acetate salts are only identified in one cluster and only associated with water. This association is due to the lack of solubility of zinc acetate in methanol (ca. 15 g/L cf. 430 g/L in water), information which can only otherwise be gained by specific knowledge of the chemistry of zinc acetate. While obvious to those who already are aware of the system, this information may otherwise be overlooked by chemists naive to the intricacies of ZIF-8 synthesis – an example of chemical intuition.[6] Therefore, clustering of similar synthesis protocols together can help users to avoid some common pitfalls when planning experiments for the first time.

Finally, to demonstrate the benefit of this approach towards synthesis optimisation, we consider the reduction in experiments that would be required to explore the identified popular sub-regions of the synthesis space. From the clustering analysis, we identified 6 sub-regions with only 3 chemicals of interest – clusters 1, 2, 3, 6, 7, and 8 in Table 4, containing only a single metal salt, 2-methylimidazole, and a single solvent – and a further 2 sub-regions with 4 chemicals of interest: clusters 5 and 8 containing either mixed salts or solvents. Accordingly, rather than requiring N^8 or $12(N^5)$ experiments, full exploration would only require $6(N^3) + 2(N^4) \approx N^{4.4}$ experiments. To illustrate the extent of dimensionality reduction in real terms, the number of experiments required to explore the synthesis space are shown in Table 5 for various values of *N*. In combination with the quantity distributions shown in Figure 3 and Figure 4, text mining and data reduction tools demonstrated in this paper will provide excellent initial values for efficient searching of chemical synthesis space, thereby accelerating methodology refinement for a range of nanomaterials.

Table 5 – Approximate number of experiments required to fully characterise the synthesis space, for various values of N.

	Full exploration (N^8)	*Limited experimental complexity ($12(N^5)$)*	*Identified clusters only ($6(N^3) + 2(N^4)$)*
N = 3	6,500	2,900	320
N = 5	390,000	37,500	2,000
N = 10	$1x10^8$	$1.2x10^6$	$2.6x10^4$

Conclusions

In this study, we applied text mining to the problem of synthesis methodology optimisation, exploring to what extent the previously accumulated collective knowledge of a particular nanomaterial can accelerate the development of reliable and scalable synthesis protocols. As the first step toward this objective, in this study, we posed three research questions: first, is it possible to use text mining tools to provide deep insight into a single synthetic target, rather than a comprehensive overview of a family of materials? Second, is it possible to standardise the synthesis details extracted as a means of performing like-for-like comparison between different studies? Finally, is it possible to use this analysis to suggest optimal synthesis conditions, thereby accelerating methodology development?

To this end, we developed software to systematically analyse nanomaterials synthesis methods based on established text mining protocols. We extracted structured data to describe the details of

each synthesis protocol, enabling large-scale statistical analysis of the synthesis parameter space and clustering of similar methods together to identify well-explored regions of the synthesis space. We believe that this progress represents the first step in creating a closed feedback loop for the automated optimisation of experimental nanomaterials synthesis, visualised in Figure 8. In this feedback loop text mined information can identify common limits to parameters as well as low-dimensional sub-regions of interest in the synthesis space. By using this information as initial conditions for iterative high-throughput experimentation, the search for synthesis protocols optimised against any target material quality metric can be greatly accelerated.

