University of Sheffield

Topological Characterisation of Metal-Organic Frameworks in the Cambridge Structural Database

Lawson Taylor Glasby

Supervisor: Dr Peyman Z. Moghadam

Supervisor: Prof. Joan L. Cordiner

Supervisor: Dr Jason C. Cole

A thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

in the

Department of Chemical and Biological Engineering

May 2024

Declaration

I, the author, confirm that the works within this document are my own except where work that has been formed as part of jointly authored publications has been included. The contribution of the authors to each of the works has been explicitly indicated below. I confirm that the appropriate credit has been given within this declaration where reference has been made to the work of others.

All sentences or passages quoted in this document from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure.

This work has not been previously presented for an award at this, or any other, university. I am aware of the university guidance on the use of unfair means. I can declare that part of this work consists of the following publications:

1. (Glasby and Moghadam, 2021) Glasby L. T., & Moghadam P. Z. (2021). Hydrogen storage in MOFs: Machine learning for finding a needle in a haystack. Patterns, 2, 7, 100305. https://doi.org/10.1016/j.patter.2021.100305

2. (Glasby et al., 2023) Glasby L. T., Whaites E. H., & Moghadam P. Z. (2023). Machine learning and digital manufacturing approaches for solid-state materials development. AI-guided design and property prediction for zeolites and nanoporous materials, 14, 377-409. https://doi.org/10.1002/9781119819783.ch14

3. (Glasby et al., 2023) Glasby L. T., Gubsch K., Bence R., Oktavian R., Isoko K., Moosavi S. M., Cordiner J. L., Cole J. C., & Moghadam P. Z. (2023). DigiMOF: A database of metal-organic framework synthesis information generated via text mining. Chemistry of Materials, 35, 11, 4510-4524. https://doi.org/10.1021/acs.chemmater.3c00788

4. (Glasby et al., 2023) Glasby L. T., Oktavian R., Zhu K., Cordiner J. L., Cole J. C., & Moghadam P. Z. (2023). Augmented reality for enhanced visualisation of MOF adsorbents. Journal of Chemical Information and Modeling, 63, 19, 5950-5955. https://doi.org/10.1021/acs.jcim.3c01190

5. Glasby L. T., Cordiner J. L., Cole J. C., & Moghadam P. Z. (2024). Topological characterisation of metal-organic frameworks: A perspective. [Manuscript under review].

6. Glasby L. T., Cordiner J. L., Cole J. C., & Moghadam P. Z. (2024). Integrating CrystalNets.jl and bench-marking its performance on metal-organic frameworks in the Cambridge Structural Database. [Manuscript in preparation for submission].

7. (Zanca et al., 2021) Zanca F., Glasby L. T., Chong S., Chen S., Kim J., Fairen-Jimenez D., Monserrat B., & Moghadam P. Z. (2021). Computational techniques for characterisation of electrically conductive MOFs: quantum calculations and machine learning approaches. Journal of Materials Chemistry C, 9, 39, 13584-13599. https://doi.org/10.1039/d1tc02543k

8. (Oktavian et al., 2022) Oktavian R., Schireman R., Glasby L. T., Huang G., Zanca F., Fairen-Jimenez D., Ruggiero M. T., & Moghadam P. Z. (2022). Computational characterization of Zr-oxide MOFs for adsorption applications. ACS Applied Materials & Interfaces, 14, 51, 56938-56947. https://doi.org/10.1021/acsami.2c13391

9. Oktavian R., Goeminne R., Glasby L. T., Song P., Huynh R., Taheri-Qazvini O., Ghaffari-Nik O., Masoumifard N., Hovington P., Van Speybroeck V., & Moghadam P. Z. (2024). Gas adsorption and framework flexibility of CALF-20 explored via experiments and simulations. [Manuscript accepted in Nature Communications].

Name: Lawson Taylor Glasby

Date: May 2024

Abstract

Metal-organic frameworks (MOFs) have become a widely studied class of porous materials over the last 30 years. They have been, and continue to be, extensively researched for applications requiring porous and adsorptive properties of materials. This thesis focuses on three main topics which sit at the forefront of MOF research: reliable topological characterisation of crystalline materials, machine learning and data driven manufacturing, and the development of new computational tools.

Firstly, we investigate topological characterisation of these crystalline materials, describing the methods and algorithms by which they can be categorised through the medium of a perspective review, followed by the integration of a newly developed open-source software with the Cambridge Crystallographic Data Centre's (CCDC) existing crystallographic data suite and Python API.

This is succeeded by the introduction of state-of-the-art machine learning (ML) and digital manufacturing techniques with the view that they can be applied to the future of the field. In this work we discuss the use of ML in solid state materials development which is followed up by work in which we developed a new method of abstracting existing synthesis information published in thousands of previous MOF studies.

Lastly, we apply new augmented reality techniques to visualise the results of topological deconstruction and adsorption studies of MOFs. The result of this work over the last 4 years has enabled researchers to use the Cambridge Structural Database (CSD) to maximum effectiveness when searching for synthesis conditions, precursors, linker types, topologies, and more whilst also integrating ML techniques such as Natural Language Processing (NLP) for data mining and introducing new ways of visualising the results.

Acknowledgements

I always endeavour to read the acknowledgement sections of the theses I come across. It's sometimes easy to forget that works such as these are only made possible thanks to the help of an army of supportive and positive people, and that the works contained within these documents are collaborative efforts made over the course of several years rather than just the contributions of one individual. Throughout a PhD programme, life also happens, and I am grateful for everyone who has stepped foot into mine, no matter how briefly, and helped to achieve my goal of making every day worth living.

Firstly, I must thank my principal supervisor, Dr Peyman Z. Moghadam, for his support throughout the last three and a half years. His rigorous and often gruelling review process has shaped me into the adept and conscientious writer I feel that I have become. It was he who guided me to produce higher and higher quality work, such that it would be considered appropriate for publication within some of the most renowned journals in our field. Perhaps at the time I didn't feel so grateful, staring at documents littered with corrections in bright red text, but hindsight is a wonderful thing, and I am so grateful to have been pushed to achieve and maintain a level of quality that was higher than I ever expected of myself.

I must also thank Professor Joan L. Cordiner for taking on a key supervisory role after Peyman left Sheffield, as my primary supervisor at the university. Throughout the time I have spent under her guidance, she has never been anything but kind, supportive, and positive, the importance of which I cannot overstate. I have never felt that I was unable to access any opportunity that was presented to me, and this is simply because Joan made it possible.

I would also like to thank Dr Jason C. Cole from the Cambridge Crystallographic Data Centre, a brilliant organisation whose employees both sponsored and mentored me throughout this journey. Jason has been a wonderful mentor to me over the course of my PhD and did not hesitate to answer any questions I had for him, no matter how trivial. He was always willing to make time to help and assist me with anything that I needed, and for that, I am truly thankful. I must also thank everyone else at the CCDC who made the funding for this PhD available, who read through my submissions and left constructive comments, and who through fruitful discussions inspired a notable proportion of the work in this document.

Next I thank everyone at the Centre for Computational Materials Discovery, or the PZM club, as it has been affectionately known to us. From the very first day, Federica Zanca was extremely helpful, supportive, and welcoming. Rama Oktavian has been, and hopefully forever will be, a beacon of positivity and encouragement within the group, and his enthusiasm has been a wonderful motivator for us all, I am sure. Chenhao Li, with whom I may have spent less time working with directly, has consistently had a positive

impact on daily life in general, even if it was just a cheerful wave or hello. I'd also like to extend my thanks to the newer group members who have joined since I started to wind down here, and those who only joined us temporarily throughout their summer research projects - their hard work and ideas are reflected in some of the publications produced and included in this document, which I consider a testament to their abilities.

Outside of strictly research life, I thank my colleagues and friends in and around the shared office within which I have spent many painstaking hours writing, but also moaning, and most importantly laughing. They are Robert Douglas, Paul Coombes, Jack Atkinson, Anna Leathard, Mahdi Ahmed, and Christopher Passmore. I also thank those of whom I worked with as a graduate teaching assistant, and I must extend further thanks to all the lecturers who offered me the opportunity to support their teaching. Leading seminars was something I grew to enjoy and I have loved the opportunity to teach, inspire, and even just converse with students around the various departments I lent my time to.

Penultimately, I begin to wrap up with immeasurable thanks to my parents Charles and Tracy who expressed unwavering belief in me even during times where I had none myself. There have been times when it has only been thanks to them that I have found the strength to continue on this journey, and I am honoured to have been able to complete it. From a very young age my dad must have seen potential in me, predicting long before I even knew what university was that I would go on to complete a doctorate. I must also thank my brothers Cameron and Morgan, my friends from back home, and my extended family who have supported me.

Finally, I express my thanks to all involved in the Sheffield University Bears Ice Hockey club throughout my time as a player and committee member, and to my friends and teammates at the Sutton Sting Ice Hockey Club who have all been a wonderful distraction from chemistry, engineering, and writing, as I spent my free time during evenings and weekends playing ice hockey. I am sure there are many more people who have influenced me on this journey, via both positive and negative experiences, that have built my character, confidence, and shaped my personality to how it is today whom I have yet to acknowledge, so consider this sentence just that.

Contents

List of Figures **x**

List of Tables **xv**

1 Introduction **1**
1.1 Metal-Organic Frameworks 1
1.2 Cambridge Structural Database and MOF Subset 3
1.3 Thesis Outline 4
1.3.1 Chapter 2 - Topological Characterisation of Metal-Organic Frameworks 4
1.3.2 Chapter 3 - Integrating CrystalNets.jl and Bench-marking Performance on the Cambridge Structural Database 5
1.3.3 Chapter 4 - Machine Learning and Digital Manufacturing Approaches for Solid-state Materials Development 5
1.3.4 Chapter 5 - DigiMOF: A Database of Metal-Organic Framework Synthesis Information Generated via Text Mining 6
1.3.5 Chapter 6 - Augmented Reality for Enhanced Visualization of MOF Adsorbents 7
1.4 Aims, Objectives and Scientific Contribution 7
1.4.1 Topological Characterisation 7
1.4.2 Machine Learning and Augmented Reality 8
1.4.3 Development of New Computational Tools 8
1.5 Summary 8
References 9

2 Topological Characterisation of Metal-Organic Frameworks: A Perspective **12**
2.1 Publication Information and Paper Contributions 12
2.2 Abstract 12
2.2.1 Keywords 13
2.3 Introduction 13
2.4 What is topology? 15
2.5 Popular Topologies and Resources 18
2.5.1 The Reticular Chemistry Structure Resource (RSCR) 18
2.5.2 Edge-Transitive Nets 20
2.6 Deconstruction Techniques 20
2.6.1 All Node and Single Node Deconstruction 21
2.6.2 Alternative Deconstruction Methods 23
2.7 MOF Databases and Design Principles 25

2.8 Topological Characterisation Software 29
2.8.1 Introduction 29
2.8.2 ToposPro and TopCryst 29
2.8.3 MOFid and web-mofid 31
2.8.4 CrystalNets.jl and CrystalNets 33
2.8.5 Guidance and Limitations 35
2.9 Recent Developments 36
2.10 Conclusions and Perspective 37
References 38

3 Integrating CrystalNets.jl and Bench-marking Performance on the Cambridge Structural Database 46
3.1 Publication Information and Paper Contributions 46
3.2 Abstract 46
3.2.1 Keywords 46
3.3 Introduction 47
3.4 Workflow and Method 48
3.4.1 Structure and Software Preparation 48
3.4.2 Software Runs 48
3.4.3 Software Outputs 49
3.5 Results and Analysis 50
3.5.1 Results and Data Cleaning 50
3.5.2 Analysis 50
3.6 Implementation of CrystalNets within CSD Mercury 54
3.7 Conclusion 58
References 59

4 Machine Learning and Digital Manufacturing Approaches for Solid-state Materials Development 63
4.1 Publication Information and Paper Contributions 63
4.2 Abstract 63
4.2.1 Keywords 64
4.3 Introduction 64
4.4 The Development of MOF Databases 67
4.5 Natural Language Processing 68
4.6 An Overview of Machine Learning Models 71
4.7 Machine Learning for Synthesis and Investigation of Solid State Materials 74
4.8 Machine Learning in Design and Discovery of MOFs 77
4.9 Current Limitations of Machine Learning for MOFs 82
4.10 Automated Synthesis and Digital Manufacturing 84
4.11 Digital Manufacturing of MOFs 91
4.12 The Future of Digital Manufacturing 95
References 97

5 DigiMOF: A Database of Metal-Organic Framework Synthesis Information Generated via Text Mining 105
5.1 Publication Information and Paper Contributions 105
5.2 Abstract 105
5.2.1 Keywords 106
5.3 Introduction 106

5.4 Property Identification and Parsing . . . 107
5.5 Methods: Automatic Generation of the DigiMOF Database . . . 108
5.5.1 Natural Language Processing . . . 108
5.5.2 Technical Validation . . . 109
5.5.3 Parser Training . . . 110
5.5.4 Obtaining Metal, Topology, and Linker Data . . . 111
5.5.5 Geometric Properties . . . 113
5.6 Results and Discussion . . . 114
5.7 Data Analysis . . . 115
5.7.1 Synthesis Methods . . . 115
5.7.2 Topology . . . 116
5.7.3 Solvent . . . 118
5.7.4 Organic Linkers . . . 119
5.7.5 Metal Precursor . . . 120
5.7.6 Temperature . . . 121
5.7.7 Building Blocks and Topology . . . 122
5.7.8 Cost Analysis . . . 123
5.8 Conclusions and Future Directions . . . 124
References . . . 126

6 Augmented Reality for Enhanced Visualization of MOF Adsorbents 133
6.1 Publication Information and Paper Contributions . . . 133
6.2 Abstract . . . 133
6.2.1 Keywords . . . 133
6.3 Introduction . . . 133
6.4 Modeling MOFs for use in AR . . . 135
6.4.1 Crystal Structure Modifications . . . 135
6.4.2 Conversion Processes from CIF/PDB to FBX . . . 136
6.4.3 Publication on p3d.in and QR Code Generation . . . 136
6.5 Applications of AR in Representing MOFs . . . 137
6.5.1 Crystal Representation . . . 137
6.5.2 Topology Representation . . . 139
6.5.3 Gas Adsorption Representation . . . 139
6.5.4 Reception . . . 140
6.6 Conclusion . . . 141
References . . . 142

7 Conclusions and Future Work 145
7.1 Conclusion . . . 145
7.2 Future Work . . . 146

Appendices 148

A Supporting Information for DigiMOF 149
A.1 Article Retrieval . . . 149
A.2 Database Overview and Performance . . . 149
A.3 Synthesis Proportionality . . . 155
A.4 Data Transformation and Visualization . . . 156
A.5 Building Blocks and Topology . . . 157
A.6 Text Mining Overview . . . 159

A.7 Parsed Articles 161
References 166

B Supporting Information for Augmented Reality 167
B.1 Visualising MOFs with Augmented Reality (AR) 167
B.1.1 Requirements: 167
B.2 AR File Creation Method 167
B.2.1 Part A – Selecting and Modifying Files: 167
B.2.2 Part B - Jmol: 169
B.2.3 Part C - Blender: 171
B.2.4 Part D - p3d.in: 173
B.2.5 Part E - QR Code: 175
B.3 RASPA AR file generation for gas adsorption visualisation 176
B.4 CrystalNets AR file generation to visualise topology 177
B.4.1 Part I: Obtaining OBJ format topological nets. 177
B.4.2 Part II: Combining OBJ nets and OBJ crystals in Blender. 179

List of Figures

1.1 A representation of one of the most renowned MOFs (MOF-5), with **pcu** topology overlaid onto the structure in CSD Mercury. 3
1.2 Summary of the seven criteria designed to build the CSD MOF subset, where QA = O, N, P, C, B, S. QB = N, P, B, S, C and superscripts "c" and "a" impose the corresponding atoms to be "cyclic" or "acyclic", respectively. Me denotes methyl groups. The dotted line refers to any of the bond types stored in the CSD (single, double, triple, quadruple, aromatic, polymeric, delocalised, and pi). The dotted line with the two lines through indicates a variable bond type (i.e., two or more of the options above). In these cases, the variable type is single, double, or delocalised. [22] 4

2.1 The distribution of MOFs within the CSD, including dimensionality breakdowns of 1D, 2D, and 3D structures. The left axis indicates the number of structures deposited per year per dimensionality, whilst the right axis keeps a cumulative total across the timeline. (Data correct to CSD 5.45 Nov 2023). 14
2.2 An example of two similarly connected crystal structures expanded 1×, 2×, and 3× from their unit cells where a. CSD OFAWAV (DUT-53(Hf)) consists of 8-connected SBUs, and b. CSD OFAWID (DUT-84(Zr)) consists of 6-connected SBUs, visualised using CCDC's Mercury [35, 36]. The latter entry is considered disjoint due to the lack of polymeric expansion sites parallel to the c-axis; however, it expands polymerically in both other dimensions. Hf (bright blue), Zr (cyan), O (red), H (white), and C (grey). 17
2.3 Example RCSR topological nets created and visualised using ToposPro [26]. Red atoms represent metal nodes, whereas green atoms represent organic nodes. 19
2.4 Schematic demonstrating crystal deconstruction techniques applied to CSD JOZWIG [59]. The distinct path taken by each algorithm for large heteroaromatic rings results in **a.** the all node approach matching the **xxv** topology, and **b.** the single node approach matching with the **ftw** topology. Wireframe structures show C(grey), O (red), N (blue), Zr (light blue), which are simplified to metal nodes (red), and organic nodes (green) connected by straight edges representative of linkers (blue). 22

2.5 Schematic demonstrating crystal deconstruction techniques applied to CSD SAHYIK [66]. The approach **a.**, standard simplification, with initial disconnection between metal atoms and the organic structural units, results in a match with **fff** topology and **b.** all/single node matches with **pcu**. Wireframe structures show C (grey), O (red), Zn (blue), which are simplified to metal nodes (red), and organic nodes (green) connected by straight edges representative of linkers (blue). 24
2.6 A metal-oxo deconstruction, shown as a schematic diagram, performed on CSD SAHYIK [66].In the original structure (left), C (grey), O (red), and Zn (violet). This technique draws many similarities to the single and all node approaches, but with a focus on structure chemistry showing the resultant (middle) Zn metals (red) and 1,3-benzenedicarboxylate linkers (green). . . 25
2.7 A timeline to show the emergence of selected experimental MOF datasets following the release of the first hypothetical MOF database (hMOF) [69] in 2012. Circle size varies to represent the relative size of the database, colour is representative of the study/research that produced the resource. 26
2.8 Distribution of selected atom-atom bonded (blue) and non-bonded (orange) contacts (out to VdW+0.0) in the CSD. a. Ag, b. Hg, c. Cd, and d. Sn. Dashed red boxes suggest contentious atom-atom bonding ranges. 28
2.9 Atomic level representations of CSD ZEHMOQ showing a. the original structure set at a 3.32 Å Ag-Ag bond distance limit and b. an auto-modified version with a 3.35 Å Ag-Ag bond distance limit, where the connectivity has been calculated using automatic bond assignment tools within CSD Mercury. 28
2.10 A snapshot of the online interface of the TopCryst web topology service used for the automatic deconstruction of CSD SAHYIK. The original CIF was modified with the use of CSD's Python API solvent removal script. . 30
2.11 A snapshot of the online interface of MOFid's web structure identification and topology tool performing a structure simplification on CSD SAHYIK by uploading the raw CIF. 32
2.12 A snapshot of the online interface of MOFid's web structure identification and topology tool, showing the options available for each uploaded CIF file. 35

3.1 A histogram depicting the most commonly obtained 2D topologies, based off the CrystalNets Input results (blue), but also showing the occurrences for CrystalNets Guess (orange), and MOFid (grey). 51
3.2 A histogram depicting the most commonly obtained 3D topologies, based off the CrystalNets Input results (blue), but also showing the occurrences for CrystalNets Guess (orange), and MOFid (grey). 53
3.3 A screenshot demonstrating the implemented CrystalNets process within CSD Mercury 56
3.4 An output file showing the result of crystal structure analysis, including the CrystalNets process, on SAHYIK launched within CSD Mercury, determining the structure has the **pcu** topology. 57

4.1 Schematic showing the applications of gas separation where CO_2 is captured and methane is separated (left), and the storage of gaseous methane (right) in MOFs. The structures here are represented in a general form where a typical metal-oxo cluster is seen as a metal node, and the organic linkers are drawn as straight connecting bars. 65

4.2 A flow diagram which shows the process of developing suitably precise parsers for data extraction by text mining [21], licenced under CC BY NC 4.0. 70
4.3 A flow diagram demonstrating the classification of some basic, and some more complex, machine learning model types. 74
4.4 A Web of Science search for trends in publications using the key words 'metal-organic framework', 'synthesis', and 'machine learning' as of July 2022. (https://www.webofscience.com/wos/woscc/citation-report/c0f28728-bd2b-4392-834e-7bc24ac6334b-474cb7be) 75
4.5 Input topologies of novel experimental MOFs for use in an inverse design algorithm targeting structures for top performance in carbon capture applications. Reprinted with permission from [19]. Copyright 2020 American Chemical Society. 78
4.6 A comparison of the trial and error approach versus training machine learning models to predict synthesis conditions [18], licenced under CC BY 4.0. 81
4.7 A schematic showing the criteria used to differentiate the positive (P) and unlabelled (U) data. Reprinted (adapted) with permission from [34]. Copyright 2022 American Chemical Society. 82
4.8 A universal system for the automatic execution of chemical synthesis from literature. Extraction of the procedure is followed by an algorithmic process for producing the code that conforms to a standard hardware and software architecture. Manual error correction and simulated execution ensure reliability and safety [53]. 83
4.9 Skeletal structures of T2, P2, T2E. These specific structures are used by Pyzer-Knapp et al. [59] as example materials in their study to accelerate the computational discovery of porous solids through improved navigation of ESF maps, licenced under CC BY 4.0. 85
4.10 Schematic of the feedback loop for data generation for the Suzuki-Miyaura reaction [60], licenced under CC BY 4.0. 86
4.11 Process flow diagram for the automatic synthesis of quantum dots from initial random experimentation, using flow synthesis and real-time data processing, to new experimental selection [64]. 89
4.12 Schematic of the reactor design for the continuous synthesis of lead halide perovskite quantum dots [65], licenced under CC BY 4.0. 90
4.13 Setup for continuous synthesis of HKUST-1 using the millifluidic reactor. The equipment as shown above: 1. Syringe pump, 2. Silicone oil in continuous phase, 3. Reactant solution in dispersed phase, 4. ETFE Tee, 5. 3D printed anchor, 6. Grooved aluminium block, 7. Hot plate, 8. Product collection vial [69], licenced under CC BY NC ND 4.0. 93
4.14 Process flow diagram for the automated synthesis of ZIF-67, showing the use of Bayesian optimisation (BO) to make continual variation to the chosen variables after the output analysis is performed. Reprinted (adapted) with permission from [71]. Copyright 2021 American Chemical Society. . . 94

5.1 Flow diagram to visualize the integration of CDE into a data-driven MOF synthesis plan: from article retrieval to text mining, computational screening, and materials discovery. 108

5.2 Flow diagram to visualize the integration of CDE into a data-driven MOF synthesis plan: from article retrieval to text mining, computational screening, and materials discovery. 109
5.3 Flow chart displaying possible outcomes when fed an input string for high-throughput MOF name parsing. 111
5.4 Collection of the top 30 organic linkers obtained via text-mining the CSD MOF subset chemical names. Hit counts (C) and CAS numbers are included for each linker. 113
5.5 (a) 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. 116
5.6 Histograms of topological types extracted from the CSD MOF subset using (a) ChemDataExtractor (CDE) (b) CrystalNets in 3D structures. (c) Top five most common 3D topologies: **pcu**, **dia**, **pts**, **rtl**, and **cds**. 117
5.7 Comparison of different topologies in the structure space for LCD as a function of void fraction for ca. 2200 porous MOFs. There are 241 structures with **pcu** topology (green); 170 **dia** (purple), 41 **stp** (red), 33 **rob** (yellow), and 32 **fsc** (orange) structures. All other structures are shown in pale blue. 118
5.8 (a) Histogram showing the most commonly occurring single linkers found in the 3D MOF subset for non-zero LCD values. (b) Box and whisker plot of linker length versus the LCD/PLD ratio across a sample of ca. 8000 MOFs. (c) Box and whisker plot of linker types against LCD for a sample of linkers with one (orange) and two or more (blue) blocks. 120
5.9 (a) Histogram of the most frequently occurring single metals found in the 3D MOF subset. (b) Comparison of the constituent metals against the LCD of structures. 121
5.10 Bar charts showing the cost per gram of organic linkers as determined by averaging the available quantities. A selection of the most prevalent linker types was chosen from the DigiMOF database for (a) low-cost and (b) high-cost linkers. Prices obtained from TCI Chemicals [54]. 123

6.1 After scanning the QR code using a smart phone, Cu-BTC is projected in augmented reality (AR), demonstrated on the table (20 cm diameter) and in much larger scale in an office with a student standing inside the pores (2 m diameter). Here, the pores contain water molecules where the structure was simulated for water adsorption. 135
6.2 Graphical representation of the AR visualization workflow for MOF adsorbents. We begin from initial structure selection in CSD ConQuest followed by exportation of the unit cell for use in RASPA via Mercury, structure cleanup, file format conversions in Jmol, modeling and export in Blender, and finally upload to the p3d.in platform and generation of a custom QR code. 137
6.3 CO2 adsorption isotherm simulated in UiO-67 at 298 K. Adsorption snapshots are highlighted at 0.15, 5.5, and 20 bar. QR codes for AR visualization are located adjacent to each snapshot. 140

A.1 MOF CDE parser performance compared with previous versions of CDE [1, 2] and the work from Park et al. [3] MOF text mining tool. Performance of individual parsers and detailed methodology for calculation of these metrics is available in the supporting information, Table A.3. 151
A.2 Proportion of synthesis methods present in the MOF Database. 156
A.3 A histogram displaying the 25 most extracted strings marked up as metal precursors. 157
A.4 Clustered columns reflecting the top five topological allocations to a. the top five linker types, and b. the top five metal clusters. 157
A.5 Top 20 topologies versus the LCD/PLD ratio in descending order of frequency for structures with PLD $>$ 0.55 Å. 158
A.6 A box and whisker plot of linker length, as categorised by the number of aromatic rings, against LCD/PLD ratio for all porous MOFs which were assigned **pcu** topology. 159
A.7 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. 160

List of Tables

3.1 A general overview of the output of each approach on 2D and 3D structures before detailed analysis was carried out. 50
3.2 A review of the output of 2D structure analysis. 51
3.3 A review of the output of 3D structure analysis. 52

5.1 Total Number of Extracted Properties and the Number of Unique Properties for Each MOF Property in the DigiMOF Database 114
5.2 Total Number of Extracted Properties and Number of Unique Properties for Structures in the 3D MOF Subset 115

6.1 Selection of MOFs Visualized in AR under Different Conditions 138

A.1 Parsing elements used to create the rule-based grammars to identify MOF names and corresponding topology, solvent, synthesis route, organic linker, and/or metal precursor [1]. 150
A.2 Examples of compound records from previous versions of CDE, the previous attempt to text mine MOF data and this work. 150
A.3 Summary of the performance of each individual parser. 152
A.4 Simplified MOF CDE Regular Expression (Regex) examples. 153
A.5 Simplified examples of organic linker exclusion list item regular expression development. Note that compound types (such as MOF names and metal precursors) were also added to this exclusion list. 155

Chapter 1

Introduction

Metal-organic frameworks (MOFs) have emerged as a widely studied class of porous materials over the last 30 years. They have been, and continue to be, extensively researched for applications requiring porous and adsorptive properties [1, 2, 3, 4]. Due to their porosity and large surface area, MOFs have been effectively deployed for adsorption of gases such as CO_2, H_2O, and H_2 [5, 6, 7]. This, coupled with the customisability of MOFs has attracted exponentially growing interest in this realm of materials science, and recent investigations have shifted from gas storage and separation to resistive sensing, electrocatalysis, and energy storage [8, 9]. Materials combining structural tunability, accessible porosity, and high surface area are highly desirable from an application point of view, however this combination of properties can be difficult to find especially when additional criteria such as conductivity, or water stability are considered [10].

Over the past several years there has been a big increase in the number of computationally theorised structures available to researchers [11], this rapid development has also led to concerns over chemical diversity and the need for screening a large number of chemically similar materials arises [12]. This has been in part due to the development of computing capabilities, but also the increasing influence of machine learning on the community [13].

This thesis focuses on three main topics which sit at the forefront of MOF research: reliable topological characterisation of crystalline materials, machine learning and data driven manufacturing, and the development of new computational tools. This introduction discusses the basics of metal-organic frameworks and some important literature, followed by an overview of the work which has been completed by the Cambridge Crystallographic Data Centre (CCDC) and collaborators resulting in the creation of the Cambridge Structural Database (CSD) MOF subset. Then, the aims and objectives of this PhD project are outlined with some contextual literature review surrounding the key areas discussed in each subsequent section.

1.1 Metal-Organic Frameworks

MOFs are crystalline materials that consist of metal clusters (Secondary Building Units, SBUs) connected by organic linkers which combine structurally to form well defined porous frameworks. Typically formed into networks of 1D-chains, 2D-sheets, or 3-D nets, MOF building units can be connected in an almost limitless number of ways [14]. When considering their formation, they can consist of almost any metal combined with a broad

catalogue of organic molecules. These materials often exhibit strong mechanical stability in conjunction with porosity, and the modular nature of structure formation enables versatile material development to tune desirable properties for a range of applications.

The origins of MOFs lie in the mid 1990s with notable contributions from Omar M. Yahgi and colleagues upon their exploration of the molecular building block approach of synthesising crystal structures [15, 16]. However, several materials that could be considered MOFs had been synthesised before this point, one example of which was the copper based $Cu[C(C_6H_4.CN_4)]BF_4$ reported by Hoskins and Robson in 1989 [17]. Since then however, the MOF field has been subject to rapid investment resulting in a significant growth in research volume, exploring various strategies used to create a vast library of MOF structures. This is in part due to the versatility of MOFs, allowing for precise control over properties, but also due to advancements in synthesis methods such as hydrothermal, solvothermal, microwave assisted, and mechano-chemical synthesis playing a vital role in accelerated development.

One key property of all crystal structures, and the focus of this thesis, is topology - explained simply as the way that a crystal has been formed in space - and it is represented by the underlying connectivity of constituent building units [18]. It holds significant importance as the underlying atomic configuration directly influences the properties and performance of the material. Understanding topology, and the ability to control specific formations, is crucial for tailoring pore structure, surface area, and mechanical stability to meet predetermined requirements. This interest in pore shape and size has been extensively studied, particularly to consider the upper bounds (or the limit) of these materials, as well as the pore shape restrictions that may be imposed [19, 20]. Achieving desired topologies can be challenging due to several factors including linker geometry, metal coordination preferences, and a variety of synthesis conditions and methods which have stimulated the development of modulator-assisted methods, ligand design strategies, and post-synthesis modifications in an attempt to control the resultant topologies.

A further geometric descriptor of MOFs and other crystalline structures are tilings, which refer to divisions of space in which repeating polyhedral structures describe 3D arrangements of MOF networks, typically used to represent the regions which are enclosed by the metal nodes and linkers. These tiles, often regular or semi-regular polyhedra, such as tetrahedra, octahedra, or more complex polyhedra, give a clearer picture of the pores or cavities present in the structure, crucial for understanding guest-accessibility, gas storage potential, or catalytic behaviour. A common MOF, MOF-5, consists of metal-oxo clusters that form cubic cages, and the organic linkers span between the nodes, such that the entire structure can be described as a tiling of cubic units.

A wide range of topological diversity can be found in the broader MOF space, typically described by mathematical nets found in the Reticular Chemistry Structure Resource (RCSR) [21], one example, MOF-5 and its underlying topology, is shown in Figure 1.1. This example, amongst thousands of others, can be found in the Cambridge Structural Database (CSD).

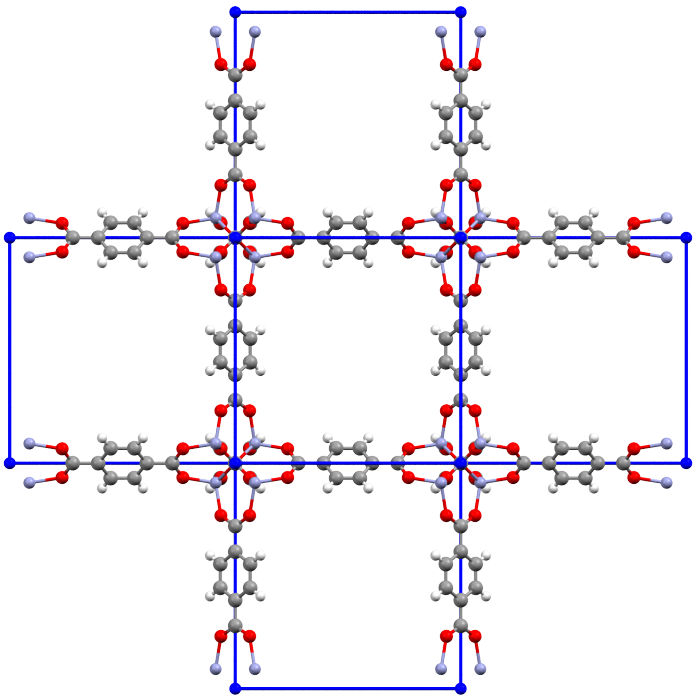


Figure 1.1: *A representation of one of the most renowned MOFs (MOF-5), with* ***pcu*** *topology overlaid onto the structure in CSD Mercury.*

1.2 Cambridge Structural Database and MOF Subset

The Cambridge Structural Database (CSD) is a comprehensive and curated repository for the experimental structural data of molecules (typically crystalline in nature) which make up organic, metal-organic, and inorganic compounds. One of the most extensive chemical databases to exist, it currently contains approximately 1.2 million entries and counting, all of which generally contain at least carbon and hydrogen (with some exceptions). The CSD serves as a valuable resource for chemists, crystallographers, and materials scientists and contains detailed structural information such as atomic coordinates, bond lengths, angles, and space groups obtained from x-ray (XRD), electron or neutron diffraction experiments.

The CSD is a widely used repository for small-molecule and metal-organic structures, all of which are typically made available for download and use at the point of publication. Researchers are able to access the CSD to analyse molecular conformations, study intermolecular interactions, and run experiments to determine structure-property relations. Targeted subsets, such as the CSD MOF subset or the COVID-19 subset, are also available to support researchers with specific focuses.

The CSD MOF subset is a specialised partition of the whole CSD, created in 2017 by Moghadam et al. [22] and it has been extensively used by almost all MOF focused research groups as a resource containing almost all experimentally produced MOFs. By focusing on a curated collection of data that is specific to a certain class of materials, it allows streamlined research into synthesis, stability, and design of MOFs. Whilst many CSD derived and/or hypothetical MOF databases exist, they are limited by their lack of periodic updates and high-level manual curation that the CCDC provides, this includes

manual verification of deposits from a variety of sources including checks for errors and inconsistencies followed by the use of periodic quality and consistency checks.

Figure 1.2 shows the seven criteria used to created the original CSD MOF subset that are still used to filter newly deposited structures into the CSD MOF database following each update.

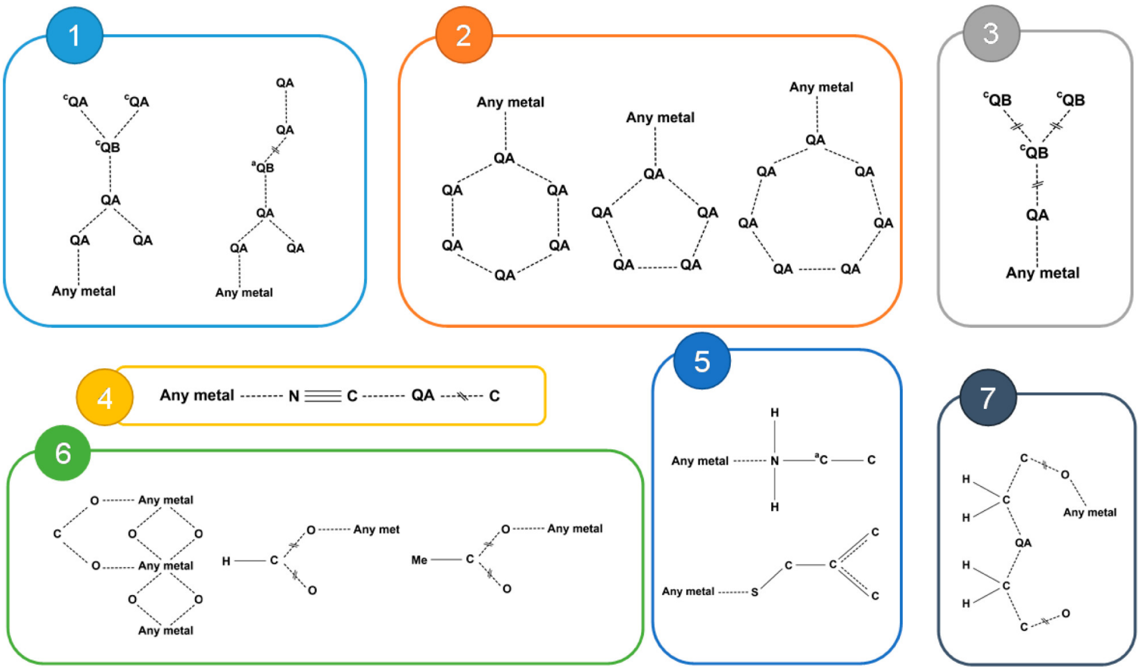


Figure 1.2: *Summary of the seven criteria designed to build the CSD MOF subset, where QA = O, N, P, C, B, S. QB = N, P, B, S, C and superscripts "c" and "a" impose the corresponding atoms to be "cyclic" or "acyclic", respectively. Me denotes methyl groups. The dotted line refers to any of the bond types stored in the CSD (single, double, triple, quadruple, aromatic, polymeric, delocalised, and pi). The dotted line with the two lines through indicates a variable bond type (i.e., two or more of the options above). In these cases, the variable type is single, double, or delocalised.* [22]

The CSD MOF subset has been used extensively throughout the work contained within this thesis as the go-to database for experimental MOF data. Whilst other MOF databases do exist, and some of them are discussed in later chapters, they are all hindered by a lack of curation that solidifies the CSD as a leading, up-to-date, and trustworthy source of information.

1.3 Thesis Outline

1.3.1 Chapter 2 - Topological Characterisation of Metal-Organic Frameworks

In chapter 2 we introduce the significance of topological analysis and how it can be used to understand metal-organic frameworks (MOFs) and their growing importance in materials science. With over 120,000 MOF-like structures deposited into the CSD database, the complexity of these structures presents challenges in characterisation. Topological analysis simplifies MOF structures by identifying their basic connectivity, which can aid

in design and synthesis. Software tools like ToposPro, MOFid, and CrystalNets assist in assigning topology descriptors to MOFs, though each has its strengths and limitations. This perspective highlights the importance of topology in MOFs, discusses available software, their algorithmic approaches and methods, limitations, and uptake within the MOF community.

In the conclusion, we emphasise that topological characterisation extends beyond MOFs to other crystalline materials like COFs and Zeolites. The choice of software depends on study requirements; CrystalNets offers speed but limited chemical insights compared to MOFid or ToposPro. A notable limitation is discrepancies in topology allocation, which should be addressed following IUPAC guidelines. The lack of a complete, freely available database of MOFs with verified topology information is highlighted, abd although initiatives like the QMOF database show promise for some significant applications more work is required. Suggestions are made to integrate topological information into established databases like the CSD, enhancing accessibility and facilitating research in MOF characterisation.

1.3.2 Chapter 3 - Integrating CrystalNets.jl and Bench-marking Performance on the Cambridge Structural Database

In this chapter we aimed to address the challenges in topological assignment of complex MOF structures by comparing two high-throughput topological assignment packages, MOFid and CrystalNets, supplemented by a custom Python workflow utilizing the CSD Python API. By analysing a large set of CSD 2D and 3D MOFs (54,473 experimental structures), the study aimed to assess the agreement between these approaches and identify the most effective method for topological assignment. This comparative analysis, believed to be the first of its kind using the CSD MOF subset, led to the development of a new Python-based approach integrated within CSD Mercury, facilitating topological assignment with a single click. The investigation aimed to determine the most suitable, presently available, workflow for high-throughput topological characterisation of MOFs.

We concluded that our results indicated that combining CrystalNets with the retention of CSD bonding information yielded the highest recall and precision rates, with minimal computational cost compared to the MOFid approach. The automated workflow achieved success rates of 75.9% for 2D MOFs and 51.25% for 3D MOFs, though room for improvement remains, particularly regarding disorder in structures and modifications due to solvents within CIFs. Additionally, the study emphasized the simplicity of integrating new tools within the CSD, providing the developed workflow publicly through CCDC's open-source GitHub repository, enhancing accessibility for researchers seeking efficient topological assignment methods.

1.3.3 Chapter 4 - Machine Learning and Digital Manufacturing Approaches for Solid-state Materials Development

Chapter 4 serves as an introduction to the integration of machine learning (ML) into the realm of material discovery and chemical manufacturing, which whilst relatively recent shows promising progress, particularly in solid-state nanomaterials. ML has been successfully applied to predict novel synthesis conditions, enhancing sustainability and economic efficiency. However, challenges such as inconsistent reporting and the lack of data on unsuccessful experiments hinder the accuracy of predictions. We discuss the shift towards open-source data repositories and collaborative efforts within the scientific

community which aim to address these challenges, making research more accessible, affordable, and environmentally conscious. With the availability of resources like journal publications, GitHub repositories, and physical experimental documentation, the use of ML in chemical space is expected to accelerate.

In this chapter we concluded that advancements such as knowledge graphs, graph-based reaction optimisation, and digital twins are facilitating the discovery of new pathways to stimulate the automation of laboratory processes. In the context of large-scale integration of ML and digitisation for synthesis of metal-organic frameworks (MOFs), high-quality algorithm outputs are crucial for continuous manufacturing and commercialisation. Predicted synthesis conditions will allow manufacturers to optimise resource usage and production processes, thereby reducing costs and environmental impact. However, challenges such as the complexity of certain MOF structures and the scarcity of certain materials pose limitations on large-scale synthesis. Events like the COVID-19 pandemic have underscored the importance of assessing the resilience of manufacturing pathways to disruptions in supply chains, highlighting the need for digitisation to increase efficiency and resilience. Nonetheless, the full potential of digitisation relies on access to sufficient, high-quality data, which remains a critical aspect for further progress.

1.3.4 Chapter 5 - DigiMOF: A Database of Metal-Organic Framework Synthesis Information Generated via Text Mining

The work in chapter 5 outlines the development of the DigiMOF database, an open-source resource generated through the adaptation of the ChemDataExtractor (CDE) tool to extract synthetic properties from over 43,000 MOF publications. A continuation from the adoption of ML within the chemical space, this database encompasses various parameters such as synthesis method, solvent, linker type, metal precursor, and topology, providing a centralised repository of valuable information for MOF researchers. Additionally, an alternative data extraction technique was employed to identify linker types and their associated costs, further enriching the database. DigiMOF offers a comprehensive resource for researchers to rapidly search for MOFs with specific properties, aiding both computational screening and experimental evaluation of MOF properties. The database and associated software are openly available, allowing for updates and modifications to ensure the continuous identification of new MOF-property relationships.

In the conclusion, the significance of DigiMOF in advancing the field of MOF research is reiterated, as one of the first automatically generated database of MOF synthesis properties, it offers a valuable resource for researchers seeking to enhance MOF production pathways and identify commercially viable synthesis routes. By providing over 15,000 unique MOF records with detailed synthetic data, DigiMOF now facilitates techno-economic assessments, life-cycle assessments, and experimental validation work. The database is poised to reduce the reliance on unsustainable synthesis routes within the MOF community to help stimulate the application of MOFs in decarbonisation technologies. With its extensive dataset and potential, subject to further work, for continual updates, DigiMOF now serves as a foundational tool for advancing the digital manufacturing of MOFs and driving innovation in materials science.

1.3.5 Chapter 6 - Augmented Reality for Enhanced Visualization of MOF Adsorbents

Finally, chapter 6 explores the innovative use of augmented reality (AR) technology to enhance the visualisation and understanding of complex 3D materials, particularly metal-organic frameworks (MOFs). The paper demonstrates a workflow for AR modeling that allows for the visualisation of MOF crystal structures, topologies, and gas adsorption sites directly on Android or iOS smartphones without the need for additional apps. The technique is not only beneficial for computational and experimental scientists in research but also serves educational purposes, offering an engaging and interactive experience for students and researchers alike.

The paper showcases the practical applications of AR modeling for MOFs and by providing freely available, no-cost methods, distributed via QR codes, this technique enables the creation and sharing of AR models globally and instantly. The ability to modify the size of AR representations adds to its versatility, allowing for educational use in various settings, such as conferences, workshops, and classroom presentations. Additionally, the paper explores potential applications of AR in other areas of materials science, including catalysis, crystal engineering, and collaboration between research groups, as well as artistic experiences of crystal structure representations. Overall, this research paper highlights the potential of AR technology to revolutionise the way we visualise, teach, and understand the field of materials science.

1.4 Aims, Objectives and Scientific Contribution

Before this project commenced, the CSD consisted of approximately 88,000 MOFs, a figure which has risen over the past 4 years to an incredible 120,000+. The goal of this project was to conduct systematic computational study into the network topology of MOFs and to implement data-driven design approaches to explore the library of MOFs within the CSD.

With unparalleled potential to investigate thousands of structures in a short time, computational high-throughput screening is extremely well suited to unravelling trends in key MOF properties, establish structure-property relationships, and guide future synthesis efforts. Whilst in the last few years CCDC's computational analysis has been primarily focused on geometric characterisation (e.g largest pore size, pore volume, surface area, and gas adsorption properties) we further these investigations using a variety of new approaches. Previous computational investigations have delivered important insights but are still yet to answer key questions about useful MOF properties such as performing reliable topological characterisation of the CSD MOF subset, and identifying constituent metals and organic linkers. To address these, we proposed this project which is the result of a combination of efforts including the expertise of CCDC's scientists in crystallography and materials science in conjunction with researchers here at the University of Sheffield and at University College London.

1.4.1 Topological Characterisation

Our initial objective was to perform topological characterisation of all MOFs within the CSD MOF subset, including investigating the methods and algorithms through which they can be categorised. We first planned to develop a new standalone software to handle

this problem that could be integrated within the CSD, however upon release of the open source CrystalNets package, only minor modifications to the run process were required to significantly increase the reliability of the current state of the art tools. This was achieved through a combination of the work detailed in Chapter 2, Chapter 3, and Chapter 5.

1.4.2 Machine Learning and Augmented Reality

Next, quantitative structure/property relationship (QSPR) analysis was used to systematically correlate certain topological descriptors to functional properties in quantitative terms. Before the commencement of this project, only a handful of QSPR studies had been reported for porous materials, and we developed new models using established methods including the use of machine learning in the form of natural language processing (NLP). We developed several publicly available resources for analysis of information related to all materials. Our state-of-the-art machine learning (ML) and digital manufacturing techniques were discussed in Chapter 4 and Chapter 5 with the view that they can be applied to the future of the field. Our focus began with the use of ML in solid state materials development, followed the developed a new method of abstracting existing synthesis information published in thousands of previous MOF studies.

1.4.3 Development of New Computational Tools

Finally, we proposed the integration of a newly developed open-source software with the Cambridge Crystallographic Data Centre's (CCDC) existing crystallographic data suite and Python API, and we applied new augmented reality techniques to visualise the results of topological deconstruction and adsorption studies of MOFs. These goals were met through a combination of the work detailed in Chapter 3 and Chapter 6.

1.5 Summary

The result of this work has enabled researchers to use the Cambridge Structural Database to maximum effectiveness when searching for synthesis conditions, precursors, linker types, topologies, and more whilst also integrating ML techniques such as Natural Language Processing (NLP) for data mining and introducing new ways of visualising the results.

References

[1] Hiroyasu Furukawa, Kyle E. Cordova, Michael O'Keeffe, and Omar M. Yaghi. The Chemistry and Applications of Metal-Organic Frameworks. *Science*, 341(6149):1230444, August 2013. Publisher: American Association for the Advancement of Science.

[2] Guorui Cai, Peng Yan, Liangliang Zhang, Hong-Cai Zhou, and Hai-Long Jiang. Metal–Organic Framework-Based Hierarchically Porous Materials: Synthesis and Applications. *Chem. Rev.*, 121(20):12278–12326, October 2021. Publisher: American Chemical Society.

[3] Gyudong Lee, Dong Kyu Yoo, Imteaz Ahmed, Hye Jin Lee, and Sung Hwa Jhung. Metal-organic frameworks composed of nitro groups: Preparation and applications in adsorption and catalysis. *Chemical Engineering Journal*, 451:138538, January 2023.

[4] Xueqin Li, Kai Chen, Ruili Guo, and Zhong Wei. Ionic Liquids Functionalized MOFs for Adsorption. *Chem. Rev.*, 123(16):10432–10467, August 2023. Publisher: American Chemical Society.

[5] Jian-Bin Lin, Tai T. T. Nguyen, Ramanathan Vaidhyanathan, Jake Burner, Jared M. Taylor, Hana Durekova, Farid Akhtar, Roger K. Mah, Omid Ghaffari-Nik, Stefan Marx, Nicholas Fylstra, Simon S. Iremonger, Karl W. Dawson, Partha Sarkar, Pierre Hovington, Arvind Rajendran, Tom K. Woo, and George K. H. Shimizu. A scalable metal-organic framework as a durable physisorbent for carbon dioxide capture. *Science*, 374(6574):1464–1469, December 2021. Publisher: American Association for the Advancement of Science.

[6] Christopher A. Trickett, Aasif Helal, Bassem A. Al-Maythalony, Zain H. Yamani, Kyle E. Cordova, and Omar M. Yaghi. The chemistry of metal–organic frameworks for CO2 capture, regeneration and conversion. *Nat Rev Mater*, 2(8):1–16, July 2017. Number: 8 Publisher: Nature Publishing Group.

[7] Mark D. Allendorf, Vitalie Stavila, Jonathan L. Snider, Matthew Witman, Mark E. Bowden, Kriston Brooks, Ba L. Tran, and Tom Autrey. Challenges to developing materials for the transport and storage of hydrogen. *Nat. Chem.*, 14(11):1214–1223, November 2022. Number: 11 Publisher: Nature Publishing Group.

[8] Lauren E. Kreno, Kirsty Leong, Omar K. Farha, Mark Allendorf, Richard P. Van Duyne, and Joseph T. Hupp. Metal–Organic Framework Materials as Chemical Sensors. *Chem. Rev.*, 112(2):1105–1125, February 2012. Publisher: American Chemical Society.

[9] Jack Gonzalez, Krishnendu Mukherjee, and Yamil J. Colón. Understanding Structure–Property Relationships of MOFs for Gas Sensing through Henry's Constants. *J. Chem. Eng. Data*, 68(1):291–302, January 2023. Publisher: American Chemical Society.