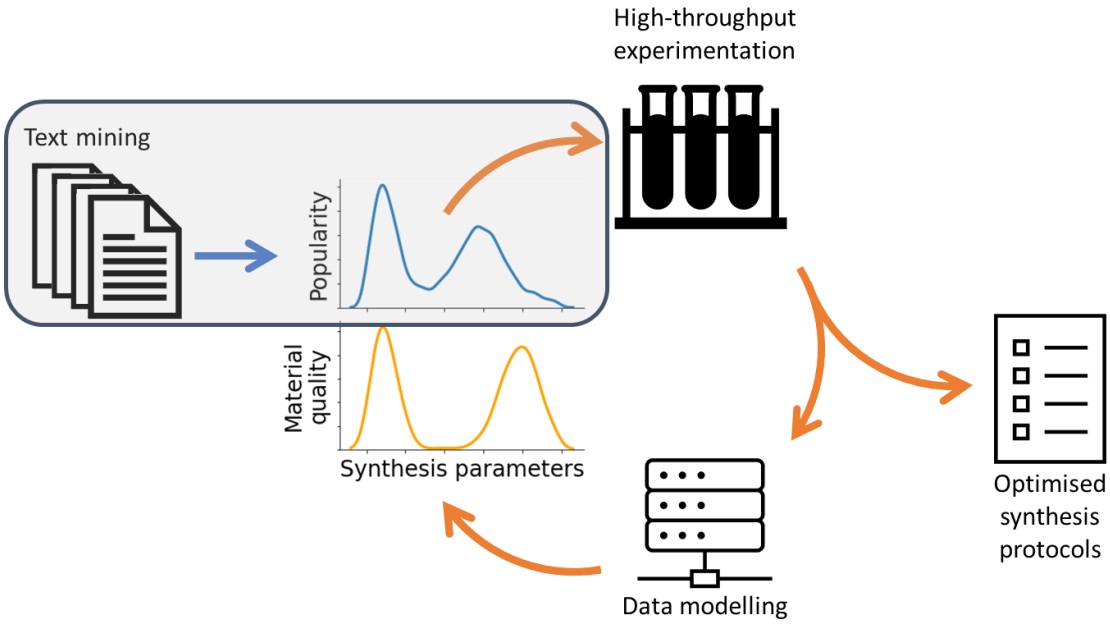

Figure 8 - Scheme of a synthesis protocol optimisation feedback loop. Work carried out in this study is shaded in grey.

As a case study to demonstrate the utility of this approach, we performed a quantitative meta-analysis of 1600 synthesis methods for the common MOF ZIF-8. Using this framework, we identified key aspects of the synthesis including the range of chemicals used as reagents, solvents, and ancillary modulators/pH modifiers. We extracted information about the quantity of each reagent used during the synthesis, enabling us to identify the distribution of synthesis scales, reagent ratios, and reaction mixture solids concentration, as well as reaction times and temperatures. Further insight was gathered by cross-referencing chemicals mentioned against the stage they were introduced into the synthesis protocol – for example identifying that ethanol is primarily used as a washing solvent rather than in the reaction medium. We demonstrated how the quantitative meta-analysis performed here can assist in systematic searches of the synthesis phase space by identifying both low-dimensional regions of interest and the distribution of synthesis parameters. As a result, we were able to reduce the number of hypothetical experiments required to optimise ZIF-8 significantly. Notably, while we considered MOF materials as a case study in this work, the methods developed here are general to any synthesis type. Particularly, we envisage they will be useful the systematising understanding of other emerging nanomaterial systems such as mesoporous (organo)silicas, covalent organic frameworks, and polymers of intrinsic microporosity.

Despite the deep insight we were able to gain into the synthesis system of ZIF-8, the current study also identified significant challenges associated with developing a true “synthetic oracle” for predicting the ideal synthesis parameters for any given material. While we were able to identify and extract information about the synthesis, we were unable to reliably connect the quality of the

material produced to the methods themselves (e.g. by identifying specific yield or surface area). A crucial next step is therefore to adopt state of the art transformer-based methods e.g. BERT or GPT-4 to better interpret the entire research article as a single unit and therefore identify implicitly described synthesis protocols (e.g. tabulated changes to individual synthesis parameters). A second challenge lies in the estimating the viability of synthesis parameters extracted during text mining or proposed by generative models, preventing automated reproduction of a synthesis protocol without human oversight and validation. Finally, as has been discussed elsewhere, the synthesis protocol extraction methods developed here can only build from published information, which is biased towards the most successful synthesis methods only. More comprehensive reporting of synthesis information using structured formats akin to the crystallographic information file format would enable far more wide-reaching analysis to be performed.