[10] Cao Xiao, Jindou Tian, Qihui Chen, and Maochun Hong. Water-stable metal-organic frameworks (MOFs): rational construction and carbon dioxide capture. *Chem Sci*, 15(5):1570–1610, January 2024.

[11] Sangwon Lee, Baekjun Kim, Hyun Cho, Hooseung Lee, Sarah Yunmi Lee, Eun Seon Cho, and Jihan Kim. Computational Screening of Trillions of Metal–Organic Frameworks for High-Performance Methane Storage. *ACS Appl. Mater. Interfaces*, May 2021. Publisher: American Chemical Society.

[12] Seyed Mohamad Moosavi, Aditya Nandy, Kevin Maik Jablonka, Daniele Ongari, Jon Paul Janet, Peter G. Boyd, Yongjin Lee, Berend Smit, and Heather J. Kulik. Understanding the diversity of the metal-organic framework ecosystem. *Nature Communications*, 11(1):4068, August 2020. Number: 1 Publisher: Nature Publishing Group.

[13] Sanggyu Chong, Sangwon Lee, Baekjun Kim, and Jihan Kim. Applications of machine learning in metal-organic frameworks. *Coordination Chemistry Reviews*, 423:213487, November 2020.

[14] Nathan W. Ockwig, Olaf Delgado-Friedrichs, Michael O'Keeffe, and Omar M. Yaghi. Reticular Chemistry: Occurrence and Taxonomy of Nets and Grammar for the Design of Frameworks. *Acc. Chem. Res.*, 38(3):176–182, March 2005. Publisher: American Chemical Society.

[15] Hailian Li, Mohamed Eddaoudi, M. O'Keeffe, and O. M. Yaghi. Design and synthesis of an exceptionally stable and highly porous metal-organic framework. *Nature*, 402(6759):276–279, November 1999.

[16] Omar M. Yaghi, Michael O'Keeffe, Nathan W. Ockwig, Hee K. Chae, Mohamed Eddaoudi, and Jaheon Kim. Reticular synthesis and the design of new materials. *Nature*, 423(6941):705–714, June 2003. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6941 Primary_atype: Reviews Publisher: Nature Publishing Group.

[17] Bernard F. Hoskins and Richard Robson. Infinite polymeric frameworks consisting of three dimensionally linked rod-like segments. *J. Am. Chem. Soc.*, 111(15):5962–5964, July 1989. Publisher: American Chemical Society.

[18] Frank Hoffmann. *Introduction to Crystallography*. Springer International Publishing, Cham, 2020.

[19] Gérard Férey, Caroline Mellot-Draznieks, Christian Serre, and Franck Millange. Crystallized frameworks with giant pores: are there limits to the possible? *Acc Chem Res*, 38(4):217–225, April 2005.

[20] Alexander P. Shevchenko, Eugeny V. Alexandrov, Andrey A. Golov, Olga A. Blatova, Alexandra S. Duyunova, and Vladislav A. Blatov. Topology versus porosity: what can reticular chemistry tell us about free space in metal–organic frameworks? *Chem. Commun.*, 56(67):9616–9619, August 2020. Publisher: The Royal Society of Chemistry.

[21] Michael O'Keeffe, Maxim A. Peskov, Stuart J. Ramsden, and Omar M. Yaghi. The Reticular Chemistry Structure Resource (RCSR) Database of, and Symbols for, Crystal Nets. *Acc. Chem. Res.*, 41(12):1782–1789, December 2008. Publisher: American Chemical Society.

[22] Peyman Z. Moghadam, Aurelia Li, Seth B. Wiggin, Andi Tao, Andrew G. P. Maloney, Peter A. Wood, Suzanna C. Ward, and David Fairen-Jimenez. Development of a

Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future. *Chem. Mater.*, 29(7):2618–2625, April 2017. Publisher: American Chemical Society.

Chapter 2

Topological Characterisation of Metal-Organic Frameworks: A Perspective

2.1 Publication Information and Paper Contributions

This paper has been submitted to the American Chemical Society's journal Chemistry of Materials for publication as a perspective and is currently under review.

In this publication, I, the candidate wrote the manuscript with supervision from Professor Joan L. Cordiner, Dr Jason C. Cole, and Dr Peyman Z. Moghadam.

2.2 Abstract

Metal-organic frameworks (MOFs) began to emerge over two decades ago, resulting in the deposition of 120,000 MOF-like structures (and counting) into the Cambridge Structural Database (CSD). Topological analysis is a critical step towards understanding periodic MOF materials, offering insight into the design and synthesis of these crystals via simplification of connectivity imposed on the complete chemical structure. Whilst some of the most prevalent topologies such as face-centred cubic (**fcu**), square lattice (**sql**), and diamond (**dia**) are simple and can be easily assigned to structures, MOFs that are built from complex building blocks, with multiple nodes of different symmetry, result in difficult to characterise topological configurations. In these complex structures representations can easily diverge where the definition of nodes and linkers are blurred, especially for cases where they are not immediately obvious in chemical terms. Currently, researchers have the option to use software such as ToposPro, MOFid, and CrystalNets to aid in the assignment of topology descriptors to new and existing MOFs. These software packages are readily available and are frequently used to simplify original MOF structures into their basic connectivity representations, before algorithmically matching these condensed representations to a database of underlying mathematical nets. These approaches often require the use of in-built bond assignment algorithms alongside the simplification and matching rules. In this perspective, we discuss the importance of topology within the field of MOFs, the methods and techniques implemented by these software packages, their availability and limitations, and review their uptake within the MOF community.

2.2.1 Keywords

Perspective, Topology, Metal-organic Frameworks, Structure Characterisation

2.3 Introduction

Metal-organic frameworks (MOFs) are an emerging class of porous materials, formed by chemical bonds between metal clusters and organic building blocks [1, 2]. MOFs are a diverse set of chemical structures often characterised by their porosity and customisability: the commercial uptake of MOFs are particularly focused towards gas adsorption [3, 4], separation [5, 6, 7], sensing [8, 9], alongside catalysis [10, 11] and quantum applications [12, 13, 14, 15]. The MOF materials space consists of many combinations of building units typically configured in a symmetrical pattern. Over time, increased importance has been placed on topology as a predictor of properties: recently investigations have been published that compare topology with porosity and mechanical stability [16, 17], but there are still areas in which potential correlations between topology and other properties have not been determined, such as electronic properties, solvent compatibility, and thermal stability [18].

The CSD MOF subset contains a staggering ca. 120,000 experimental crystal structures of MOFs (CSD release April 2023), representative of the input of the worldwide research community, with updates to the total number of synthesised structures being made quarterly [19, 20, 21]. Figure 2.1 shows the distribution of MOFs within the CSD from 1981 to present day, including a breakdown of their structural dimensionalities. Whilst there appears to have been a clear preference towards the synthesis of 1D MOF-like structures from the inception of the CSD until 2011, there has been a recent increase in the popularity of 3D structures compared to the initial high proportion of 1D deposits. The initial prevalence of 1D MOFs could be explained by the cost-effective formation of simple structures consisting of basic pyridyl and chelate ligands, typically synthesised with the intention to study these ligands and their interactions with metal centres. These 1D chains have interesting applications in magnetism, proton conductivity, and ferroelectricity and can often form larger crystals than equivalent 2D and 3D structures under ambient conditions. We note that, despite their dimensionality, these structures can exhibit porosity when linked by hydrogen bonds or other interactions, when woven together/interpenetrating (1D+1D), or they could potentially exhibit porosity on desolvation [22]. 3D MOFs are typically considered to be the ideal candidates for adsorption applications and the increasing focus on 3D MOFs can be seen in the cumulative 3D structure deposits (red line in Figure 2.1) where they begin to overtake 2D submissions in 2015. The number of 3D MOF submissions to the CSD has consistently exceeded 1000 accepted annual deposits for the last 15 years.

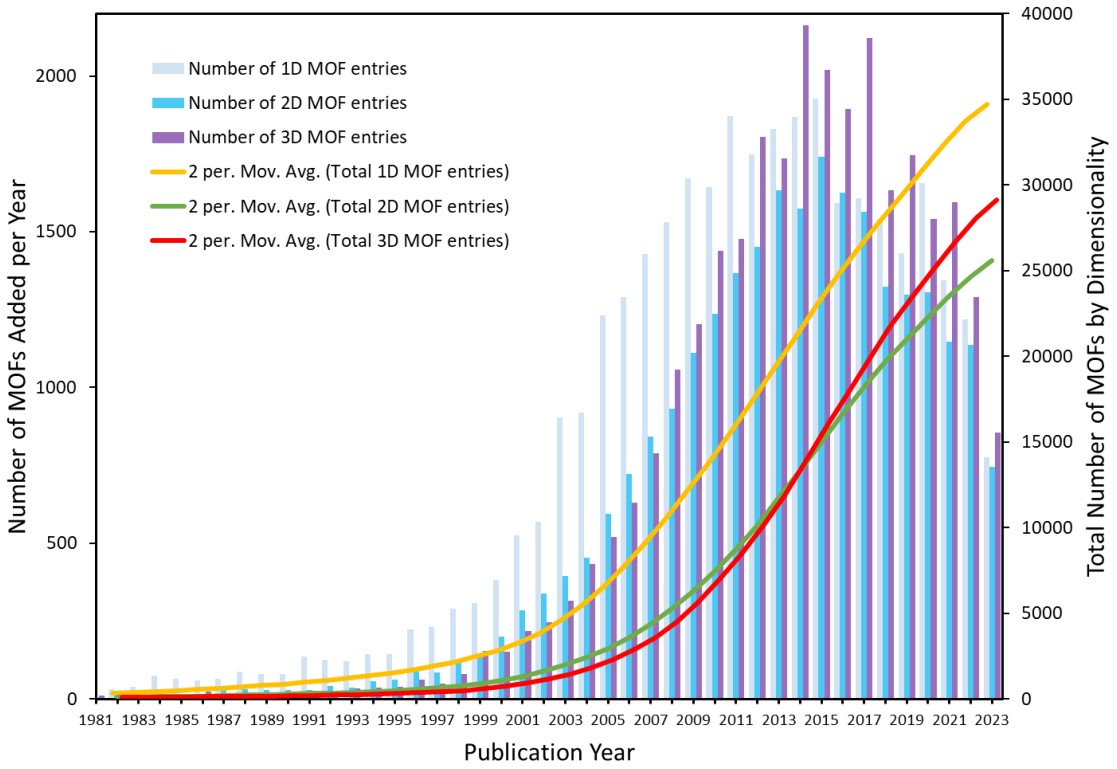


Figure 2.1: *The distribution of MOFs within the CSD, including dimensionality breakdowns of 1D, 2D, and 3D structures. The left axis indicates the number of structures deposited per year per dimensionality, whilst the right axis keeps a cumulative total across the timeline. (Data correct to CSD 5.45 Nov 2023).*

Following the International Union for Pure and Applied Chemistry (IUPAC) recommendations, published in 2013, suggesting that all MOF structures are assigned topological representations, a significant number of these materials should now be published and deposited with accurate topological information [23]. Ohrstrom et al. [24] released an informative review in 2015 following the publication of these IUPAC recommendations, where they offered guidance to researchers working in the field of MOFs surrounding identification of nets and network topologies. At present, the CSD does not report network topologies of its deposited structures, although for many materials submitted since 2013, this information may be available within the corresponding manuscripts as evidenced by our previous study which included the text-mining of MOF topologies [25]. The suggested procedure for reporting MOF network topologies is using a unique three letter code taken from the Reticular Chemistry Structure Resource (RCSR), printed in bold lowercase letters [2]. The RCSR is an open source, online database consisting of 2,929 3-periodic, and 200 2-periodic network representations. It is self-described as a collection of spatial information, and corresponding diagrams, which can be used to map networks that are built using straight, non-intersecting linkers.

Additional alternative databases for topological descriptions do exist, these primarily include the Topological Types Database (TTD) [26] and Euclidean Patterns In Non-Euclidean Tilings (EPINET) [27] theoretical database. Whilst there is often some overlap between these collections, it is very common to see newly reported structures represented

in literature by their corresponding RCSR identifiers. Where the RCSR representation is not present and if the topology has been determined by the authors, the alternative EPINET or TTD terminology may be seen. Typically, topological identification software packages refer to the RCSR labels with a preference over other representations wherever it is possible to do so, although RCSR and EPINET topologies are sometimes reported together. It is worth noting that RCSR topologies appear in the EPINET database with a different unique reference, for example the RCSR **pcu** is also represented by the EPINET s-net name sqc1, and likewise bcu can be reported as sqc3.

As the CSD does not contain topological information, and there is at present no publicly available complete MOF topology database, to obtain the topology for a given MOF structure one would need to search for the corresponding topology in the respective publication, or if this was not available, determine the topology for the structure by using one of the existing software packages. This article discusses the use of three readily available MOF topology identification programmes: Topos Pro [26], hosted by Blatov and colleagues from the Samara Topological Data Centre, MOFid [28] published by the Snurr Group at Northwestern University, and finally CrystalNets [29] a Julia based software from Chimie ParisTech published by the Coudert lab. Each of these approaches differ, sometimes subtly, in the structure connectivity, deconstruction, and identification stages. We also explain the important challenge of bond assignment and different approaches to topological identification, and compare different software features that are currently available. We also discuss the techniques used to obtain deconstructed or underlying nets, and current examples of datasets created using these packages.

2.4 What is topology?

A long-recognised feature of crystal chemistry is that the connectivity between atoms can be represented as a simple periodic graph with linkers being considered as edges, and metals or metallic clusters being treated as nodes, famously summarised by A.F Wells in his 1977 book on Three-dimensional Nets and Polyhedra [30]. Topological analysis provides deeper understanding of the synthesised materials and their properties, enabling comparisons of new materials with existing literature, and effectively communicating the networks of new materials. Topology holds significance beyond the simplest natural structures such as diamonds, zeolites, and quartz to describe and understand the variety of crystalline materials. Even in these simple one atom type configurations, the structural connectivity at atomic scale can affect the properties of the macrostructure. If we consider only carbon, whilst diamond, with its instantly recognisable cubic lattice construction registers at the peak of the hardness scale, lonsdaleite is built using a hexagonal lattice configuration and is potentially up to 58% harder than its cubic counterpart when measured across the <100> face [31].

In 2019, Moghadam et al. [19] reported the correlation between structure-mechanical stability and topology for 3,385 MOFs and 41 distinct topologies. In this context, they identified the top robust network topologies and emphasised the importance of building blocks, coordination numbers, and linker lengths. Later, in 2022, Li et al. [32] experimented with different synthesis conditions and concluded that it is possible to control the formation of specific topologies for a set of identical building blocks which can be useful to consider if a certain pore shape, size, or stability is desirable. The formation of distinct MOF nets from the same building blocks is an important insight to consider as it demonstrates the remarkable structural diversity and flexibility of MOFs and underlines

the importance of the principles of MOF formation.

In 2018, Bonneau et al. [33] published terminology guidelines to aid in the deconstruction of crystalline networks into their underlying nets. Their estimation suggested that 40,000 MOFs would be synthesised and published by 2025, a result that seems almost achievable given the 28,729 3D MOFs offered in the CSD release of April 2023, or one that already has been achieved if we include 2D MOFs within the prediction. One important focus of these guidelines was to address the ambiguity of node assignment. The method through which the nodes are chosen can have a significant impact on the outcome of topological assignment, depending on the constituent building blocks. If, for example, large linkers with porphyrin rings are present, the style of deconstruction approach can offer different outcomes to the most basic structure form. The general goal is to represent the connectivity of a structure using an underlying net which is mathematically defined as a simple periodic graph, consisting of vertices and edges. A simple graph is made suitable for modelling topological representations of MOFs by four important criteria:

1. Edges are non-directional, only a Boolean result when questioning connectivity between two nodes is required.

2. Nodes cannot exist which have only 1-connection, they must be considered 'loose ends' and removed. Elements such as hydrogen cannot become nodes.

3. A node cannot be connected to itself, there are no loops, and although this is not expected when approaching MOFs, it must be considered.

4. Each node connects only once to another node, additional connections between two of the same nodes are discarded. In some instances, where for example a MOF has a double linker between two nodes[34], these must be simplified into a single edge.

A net must be connected, periodic, and simple; this is the minimum information required to construct a good topological representation. Topology can be represented for any periodic crystal structure in both 2D and 3D planes, and for both cases the same rules apply. Structures that are 3D but only 'grow' into two planes (2-periodic) are known as disjoint, and do not have a true topological representation when considering RCSR criteria, although some representations for these types of crystal can be found in the TTD. Figure 2.2a. demonstrates the 3-periodic **bcu** topology CSD OFAWAV (DUT-53(Hf)) structure expanding polymerically from its 8-connected SBU in all 3 planes of space, yet Figure 2.2b. shows the existence of 'stunted' nodes on CSD OFAWID (DUT-84(Zr)), a derivative of the **bcu** based structure, where we see expansion in only two of the possible three planes originating from the now 6-connected SBU [35]. Here, two atomic scale sheets have been layered and are bonded by a linker, but in this case, there is no potential for expansion via further bonded sheets in the **c** plane for this structure, and therefore any subsequent layers would be treated as separate structures, like stacking sheets of corrugated cardboard. For this structure, the disjoint configuration is due to the deliberate replacement of linker molecules on the 8-connected SBU metal clusters with acetic acid molecules, resulting in a 6-connected SBU leading to a restricted 2D structure consisting of double layers. Interestingly, the pore limiting diameter (PLD), and the maximum pore diameter are not drastically changed between each configuration, and when shifting from **bcu** to the disjoint structure we see them reducing from 8.5 Å to 7.6 Å, and 11.2 Å to 11.1 Å, respectively [35]. As a result, we might expect to find several deliberately disjointed structures within the CSD's 2D MOF subset that demonstrate a comparable level of porosity to 3D structures.

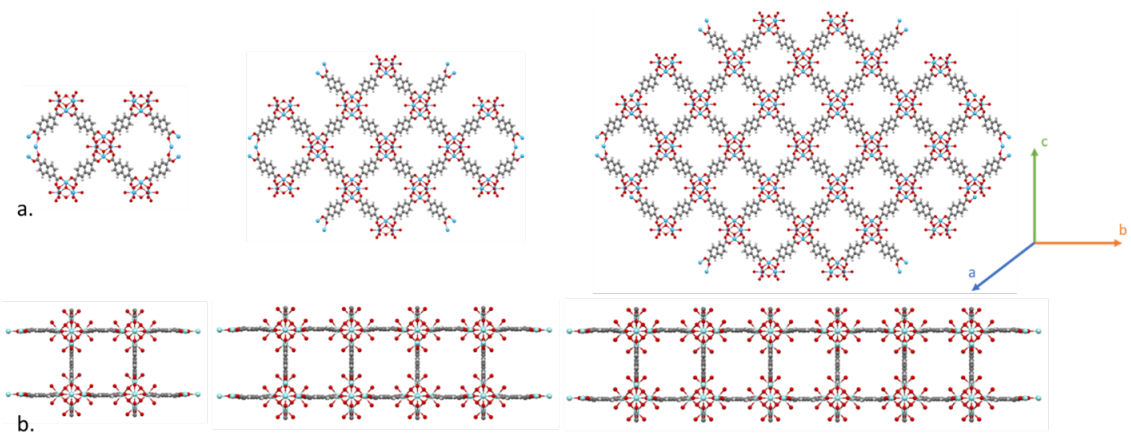


Figure 2.2: *An example of two similarly connected crystal structures expanded 1×, 2×, and 3× from their unit cells where a. CSD OFAWAV (DUT-53(Hf)) consists of 8-connected SBUs, and b. CSD OFAWID (DUT-84(Zr)) consists of 6-connected SBUs, visualised using CCDC's Mercury [35, 36]. The latter entry is considered disjoint due to the lack of polymeric expansion sites parallel to the c-axis; however, it expands polymerically in both other dimensions. Hf (bright blue), Zr (cyan), O (red), H (white), and C (grey).*

Clearly, there is a requirement for a rigorous and well-defined way to describe the symmetry demonstrated in MOFs, which could be extended to other crystal structures that consist of repeating units. This is generally accepted to be best represented by repeating the structure according to one of the 230 space groups found in the International Table for Crystallography Volume A [37]. After the space group of a structure has been determined, it is typically followed by the allocation of coordinates for each unique metal node in a unit cell, designed to create an infinitely expandable 2D or 3D network representation of a structure where there is little room for ambiguity.

The next, and truly key, step in the topology identification process is defining the positions of atoms that make up the nodes and linkers of the structure. Once coordinates are assigned to a vertex it is then designated as a node and the same applies to edges and their distinction as linkers. Although coordinates may be assigned by a variety of methods, the topology can be identical for structures that have different geometry. The creation of several nets may lead to a group of isomorphic representations, although it is often recommended that the network with the highest symmetry should (in these cases) be chosen as the universal net. This is somewhat subjective as it is often the whim of the crystallographer that decides the outcome as there are currently no set rules or absolutes for topological assignment, and it appears likely that will remain the case for the foreseeable future. There are several valuable discussions available for further reading that focus on the assignment of topology based on metal-organic polyhedra, such as the contributions from Goesten et al. in 2013 [38] followed by Kim et al. in 2015 [39].

Additionally, our discussion here must mention the existence of interpenetrating structures in which the empty space between nodes may accommodate one or more additional networks. Whilst the description and relationship between two 3D nets is quite straightforward, the complexity of possible relations between 2D sheets, or 1D chains, is significantly increased [40, 41, 42]. Interpenetrating MOFs, often referred to as IMOFs, can display some fascinating topologies and architectures and they often exhibit improved functions for certain applications. The existence of homo- and hetero- IMOFs can make for interest-

ing discussion surrounding the topology of these structures and the representations that are allocated to them, particularly those created using two or more underlying structures that results in a change of dimensionality for the macroscale material. Typically, each separate structure is considered during topological assignment rather than considering the interpenetrating nets as a single material, IMOFs do not contain bonds between the nets that are interpenetrated as they typically form independent structures inside the pores of each other. As an example, some MOFs can consist of many layers of the same 2D sheets interpenetrated throughout the entire structure to give an infinite number of 2D sheets where only one topological assignment needs to be made. An identical procedure is followed where these simplified nets are then matched to pre-existing representations found within the RCSR. We note that the interpretability of topology can also create barriers towards having exact solutions for each structure where additional representations are arguably equally suitable for an underlying representation.

2.5 Popular Topologies and Resources

2.5.1 The Reticular Chemistry Structure Resource (RSCR)

The RCSR was developed as a database to aid in both the design of new structures and the analysis of existing structures [2]. The latter being particularly useful as a considerable number of materials in the CSD were deposited before the popularity of MOFs began to boom, and in fact before the distinction of these structures was made in the early 2000s.

The RCSR consists of four sections, 0-, 1-, 2-, and 3-periodic nets. These are also split into two subsections of default or woven nets. Woven nets contain tangled polyhedra, chains, interlocked components, weaving and interpenetrating nets, and multi-component structures. For the default setting, the 0- periodic set contains structures consisting of convex polyhedra, including cages with 2-coordinated vertices. The 1-periodic list consists of cylindrical tilings and unsurprisingly, the 2-periodic set consists of plane tilings. Finally, the bulk of the RCSR, and the most interesting collection for those with an interest in gas adsorption, separation, and other porous applications of MOFs, is the 3-periodic set containing embeddings of periodic graphs. These structure definitions have been collected over a period from 2003 to present day in a series of important works [43, 44, 45, 46, 47, 48, 49, 50, 51].

In the RCSR, each topology is given a unique 3-letter identifier, typically reported in bold. These are sometimes presented with a simple suffix providing additional information. Each entry contains information regarding the vertices and their symmetry, coordinates, coordination, and order, with the same provided for edges, besides coordination. It is this data which is necessary to match these representations to simplified MOF structures, and these representations that are often reported in catalogues of MOF data. Figure 2.3 shows a collection of 10 of the most commonly occurring 3-periodic RCSR nets found in the CSD 3D MOF subset [25].

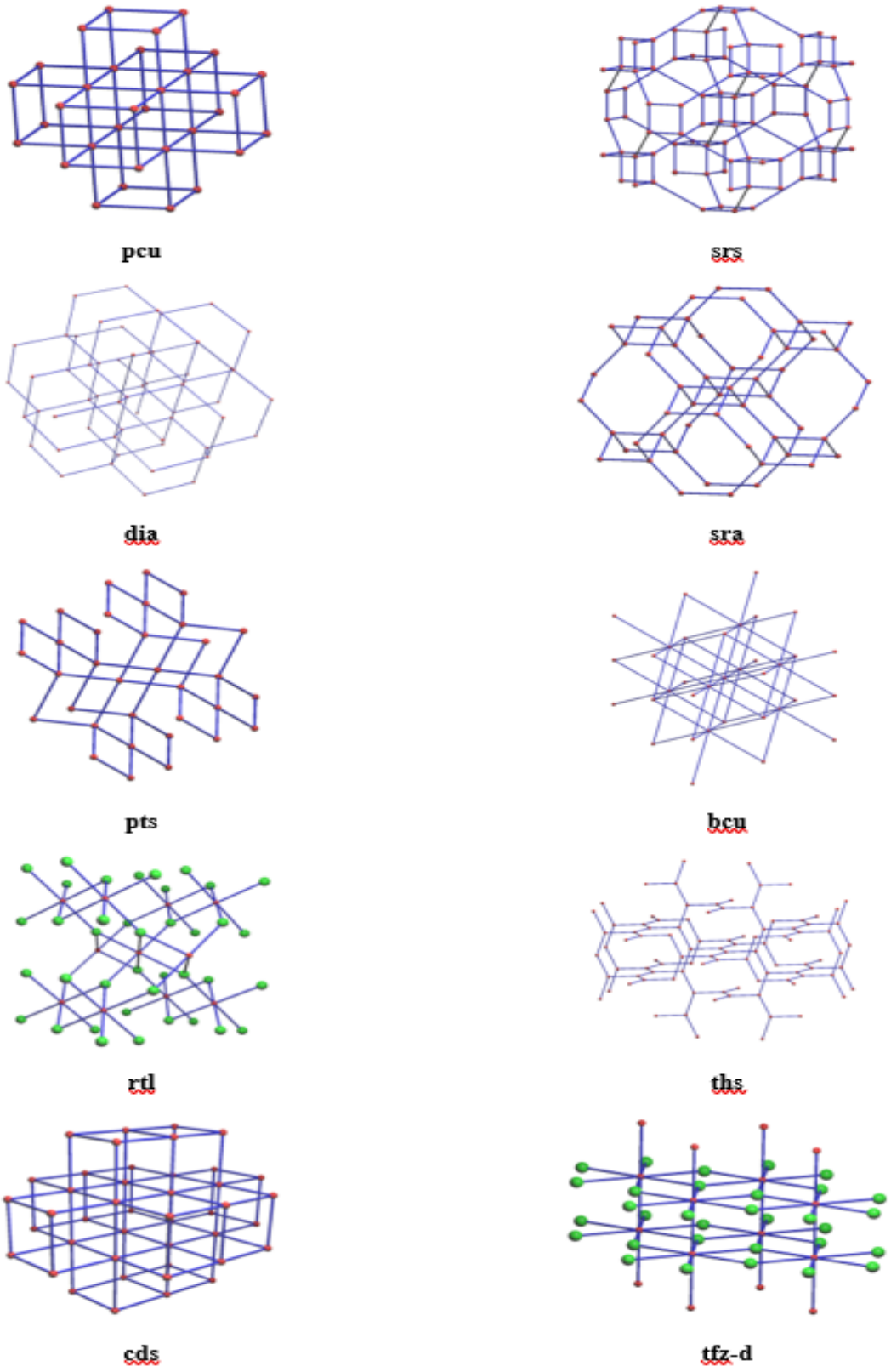


Figure 2.3: *Example RCSR topological nets created and visualised using ToposPro [26]. Red atoms represent metal nodes, whereas green atoms represent organic nodes.*

In this section, it is also worth mentioning the existence of the TTD, and the EPINET resource as other examples of topological collections that are notably relevant to the underlying connectivity of MOF structures. However, due to the limited availability of the TTD database without a licence, we focus our discussion on the RCSR collection. This is solely to ensure that fair comparison can be made between the topology assignment software packages detailed in Section 2.8 For completeness, all the structure representations in Figure 2.3 can also be found in the EPINET collection by searching the related RCSR names to find the corresponding sqc'xxx' style reference codes.

Here, we point out the existence of zeolite framework type descriptors that are also represented by 3 letter reference codes [52]. These are an older resource than MOF topologies with rules on nomenclature dating back to 1979 [53], and are unrelated to the RCSR. The 3-letter codes are typically derived from the material or institution origins, for example faujasite becomes FAU, and a complete list can be viewed here `https://europe.iza-structure.org/IZA-SC/Zeolite_names.html`. However, this is not to say that RCSR topologies could not be assigned to zeolites, and the use of capitalisation should set them clearly apart from lowercase RCSR references.

2.5.2 Edge-Transitive Nets

Whilst edge-transitive nets are often reported for many MOF structures, they may not necessarily be considered as the underlying topology of a structure. Edge-transitive nets are typically used to describe the structural symmetry, as opposed to the connectivity of the nodes and linkers. By selecting any edge in an edge-transitive net it is possible to rotate or reflect the structure around that edge and observe the arrangement of linkers and nodes remains unchanged. The nets represent a particular structure symmetry and can be used to design and synthesise MOFs with specific properties.

On the contrary, underlying nets are not restricted by the specific arrangement of linkers and represent only the spatial arrangements of nodes and connections. Edge-transitive nets are typically derived from the underlying nets, for example the underlying basic **nts** net can be obtained from simplifying further a derived net **ntt** structure. The derivations often consist of assigning geometric polyhedra to the nodes, and across some linkers, to have further influence on the exact shapes that can be obtained from a certain net. Chen et al. [54, 55] have worked on reviewing minimal edge-transitive nets specifically for the design and development of MOFs, and Hoffmann's Introduction to Crystallography [56] discusses details surrounding the basic and derived nets found in the RCSR, supplemented by an online resource [57]. A recent contribution from Delgado-Friedrichs et al. [58] discusses some new results and contains a concise review on 3D tilings and surfaces.

2.6 Deconstruction Techniques

Embedded within the topological identification software packages are several algorithms that are typically applied to a basic (i.e. containing no additional information such as atomic bonding) CIF to determine the simple underlying connectivity of the structure provided. Each algorithm takes a slightly different approach to simplification, and as metal nodes can be assigned subjectively, it is important to understand the differences between the techniques and how they operate. All methods first define which groups of atoms should be considered as nodes, and subsequently which connecting branches become the linkers. It is worth noting that some linkers may contain metals which are

not necessarily assigned as nodes, for example in a metallic porphyrin ring (CSD BEDYEQ [59]), and conversely a linker may contain an organic ring which is best represented by a node, albeit an organic one (CSD JOZWIG [60]). It must also be considered that, for a topological representation, there is no difference between the types of nodes which exist in a simple periodic graph as there is no absolute distinction between metals and organics in these underlying representations.

The typical algorithms employed in MOF deconstruction include, all node [61, 62, 63], single node, standard representation [26], and metal-oxo [28]. An additional cluster representation method is a partial but chemically reasonable deconstruction technique that requires the division of all bonds into inter-cluster and intra-cluster criteria. In what follows, we outline the steps performed by each of these algorithms and include schematic diagrams to aid understanding via visual representation of these stages.

2.6.1 All Node and Single Node Deconstruction

The most recent publication describing the all node algorithm was from the work of Li et al. in 2014 [61]. However, earlier examples have been published as far back as 2006 [62, 63]. This algorithm works by considering inorganic nodes and organic linkers as abstract shapes (polygons and polyhedra) connected in a simplified net. Connected carboxylates and heteroaromatic rings are considered to constitute part of the node. After the nodes and linkers have been assigned, these clusters are simplified via replacement with pseudo-atoms at geometric centres. Any isolated pseudo-atoms are considered free solvents and are removed from this simplified net. Figure 2.4a demonstrates the steps undertaken to assign an all node net for an atomic level crystal structure. Here, the metal clusters, formed of polygons, are treated as a single polyhedron and simplified to a single inorganic node. Similarly, the porphyrin ring is considered also to have been built with polygons, which are used to create a single polyhedron with four pseudo-atom connecting points on the vertices.

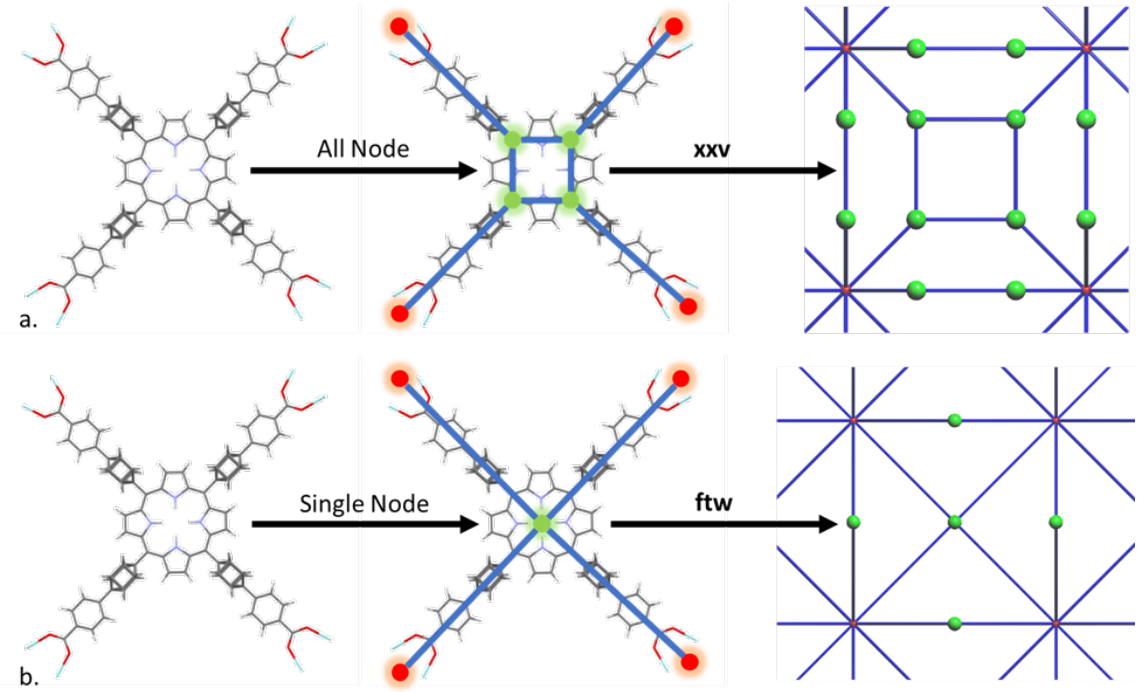


Figure 2.4: *Schematic demonstrating crystal deconstruction techniques applied to CSD JOZWIG [59]. The distinct path taken by each algorithm for large heteroaromatic rings results in* ***a.*** *the all node approach matching the* ***xxv*** *topology, and* ***b.*** *the single node approach matching with the* ***ftw*** *topology. Wireframe structures show C(grey), O (red), N (blue), Zr (light blue), which are simplified to metal nodes (red), and organic nodes (green) connected by straight edges representative of linkers (blue).*

This approach specifically identifies branching points within the linkers of a MOF to provide additional information about the underlying structure, but this allows for the creation of ambiguous branching nodes. Typically, the all node algorithm creates a more complex structure which can be matched to non-parent nets in the RCSR. For example, for the structure shown in Figure 2.5, the **xxv** net can be considered a derivative of the **ftw** net. O'Keefe et al. [64] explains there are many situations in which retained information takes precedence over reporting only the most simplified parent net. Using these non-parent nets can often be useful for comparing similar structures because of the retention of this important higher-level connectivity information and it makes the discovery of closely geometrically related structures much easier.

The single node approach is very similar to that of the all node approach, however pseudo-atoms with only one neighbour are dealt with based on their identity. Either metal containing linker molecules show up as pseudo-atoms with non-redundant connections to a linker and therefore are merged, or linkers with a single connection, except for single non-oxygen atoms such as halogens, are removed as unnecessary bound solvent molecules. This approach is demonstrated in Figure 2.4b where the difference between the all node algorithm above can be noted for the simplification of the large aromatic ring structure. Here, the metal clusters are treated the same way as above, but the porphyrin ring is instead considered to be a single point, rather than a polyhedron with separate vertices and edges.

The single node approach is often considered the preferred technique to determine the most basic nets in MOF chemistry as it typically reports the parent net of structures that may also have alternative complex representations. It is anticipated that most reported topologies are obtained using the single node approach, and this allows for easier categorisation of structures into broader topology groups. The allocation of **xxv** and **ftw** topologies to this same structure can both be considered correct; we must remember that one is only a more complex net that has been derived from the other. As the simplifications to the structure are only being conducted differently due to the choice of algorithm used, either representation is permitted.

Overall, the single node method describes the most basic form, whereas the all node algorithm retains complexity. It is essentially down to the preference of the researcher to determine which outcome they consider more favourable, although it is worth noting that for many materials both algorithms will report the same result as they have only one valid representation. The IUPAC recommends that researchers should report multiple topologies if appropriate, in this case when reporting the all node result, we would expect to see a statement like "the **ftw**-derived net **xxv**" which should be stated alongside the **ftw** single node outcome [23].

2.6.2 Alternative Deconstruction Methods

Standard Representation (Standard Simplification)

This is perhaps the simplest of all the algorithms mentioned in this list, it is concerned with disconnecting any bonds to metal atoms and leaving the remaining molecular graph intact [65]. Metal atoms and organic ligands are the only structural units, and all atoms of each ligand are substituted by a pseudo-atom. More generally, anything classed as non-metal will be contracted to a single atom at the centre of mass including but not limited to single non-metal atoms such as oxygen, halogens, or multi-atomic non-coordinated species.

For the case demonstrated in Figure 2.5a where this simple technique is applied to MOF-5 (SAHYIK) from the CSD 3D MOF subset, we can see that a significant number of bonds are retained. This method is shown in parallel to the previously described all or single node approaches shown in Figure 2.5b, proving a distinct difference in outcome. Where standard representation here assigns a more complex **fff** topology consisting of significantly more pseudo-atoms, the all or single node approach selects only the metal nodes in a more extreme simplification represented by the **pcu** topology that could be considered a loss of key information.

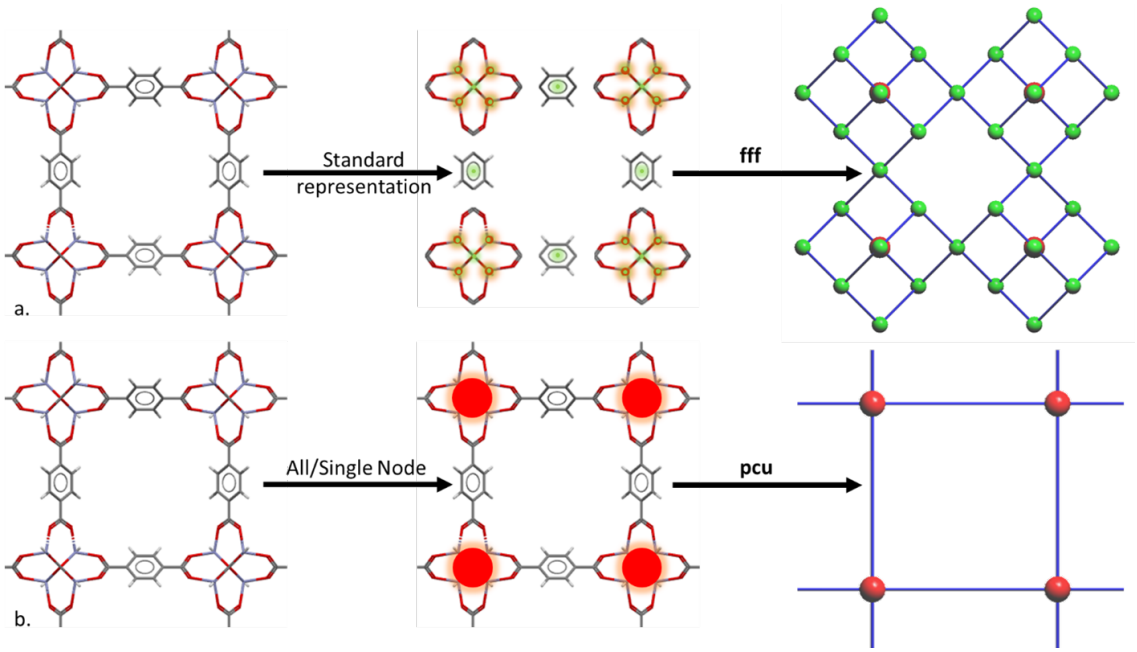


Figure 2.5: *Schematic demonstrating crystal deconstruction techniques applied to CSD SAHYIK [66]. The approach* ***a.****, standard simplification, with initial disconnection between metal atoms and the organic structural units, results in a match with* ***fff*** *topology and* ***b.*** *all/single node matches with* ***pcu****. Wireframe structures show C (grey), O (red), Zn (blue), which are simplified to metal nodes (red), and organic nodes (green) connected by straight edges representative of linkers (blue).*

In addition to this approach, there is a second method detailed by Barthel et al. [65] called cluster simplification which recognises clusters of atoms with high connectivity. This technique draws many similarities to the all node and single node algorithms, and has been used to determine if two separately deposited structures are the same. For example, rotating a linker of a specific MOF may not change the material, but it could have an impact on the space group which in some circumstances would allow the structure to be redeposited into the same database. In this technique, the smallest ring of bonds is found for each bond. Next, the ring sizes, a, are sorted by increasing value from a_1 to a_N, where N is the number of bonds in the structure, in the sequence $a_1 \leq a_2 \leq \ldots a_N$. If the sequence contains a pair a_j, a_{j+1} such that $a_j - a_{j+1} > 2$, the bonds where the smallest rings are formed by less than $i+1$ bonds belong to a cluster, and the others connect two clusters together. Each cluster is substituted by a pseudo-atom to obtain i and the bonds are preserved between clusters.

Metal-Oxo

The metal-oxo algorithm is a more recently developed technique, created by the Snurr Group to describe MOF chemistry by dividing structures into distinct organic and inorganic building blocks - retaining organic linkers as discrete building blocks (including carboxylate groups) [28]. Compared to the more topologically inclined single and all node algorithms, the metal-oxo approach is a more chemistry focused approach to describe the targeted structure, although it draws some comparisons with the single node approach. The result is achieved by keeping organic linkers intact and therefore it provides alternative information to the other methods. MOF structures are divided into distinct inorganic

and organic building blocks via a bond adjacency matrix using a distance cut-off method that adopts the InChI convention of classifying metals and non-metals. Typically, the inorganic blocks consist of metal-oxo clusters including oxides and bound hydroxide, peroxide and water species with the remaining fragments considered organic building blocks and described as larger non-metal clusters. These building blocks, represented as SBUs, are characterised by their points of extension, through which they connect to other building blocks in the underlying net [67]. This distinction between the metal-oxo algorithm, and the single and all node algorithms which consider carboxylates part of the node, can be an important distinction in cases where, for example, five discrete metal atoms are instead represented by a pentametallic SBU [68]. The metal-oxo approach is shown as a schematic in Figure 2.6, where it is used to simplify the structure into a complex, metal independent form.

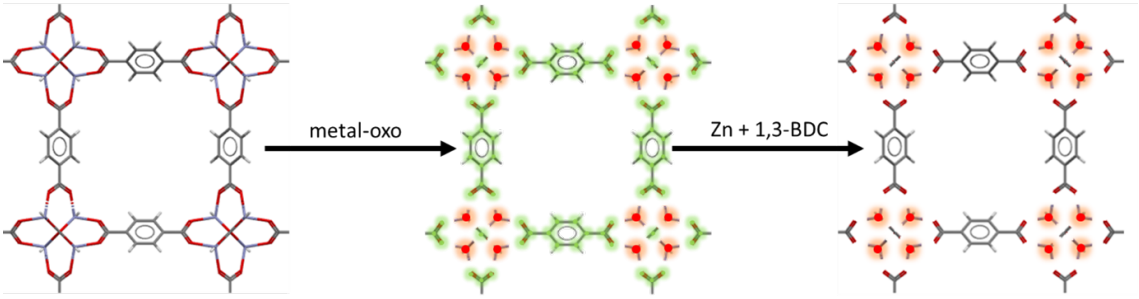


Figure 2.6: *A metal-oxo deconstruction, shown as a schematic diagram, performed on CSD SAHYIK [66].In the original structure (left), C (grey), O (red), and Zn (violet). This technique draws many similarities to the single and all node approaches, but with a focus on structure chemistry showing the resultant (middle) Zn metals (red) and 1,3-benzenedicarboxylate linkers (green).*

Whilst the metal-oxo method is not typically employed to determine the topology of a structure, due to being primarily developed to offer insight into the constituent metals and linkers of a crystal structure, it is both important and interesting nonetheless to consider alternative approaches to structure simplification.

2.7 MOF Databases and Design Principles

Over the past decade, significant research has been conducted via large-scale high throughput computational screening of structures from various databases containing key information regarding thousands of lab synthesised MOFs or hypothetical materials. Continuous improvement in MOF synthesis practices have led to a greater ability to control key properties of newly created structures, including topology. Over the past 10 years, several databases containing hypothetical and experimental structures have emerged. Figure 2.7 shows a timeline noting the release date of a handful of key MOF datasets.

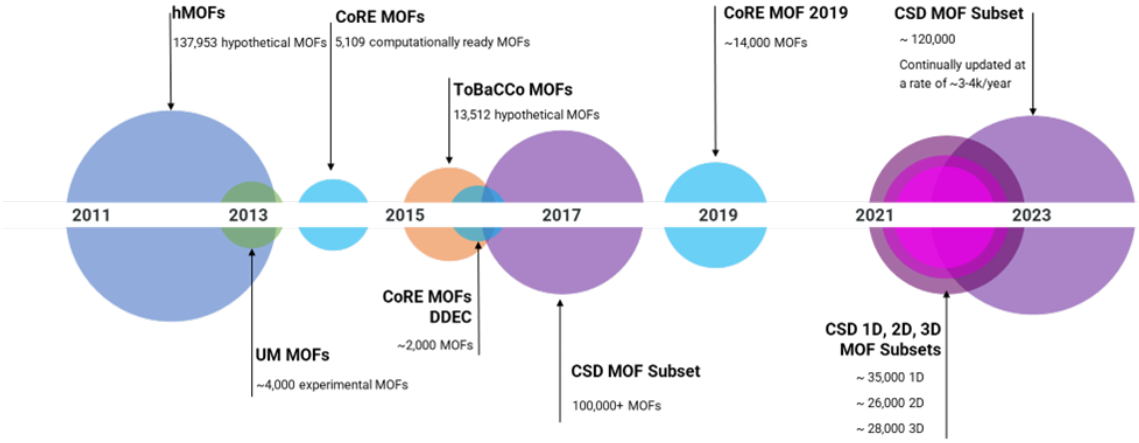


Figure 2.7: *A timeline to show the emergence of selected experimental MOF datasets following the release of the first hypothetical MOF database (hMOF) [69] in 2012. Circle size varies to represent the relative size of the database, colour is representative of the study/research that produced the resource.*

The first categorisation of large sets of experimental MOF structures began with the creation of the UM MOF database in 2013 [70], this study was focused on the identification of porous MOFs from the CSD, selected to calculate theoretical limits of H2 storage, a study that was completed for 4000 MOF compounds out of around 22,000 'computationally ready' candidates. This was closely followed by the development of the Computationally Ready Experimental (CoRE) MOF database in 2014, as part of the Materials Genome Initiative [71]. Consisting of modified CSD entries, it had been specifically created for use in molecular simulations. Only 3D structures with pore sizes exceeding 2.4 Å were considered, and over 4,700 porous materials were collected in a computationally ready database. Later, in 2019, the CoRE MOF database saw the completion of an update, increasing the total of porous 3D MOF structures, reported in published literature sources, to 14,000. This new update also added further value to the data set by offering new pore analytics and physical property data alongside the correction and reconstruction of many disordered structures [72].

In 2017, Moghadam et al. [19] developed the CSD MOF subset, a searchable database of MOFs that is continually and automatically updated, with additions to the collection every quarter, as new materials are deposited and accepted as part of the CSD. This work created the largest collection of experimentally synthesised MOF-like structures to date (now numbering ca. 120,000 as of April 2023) but was done so using loose definitions to avoid omitting potentially useful or interesting structures, and to allow for an all-encompassing data set that can be further scrutinised by the user depending on their interests. Containing an initial ca. 70,000 1D, 2D, and 3D structures combined, the size of the CSD MOF subset has almost doubled in just seven years. Additional developments to the CSD MOF subset reported in 2020, resulted in the creation of 1D, 2D, and 3D MOF subsets20. Whilst at present there is no option available when browsing CSD structures to easily identify a material's topology, we are developing methods to perform reliable high-throughput topological allocation on these new sub-categories of structures to be included within the CCDC's database. The CSD 3D MOF subset is an ideal candidate for development into a resource where the inclusion of topological characterisations would become most readily available.

Further to this, we note that the distinction between a set containing all structures and those without disorder is significant in this field where the exact connectivity of atoms is of upmost importance for producing reliable high-throughput topological analysis. It is imperative then, that the first step towards topological identification of any structure found in the CSD MOF subset using these approaches is to determine whether the structure is crystalline, and what level of periodicity it demonstrates. 1D structures, known in the CSD as 1D chains, are not expected to be assigned topology using the techniques outlined in this article. 2D structures, known as 2D sheets, are restricted in their allocation to a limited set of 200 configurations as specified in the RCSR, and due to the limited range and complexity, we expect a significant proportion of these should be identifiable, via the use of software. This distinction into periodic categories enables even the novice crystallographer to quickly determine, by knowing its dimensionality, as to whether an incorrect topological net has been allocated to their structure.

One reason for mismatched topological assignment between dimensionalities could occur due to incorrect bonding determination, for example a 3D structure may be assigned a 2D topology if atom connectivity between 2D layers had not been correctly interpreted – a possible outcome when using automatic bonding assignment software, and one that is particularly prevalent for structures that contain metal-metal bonds. Bond assignments are typically entered by the CSD editorial team with a view to represent the original experimental publication as closely as possible, this is to ensure that the process of assigning bonds is not done entirely on distance - particularly for bridging O or H.

Figure 2.8 compares the distribution of a variety of metallic (X) X-X bonds and non-bonded interactions within the CSD. In Figure 2.8a bonded (blue) and non-bonded interactions (orange) for Ag-Ag fall within a range primarily between 2.7 – 3.5 Å (represented within the dashed red box). Figures 2.8b—d show more examples of metals that either have potential atom-atom bond misalignments or metals where this may be of no concern. Hg-Hg shows a similar pattern to Ag where ambiguity may lie (also within the red dashed box) for structures such as the bonded CSD GIZPIP [73] for interactions between 3.5 – 4 Å. An apparent lack of data surrounding Cd-Cd bonds here suggests a lack of Cd-Cd based SBUs (highly likely given the bond order calculations for creating Cd-Cd bonds [74]) with only 6 non-disordered MOFs containing a Cd-Cd bond, and lastly the Sn-Sn data shows an example of clear delineation at approximately 3.75 Å between bonded and non-bonded contacts. In the dashed red regions, we expect to see examples of both bonded and non-bonded layers in 3D Ag and Hg containing structures, a highly important detail when we consider the use of auto-bonding software in the topological assignment process, but conversely structures such as Sn-Sn would be ideal candidates for investigation where the use of automatic-bond assignment software could be considered less troublesome. Subsequently, the 6 Cd-Cd bonded structures identified here were investigated and manually corrected as a result of this study. Whilst we have mentioned only a select few examples here, metallic bonding data is available for all structures in the CSD.

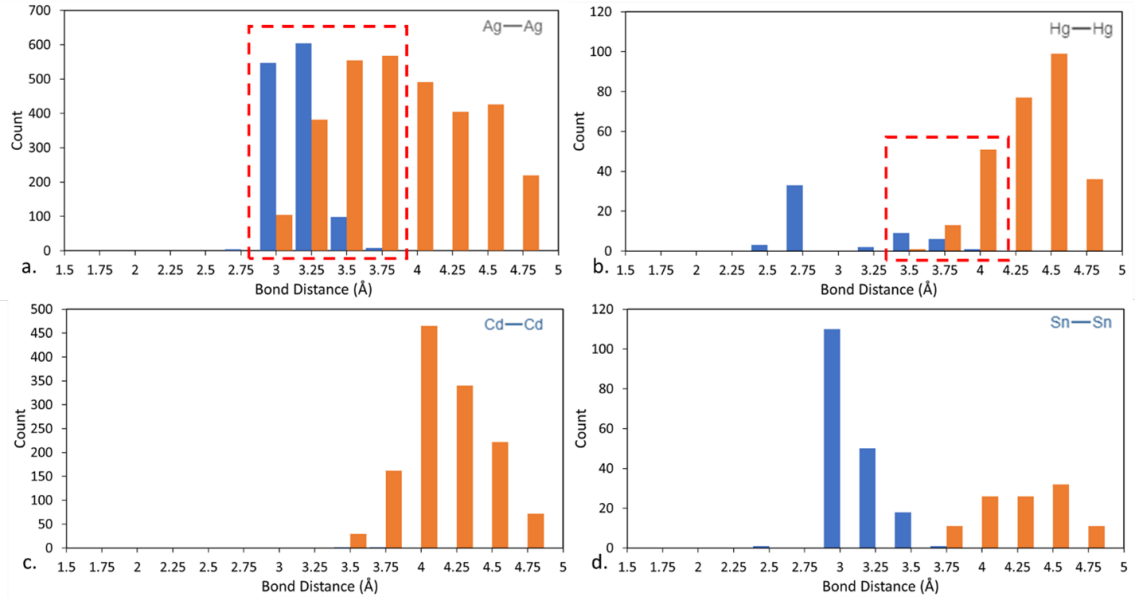


Figure 2.8: *Distribution of selected atom-atom bonded (blue) and non-bonded (orange) contacts (out to VdW+0.0) in the CSD. a. Ag, b. Hg, c. Cd, and d. Sn. Dashed red boxes suggest contentious atom-atom bonding ranges.*

To highlight the importance of bond assignment in determining structure dimensionality, let us investigate an example. Figure 2.9, CSD ZEHMOQ [75] is a 2D MOF containing some Ag-Ag bonding at 3.32 Å, however, extending the bonding limit just slightly to 3.35 Å (which could be considered a possible bonded or non-bonded distance) transforms the 2D sheets into a single 3D crystal structure. Therefore, taking atom connectivity data directly from the CSD before modification offers a more chemically aware insight into the structure of a crystal, as determined by the experimentalists themselves when depositing structure information, rather than risking miscalculation by automatic bond assignment software with algorithms deciding bonding based on atomic distance.

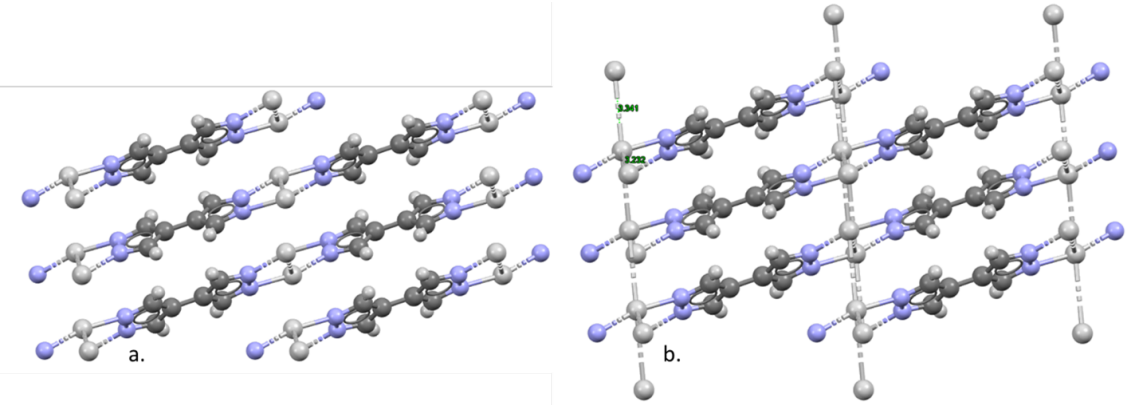


Figure 2.9: *Atomic level representations of CSD ZEHMOQ showing a. the original structure set at a 3.32 Å Ag-Ag bond distance limit and b. an auto-modified version with a 3.35 Å Ag-Ag bond distance limit, where the connectivity has been calculated using automatic bond assignment tools within CSD Mercury.*

We would recommend that when attempting to assign topology to a structure that

the original chemical bonding is considered (wherever possible), as opposed to removing/omitting the existing bonding data and attempting to reassign it using additional software such as OpenBabel [76]. Therefore, whilst most topological characterisation software is packaged with some form of bonding assignment tool to calculate atomic bonding for imported CIFs, we recommend inclusion of the CSD's atomic bonding data in all generated CIFs. Although this is available for structures obtained through the CSD's Python API, typical CIFs do not contain atom-atom bonding information. Further to this, even if the bonding data is present, it is not always possible to upload a CIF to these software packages and retain the relevant CSD bonding data as the only option available may be to re-calculate bond types and distances, and whilst these may be manually edited later, structures requiring manual bond modification may restrict the capability for high-throughput calculations.

2.8 Topological Characterisation Software

2.8.1 Introduction

At present, a handful of topological identification tools exist, aside from the painstakingly slow and perhaps unreliable method of performing manual structure-net matching. The most well established and frequently cited package is ToposPro [26]. The developers at Samara continue to maintain this software, have published many video guides for inexperienced users, and even offer a topological identification service for a fee. A more recent development, which has seen some updates this year for use in high-throughput topological assignment approaches is MOFid [28]. MOFid has been used as a topological identification software for the CoRE MOF database so that topology can be searched for within the data set, but its primary use is focused on obtaining unique identifiers for MOF linkers. Finally, and most recently published is the CrystalNets package [29], and although this software has been published and is available, not enough opportunity has been given since its release to judge the uptake of this approach within the community, aside from a small number of interesting citations. These software packages have all been built using different programming languages and offer the user multiple approaches to verify the output of their structure's topological identification.

2.8.2 ToposPro and TopCryst

ToposPro is a licenced downloadable programme that it is frequently maintained and updated, with the latest version 5.5.2.2 available at `https://topospro.com/`, that can be activated using a free licence provided for academic users. An entirely automated version can be implemented for single structure analysis without requiring any installation by uploading a CIF online at `https://www.topcryst.com`. The topology of a single structure can also be quickly obtained by searching the TTD database. An added, and useful, feature of this online tool allows a user to search for any 3-letter RCSR topological representation and view this in a JSmol window at various dimensions of unit cell, with several example structures from the CSD also shown in a table below the topological search. This is not a complete open-source online database of structures as the free version does not allow the user to download the CSD refcodes of any specific topology, but instead offers five random examples of structures which meet the criteria of the searched topology and notes the technique through which they were obtained. An example is shown in Figure 2.10 below using CSD SAHYIK, more commonly known as MOF-5, with the TopCryst online interface after uploading a CIF file where the unbound solvents have

been removed. Here we can see the allocation of three distinct RCSR topologies, **mof**, **fff**, and **pcu**, with a clear indication of the methods used to obtain each underlying net.

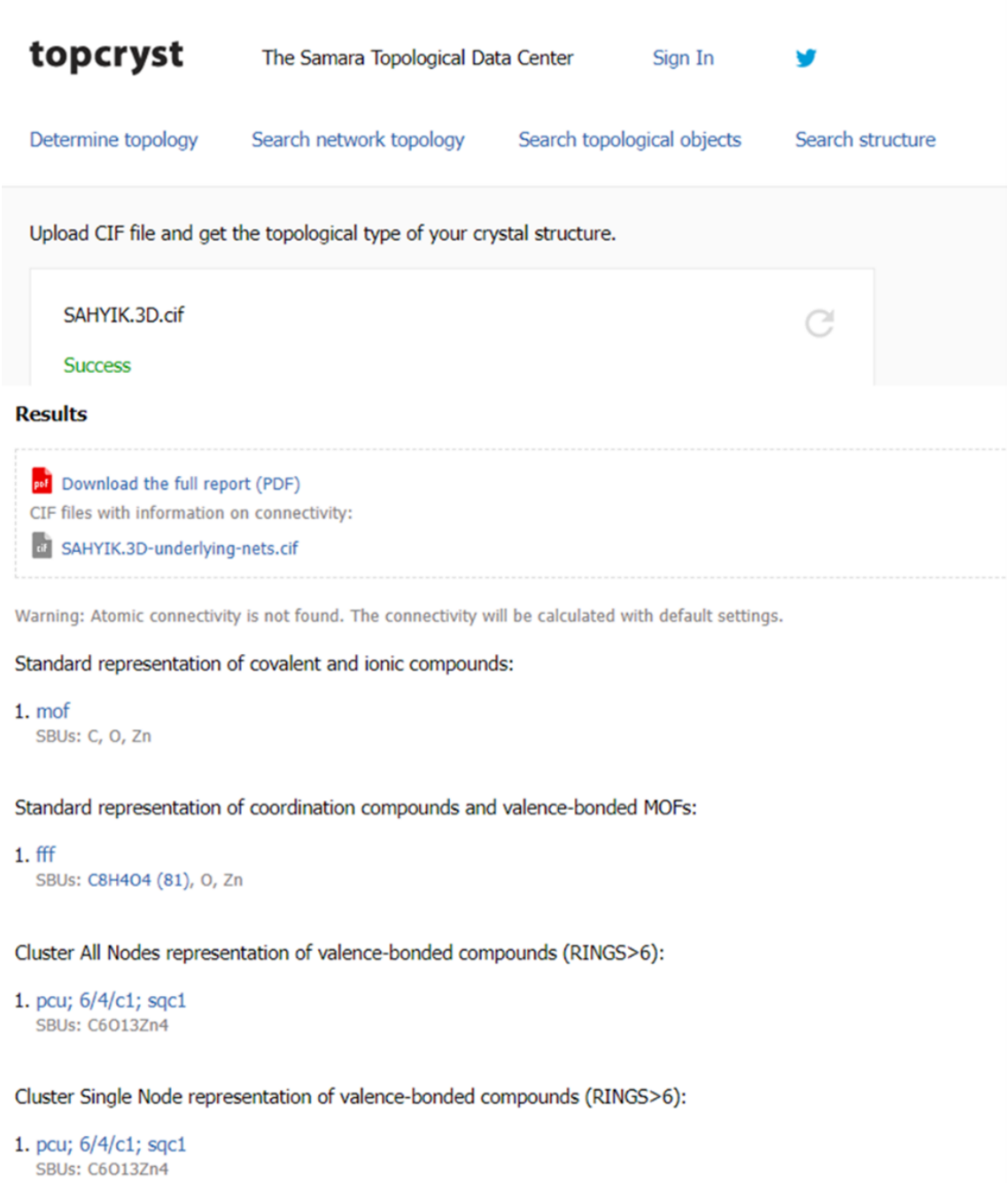


Figure 2.10: *A snapshot of the online interface of the TopCryst web topology service used for the automatic deconstruction of CSD SAHYIK. The original CIF was modified with the use of CSD's Python API solvent removal script.*

With regards to the software itself, ToposPro is a program package for comprehensive analysis of geometric and topological properties of periodic structures such as, but not limited to, MOFs. The techniques contained within can be applied to almost any structure of a chemical nature. It has been developed to process large crystallographic data

samples and correlate structure property parameters. The principles behind this software package aim to achieve a human independent crystallographic data processing tool which approaches materials that have a variety of complexity levels with universal algorithms in contrast to traditional crystallochemical visual analysis. The aim of separating structures using universal algorithms is an effort to avoid the difficult nature of topological assignment and offer consistent topological representation of structures by minimising any errors. This method is known as the Domains algorithm which uses atomic Voronoi polyhedra as geometrical parameters of atoms and bonds [26].

All methods contained within ToposPro can be divided into geometric or topological groups, respectively. The first group is concerned with routine geometric calculations and crystal structure visualisation, and the second contains the procedures required for studying connectivity of the whole crystal environment. A database is created upon the importation of a CIF, and bonding must be assigned to structures added to the database before topological assignment can occur. This is performed using the AutoCN programme, the details of which can be found in the ToposPro manual. It has been tested on thousands of structures from the CSD and has showed good agreement with chemical models [64, 77]. For structure deconstruction, the use of cluster representation is possible in three different ways, using the chemistry mapping single node, the geometry mapping all node, and the tertiary building unit (TBU) cluster mode. There is the additional possibility, which is applicable to all structures, called the ToposPro standard, or standard representation, mode. It should be noted that this is not always the most descriptive method, and typically more information can be obtained using other approaches. Additional features of ToposPro include the ability to detect duplication of structures, investigate entanglements and interpenetration, and the modification of structure bonding following the use of AutoCN. The software is noted for its high accuracy when implemented on suitable structures following the AutoCN stage.

The limitations of the software include the application of the program on large datasets, and whilst it is possible to run continuous calculations on tens of structures at once, the nature of the programme restricts the use of true high throughput operation. The ToposPro package is best suited to investigating individual structures on a case-by-case basis, and when using this approach it is a powerful tool for topological assignment, particularly when focused on rod-like MOFs as other packages struggle to handle these difficult to interpret materials.

2.8.3 MOFid and web-mofid

MOFid [28] is a freeware Github hosted identification software available at `https://github.com/snurr-group/mofid`. The primary MOFid package can be downloaded and installed using a make file directly into a virtual environment. Any CIF located in an accessible directory can be parsed using the cif2mofid function of the MOFid programme for topological analysis directly from the command window within a python environment. It is perhaps worth noting that the software package has had a larger focus on the identification of linkers than topology and is primarily designed to offer insights into MOF building blocks by assigning the linkers with unique identities to improve the cross referencing of linkers between MOF structures that share some of the same building blocks.

Similarly to TopCryst there is a single structure web based analysis feature into which CIF files can be uploaded for topological analysis as well as deconstruction into individual building blocks followed by the allocation of identifiers, this can be found at `https:`

//snurr-group.github.io/web-mofid/. Not only does MOFid return a topology parameter, but it also returns a MOFid or MOFkey string. A MOFid is based upon SMILES strings and takes the form of inorganic building block, organic building block, format, topology code, catenation, comment. A typical example of a MOFid for Cu-BTC would be: [Cu][Cu].[O-]C(=o)c1cc(cc(c1)C(=O)[O-])C(=O)[O-] MOFid.tbo.cat0;Cu-BTC. This can be pasted into any software package that recognises SMILES, such as ChemDraw, and it should render for visualisation. The alternative output is the MOFkey which takes a similar form as above except with the catenation and comments no longer present, and the organic building blocks now represented by a unique alphabetised code.. The same Cu-BTC structure as above has the MOFkey as follows: Cu.QMKYBPDZANOJGF.MOFkey-v1.tbo. Figure 2.11 shows the output of CSD SAHYIK uploaded in CIF format to the web interface of MOFid, displaying the options for algorithm visualisation in a drop-down box, and the corresponding MOFid text-string below.

A final web-based feature is the CoRE MOF database search tool [28] which allows a user to search over 15,000 MOFs by SMILES/SMARTS, topology, or catenation. A simple text-based search in this dataset for **pcu** reveals 749 MOFs and their SMILES string, catenation, and where applicable their CSD refcode. If a user's chosen refcode matches a structure in CoRE MOF, there is no requirement for the user to re-run any structures found in the database to obtain these parameters.

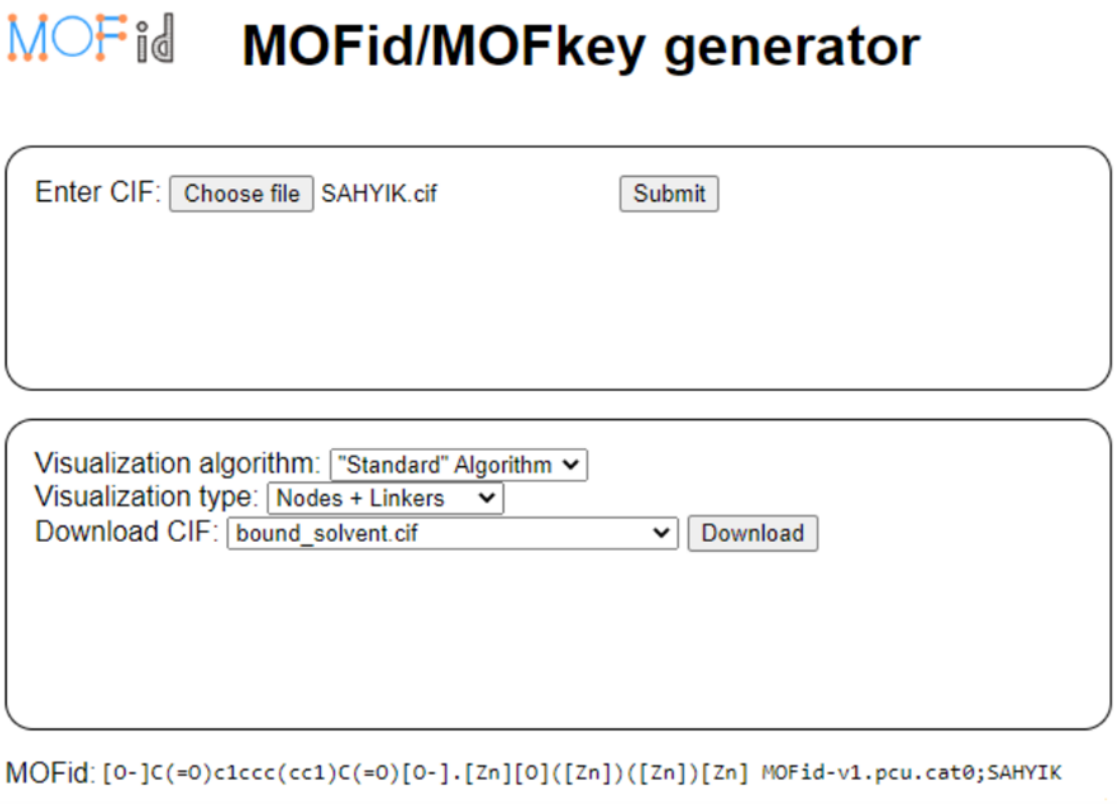


Figure 2.11: *A snapshot of the online interface of MOFid's web structure identification and topology tool performing a structure simplification on CSD SAHYIK by uploading the raw CIF.*

The MOFid Github package also contains shell scripts to run a directory of CIFs on a high-performance computing cluster, and it is possible to process a folder containing

thousands of MOFs, provided that the input files are suitable for the software. Bonding is assigned using the open source OpenBabel chemical toolbox that was designed for use with molecular modelling, chemistry, solid-state materials, or related applications [76]. OpenBabel can implement a wide range of cheminformatics algorithms including bond order perception, once the unit cell information is extracted from a CIF file.

Simplification is performed by the metal-oxo, single node, and all node algorithms with the output of each technique available to visualise via the dropdown box. This feature is particularly useful to compare the different methods, although the output string containing the topology reports only one underlying net even if several have been detected.

The simplified net is exported to Java based net matching programme Systre [51], where the RCSR nets are pre-loaded, and the new simplified net is matched to one of the existing configurations within this data set. The use of this programme within MOFid is key to the topological identification stage and the speed at which this matching is performed can be a limiting factor in the high-throughput use of this software when compared with CrystalNets which does not require the use of Systre.

It is possible when using the MOFid python package to modify the output desired by the user by editing a few simple lines of Python code. By performing this modification, a user can report topology based on whatever criteria they so choose, and for example might only be interested in structures where the topology obtained via the single and all node algorithms are the same. It would be equally as simple to report topology for only structures where the output between the two techniques is different, or for all three methods contained within this software.

2.8.4 CrystalNets.jl and CrystalNets

CrystalNets.jl [29] is an open-source Julia based software package hosted in Github, that can be obtained from `https://github.com/coudertlab/CrystalNets.jl`. The installation can be performed quickly and easily after opening Julia, by entering the package manager and adding CrystalNets, and the integration of the programme within a Python environment can be enabled with relative ease. It is possible to install the package as an executable for a handful of structures, but for high-throughput approaches the use of CrystalNets as a Julia module is recommended. This software is specifically designed for the automatic detection and identification of underlying topological nets of crystalline materials, and the input format can follow any file type that is recognised by chemfiles [78].

Upon installation, there are a variety of settings available to the user, the most basic of these includes the ability to select the deconstruction algorithm used, whether to use the bonds that are input in the file or to guess them, and the type of structure that is being investigated. In this package the standard, all node, and single node approaches are available, so for example, it is possible to select MOFs, deconstructed using the all node algorithm, with the guess function enabled for bonding if they were not included in the original input file, or auto if some files contain bonds and others do not. This feature is particularly useful for defining the topology of MOFs where there are bonding parameters contained within the input file should the CIF have been taken directly from the CSD with care taken to ensure that the bond lengths have been retained. There is also the availability of a MOF option which modifies the approach to enable the detection of organic and inorganic clusters, allowing them to be subdivided using either all node or

single node algorithms to identify the underlying nets. Other choices for this parameter also include Zeolite, Cluster, Auto, and Guess. The CrystalNets manual is a good accompanying resource that contains all the available options for each function and further explanations surrounding exactly what each of the changes to these input parameters makes to the process.

The use of a Julia module allows for some extremely fast structure deconstruction compared to the other methods available, and this is amplified by the availability of a multi-threaded implementation for a large set of structures. The CrystalNets programme is orders of magnitude faster on a typical laptop running a few threads compared to the automated and high throughput MOFid approach even when it is performed on several nodes of a high-performance computing cluster, and of course quicker still than the more user dependent ToposPro approach that requires much more user interaction than the other techniques. CrystalNets has the power to perform topological identification on tens of thousands of MOF and MOF-like structures with notable reliability, in a recent study by Burner et al. [79] this software was used to identify the topology of 72,257 MOFs, for a new database ARC-MOF, with a match to file name 93% of the time and at a rate that can outperform a competent and experienced researcher investigating a single structure in ToposPro. The entire database of ARC-MOF could be assessed within a single afternoon on a regular computer using a multi-threading approach [79].

In addition to the Julia module there is also an online web interface which allows for the upload of CIF files, with a more user-friendly process for topological identification of individual structures than running them in the Julia interface. The available options using the newly released online version of the CrystalNets software can be seen in Figure 2.12.

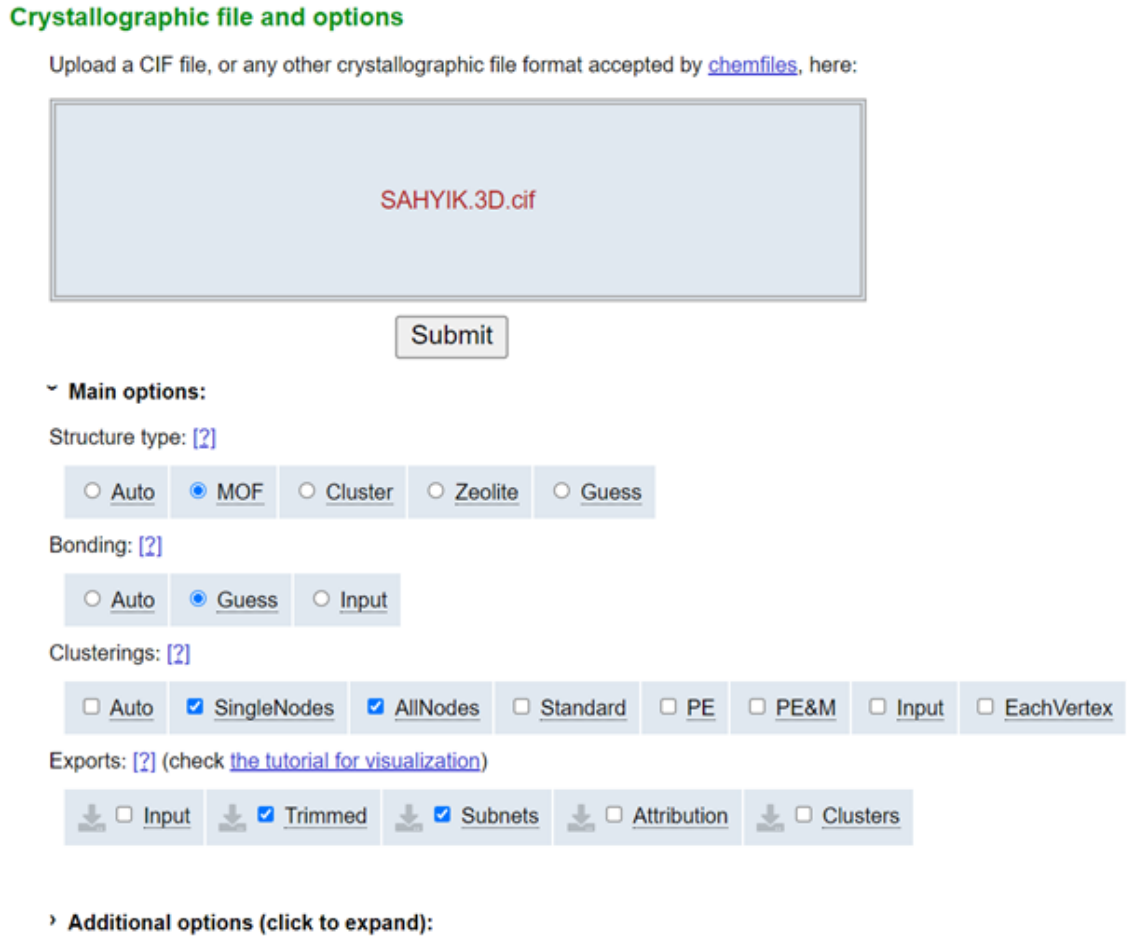


Figure 2.12: *A snapshot of the online interface of MOFid's web structure identification and topology tool, showing the options available for each uploaded CIF file.*

The online version of CrystalNets is not dissimilar to the online interfaces of MOFid or TopCryst, boasting a visualisation tool that shows a simplified net overlaid on the original structure. One major difference is the ability to select the structure style and bonding settings before the structure is uploaded. This is useful for a user who may know specifically which algorithm to select, whereas the reporting of all potential nets by TopCryst via each technique is perhaps more suited to a more inexperienced crystallographer.

2.8.5 Guidance and Limitations

One major limitation of these high-throughput automated approaches for topological assignment of crystalline materials via the medium of CIFs is the lack of verifiability of results returned using these software packages given the subjective nature of topology assignment, something which is only addressed using a manual topological assignment tool. However, the possibility to analyse a prospective structure within several different programmes allows for more certainty surrounding the identification process than using a single approach, particularly when considering the similarity in deconstruction algorithms used across these platforms. We recommend that topological analysis is performed using at least two software packages, running the same algorithm, to verify the results. Should the case arise that the two results disagree then further, more detailed investigation must

take place.

We must consider that some key differences between these software packages exist, the most notable being the technique used to assigned bonds between atoms in non-bonded CIFs. These bonding approaches, despite their apparent similarity contain subtle differences in their approach and it is these subtle differences that can create major changes in the outcome of topological assignment software. To check that topology for a large set of structures has not been incorrectly assigned, it is possible to cross reference structure refcodes with the CSD's 1D, 2D, and 3D structure subset, and the resultant topology with the RCSR's 1-periodic, 2-periodic, and 3-periodic net database, however some errors may persist.

2.9 Recent Developments

In recent years, many groups around the world have employed ToposPro to identify the topology of individual structures, or larger sets of crystal data, and used this information alongside other properties to create data sets for MOF and MOF like materials. In a recent study by Cheng et al. [80], Topos software was used for the topological classification of coordination polymers which were generated in the exploration of H2pdba, an adaptable linker. It was used to assemble a diversity of new Mn, Co, Ni, and Cu coordination polymers into 2D metal–organic layers and 3D MOFs which disclosed several types of topologies including **sql**, **hcb**, and **tfk**. There are many examples of the implementation of Topos for structure analysis within the community and these can be found within the 2000+ citations of the ToposPro software package, although not all of these publications are exclusively MOF related. It is imperative in this review that we should include the introduction of the TopCryst online package [81] which was made available for use only in March 2022, followed very quickly by that of the CrystalNets web interface that came online just six months later in October 2022. The TopCryst service has already been cited several times in significant journal publications within the first six months of its release.

There have also been several examples of recent implementations of MOFid to explore the importance of structure topology. One primary example is the recent publication of the Automated Reticular Framework (RF) Discovery platform by Pollice et al. [82] in 2021 where they implement data obtained using the tools published in MOFid for a data-driven strategy focused on accelerated materials design. Knowing the physically feasible topologies for structures based on chosen linkers has also been useful for bottom-up MOF building approaches where the topologies and linkers of previously synthesised MOFs had been extracted from the CoRE MOF database using MOFid [45, 28].

In another study, MOFid was used to identify Cu paddlewheel MOFs from a set of 1172 non-disordered MOFs to investigate structural collapse during activation [83]. Once these structures were gathered it was possible to perform high-throughput computational analysis to investigate the effect of various mechanical properties.

Lastly, the CrystalNets publication, despite its recent publication, has already received several citations from studies focused on the topological identification of MOF structures. It was first used in print to characterise the topology of 100 Zr-oxide MOFs, before it was then applied to much larger sets of data by Burner et al. on a group of approximately 72,000 MOFs that included previously known topologies [84, 79]. Later, Glasby et al. ran CrystalNets on ca. 28,000 experimental MOFs from the CSD 3D MOF subset for the first

time during the development of the DigiMOF database [25]. Mourino et al. also used CrystalNets to characterise the topology of over 300 COFs as prospective candidates for photocatalysis, showing that the use of this software is not limited to MOFs alone [85].

2.10 Conclusions and Perspective

The availability of these software packages shows that topological characterisation of crystal structures is important, not only to MOF researchers but also to those interested in COFs, Zeolites, and other crystals that form periodic networks in their atomic structure. MOF synthesis can play a major role in topological determination as different conditions lead to the formation of topologically different structures, influencing not only the resultant mechanical stability but also the pore shape and sizes of a crystal depending on the SBUs and linker types that have been selected for their synthesis.

The choice of topological assignment software is highly likely to depend on the requirements of the individual study, as each different tool has its own strengths and limitations. CrystalNets is more notable for its speed and its ability to read in atom bonding information, but it does not offer the same chemical structure insights as MOFid for example, and its choice of topological representations is limited compared with ToposPro. However, ToposPro has an advantage in that any structure can be manually modified during the deconstruction process increasing accuracy when used by experienced crystallographers compared to fully automated methods.

A notable limitation of all software approaches is when comparison between single node and all node topology allocation differ from each other. Following IUPAC guidelines as outlined in this article, any cases where a different net is reported the result should be designated as "the **xxx**-derived net **yyy**", something we note is seldom seen. A simple change in the software output to reflect this might help researchers to ensure they are reporting in line with the guidelines.

Lastly, we reiterate that to date there is not yet a complete, freely available database of MOFs that contains the relevant RCSR or other topology type for all structures that has been proven and adequately verified. The introduction of resources such as the QMOF database [86], which contains over 20,000 MOFs and their quantum-chemical properties serves as an example of the importance of publishing key data to limit the need to repeat computational calculations between research groups. Once a database of MOF topologies has been properly curated and confirmed it can prevent the need for repetition. Whilst the topologies reported in the CoRE MOF database has been a good start, there are still improvements to be made.

The CSD is an ideal target for a database that could include topological information published during deposition given its manual curation, continuous quarterly updates, and extensive searching tools. Whilst we note here that the CSD system itself is not freely available, individual structures are through the CCDC's access structures service, and should the relevant topology be contained within a deposited CIF then that information would become freely available, as the individual deposited CIFs can be downloaded from the respective entries.

References

[1] Bernard F. Hoskins and Richard Robson. Infinite polymeric frameworks consisting of three dimensionally linked rod-like segments. *J. Am. Chem. Soc.*, 111(15):5962–5964, July 1989. Publisher: American Chemical Society.

[2] Omar M. Yaghi, Michael O'Keeffe, Nathan W. Ockwig, Hee K. Chae, Mohamed Eddaoudi, and Jaheon Kim. Reticular synthesis and the design of new materials. *Nature*, 423(6941):705–714, June 2003. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6941 Primary_atype: Reviews Publisher: Nature Publishing Group.

[3] Christopher A. Trickett, Aasif Helal, Bassem A. Al-Maythalony, Zain H. Yamani, Kyle E. Cordova, and Omar M. Yaghi. The chemistry of metal–organic frameworks for CO2 capture, regeneration and conversion. *Nat Rev Mater*, 2(8):1–16, July 2017. Number: 8 Publisher: Nature Publishing Group.

[4] Jian-Bin Lin, Tai T. T. Nguyen, Ramanathan Vaidhyanathan, Jake Burner, Jared M. Taylor, Hana Durekova, Farid Akhtar, Roger K. Mah, Omid Ghaffari-Nik, Stefan Marx, Nicholas Fylstra, Simon S. Iremonger, Karl W. Dawson, Partha Sarkar, Pierre Hovington, Arvind Rajendran, Tom K. Woo, and George K. H. Shimizu. A scalable metal-organic framework as a durable physisorbent for carbon dioxide capture. *Science*, 374(6574):1464–1469, December 2021. Publisher: American Association for the Advancement of Science.

[5] Eyas Mahmoud, Labeeb Ali, Asmaa El Sayah, Sara Awni Alkhatib, Hend Abdulsalam, Mouza Juma, and Ala'a H. Al-Muhtaseb. Implementing Metal-Organic Frameworks for Natural Gas Storage. *Crystals*, 9(8):406, August 2019. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[6] Lerato Y. Molefe, Nicholas M. Musyoka, Jianwei Ren, Henrietta W. Langmi, Mkhulu Mathe, and Patrick G. Ndungu. Effect of Inclusion of MOF-Polymer Composite onto a Carbon Foam Material for Hydrogen Storage Application. *J Inorg Organomet Polym*, 31(1):80–88, January 2021.

[7] Mark D. Allendorf, Vitalie Stavila, Jonathan L. Snider, Matthew Witman, Mark E. Bowden, Kriston Brooks, Ba L. Tran, and Tom Autrey. Challenges to developing materials for the transport and storage of hydrogen. *Nat. Chem.*, 14(11):1214–1223, November 2022. Number: 11 Publisher: Nature Publishing Group.

[8] Mehdi Ghommem, Vladimir Puzyrev, Rana Sabouni, and Fehmi Najar. Deep learning for gas sensing using MOFs coated weakly-coupled microbeams. *Applied Mathematical Modelling*, 105:711–728, May 2022.

[9] Jack Gonzalez, Krishnendu Mukherjee, and Yamil J. Colón. Understanding Structure–Property Relationships of MOFs for Gas Sensing through Henry's Constants. *J. Chem. Eng. Data*, 68(1):291–302, January 2023. Publisher: American Chemical Society.

[10] Vlad Pascanu, Greco González Miera, A. Ken Inge, and Belén Martín-Matute. Metal–Organic Frameworks as Catalysts for Organic Synthesis: A Critical Perspective. *J. Am. Chem. Soc.*, 141(18):7223–7234, May 2019. Publisher: American Chemical Society.

[11] Yu Shen, Ting Pan, Liu Wang, Zhen Ren, Weina Zhang, and Fengwei Huo. Programmable Logic in Metal–Organic Frameworks for Catalysis. *Advanced Materials*, 33(46):2007442, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adma.202007442.

[12] Shaoyang Lin, Pavel M. Usov, and Amanda J. Morris. The role of redox hopping in metal–organic framework electrocatalysis. *Chemical communications (Cambridge, England)*, 54(51):6965–6974, 2018. Place: CAMBRIDGE Publisher: Royal Society of Chemistry RSC, ROYAL SOC CHEMISTRY.

[13] Jorge Gascon, María D. Hernández-Alonso, Ana Rita Almeida, Gerard P. M. van Klink, Freek Kapteijn, and Guido Mul. Isoreticular MOFs as Efficient Photocatalysts with Tunable Band Gap: An Operando FTIR Study of the Photoinduced Oxidation of Propylene. *ChemSusChem*, 1(12):981–983, 2008. _eprint: https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cssc.200800203.

[14] Nazario Lopez, Hanhua Zhao, Akira Ota, Andrey V. Prosvirin, Eric W. Reinheimer, and Kim R. Dunbar. Unprecedented Binary Semiconductors Based on TCNQ: Single-Crystal X-ray Studies and Physical Properties of Cu(TCNQX2) X=Cl, Br. *Advanced Materials*, 22(9):986–989, 2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adma.200903217.

[15] Lei Sun, Michael G. Campbell, and Mircea Dincă. Electrically Conductive Porous Metal-Organic Frameworks. *Angewandte Chemie (International ed.)*, 55(11):3566–3579, 2016. Place: WEINHEIM Publisher: Wiley, WILEY-V C H VERLAG GMBH, Wiley Subscription Services, Inc.

[16] Alexander P. Shevchenko, Eugeny V. Alexandrov, Andrey A. Golov, Olga A. Blatova, Alexandra S. Duyunova, and Vladislav A. Blatov. Topology versus porosity: what can reticular chemistry tell us about free space in metal–organic frameworks? *Chem. Commun.*, 56(67):9616–9619, August 2020. Publisher: The Royal Society of Chemistry.

[17] Peyman Z. Moghadam, Sven M. J. Rogge, Aurelia Li, Chun-Man Chow, Jelle Wieme, Noushin Moharrami, Marta Aragones-Anglada, Gareth Conduit, Diego A. Gomez-Gualdron, Veronique Van Speybroeck, and David Fairen-Jimenez. Structure-Mechanical Stability Relations of Metal-Organic Frameworks via Machine Learning. *Matter*, 1(1):219–234, July 2019.

[18] N. Scott Bobbitt, Andrew S. Rosen, and Randall Q. Snurr. Topological effects on separation of alkane isomers in metal-organic frameworks. *Fluid Phase Equilibria*, 519:112642, September 2020.

[19] Peyman Z. Moghadam, Aurelia Li, Seth B. Wiggin, Andi Tao, Andrew G. P. Maloney, Peter A. Wood, Suzanna C. Ward, and David Fairen-Jimenez. Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future. *Chem. Mater.*, 29(7):2618–2625, April 2017. Publisher: American Chemical Society.

[20] P. Z. Moghadam, A. Li, X.-W. Liu, R. Bueno-Perez, S.-D. Wang, S. B. Wiggin, P. A. Wood, and D. Fairen-Jimenez. Targeted classification of metal–organic frameworks in the Cambridge structural database (CSD). *Chemical Science*, 11(32):8373–8387, 2020. Publisher: Royal Society of Chemistry RSC.

[21] Aurelia Li, Rocio Bueno Perez, Seth Wiggin, Suzanna C. Ward, Peter A. Wood, and David Fairen-Jimenez. The launch of a freely accessible MOF CIF collection from the CSD. *Matter*, 4(4):1105–1106, April 2021.

[22] Brendan F. Abrahams, A. David Dharma, Brendan Dyett, Timothy A. Hudson, Helen Maynard-Casely, Christopher J. Kingsbury, Laura J. McCormick, Richard Robson, Ashley L. Sutton, and Keith F. White. An indirect generation of 1D MII-2,5-dihydroxybenzoquinone coordination polymers, their structural rearrangements and generation of materials with a high affinity for H2, CO2 and CH4. *Dalton Trans.*, 45(4):1339–1344, January 2016. Publisher: The Royal Society of Chemistry.

[23] Stuart R. Batten, Neil R. Champness, Xiao-Ming Chen, Javier Garcia-Martinez, Susumu Kitagawa, Lars Öhrström, Michael O'Keeffe, Myunghyun Paik Suh, and Jan Reedijk. Terminology of metal–organic frameworks and coordination polymers (IUPAC Recommendations 2013). *Pure and Applied Chemistry*, 85(8):1715–1724, July 2013. Publisher: De Gruyter Section: Pure and Applied Chemistry.

[24] Lars Öhrström. Let's Talk about MOFs—Topology and Terminology of Metal-Organic Frameworks and Why We Need Them. *Crystals*, 5(1):154–162, March 2015. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[25] Lawson T. Glasby, Kristian Gubsch, Rosalee Bence, Rama Oktavian, Kesler Isoko, Seyed Mohamad Moosavi, Joan L. Cordiner, Jason C. Cole, and Peyman Z. Moghadam. DigiMOF: A Database of Metal–Organic Framework Synthesis Information Generated via Text Mining. *Chem. Mater.*, 35(11):4510–4524, June 2023. Publisher: American Chemical Society.

[26] Vladislav A. Blatov, Alexander P. Shevchenko, and Davide M. Proserpio. Applied Topological Analysis of Crystal Structures with the Program Package ToposPro. *Crystal Growth & Design*, 14(7):3576–3586, July 2014. Publisher: American Chemical Society.

[27] S. J. Ramsden, V. Robins, and S. T. Hyde. Three-dimensional Euclidean nets from two-dimensional hyperbolic tilings: kaleidoscopic examples. *Acta Cryst A*, 65(2):81–108, March 2009. Number: 2 Publisher: International Union of Crystallography.

[28] Benjamin J. Bucior, Andrew S. Rosen, Maciej Haranczyk, Zhenpeng Yao, Michael E. Ziebel, Omar K. Farha, Joseph T. Hupp, J. Ilja Siepmann, Alán Aspuru-Guzik, and Randall Q. Snurr. Identification Schemes for Metal–Organic Frameworks To Enable Rapid Search and Cheminformatics Analysis. *Crystal Growth & Design*, 19(11):6682–6697, November 2019. Publisher: American Chemical Society.

[29] Lionel Zoubritzky and François-Xavier Coudert. CrystalNets.jl: Identification of Crystal Topologies. *SciPost Chem.*, 1(2):005, June 2022.

[30] Alexander Frank Wells. *Three Dimensional Nets and Polyhedra*. Wiley, 1977.

[31] Zicheng Pan, Hong Sun, Yi Zhang, and Changfeng Chen. Harder than Diamond: Superior Indentation Strength of Wurtzite BN and Lonsdaleite. *Phys. Rev. Lett.*, 102(5):055503, February 2009. Publisher: American Physical Society.

[32] Xingyu Li, Jiaqi Liu, Kang Zhou, Saif Ullah, Hao Wang, Jizhao Zou, Timo Thonhauser, and Jing Li. Tuning Metal–Organic Framework (MOF) Topology by Regulating Ligand and Secondary Building Unit (SBU) Geometry: Structures Built on

8-Connected M6 (M = Zr, Y) Clusters and a Flexible Tetracarboxylate for Propane-Selective Propane/Propylene Separation. *J. Am. Chem. Soc.*, 144(47):21702–21709, November 2022. Publisher: American Chemical Society.

[33] Charlotte Bonneau, Michael O'Keeffe, Davide M. Proserpio, Vladislav A. Blatov, Stuart R. Batten, Susan A. Bourne, Myoung Soo Lah, Jean-Guillaume Eon, Stephen T. Hyde, Seth B. Wiggin, and Lars Öhrström. Deconstruction of Crystalline Networks into Underlying Nets: Relevance for Terminology Guidelines and Crystallographic Databases. *Crystal Growth & Design*, 18(6):3411–3418, June 2018. Publisher: American Chemical Society.

[34] Dayoung Ryu and Jihan Kim. Database Design for Double-Linker Metal–Organic Frameworks. *J. Phys. Chem. C*, 127(21):10384–10390, June 2023. Publisher: American Chemical Society.

[35] Volodymyr Bon, Irena Senkovska, Manfred S. Weiss, and Stefan Kaskel. Tailoring of network dimensionality and porosity adjustment in Zr- and Hf-based MOFs. *CrystEngComm*, 15(45):9572–9577, October 2013. Publisher: The Royal Society of Chemistry.

[36] C. F. Macrae, I. Sovago, S. J. Cottrell, P. T. A. Galek, P. McCabe, E. Pidcock, M. Platings, G. P. Shields, J. S. Stevens, M. Towler, and P. A. Wood. Mercury 4.0: from visualization to analysis, design and prediction. *J Appl Cryst*, 53(1):226–235, February 2020. Number: 1 Publisher: International Union of Crystallography.

[37] T. Hahn. *International Tables for Crystallography Volume A: Space-group Symmetry*, volume A. Wiley, 5th edition, 2006. Publisher: International Union of Crystallography.

[38] Maarten G. Goesten, Freek Kapteijn, and Jorge Gascon. Fascinating chemistry or frustrating unpredictability: observations in crystal engineering of metal–organic frameworks. *CrystEngComm*, 15(45):9249–9257, October 2013. Publisher: The Royal Society of Chemistry.

[39] Dongwook Kim, Xinfang Liu, and Myoung Soo Lah. Topology analysis of metal–organic frameworks based on metal–organic polyhedra as secondary or tertiary building units. *Inorg. Chem. Front.*, 2(4):336–360, March 2015. Publisher: The Royal Society of Chemistry.

[40] Lucia Carlucci, Gianfranco Ciani, and Davide M. Proserpio. Polycatenation, polythreading and polyknotting in coordination network chemistry. *Coordination Chemistry Reviews*, 246(1):247–289, November 2003.

[41] Tatiana G. Mitina and Vladislav A. Blatov. Topology of 2-Periodic Coordination Networks: Toward Expert Systems in Crystal Design. *Crystal Growth & Design*, 13(4):1655–1664, April 2013. Publisher: American Chemical Society.

[42] Lucia Carlucci, Gianfranco Ciani, Davide M. Proserpio, Tatiana G. Mitina, and Vladislav A. Blatov. Entangled Two-Dimensional Coordination Networks: A General Survey. *Chem. Rev.*, 114(15):7557–7580, August 2014. Publisher: American Chemical Society.

[43] O. Delgado Friedrichs, M. O'Keeffe, and O. M. Yaghi. Three-periodic nets and tilings: regular and quasiregular nets. *Acta Cryst A*, 59(1):22–27, January 2003. Number: 1 Publisher: International Union of Crystallography.

[44] Charlotte Bonneau, Olaf Delgado-Friedrichs, Michael O'Keeffe, and Omar M. Yaghi. Three-periodic nets and tilings: minimal nets. *Acta Crystallogr A*, 60(Pt 6):517–520, November 2004.

[45] Olaf Delgado-Friedrichs and Michael O'Keeffe. Crystal nets as graphs: Terminology and definitions. *Journal of Solid State Chemistry - J SOLID STATE CHEM*, 178:2480–2485, August 2005.

[46] Olaf Delgado-Friedrichs, Michael O'Keeffe, and Omar M. Yaghi. Three-periodic nets and tilings: edge-transitive binodal structures. *Acta Crystallogr A*, 62(Pt 5):350–355, September 2006.

[47] Olaf Delgado-Friedrichs, Martin D. Foster, Michael O'Keeffe, Davide M. Proserpio, Michael M. J. Treacy, and Omar M. Yaghi. What do we know about three-periodic nets? *Journal of Solid State Chemistry*, 178(8):2533–2554, August 2005.

[48] Olaf Delgado-Friedrichs, Michael O'Keeffe, and Omar M. Yaghi. Taxonomy of periodic nets and the design of materials. *Physical Chemistry Chemical Physics*, 9(9):1035–1043, 2007. Publisher: Royal Society of Chemistry.

[49] V. A. Blatov, O. Delgado-Friedrichs, M. O'Keeffe, and D. M. Proserpio. Three-periodic nets and tilings: natural tilings for nets. *Acta Cryst A*, 63(5):418–425, September 2007. Number: 5 Publisher: International Union of Crystallography.

[50] Nathan W. Ockwig, Olaf Delgado-Friedrichs, Michael O'Keeffe, and Omar M. Yaghi. Reticular Chemistry: Occurrence and Taxonomy of Nets and Grammar for the Design of Frameworks. *Acc. Chem. Res.*, 38(3):176–182, March 2005. Publisher: American Chemical Society.

[51] O. Delgado-Friedrichs and M. O'Keeffe. Identification of and symmetry computation for crystal nets. *Acta Cryst A*, 59(4):351–360, July 2003. Number: 4 Publisher: International Union of Crystallography.

[52] Ch. Baerlocher and L.B. McCusker. Database of Zeolite Structures, 1996.

[53] R.M. Barrer. Chemical Nomenclature and Formulation of Compositions of Synthetic and Natural Zeolites. *Pure Appl. Chem.*, 51:1091–1100, 1979.

[54] Zhijie Chen, Hao Jiang, Michael O'Keeffe, and Mohamed Eddaoudi. Minimal edge-transitive nets for the design and construction of metal–organic frameworks. *Faraday Discuss.*, 201(0):127–143, September 2017. Publisher: The Royal Society of Chemistry.

[55] Zhijie Chen, Hao Jiang, Mian Li, Michael O'Keeffe, and Mohamed Eddaoudi. Reticular Chemistry 3.2: Typical Minimal Edge-Transitive Derived and Related Nets for the Design and Synthesis of Metal–Organic Frameworks. *Chem. Rev.*, 120(16):8039–8065, August 2020. Publisher: American Chemical Society.

[56] Frank Hoffmann. *Introduction to Crystallography*. Springer International Publishing, Cham, 2020.

[57] Basic and derived nets. *The Fascination of Crystals and Symmetry*, August 2015.

[58] O. Delgado-Friedrichs, M. O'Keeffe, D. M. Proserpio, and M. M. J. Treacy. Three-periodic nets, tilings and surfaces. A short review and new results. *Acta Cryst A*, 79(2):192–202, March 2023. Number: 2 Publisher: International Union of Crystallography.

[59] Dawei Feng, Zhi-Yuan Gu, Jian-Rong Li, Hai-Long Jiang, Zhangwen Wei, and Hong-Cai Zhou. Zirconium-Metalloporphyrin PCN-222: Mesoporous Metal–Organic Frameworks with Ultrahigh Stability as Biomimetic Catalysts. *Angewandte Chemie International Edition*, 51(41):10307–10310, 2012. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201204475.

[60] Qipu Lin, Xianhui Bu, Aiguo Kong, Chengyu Mao, Xiang Zhao, Fei Bu, and Pingyun Feng. New Heterometallic Zirconium Metalloporphyrin Frameworks and Their Heteroatom-Activated High-Surface-Area Carbon Derivatives. *J. Am. Chem. Soc.*, 137(6):2235–2238, February 2015. Publisher: American Chemical Society.