In summary, the methods developed in this study acts as a preliminary approach for the large-scale standardisation and analysis of experimental synthesis data, representing the first step in creating a closed feedback loop for the automated optimisation of experimental nanomaterials synthesis. By interfacing with automated and high throughput reactionware e.g. through integration of the XDL chemical programming language, methodology development will be significantly accelerated thereby easing the adoption of nanomaterials at larger scales and in new settings.

References

1. Kononova, O. *et al.* Opportunities and challenges of text mining in materials research. *iScience* **24**, 102155 (2021).
2. Kim, E., Huang, K., Kononova, O., Ceder, G. & Olivetti, E. Distilling a Materials Synthesis Ontology. *Matter* **1**, 8–12 (2019).
3. Baer, D. R. & Gilmore, I. S. Responding to the growing issue of research reproducibility. *J. Vac. Sci. Technol. A* **36**, 068502 (2018).
4. Baer, D. R., Munusamy, P. & Thrall, B. D. Provenance information as a tool for addressing engineered nanoparticle reproducibility challenges. *Biointerphases* **11**, 04B401 (2016).
5. Burger, B. *et al.* A mobile robotic chemist. *Nature* **583**, 237–241 (2020).
6. Moosavi, S. M. *et al.* Capturing chemical intuition in synthesis of metal-organic frameworks. *Nat. Commun.* **10**, 539 (2019).
7. Wilbraham, L., Mehr, S. H. M. & Cronin, L. Digitizing Chemistry Using the Chemical Processing Unit: From Synthesis to Discovery. *Acc. Chem. Res.* **54**, 253–262 (2021).
8. Hammer, A. J. S., Leonov, A. I., Bell, N. L. & Cronin, L. Chemputation and the Standardization of Chemical Informatics. *JACS Au* **1**, 1572–1587 (2021).
9. Jessop, D. M., Adams, S. E., Willighagen, E. L., Hawizy, L. & Murray-Rust, P. OSCAR4: A flexible architecture for chemical textmining. *J. Cheminform.* **3**, 1–12 (2011).
10. Mavračić, J., Court, C. J., Isazawa, T., Elliott, S. R. & Cole, J. M. ChemDataExtractor 2.0: Autopopulated Ontologies for Materials Science. *J. Chem. Inf. Model.* **61**, 4280–4289 (2021).
11. Hawizy, L., Jessop, D. M., Adams, N. & Murray-Rust, P. ChemicalTagger: A tool for semantic text-mining in chemistry. *J. Cheminform.* **3**, 1–13 (2011).
12. Vazquez, M., Krallinger, M., Leitner, F. & Valencia, A. Text mining for drugs and chemical compounds: Methods, tools and applications. *Mol. Inform.* **30**, 506–519 (2011).

13. Guo, J. *et al.* Automated Chemical Reaction Extraction from Scientific Literature. *J. Chem. Inf. Model.* **62**, 2035–2045 (2022).

14. Beard, E. J., Sivaraman, G., Vázquez-Mayagoitia, Á., Vishwanath, V. & Cole, J. M. Comparative dataset of experimental and computational attributes of UV/vis absorption spectra. *Sci. Data* **6**, 1–11 (2019).

15. Cooper, C. B. *et al.* Design-to-Device Approach Affords Panchromatic Co-Sensitized Solar Cells. *Adv. Energy Mater.* **9**, 1–10 (2019).

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. Beard, E. J. & Cole, J. M. Perovskite- and Dye-Sensitized Solar-Cell Device Databases Auto-generated Using ChemDataExtractor. *Sci. Data* **9**, 1–19 (2022).

18. Zhao, J. & Cole, J. M. A database of refractive indices and dielectric constants auto-generated using ChemDataExtractor. *Sci. Data* **9**, 1–11 (2022).

19. Kononova, O. *et al.* Text-mined dataset of inorganic materials synthesis recipes. *Sci. Data* **6**, 1–11 (2019).

20. Isazawa, T. & Cole, J. M. Single Model for Organic and Inorganic Chemical Named Entity Recognition in ChemDataExtractor. *J. Chem. Inf. Model.* **62**, 1207–1213 (2022).

21. Klinger, R., Kolářik, C., Fluck, J., Hofmann-Apitius, M. & Friedrich, C. M. Detection of IUPAC and IUPAC-like chemical names. *Bioinformatics* **24**, 268–276 (2008).

22. Majumdar, S., Moosavi, S. M., Jablonka, K. M., Ongari, D. & Smit, B. Diversifying Databases of Metal Organic Frameworks for High-Throughput Computational Screening. *ACS Appl. Mater. Interfaces* **13**, 61004–61014 (2021).

23. Boyd, P. G. & Woo, T. K. A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. *CrystEngComm* **18**, 3777–3792 (2016).

24. Wilmer, C. E. *et al.* Large-scale screening of hypothetical metal-organic frameworks. *Nat. Chem.* **4**, 83–89 (2012).

25. Lee, S. Y. S. *et al.* Computational Screening of Trillions of Metal-Organic Frameworks for High-Performance Methane Storage. *ACS Appl. Mater. Interfaces* **13**, 23647–23654 (2021).

26. Moosavi, S. M. *et al.* Understanding the diversity of the metal-organic framework ecosystem. *Nat. Commun.* **11**, 4068 (2020).

27. Moghadam, P. Z. *et al.* Targeted classification of metal-organic frameworks in the Cambridge structural database (CSD). *Chem. Sci.* **11**, 8373–8387 (2020).

28. Moghadam, P. Z. *et al.* Development of a Cambridge Structural Database Subset: A Collection of Metal-Organic Frameworks for Past, Present, and Future. *Chem. Mater.* **29**, 2618–2625 (2017).

29. Ongari, D., Talirz, L. & Smit, B. Too Many Materials and Too Many Applications: An Experimental Problem Waiting for a Computational Solution. *ACS Cent. Sci.* **6**, 1890–1900 (2020).

30. Bucior, B. J. *et al.* Identification Schemes for Metal–Organic Frameworks To Enable Rapid Search and Cheminformatics Analysis. *Cryst. Growth Des.* **19**, 6682–6697 (2019).

31. Groom, C. R. & Allen, F. H. The Cambridge Structural Database in retrospect and prospect.

Angew. Chemie - Int. Ed. **53**, 662–671 (2014).

32. Luo, Y. *et al.* MOF Synthesis Prediction Enabled by Automatic Data Mining and Machine Learning**. *Angew. Chemie Int. Ed.* **61**, (2022).

33. Gubsch, K., Bence, R., Glasby, L. T. & Moghadam, P. Z. DigiMOF: A Database of MOF Synthesis Information Generated via Text Mining. *ChemRxiv* 1–26 (2022) doi:10.26434/chemrxiv-2022-41t70.

34. Cox, C. S., Slavich, E., Macreadie, L. K., McKemmish, L. K. & Lessio, M. Understanding the Role of Synthetic Parameters in the Defect Engineering of UiO-66: A Review and Meta-analysis. *Chem. Mater.* **35**, 3057–3072 (2023).

35. Zhang, H. & Snurr, R. Q. Computational Study of Water Adsorption in the Hydrophobic Metal-Organic Framework ZIF-8: Adsorption Mechanism and Acceleration of the Simulations. *J. Phys. Chem. C* **121**, 24000–24010 (2017).

36. Bhattacharyya, S. *et al.* Acid gas stability of zeolitic imidazolate frameworks: Generalized kinetic and thermodynamic characteristics. *Chem. Mater.* **30**, 4089–4101 (2018).

37. De Lange, M. F. *et al.* Metal-Organic Frameworks in Adsorption-Driven Heat Pumps: The Potential of Alcohols as Working Fluids. *Langmuir* **31**, 12783–12796 (2015).

38. Kinoshita, M., Yanagida, S., Gessei, T. & Monkawa, A. Precursor concentration effects on crystallite size and enzyme immobilization efficiency of Enzyme@ZIF-8 composite. *J. Cryst. Growth* **600**, 126877 (2022).

39. Lee, Y. R., Do, X. H., Hwang, S. S. & Baek, K. Y. Dual-functionalized ZIF-8 as an efficient acid-base bifunctional catalyst for the one-pot tandem reaction. *Catal. Today* **359**, 124–132 (2021).