[61] Mian Li, Dan Li, Michael O'Keeffe, and Omar M. Yaghi. Topological Analysis of Metal–Organic Frameworks with Polytopic Linkers and/or Multiple Building Units and the Minimal Transitivity Principle. *Chem. Rev.*, 114(2):1343–1370, January 2014. Publisher: American Chemical Society.

[62] Morsy A. M. Abu-Youssef, Vratislav Langer, and Lars Öhrström. A unique example of a high symmetry three- and four-connected hydrogen bonded 3D-network. *Chem. Commun.*, (10):1082–1084, March 2006. Publisher: The Royal Society of Chemistry.

[63] Jarrod F. Eubank, Rosa D. Walsh, Pankaj Poddar, Hariharan Srikanth, Randy W. Larsen, and Mohamed Eddaoudi. MetalOrganic Framework Diversity via Heterocoordination of a Multifunctional Ligand: SrAl2 and a Novel (3,4)-Connected Network. *Crystal Growth & Design*, 6(6):1453–1457, June 2006. Publisher: American Chemical Society.

[64] Michael O'Keeffe and Omar M. Yaghi. Deconstructing the Crystal Structures of Metal–Organic Frameworks and Related Materials into Their Underlying Nets. *Chem. Rev.*, 112(2):675–702, February 2012. Publisher: American Chemical Society.

[65] Senja Barthel, Eugeny V. Alexandrov, Davide M. Proserpio, and Berend Smit. Distinguishing Metal–Organic Frameworks. *Crystal Growth & Design*, 18(3):1738–1747, March 2018. Publisher: American Chemical Society.

[66] Hailian Li, Mohamed Eddaoudi, M. O'Keeffe, and O. M. Yaghi. Design and synthesis of an exceptionally stable and highly porous metal-organic framework. *Nature*, 402(6759):276–279, November 1999.

[67] Markus J. Kalmutzki, Nikita Hanikel, and Omar M. Yaghi. Secondary building units as the turning point in the development of the reticular chemistry of MOFs. *Science Advances*, 4(10):9180, October 2018. Publisher: American Association for the Advancement of Science.

[68] Dmytro Denysenko, Maciej Grzywa, Markus Tonigold, Barbara Streppel, Ivana Krkljus, Michael Hirscher, Enrico Mugnaioli, Ute Kolb, Jan Hanss, and

Dirk Volkmer. Elucidating Gating Effects for Hydrogen Sorption in MFU-4-Type Triazolate-Based Metal–Organic Frameworks Featuring Different Pore Sizes. *Chemistry – A European Journal*, 17(6):1837–1848, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/chem.201001872.

[69] Christopher E. Wilmer, Michael Leaf, Chang Yeon Lee, Omar K. Farha, Brad G. Hauser, Joseph T. Hupp, and Randall Q. Snurr. Large-scale screening of hypothetical metal–organic frameworks. *Nature Chem*, 4(2):83–89, February 2012.

[70] Jacob Goldsmith, Antek G. Wong-Foy, Michael J. Cafarella, and Donald J. Siegel. Theoretical Limits of Hydrogen Storage in Metal–Organic Frameworks: Opportunities and Trade-Offs. *Chem. Mater.*, 25(16):3373–3382, August 2013. Publisher: American Chemical Society.

[71] Yongchul G. Chung, Jeffrey Camp, Maciej Haranczyk, Benjamin J. Sikora, Wojciech Bury, Vaiva Krungleviciute, Taner Yildirim, Omar K. Farha, David S. Sholl, and Randall Q. Snurr. Computation-Ready, Experimental Metal–Organic Frameworks: A Tool To Enable High-Throughput Screening of Nanoporous Crystals. *Chem. Mater.*, 26(21):6185–6192, November 2014. Publisher: American Chemical Society.

[72] Yongchul G. Chung, Emmanuel Haldoupis, Benjamin J. Bucior, Maciej Haranczyk, Seulchan Lee, Hongda Zhang, Konstantinos D. Vogiatzis, Marija Milisavljevic, Sanliang Ling, Jeffrey S. Camp, Ben Slater, J. Ilja Siepmann, David S. Sholl, and Randall Q. Snurr. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data*, 64(12):5985–5998, December 2019. Publisher: American Chemical Society.

[73] Cheryl D. L. Saunders, Neil Burford, Ulrike Werner-Zwanziger, and Robert McDonald. Preparation and Comprehensive Characterization of [Hg6(Alanine)4(NO3)4]·H2O. *Inorg. Chem.*, 47(9):3693–3699, May 2008. Publisher: American Chemical Society.

[74] A. A. Ashcheulov, O. N. Manyk, T. O. Manyk, S. F. Marenkin, and V. R. Bilynskiy-Slotylo. Chemical bonding in cadmium. *Inorg Mater*, 47(9):952–956, September 2011.

[75] Aurel Tăbăcaru, Claudio Pettinari, Fabio Marchetti, Corrado di Nicola, Konstantin V. Domasevitch, Simona Galli, Norberto Masciocchi, Stefania Scuri, Iolanda Grappasonni, and Mario Cocchioni. Antibacterial Action of 4,4-Bipyrazolyl-Based Silver(I) Coordination Polymers Embedded in PE Disks. *Inorg. Chem.*, 51(18):9775–9788, September 2012. Publisher: American Chemical Society.

[76] Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, October 2011.

[77] Xiaonan Gao, Ai-Yun Fu, Bo Liu, Jun-cheng Jin, Long-Tao Dou, and Li-Xia Chen. Unique Topology Analysis by ToposPro for a Metal–Organic Framework with Multiple Coordination Centers. *Inorg. Chem.*, 58(5):3099–3106, March 2019. Publisher: American Chemical Society.

[78] Guillaume Fraux, Jonathan Fine, Len Kimms, German P. Barletta, Mykola Dimura, F. X. Coudert, pelsa, Maximilien Levesque, Shoubhik Maiti, Simon Guionniere, and jmintser. chemfiles/chemfiles: Version 0.10.2, December 2021.

[79] Jake Burner, Jun Luo, Andrew White, Adam Mirmiran, Ohmin Kwon, Peter G. Boyd, Stephen Maley, Marco Gibaldi, Scott Simrod, Victoria Ogden, and Tom K. Woo. ARC–MOF: A Diverse Database of Metal-Organic Frameworks with DFT-Derived Partial Atomic Charges and Descriptors for Machine Learning. *Chem. Mater.*, 35(3):900–916, February 2023. Publisher: American Chemical Society.

[80] Xiaoyan Cheng, Lirong Guo, Hongyu Wang, Jinzhong Gu, Ying Yang, Marina V. Kirillova, and Alexander M. Kirillov. Coordination Polymers Constructed from an Adaptable Pyridine-Dicarboxylic Acid Linker: Assembly, Diversity of Structures, and Catalysis. *Inorg. Chem.*, 61(45):17951–17962, November 2022. Publisher: American Chemical Society.

[81] Alexander P. Shevchenko, Aleksandr A. Shabalin, Igor Yu. Karpukhin, and Vladislav A. Blatov. Topological representations of crystal structures: generation, analysis and implementation in the TopCryst system. *Science and Technology of Advanced Materials: Methods*, 2(1):250–265, December 2022. https://doi.org/10.1080/27660400.2022.2088041.

[82] Robert Pollice, Gabriel dos Passos Gomes, Matteo Aldeghi, Riley J. Hickman, Mario Krenn, Cyrille Lavigne, Michael Lindner-D'Addario, AkshatKumar Nigam, Cher Tian Ser, Zhenpeng Yao, and Alán Aspuru-Guzik. Data-Driven Strategies for Accelerated Materials Design. *Acc. Chem. Res.*, 54(4):849–860, February 2021. Publisher: American Chemical Society.

[83] Saad Aldin Mohamed, Yeongjin Kim, Junkee Lee, Wonyoung Choe, and Jihan Kim. Understanding the Structural Collapse during Activation of Metal–Organic Frameworks with Copper Paddlewheels. *Inorg. Chem.*, 61(25):9702–9709, June 2022. Publisher: American Chemical Society.

[84] Rama Oktavian, Raymond Schireman, Lawson T. Glasby, Guanming Huang, Federica Zanca, David Fairen-Jimenez, Michael T. Ruggiero, and Peyman Z. Moghadam. Computational Characterization of Zr-Oxide MOFs for Adsorption Applications. *ACS Appl. Mater. Interfaces*, 14(51):56938–56947, December 2022. Publisher: American Chemical Society.

[85] Beatriz Mourino, Kevin Maik Jablonka, Andres Ortega-Guerrero, and Berend Smit. In Search of Covalent Organic Framework Photocatalysts: A DFT-Based Screening Approach. *Advanced Functional Materials*, 33(32):2301594, 2023.

[86] Andrew S. Rosen, Shaelyn M. Iyer, Debmalya Ray, Zhenpeng Yao, Alán Aspuru-Guzik, Laura Gagliardi, Justin M. Notestein, and Randall Q. Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, 2021.

Chapter 3

Integrating CrystalNets.jl and Bench-marking Performance on the Cambridge Structural Database

3.1 Publication Information and Paper Contributions

This paper has been prepared for publication as short article and is currently under review.

In this publication I, the candidate, wrote the manuscript under the supervision of Professor Joan L. Cordiner, Dr Jason C. Cole, and Dr Peyman Z. Moghadam.

3.2 Abstract

Obtaining and verifying the results of topological assignment is not an easy task for the many complex and confusing MOF structures that exist in materials space, and even manually attempting to visualise the assigned underlying nets can be a challenge. We initially compared the use of two high throughput topological assignment packages, MOFid and CrystalNets on a large set of CSD 2D and 3D MOFs to gauge the agreement between approaches and then ran these same structures after implementing a custom CSD Python API workflow that prepared CSD structures prior to the use of CrystalNets for topological assignment. A total of 54,473 experimental structures, consisting of 28,962 3D MOFs supplemented by an additional 25,511 2D MOFs, were topologically analysed using the three approaches. We believe this is the first comparison of topological assignment tools using the CSD MOF subset. Additionally, we developed a new Python based approach that allows a user to perform topological assignment in CSD Mercury using the CSD Python API at the click of a button.

3.2.1 Keywords

Metal-organic frameworks, Topology, Crystallography, Software Analysis

3.3 Introduction

Metal-organic frameworks (MOFs) are an intensely studied area of crystal chemistry. Known for their porosity, MOFs are primarily considered for applications in areas such as gas sensing [1, 2, 3], storage [4, 5, 6], and separation [7, 8, 9], as well as catalysis [10, 11], and drug-delivery [12, 13]. Consisting primarily of metal primary building units (PBUs) joined by branches of organic molecules, MOFs can form some beautiful and unique crystallographic configurations with almost limitless possibilities. The underlying connectivity representations of these materials form the basis of their topology, and to identify these configurations it is often necessary to analyse their unit cells using a powerful computational tool such as ToposPro [14], Systre [15], MOFid [16], or CrystalNets [17]. Once complex linkers have been broken down into more simplistic branches, the most basic representations can be matched to an underlying mathematically described net obtained from a database such as the Reticular Chemistry Structural Resource (RCSR) [18], the Topological Types Database (TTD) [19, 20], or the Euclidean Patterns In Non-Euclidean Tilings (EPINET) collection [21].

The topology of the structure, once determined, is typically reported within the original experimental manuscript alongside other details surrounding the synthesis of the material [22]. Structures that have been assigned the same RCSR descriptors can be considered to belong to a subset of structures that have the same underlying connectivity, and as a result are likely to share some properties through structure-property relationships. Structures that are considered constituents of the same family, such as ZrO based MOFs, can often demonstrate different physical properties but their topology can be used to help categorise physically similar structures into groups that might show higher mechanical stability, have larger pores, or demonstrate more interesting pore shapes than those that only share a constituent metal type [23, 24]. This is particularly true in cases where the connectivity of the SBUs can vary between 2 to 12 carboxylates.

Furthermore, topological insights are often used to develop hypothetical structures using a bottom-up approach to their synthesis by creating physically different materials using the same building blocks, or creating similar crystal structures using alternate building blocks, and there are several services available that can perform this task such as ToBaCCo [25] or AuToGraFS [26]. There are also many examples of hypothetical MOF databases such as Majumdar et al.'s set of ∼20,000 hMOFs [27], and a more recent ML-focused ultra-stable MOF database developed by Nandy et al. of ∼50,000 MOFs [28], for which the majority report structure topology.

Whilst there are a considerable number of experimentally synthesised structures available from the CSD MOF subset (∼120,000) [29], the CoRE MOF database (∼14,000) [30, 31, 32], and the QMOF database (∼20,000) [33] (the last of which also contains some hypothetical materials), none of these collections contain a complete data set of topological identities for each structure. Whilst the CoRE MOF database [30] can be accessed through the MOFid web service [16], and it is possible to search through these structures using keywords to obtain a set of topologies, we note that searching for NA, ERROR, UNKNOWN, TIMEOUT, or MISMATCH, also returns a long list of constituent MOFs for which there are no results. It is clear then, that there is the need for an improved, accessible database of MOF topologies.

Given the nature of topological assignment software, as discussed in our recent perspective publication [34], we would expect that new and successful results should be able

to be obtained from a proportion of structures labelled as ERROR or MISSING LINKERS by using alternative approaches to assigning topology for these materials. However, at present there is no method to ensure the verifiability for the software approaches used in assignment besides individual interpretation of the results by the investigating scientist performing the runs, and we note further that some of the assigned results in the CoRE MOF database may not agree with the results obtained using an alternative approach.

We might anticipate that a small percentage of structures are constructed using unique or non-verifiable connectivity, but we should in general expect that the majority of 2D and 3D nets (not including those which may be considered disjoint) should be assigned an RCSR [35] topology (or acceptable alternative). Whilst we note that both the MOFid [16] and CrystalNets [17] packages have been used to identify topologies for all MOFs in the CoreMOF [30] database there has yet to have been, to our knowledge, a publication comparing software outputs between all topological types in the CSD MOF subset, besides some comparisons made between text-mined topologies and CrystalNets outputs in the creation of the DigiMOF database [36].

Following the work of Zoubritsky and Coudert (2023) [17], where they compared the use of MOFid and CrystalNets on the CoreMOF database, resulting in a 25% increase in identified structures from an original 36% assignment rate using MOFid to 61% using CrystalNets, we conducted a similar investigation performing high throughput topological identification using both MOFid and CrystalNets on 28,962 MOFs in the 3D MOF subset followed by an additional run on the 25,511 structures of the 2D MOF subset. We employed the single node approach to topological assignment to ensure that the analysis obtained the simplest representations of the crystal structures in question, but we also compared the effect of using a custom CSD API approach, involving the inclusion of bond data, as a step towards preparing structures specifically for topological evaluation in conjunction with the CrystalNets software.

3.4 Workflow and Method

3.4.1 Structure and Software Preparation

To evaluate the outputs of two programs, MOFid and CrystalNets, we downloaded the 2D and 3D MOF subsets (Version 5.43 Apr 2023) using the subset search tool in CSD ConQuest, saving the database files in CIF format and ensuring that the "_geom_bond_" parameters were retained within the files. We also obtained the latest versions of both software suites (as of June 2023), from their respective Github pages.

3.4.2 Software Runs

We initially ran two separate directories of CIFs (2D MOFs and 3D MOFs) on the CrystalNets platform, specifying the structure type as MOF, the clustering as SingleNode, and selecting additional settings so that the software would auto-assign bonding information by using the Guess input parameter. We made no other initial preparation of these structures. Each run took sub-30 minutes using a single thread on a Dell XPS 8940 desktop computer, but it would be possible to use multi-threading with relative ease to further reduce the overall time taken to assess the structures.

Next, we performed a similar investigation but instead used the retained atom bonding information obtained from the CSD. To execute this, we simply altered the input

parameter for bonding, changing it from Guess to Input and performed the structure analysis again.

Lastly, to compare the results of CrystalNets against another approach we submitted these directories of structures to MOFid. The Python based tool takes considerably longer than Julia based CrystalNets and necessitated the use of high-performance computing (HPC) resources. The directories were split into 20 sub-folders containing approximately 1000-1500 structures and submitted as batch jobs to the CSD3 HPC resource. Each folder had an average compute time between 90 and 120 minutes, with an upper bound of almost 4 hours. The overall compute time for all structures in MOFid sat in the region of 50 hours, a significant increase on the 1 hour required to run CrystalNets on an identical set of structures. One minor difference to this approach was that we retained the original setup of MOFid such that if the Single Node algorithm output was not available, the All Node output was printed, rather than printing an UNKNOWN or ERROR output. MOFid uses the open-source chemical toolbox Open Babel [37] to assign bonding information and at present does not take bonding data as an input, so only one analysis run was performed with this approach.

We note that selecting CIFs as an input file could be considered a flawed approach, typically a potentially significant quantity of interesting information is lost when default CIFs are written out from various visualisation programmes. CIFs are not required to contain bonding information by default, merely the positions of the atoms within the crystal's unit cell, but a key feature of obtaining them from the CSD is the ability to retain this key information.

3.4.3 Software Outputs

The data was gathered from each software output and collated into a single TSV file for each dimensionality (2D and 3D). From these collated spreadsheets we were easily able to perform matching and analysis of structures and results, which have been detailed in the results and analysis section of this paper.

The general results of these three runs can be found in Table 3.1. Initially we considered "result" for 2D MOFs to be any output that matched with references from the RCSR database. However, after some initial investigations, the criteria for "result" were modified to consider that many EPINET codes, such as sqc2075, were likely to be found in the 3D subset for results that were obtained from CrystalNets analysis due to the inclusion of the EPINET database in this software. A list was taken from the respective resources (as of June 2023) to allow for matching to be considered. We note later in the analysis section that some of these assumptions did include errors, so the total of "successful results" is slightly lower, however we do note there is benefit to keeping additional data in the preliminary stages to discard later as opposed to discarding potnetially interesting/useful results in the initial analysis stage.

Table 3.1: *A general overview of the output of each approach on 2D and 3D structures before detailed analysis was carried out.*

Condition (2D)	**CrystalNets (Guess)**	**CrystalNets (Input)**	**MOFid**
"Result"	19063	19363	7177
UNKNOWN	4493	4170	3142
ERROR/unstable	1919	1956	12052
Other	36	21	3140
Total	25511	25511	25511
Unique Outputs	112	65	66
Condition (3D)	**CrystalNets (Guess)**	**CrystalNets (Input)**	**MOFid**
"Result"	15441	15919	6911
UNKNOWN	12970	12626	6258
ERROR/unstable	453	405	13722
Other	98	12	2071
Total	28962	28962	28962
Unique Outputs	490	497	301

From the preliminary results we can see a significant increase in the number of "results" from MOFid to both runs in CrystalNets, as well as some fluctuations in the number of unique outputs from the software when we switch from the built in Guess algorithm to opting to include CSD bonding information in the Input category.

3.5 Results and Analysis

3.5.1 Results and Data Cleaning

The data was compiled after each software run in an XLS workbook for cleaning and analysis. These initial output workbooks were cleaned using a custom Python script to extract the relevant CSD refcodes, topology information, and degree of interpenetration. The outputs were sorted into columns based on the software used and a list of RCSR topologies was used to determine which results were 2D nets and which were 3D nets. The entire results workbook is available as Supporting Information.

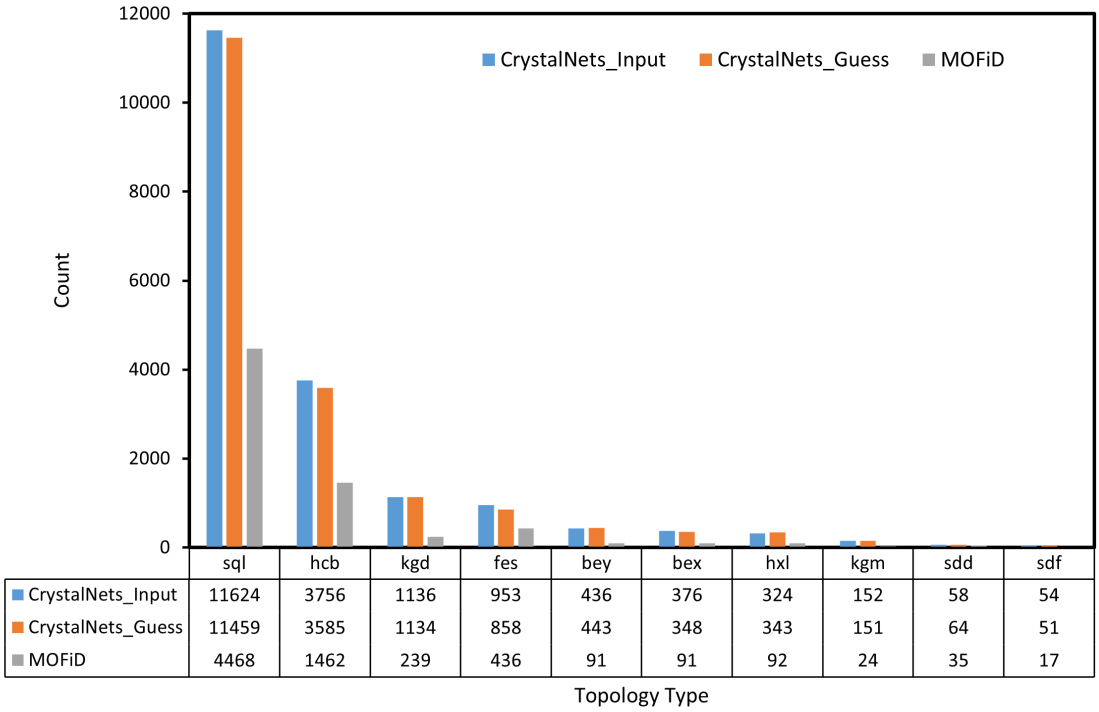
3.5.2 Analysis

First, we investigated the results of running each approach on 2D MOFs. The implementation of MOFid saw a result assigned to 7,177 2D MOFs, and following our analysis it was determined that infact 43 of those results were incorrectly assigned 3D topologies giving a final success rate of 27.96% from a possible 25,511 total structures. Similar analysis was performed for CrystalNets (Guess) which uncovered 150 3D assignments, lowering the successful results to 18,898 and giving us an assignment success rate of 74.08%. Lastly, within the 2D category, CrystalNets (Input) contained only two 3D assignments, dropping the success total to 19,361 and 75.89%. Here, we can show that the retention of bond information has not only increased the number of successful assignments from 18,898 to 19,361 it has also reduced the number of misaligned dimensionalities from 0.01% of assignments to 0.0001%. This is better visualised in Table 3.2.

Table 3.2: *A review of the output of 2D structure analysis.*

Condition (2D)	CrystalNets (Guess)	CrystalNets (Input)	MOFid
Initial 2D Results	19048	19363	7167
Identified 3D Results	150	2	43
Final 2D Results	18898	19361	7134
Success Rate	74.08%	75.89%	27.96%

More investigations led to identification of the frequency of topological types. A histogram, shown inFigure 3.1, noted the top 10 most prevalent 2D MOF topologies identified in our study by the CrystalNets Input approach compare besides the same results from our other two studies. This figure provides a visual representation of the diversity and prevalence of 2D MOF topologies as histogram illustrates the frequency of occurrence of each topology within the dataset. Each bar corresponds to a specific topology, with the height of the bar indicating the number of occurrences observed.



	sql	hcb	kgd	fes	bey	bex	hxl	kgm	sdd	sdf
CrystalNets_Input	11624	3756	1136	953	436	376	324	152	58	54
CrystalNets_Guess	11459	3585	1134	858	443	348	343	151	64	51
MOFiD	4468	1462	239	436	91	91	92	24	35	17

Figure 3.1: *A histogram depicting the most commonly obtained 2D topologies, based off the CrystalNets Input results (blue), but also showing the occurrences for CrystalNets Guess (orange), and MOFid (grey).*

From this histogram we can see that **sql** style structures make up a significant proportion of the total number of 2D MOFs, with a return of 45.5% in the 2D MOF subset. Additionally, **hcb** also make up a significant proportion at almost 15% meaning there is a 3 in 5 chance that any given 2D MOF is configured in one of these two ways.

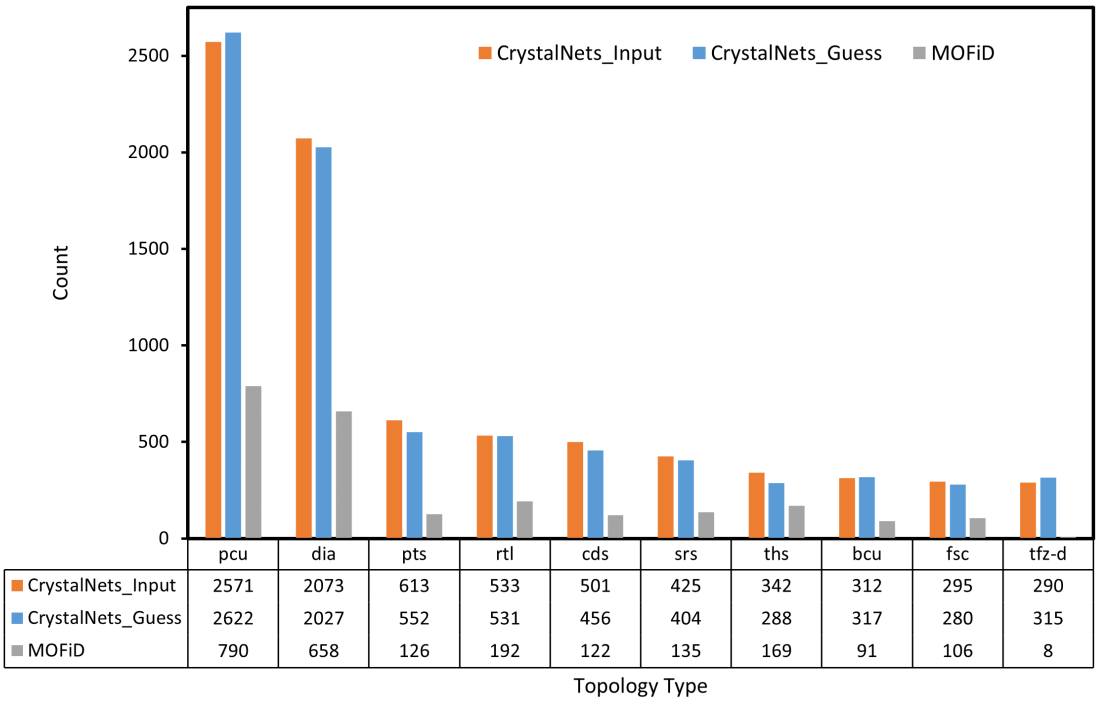
Next, we performed a similar analysis on the 3D set of structures. Beginning again with MOFid, we saw a result assigned to 6,911 structures and following analysis it was determined that 2,078 of these were assigned 2D topologies, leaving a total of 4,833 successful assignments at a rate of 16.69%. Moving on to CrystalNets (Guess) we went

from an initial 15,441 results to 15,346 after 95 2D assignments were removed giving a rate of 50.07%. The final analysis conducted at this stage saw the CrystalNets (Input) total shift from 15,919 to 15,798 resulting in a 3D assignment rate of 51.25% and 121 2D assigned nets. Interesting, the influence of retained bonding information appears to have had less of an effect on 3D MOFs than the initial 2D analysis. This can be seen in Table 3.3 where we also include the EPINET results, however these were not analysed in signficant detail due to an inability to make comparisons across the board and so they have all been considered successful results in this instance.

Table 3.3: *A review of the output of 3D structure analysis.*

Condition (3D)	**CrystalNets (Guess)**	**CrystalNets (Input)**	**MOFid**
Initial 3D Results	14956	14963	6911
Identified 2D Results	95	121	2078
Final 3D Results	14501	14842	4833
EPINET Results	845	956	NA
Success Rate	50.07%	51.25%	16.69%

A deeper investigation into the most commonly occurring topologies within the 3D subset was also conducted, and the results of the top 10 topologies from the CrystalNets Input set are shown in Figure 3.2, a histogram depicting the distribution of the top 10 most prevalent 3D MOF topologies, with comparisons made to the occurrences from the additional sets. The histogram illustrates the frequency of occurrence of each topology, providing insights into the relative abundance and diversity of 3D MOF structures. This visualisation offers a concise overview of the dominant 3D MOF topologies, offering new insight into the structural landscape of these materials.



	pcu	dia	pts	rtl	cds	srs	ths	bcu	fsc	tfz-d
CrystalNets_Input	2571	2073	613	533	501	425	342	312	295	290
CrystalNets_Guess	2622	2027	552	531	456	404	288	317	280	315
MOFiD	790	658	126	192	122	135	169	91	106	8

Figure 3.2: *A histogram depicting the most commonly obtained 3D topologies, based off the CrystalNets Input results (blue), but also showing the occurrences for CrystalNets Guess (orange), and MOFid (grey).*

We found that the most common structure representations are **pcu**, **dia**, **pts**,**rtl**, and **cds**, however it is noted that this order is not necessarily true of the MOFid results which shows some mildly significant discrepancies - particularly for the **pts** and **cds** topologies. In contrast to the 2D subset which shows relatively low topological diversity, **pcu** is the most prevalent topology at approximately 9%, with the top 10 structures combined making up 27% of the 3D MOF subset. This result confirms that a much greater topological diversity exists within the 3D MOF space, as expected of a more complex space with an additional dimension in which the crystals can occupy.

These differences in assignment we see here have been attributed to the atomic bonding approaches that are taken in each respective run. We allow MOFid to assign bonds using OpenBabel, we also allow CrystalNets to assign bonds using its inbuilt CrystalNN approach, and then we finally use the CSD to provide the bonding information as CIFs are entered into the CrystalNets programme. Whilst this could clearly explain the differences between using Input and Guess, it does not offer conclusive evidence that bonding is the sole reason for discrepancy between MOFid and CrystalNets, likely the subtle differences between algorithms within the software are also a contributing factor. OpenBabel uses chemical rules and heuristics to infer bonding, relying on predefined chemical knowledge, whereas CrystalNets can be tailored to focus on topology and connectivity and may disregard bond orders or chemical valence rules, as a result the former may struggle with unconventional bonding scenarios that are not well represented in its database whereas the latter is designed for use in creating a simplified network representation to be used for topological characterisation. An example of a structure that is given different topologies is CSD BUSSOZ which was assigned **tfi** using MOFid and **fes** using CrystalNets.

RCSR Topologies and Diversity

In total, the RCSR contains 3,132 unique nets (excluding 0-D and 1-D nets), of which 200 are 2D with the remainder being assigned to 3D structures, as of May 2024. Additionally, it partially overlaps with the 14,532 3D nets of the EPINET database. We analysed the results for the number of unique RCSR topologies found within the dataset.

For 2D MOFs, MOFid identified 49 unique nets, CrystalNets (Guess) identified 63 unique nets, and CrystalNets (Input) identified 59. These figures differ from the unique outputs stated in Table 3.1 due to the inclusion of only RCSR outputs and not just unique outputs. Similarly, for 3D MOFs, MOFid identified 257 unique nets, CrystalNets (Guess) identified 385, and CrystalNets (Input) identified 388. These figures do not include the misidentified dimensionality nets as stated in the section above.

Further, in terms of the mismatched nets, in the 2D run, MOFid identified 12 3D nets, CrystalNets (Guess) identified 41 3D nets, and CrystalNets (Input) identified 2 3D nets in keeping with the 2 misidentified structures. Similarly, for mismatched nets in the 3D space, MOFid identified 38 unique nets that were 2D in nature, CrystalNets (Guess) identified 8 2D nets, and CrystalNets (Input) identified a total of 9 unique 2D nets within the set. Perhaps also interestingly, for MOFid 76 nets appeared only once, whereas for CrystalNets (Guess) we saw 97 nets just once, and 100 unique nets occurred only once for CrystalNets (Input).

Whilst there is the potential for further discussion of these occurrences to take place, we do not feel they would add anything of significant value to the discussion other than becoming an interesting anecdote to the more pressing research questions we try to address, i.e why are these discrepancies occurring, and how can they be solved?

Fingerprints

It is worth also mentioning a unique feature of CrystalNets, which is that when analysis of an unidentifiable topology is performed, the software produces a 'fingerprint' consisting of a string of numbers that represents the topological configuration of the structure due to failure to match it to the database. These fingerprints can be used to match topologically similar structures that do not currently have entries in any nets database for their configurations. Analysis of these fingerprints was not performed in this study, however future work could include further analysis to be performed on this information to discover new unique nets, or match several unknown structures together for further investigation. One of the main reasons behind this is their current lack of recognition by the IUPAC committee, they are not a recommended format through which to report structures, and we have seldom, if at all, seen any new structures being reported with topology fingerprints thus far. This is not to say that it would not be a recognised and adopted feature in the future but at present it is not common practice.

3.6 Implementation of CrystalNets within CSD Mercury

As a notable part of this study we demonstrated that there is a preference towards the implementation of a CrystalNets tool within CSD Mercury over MOFid, both in part due to the computational advantage of using a Julia package when compared to Python, but also due to the increased reliability of results for experimental structures analysed from their respective CIFs. As a result of this study we developed a Python script that

enables a user to call CrystalNets.jl whilst working within both CSD Mercury, or in the CSD Python API. This script began as a standalone topology analyser but we decided to import some additional functionality from other existing Python API scripts to avoid the need to duplicate existing workflows, but also to prevent users interested in additional structure properties from having to duplicate analysis using two or more different scripts on the same materials.

The ease of implementing the CrystalNets package compared with other available software such as ToposPro, or Systre, was a significant factor in its selection. We initially attempted to output files in a format that would be suitable to implement Systre in the analysis (or net-matching) stage, however as a result of the investigations conducted in this article we instead decided to select CrystalNets due to its significant increase in analysis speed and its ease of use. The CSD's Python API can easily be modified to call a Julia package with just a few lines of additional code, and this also makes it easily transferable for use within CSD Mercury where the .py file only needs to be placed within a relevant directory, selected within the software settings, to enable it to produce HTML structure analysis files. Figure 3.3 demonstrates the recommended procedure for using this new structure analysis tool within Mercury, highlighting its ease of use with any selected structure.

To add a script location, simply select CSD Python API in the navigation bar, select Options... and in this menu it is possible to add a directory to a list of locations Mercury can search for python scripts. Here it is also possible to select a new output directory if desired, or change your installed python version. Once set-up in Mercury simply go to your desired structure, then find the CSD Python API tab in the navigation bar again and select mof_crystal_structure_report_v2.py to run the analysis.

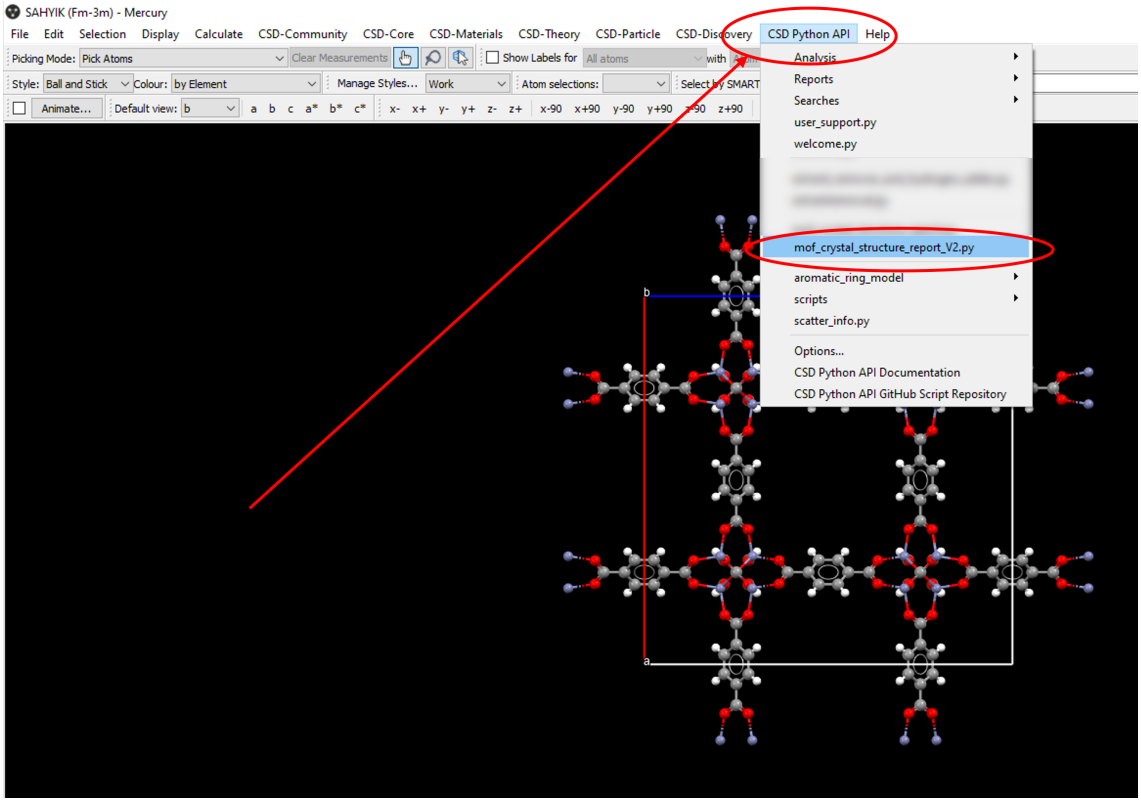


Figure 3.3: *A screenshot demonstrating the implemented CrystalNets process within CSD Mercury*

Whilst implementing the CrystalNets approach to CSD structures we also decided it was necessary to modify the output script to include some other potentially desirable parameters that can be obtained from the CSD, specifically for researchers who are interested in structural features of MOFs or other materials. An example of the HTML output can be seen in Figure 3.4.

This script is a great tool for those with an interest in adsorption applications as a recent update to the CCDC software suite included the PoreBlazer tools [38] an open-source, and open access Fortran 90 code used to determine the pore properties of MOFs, including the calculated void percentage and void volume. Here, the void calculations were performed using a probe radius of 1.2 Å, and a grid spacing of 0.7 Å, but these can be easily changed within the python script to custom values. We decided to package this in the same output with other results such as the space group, cell dimensions, experimental volume, and of course the topology. This is also supplemented by a diagram to show the atomic configuration, as well as the refcode and chemical formula. This workflow, when implemented, should eliminate the need for users to download CIFs from the CSD to determine pore volumes and topological configurations.

MOF Crystal Structure Report for SAHYIK

CCDC
advancing structural science

Crystal Structure Analysis

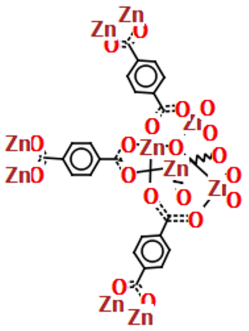


Figure 1. Diagram for SAHYIK

Table 1. Selected Crystal Structure Information

Identifier	SAHYIK
Formula	C24 H12 O13 Zn4
Space Group	Fm-3m
Cell Lengths (Å)	**a** 25.6690 **b** 25.6690 **c** 25.6690
Cell Angles (°)	**α** 90.00 **β** 90.00 **γ** 90.00
Cell Volume (Å)	16913.24
R-Factor	10.99

Volume and Packing Analysis

Table 2. Crystal Packing Information

Estimated volume from 18 Å^3 rule	11448.0
Experimental volume	16913.241
Packing coefficient	0.195
CSD average packing coefficient for organometallic molecules	0.67(5)
Calculated void percentage*	78.49683138214914
Calculated void volume*	13276.358656258448

Topology Information

MOF File: AllNodes, SingleNodes: pcu

Appendix

* **Void calculations** performed with a probe radius of 1.2 Å and a grid spacing of 0.7 Å. [return]

Figure 3.4: *An output file showing the result of crystal structure analysis, including the CrystalNets process, on SAHYIK launched within CSD Mercury, determining the structure has the* ***pcu*** *topology.*

3.7 Conclusion

This study investigated and compared the use of different approaches when considering high-throughput topological characterisation of 2D and 3D MOFs in the CSD MOF subset. We tested three approaches to determine the highest recall, and highest precision method so as to select one as the most appropriate workflow for integration within our new Python approach to topological assignment of CSD materials. We determined that the retention of bonding information, and the use of CrystalNets combined to provide the highest return rate, whilst also not compromising on computation expense (the computational cost of analysis in this case was almost negligible when compared with the MOFid approach). With a success rate of 75.9% on 2D MOFs, and 51.25% on 3D MOFs using an entirely automated workflow there is certainly some scope for improvement. Lastly, we note that this analysis has not considered the presence of disorder in structures or the methods through which modifications were made due to the presence of solvents within the CSD's MOF subset CIFs.

We showed in this article that it is a simple process to integrate new tools within the CSD by a user for their own use, and we make this workflow publicly available hosted in the CCDC's open source GitHub folder at `github.com/ccdc-opensource`. The modification of the CrystalNets output script includes additional parameters catering to researchers interested in structural features of MOFs or other materials, such as pore properties determined using the PoreBlazer tools, space group, cell dimensions, experimental volume, topology, and chemical formula, thereby eliminating the need for users to download CIFs from the CSD for pore volume and topology determination.

References

[1] Jack Gonzalez, Krishnendu Mukherjee, and Yamil J. Colón. Understanding Structure–Property Relationships of MOFs for Gas Sensing through Henry's Constants. *J. Chem. Eng. Data*, 68(1):291–302, January 2023. Publisher: American Chemical Society.

[2] Lauren E. Kreno, Kirsty Leong, Omar K. Farha, Mark Allendorf, Richard P. Van Duyne, and Joseph T. Hupp. Metal–Organic Framework Materials as Chemical Sensors. *Chem. Rev.*, 112(2):1105–1125, February 2012. Publisher: American Chemical Society.

[3] Young-Moo Jo, Yong Kun Jo, Jong-Heun Lee, Ho Won Jang, In-Sung Hwang, and Do Joon Yoo. MOF-Based Chemiresistive Gas Sensors: Toward New Functionalities. *Advanced Materials*, n/a(n/a):2206842, August 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adma.202206842.

[4] Omar K. Farha, A. Özgür Yazaydın, Ibrahim Eryazici, Christos D. Malliakas, Brad G. Hauser, Mercouri G. Kanatzidis, SonBinh T. Nguyen, Randall Q. Snurr, and Joseph T. Hupp. De novo synthesis of a metal-organic framework material featuring ultrahigh surface area and gas storage capacities. *Nat Chem*, 2(11):944–948, November 2010.

[5] Shengqian Ma and Hong-Cai Zhou. Gas storage in porous metal–organic frameworks for clean energy applications. *Chem. Commun.*, 46(1):44–53, January 2010. Publisher: The Royal Society of Chemistry.

[6] Bin Li, Hui-Min Wen, Wei Zhou, Jeff Q. Xu, and Banglin Chen. Porous Metal-Organic Frameworks: Promising Materials for Methane Storage. *Chem*, 1(4):557–580, October 2016.

[7] Jian-Rong Li, Julian Sculley, and Hong-Cai Zhou. Metal–Organic Frameworks for Separations. *Chem. Rev.*, 112(2):869–932, February 2012. Publisher: American Chemical Society.

[8] Qihui Qian, Patrick A. Asinger, Moon Joo Lee, Gang Han, Katherine Mizrahi Rodriguez, Sharon Lin, Francesco M. Benedetti, Albert X. Wu, Won Seok Chi, and Zachary P. Smith. MOF-Based Membranes for Gas Separations. *Chem. Rev.*, 120(16):8161–8266, August 2020. Publisher: American Chemical Society.

[9] Rui-Biao Lin, Shengchang Xiang, Wei Zhou, and Banglin Chen. Microporous Metal-Organic Framework Materials for Gas Separation. *Chem*, 6(2):337–363, February 2020.

[10] Vlad Pascanu, Greco González Miera, A. Ken Inge, and Belén Martín-Matute. Metal–Organic Frameworks as Catalysts for Organic Synthesis: A Critical Perspective. *J. Am. Chem. Soc.*, 141(18):7223–7234, May 2019. Publisher: American Chemical Society.

[11] Yu Shen, Ting Pan, Liu Wang, Zhen Ren, Weina Zhang, and Fengwei Huo. Programmable Logic in Metal–Organic Frameworks for Catalysis. *Advanced Materials*, 33(46):2007442, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adma.202007442.

[12] Michelle H. Teplensky, Marcus Fantham, Peng Li, Timothy C. Wang, Joshua P. Mehta, Laurence J. Young, Peyman Z. Moghadam, Joseph T. Hupp, Omar K. Farha, Clemens F. Kaminski, and David Fairen-Jimenez. Temperature Treatment of Highly Porous Zirconium-Containing Metal–Organic Frameworks Extends Drug Delivery Release. *J. Am. Chem. Soc.*, 139(22):7522–7532, June 2017. Publisher: American Chemical Society.

[13] Harrison D. Lawson, S. Patrick Walton, and Christina Chan. Metal–Organic Frameworks for Drug Delivery: A Design Perspective. *ACS Appl. Mater. Interfaces*, 13(6):7004–7020, February 2021. Publisher: American Chemical Society.

[14] Vladislav A. Blatov, Alexander P. Shevchenko, and Davide M. Proserpio. Applied Topological Analysis of Crystal Structures with the Program Package ToposPro. *Crystal Growth & Design*, 14(7):3576–3586, July 2014. Publisher: American Chemical Society.

[15] O. Delgado-Friedrichs and M. O'Keeffe. Identification of and symmetry computation for crystal nets. *Acta Cryst A*, 59(4):351–360, July 2003. Number: 4 Publisher: International Union of Crystallography.

[16] Benjamin J. Bucior, Andrew S. Rosen, Maciej Haranczyk, Zhenpeng Yao, Michael E. Ziebel, Omar K. Farha, Joseph T. Hupp, J. Ilja Siepmann, Alán Aspuru-Guzik, and Randall Q. Snurr. Identification Schemes for Metal–Organic Frameworks To Enable Rapid Search and Cheminformatics Analysis. *Crystal Growth & Design*, 19(11):6682–6697, November 2019. Publisher: American Chemical Society.

[17] Lionel Zoubritzky and François-Xavier Coudert. CrystalNets.jl: Identification of Crystal Topologies. *SciPost Chem.*, 1(2):005, June 2022.

[18] Michael O'Keeffe, Maxim A. Peskov, Stuart J. Ramsden, and Omar M. Yaghi. The Reticular Chemistry Structure Resource (RCSR) Database of, and Symbols for, Crystal Nets. *Acc. Chem. Res.*, 41(12):1782–1789, December 2008. Publisher: American Chemical Society.

[19] V. A. Blatov, O. Delgado-Friedrichs, M. O'Keeffe, and D. M. Proserpio. Three-periodic nets and tilings: natural tilings for nets. *Acta Cryst A*, 63(5):418–425, September 2007. Number: 5 Publisher: International Union of Crystallography.

[20] V. A. Blatov and D. M. Proserpio. Topological relations between three-periodic nets. II. Binodal nets. *Acta Cryst A*, 65(3):202–212, May 2009. Number: 3 Publisher: International Union of Crystallography.

[21] S. J. Ramsden, V. Robins, and S. T. Hyde. Three-dimensional Euclidean nets from two-dimensional hyperbolic tilings: kaleidoscopic examples. *Acta Cryst A*, 65(2):81–108, March 2009. Number: 2 Publisher: International Union of Crystallography.

[22] Stuart R. Batten, Neil R. Champness, Xiao-Ming Chen, Javier Garcia-Martinez, Susumu Kitagawa, Lars Öhrström, Michael O'Keeffe, Myunghyun Paik Suh, and Jan Reedijk. Terminology of metal–organic frameworks and coordination polymers (IUPAC Recommendations 2013). *Pure and Applied Chemistry*, 85(8):1715–1724, July 2013. Publisher: De Gruyter Section: Pure and Applied Chemistry.

[23] Alexander P. Shevchenko, Eugeny V. Alexandrov, Andrey A. Golov, Olga A. Blatova, Alexandra S. Duyunova, and Vladislav A. Blatov. Topology versus porosity: what can reticular chemistry tell us about free space in metal–organic frameworks? *Chem. Commun.*, 56(67):9616–9619, August 2020. Publisher: The Royal Society of Chemistry.

[24] Peyman Z. Moghadam, Sven M. J. Rogge, Aurelia Li, Chun-Man Chow, Jelle Wieme, Noushin Moharrami, Marta Aragones-Anglada, Gareth Conduit, Diego A. Gomez-Gualdron, Veronique Van Speybroeck, and David Fairen-Jimenez. Structure-Mechanical Stability Relations of Metal-Organic Frameworks via Machine Learning. *Matter*, 1(1):219–234, July 2019.

[25] Yamil J. Colón, Diego A. Gómez-Gualdrón, and Randall Q. Snurr. Topologically Guided, Automated Construction of Metal–Organic Frameworks and Their Evaluation for Energy-Related Applications. *Crystal Growth & Design*, 17(11):5801–5810, November 2017. Publisher: American Chemical Society.

[26] Matthew A. Addicoat, Damien E. Coupry, and Thomas Heine. AuToGraFS: Automatic Topological Generator for Framework Structures. *J. Phys. Chem. A*, 118(40):9607–9614, October 2014. Publisher: American Chemical Society.

[27] Sauradeep Majumdar, Seyed Mohamad Moosavi, Kevin Maik Jablonka, Daniele Ongari, and Berend Smit. Diversifying Databases of Metal Organic Frameworks for High-Throughput Computational Screening. *ACS Appl. Mater. Interfaces*, 13(51):61004–61014, December 2021. Publisher: American Chemical Society.

[28] Aditya Nandy, Shuwen Yue, Changhwan Oh, Chenru Duan, Gianmarco G. Terrones, Yongchul G. Chung, and Heather J. Kulik. A database of ultrastable MOFs reassembled from stable fragments with machine learning models. *Matter*, 6(5):1585–1603, May 2023.

[29] Peyman Z. Moghadam, Aurelia Li, Seth B. Wiggin, Andi Tao, Andrew G. P. Maloney, Peter A. Wood, Suzanna C. Ward, and David Fairen-Jimenez. Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future. *Chem. Mater.*, 29(7):2618–2625, April 2017. Publisher: American Chemical Society.

[30] Yongchul G. Chung, Jeffrey Camp, Maciej Haranczyk, Benjamin J. Sikora, Wojciech Bury, Vaiva Krungleviciute, Taner Yildirim, Omar K. Farha, David S. Sholl, and Randall Q. Snurr. Computation-Ready, Experimental Metal–Organic Frameworks: A Tool To Enable High-Throughput Screening of Nanoporous Crystals. *Chem. Mater.*, 26(21):6185–6192, November 2014. Publisher: American Chemical Society.

[31] Yongchul G. Chung, Emmanuel Haldoupis, Benjamin J. Bucior, Maciej Haranczyk, Seulchan Lee, Hongda Zhang, Konstantinos D. Vogiatzis, Marija Milisavljevic, Sanliang Ling, Jeffrey S. Camp, Ben Slater, J. Ilja Siepmann, David S. Sholl, and Randall Q. Snurr. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data*, 64(12):5985–5998, December 2019. Publisher: American Chemical Society.

[32] Yongchul G. Chung, Emmanuel Haldoupis, Benjamin J. Bucior, Maciej Haranczyk, Seulchan Lee, Konstantinos D. Vogiatzis, Sanliang Ling, Marija Milisavljevic,

Hongda Zhang, Jeff S. Camp, Ben Slater, J. Ilja Siepmann, David S. Sholl, and Randall Q. Snurr. Computation-Ready Experimental Metal-Organic Framework (CoRE MOF) 2019 Dataset, February 2020.