40. Paul, A., Banga, I. K., Muthukumar, S. & Prasad, S. Engineering the ZIF-8 Pore for Electrochemical Sensor Applications-A Mini Review. *ACS Omega* **7**, 26993–27003 (2022).

41. Lewis, A. *et al.* Crystallization and phase selection of zeolitic imidazolate frameworks in aqueous cosolvent systems: The role and impacts of organic solvents. *Results Eng.* **17**, 100751 (2023).

42. Tsai, C.-W. & Langner, E. H. G. The effect of synthesis temperature on the particle size of nano-ZIF-8. *Microporous Mesoporous Mater.* **221**, 8–13 (2016).

43. Kida, K., Okita, M., Fujita, K., Tanaka, S. & Miyake, Y. Formation of high crystalline ZIF-8 in an aqueous solution. *CrystEngComm* **15**, 1794–1801 (2013).

44. Cravillon, J. *et al.* Formate modulated solvothermal synthesis of ZIF-8 investigated using time-resolved in situ X-ray diffraction and scanning electron microscopy. *CrystEngComm* **14**, 492–498 (2012).

45. Nordin, N. A. H. M. *et al.* Aqueous room temperature synthesis of zeolitic imidazole framework 8 (ZIF-8) with various concentrations of triethylamine. *RSC Adv.* **4**, 33292–33300 (2014).

46. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual*. (CreateSpace, 2009).

47. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018) doi:10.48550/arXiv.1810.04805.

48. McKinney, W. Data Structures for Statistical Computing in Python. in *Proceedings of the 9th*

Python in Science Conference (eds. van der Walt, S. & Millman, J.) vol. 1 56–61 (2010).

49. The pandas development team. pandas-dev/pandas: Pandas. (2023) doi:10.5281/zenodo.3509134.

50. Kim, S. *et al.* PubChem 2023 update. *Nucleic Acids Res.* **51**, D1373–D1380 (2023).

51. Landrum, G. RDKit: Open-source cheminformatics. (2010).

52. Bell, C. & Cortes-Pena, Y. R. Chemicals: Chemical properties component of Chemical Engineering Design Library (ChEDL). (2016).

53. Thomson, G. H., Hankinson, R. W. & Thomson, G. H. A new correlation for saturated densities of liquids and their mixtures. *AIChE J.* **25**, 653–663 (1979).

54. Zhang, Z. *et al.* Enhancement of CO 2 Adsorption and CO 2 /N 2 Selectivity on ZIF-8 via Postsynthetic Modification. *AIChE J.* **59**, 2195–2206 (2013).

55. Wang, Z. *et al.* ULSA: unified language of synthesis actions for the representation of inorganic synthesis protocols. *Digit. Discov.* **1**, 313–324 (2022).

56. Sparck Jones, K. A Statistical Interpretation of Term Specificity And Its Application In Retrieval. *J. Doc.* **28**, 11–21 (1972).

57. Esther, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* 226–231 (1996).

58. Maaten, L. van der & Hinton, G. Visualizing High-Dimensional Data Using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

59. Pan, Y. *et al.* Tuning the crystal morphology and size of zeolitic imidazolate framework-8 in aqueous solution by surfactants. *CrystEngComm* **13**, 6937 (2011).

60. Swain, M. C. & Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **56**, 1894–1904 (2016).

61. Liu, J. *et al.* The active roles of ZIF-8 on the enhanced visible photocatalytic activity of Ag/AgCl: Generation of superoxide radical and adsorption. *J. Alloys Compd.* **693**, 543–549 (2017).

62. Butova, V. V. *et al.* Modification of ZIF-8 with triethylamine molecules for enhanced iodine and bromine adsorption. *Inorganica Chim. Acta* **509**, 119678 (2020).

63. Samadi-Maybodi, A., Ghasemi, S. & Ghaffari-Rad, H. Ag-doped zeolitic imidazolate framework-8 nanoparticles modified CPE for efficient electrocatalytic reduction of H2O2. *Electrochim. Acta* **163**, 280–287 (2015).

64. Schejn, A. *et al.* Controlling ZIF-8 nano- and microcrystal formation and reactivity through zinc salt variations. *CrystEngComm* **16**, 4493–4500 (2014).

65. Jian, M. *et al.* Water-based synthesis of zeolitic imidazolate framework-8 with high morphology level at room temperature. *RSC Adv.* **5**, 48433–48441 (2015).

66. Öztürk, Z., Filez, M. & Weckhuysen, B. M. Decoding Nucleation and Growth of Zeolitic Imidazolate Framework Thin Films with Atomic Force Microscopy and Vibrational Spectroscopy. *Chem. - A Eur. J.* **23**, 10915–10924 (2017).

67. Hamidon, N. F., Tahir, M. I. M., Latif, M. A. M. & Abdul Rahman, M. B. Effect of altering linker ratio on nano-ZIF-8 polymorphisms in water-based and modulator-free synthesis. *J. Coord. Chem.* **75**, 1180–1192 (2022).

68. Yamamoto, D. *et al.* Synthesis and adsorption properties of ZIF-8 nanoparticles using a micromixer. *Chem. Eng. J.* **227**, 145–150 (2013).

69. Chen, B., Bai, F., Zhu, Y. & Xia, Y. A cost-effective method for the synthesis of zeolitic imidazolate framework-8 materials from stoichiometric precursors via aqueous ammonia modulation at room temperature. *Microporous Mesoporous Mater.* **193**, 7–14 (2014).

70. Zhang, Y., Jia, Y., Li, M. & Hou, L. Influence of the 2-methylimidazole/zinc nitrate hexahydrate molar ratio on the synthesis of zeolitic imidazolate framework-8 crystals at room temperature. *Sci. Rep.* **8**, 1–7 (2018).

71. Albright, P. S. & Gosting, L. J. Dielectric Constants of the Methanol-Water System from 5 to 55°. *J. Am. Chem. Soc.* **68**, 1061–1063 (1946).

72. Lee, Y. R. *et al.* ZIF-8: A comparison of synthesis methods. *Chem. Eng. J.* **271**, 276–280 (2015).

73. Xia, W., Lau, S. K. & Yong, W. F. Comparative life cycle assessment on zeolitic imidazolate framework-8 (ZIF-8) production for CO2 capture. *J. Clean. Prod.* **370**, 133354 (2022).

74. Tricker, A. W., Samaras, G., Hebisch, K. L., Realff, M. J. & Sievers, C. Hot spot generation, reactivity, and decay in mechanochemical reactors. *Chem. Eng. J.* **382**, 122954 (2020).

75. Torad, N. L. *et al.* Nanoarchitectured porous carbons derived from ZIFs toward highly sensitive and selective QCM sensor for hazardous aromatic vapors. *J. Hazard. Mater.* **405**, 124248 (2021).

76. Zhou, C. *et al.* Epoxy composite coating with excellent anticorrosion and self-healing performances based on multifunctional zeolitic imidazolate framework derived nanocontainers. *Chem. Eng. J.* **385**, (2020).

77. Zheng, G. *et al.* Shape control in ZIF-8 nanocrystals and metal nanoparticles@ZIF-8 heterostructures. *Nanoscale* **9**, 16645–16651 (2017).

78. Cravillon, J. *et al.* Controlling zeolitic imidazolate framework nano- and microcrystal formation: Insight into crystal growth by time-resolved in situ static light scattering. *Chem. Mater.* **23**, 2130–2141 (2011).

79. Tuffnell, J. M. *et al.* Comparison of the ionic conductivity properties of microporous and mesoporous MOFs infiltrated with a Na-ion containing IL mixture. *Dalt. Trans.* **49**, 15914–15924 (2020).

80. Huang, S. & Cole, J. M. BatteryDataExtractor: battery-aware text-mining software embedded with BERT models. *Chem. Sci.* **13**, (2022).

81. Jablonka, K. M., Schwaller, P. & Ortega-guerrero, A. Is GPT-3 all you need for low-data discovery in chemistry? *ChemRxiv* 1–32 (2023).