[33] Andrew Rosen, Shaelyn Iyer, Debmalya Ray, Zhenpeng Yao, Alan Aspuru-Guzik, Laura Gagliardi, Justin Notestein, and Randall Q. Snurr. Machine Learning the Quantum-Chemical Properties of Metal–Organic Frameworks for Accelerated Materials Discovery with a New Electronic Structure Database. *Matter*, 4(5):1578–1597, October 2020. Publisher: ChemRxiv.

[34] Lawson T. Glasby, Joan L. Cordiner, Jason C. Cole, and Peyman Z. Moghadam. Topological Characterisation of Metal-Organic Frameworks: A Perspective. *In Review*, 2024.

[35] Omar M. Yaghi, Michael O'Keeffe, Nathan W. Ockwig, Hee K. Chae, Mohamed Eddaoudi, and Jaheon Kim. Reticular synthesis and the design of new materials. *Nature*, 423(6941):705–714, June 2003. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6941 Primary_atype: Reviews Publisher: Nature Publishing Group.

[36] Lawson T. Glasby, Kristian Gubsch, Rosalee Bence, Rama Oktavian, Kesler Isoko, Seyed Mohamad Moosavi, Joan L. Cordiner, Jason C. Cole, and Peyman Z. Moghadam. DigiMOF: A Database of Metal–Organic Framework Synthesis Information Generated via Text Mining. *Chem. Mater.*, 35(11):4510–4524, June 2023. Publisher: American Chemical Society.

[37] Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, October 2011.

[38] Lev Sarkisov, Rocio Bueno-Perez, Mythili Sutharson, and David Fairen-Jimenez. Materials Informatics with PoreBlazer v4.0 and the CSD MOF Database. *Chem. Mater.*, 32(23):9849–9867, December 2020. Publisher: American Chemical Society.

Chapter 4

Machine Learning and Digital Manufacturing Approaches for Solid-state Materials Development

4.1 Publication Information and Paper Contributions

This work has been published as Chapter 14 of the book AI-Guided Design and Property Prediction for Zeolites and Nanoporous Materials published by John Wiley & Sons Ltd, and edited by German Sastre and Frits Daeyaert.

In this publication I, the candidate, wrote the chapter with contributions from Emily H. Whaites under the supervision of Dr Peyman Z. Moghadam.

4.2 Abstract

Novel solid materials are urgently needed for energy applications including carbon capture and energy storage. The increasing breadth and availability of large and complex databases of such materials have made their design space too vast, and clearly, relying on trial-and error and serendipity to accelerate their discovery is costly, slow, and unreliable. The synthesis of solid materials has been largely a manual process and, due to the intrinsically high dimensionality of the experiments, such processes are inefficient in the exploration of realisable materials. To tackle these pressing challenges, there is a clear need for the deployment of new technologies to speed up the way solid materials are developed in a repeatable, reproducible, and traceable manner.

The emerging introduction of state-of-the-art data science and digital technologies will help chemists to work more efficiently and advance materials discovery. A key challenge in the adoption of such technologies is to overcome numerous "data-poor" situations. Furthermore, the integration between pre-existing physico-chemical knowledge and synthesis is recognised as critical to success given the complex chemistry that characterises many classes of solid materials. The concept of digital manufacturing itself is developing as a highly sought after technique in both academia and industry, from the initial stages of materials discovery to the final stages of material synthesis and performance analysis.

This chapter focuses on the challenges faced by researchers in novel materials discovery with a particular focus on the collection and the use of data to drive digital synthesis. We

specifically discuss databases of nanoporous materials synthesis data and how they can be explored via text mining, machine learning and artificial intelligence, as well as robotics to develop self-adaptive systems to automate and control the chemistry of solid structures. We also investigate the topic of natural language processing within text mining, and a variety of machine learning models are discussed with a look at some of the most recent models that have been used to enable the prediction of synthesis conditions.

Finally, we discuss digitisation of production and the recent developments that aim to shift the current practices within synthesis systems. Advances have been made in the conversion of material synthesis to flow chemistry, enabling the application of robotic systems for continual improvement. After a thorough dig into the work currently published in digital MOF synthesis and adjacent fields, the importance of sufficient, high quality, unbiased data is highlighted as a key factor for the next stages of development, forming the basis for future work regarding synthesis and digitisation of MOFs.

4.2.1 Keywords

Solid State Materials, Data Science, Digital Technologies, Materials Discovery, Synthesis

4.3 Introduction

Solid state chemistry, often referred to as materials chemistry, is a field of chemistry concerned with studying the synthesis, structure, and properties of materials in the solid phase. These solids are often classified as crystalline, amorphous, organic, inorganic, or nano-materials depending on the type, and the arrangements of their constituent atoms. Some notable examples include zeolites, covalent organic frameworks (COFs), metal organic cages (MOCs), and metal organic nano-sheets (MONs).

One intensely studied class of solid state materials, and the primary example used throughout this chapter, are metal-organic frameworks (MOFs), crystalline structures synthesised from organic and inorganic building blocks to form an extended framework material. The building-block approach creates the opportunity for the synthesis of tens of thousands of combinations where they can be tailored to achieve particular properties for a multitude of applications, and since the start of the 1990s, thousands of MOF materials have been synthesized at laboratory scales [1, 2, 3, 4, 5, 6]. However, despite their great promise for a wide range of applications, only a handful have been successfully commercialised [7].

In general, the production of MOFs is largely a manual process and because of the complex multi-dimensional nature of their synthesis, the development process can be time-consuming and inefficient when exploring the entire MOF synthesis space. To tackle these challenges, there is a clear need for the adoption of technologies that can expedite the way MOF materials are designed and developed with optimum properties. One way to address these challenges involves the deployment of state-of-the-art computer simulations and digital technologies. This approach includes a wide range of techniques from database generation, to high-throughput screening, machine learning (ML), and the use of novel digitalisation tools to overcome “data-poor” processes that characterise the complex chemistry of MOFs. Figure 4.1 demonstrates gas adsorption applications in MOFs.

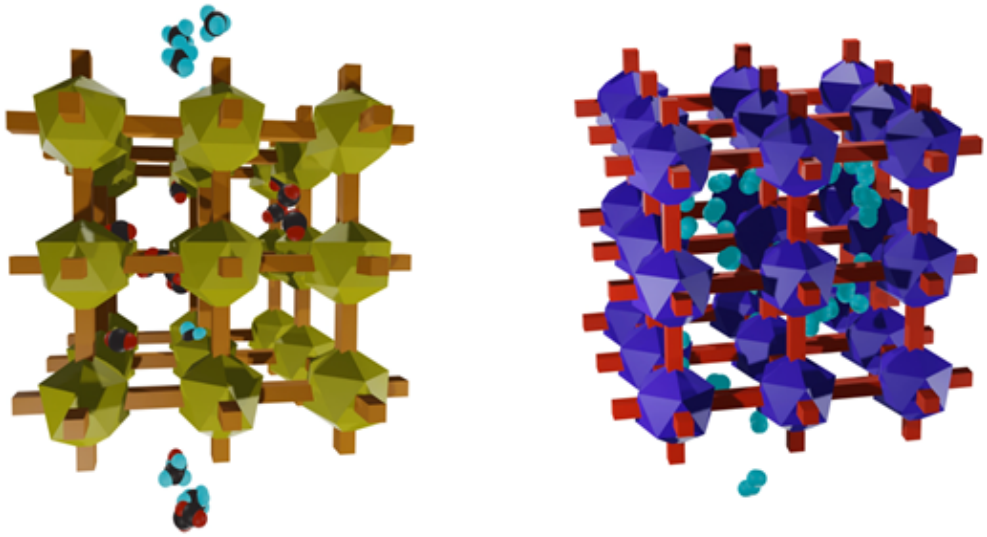


Figure 4.1: *Schematic showing the applications of gas separation where CO_2 is captured and methane is separated (left), and the storage of gaseous methane (right) in MOFs. The structures here are represented in a general form where a typical metal-oxo cluster is seen as a metal node, and the organic linkers are drawn as straight connecting bars.*

To fully understand the current size of the MOF space, Moghadam et al. (2017) reported nearly 70,000 experimentally synthesised MOF structures in the Cambridge Structural Database (CSD) upon the creation of the CSD MOF subset [8]. Moving forward to the present day, based on data published in the latest CSD release (version 5.3, 2021), there are now 114,373 experimentally synthesised MOFs. There has been a rapid increase in the quantity of submitted MOF structures within the CSD with an estimated 10% of all deposited structures meeting the criteria of a MOF, and an incredible 15,998 experimentally synthesised materials were deposited and approved last quarter (June 2022). This large, rapidly expanding chemical space continually adds to the difficulty faced by chemists in identifying useful MOFs for their chosen applications. Many research groups within the materials field have begun to quickly adopt ML techniques in the quest to enable rapid and reliable materials discovery and synthesis. The development of ML tools should be seen as a means to ease the process at every stage, from the identification of useful MOFs, to reducing the cost of synthesis for novel structures.

To maximise the potential for MOF discovery, it is essential to develop robust data collection techniques to obtain crucial information from the large volume of published work on MOFs and their properties. Moghadam et al. (2017) used keyword searching to produce the CSD MOF subset, an automatically updating database, with specific criteria for material selection from the Cambridge Crystallographic Data Centre's 1,000,000+ structures [8]. This database, alongside the Computation-Ready Experimental (CoRE) MOF database, developed by Chung et al. (2014) [9], are two of the largest and highly curated collections of experimental MOF materials, and both allow for high-throughput computational screening of MOFs for a wide range of applications. Due to the nature of these porous structures, particularly those that are constructed in 3D, there are many applications for in gas storage and separation but many other studies have focused on sensing, catalysis, drug delivery and chemical removal [10, 11, 12, 13, 14, 15].

Often, large-scale high-throughput screening is not thorough enough to filter large databases into small subsets of interesting structures without significant input from more computationally expensive studies. The development of ML tools for property prediction has commonly been touted as the next step towards a rapid and inexpensive computational approach. ML usage has seen a significant increase in all fields of scientific discovery and data development over recent years, in part due to the popularity of user-friendly tool kits such as RDKit [16]. ML models can be created and developed to make highly accurate predictions, although the success of such models is dependent on a sufficiently large amount of data taken from materials databases for use as training sets.

Structure-property data is often used as a foundation for investigating the synthesis requirements of novel structures. One of the most promising applications of new ML approaches is to predict optimal reaction synthesis conditions for stable structures with sufficient crystallinity. This approach however is not limited to MOFs and can, and has been applied to a range of different materials, examples of which are discussed in this chapter.

Typically, the time required to obtain optimal MOF synthesis is long and expensive due to the highly diverse synthesis conditions required in the creation of these materials. This diversity means there is a lack of general synthesis recipes, and for many cases predicted synthesis conditions are non-transferable. Key parameters for synthesis include solvents, reaction temperature and time [17]. Due to the novelty of using ML for synthesis of MOFs, only a handful of studies have been reported to date. The use of Natural Language Processing (NLP), a sub field of ML and artificial intelligence (AI), has been investigated to augment chemists' expertise when approaching experimental design. Luo et al. (2022) [18] and Zhang et al. (2020) [19] have utilised NLP to study the MOF databases, CoRE MOF and CSD MOF subset, respectively in an effort to train models that are able to predict the optimal synthesis conditions for new MOF structures.

If we consider the use of a digital manufacturing approach when investigating MOF synthesis, a considerable bottleneck in synthesis improvement is the availability of synthesis data. Digital manufacturing techniques, such as digital twins, rely on the constant availability of high quality data to feedback to the synthesis process. A digital twin is a virtual representation of a physical object or process, in this case a synthesis process, used for the continual improvement of engineering activities. When developing an initial synthesis path to create for example, a MOF in a laboratory, a similar ML approach can be used despite the requirement for greater flexibility in experimental design. By implementing ML models for small quantity experiments to predict synthesis conditions, the number of real experimental tests can be reduced as only the most viable reactions are chosen for experimental trial, and although this process is perhaps more suited to the improvement of larger scale materials synthesis, it is still applicable at lab scale.

A shift from batch production to a closed loop process is a significant driver towards the introduction of digital manufacturing. This approach, when combined with ML prediction techniques described previously, can be used to redevelop existing larger scale synthesis techniques. By increasing the available information about the synthesis of a particular material, including accurate ML predicted data such as approximate cost, reaction time, and required reagents, the implementation of a digital twin becomes easier. Considering the savings made due to the reduction in labour and time when attempting to optimise the synthesis process, focus can be directed towards the development of other structures and their use in desired applications. The use of ML should be seen as a means to achieve

the goal of automated synthesis, and as a tool to drive continuous improvement.

Improving the manufacturing pathway of novel or complex materials is often seen as a large hurdle for many chemists. Accessing the relevant data and hosting the technology for synthesis development can be expensive and time consuming. Digital manufacturing opens the doors for significant changes in the development of material production. With digital infrastructure such as synthesis servers and synthesis databases, high cost equipment can be accessed remotely, making the research process much less cost intensive, expanding the research field further to groups without a wide array of computational or experimental resources. New tools and databases designed for digitisation can increase collaboration of computational researchers and experimentalists to aid in their development of novel material manufacturing techniques for new and existing solid state materials.

4.4 The Development of MOF Databases

The progress of ML would begin to stagnate without the access to sufficient training data. However, the field of solid state chemistry sees thousands of materials synthesised each year, with tens of thousands published in the past decade. However, due to the substantial amount of text accompanying the interesting data in each subsequent publication, obtaining large and reliable databases of ML training material has the potential to become a significant bottleneck to development without continuous addition of new information. The number of MOFs in the CSD MOF subset alone has increased by 30,000 structures since 2017, and additional publications with information about synthesis of new structures are published regularly [8]. The latest CSD update contained over 16,000 additions to the database with an estimated 10% of structures being MOF or MOF like. To allow for chemists to efficiently process information, data must be presented and submitted in a fashion that is accurate, organised, and machine readable.

In 2014, Chung et al., as part of the Materials Genome Initiative, developed the first MOF database CoRE MOF, containing 4,700 porous structures [9]. The data originated from the CSD, and included some additional properties to ensure these entries were suitable for molecular simulations. The conditions for entry into the database included only 3D structures with pore sizes larger than 2.4 Å. These criteria were chosen specifically to allow for the screening of MOF structures for use in gas storage, separation, and catalysis. In 2019, the CoRE MOF database was manually updated to include over 14,000 porous 3D structures that have been reported in published literature [20]. Value was also added to the database with the introduction of reconstructed disordered structures, and new pore analytics and physical property data.

Following the release of the CoRE MOF database, the CSD MOF subset was developed in 2017. Moghadam et al. expanded the criteria used to identify MOFs by allowing for the inclusion of materials of other dimensionalities and pore sizes [8]. One notable advantage of the CSD MOF subset over the CoRE MOF database is that the CSD MOF subset has been designed to be automatically updated, adding new materials quarterly to ensure all structural data is current. The CSD MOF subset includes 1D, 2D and 3D structures, and can be organised into these specific categories at the click of a button. The CSD MOF subset is primarily accessed as part of the CCDC's software suite and is compatible with their structure search tool ConQuest, and the CSD PythonAPI.

The CSD MOF subset has been used in a significant number of studies, and has

formed the primary data source for the development of DigiMOF, a text-mined structure information database created by Glasby et al. (2023) [21]. The aim of the DigiMOF database is to provide MOF synthesis data alongside other key properties which are lacking within previous databases. Synthesis data, when collated within a database, will allow for synthesis routes to be compared and assessed, leading to the minimisation of failed or inefficient experiments, and increased likelihood of scalability and profitability when compared with current trial and error synthesis development. Due to the diverse range of applications for MOF structures, several other significant contributions have been developed and published which focus on specific applications due to certain properties. Notable examples include the CoRE MOF 2014-DDEC database by Nazarian et al. (2016) [22] in which DFT-derived partial atomic charges were determined, and the quantum (QMOF) database by Rosen et al. (2022) [23]. The QMOF database enables the searching of structures based on properties such as charges, bond orders, or band gaps for use in photo-catalysis, and other similar applications. As part of their works carried out with the use of ML for MOF synthesis, Luo et al. (2022) [18] and Nandy et al. (2022) [24] have also created small publicly accessible databases to investigate synthesis conditions of MOF structures and solvent extraction data, respectively. Aside from databases containing 'already-synthesized' structures, there are several examples of collections of hypothetical structures such as the hMOF database by Wilmer et al. (2012) [25] containing 137,953 hypothetical MOFs, the 13,512 MOF structures created in specific topologies using a material generation algorithm called ToBaCCo by Colón et al. (2017) [26], and more recently by Majumdar et al. (2021) [27] with ca. 20,000 hypothetical MOFs designed specifically for a diverse chemical design space. The use of ML in synthesis prediction is a step towards finding feasible pathways for the eventual creation of many of these hypothesised materials without the need for expensive trial and error experimentation.

4.5 Natural Language Processing

One of the biggest limitations of ML is the requirement for data availability, accurate deep generative models typically require training data sets with a size of the order of 10^6 [28]. In terms of MOF structures, properties, and applications, an abundance of data is collated within currently available databases, however they rarely contain information regarding the synthesis conditions and parameters required to prepare them. Alternatives to creating new or real-time data, include NLP algorithms that are frequently being used to extract synthesis data from published scientific literature. NLP can be structured into a series of four steps: article retrieval, conversion and paragraph classification, word tokenisation and, extraction and manual verification.

Datasets for ML training can be generated through NLP of new experimentation or by accessing existing data held within experimental logbooks. One example where NLP would have been useful is Xie et al. (2020) [29], who collected the synthesis parameters of 486 reactions from archived experimental notebooks of both successful and failed experiments. This data was used to train models for synthesis condition prediction of metal-organic nanocapsules (MONCs). An eXtreme gradient boost (XGBoost) algorithm topped the table of prediction accuracy, using 17 descriptors, at 91% with solvents, modulators (molar mass and mole), and cations emerging as the dominant factors in the formation of single-crystal MONCs.

Additionally, it is worth noting here that a recent perspective contribution by Jablonka et al. (2022) [30] has sought to address the issue of inaccessible, non-digital, and non-

reported experimental practices. As paper based lab records are often still the norm in many institutions, the authors suggested that the development of a modular open science platform would benefit not only the data mining studies highlighted in this chapter, but also beyond that. In recent years the introduction of electronic lab notebooks (ELNs) aimed to address these concerns surrounding data management, and increase the reusability of experimentally gathered data. However, although the argument that the technology is already available to begin the collection of this work, the adoption of these ELNs must be suitable for the synthetic work performed by chemists and materials scientists, and developed as an easily accessible and open source resource that demonstrates fair principles and practices. Only, once an acceptable and agreed format is chosen, would such a tool become useful for the gathering of data for the application of ML.

The use of NLP can be found throughout the field, with a heavy focus on using ML for the prediction of synthesis parameters. Kim et al. (2017) [31] used a cross-reference application programming interface (API) to retrieve a list of articles focused on the synthesis of titanium nanotubes. From approximately 100 different journal articles, several hundred paragraphs were manually labelled as either synthesis paragraphs, or other, before being fed into a logistic regression classifier. Paragraphs underwent a word embedding approach such that they could be represented as real-value vectors, followed by binary labels, with 1 indicating synthesis information present, and 0 as unrelated. The post logistic regression data had an overall accuracy of 95% on unseen test data.

The next stage saw relevant synthesis paragraphs undergo transformation into dependency parse trees using ChemDataExtractor and SpaCy parsers [32]. Word tokenisation and speech tagging are performed to split sentences into constituent words, and grammatical labels are added to each word token. Synthesis verbs of interest are detected by a neural network (NN) approach upon the traversal of these dependency parse trees, these are then iterated along to find operating parameters. Nouns are then scanned and matched against the PubChem database and validated against the ChemDataExtractor model to confirm meaningfulness [33]. Trained on 5000 human annotated words, this NN approach yielded an overall accuracy of 86% as measured against a set of 100 human-annotated synthesis articles.

In recent years, NLP has seen increasing use in the field of MOF synthesis data extraction. Luo et al. (2022) [18] used NLP to extract data for use in predicting MOF synthesis conditions; with the structure selection based on those found exclusively in the CoRE MOF database. The NLP successfully extracted synthesis conditions for 983 MOF structures. The parameters have been collected and made available within an open source synthesis database, however due to the lack of standardisation when reporting reaction conditions, some key parameters are missed due to ambiguity. Therefore, accurate and reliable NLP currently requires parallel manual extraction. A total of six relevant parameters were extracted from the CoRE MOF into the new automatically created SynMOF-A database, alongside metal and linker information taken directly from the associated crystallographic information files (CIFs) of each material. Additional manual versions, SynMOF-M and SynMOF-ME were also created to ensure accurate data sources before being used to train ML models to discover similarity patterns in the synthesis conditions.

Park et al. (2022) [34] extracted data for 46,071 MOF synthesis reactions from 28,565 papers using a newly developed data extraction code. This study categorised paragraphs in papers much in the same way as Kim et al. (2017) [31] with a binary approach to synthesis information, but with a notably larger test set of 180 papers. However,

this study extracted synthesis conditions from solid state materials found only within the CSD MOF subset database. The NLP method here used named entity recognition (NER) to extract chemical names and then categorised them using neural networks, a 100-dimensional bi-directional-LSTM which is able to consider the forward and backwards context, alongside a conditional random field (CRF) layer used to predict each label of sequence data. Here the NLP was used to extract MOF names, precursor, and solvents, with a high precision (98%) but significantly varying recall. The data collected by Park et al. (2022) [34] focused on single-step reactions as their NLP algorithm was unable to differentiate between multiple steps within MOF synthesis.

After a review of current text-mining and MOF synthesis literature, Glasby et al. (2023) [21] chose to develop parsers that would extract the information on the solvents used, the inorganic and organic precursors, and their synthesis methods. The parser training technique used in this study has been visualised in Figure 4.2.

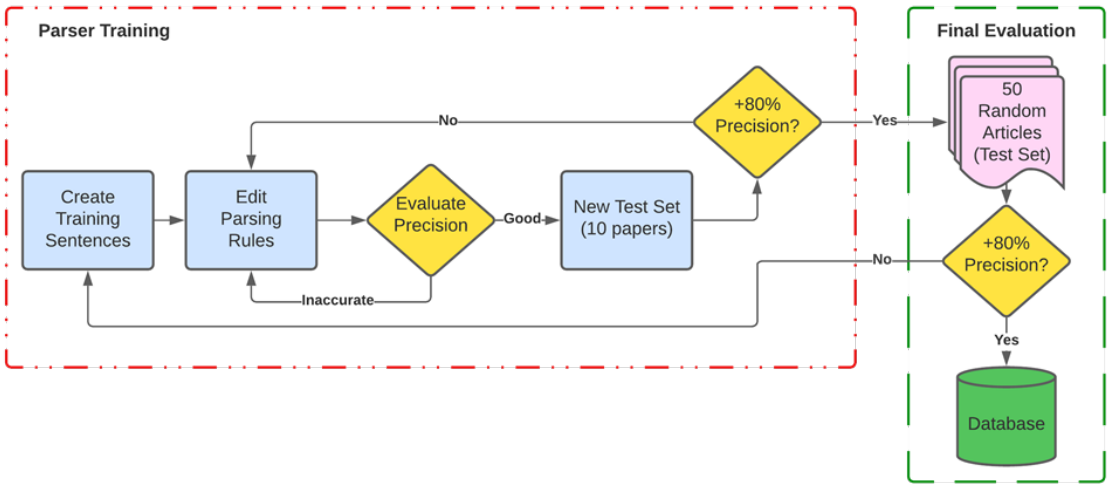


Figure 4.2: *A flow diagram which shows the process of developing suitably precise parsers for data extraction by text mining [21], licenced under CC BY NC 4.0.*

DigiMOF database is the first within the MOF field to utilise the tool ChemDataExtractor, in conjunction with the CSD PythonAPI, to produce a property database using text mining software. MOF topologies were also extracted for further synthesis route analysis, and compared with the building blocks to investigate potential trends. An exclusion list was employed to filter frequently found misidentifications within the published text. The DigiMOF project is made open source to encourage collaboration for further improvements in the gathering of useful synthesis information. The current version of the DigiMOF database consists of 43,281 unique MOF properties with a precision of 77%. This work, amongst other attempts such as that of Luo et al. (2022) [18] are the foundations of the future of digital manufacturing, hoping to provide a searchable database for key MOF properties that can allow for assessment of viable reactions.

Despite these promising studies into the utilisation of NLP to create useful synthesis data sets, a bottleneck associated with extracting data from published journals is the bias towards successful reactions. Without proper weighting or failed reaction data, training sets for ML models will include bias. However, the data extracted in these studies still has multiple uses, from collation into a searchable database for material evaluation, to use as

training data for further development of ML models. The availability of more and more extracted data helps to shift ML development towards overcoming frustrating "data-poor" situations that have previously been a hindrance for the development of digital synthesis.

4.6 An Overview of Machine Learning Models

ML models are typically chosen based on the available data and the desired result. Certain models are able to handle higher levels of complexity and often have variable levels of interpretability. ML tools have become increasingly accessible to the general researcher in most scientific fields, with the emergence of downloadable software packages including Scikit-learn [35] and TensorFlow [36], combined with an abundance of online help, ML training courses, and video guides. These packages provide access to a multitude of models from the simplest to increasingly intricate.

Arguably the simplest model is linear regression, a linear approach used to model the relationship between a scalar response and one variable. Conroy et al. (2022) [37] used linear regression as part of their work to identify key property descriptors required to predict synthesis routes for zeolite LTA synthesis, and predict the quantitative output of synthesis routes. Linear regression works on the basis that the output is linearly relative to the feature inputs, and the predicted outcome is a sum of the weighted features [38]. The assumption of linearity limits both the potential uses and results obtained from this method, although due to the simplicity of the model, the results are often transparent and can be easily interpreted.

Non-linear models include random forest regression, a supervised learning algorithm based on selecting decision trees after being trained on a dataset, and averaging the results. This technique makes predictions based on the outputs of merging multiple decision trees, combining multiple predictions to make a more accurate prediction than a single model. The results of this model are typically easily interpreted, with easy to visualise results due to the simple nature and easy visualisation of a decision tree; a tree with a depth of five, for example, would be easy for most people to follow after a brief explanation. Limitations however, include overfitting and slow prediction speed when using a large number of trees, and the combination with random forest regression reduces the interpretability, particularly as the maximum number of trees can begins to exceed several hundreds.

Jensen et al. (2019) [39] used random forest regression to predict the densities of zeolite based on their synthesis conditions, following the use of NLP and text markup parsing tools to automatically extract information from 70,000 zeolite journal articles. They trained a random forest regression model using sci-kit learn, across 100 decision trees with splits determined by mean squared error. The model was cross validated on syntheses that resulted in a pure phase zeolite, including 898 synthesis routes. Support vector regression, simple neural network, and Gaussian process regression models were also used and compared to random forest, with the random forest model exhibiting the highest accuracy with the added benefit of human interpretability.

Support vector machines (SVM) are a popular example of kernel models, a class of algorithms used for pattern analysis. The general task of kernel models is to compare new data with the data found in training datasets to make predictions. An SVM training algorithm builds a model that assigns examples to one category or another in a non-

probabilistic binary classification. The SVM maps training examples to points in space, creating a gap between two categories, then new examples are mapped into the same space and predicted to belong to either side, depending on which side of the gap they fall.

Raccuglia et al. (2016) [40] employed SVM when exploring chemical space, focusing particularly on inorganic-organic hybrid materials, based on failed experimental data found in archived lab notebooks. The vast majority of unreported failed reactions are archived in notebooks that are typically inaccessible, to guide future efforts towards successful synthesis a web-accessible public database was created to enable initial data entry from existing notebooks and current experimental data. The dataset, 3955 unique and complete reactions, was split into groups, 1/3 as test and 2/3 as training. A single SVM model was used to predict the likelihood of crystallisation based on synthesis parameters, it had an accuracy of 78% in describing all reaction types, and 79% when considering only vanadium-selenite reactions.

When dealing with larger data sets, it is more common to see neural networks being used as the primary ML models. Neural networks are designed to mimic the process through which the human brain operates, with a combination of hidden layers with input and output layers. The node connection between each layer forms a network where each node has an associated weight. The capability of a neural network is highly dependent on the quality and size of the dataset used, but these models can adapt to a changing input so that the network can generate the best possible result without requiring any redefinition of the output criteria.

Park et al. (2022) [34] trained their artificial neural network (ANN) model using a positive-unlabelled learning (PU learning) algorithm. This model was trained to predict the synthesisability of MOFs based on given input synthesis parameters, and was able to differentiate between amorphous and crystalline forms of the same MOF material. This research has been some of the first in the field of MOFs to use 'big data' to achieve meaningful insights into ideal synthesis conditions.

XGBoost is an open source software library composed of gradient boosting ML algorithms. The aim of gradient boosting is to find patterns within the data and make predictions based on these relationships. It gives a prediction model in the form of ensemble learning, where multiple learning algorithms are used to obtain better predictive performance, typically a fixed set of alternative models which allows some flexibility. The gradient boosting prediction model most often consists of decision trees, and where a decision tree is the weak learner, the result is a gradient-boosted trees algorithm which usually outperforms random forest. Whilst this boosting can increase the accuracy of linear regression or a decision tree, it may sacrifice intelligibility and interpretability. To recover performance and interpretability, some model compression techniques exist which allow for the transformation of an XGBoost into a single decision tree that can approximate the original decision function [41].

For prediction of the likelihood of crystallisation of MONCs, Xie et al. (2020) [29] utilised a multi-model method to allow for cross validation and comparison of results. The models included linear regression, Gaussian Naïve Bayes (GNB), k-nearest neighbours (KNN), SVM, decision tree, random forest, XGBoost, and multilayer perceptron (MLP). All of the tested models achieved an accuracy of 82% or higher, and an F1 score which exceeds 81%. The result of training and evaluation of these nine ML models found XGBoost to have the highest accuracy.

It is important to note that all ML models offer individual advantages and limitations, and when choosing a ML model, it is essential to consider the importance of high accuracy and interpretability as well as ease of operation. When choosing a model that best fits the input dataset, the following factors must be taken into consideration. Firstly, the runtime of models such as random tree will become progressively longer as the number of trees increases, and the output will be achieved much more slowly. This can be impractical when using large, and complex datasets. Some models also lack interpretability, making it difficult to know exactly what the results of the study are, and the user may be unable to interpret the factors that have caused the predicted output. In a synthesis context, the relationships within the data provide the essential information required to aid process and material development. Without interpretability of the data's relationships, it becomes progressively difficult, as models increase in complexity, to further understand low accuracies and slow runtimes.

Evaluation is a key stage in ensuring the return of accurate ML model results, and evaluation metrics allow chemists to quantify the performance of models on data that has yet to be seen. Accuracy, the ratio between correct predictions and total predictions, is most useful with balanced data. For data with imbalances, F1 scores are used to evaluate the outcome, combining the precision of the results with recall. Additionally, R^2 represents the proportion of variance from the original data set, and is best used with regression models such as the random forest regression.

Another approach, Bayesian optimisation is becoming increasingly popular in the development of synthesis digitisation [42, 43, 44]. This algorithm allows for continued optimisation of a closed loop process when used in conjunction with automated physical systems, helping to optimise products and synthesis conditions, and increase the feasibility of reactions. Using input data to take an initial guess about a chosen function, it continues to refine this first prediction as data is added with each iteration, and builds a probable model of the objective function that is being explored. It is composed of two models; a surrogate model and an acquisition function. The surrogate model defines the probability distributions over the chosen function, built using the sample data provided. The acquisition function selects the next samples from the search space and these new sample points are used to update the surrogate model, increasing the accuracy. The cycle of update and optimise continues to produce an accurate model of the chosen objective function. The surrogate model most used in Bayesian optimisation is the Gaussian process, due to its flexibility to fit wide ranges of data, and the construction of a Gaussian distribution. Other models for the surrogate include the tree-structured parzen estimator (TPE), a sequential model based optimisation approach which constructs models to approximate the performance of parameters based on historical measurements, and chooses new parameters to test based on this model.

A simple, flow-style overview of the different approaches to ML models that have been discussed in this section can be seen in Figure 4.3.

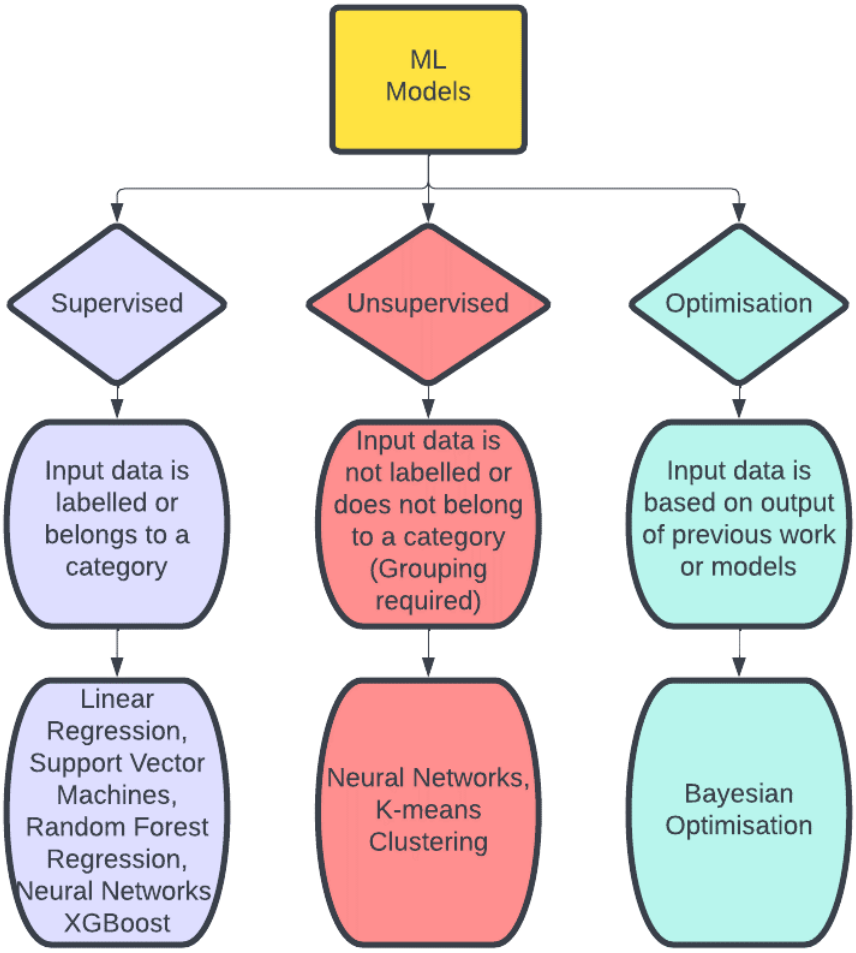


Figure 4.3: *A flow diagram demonstrating the classification of some basic, and some more complex, machine learning model types.*

4.7 Machine Learning for Synthesis and Investigation of Solid State Materials

The application of ML in the realm of solid state chemistry is becoming commonplace for predicting realisable structures as well as their synthesis and intrinsic properties. In addition to benefiting from lower research and production costs as the sector moves away from trial-and-error experimentation, improvements to the field include: smaller starting material requirements, fewer failed structures, and less reliance on the intuition of chemists.

The availability of pre-generated data that can currently be found in journals and databases has made it easier to train models for property predictions. This, combined with user-friendly models that can be used by chemists with limited coding experience means that new developments in ML for material synthesis are being made every day. The year 2022 has so far seen a significant increase in publications which employ ML for material

synthesis predictions, with an expectation that this trend will continue throughout the decade. The Thomson Reuters' Web of Science offers researchers the ability to search for key words and subjects, a useful trend analysis tools which can be used to confirm the increase in popularity of certain topics, (found here: https://www.webofscience.com/wos/woscc/basic-search), the tool was used to confirm the increasing trend in ML for MOF synthesis, and the results can be seen in Figure 4.4.

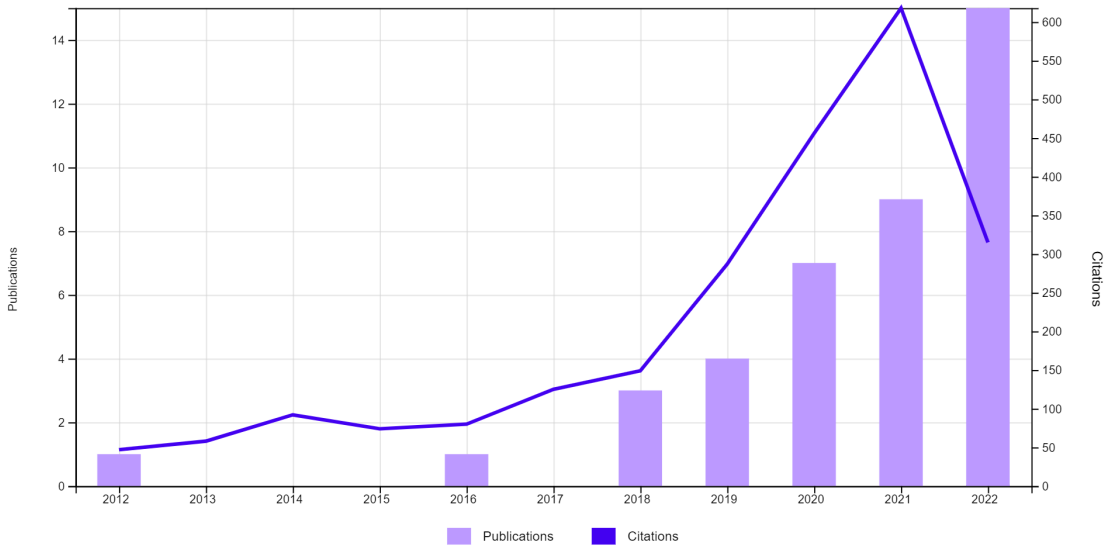


Figure 4.4: *A Web of Science search for trends in publications using the key words 'metal-organic framework', 'synthesis', and 'machine learning' as of July 2022. (https://www.webofscience.com/wos/woscc/citation-report/c0f28728-bd2b-4392-834e-7bc24ac6334b-474cb7be)*

ML tools began with property prediction, enhancement, and analysis of the structure-property relationships. Kennicutt et al. (2016) [45] used SVM, a supervised machine learning algorithm, to predict the adsorption rate of activated carbon, specifically choosing SVM to reduce the risk of overfitting and for ease of use. The model was trained on 95 compounds with 23 structure descriptors, and reached a final training dataset R2 accuracy of 0.932. Although there is a high R2 value, it may not necessarily be a suitable predictive model for compounds outside of the training domain. Efforts were focused on data for Calgon Filtrasorb 400 (F400), a well-studied, microporous, coal based adsorbent and so descriptors for carbon surface chemistry and pore properties were not considered.

More recently, Dico et al. (2021) [46] used several independent ML models to assess natural nano-porous clays for their use in adsorption and catalysis. An extremely randomised trees (Extra Trees) regression algorithm was used to characterise the raw clay characteristics, additive characteristics, and processing conditions assessed against simple decision tree, random forest, and MLP models. The final R^2 for the extra tree regressor of 0.77 is reasonable for a data set of this size, based on 41 feature representations, with a variation in R^2 values when reduced to 20 descriptors. The data has sufficient accuracy for its predictive ability of most parameters, which will enable its use as a tool for further assessment of the viability for parameters required in the processing of natural porous materials, particularly porous clays. The analysis of the models sheds light onto the features of raw minerals which significantly affect the internal and external surface area, and pore volume.

Following the rise in models aimed at finding the relationships between property and structure, chemists began to develop models that would look to find and confirm synthesis-structure relationships. These models work to extract significant synthesis descriptors from substantial amounts of data and realise patterns unable to be seen by humans. For example, Jensen et al. (2019) [39] used ML models to investigate less intuitive relationships between zeolite synthesis parameters and the final structure. They trained a single decision tree regression model on 898 pure phase zeolite systems specifically to demonstrate ML synthesis intuition, rather than predictive ability. This model was able to reproduce the framework density of training data with an impressive R2 value of 0.97.

Raccuglia et al. (2016) [40] developed an algorithm focused on exploring the chemical space to predict a range of reaction conditions that will result in crystallisation of template vanadium-selenites. Information gathered from 'dark' reactions – failed or unsuccessful hydrothermal syntheses, trained an ML model to predict reaction success. The model successfully predicted conditions for new products with a final success rate of 89%. The authors concluded that the model may be used to reveal further relationships between reaction conditions and product formation.

Muraoka et al. (2019) [47] also investigated the relationships between reaction conditions and products using ML synthesis prediction techniques. The data set was formed using NLP from zeolite synthesis records, alongside a similarity network of the crystal structure and synthesis descriptors, producing a dataset of 686 synthesis conditions. Multiple ML models including random forest and XGBoost were used to predict the synthesis results from these custom descriptors. The XGBoost was found to have the highest accuracy at 75-80%, and while not within the scope of the work, they envisioned that the XGBoost model could be used to provide further informatics into the likelihood of formation of specific zeolites. The combination of similarity networks with synthesis prediction allowed for unexplored areas of the chemical space to be found and populated, increasing the diversity of the products formed, a major challenge with zeolites and MOFs alike.

XGBoost is a popular ML model used for predicting synthesis conditions for reactions. Xie et al. (2020) [29] used 9 models to predict the synthesis conditions of MONCs and successfully synthesise a new set of crystalline MONCs. The data originated from archived lab notebooks, similar to Raccuglia et al (2016) [40], and provided a dataset of 486 reactions. The XGBoost was found to have the highest prediction accuracy at 91% and was able to quantify "chemical intuition". By providing quantified importance values, reaction parameters were ranked for future reactions. These results also compared the model's synthesis prediction against a chemist, achieving a prediction accuracy 5% higher than the chemist, at 80%.

The continuous development of ML models lays the foundation of new techniques that can enable accurate and inexpensive synthesis paths for solid state materials. The reaction data gathered for synthesis processes will provide a starting point for further analysis by chemists and data scientists. Combining new data produced via ML models, with improvements in technique for design of faster, more sustainable, and more economic synthesis pathways will see the field shift more quickly to full digitisation for a wide range of material requirements.

4.8 Machine Learning in Design and Discovery of MOFs

Developments in the field of ML for structurally similar materials, such as zeolites and organic porous structures, have also inspired MOF scientists to perform synthesis prediction studies. Due to the large volumes of data produced, ML is quickly becoming a necessity for efficient exploration of the MOF material space. For the large library of existing MOF structures and considering its' continued growth, high throughput screening techniques on their own are no longer fast enough to identify promising materials for synthesis. It is very difficult to computationally screen the vast material space for a single application, restricting the likelihood of the most effective structure being identified. The ML models which are discussed in this section have contributed to recent advances in the shift towards a new data-driven and digitalised paradigm to design and discover new MOFs.

Inverse design is a tool that is seeing a growth in popularity for chemists with advanced ML tools, particularly deep learning, being developed to aid this process. Zhang et al. (2020) [19] developed an inverse design algorithm to directly create novel MOFs for use in carbon capture. In this contribution, the authors combined Monte Carlo tree search with recurrent neural networks (RNN) based on input from 10 different combinations of metal nodes and topologies from previously reported experimental MOFs, these can be seen in Figure 4.5. Using the criteria obtained by the ML study, the algorithm hypothesised a novel MOF based on IRMOF-15, which was expected to perform in carbon capture applications with a CO_2 adsorption capacity up to 543% higher than the input structure. Other examples include a 165% increase on the current capability of MOF-118, and 11% increase for MOF-119. A small increase in CO_2 adsorption of MOF-119 is expected considering the already high loading capacity of 8.18 mmol/g, particularly compared with the increase in loading from 1.29 to 8.30 mmol/g hypothesised for IRMOF-15.

This large increase in CO_2 uptake can be attributed to The interaction between CO_2 molecules and the MOF itself, which are governed by Van der Waal forces as well as electrostatic and dispersion interactions. Adjusting linker design provides more degrees of freedom for adsorption which the algorithm designed in this paper explores. It is known that there is a trade-off between density of adsorption sites and density of material, too large pore spaces often lead to a low density of adsorption sites, short linkers with side function groups were theorised to increase adsorption site density. Further, linkers which contain more non-carbon atoms (such as S and N) provide strong adsorption sites due to their negative atomic charges.

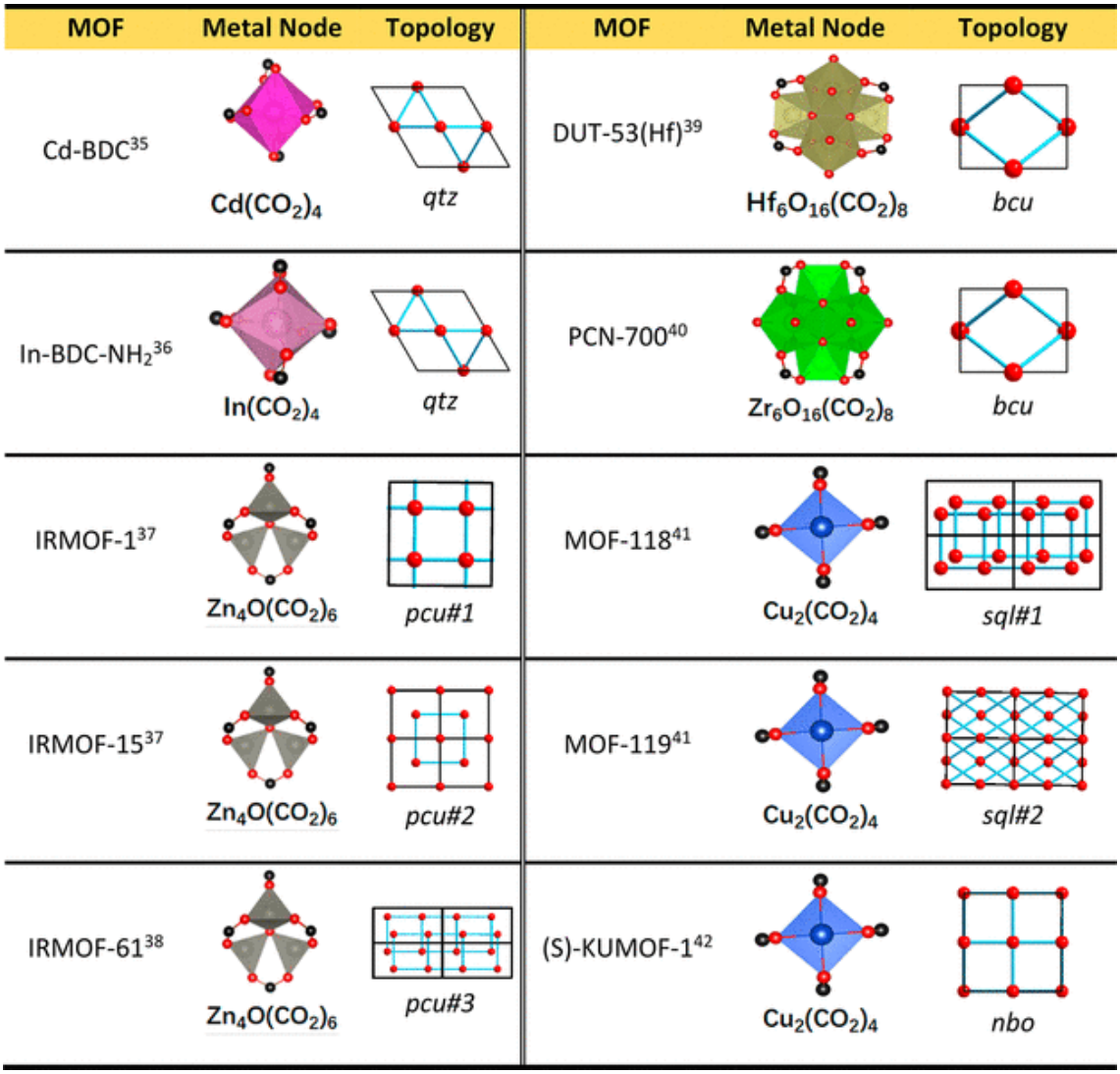


Figure 4.5: *Input topologies of novel experimental MOFs for use in an inverse design algorithm targeting structures for top performance in carbon capture applications. Reprinted with permission from [19]. Copyright 2020 American Chemical Society.*

Hardian et al. (2020) [48] developed a new ML module that combines a design of experiments (DoE), a SVM, an evolutionary algorithm, and a desirability function to predict the optimal conditions for sustainable ZIF-8 synthesis. DoE is an alternative systematic approach to achieve good balance between a reduced number of experiments and efficiency, allowing different factors to be varied and investigated simultaneously, in the hopes of accelerating the process of discovery and optimisation.

The data was first obtained by experimental face centred central composite design, completing 27 runs. The structures' crystallinity was analysed using x-ray diffraction (XRD), and after confirmation of sharp XRD peaks which well matched the calculated XRD pattern, AI 1 was used with the experimental data output from the DoE as input for the SVM algorithm, followed by a grid search to generate 456,976 virtual data points. The results found that highest product quality was achieved at high voltage, long reaction time, low electrolyte concentration, and high linker concentration. AI 2 followed a similar

method to AI 1, followed by 50 random virtual data generations as the initial population for the evolutionary algorithm, followed by the implementation of a desirability function for the last optimisation step. Two optimisation steps were considered, firstly to maximise the product quality, and secondly to maximise product quality and process sustainability.

DoE results were insufficient to identify optimal conditions, AI 1 created an unnecessarily large data set which increased the computational cost, and AI 2 with the incorporation of the evolutionary algorithm made it possible to screen only the best data. These predicted conditions were used to synthesise ZIF-8, and the resulting structure had 100% purity, 88% yield with 86% crystallinity. A major argument towards using ML for material synthesis is to improve the sustainability of the process, with in-silico design allowing for a reduction in reagents and energy. The assessed environmental sustainability and provided a final E-factor for the synthesis reaction of ZIF-8 at 11 kg/kg, and a carbon footprint of 27 kg/kg using 7 kWh/kg.

As expected, most current synthesis approaches have a heavy reliance on the chemical intuition of a chemist with experience of previous synthesis routes. This chemical intuition is lacking in ML tools, and can often be found to be slowing progress. Chemists have been looking to develop this chemical intuition in their algorithms by evaluating the effects of synthesis conditions and their relationship with the final structure. Kitamura et al. (2021) [49] used cluster analysis to classify powder XRD patterns of products and determine experimental success, and decision tree analysis to visualise the experimental results to determine dominant synthesis factors for the production of KGF-3, building unit used in the synthesis of MOFs containing lanthanides. Lanthanide based MOFs are particularly difficult to synthesise and predict, due to a high sensitivity to condition changes, leading to overall poor reproducibility. Data collected from 108 experiments, focusing on lanthanide ions, concentration of metal ion and/or ligand solution, reaction temperature and time, cooling time, and type of reaction vessel found that the synthesis results are highly affected by the lanthanide ion. After difficulty in isolating KGF-3 from initial screening experiments, dominant factors were extracted by evaluating both successful and failed procedures using ML. With this information, the experimentalists successfully synthesised a series of novel pillar-layered lanthanide MOFs containing the double-layer-based building units KGF-3.

Huelsenbeck et al. (2021) [50] developed an active learning algorithm to aid in the synthesis of HKUST-1 thin film. Some MOFs can be grown on multiple substrates using a roll-to-roll process, however this process often lacked full coverage. Other techniques include layer-by-layer growth, solvothermal growth, and gel-layer growth, with a greater success rate for full coverage. Drawbacks for these techniques include slow crystallisation, lack of orientation, and poor thickness control. HKUST-1 thin film is used in transistors and sensors, and after preparation the final product must have a full coverage of the substrate with no void spaces. When manufacturing thin films, the synthesis process must also take coating speed, substrate temperature and the number of coating passes into account. A pool-based active learning (PAL) and regression method was used to efficiently guide the solution-shearing synthesis. Each iteration chose 18 diverse and representative solution-shearing process parameters for validation and feedback, these were composed of samples created using parameters determined by a generalised subset design (GSD). Each sample was replicated three times and characterised using optical microscopy to label the samples as “fully covered” or “not fully covered”, where the substrate is visible between HKUST-1 particles. The results showed 22% of the 18 initial

experimentally synthesised samples were fully covered. An ensemble SVM model was trained to classify the coverage into the same two tiers, which was then used to predict coverage for over 11 million parameter combinations in the unexplored data set. A second set of experiments were performed to obtain thin films based on the 18 parameter sets, with optical micrographs showing 67% of conditions having full coverage. These were used to retrain the SVM ensemble, resulting in a final virtual parameter space predicting 13% of parameter combinations to have full coverage. The use of active learning combined with a solution-shearing process resulted in a final product with a large full coverage and a minimum thickness of 2.2 μm. This work not only uses ML to predict synthesis conditions, but also the results from each experimental stage to optimise performance, allowing for an increased pace of material development; a key step for automating and digitising manufacturing.

Luo et al. (2022) [18] collected data based on the CoRE MOF database, with a training data set composed of MOFs extracted from published and predominantly successful experimental synthesis data, focusing on six relevant parameters. Random forest and neural network were the best performers used to predict the synthesis time and temperature of MOF structures. The results were compared, showing the random forest approach had the higher accuracy across all predicted parameters, however, the neural network predictions will become more accurate as datasets grow, and additionally may be able to exploit correlations between different synthesis parameters. Therefore, the neural network has greater potential as the field develops where more complex models are expected to outperform random forest in the not so distant future. A comparison between the expert trial and error approach, and a ML model implementation is visualised in Figure 4.6.

Prediction of MOF synthesis conditions is a difficult task as for many cases there is not one true solution, a whole range of conditions can lead to a successful synthesis. Some reactions may be optimised for yield; others may be chosen for environmental or financial cost. Despite what can be described as low R^2 values for reaction temperature prediction at 0.286, and predicted reaction time at 0.076, comparisons between the predictions of 11 human MOF synthesis experts revealed that the ML model out predicted all experts who had an R2 value much closer to zero, even after averaging across all human estimates. Even small correlations learned and exploited by ML modelling can help to better estimate synthesis conditions without the availability of big data.

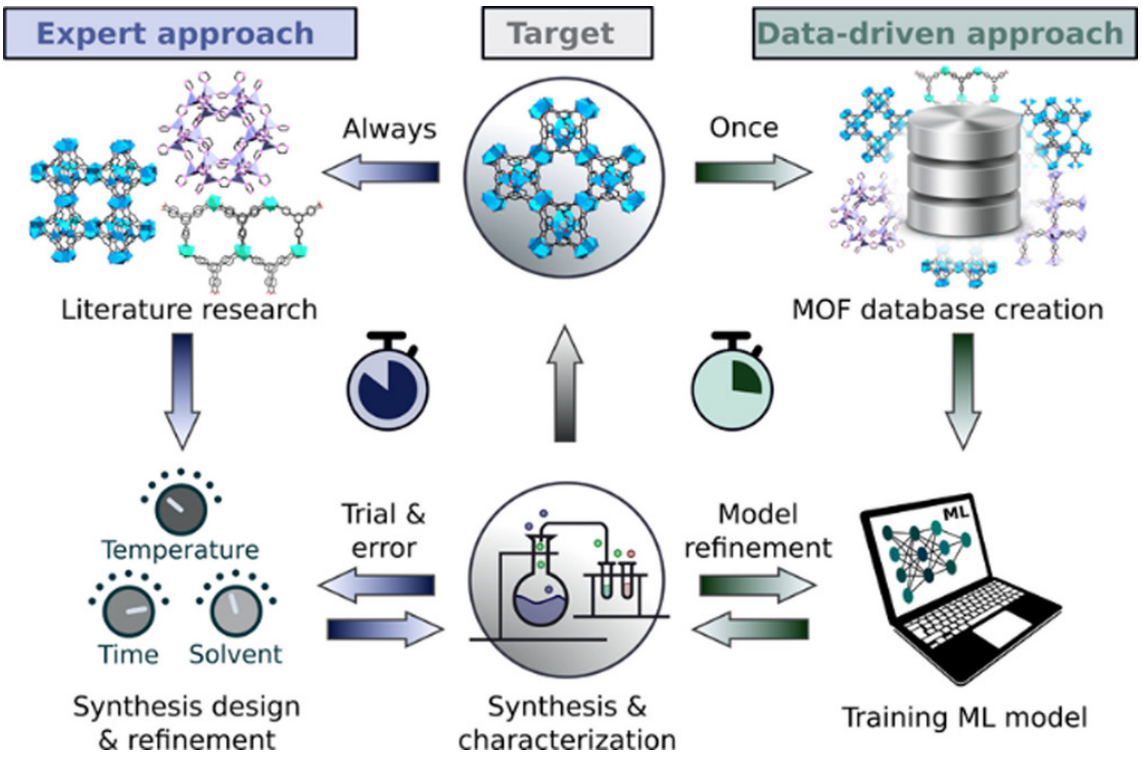


Figure 4.6: *A comparison of the trial and error approach versus training machine learning models to predict synthesis conditions [18], licenced under CC BY 4.0.*

Park et al. (2022) [34] created a database of 46,701 structures from 28,565 published papers found in the CSD MOF subset, and a PU learning algorithm was chosen to reduce the positive bias found in the dataset. PU learning algorithms are typically used in cases where the proportion of positive to negative data is heavily skewed, and this approach has been used previously to predict synthesis information in inorganic materials [51]. Extracted synthesis conditions were used as positive data, and to be regarded as positive data the entry must include composition or temperature, plus one other parameter, this is shown in Figure 4.7. In total 3,748 pieces of extracted MOF synthesis information were considered as the positive data, with 2,998 positive data points used in in the training set and 750 in the test set. In-silico data generated by randomly sampling parameters for the extracted data, was classified as unlabelled data. Randomly generated unlabelled data totalled 1,000,000 pieces, with 900,000 used in the training set and 100,000 used in the test set. Often, ML methods such as decision tree or SVM are used as the binary classifier for PU learning, but in this case that approach was not appropriate and a neural network with simple dense layers was used instead. At each iteration the binary classifier is trained, with positive and unlabelled data in a ratio of 1:10, using the Adam optimiser implemented using Tensorflow [52].

The model used the input synthesis conditions to predict the output crystallinity of a structure. Final output crystallinity scores, with 1 representing conditions for high crystalline structure and 0 indicating conditions for low crystalline structures, were determined by averaging the prediction scores of the binary classifier over the total number of iterations. The final score is denoted as 'crystal score' and crystallinity is classified based on the output of the PU learning algorithm exceeding 0.5. A positive-negative (PN learning) learning model was implemented as an alternative method to evaluate the

performance of the PU learning model. For the test set, the recalled scores were: PU learning at 83.1% and PN learning at 50.3%. False negatives are also low for PU learning, but higher for PN learning.

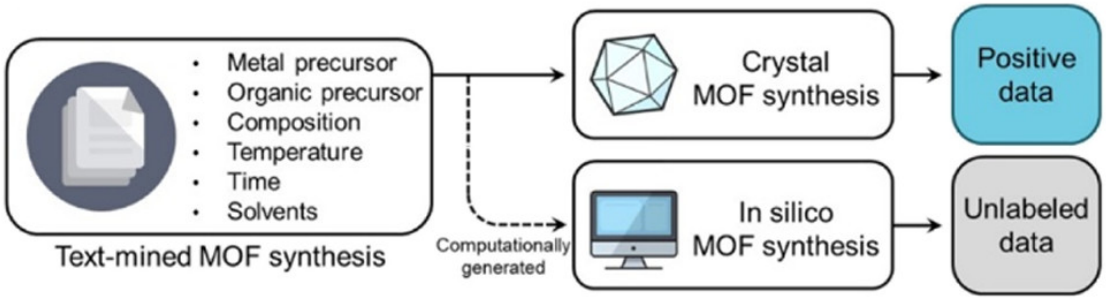


Figure 4.7: *A schematic showing the criteria used to differentiate the positive (P) and unlabelled (U) data. Reprinted (adapted) with permission from [34]. Copyright 2022 American Chemical Society.*

ML models are used in conjunction with MOF synthesis for all stages of material design, discovery, and manufacture. These approaches highlighted in this section are concerned with not only discovering new materials, but with discovering new synthesis techniques for existing materials, and the outcomes of new techniques such that the yield and crystallinity can be determined before experimental synthesis even begins to take place. However, despite the ongoing adoption of ML there are still many limitations to the process of applying ML to MOFs at every stage.

4.9 Current Limitations of Machine Learning for MOFs

A major factor in the chance of success for all ML is meeting sufficient data requirements. NLP has been used to access thousands of documents within the published space but at present, due to the lack of systematic labelling when reporting experimental synthesis, NLP often misses critical information. With all ML algorithms, poor quality data cannot be used to train the models, and the processed can often be referred to as "rubbish in equals rubbish out". Where data is omitted or incorrectly parsed, it might require chemists to find the missing values within the literature, taking up valuable time and resources to correct a data set so that it is suitable for use.

Mehr et al. (2020) [53] have proposed a solution to the lack of systematic reporting of synthesis conditions. They developed a software platform that uses NLP to translate organic chemistry literature directly into an editable code which may be used to drive automated synthesis within a laboratory setting. Automatic literature reading has the ability to create a universal autonomous workflow, this is demonstrated in Figure 4.8. There is currently a plethora of robotic solutions, but they lack a common standard architecture, and often struggle to adapt to new synthetic methods. The standard of recording and reporting of new chemical compound synthesis varies greatly and procedures are typically explained in ambiguous and incomplete passages of text, relying on expert chemical intuition to bridge the gaps. The quality of reaction database data therefore can be sporadic, and this presents many problems in terms of reproducibility and for the development of reliable digital methods that aim to predict synthesis properties for new materials and functionalisation.

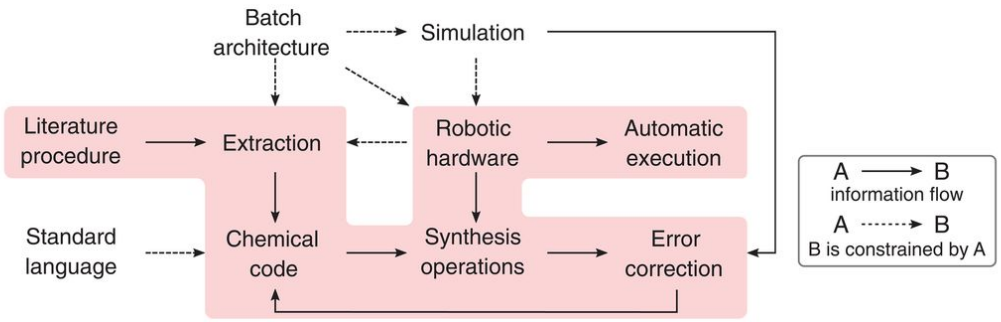


Figure 4.8: *A universal system for the automatic execution of chemical synthesis from literature. Extraction of the procedure is followed by an algorithmic process for producing the code that conforms to a standard hardware and software architecture. Manual error correction and simulated execution ensure reliability and safety [53].*

This leads the discussion to the importance of ensuring that there is an availability of high quality training data to give a ML model the best chance of making reliable and accurate predictions. Most published work reports only successful experiment and synthesis data, particularly in the field of MOFs, however zeolite studies are noted as being particularly good for reporting both successful and failed experiments [39]. In cases where most failed procedures are confined to personal notes and lab books, the use of purely successful data found in published literature and databases may lead to bias within training sets.

Moosavi et al. (2019) investigated a robotic synthesis approach guided by ML for HKUST-1 synthesis optimisation [54]. This approach enabled the variation of multiple synthesis conditions such as solvent composition, temperature, and method (e.g. conventional heating, microwave, electrochemistry) and enabled the recording of both successful and failed data. This regenerated experimental data was used to optimise the ML models with consideration towards the removal of potential bias. Xie et al. (2020) [29] and Raccuglia et al. (2016) [40], took advantage of archived lab notebooks holding all experimental results in an attempt to overcome bias, however this approach used real time failed experimental output data. These techniques worked well on a small scale, but for larger datasets, alternatives must be proposed. Park et al. (2022) [34] worked to overcome the lack of negative data available, with regards to the crystallinity of MOFs, by using PU learning algorithms. Although these are intuitive solutions to maximise ML potential, new publications should endeavour to ensure the suitability of data for digitisation and chemists should begin to realise the importance of all experimental data and begin to publish all results, including those of failed synthesis. Moosavi et al. (2019) [54] produced a public webpage on the MaterialsCloud as part of the work on HKUST-1. The aim of this web application is to involve a large number of groups involved in MOF synthesis to collective report failed or partially successful experiments, offering the potential to change the way the community approaches synthetic chemistry. Chemists can document all synthesis reactions, with public access to all reaction data, increasing the amount of negative data available for training.

Lastly, the development of automated systems that may run reactions and create molecules are often hindered because of a lack of a machine readable standard within experimental publications. Glasby et al. (2023) [21] also suggested that publications introduce new standardised submission templates for material synthesis articles so that

the data is presented in a suitable manner for reliable and accurate text mining and parsing tools. Taking this approach would significantly simplify the parsing techniques for all NLP approaches, and simple methods can be used to extract key parameters from tables of well presented data without the need for continuous re-configuration of parsers or the current requirement for blacklists.

4.10 Automated Synthesis and Digital Manufacturing

Digital manufacturing aims to use previously collected data to streamline synthesis, reduce risk, and locate the most viable reaction pathways and cost-effective materials for a given material. This involves the use of a database in conjunction with the material property data, NLP for accessing published reactions, and ML for property and synthesis prediction. King et al. (2009) [55] was one of the first publications to acknowledge the influence of computation in the scientific process, and their development of Robot Scientist "Adam" was used to generate genomic hypotheses about yeast, which were confirmed with manual experiment. In the years following, there has been a significant influx in the number of publications focused on automated synthesis, these include: Burger et al. (2020) [56], who created a mobile robot that ran autonomously for eight days, driven by a Bayesian search algorithm to search for improved photocatalysis for hydrogen production from water, Sun et al. (2021) [57], who developed a meta-learning model to predict the adsorption loading of materials over a range of temperatures and pressures, and Domingues et al. (2022) [58], who chose genetic algorithms to obtain conditions that provide excellent crystallinity and yield for the microwave based high-throughput robotic synthesis of Al-PMOF.

Pyzer-Knapp et al. (2021) [59] used Bayesian optimisation, a branch of ML, in combination with an energy structure function map (ESF) to aid in the discovery of porous crystals for methane capture. The molecules T2, P2, and T2E, as seen in Figure 4.9 were chosen and screened for methane deliverable capacity, as they have been predicted to have stable crystal structures. ESF maps are very computationally expensive for this particular application due to the large energy range of predicted crystal structures, plus the effect of solvent stabilisation and methane adsorption calculations. It took around 800,000 CPU hours to compute an ESF map for a single molecule (T2E) in this study.

Bayesian optimisation was used to selectively acquire energy and property data to generate the same levels of insight as ESF at a fraction of the computational cost. Without Bayesian optimization, the generation of the energy structure maps are highly computationally expensive and computational cost increases with complexity, a particular drawback for porous materials where the energy range across the crystal structure is extended by solvent templating. Bayesian optimisation techniques including Thompson sampling for parallel optimisation, and greedy sampling, were compared across the three systems, with a clear preference for the Bayesian approach in the T2E and T2 systems. Using this technique, an enormous 544,955 hours of computational time were saved. In cases where density functional theory (DFT) calculations would be required for lattice energy rankings, then the savings would become even greater. During this time saved, many more candidate molecules can be screened, increasing the likelihood of finding better candidates for methane uptake in the same duration, although it is important to remember that as with all computational accelerations using an ML approach, there may not be the same completeness to the study compared with using the ESF mapping approach as some parameters may not have been calculated by ML models.

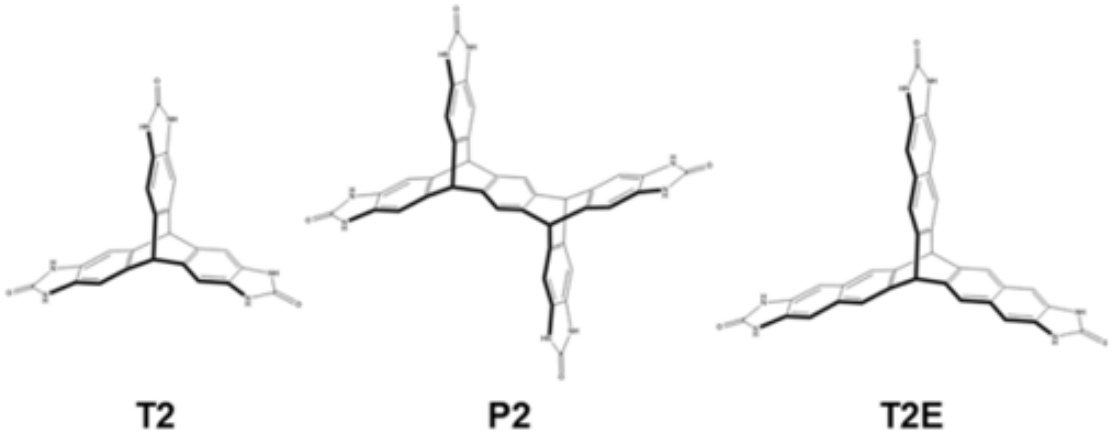


Figure 4.9: *Skeletal structures of T2, P2, T2E. These specific structures are used by Pyzer-Knapp et al. [59] as example materials in their study to accelerate the computational discovery of porous solids through improved navigation of ESF maps, licenced under CC BY 4.0.*

The application of ML in synthesis and digitisation of materials production are relatively new and exciting fields of research. One particular area of interest is focused on automated synthesis improvement, also referred to as flow chemistry, where the design of synthesis systems is continually improved using synthesis steps extracted from literature as inputs, and an ML approach repeatedly alters the inputs based on the output data. This continual improvement approach centred around ML data allows for minimal physical input from chemists during reactions, and it is hoped that this technique can outperform the reaction modifications made by chemists from their intuition alone. This powerful new approach has the added bonus of freeing up time for chemists to continue research, instead of being required to complete laborious and repetitive experiments.

Granda et al. (2018) [60] noticed the progress in automated chemistry, online analytics, and real-time optimisation, suggesting it was possible to construct robots which can autonomously explore chemical reactivity. They designed, built, and programmed an organic synthesis robot to autonomously perform reactions based on the Suzuki-Miyaura reaction, comprising of inline spectroscopy, real-time data analysis, and feedback mechanisms. In this experimental setup, the robot was configured to perform up to six experiments in parallel, producing up to 36 sets of successful and failed experimental data each day for use with ML. A schematic of the feedback loop which was used in conjunction with the experimental setup can be seen in Figure 4.10, showing the use of ML to aid in the process of continual reaction improvement. For almost all experimental techniques investigating chemical reactivity, generating data is time-consuming and cost intensive, employing ML to make more educated guesses at each iteration is a significant step to better discovering new pathways.

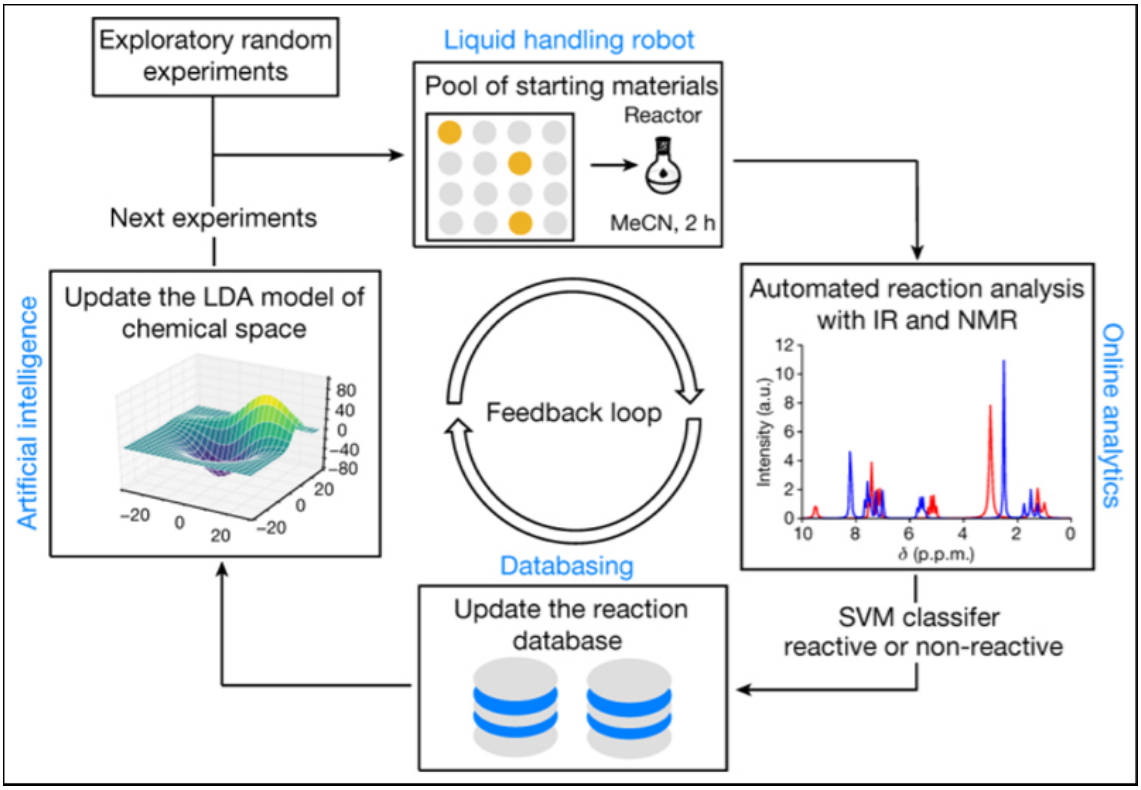


Figure 4.10: *Schematic of the feedback loop for data generation for the Suzuki-Miyaura reaction [60], licenced under CC BY 4.0.*

Moosavi et al. (2019) [54] created a ML methodology focused on capturing chemical intuition from a set of partially failed MOF synthesis attempts, to find the optimal synthesis conditions for yielding the highest surface area HKUST-1 product. Synthesis data was gathered using a genetic algorithm (GA), a robust global optimisation algorithm for searching complex space, and the optimal conditions were synthesised, including the largest BET area HKUST-1 to date. After 120 failed and partly successful experiments which did not achieve the largest BET area, a random decision forest was used to assess the relative importance of synthesis variables to determine their impact on crystallinity and phase purity. The results found that temperature change had up to three times the impact when compared to adjusting the reactant ratio, this is considered to be 'chemical intuition' and ML allows for the transfer of this knowledge into subsequent experiments.

Weighting of the 9 model parameters using the previously determined chemical knowledge shrinks the chemical space of HKUST-1, and can be transferred to a new synthesis. Conditions for synthesis of Zn-HKUST-1 were predicted across a weighed set of 20 diverse conditions, and 2 methods of synthesis for Zn-HKUST-1 that resulted in crystals were revealed. These 20 intuition based samples would need to be replaced by an estimated 5,000 random samples to maintain the same sampling accuracy. While the work was limited to a small subset of MOFs (HKUST-1), the quantification of synthesis variables can be applied to other future synthesis, particularly in the case where the chemistry is too specific for exact conditions be transferable.

Perhaps sensibly, Wilbraham et al. (2020) [61] suggest that the digitisation of chemistry is not simply about implementing ML or AI to process chemical data, nor is it about the development of increasingly capable automation hardware, but that it should

be focused on unambiguous development of a chemical state machine that uses ontology to connect precise instruction sets with hardware performing chemical transformations. Setting a universal standard should result in an increase in collaboration, reproducibility, and safety while decreasing the labour required to make new compounds and broaden chemical space. Similar to the proposal by Mehr et al. (2020) [53], the authors seek to create a universal programming language that is machine readable, with the ability to be exported and executed on robotic platforms, and it should facilitate the unambiguous dissemination of these procedures. A shift from fixed-configuration synthesis machines to a robotic platform is required to enable the processing of reactions whilst collecting real time data. In fact, the most recent developments within automation have even reached beyond flow chemistry with a focus on bespoke workflows no longer exclusive to synthesis procedures. As technology in this field develops, digital chemical robot systems will require feedback from simple sensors as well real time online analytics to navigate process space autonomously and enable efficient synthesis optimisation and novel reaction discovery.

Mehr et al. (2020) [53] also investigated a system for autonomous workflow combined with NLP. To use batch synthesis for digitisation of chemistry, the robot's hardware must be connected to practical synthesis by an executable hardware-independent programming language, this allows for the execution of laboratory synthetic procedures without manual adaptation or modification. The developed system was designed to be accessible to all, with the goal that instructions should be able to be translated between chemistry and robot without loss of information. This is achieved by allowing users to directly execute procedures imported from literature on an automated synthesis platform such as the Chemputer, a potentially crucial step for large scale digitisation of solid state material production. The authors devised a new chemical programming language, Chemical Description Language (XDL), allowing users to encode procedures without ambiguity, and which represents syntheses as sequences of processes. The system includes a chemical integrated development environment (ChemIDE) which enables the importation of procedures from literature using a NLP called SynthReader, and although this is useful for mining vast literature datasets, a machine-readable representation of procedures with unambiguous details is required, including strict tagging of chemical entities, locations of reagents, implicit process details, and an environment in which the user can manually edit the output. Currently, this system is constrained by the capability of SynthReader, however it has successfully tagged relevant text entries, converted them to a list of actions, added process information in an XDL format and synthesised target modules upon execution of the aforementioned XDL file on an automated platform.

The Crystputer, in a similar vein of digitisation to the Chemputer, is a cyber-physical system developed by Zhao et al. (2021) [62] which has been developed to enable the digital manufacture of nanocrystals via convergence between digital and physical systems. The synthesis process explored in this work is the creation of colloidal Au nanocrystals. The system combines the physical modular set up containing pipette modules and a 6-axis robotic arm, with an ultra-sensitive camera, performing over 2,300 experiments autonomously to develop an Au nanocrystals genome. The process begins with NLP, and a python-based algorithm scans literature to design the experiments based on reported parameters and conditions, these are then exported to the robotic arm and the parsed experiments are conducted on the autonomous physical system. Data is collected and used to train a ML model to pinpoint the relationships between the original synthesis conditions and the product's properties, which can then be used to aid in the retrosynthesis and

scale up of targeted Au nanorods. The Crystputer is yet another development which contributes further to the advancement into automation of production, facilitating the shift of data-driven materials innovation to intelligent manufacturing.

Salley et al. (2020) [63] developed an automated robotic platform for the synthesis of gold nanoparticles based on a Darwinian approach. Genetically inspired optimisation has been used in a range of applications already such as catalysis and in light emitting materials, although not for autonomous synthesis. Here, a genetic algorithm approach was used to mimic natural material evolution for a robotic platform in an attempt to optimise the production of gold nanoparticles over many cycles by discovering new synthesis conditions for known nanoparticle shapes. Over three independent cycles of material evolution the system produced spherical nanoparticles, rods, and octahedral nanoparticles by using optimised rods as seeds.

The system begins with an established spectral target for spherical nanoparticles. Synthesis conditions are extracted from published literature, and these spherical particles are synthesised to obtain a target for the automated system. These spheres were analysed using in-line UV spectroscopy and the platform was given the next set of reagents (which are estimated and optimised by the genetic algorithm) to synthesise nanorods, alongside a new spectral target. These materials can be used as seeds for further cycles, in this study the authors set their own targets for this stage rather than choosing a literature value. This automated closed loop approach has created reliable known materials without bias, and can be used to discover complex nano-constructs using desired spectroscopic responses. This methodology offers many benefits compared with a manual approach including automation, speed, safety, and reproducibility (via the use of a digital code in an automatic platform) in addition to providing researchers with a new tool to aid in the understanding of nanoparticle formation and in the development of new application areas.

Epps et al. (2020) [64] integrated ML with flow chemistry to automate the synthesis of inorganic perovskite quantum dots (QDs). By digitising the process, the self-driving Artificial Chemist is able to create made to measure inorganic QDs from eleven precision tailored QD synthesis compositions that were obtained without prior knowledge, within 30 hours, and using less than 210 mL of QD starting solution. Artificial Chemist was pre-trained to use new precursors to further accelerate the synthetic discovery of QD compositions without user selection of experiments, and further enhance the optoelectronic properties of the in-flow synthesised QDs. This fully autonomous closed loop experiment selection method expedited the tedious process of synthetic path discovery at a fraction of time and material cost when compared with user dependent experiment selection, the full process is shown in Figure 4.11.

The Artificial Chemist uses plug-and-play fluidic micro-reactors, which are capable of autonomous synthesis across multiple target parameters simultaneously, to explore the chemical space of colloidal QDs, learn synthesis pathways, identify composition and relevant routes, transfer knowledge to experiments, and continually synthesise rapidly optimised QDs on demand. This QD technology utilises UV absorption and photoluminescence monitoring alongside a real-time ML based Bayesian optimisation approach. Artificial Chemist studied over 1400 reactions with eleven target values across eight different optimisation algorithms. The final technique comprises of a central control system which responds to a constantly updated ensemble neural network-based Bayesian optimisation algorithm with intelligent decision making.

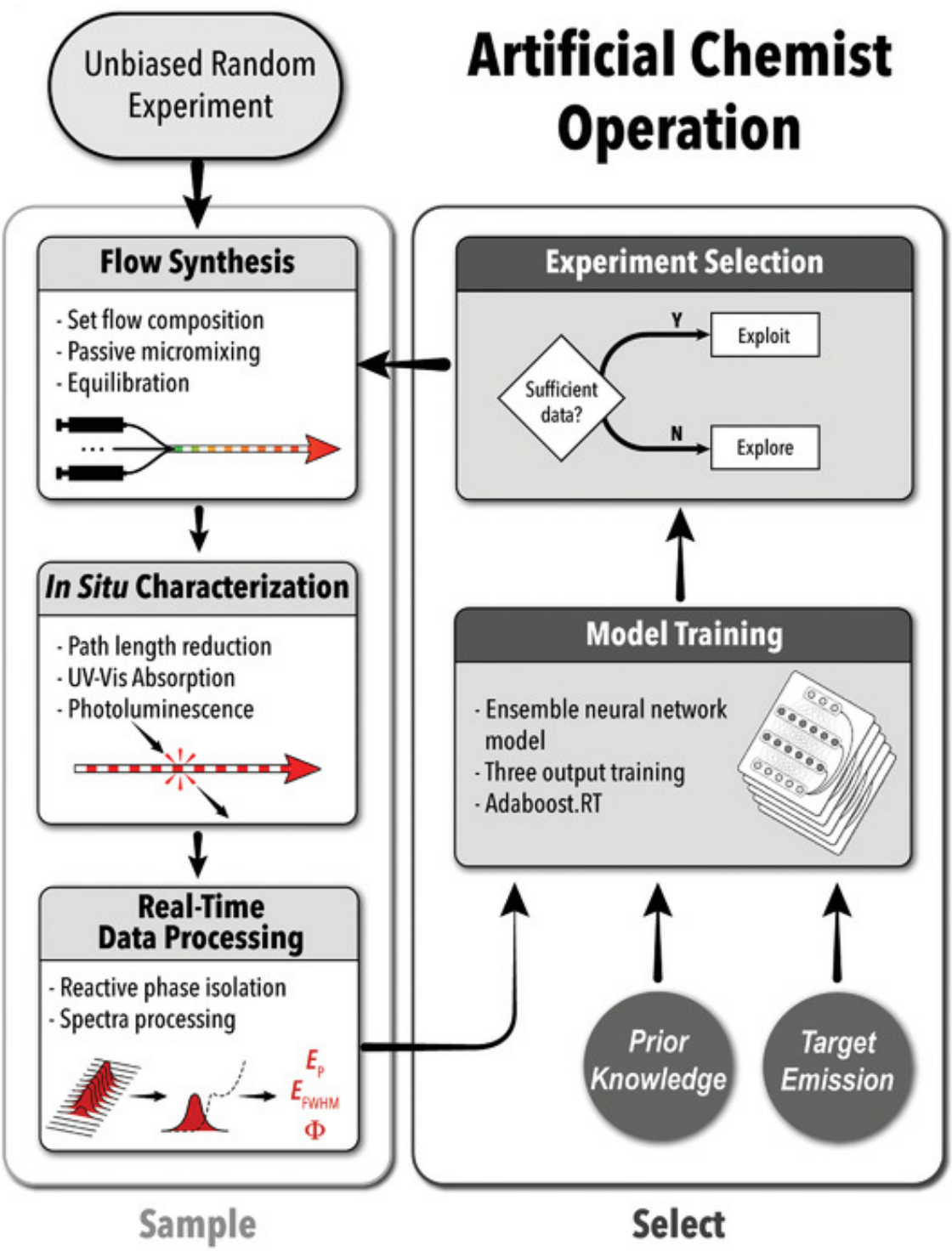


Figure 4.11: *Process flow diagram for the automatic synthesis of quantum dots from initial random experimentation, using flow synthesis and real-time data processing, to new experimental selection [64].*

Abdel-Latif et al. (2021) [65] continued the focus on digital and autonomous manufacturing, with experimentation into nanocrystal lead halide perovskite (LHP) QD synthesis. They identified the optimal formulation of emerging inorganic LHP QDs, which with their vast colloidal synthesis universe and multiple synthesis/post synthesis processing param-

eters, was previously a challenging undertaking for material- and time-intensive batch synthesis strategies. A modular microfluidic synthesis strategy, integrated with an AI-guided decision-making agent for intelligent navigation through the complex synthesis universe of LHP QDs with 10 individually controlled synthesis parameters and an accessible parameter space exceeding 2 × 107, was introduced. The developed autonomous microfluidic experimentation strategy rapidly identified the optimal formulation of LHP QDs through a two-step colloidal synthesis and post synthesis halide exchange reaction. In this study, the use of two in-series microfluidic reactors enabled continuous bandgap engineering of LHP QDs via in-line halide exchange reactions, and using an inert gas within a three-phase flow format resulted in accelerated closed-loop formulation optimisation and end-to-end continuous manufacturing of LHP QDs. These QD crystals have similar applications to certain MOFs, with fields including optoelectronics and photovoltaic devices, as a result of their high photoluminescence yield. An example of the reactor design used to produce these LHP QDs in closed loop formulation is shown in Figure 4.12.

This project improved yield by varying the starting concentration, volumetric injection ratio, halide salt concentrations, and compositions to improve peak emission energy. Current production of the crystals is achieved via a batch flask colloidal synthesis process which can lead to slow production as well as a lack of consistency within final products. Large scale production of nanocrystals requires large reactor designs, synthesis modifications, and oversight for prevention of defects at nanoscale. By integrating AI with modular microfluidic reactors for synthesis of the crystals, including 3 precursor models, 2 in-series microfluidic reactors, and an online spectral characterisation model, the system was designed to have an autonomous run time of 24 hours. To allow for the closed loop manufacturing, optical properties were monitored in real time, and fed forward to an ensemble neural network. After multiple runs were conducted, ten optimised products were identified within a total of 250 experiments, demonstrating end to end manufacturing.

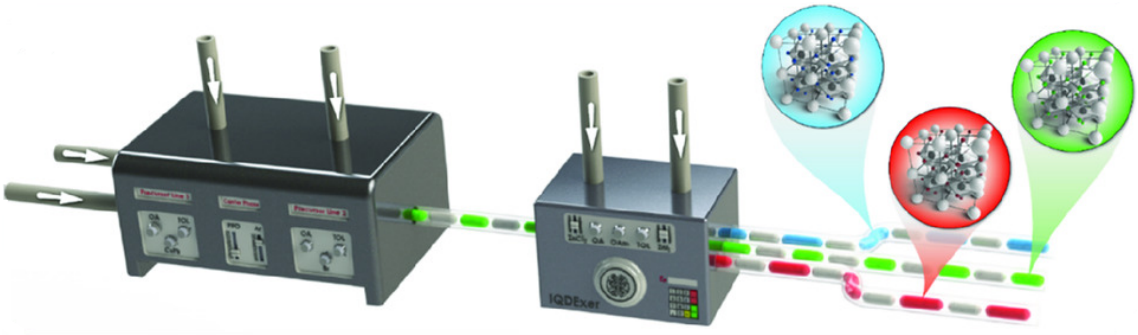


Figure 4.12: *Schematic of the reactor design for the continuous synthesis of lead halide perovskite quantum dots [65], licenced under CC BY 4.0.*

Chang et al. (2020) [66] also understood the necessity of closed loop systems for digitisation, and they developed a method using a system called the autonomous research system (ARES) in combination with Bayesian optimization to improve the growth rate of carbon nanotubes (CNT). Two comparable Bayesian optimization models, to allow for full evaluation, were used to validate the use of ML in CNT growth rates. In the initial stage, seed experiments were manually conducted and analysed, these consisted of a series of input and output variables: total system pressure, flow rate of ethylene, flow rate of hydrogen, total water vapour, and growth temperature. Following this stage, more meaningful variables were calculated for the algorithm including partial pressures of ethylene

and partial pressures of hydrogen, followed by a final critical output of maximum growth rate. Model BO-1 was seeded using 25 experiments which were manually selected and had been confirmed to produce successful growth. In comparison, model BO-2 was randomly seeded from a random selection of 48 unbiased growth conditions. After receiving seed data, the ARES conducted the experiment and the BO-1 and BO-2 models then suggested new growth conditions. ARES executed the new experiment and updated the data set, performing autonomous improvement in a closed-loop fashion. Overall, BO-1 and BO-2 met the goal of improved growth with both converging within 100 experiments, with a growth rate increase of up to a factor of eight.

Moving on from the conversion of synthesis processes from batch approaches to continuous manufacturing, it must be noted that modular systems allow for the easy conversion to adaptable industrial and commercial scale production. However, scaling up a process, if ill-designed, may lead to inefficient production and a larger than desirable proportion of defective materials. Published literature contains thousands of synthesis processes for ZnO, although very few are able to be scaled up to industrial scale without high costs or a compromise in quality. These two key parameters; cost and performance are a necessity when looking at the feasibility of manufacturing materials.

Jose et al. (2021) [67] considered cost and performance when designing their annular micro reactor synthesis (AMS) system for the large scale (kilograms per day) production of ZnO. The Thompson sampling efficient multi-objective (TSEMO) algorithm was used to increase the quality of the product and process using a simultaneous optimisation approach with limited experimental evaluations. Mechanistic insights were determined following the characterisation of post-optimisation materials, which was assessed by comparing development time, safety, complexity, and scalability to known continuous and batch processes. This algorithm used the data obtained from 25 papers, which were surveyed for wet-chemical precipitation methods compatible with the AMS approach, to determine the synthesis variables. These variables were screened to reduce the number of redundant variables and establish conditions for optimisation, producing a total of 26 different conditions for synthesis. Three iterations of TSEMO were performed and 20 experimental conditions were generated. TSEMO then fitted a Gaussian surrogate model for each objective and the next set of experimental conditions were computed to maximise the objective, these conditions were repeated until the maximum number of iterations had been reached. If optimal conditions had not been reached, then the previous steps must be repeated. TSEMO required six experimental steps per iteration, a total of 18 experimental conditions. The molar concentration reached an optimal condition after only one iteration, this indicated that high concentration can produce high performance and yield, a fact not previously realised in the literature.

It is clear that the trend of combining real-time data analysis, ML, and AI into synthesis processes has spread through many areas of materials science. Chemists in all fields from QD manufacture to MOFs are utilising the power of closed-loop processes and implementing systems of continual improvement combined with product feedback data.

4.11 Digital Manufacturing of MOFs

The previous studies worked to improve the real-time data collection for a variety of different materials, this can be seen to have provided a foundation for progression in the digitisation of MOF production. For many chemists looking to produce MOFs au-

tonomously, the first step is often a conversion from the batch process to a continuous one for use in flow chemistry. Batch processing may be difficult to scale up, due to problems with hazards and increased costs and further possibilities of batch to batch error and low reproducibility.

In a contribution focused on the scalability of MOF nanosheets, Jose et al. (2020) [68] suggested that for MOF synthesis to be scalable, while maintaining sufficient precision, improvements to continuous reactors must be made. Current techniques are not scalable nor precise enough to use at industry scale and the characterisation of 2D MOF nanostructures is problematic due to post-processing methods. In this work, copper benzene dicarboxylic acid (CuBDC) nanosheets were synthesised using an annular flow micro reactor with accelerated precipitation kinetics. Previous methods use liquid-phase, hydrothermal conditions in batch reactors requiring long reaction times and elevated temperatures, this method is often hazardous, expensive, and imprecise above kilogram scales. Efficient synthesis is challenging due to difficulty in mixing, the fast kinetics of particle growth, and anisotropic growth. Micro reactors can control mixing conditions tightly and provide fast and continuous mixing, although at high saturation reactor clogging can become a problem. This study overcame the challenges of ton scale MOF nanosheet synthesis and developed a more scalable technique for CuBDC, using trimethylamine at ambient temperature and pressure, by utilising a continuous approach in recently developed AMS which enabled rapid mixing and uniform shear. The synthesised monodispersed CuBDC nanosheets were analysed using XRD and infrared spectroscopy to determine particle size distributions. The final process showed an improved efficiency of up to 105 times the previous batch production techniques.

In another MOF synthesis contribution, Shukre et al. (2022) [69] studied the crystallinity, yield, and precipitation of 45 sample variations of HKUST-1 with the aim to convert the synthesis method from a batch process to a continuous one. The batch process was initially used to pinpoint the conditions for optimal synthesis of HKUST-1, and an optimised reaction was found which could be used in flow conditions with a millifluidic droplet based reactor. Based on a detailed comparison of samples using both batch and flow techniques, the primary conditions identified for optimisation were the residence time, temperature, and the diameter of the inner tubing. The flow process was able to continuously synthesise HKUST-1, with high quality crystals that were comparable to the output of a traditional batch synthesis process. The novel configuration, a variation on the millifluidic droplet reactor, was able to synthesise HKUST-1 within a few minutes of residence time, and has shown there is great potential in the scale-up synthesis of this MOF alongside the opportunity to investigate the process of other MOFs in a similar reactor setup. Figure 4.13 shows the experimental setup of the millifluidic reactor and the additional equipment required to set up the new synthesis technique for HKUST-1.

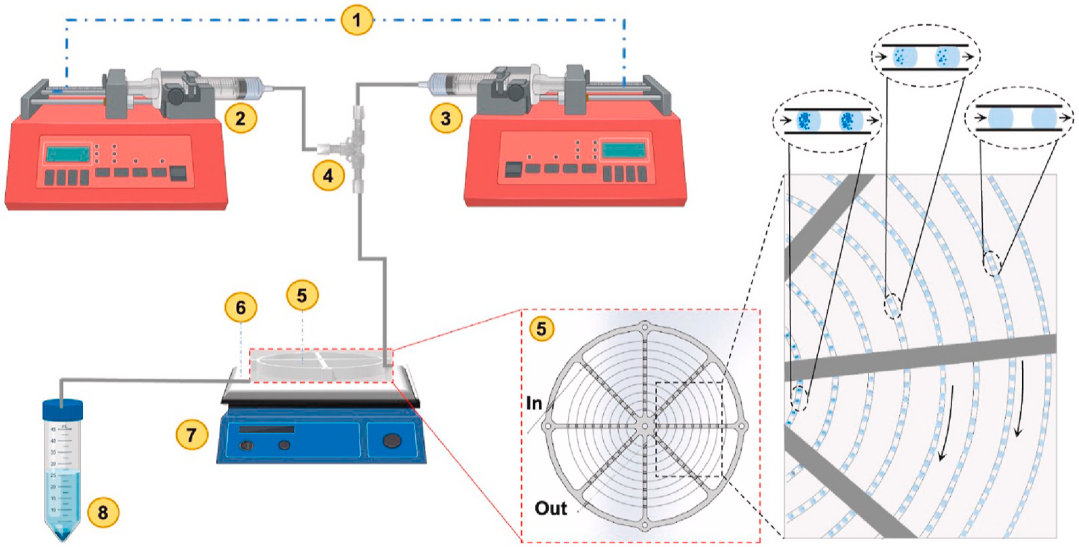


Figure 4.13: *Setup for continuous synthesis of HKUST-1 using the millifluidic reactor. The equipment as shown above: 1. Syringe pump, 2. Silicone oil in continuous phase, 3. Reactant solution in dispersed phase, 4. ETFE Tee, 5. 3D printed anchor, 6. Grooved aluminium block, 7. Hot plate, 8. Product collection vial [69], licenced under CC BY NC ND 4.0.*

Although the previous studies are not strictly focused on the digitisation of the MOF synthesis process, the conversion from batch to flow chemistry opens the door for future studies to implement a digital approach more easily. To shift MOF synthesis towards digitisation of production, availability of data is paramount. A review of MOF sustainability by Julien et al. (2017) [70] praised the work completed by various chemists on their adaptation of ML for data prediction, particularly using the predictive power when approaching the development of sustainable manufacturing procedures that are conscious of environmental impact. Whilst MOFs may often be seen as the key to a sustainable future, particularly in the field of carbon capture, the synthesis of these structures is not free from environmental issue. As the demand for MOFs for use in "green" applications increases, these environmental factors will become amplified as there are several stages of synthesis that have the potential to cause difficulty in scale-up for commercialisation. High energy inputs, use of water as a reaction media, unsafe building blocks, cost of raw materials, and a requirement for bulk petrochemically-derived solvents all have the potential to dismount any MOF scale-up process. The ongoing desire to commercialise MOFs requires an urgent address of these challenges, introduction of ML and the digitisation of processes for non-MOF materials has already proved that a synthesis pathway can be monitored for product yield alongside developing sustainable synthesis properties, and this should be extended to MOFs.

In one of currently very few explicit MOF digitisation approaches, Xie et al. (2021) [71] combined robotic synthesis with a Bayesian optimisation algorithm to accelerate the synthesis of ZIF-67 using direct laser writing apparatus, precursor injecting, and Joule-heating components. The MOF synthesis reaction was automated upon the feeding of Bayesian recommended reaction parameters without prior knowledge, and the platform continually improved the crystallinity of ZIF-67 within limited iterations. Figure 4.14 shows the approach pathway of semi-automated robotic synthesis within this study. The dependencies showed molar ratio, precursor volume, and DC voltage were much more

significant factors in improving crystallinity than the duration of reaction. This study resulted in the creation of a robotic platform that enabled semi-autonomous synthesis, with lower reagent consumption and time, and involved minimal human intervention during production. The process was not completely autonomous due to manual composition measurement although this could be overcome through the development of autonomous X-ray analysis and real time feedback sensors throughout the platform. Other limitations of this approach included the batch-to-batch manufacturing technique, and a non-closed loop optimisation which restricts the development of the robotic platform. To transfer these techniques to the future of automation requires the incorporation of a roll-to-roll approach, accompanied by in situ characterisation, automatic data analysis, and self-optimisation.

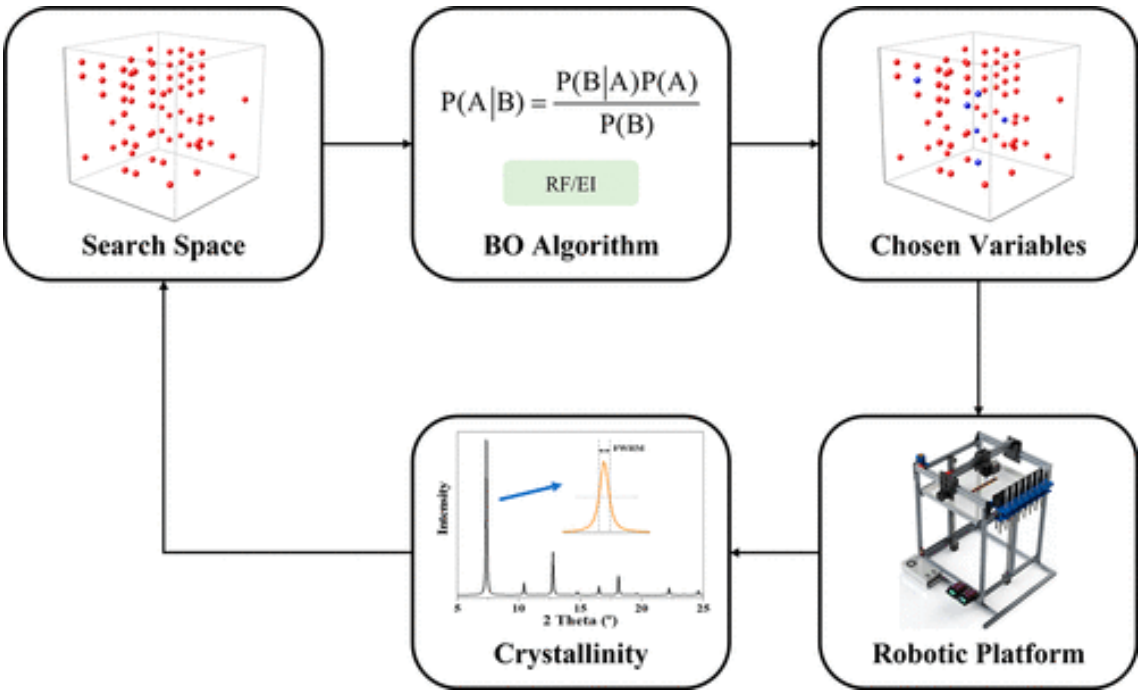


Figure 4.14: *Process flow diagram for the automated synthesis of ZIF-67, showing the use of Bayesian optimisation (BO) to make continual variation to the chosen variables after the output analysis is performed. Reprinted (adapted) with permission from [71]. Copyright 2021 American Chemical Society.*

In another MOF based study, a variation of ZIF-8, after mineralisation with poly(ethylene glycol) (PEG) in the presence of bio-macromolecules, was synthesised through an autonomous computer controlled system by Wu et al. (2022) [72]. An automatic synthesis system prepared PEG mineralised ZIF-8 composite particles based on flow chemistry with microfluidic chips produced using femtosecond laser micromachining. Ideally designed for use by non-specialists, who will be able to obtain target ZIF composites by selecting an input parameter, as the process is able to monitor and regulate itself. After inputting the target size, the system calculates required concentration of reagents, and pumps are instructed to mix the set reactants. The synthesis reaction is monitored automatically by an in situ UV- visual spectrometer to check the size of synthesised ZIF-8 particles and when the calculated results are equal to the experimental results, the reaction is ended and the crystals can be harvested as final product. The authors also assessed the possibility

of remote control of the system, with successful control from a building 20 km away. Remote control of manufacturing is an advantage of production digitisation, opening up the possibility of manufacturing server rooms that would allow small, lower budget research groups without access to expensive equipment to take advantage of digital manufacturing in their research.

4.12 The Future of Digital Manufacturing

Although ML is now a well-established field in most areas of data science and computing —with many commercial uses that people around the world benefit from every day— its use within the realm of material discovery and chemicals manufacturing is relatively new. Integration of ML within the solid state nanomaterials field has seen some significant progress, particularly since the beginning of the decade with multiple articles in high impact publications reporting the successful prediction of novel, sustainable, or economically improved synthesis conditions. To see continued advancement of these increasingly valuable synthesis predictions, future work must focus on overcoming certain limitations. For example, the lack of consistency in the reporting of results and synthesis conditions, alongside the lack of data for unsuccessful experiments, poses one of the most significant challenges. Without an easily accessible and abundant high-quality data source, the accuracy of predictions is hampered and the ML discovery process slows down. The shift towards open-source data repositories within the scientific community has seen the idea of collaboration begin to take off and boost the progress of computational investigation, making this area of research much more accessible, affordable, and environmentally conscious. With websites, GitHub resources, video guides, and helpful documentation all developed to increase the access to data, algorithms, and prediction tools, the use of ML within chemical space is set to accelerate.

Additionally, the introduction of knowledge graphs, graph-based reaction optimisation, and digital twins are a good step forward in producing the knowledge required to discover new or missing knowledge, enable rapid pathway predictions, and evolve the automation of the laboratory [73, 74, 75]. For the large scale integration of ML and digitisation for synthesis of MOFs, ML algorithms should begin to provide high quality outputs for the next stage of production. These high quality and reliable algorithm outputs are key to provide sufficient data for continuous manufacturing and commercialisation of otherwise inaccessible materials, offering essential information to maximise the potential of the input resources, synthesis process, and final product. The predicted synthesis conditions will allow manufacturers to compare the resources required for each structure from input to output, helping to ease the monetary and environmental cost of production. Automated analysis of structure properties, crystallinity, cost, and environmental effects will play a key role in the sustainable manufacturing of these highly functional materials.

Many of the concepts in this chapter can be combined to achieve processes that incorporate real-time adaptability to enable dynamic responses to experimental data and variation in manufacturing conditions. This approach is essential to optimise future synthesis pathways and ensure reproducibility. Flow chemistry systems with automated feedback loops, monitored by artificial intelligence, allow for data-driven optimisation by implementing changes in rates, concentrations, and reaction times instantaneously. The ML models can predict outcomes of a synthesis step given the feedback loops and suggest these adjustments before failures are likely to occur, as can the virtual replication using a digital twin to allow pre-emptive adjustments to avoid suboptimal production. Real-time

adaptability in MOF manufacturing can transform the field by combining the precision of automation with the flexibility of dynamic decision-making, driving more efficient and more scalable production of advanced materials such as MOFs.

It should be noted that as novel manufacturing techniques are explored with an increased use of digital processes, not all MOFs will find suitable pathways for large scale synthesis. Due to the complexity of the nanocrystal structures of a subsection of MOFs, large scale synthesis is unfeasible and unnecessary. Additionally, the rarity of certain transition metals on Earth limits the scale at which certain materials could be produced, these MOFs experience significant costs for materials alone even at kilogramme levels with certain physical constraints limiting the potential of more complex combinations. Recent world events have highlighted many problems regarding the adaptability of production processes when facing fluctuations in demand and unreliable supply chains, it is imperative to assess the vulnerability of the manufacturing pathway such that essential materials are not restricted so that when digitising the synthesis processes it is possible to increase resilience and efficiency. However, the full potential of digitisation cannot be reached until sufficient, high-quality data is accessible, although large-scale global events, such as COVID-19 and the subsequent disruption to global supply lines cannot be predicted.

References

[1] Shilun Qiu and Guangshan Zhu. Molecular engineering for synthesizing novel structures of metal–organic frameworks with multifunctional properties. *Coordination Chemistry Reviews*, 253(23):2891–2911, December 2009.

[2] Norbert Stock and Shyam Biswas. Synthesis of Metal-Organic Frameworks (MOFs): Routes to Various MOF Topologies, Morphologies, and Composites. *Chem. Rev.*, 112(2):933–969, February 2012. Publisher: American Chemical Society.

[3] Arineh Tahmasian and Ali Morsali. Ultrasonic synthesis of a 3D Ni(II) Metal–organic framework at ambient temperature and pressure: New precursor for synthesis of nickel(II) oxide nano-particles. *Inorganica Chimica Acta*, 387:327–331, May 2012.

[4] Seth M. Cohen. Postsynthetic Methods for the Functionalization of Metal–Organic Frameworks. *Chem. Rev.*, 112(2):970–1000, February 2012. Publisher: American Chemical Society.

[5] Sophie E. Miller, Michelle H. Teplensky, Peyman Z. Moghadam, and David Fairen-Jimenez. Metal-organic frameworks as biosensors for luminescence-based detection and imaging. *Interface Focus*, 6(4):20160027, August 2016. Publisher: Royal Society.

[6] Hannelore Konnerth, Babasaheb M. Matsagar, Season S. Chen, Martin H. G. Prechtl, Fa-Kuen Shieh, and Kevin C. W. Wu. Metal-organic framework (MOF)-derived catalysts for fine chemical production. *Coordination Chemistry Reviews*, 416:213319, August 2020.

[7] Zhijie Chen, Megan C. Wasson, Riki J. Drout, Lee Robison, Karam B. Idrees, Julia G. Knapp, Florencia A. Son, Xuan Zhang, Wolfgang Hierse, Clemens Kühn, Stefan Marx, Benjamin Hernandez, and Omar K. Farha. The state of the field: from inception to commercialization of metal–organic frameworks. *Faraday Discuss.*, 225(0):9–69, February 2021. Publisher: The Royal Society of Chemistry.

[8] Peyman Z. Moghadam, Aurelia Li, Seth B. Wiggin, Andi Tao, Andrew G. P. Maloney, Peter A. Wood, Suzanna C. Ward, and David Fairen-Jimenez. Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future. *Chem. Mater.*, 29(7):2618–2625, April 2017. Publisher: American Chemical Society.

[9] Yongchul G. Chung, Jeffrey Camp, Maciej Haranczyk, Benjamin J. Sikora, Wojciech Bury, Vaiva Krungleviciute, Taner Yildirim, Omar K. Farha, David S. Sholl, and Randall Q. Snurr. Computation-Ready, Experimental Metal–Organic Frameworks: A Tool To Enable High-Throughput Screening of Nanoporous Crystals. *Chem. Mater.*, 26(21):6185–6192, November 2014. Publisher: American Chemical Society.

[10] Jorge Gascon, María D. Hernández-Alonso, Ana Rita Almeida, Gerard P. M. van Klink, Freek Kapteijn, and Guido Mul. Isoreticular MOFs as Efficient Photocatalysts with Tunable Band Gap: An Operando FTIR Study of the Photoinduced Oxidation of Propylene. *ChemSusChem*, 1(12):981–983, 2008. _eprint: https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cssc.200800203.

[11] Nazario Lopez, Hanhua Zhao, Akira Ota, Andrey V. Prosvirin, Eric W. Reinheimer, and Kim R. Dunbar. Unprecedented Binary Semiconductors

Based on TCNQ: Single-Crystal X-ray Studies and Physical Properties of Cu(TCNQX2) X=Cl, Br. *Advanced Materials*, 22(9):986–989, 2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adma.200903217.

[12] François-Xavier Coudert. Water Adsorption in Soft and Heterogeneous Nanopores. *Acc. Chem. Res.*, 53(7):1342–1350, July 2020. Publisher: American Chemical Society.

[13] Timur Islamoglu, Zhijie Chen, Megan C. Wasson, Cassandra T. Buru, Kent O. Kirlikovali, Unjila Afrin, Mohammad Rasel Mian, and Omar K. Farha. Metal–Organic Frameworks against Toxic Chemicals. *Chem. Rev.*, 120(16):8130–8160, August 2020. Publisher: American Chemical Society.

[14] Harrison D. Lawson, S. Patrick Walton, and Christina Chan. Metal–Organic Frameworks for Drug Delivery: A Design Perspective. *ACS Appl. Mater. Interfaces*, 13(6):7004–7020, February 2021. Publisher: American Chemical Society.

[15] Federica Zanca, Lawson T. Glasby, Sanggyu Chong, Siyu Chen, Jihan Kim, David Fairen-Jimenez, Bartomeu Monserrat, and Peyman Z. Moghadam. Computational techniques for characterisation of electrically conductive MOFs: quantum calculations and machine learning approaches. *Journal of Materials Chemistry C*, 9(39):13584–13599, 2021. Publisher: Royal Society of Chemistry.

[16] Greg Landrum, Paolo Tosco, Brian Kelley, Ric, sriniker, gedeck, Riccardo Vianello, NadineSchneider, Eisuke Kawashima, Andrew Dalke, David Cosgrove, Dan N, Gareth Jones, Brian Cole, Matt Swain, Samo Turk, AlexanderSavelyev, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, Daniel Probst, Kazuya Ujihara, Vincent F. Scalfani, guillaume godin, Axel Pahl, Francois Berenger, JLVarjo, strets123, JP, and DoliathGavid. rdkit/rdkit: 2022_03_4 (Q1 2022) Release, July 2022.

[17] Seyed Mohamad Moosavi, Aditya Nandy, Kevin Maik Jablonka, Daniele Ongari, Jon Paul Janet, Peter G. Boyd, Yongjin Lee, Berend Smit, and Heather J. Kulik. Understanding the diversity of the metal-organic framework ecosystem. *Nature Communications*, 11(1):4068, August 2020. Number: 1 Publisher: Nature Publishing Group.

[18] Yi Luo, Saientan Bag, Orysia Zaremba, Adrian Cierpka, Jacopo Andreo, Stefan Wuttke, Pascal Friederich, and Manuel Tsotsalas. MOF Synthesis Prediction Enabled by Automatic Data Mining and Machine Learning**. *Angewandte Chemie International Edition*, 61(19):e202200242, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.202200242.

[19] Xiangyu Zhang, Kexin Zhang, and Yongjin Lee. Machine Learning Enabled Tailor-Made Design of Application-Specific Metal-Organic Frameworks. *ACS Applied Materials and Interfaces*, 12(1):734–743, 2020.

[20] Yongchul G. Chung, Emmanuel Haldoupis, Benjamin J. Bucior, Maciej Haranczyk, Seulchan Lee, Hongda Zhang, Konstantinos D. Vogiatzis, Marija Milisavljevic, Sanliang Ling, Jeffrey S. Camp, Ben Slater, J. Ilja Siepmann, David S. Sholl, and Randall Q. Snurr. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data*, 64(12):5985–5998, December 2019. Publisher: American Chemical Society.

[21] Lawson T. Glasby, Kristian Gubsch, Rosalee Bence, Rama Oktavian, Kesler Isoko, Seyed Mohamad Moosavi, Joan L. Cordiner, Jason C. Cole, and Peyman Z. Moghadam. DigiMOF: A Database of Metal–Organic Framework Synthesis Information Generated via Text Mining. *Chem. Mater.*, 35(11):4510–4524, June 2023. Publisher: American Chemical Society.

[22] Dalar Nazarian, Jeffrey S. Camp, and David S. Sholl. A Comprehensive Set of High-Quality Point Charges for Simulations of Metal–Organic Frameworks. *Chemistry of Materials*, 28(3), January 2016. Institution: Univ. of Minnesota, Minneapolis, MN (United States) Publisher: American Chemical Society (ACS).

[23] Andrew S. Rosen, Shaelyn M. Iyer, Debmalya Ray, Zhenpeng Yao, Alán Aspuru-Guzik, Laura Gagliardi, Justin M. Notestein, and Randall Q. Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, 2021.

[24] Aditya Nandy, Gianmarco Terrones, Naveen Arunachalam, Chenru Duan, David W. Kastner, and Heather J. Kulik. MOFSimplify, machine learning models with extracted stability data of three thousand metal–organic frameworks. *Scientific Data*, 9(1):1–11, 2022. ISBN: 4159702201 Publisher: Springer US.

[25] Christopher E. Wilmer, Michael Leaf, Chang Yeon Lee, Omar K. Farha, Brad G. Hauser, Joseph T. Hupp, and Randall Q. Snurr. Large-scale screening of hypothetical metal–organic frameworks. *Nature Chem*, 4(2):83–89, February 2012.

[26] Yamil J. Colón, Diego A. Gómez-Gualdrón, and Randall Q. Snurr. Topologically Guided, Automated Construction of Metal–Organic Frameworks and Their Evaluation for Energy-Related Applications. *Crystal Growth & Design*, 17(11):5801–5810, November 2017. Publisher: American Chemical Society.

[27] Sauradeep Majumdar, Seyed Mohamad Moosavi, Kevin Maik Jablonka, Daniele Ongari, and Berend Smit. Diversifying Databases of Metal Organic Frameworks for High-Throughput Computational Screening. *ACS Appl. Mater. Interfaces*, 13(51):61004–61014, December 2021. Publisher: American Chemical Society.

[28] Zhenpeng Yao, Benjamín Sánchez-Lengeling, N. Scott Bobbitt, Benjamin J. Bucior, Sai Govind Hari Kumar, Sean P. Collins, Thomas Burns, Tom K. Woo, Omar K. Farha, Randall Q. Snurr, and Alán Aspuru-Guzik. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Machine Intelligence*, 3(1):76–86, 2021. Publisher: Springer US.

[29] Yunchao Xie, Chen Zhang, Xiangquan Hu, Chi Zhang, Steven P. Kelley, Jerry L. Atwood, and Jian Lin. Machine Learning Assisted Synthesis of Metal-Organic Nanocapsules. *Journal of the American Chemical Society*, 142(3):1475–1481, 2020.

[30] Kevin Maik Jablonka, Luc Patiny, and Berend Smit. Making the collective knowledge of chemistry open and machine actionable. *Nat. Chem.*, 14(4):365–376, April 2022. Number: 4 Publisher: Nature Publishing Group.

[31] Edward Kim, Kevin Huang, Adam Saunders, Andrew McCallum, Gerbrand Ceder, and Elsa Olivetti. Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chemistry of Materials*, 29(21):9436–9444, 2017.

[32] Matthew C. Swain and Jacqueline M. Cole. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.*, 56(10):1894–1904, October 2016. Publisher: American Chemical Society.

[33] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research*, 49(D1):D1388–D1395, January 2021.

[34] Hyunsoo Park, Yeonghun Kang, Wonyoung Choe, and Jihan Kim. Mining Insights on Metal–Organic Framework Synthesis from Scientific Literature Texts. *J. Chem. Inf. Model.*, 62(5):1190–1198, March 2022. Publisher: American Chemical Society.

[35] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[36] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, March 2016. arXiv:1603.04467 [cs].

[37] Bethany Conroy, Richi Nayak, Andrea Lucia Rocha Hidalgo, and Graeme J. Millar. Evaluation and application of machine learning principles to Zeolite LTA synthesis. *Microporous and Mesoporous Materials*, 335(February):111802, 2022. Publisher: Elsevier Inc.

[38] C. Molnar. *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable*. Independently Published, 2022.

[39] Zach Jensen, Edward Kim, Soonhyoung Kwon, Terry Z. H. Gani, Yuriy Román-Leshkov, Manuel Moliner, Avelino Corma, and Elsa Olivetti. A Machine Learning Approach to Zeolite Synthesis Enabled by Automatic Literature Data Extraction. *ACS Central Science*, 2019.

[40] Paul Raccuglia, Katherine C. Elbert, Philip D. F. Adler, Casey Falk, Malia B. Wenny, Aurelio Mollo, Matthias Zeller, Sorelle A. Friedler, Joshua Schrier, and Alexander J. Norquist. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533(7601):73–76, 2016. Publisher: Nature Publishing Group.

[41] Omer Sagi and Lior Rokach. Approximating XGBoost with an interpretable decision tree. *Information Sciences*, 572:522–542, September 2021.

[42] Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. The Application of Bayesian Methods for Seeking the Extremum. *Towards Global Optimization*, 2(117-129):2, 1978. Publisher: Amsterdam: Elsevier.

[43] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical Bayesian Optimization of Machine Learning Algorithms, August 2012. arXiv:1206.2944 [cs, stat].

[44] F. Nogueira. Bayesian Optimization: Open source constrained global optimization tool for Python. *URL https://github. com/fmfn/BayesianOptimization*, 2014.

[45] A. R. Kennicutt, L. Morkowchuk, M. Krein, C. M. Breneman, and J. E. Kilduff. A quantitative structure–activity relationship to predict efficacy of granular activated carbon adsorption to control emerging contaminants. *SAR and QSAR in Environmental Research*, 27(8):653–676, 2016. Publisher: Taylor & Francis.

[46] Giulia Lo Dico, Álvaro Peña Nuñez, Verónica Carcelén, and Maciej Haranczyk. Machine-learning-accelerated multimodal characterization and multiobjective design optimization of natural porous materials. *Chem. Sci.*, 12(27):9309–9317, July 2021. Publisher: The Royal Society of Chemistry.

[47] Koki Muraoka, Yuki Sada, Daiki Miyazaki, Watcharop Chaikittisilp, and Tatsuya Okubo. Linking synthesis and structure descriptors from a large collection of synthetic records of zeolite materials. *Nature Communications*, 10(1):1–11, 2019. Publisher: Springer US.

[48] R. Hardian, Z. Liang, X. Zhang, and G. Szekely. Artifical Intelligence: the silver bullet for sustaimable materials development. *Green Chemistry*, 22(7521), 2020.

[49] Yu Kitamura, Emi Terado, Zechen Zhang, Hirofumi Yoshikawa, Tomoko Inose, Hiroshi Uji-i, Masaharu Tanimizu, Akihiro Inokuchi, Yoshinobu Kamakura, and Daisuke Tanaka. Failure-Experiment-Supported Optimization of Poorly Reproducible Synthetic Conditions for Novel Lanthanide Metal–Organic Frameworks with Two-Dimensional Secondary Building Units. *Chemistry – A European Journal*, 27(66):16274–16274, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/chem.202104014.

[50] Luke Huelsenbeck, Sangeun Jung, Roberto Herrera Del Valle, Prasanna V. Balachandran, and Gaurav Giri. Accelerated HKUST-1 Thin-Film Property Optimization Using Active Learning. *ACS Applied Materials and Interfaces*, 13(51):61827–61837, 2021.

[51] Jidon Jang, Geun Ho Gu, Juhwan Noh, Juhwan Kim, and Yousung Jung. Structure-Based Synthesizability Prediction of Crystals Using Partially Supervised Learning. *J. Am. Chem. Soc.*, 142(44):18836–18843, November 2020. Publisher: American Chemical Society.

[52] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. arXiv:1412.6980 [cs].

[53] S. Hessam M. Mehr, Matthew Craven, Artem I. Leonov, Graham Keenan, and Leroy Cronin. A universal system for digitization and automatic execution of the chemical synthesis literature. *Science*, 370(6512):101–108, October 2020.

[54] Seyed Mohamad Moosavi, Arunraj Chidambaram, Leopold Talirz, Maciej Haranczyk, Kyriakos C. Stylianou, and Berend Smit. Capturing chemical intuition in synthesis of metal-organic frameworks. *Nat Commun*, 10(1):539, February 2019. Number: 1 Publisher: Nature Publishing Group.

[55] Ross D. King, Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N. Soldatova, Andrew Sparkes, Kenneth E. Whelan, and Amanda Clare. The Automation of Science. *Science*, 324(5923):85–89, April 2009. Publisher: American Association for the Advancement of Science.

[56] Benjamin Burger, Phillip M. Maffettone, Vladimir V. Gusev, Catherine M. Aitchison, Yang Bai, Xiaoyan Wang, Xiaobo Li, Ben M. Alston, Buyi Li, Rob Clowes, Nicola Rankin, Brandon Harris, Reiner Sebastian Sprick, and Andrew I. Cooper. A mobile robotic chemist. *Nature*, 583(7815):237–241, July 2020. Number: 7815 Publisher: Nature Publishing Group.

[57] Yangzesheng Sun, Robert F. DeJaco, Zhao Li, Dai Tang, Stephan Glante, David S. Sholl, Coray M. Colina, Randall Q. Snurr, Matthias Thommes, Martin Hartmann, and J. Ilja Siepmann. Fingerprinting diverse nanoporous materials for optimal hydrogen storage conditions using meta-learning. *Science Advances*, 7(30):eabg3983, July 2021. Publisher: American Association for the Advancement of Science.

[58] Nency P. Domingues, Seyed Mohamad Moosavi, Leopold Talirz, Christopher P. Ireland, Fatmah Mish Ebrahim, and Berend Smit. Using genetic algorithms to systematically improve the synthesis conditions of Al-PMOF. *Communications Chemistry*, 5, June 2022.

[59] Edward O. Pyzer-Knapp, Linjiang Chen, Graeme M. Day, and Andrew I. Cooper. Accelerating computational discovery of porous solids through improved navigation of energy-structure-function maps. *Science Advances*, 7(33), 2021.

[60] Jarosław M. Granda, Liva Donina, Vincenza Dragone, De Liang Long, and Leroy Cronin. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature*, 559(7714):377–381, 2018.

[61] Liam Wilbraham, S. Hessam M. Mehr, and Leroy Cronin. Digitizing Chemistry Using the Chemical Processing Unit: From Synthesis to Discovery. *Accounts of Chemical Research*, 54(2):253–262, 2020.

[62] H. Zhao, W. Chen, Z. Wang, Z. Sun, C. Wang, F. Lai, H. Huang, O. Moses, M. Adam, Z. Chen, Y. He, C. Pang, Y. Lu, P. Chu, Z. Yin, and X. Yu. Cyber-Physical System Enabled Digital Manufacturing of Nanocrystals: A Crystputer. *SSRN*, 2021.

[63] Daniel Salley, Graham Keenan, Jonathan Grizou, Abhishek Sharma, Sergio Martín, and Leroy Cronin. A nanomaterials discovery robot for the Darwinian evolution of shape programmable gold nanoparticles. *Nat Commun*, 11(1):2771, June 2020. Number: 1 Publisher: Nature Publishing Group.

[64] Robert W. Epps, Michael S. Bowen, Amanda A. Volk, Kameel Abdel-Latif, Suyong Han, Kristofer G. Reyes, Aram Amassian, and Milad Abolhasani. Artificial Chemist: An Autonomous Quantum Dot Synthesis Bot. *Advanced Materials*, 32(30):2001626, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adma.202001626.

[65] Kameel Abdel-Latif, Robert W. Epps, Fazel Bateni, Suyong Han, Kristofer G. Reyes, and Milad Abolhasani. Self-Driven Multistep Quantum Dot Synthesis Enabled by Autonomous Robotic Experimentation in Flow. *Advanced Intelligent Systems*, 3(2):2000245, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aisy.202000245.

[66] Jorge Chang, Pavel Nikolaev, Jennifer Carpena-Núñez, Rahul Rao, Kevin Decker, Ahmad E. Islam, Jiseob Kim, Mark A. Pitt, Jay I. Myung, and Benji Maruyama. Efficient Closed-loop Maximization of Carbon Nanotube Growth Rate using Bayesian Optimization. *Scientific Reports*, 10(1):1–9, 2020.

[67] Nicholas A. Jose, Mikhail Kovalev, Eric Bradford, Artur M. Schweidtmann, Hua Chun Zeng, and Alexei A. Lapkin. Pushing nanomaterials up to the kilogram scale – An accelerated approach for synthesizing antimicrobial ZnO with high shear reactors, machine learning and high-throughput analysis. *Chemical Engineering Journal*, 426(July):131345, 2021. Publisher: Elsevier B.V.

[68] Nicholas A. Jose, Hua Chun Zeng, and Alexei A. Lapkin. Scalable and precise synthesis of two-dimensional metal organic framework nanosheets in a high shear annular microreactor. *Chemical Engineering Journal*, 388(January):124133, 2020. Publisher: Elsevier.

[69] R. Shukre, T. Ericson, D. Unruh, H. Harbin, A. Cozzolino, Chau-chyun Chen, and Siva Vanapalli. Batch-Screening Guided Continuous Flow Synthesis of the Metal-organic 2 Framework HKUST-1 in a Millifluidic Droplet Reactor. *ChemRxiv*, pages 1–18, 2022.

[70] Patrick A. Julien, Cristina Mottillo, and Tomislav Friščić. Metal-organic frameworks meet scalable and sustainable synthesis. *Green Chemistry*, 19(12):2729–2747, 2017.

[71] Yunchao Xie, Chi Zhang, Heng Deng, Bujingda Zheng, Jheng Wun Su, Kenyon Shutt, and Jian Lin. Accelerate Synthesis of Metal-Organic Frameworks by a Robotic Platform and Bayesian Optimization. *ACS Applied Materials and Interfaces*, 13(45):53485–53491, 2021.

[72] Miao Wu, Lingling Xia, Yucen Li, Difeng Yin, Jianping Yu, Wenbo Li, Ning Wang, Xin Li, Jiwei Cui, Wei Chu, Ya Cheng, and Ming Hu. Automated and remote synthesis of poly(ethylene glycol)-mineralized ZIF-8 composite particles via a synthesizer assisted by femtosecond laser micromachining. *Chinese Chemical Letters*, 33(1):497–500, 2022. Publisher: Elsevier B.V.

[73] Matthew J. McDermott, Shyam S. Dwaraknath, and Kristin A. Persson. A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis. *Nat Commun*, 12(1):3097, May 2021. Number: 1 Publisher: Nature Publishing Group.

[74] Jiaru Bai, Liwei Cao, Sebastian Mosbach, Jethro Akroyd, Alexei A. Lapkin, and Markus Kraft. From Platform to Knowledge Graph: Evolution of Laboratory Automation. *JACS Au*, 2(2):292–309, February 2022. Publisher: American Chemical Society.

[75] Yuan An, Jane Greenberg, Xintong Zhao, Xiaohua Hu, Scott McCLellan, Alex Kalinowski, Fernando J. Uribe-Romo, Kyle Langlois, Jacob Furst, Diego A. Gómez-Gualdrón, Fernando Fajardo-Rojas, and Katherine Ardila. Building Open Knowledge Graph for Metal-Organic Frameworks (MOF-KG): Challenges and Case Studies, July 2022. arXiv:2207.04502 [cs].

Chapter 5

DigiMOF: A Database of Metal-Organic Framework Synthesis Information Generated via Text Mining

5.1 Publication Information and Paper Contributions

This paper has been published as an article in the American Chemical Society's journal Chemistry of Materials.

In this publication I, the candidate, wrote the manuscript with equal contributions from Kristian Gubsch, Rosalee Bence and contributions from Rama Oktavian, Kesler Isoko, and Seyed Mohamad Moosavi, under the supervision of Professor Joan L. Cordiner, Dr Jason C. Cole, and Dr Peyman Z. Moghadam.

5.2 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 (ML), 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 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. Additionally, we developed an alternative data extraction technique to obtain and transform the chemical names assigned to each CSD structure in order to determine linker types for each structure in the CSD MOF subset. This data enabled us to match MOFs to a list of known linkers provided by

Tokyo Chemical Industry UK Ltd. (TCI) and analyse the cost of these important chemicals. This centralised, structured database reveals the MOF synthetic data embedded within thousands of MOF publications and contains further topology, metal type, accessible surface area (ASA), largest cavity diameter (LCD), pore limiting diameter (PLD), open metal sites (OMS), and density calculations for all 3D MOFs in the CSD MOF subset. The DigiMOF database and associated software are publicly available for other researchers to rapidly search for MOFs with specific properties, conduct further analysis of alternative MOF production pathways and create additional parsers to search for other desirable properties.

5.2.1 Keywords

Biological databases, Chemical synthesis, Mathematical methods, Metal-organic frameworks, Metals

5.3 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 high surface areas and exceptionally tunable properties, which enable their potential application in areas including gas storage [1, 2, 3, 4, 5, 6], sensing [7, 8, 9, 10], separations [11, 12, 13, 14, 15], drug delivery [16, 17, 18], and catalysis [19, 20, 21, 22, 23]. Since the first MOFs were synthesized in the 1990s, thousands of MOFs have been produced at a laboratory scale. As of 2023, more than 100,000 MOF structures have been reported in the Cambridge Structural Database (CSD) [24, 25]. 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 (ML) for the elucidation of structure–property relationships, in particular for gas adsorption and separation properties of MOFs [26, 27, 28, 29, 30, 31, 32]. 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. While ML has so far been successfully applied to MOF synthesis using failed experimental data [33], 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 valorizing 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 [34]. To date, CDE has been deployed to automatically assemble databases of magnetic materials [35, 36], battery materials [37], UV/vis absorption spectra [38], hydrogen storage and synthesis applications [39], and nanomaterial synthesis [40, 41, 42]. While CDE has been used to text mine both organic and inorganic chemistry literatures, 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 [43]. More recently, Luo et al. [44] developed an automatic

data mining tool using the CoRE MOF database [45], alongside the web-scraping tool Puppeteer (https://pptr.dev) to text mine 6099 journal articles. These were then analyzed using ChemicalTagger software [46] to extract metal sources, linker(s), solvent(s), additives, synthesis time, and temperature. A further recent submission from Park et al. data mined 46,701 MOFs to extract synthesis information from 28,565 papers using a joint ML/rule-based algorithm [47].

The CSD MOF subset contains comprehensive structural information about MOFs; however, the data related to their synthesis is scarce and inconsistent. Here, we text-mined the CSD MOF subset and 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 ML 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.

5.4 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 the 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. [48] 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 costs in comparison to using a less solvent-intensive pathway. In another study, Luo et al. [49] compared traditional solvothermal synthesis with an aqueous pathway to produce UiO-66-NH2 and found that omitting solvents from the synthesis of this MOF resulted in an 84% reduction in production cost. The key properties that influenced the production cost were solvents, organic linkers, 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 pelletization and industrial application of MOFs [50]. 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. The CSD Python API was also used to extract the chemical names that corresponded to each MOF refcode in the 3D MOF subset for linker matching.

5.5 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 5.1 demonstrates how the DigiMOF database and the adapted CDE parsers can be integrated into a data-driven pipeline for MOF design and discovery.

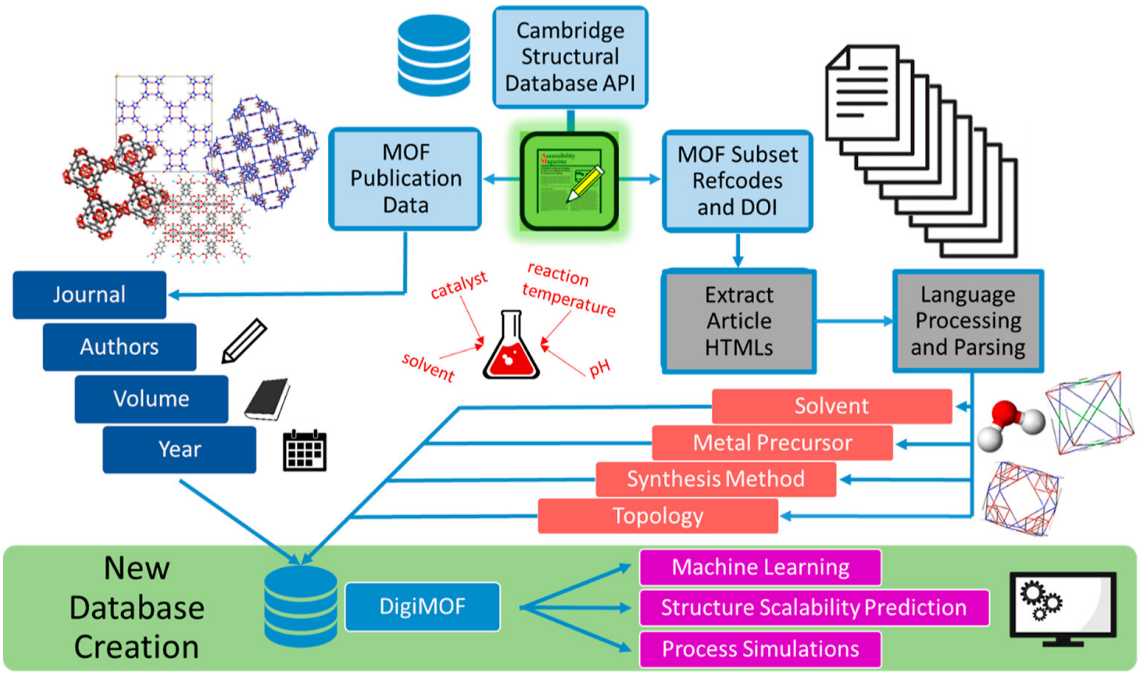


Figure 5.1: *Flow diagram to visualize the integration of CDE into a data-driven MOF synthesis plan: from article retrieval to text mining, computational screening, and materials discovery.*

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 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 data sets for further processing.

Note that a user can create new and personalized databases for text mining by modifying the provided CDE web scraping script to obtain any collection of online files saved into HTML format, i.e., patents, webpages, and journal articles, from other sources.

5.5.1 Natural Language Processing

To identify specific MOF properties using CDE-based classes and variables, we created customized parsers which use part-of-speech (POS) taggers and chemical entity recognizers. These parsers contain specific regular expressions for the identification of MOF compound names. The natural language processing (NLP) pipeline in CDE first iden-

tifies a sentence, which is then tokenized into individual words and punctuation known as tokens [34]. These tokens are marked up by POS tagging to reflect their syntactical functions, such as a noun, a verb, a chemical mention, and an adjective [34]. Entity recognition of the chemical species allows relationships to be extracted and merged with their corresponding compounds by interdependency resolution [34]. Our rule-based parsers used 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 created exclusion lists to exclude words which were frequently misidentified as these variables. The use of regular expressions and parsing elements, as shown in Table A.1, was crucial to improving performance.

The process of building and refining the parsers is shown in Figure 5.2 following a similar process used by Huang and Cole [51]. 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.

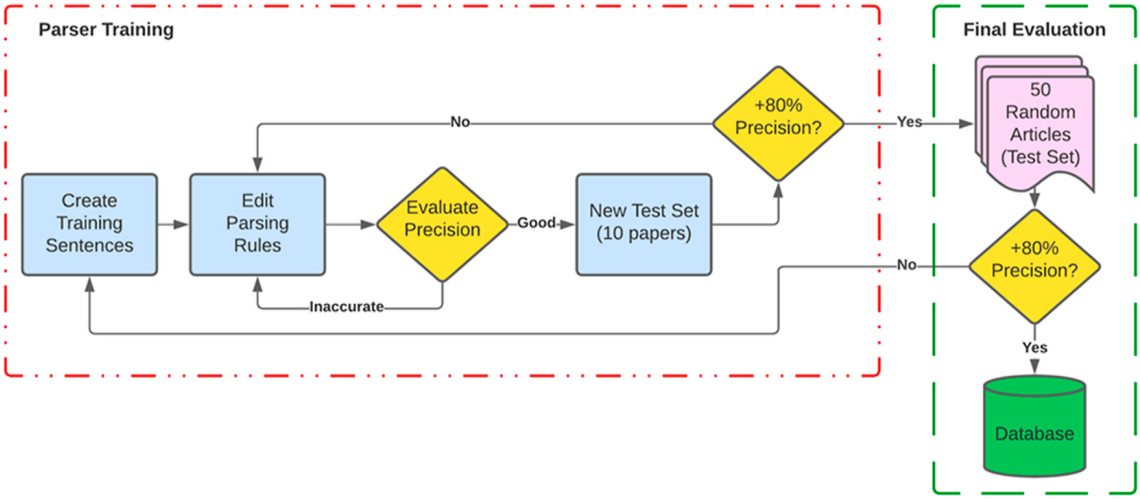


Figure 5.2: *Flow diagram to visualize the integration of CDE into a data-driven MOF synthesis plan: from article retrieval to text mining, computational screening, and materials discovery.*

5.5.2 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 Eqs. 5.1–5.3, 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.

$$Precision = \frac{TP}{TP + FP} \tag{5.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

$$F - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5.3}$$

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 prioritize 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 improve recall, but they also show a reduction in specificity, which reduces precision. Generally, high precision should be given precedence over recall; low recall is acceptable provided that a large enough data set 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 using CDE are shown in Table A.2. 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 maximizing precision, extracting common and unambiguous sentences observed in MOF literature was prioritized, although it was expected that lower recall would be obtained compared to previous iterations of CDE. Figure A.1 summarizes the overall performance of our parsers compared to previous CDE projects and the MOF text mining tool from Park et al. [47] 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 [51]. A breakdown of individual parser results for the synthesis route, topology, linkers, and metal precursors can be found in Table A.3.

5.5.3 Parser Training

During parser training, precision was substantially improved by employing exclusion lists to filter out frequently observed misidentifications. The addition of common abbreviations, names, and exclusion list items for metal precursors, linkers, MOFs, and topologies to the regular expressions helped to improve both precision and recall. As MOF terminology and literature are 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 exclusion lists and compound regular expressions, which will allow the tool to evolve and improve over time.

Figure 5.3 shows the process for the selection of regular expressions that can be incorporated into CDE. Here, we demonstrate how regular expressions (regex) may be developed iteratively to achieve more TPs and eliminate FPs and negatives. Table A.4 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.

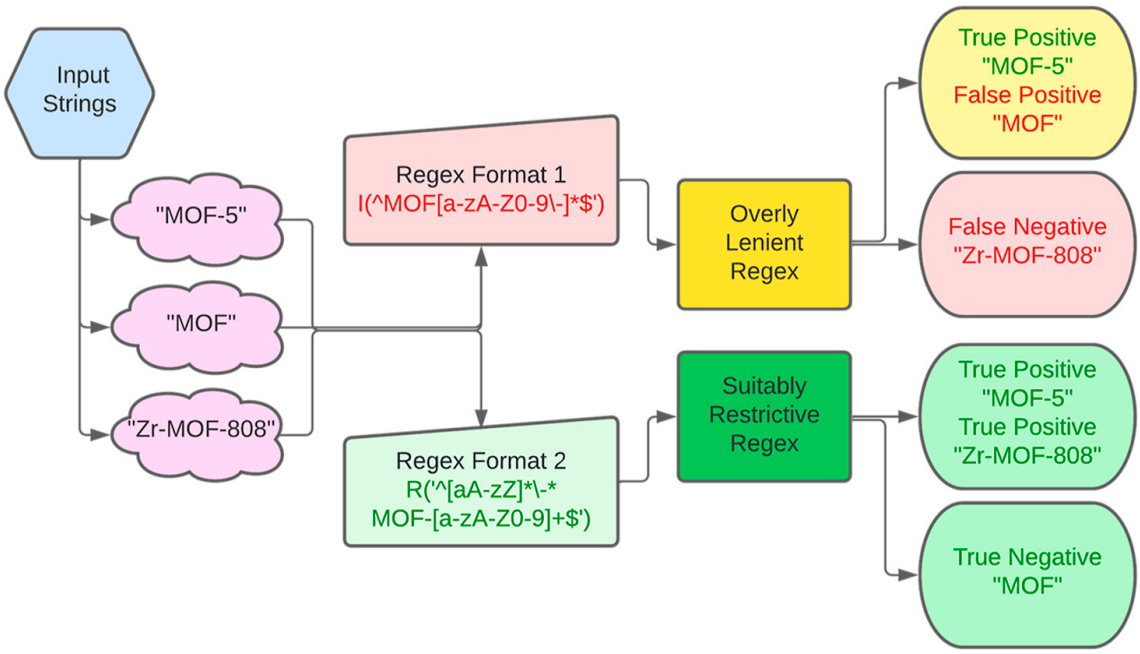


Figure 5.3: *Flow chart displaying possible outcomes when fed an input string for high-throughput MOF name parsing.*

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 FPs, as demonstrated by expression number 4 in Table A.4 which is the lenient regular expression for common linker abbreviations. To accommodate a wider variety of sentence structures to help recognize MOF names, an exclusion list was integrated into the regular expression rules to exclude FPs, as with expression 9 in Table A.4. Regular expressions within the context of exclusion listing are further detailed in the Supporting Information in Table A.5.

5.5.4 Obtaining Metal, Topology, and Linker Data

After parsing was complete, to obtain further, more detailed information surrounding the metal elements contained with each MOF, we used a high-throughput approach that involved obtaining the relevant crystallographic information files (CIFs) for use in the MOFid software suite [52]. Each CIF was entered into the program where it was then deconstructed, and the metals present in the MOF were extracted. For topological representations of these structures, we used the Julia-based CrystalNets [53] program to

automatically assign network topologies to all CIFs. This enabled the comparison of algorithmically assigned values from these software packages with the text-mined data for verification purposes.

Obtaining linker information proved to be more challenging. We created "rules" in the CSD Python API to extract linker names which enabled the simplification of CSD's long text-based chemical names into distinct repeating units. For example, the chemical name for SAHYIK within the CSD is *"catena-(tris(4-1,4-Benzenedicarboxylato)-(4-oxo)-tetrazinc octakis(dimethylformamide) chlorobenzene clathrate)"*. These names were initially treated by extracting the metal names, in this case zinc, and adding them to the list of metals for each structure. Then, the remaining text is split based upon the names which succeed , indicating that there are repeating units; the remaining non-chemical items such as *"catena-"* and *"tris"* are also discarded here. These repeating units are then transformed to match the chemical names found in the list provided by TCI Chemicals [54] for common MOF linkers. For this first entry, e.g., "1,4-benzenedicarboxylato" is modified to "1,4-benzenedicarboxylate", which can also be represented by its alias terephthalic acid and is then matched to the TCI Chemicals list. The second corresponds to the string "oxo", which is discarded as it refers to the repeating oxygen molecules in the zinc oxide node. Anything that succeeds the metal in the chemical name and is separated by a space is removed and retained for further processing as possible solvents used in the synthesis. Figure 5.4 shows the outcome of this process for the 30 most frequently extracted records taken from a list of 149 unique chemical names and matched after both a manual and an automatic transformation process were performed. The matching list, which includes linker synonyms and chemical prices, can be found in the Supporting Information TCI_Chemicals (XLS) document.

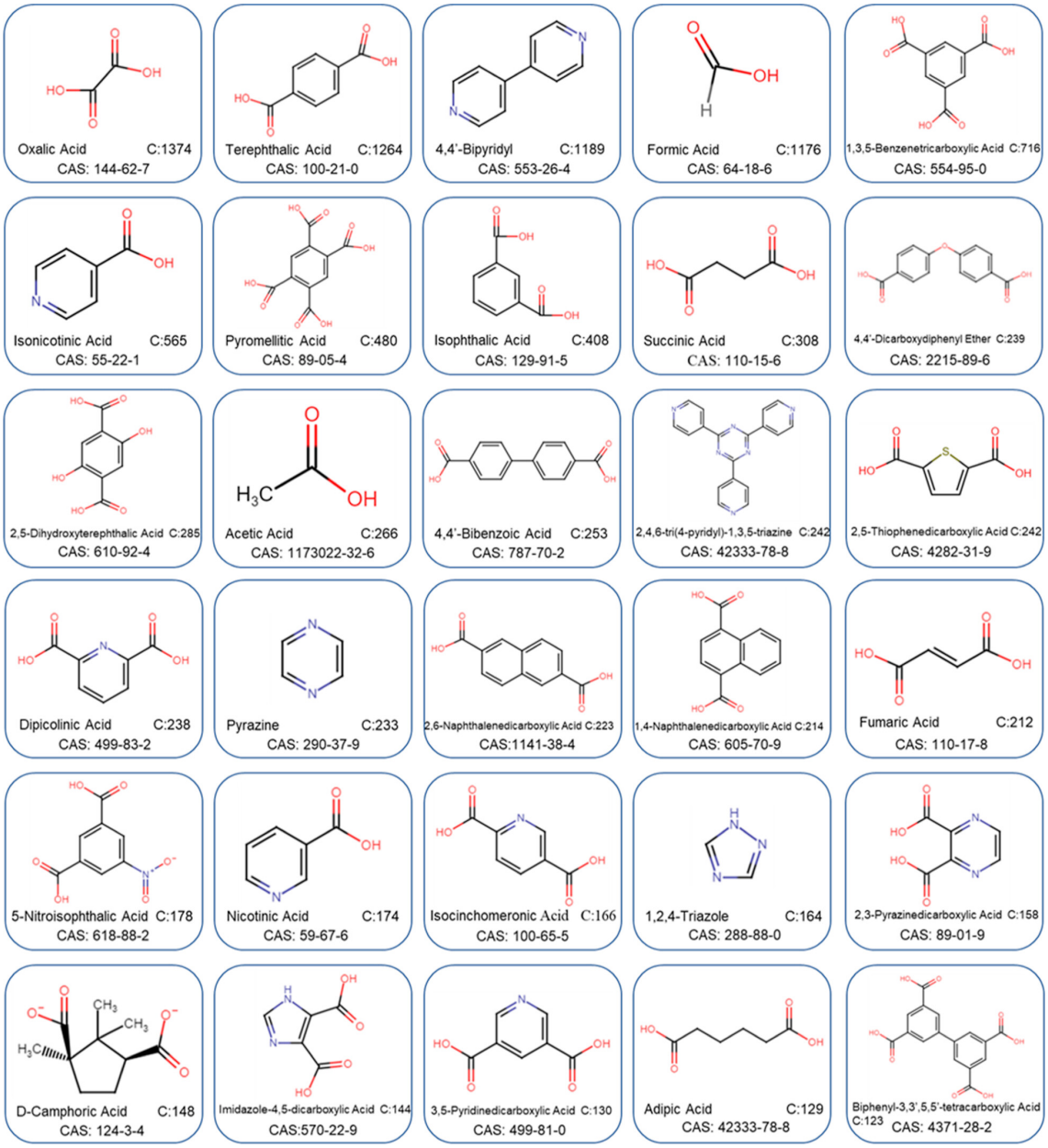


Figure 5.4: *Collection of the top 30 organic linkers obtained via text-mining the CSD MOF subset chemical names. Hit counts (C) and CAS numbers are included for each linker.*

5.5.5 Geometric Properties

By analyzing the text-mined data, correlations between different MOF topologies and structural properties were unveiled by determining a complete set of geometric properties and investigating the patterns which emerged from known and unknown relationships. The largest cavity diameter (LCD), pore limiting diameter (PLD), accessible surface area (ASA), frameworks density, the presence of open metal sites, and void fraction of all 3D MOFs in the subset were calculated using Zeo++ software [55] to quantitatively characterize their structural properties. A probe radius of 1.86 Å, corresponding to the kinetic radius of N_2, was applied for ASA calculations. The results of these calculations can be found in the DigiMOF3DSubset (CSV) document of the Supporting Information.

5.6 Results and Discussion

We note that for a MOF compound name and the 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 5.1 displays the total number of each type of synthesis property associated with 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 5.1: *Total Number of Extracted Properties and the 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	9705	8
solvents	1211	81
topologies	6680	154
linkers	24,166	10,690
metals including ions	10,968	1803
metals excluding ions and element names	5163	1476

The DigiMOF database contains a MOF compound name and corresponding topology, organic linkers, metal precursors, synthesis methods, 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 labeled 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 summarize our key findings after text mining the CSD MOF subset.

To enrich the database of 10,696 3D structures extracted with CDE, we also gathered additional information using alternative computational methods as detailed in Section 5.5.4. Table 5.2 shows a breakdown of the parameters we extracted and calculated to supplement the text-mined data set. A total of 24,784 3D MOFs were admitted to the calculation stages, where the constituent metal was identified for 23,832 structures, and either an RCSR or EPINET topology was assigned for 13,816 3D structures.

Table 5.2: *Total Number of Extracted Properties and Number of Unique Properties for Structures in the 3D MOF Subset*

property	total extracted	total unique properties extracted
MOF compound names	24,784	
topologies	13,816	460
linkers	15,901	129
elemental metals	23,832	716
LCD and PLD	22,104	
density	24,587	
open metal sites	763	
geometric properties	>6474	

Here, we note that despite obtaining 10,690 unique linker names in the text mining stage for journal articles, once we take the more uniform CSD chemical names and match synonymous chemicals together, we collected information for at least one linker type for 40% of materials that have suitable chemical names for the matching process. The complete data for linker names, metals, and topologies can be found in the Supporting Information DigiMOF3D subset (CSV).

5.7 Data Analysis

5.7.1 Synthesis Methods

When analyzing the data for synthesis methods, we first investigated how synthesis methods have changed over time. A total of 9705 synthesis route records were extracted from 43,281 papers. Figure 5.5 shows the cumulative sum of records extracted for various types of synthesis routes 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 a 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 laboratory-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 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 A.7a also shows that hydrothermal synthesis was the most common alternative/low-solvent synthesis route extracted by the parsers.

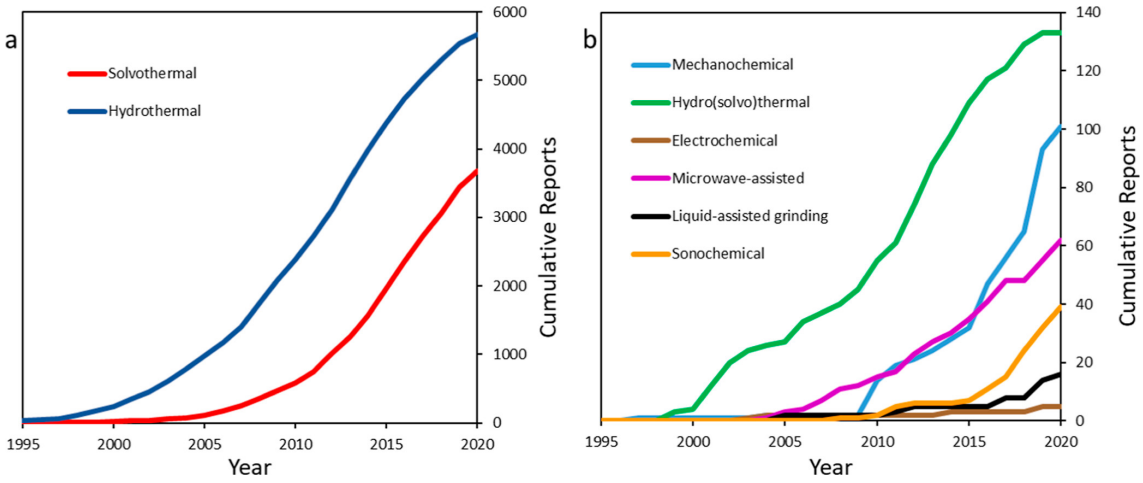


Figure 5.5: *(a) Cumulative sum of the two main MOF synthesis methods from 1995 to 2020. (b) Cumulative sum of alternative and emerging synthesis methods showing periods where these techniques were first introduced for MOF synthesis.*

We also note that the majority of 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, 6033 (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 cross-over 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 utilizing greener synthesis methods is also reflected in innovative MOF commercialization efforts such as the ton-scale water-based processes that BASF has developed [56] and the mechanochemical process from MOF Technologies [57].

The DigiMOF database allows users to search for potentially scalable MOFs via the synthesis method to discover MOFs that can be more easily synthesized 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.

5.7.2 Topology

Topological characterization of MOFs is important as it can constrain key structural properties such as pore shape, size, and chemistry, and it is directly related to mechanical

stability [50]. Figure 5.6a. 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 5.6a, **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 also performed topological characterization of the 3D MOF subset using CrystalNets [53] and achieved a return of 55.8% across 460 unique topologies. We note here that the CrystalNets calculations allowed for the extraction of topological types that matched the EPINET [58] database, whereas our text-mining approach was specifically developed to seek out RCSR-type topologies [59]. Figure 5.6b shows the occurrence of the top ten topological nets with **pcu** as the most frequently occurring topology, followed solely by 3D representations in **dia**, **pts**, **rtl**, and **cds** rounding out the top five. Figure 5.6c shows examples of commonly occurring 3D underlying nets. An additional outcome of this study was that 2375 structures in the 3D MOF subset were built from two or more interpenetrating nets. We anticipate that this topological characterization of MOFs will also guide future efforts to identify mechanically stable MOFs.

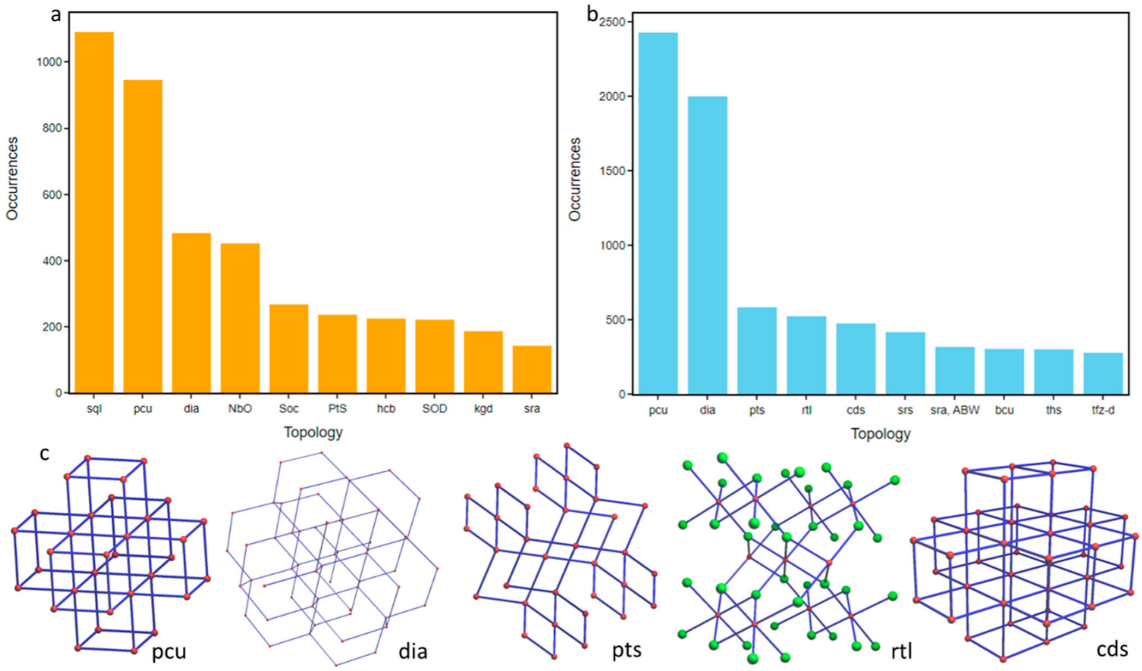


Figure 5.6: *Histograms of topological types extracted from the CSD MOF subset using (a) ChemDataExtractor (CDE) (b) CrystalNets in 3D structures. (c) Top five most common 3D topologies:* ***pcu****,* ***dia****,* ***pts****,* ***rtl****, and* ***cds****.*

We used the topological data to investigate the topology–structure relationships for certain geometric properties. Figure 5.7 shows the different regions that are occupied by a selection of five topological types. For some representations, there does not appear to be any restriction on the types of pores that can be formed with a wide variety of void fractions seen for **pcu** and **dia**. Both representations span a range of void fractions between 0 and 0.85 across and the LCD range of 3.7–15 Å. On the contrary, there are some slightly more distinct linear patterns between the LCD and the void fraction for

other representations, which are particularly noticeable for **stp** and **rob**. The former shows a distinct linear pattern within the region of 5 to 10 Å and 0.2 to 0.35 void fraction and displays a similar linearity into the 15 to 20 Å range.

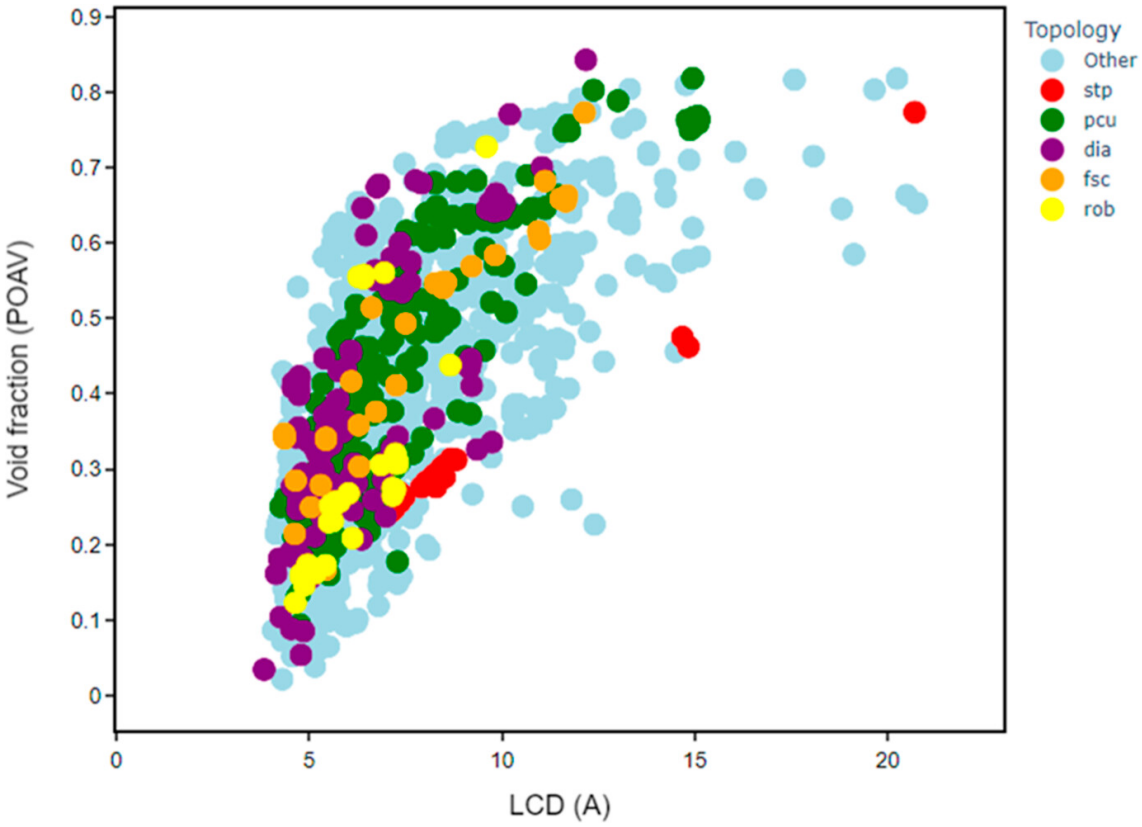


Figure 5.7: *Comparison of different topologies in the structure space for LCD as a function of void fraction for ca. 2200 porous MOFs. There are 241 structures with* ***pcu*** *topology (green); 170* ***dia*** *(purple), 41* ***stp*** *(red), 33* ***rob*** *(yellow), and 32* ***fsc*** *(orange) structures. All other structures are shown in pale blue.*

5.7.3 Solvent

Dimethylformamide (DMF) is the most frequently extracted solvent, representing 469 of the 1211 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/H2O". The additional top hits for solvent extraction can be seen in Figure A.7c.

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 commercialization. It should be noted that while 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.

5.7.4 Organic Linkers

Histograms in Figure A.7d show that carboxylate-type linkers were the most frequently extracted type of organic linkers, 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 [60, 61]. While 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 transformation 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 transformations were 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 does not refer to a specific chemical structure; instead, it is a generic label used within the MOF literature to refer to a number of organic linkers [62]. 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 a 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 toward more complex and functionalized linkers in MOF synthesis, which may make consistent identification and naming of these structures more challenging [63]. The chemical diversity of MOF linkers is an important factor, particularly when considering the application of ML on these data sets [64].

To combat this ambiguity, we developed a new approach to text extracting MOF linker names using the chemical names found in the CSD, as these are available for over 99% of all deposited structures. The result of this text mining required some manual intervention as CSD chemicals can have different naming protocols; for example, one might find 1,3,5-benzenetricarboxylate or the synonymous benzene-1,3,5-tricarboxylate both used within this data set. There are in fact tens of examples of similar synonymous chemical names being used across the 149 linker names we used as our match list. Overall, this new method had a significantly higher accuracy given the strict designations for similar linker molecules. For example, the distinction between 4,4-bipyridine and 2,2-bipyridine, when compared to the CDE text mining results, avoids the need to note "generic bipyridine-type linkers" and enables deeper analysis of similarly named but chirally different molecules. Figure 5.8a shows the frequency at which a linker type was reported for structures that

contained reference to only a single linker but also had a non-zero cavity diameter. This data was then used to separate linkers depending on their length, which was determined by the number of consecutive blocks, e.g., number of benzene or pyridyl rings, into categories of 1 or 2+ blocks. The difference between the linker length and their respective MOF LCD ranges is shown in Figure 5.8b. We note here that the longer 2+ block linkers have a larger LCD range from 1.3 to 12.3 Å, whereas shorter one-block linkers span a slightly smaller range of LCD values from just above 0 to 8.5 Å. Interestingly, despite the mean LCD following the pattern of increasing with linker length, there are several one-length linker structures that far exceed the average LCD of MOFs built with two or greater length linkers. Once the linkers had been categorized with respect to their length, it was possible to investigate the pore morphology, as shown in Figure 5.8c, a box and whisker plot of linker length against the LCD/PLD ratio. The results here suggest that shorter linkers with one block can generate structures with a wide range of LCD/PLD ratios, whereas longer linkers containing 2+ blocks generate structures on lower ranges of LCD/PLD ratios of <2.5: a finding which is dominantly due to larger PLD values in these structures.

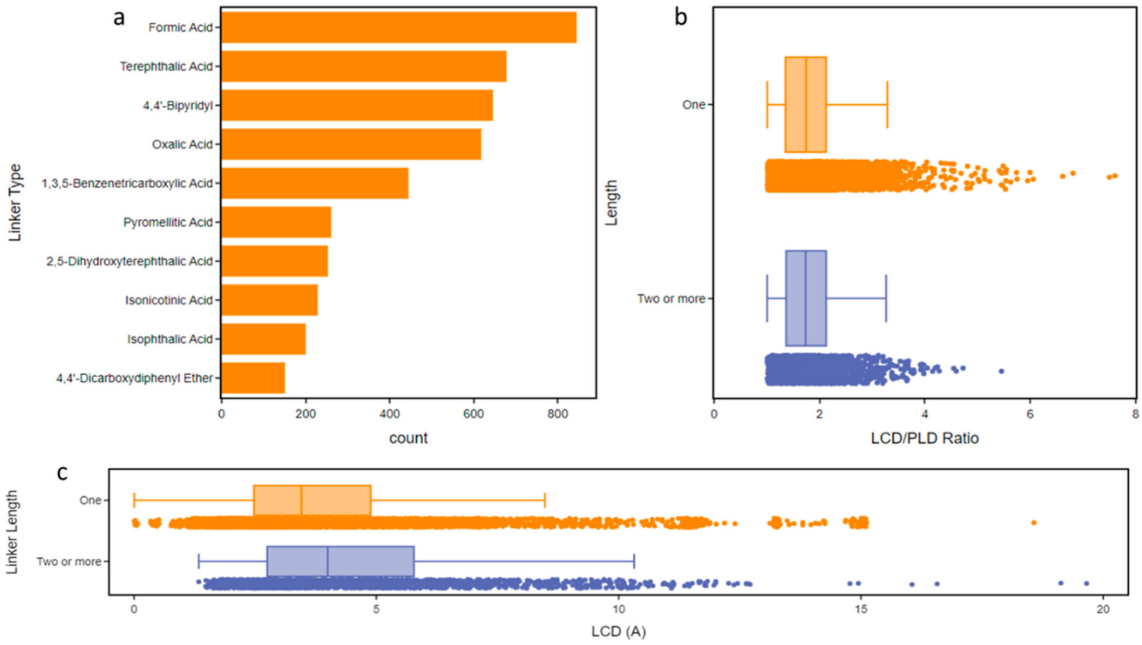


Figure 5.8: *(a) Histogram showing the most commonly occurring single linkers found in the 3D MOF subset for non-zero LCD values. (b) Box and whisker plot of linker length versus the LCD/PLD ratio across a sample of ca. 8000 MOFs. (c) Box and whisker plot of linker types against LCD for a sample of linkers with one (orange) and two or more (blue) blocks.*

5.7.5 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 A.7e. Zinc-based precursors were most frequently extracted, with "Zn(NO3)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, the 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 A.3. Additionally, compared to zinc and copper-based MOFs, Zr-based MOFs were not widely produced until after 2012 [65]. The second most frequently extracted metal salt was "Cd(NO3)2·4H2O" 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 precursors 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.

Figure 5.9a shows the most frequently occurring single metal types in the MOF subset as identified using MOFid [52]. Figure 5.9b shows the relationship between metal types and the typical LCD values expected for each MOF containing that metal. The most common metal, Zn, contains over 1200 entries in the database, for which 752 or 60% can be considered porous such that they have an LCD which exceeds the probe diameter of 3.7 Å. For Co and Cd, this ratio decreases to 51 and 41%, respectively. The lowest proportion of porous MOFs from the metals can be found for structures containing Na, where only 34% of entries have LCDs greater than 3.7 Å. Na-containing MOFs have the lowest mean LCD of all metals at 1.5 Å, whereas Cu-MOFs have the highest at 4.6 Å, with the average LCD across all metals sitting at 3.5 Å.

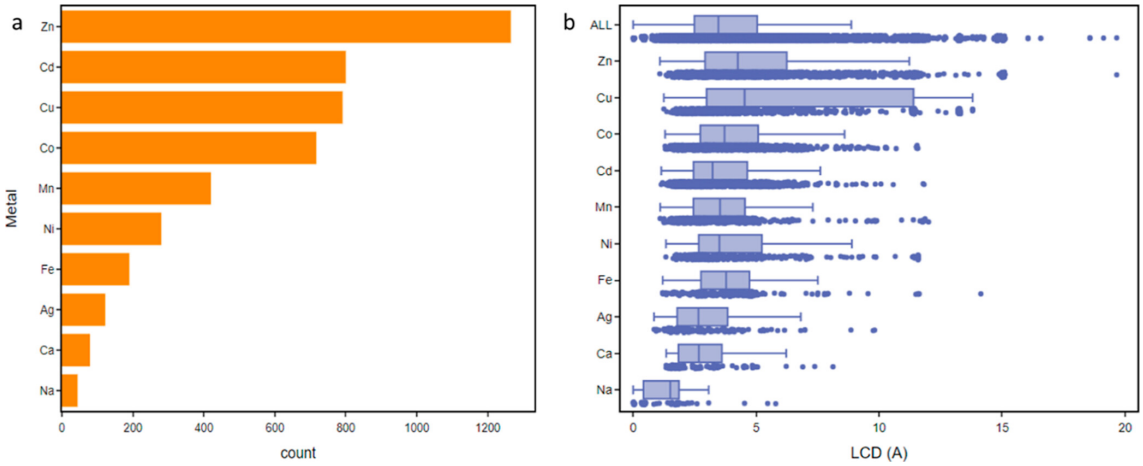


Figure 5.9: *(a) Histogram of the most frequently occurring single metals found in the 3D MOF subset. (b) Comparison of the constituent metals against the LCD of structures.*

5.7.6 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 without error. The results of these extractions, which have been rounded to the nearest whole degree Kelvin, can be seen in Figure A.7f it is important to note that these values are not the synthesis temperatures of the materials but are of the variable-temperature crystallographic studies. These are the temperatures used

in post-synthesis investigations at which the results of certain experimental procedures in each manuscript have been reported, specific to each material. This data does not guarantee the stability of MOFs at these temperatures. Typically, an experimental structure is tested and reported at or around room temperature, explaining the spike in records at 293 K. It is also common that a Cryostream or other device is used to cool a sample for low-temperature crystallographic testing. We would recommend the introduction of more useful temperature data fields, such as activation temperature, destabilization temperature, or solvent/synthesis/reaction temperature, alongside the existing crystallographic study temperatures.

5.7.7 Building Blocks and Topology

The underlying networks of the extracted MOF structures can be investigated using insight gained from the data presented in Figure A.4. There are 4972 linker hits for which there was a corresponding topology and a further 1424 results for metal clusters. Taking into consideration the top five most frequently parsed linkers and metal precursors from Figure A.7d,e respectively, we can deduce the top five topologies for each MOF building block. These results are represented in a clustered column graph, Figure A.4. Furthermore, additional data obtained via CrystalNets [53] has offered insight into the topological configuration of 3D MOFs in the DigiMOF database, with a return rate of 55%. A filter can be applied to this data set to select all matched linker types for a given topology.

The top linker type extracted using CDE, ["carboxylate"] corresponded to a total of 100 topologies, the most frequent being **sql** (12), and **pcu** (12). These two topological types emerged as the most frequent for almost all investigated linkers and metal clusters, an unsurprising result considering the high frequency of these two representations across the whole study. These are two of the simplest underlying structure representations, which may explain their abundance; more complex structures are less likely to have topology reports due to potential errors, and additionally, it is common to report the most simplified underlying net even where a more complex representation exists. For the 3D data set, the highest linker type ["oxalic acid"] corresponded to a total of 66 unique topologies, with the most frequent being **dia** (84), followed by **pcu** (50).

In 2014, a study by Cai et al. investigated the crystal structures of derivatives of HKUST-1, which notes that for H-BTC (the 5th most common linker type), the predicted topological type is **tbo**; however, variations in the functionalization of this same linker can give rise to a preference for **fmj** connectivity using the same building blocks [66].

Perhaps more interesting than the results for linkers is that of metal clusters; typically, linkers are connected only at each edge, although in some less common cases (e.g., where linkers consist of porphyrins and derivatives), there can be a higher number of connections. Depending on the coordination of certain metal clusters, it can be impossible to achieve some topological types, making the choice of the metal cluster more restrictive than the choice of the linker; a significant influence on the potential underlying network of a crystal structure. From these metal cluster results, we can deduce that transition-metal nitrate structures form some of the simplest underlying nets with **sql**, **pcu**, **dia**, and **kgd** being frequently reported in synthesis papers. This variety of 2 and 3 dimensional, and 4-connected, 6-connected, and 6 plus 3-connected clusters suggests flexibility in the coordination number of these transition-metal building blocks.

Further to this point, it is worth noting the influence of temperature on the dimensionality of MOF structures. Reaction temperature has been found to have a remarkable influence on the formation and structure of MOFs, especially toward the control of topology [67]. Increasing the hydro/solvothermal reaction temperature has the potential to increase the coordination number of the central metal ion [68]. Anderson et al. suggested that a temperature-dependent quantity such as free energy, which would have a notable influence toward the topological selectivity of MOF synthesis, should be considered in MOF synthetic accessibility predictions [69].

5.7.8 Cost Analysis

As a result of improving the accuracy in linker designation from Section 5.4, and from the use of a matching list modified from the publicly available TCI Chemical list, it was possible to add an approximate linker cost analysis to our data set, given the availability of pricing data for these chemicals [54]. We took the TCI Chemicals list and added several other commonly used organic materials, followed by the inclusion of live online prices, these costs are typically for quantities of 99%+ purity precursor chemicals. Due to the inclusion of additional listings, it was necessary to obtain some missing cost values from Sigma-Aldrich to get a complete list of approximate linker "raw chemical" costs [70]. The available quantities varied between all linker types, and so the prices in this list were determined by taking into consideration all of the possible prices and finding the mean cost per gram. Figure 5.10 shows the results of the linker cost analysis on some of the most prevalent linkers detailed in Section 5.5.4. As the structures obtained from the CSD MOF subset are experimental, we expected to see most of the structures containing lower cost linkers for the simple reason that they would be more economical to produce. While the range of linker costs across the chemical list spans £0.05 to £830 per gram, out of the top 45 linkers, 40 of them had a cost per gram under £10, as can be seen in Figure 5.10a. This sample of linkers in the "low-cost" range spans a total of 6643 structures. Figure 5.10b also shows a total of 33 linkers that exceed a cost of £10 per gram, although they make up a much smaller proportion of the total structures that have been identified as linkers in this study.

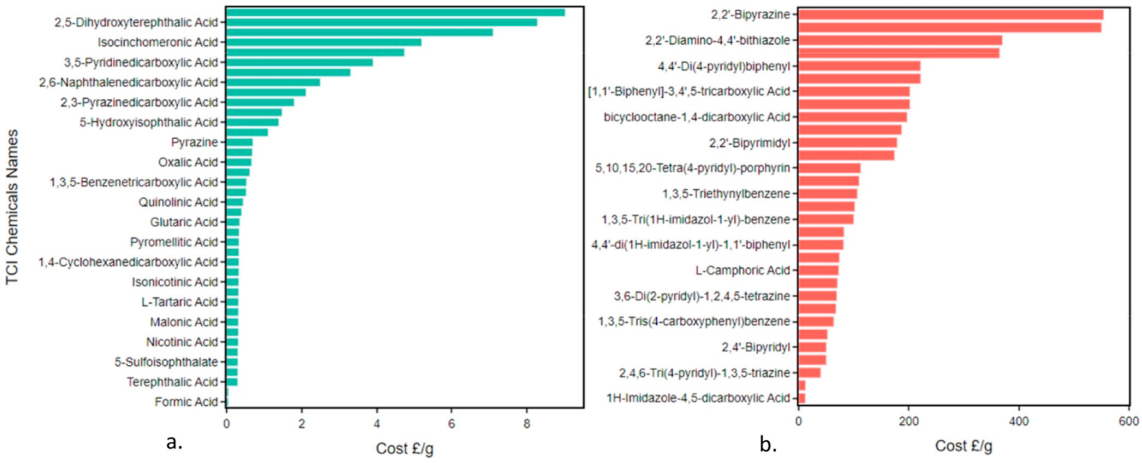


Figure 5.10: *Bar charts showing the cost per gram of organic linkers as determined by averaging the available quantities. A selection of the most prevalent linker types was chosen from the DigiMOF database for (a) low-cost and (b) high-cost linkers. Prices obtained from TCI Chemicals [54].*

The results of this cost analysis can be used to select specific linker types for techno-economic assessment in conjunction with limiting solvent quantities, finding optimal reaction temperatures, selecting suitable catalysts, and selecting low-cost metals. The cost per mole of each linker type can also be found in the Supporting Information document, TCI_Chemicals (XLS).

5.8 Conclusions and Future Directions

To the best of our knowledge, the DigiMOF database is the first automatically generated database of MOF synthesis properties using ChemDataExtractor to text-mine 43,281 MOF publications. After an iterative training process, the parsers yielded an overall precision of 77% to extract 52,680 associated MOF synthesis properties. This initial text-mined data was supplemented with additional data mined from the CSD MOF subset, which enabled the identification of linker types and their corresponding costs. DigiMOF will allow researchers to search for key properties related to implementing large-scale MOF production, e.g., synthesis routes and solvents, organic linkers, metal precursors, structure topology, constituent metals, and linker cost. 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 to allow researchers to update our regular expressions as new compounds emerge, ensuring these algorithms can continue to identify new MOF-property relationships. 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 by making 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 primarily focused on the production of MOF compounds but also includes basic geometric properties to offer an additional level of insight. 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 timeall essential parameters for enhancing MOF synthesis pathways. We also recommend that future MOF synthesis publications contain specifically formatted tables of key information as an appendix to the article, presented in a way that is friendly to text mining algorithms to enable the scraping of data using a high-throughput screening approach, improving both the precision and recall of any chemical journal parser. By improving the precision and recall of structure property parsing beyond the levels we see today, there is the potential to enable an accurate and reliable database of synthesis data to be created in the public domain that can be continually and accurately updated following new publications.

We envisage that this work will lay the foundation for enabling digital manufacturing of MOFs and facilitate the 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 decarbonization technologies that motivate many contemporary MOF research proposals. With thousands of entries for each parameter parsed in this study, DigiMOF augments MOF scientists'

expertise, allowing them to design more efficient MOF discovery pathways and advance the synthesis of these fascinating materials.

References

[1] Bin Li, Hui-Min Wen, Wei Zhou, Jeff Q. Xu, and Banglin Chen. Porous Metal-Organic Frameworks: Promising Materials for Methane Storage. *Chem*, 1(4):557–580, October 2016.

[2] Omar K. Farha, A. Özgür Yazaydın, Ibrahim Eryazici, Christos D. Malliakas, Brad G. Hauser, Mercouri G. Kanatzidis, SonBinh T. Nguyen, Randall Q. Snurr, and Joseph T. Hupp. De novo synthesis of a metal-organic framework material featuring ultrahigh surface area and gas storage capacities. *Nat Chem*, 2(11):944–948, November 2010.

[3] Shengqian Ma and Hong-Cai Zhou. Gas storage in porous metal–organic frameworks for clean energy applications. *Chem. Commun.*, 46(1):44–53, January 2010. Publisher: The Royal Society of Chemistry.

[4] Jarad A. Mason, Mike Veenstra, and Jeffrey R. Long. Evaluating metal–organic frameworks for natural gas storage. *Chem. Sci.*, 5(1):32–51, November 2013. Publisher: The Royal Society of Chemistry.

[5] Eyas Mahmoud, Labeeb Ali, Asmaa El Sayah, Sara Awni Alkhatib, Hend Abdulsalam, Mouza Juma, and Ala'a H. Al-Muhtaseb. Implementing Metal-Organic Frameworks for Natural Gas Storage. *Crystals*, 9(8):406, August 2019. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[6] Lerato Y. Molefe, Nicholas M. Musyoka, Jianwei Ren, Henrietta W. Langmi, Mkhulu Mathe, and Patrick G. Ndungu. Effect of Inclusion of MOF-Polymer Composite onto a Carbon Foam Material for Hydrogen Storage Application. *J Inorg Organomet Polym*, 31(1):80–88, January 2021.

[7] Sophie E. Miller, Michelle H. Teplensky, Peyman Z. Moghadam, and David Fairen-Jimenez. Metal-organic frameworks as biosensors for luminescence-based detection and imaging. *Interface Focus*, 6(4):20160027, August 2016. Publisher: Royal Society.

[8] Lauren E. Kreno, Kirsty Leong, Omar K. Farha, Mark Allendorf, Richard P. Van Duyne, and Joseph T. Hupp. Metal–Organic Framework Materials as Chemical Sensors. *Chem. Rev.*, 112(2):1105–1125, February 2012. Publisher: American Chemical Society.

[9] Jack Gonzalez, Krishnendu Mukherjee, and Yamil J. Colón. Understanding Structure–Property Relationships of MOFs for Gas Sensing through Henry's Constants. *J. Chem. Eng. Data*, 68(1):291–302, January 2023. Publisher: American Chemical Society.

[10] Mehdi Ghommem, Vladimir Puzyrev, Rana Sabouni, and Fehmi Najar. Deep learning for gas sensing using MOFs coated weakly-coupled microbeams. *Applied Mathematical Modelling*, 105:711–728, May 2022.

[11] Jian-Rong Li, Julian Sculley, and Hong-Cai Zhou. Metal–Organic Frameworks for Separations. *Chem. Rev.*, 112(2):869–932, February 2012. Publisher: American Chemical Society.

[12] Shotaro Hiraide, Yuta Sakanaka, Hiroshi Kajiro, Shogo Kawaguchi, Minoru T. Miyahara, and Hideki Tanaka. High-throughput gas separation by flexible metal–organic frameworks with fast gating and thermal management capabilities. *Nat Commun*, 11(1):3867, August 2020. Number: 1 Publisher: Nature Publishing Group.

[13] Mickaele Bonneau, Christophe Lavenn, Patrick Ginet, Ken-ichi Otake, and Susumu Kitagawa. Upscale synthesis of a binary pillared layered MOF for hydrocarbon gas storage and separation. *Green Chem.*, 22(3):718–724, February 2020. Publisher: The Royal Society of Chemistry.

[14] Weidong Fan, Xiurong Zhang, Zixi Kang, Xiuping Liu, and Daofeng Sun. Isoreticular chemistry within metal–organic frameworks for gas storage and separation. *Coordination Chemistry Reviews*, 443:213968, September 2021.

[15] Rama Oktavian, Raymond Schireman, Lawson T. Glasby, Guanming Huang, Federica Zanca, David Fairen-Jimenez, Michael T. Ruggiero, and Peyman Z. Moghadam. Computational Characterization of Zr-Oxide MOFs for Adsorption Applications. *ACS Appl. Mater. Interfaces*, 14(51):56938–56947, December 2022. Publisher: American Chemical Society.

[16] Michelle H. Teplensky, Marcus Fantham, Peng Li, Timothy C. Wang, Joshua P. Mehta, Laurence J. Young, Peyman Z. Moghadam, Joseph T. Hupp, Omar K. Farha, Clemens F. Kaminski, and David Fairen-Jimenez. Temperature Treatment of Highly Porous Zirconium-Containing Metal–Organic Frameworks Extends Drug Delivery Release. *J. Am. Chem. Soc.*, 139(22):7522–7532, June 2017. Publisher: American Chemical Society.

[17] Isabel Abánades Lázaro, Salame Haddad, Sabrina Sacca, Claudia Orellana-Tavra, David Fairen-Jimenez, and Ross S. Forgan. Selective Surface PEGylation of UiO-66 Nanoparticles for Enhanced Stability, Cell Uptake, and pH-Responsive Drug Delivery. *Chem*, 2(4):561–578, April 2017.

[18] Harrison D. Lawson, S. Patrick Walton, and Christina Chan. Metal–Organic Frameworks for Drug Delivery: A Design Perspective. *ACS Appl. Mater. Interfaces*, 13(6):7004–7020, February 2021. Publisher: American Chemical Society.

[19] Minyoung Yoon, Renganathan Srirambalaji, and Kimoon Kim. Homochiral Metal–Organic Frameworks for Asymmetric Heterogeneous Catalysis. *Chem. Rev.*, 112(2):1196–1231, February 2012. Publisher: American Chemical Society.

[20] A. Corma, H. García, and F. X. Llabrés i Xamena. Engineering Metal Organic Frameworks for Heterogeneous Catalysis. *Chem. Rev.*, 110(8):4606–4655, August 2010. Publisher: American Chemical Society.

[21] Liqing Ma, Carter Abney, and Wenbin Lin. Enantioselective catalysis with homochiral metal–organic frameworks. *Chem. Soc. Rev.*, 38(5):1248–1256, April 2009. Publisher: The Royal Society of Chemistry.

[22] Vlad Pascanu, Greco González Miera, A. Ken Inge, and Belén Martín-Matute. Metal–Organic Frameworks as Catalysts for Organic Synthesis: A Critical Perspective. *J. Am. Chem. Soc.*, 141(18):7223–7234, May 2019. Publisher: American Chemical Society.

[23] Yu Shen, Ting Pan, Liu Wang, Zhen Ren, Weina Zhang, and Fengwei Huo. Programmable Logic in Metal–Organic Frameworks for Catalysis. *Advanced Materials*, 33(46):2007442, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adma.202007442.

[24] P. Z. Moghadam, A. Li, X.-W. Liu, R. Bueno-Perez, S.-D. Wang, S. B. Wiggin, P. A. Wood, and D. Fairen-Jimenez. Targeted classification of metal–organic frameworks in the Cambridge structural database (CSD). *Chemical Science*, 11(32):8373–8387, 2020. Publisher: Royal Society of Chemistry RSC.

[25] Peyman Z. Moghadam, Aurelia Li, Seth B. Wiggin, Andi Tao, Andrew G. P. Maloney, Peter A. Wood, Suzanna C. Ward, and David Fairen-Jimenez. Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future. *Chem. Mater.*, 29(7):2618–2625, April 2017. Publisher: American Chemical Society.

[26] Peyman Z. Moghadam, Timur Islamoglu, Subhadip Goswami, Jason Exley, Marcus Fantham, Clemens F. Kaminski, Randall Q. Snurr, Omar K. Farha, and David Fairen-Jimenez. Computer-aided discovery of a metal–organic framework with superior oxygen uptake. *Nat Commun*, 9(1):1378, April 2018. Publisher: Nature Publishing Group.

[27] Peyman Z. Moghadam, David Fairen-Jimenez, and Randall Q. Snurr. Efficient identification of hydrophobic MOFs: application in the capture of toxic industrial chemicals. *J. Mater. Chem. A*, 4(2):529–536, December 2015. Publisher: The Royal Society of Chemistry.

[28] Christopher E. Wilmer, Michael Leaf, Chang Yeon Lee, Omar K. Farha, Brad G. Hauser, Joseph T. Hupp, and Randall Q. Snurr. Large-scale screening of hypothetical metal–organic frameworks. *Nature Chem*, 4(2):83–89, February 2012.

[29] Nakul Rampal, Abdulmalik Ajenifuja, Andi Tao, Christopher Balzer, Matthew S. Cummings, Arwyn Evans, Rocio Bueno-Perez, David J. Law, Leslie W. Bolton, Camille Petit, Flor Siperstein, Martin P. Attfield, Megan Jobson, Peyman Z. Moghadam, and David Fairen-Jimenez. The development of a comprehensive toolbox based on multi-level, high-throughput screening of MOFs for CO/N 2 separations. *Chemical Science*, 12(36):12068–12081, 2021. Publisher: Royal Society of Chemistry.

[30] Justyna Rogacka, Agnieszka Seremak, Azahara Luna-Triguero, Filip Formalik, Ismael Matito-Martos, Lucyna Firlej, Sofia Calero, and Bogdan Kuchta. High-throughput screening of metal – Organic frameworks for CO2 and CH4 separation in the presence of water. *Chemical Engineering Journal*, 403:126392, January 2021.

[31] Gokay Avci, Sadiye Velioglu, and Seda Keskin. High-Throughput Screening of MOF Adsorbents and Membranes for H2 Purification and CO2 Capture. *ACS Appl. Mater. Interfaces*, 10(39):33693–33706, October 2018. Publisher: American Chemical Society.

[32] Prosun Halder and Jayant K. Singh. High-Throughput Screening of Metal–Organic Frameworks for Ethane–Ethylene Separation Using the Machine Learning Technique. *Energy Fuels*, 34(11):14591–14597, November 2020. Publisher: American Chemical Society.

[33] Seyed Mohamad Moosavi, Arunraj Chidambaram, Leopold Talirz, Maciej Haranczyk, Kyriakos C. Stylianou, and Berend Smit. Capturing chemical intuition in synthesis of metal-organic frameworks. *Nat Commun*, 10(1):539, February 2019. Number: 1 Publisher: Nature Publishing Group.

[34] Matthew C. Swain and Jacqueline M. Cole. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.*, 56(10):1894–1904, October 2016. Publisher: American Chemical Society.

[35] Callum J. Court, Apoorv Jain, and Jacqueline M. Cole. Inverse Design of Materials That Exhibit the Magnetocaloric Effect by Text-Mining of the Scientific Literature and Generative Deep Learning. *Chem. Mater.*, 33(18):7217–7231, September 2021. Publisher: American Chemical Society.

[36] Yan Duan, Lorena E. Rosaleny, Joana T. Coutinho, Silvia Giménez-Santamarina, Allen Scheie, José J. Baldoví, Salvador Cardona-Serra, and Alejandro Gaita-Ariño. Data-driven design of molecular nanomagnets. *Nat Commun*, 13(1):7626, December 2022. Number: 1 Publisher: Nature Publishing Group.

[37] Shu Huang and Jacqueline M. Cole. BatteryDataExtractor: battery-aware text-mining software embedded with BERT models. *Chem. Sci.*, 13(39):11487–11495, October 2022. Publisher: The Royal Society of Chemistry.

[38] Edward J. Beard, Ganesh Sivaraman, Álvaro Vázquez-Mayagoitia, Venkatram Vishwanath, and Jacqueline M. Cole. Comparative dataset of experimental and computational attributes of UV/vis absorption spectra. *Sci Data*, 6(1):307, December 2019. Number: 1 Publisher: Nature Publishing Group.

[39] Avan Kumar, Swathi Ganesh, Divyanshi Gupta, and Hariprasad Kodamana. A text mining framework for screening catalysts and critical process parameters from scientific literature - A study on Hydrogen production from alcohol. *Chemical Engineering Research and Design*, 184:90–102, August 2022.

[40] Jacqueline M. Cole. How the Shape of Chemical Data Can Enable Data-Driven Materials Discovery. *Trends in Chemistry*, 3(2):111–119, February 2021.

[41] Anna M. Hiszpanski, Brian Gallagher, Karthik Chellappan, Peggy Li, Shusen Liu, Hyojin Kim, Jinkyu Han, Bhavya Kailkhura, David J. Buttler, and Thomas Yong-Jin Han. Nanomaterial Synthesis Insights from Machine Learning of Scientific Articles by Extracting, Structuring, and Visualizing Knowledge. *J. Chem. Inf. Model.*, 60(6):2876–2887, June 2020. Publisher: American Chemical Society.

[42] Ming Wang, Ting Wang, Pingqiang Cai, and Xiaodong Chen. Nanomaterials Discovery and Design through Machine Learning. *Small Methods*, 3(5):1900025, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/smtd.201900025.

[43] Sanghoon Park, Baekjun Kim, Sihoon Choi, Peter G. Boyd, Berend Smit, and Jihan Kim. Text Mining Metal–Organic Framework Papers. *J. Chem. Inf. Model.*, 58(2):244–251, February 2018. Publisher: American Chemical Society.

[44] Yi Luo, Saientan Bag, Orysia Zaremba, Adrian Cierpka, Jacopo Andreo, Stefan Wuttke, Pascal Friederich, and Manuel Tsotsalas. MOF Synthesis Prediction Enabled by Automatic Data Mining and Machine Learning**.

Angewandte Chemie International Edition, 61(19):e202200242, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.202200242.

[45] Yongchul G. Chung, Emmanuel Haldoupis, Benjamin J. Bucior, Maciej Haranczyk, Seulchan Lee, Hongda Zhang, Konstantinos D. Vogiatzis, Marija Milisavljevic, Sanliang Ling, Jeffrey S. Camp, Ben Slater, J. Ilja Siepmann, David S. Sholl, and Randall Q. Snurr. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data*, 64(12):5985–5998, December 2019. Publisher: American Chemical Society.

[46] Lezan Hawizy, David M. Jessop, Nico Adams, and Peter Murray-Rust. ChemicalTagger: A tool for semantic text-mining in chemistry. *Journal of Cheminformatics*, 3(1):17, May 2011.

[47] Hyunsoo Park, Yeonghun Kang, Wonyoung Choe, and Jihan Kim. Mining Insights on Metal–Organic Framework Synthesis from Scientific Literature Texts. *J. Chem. Inf. Model.*, 62(5):1190–1198, March 2022. Publisher: American Chemical Society.

[48] Daniel DeSantis, Jarad A. Mason, Brian D. James, Cassidy Houchins, Jeffrey R. Long, and Mike Veenstra. Techno-economic Analysis of Metal–Organic Frameworks for Hydrogen and Natural Gas Storage. *Energy Fuels*, 31(2):2024–2032, February 2017. Publisher: American Chemical Society.

[49] Hongxi Luo, Fangwei Cheng, Luke Huelsenbeck, and Natalie Smith. 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*, 9(2):105159, April 2021.

[50] Peyman Z. Moghadam, Sven M. J. Rogge, Aurelia Li, Chun-Man Chow, Jelle Wieme, Noushin Moharrami, Marta Aragones-Anglada, Gareth Conduit, Diego A. Gomez-Gualdron, Veronique Van Speybroeck, and David Fairen-Jimenez. Structure-Mechanical Stability Relations of Metal-Organic Frameworks via Machine Learning. *Matter*, 1(1):219–234, July 2019.

[51] Shu Huang and Jacqueline M. Cole. A database of battery materials auto-generated using ChemDataExtractor. *Sci Data*, 7(1):260, August 2020. Number: 1 Publisher: Nature Publishing Group.

[52] Benjamin J. Bucior, Andrew S. Rosen, Maciej Haranczyk, Zhenpeng Yao, Michael E. Ziebel, Omar K. Farha, Joseph T. Hupp, J. Ilja Siepmann, Alán Aspuru-Guzik, and Randall Q. Snurr. Identification Schemes for Metal–Organic Frameworks To Enable Rapid Search and Cheminformatics Analysis. *Crystal Growth & Design*, 19(11):6682–6697, November 2019. Publisher: American Chemical Society.

[53] Lionel Zoubritzky and François-Xavier Coudert. CrystalNets.jl: Identification of Crystal Topologies. *SciPost Chem.*, 1(2):005, June 2022.

[54] TCI Chemicals.

[55] Thomas F. Willems, Chris H. Rycroft, Michaeel Kazi, Juan C. Meza, and Maciej Haranczyk. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials*, 149(1):134–141, February 2012.

[56] Alexander U. Czaja, Natalia Trukhan, and Ulrich Müller. Industrial applications of metal–organic frameworks. *Chem. Soc. Rev.*, 38(5):1284–1293, April 2009. Publisher: The Royal Society of Chemistry.

[57] Stuart Lloyd James, Ana Lazuen-Garay, and Anne Pichon. Use of grinding in chemical synthesis, May 2007.

[58] S. T. Hyde, O. Delgado Friedrichs, S. J. Ramsden, and V. Robins. Towards enumeration of crystalline frameworks: the 2D hyperbolic approach. *Solid State Sciences*, 8(7):740–752, July 2006.

[59] Michael O'Keeffe, Maxim A. Peskov, Stuart J. Ramsden, and Omar M. Yaghi. The Reticular Chemistry Structure Resource (RCSR) Database of, and Symbols for, Crystal Nets. *Acc. Chem. Res.*, 41(12):1782–1789, December 2008. Publisher: American Chemical Society.

[60] Honghan Fei and Seth M. Cohen. A robust, catalytic metal–organic framework with open 2,2-bipyridine sites. *Chem. Commun.*, 50(37):4810–4812, April 2014. Publisher: The Royal Society of Chemistry.

[61] Jack Y. Lu, Brenda R. Cabrera, Ru-Ji Wang, and Jing Li. 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.*, 38(20):4608–4611, October 1999. Publisher: American Chemical Society.

[62] Alexander J. Tansell, Corey L. Jones, and Timothy L. Easun. MOF the beaten track: unusual structures and uncommon applications of metal–organic frameworks. *Chemistry Central Journal*, 11(1):100, October 2017.

[63] Giulia E. M. Schukraft, Sergio Ayala, Benjamin L. Dick, and Seth M. Cohen. Isoreticular expansion of polyMOFs achieves high surface area materials. *Chem. Commun.*, 53(77):10684–10687, September 2017. Publisher: The Royal Society of Chemistry.

[64] Seyed Mohamad Moosavi, Aditya Nandy, Kevin Maik Jablonka, Daniele Ongari, Jon Paul Janet, Peter G. Boyd, Yongjin Lee, Berend Smit, and Heather J. Kulik. Understanding the diversity of the metal-organic framework ecosystem. *Nature Communications*, 11(1):4068, August 2020. Number: 1 Publisher: Nature Publishing Group.

[65] Yan Bai, Yibo Dou, Lin-Hua Xie, William Rutledge, Jian-Rong Li, and Hong-Cai Zhou. Zr-based metal–organic frameworks: design, synthesis, structure, and applications. *Chem. Soc. Rev.*, 45(8):2327–2367, April 2016. Publisher: The Royal Society of Chemistry.

[66] Yang Cai, Ambarish R. Kulkarni, You-Gui Huang, David S. Sholl, and Krista S. Walton. Control of Metal–Organic Framework Crystal Topology by Ligand Functionalization: Functionalized HKUST-1 Derivatives. *Crystal Growth & Design*, 14(11):6122–6128, November 2014. Publisher: American Chemical Society.

[67] Yin-Xia Sun and Wei-Yin Sun. Influence of temperature on metal-organic frameworks. *Chinese Chemical Letters*, 25(6):823–828, June 2014.

[68] Min Chen, Man-Sheng Chen, Taka-aki Okamura, Mei-Fang Lv, Wei-Yin Sun, and Norikazu Ueyama. A series of silver(I)–lanthanide(III) heterometallic coordination polymers: syntheses, structures and photoluminescent properties. *CrystEngComm*, 13(11):3801–3810, May 2011. Publisher: The Royal Society of Chemistry.

[69] Ryther Anderson and Diego A. Gómez-Gualdrón. Large-Scale Free Energy Calculations on a Computational Metal–Organic Frameworks Database: Toward Synthetic Likelihood Predictions. *Chem. Mater.*, 32(19):8106–8119, October 2020. Publisher: American Chemical Society.

[70] Sigma Aldrich.

Chapter 6

Augmented Reality for Enhanced Visualization of MOF Adsorbents

6.1 Publication Information and Paper Contributions

This paper has been published as an application note in American Chemical Society's Journal of Chemical Information and Modeling.

In this publication I, the candidate, wrote the application note with contributions from Rama Oktavian and Kewei Zhu, under the supervision of Professor Joan L. Cordiner, Dr Jason C. Cole, and Dr Peyman Z. Moghadam.

6.2 Abstract

Augmented reality (AR) is an emerging technique used to improve visualisation and comprehension of complex 3D materials. This approach has been applied not only in the field of chemistry but also in real estate, physics, mechanical engineering, and many other areas. Here, we demonstrate the workflow for an app-free AR technique for visualisation of metal–organic frameworks (MOFs) and other porous materials to investigate their crystal structures, topology, and gas adsorption sites. We think this workflow will serve as an additional tool for computational and experimental scientists working in the field for both research and educational purposes.

6.2.1 Keywords

Adsorption, Crystal structure, Metal-organic frameworks, Molecular structure, Surface chemistry

6.3 Introduction

Porous materials such as metal–organic frameworks (MOFs), covalent organic frameworks (COFs), zeolites, silicates, polymers, and aerogels are characterized by their pore space and functionality. These properties make them desirable for a broad range of applications within the fields of chemistry, materials science, and engineering. One particularly popular subclass, MOFs, are highly ordered porous materials comprised of metal ions or clusters connected by organic ligands. MOFs have received considerable interest over the past 25 years due to their structural diversity, high surface area, and tunable properties

making them suitable materials for a broad range of adsorption applications including gas storage [1, 2, 3], separation [4, 5, 6], sensing [7, 8], and catalysis [9, 10]. As of April 2023, the Cambridge Structural Database (CSD) has seen the addition of over 27,000 3D experimentally synthesized MOFs and, due to the porous nature of these structures, many are studied as potential candidates for gas adsorption and separation applications [11].

Details regarding MOFs' structural network, pores, surface chemistry, and adsorption sites are critical pieces of information when investigating the adsorption properties of these structurally complex materials, and this information is often used in conjunction with simulation software to predict gas adsorption properties [12, 13, 14]. Adsorption simulation snapshots can be used to analyze energetically favorable adsorption sites in porous materials, and AR can create aesthetic representations of pores along with the adsorbed molecules. AR enhances spatial understanding by enabling visualization and manipulation of complex structures under specific conditions in 3D, offering additional insights when studying structure–property relationships and guest–host interactions. The use of AR has been previously reported for molecular structures using an app [15], for polymers using an app-free technique [16], as well as published workflows that build VR models which can also be viewed in AR [17]: these are typically used for educational purposes as a teaching medium [18, 19, 20]. Here, following the work of Roshandel et al. [16], we developed a protocol for app-free AR models that can be used to display MOFs under adsorption conditions or represent crystal structures in conjunction with their topology which can be viewed using a smartphone running Android or iOS.

Before we discuss the workflow for creating AR models of MOFs, let us demonstrate an example for AR gas adsorption visualization in an educational setting. Figure 6.1 shows the application of AR for water adsorption visualization in a prototypical MOF called Cu-BTC (Cu and benzene-1,3,5-tricarboxylate (BTC)), from the point of scanning the QR code to scaling the structure to fill the room. The modeling software enables the user to view the crystal structure online in 3D, or to project AR into the room either using small dimensions as shown on a desk or using large dimensions such that a person can easily fit inside the pores. Another great immersive feature is the shadowing and depth perception capabilities within the AR platform that make it an ideal tool for use in education, enabling the demonstration of complex structures in a classroom or lecture setting.

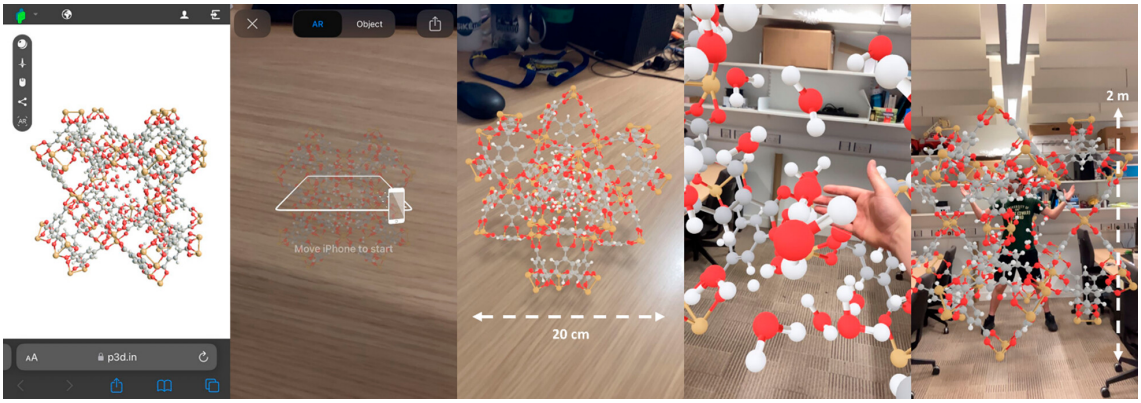


Figure 6.1: *After scanning the QR code using a smart phone, Cu-BTC is projected in augmented reality (AR), demonstrated on the table (20 cm diameter) and in much larger scale in an office with a student standing inside the pores (2 m diameter). Here, the pores contain water molecules where the structure was simulated for water adsorption.*

6.4 Modeling MOFs for use in AR

This paper explains how to create simple but aesthetic AR representations of MOFs and other crystalline materials, without the need for any significant coding knowledge. Most of the tools we use are freely available, and where licensed software packages are used, freeware alternatives can be obtained. While the workflow in this project consisted primarily of obtaining CIF or PDB files from the CSD 3D MOF subset [21], it is possible to create these AR models using any CIF or PDB file from other sources. A detailed guide showing step-by-step instructions on how to obtain and process these files can be found in the Supporting Information, followed by further explanations regarding the conversion from these chemical data files into file formats that can be used for the visualization stages, the generation of quick response (QR) codes, and the subsequent hosting of the AR maps. From the initial stage of obtaining the desired structure from the CSD, to distribution of a QR code directing your audience to the online AR resource, creating these representations can take less than 1 h per structure.

6.4.1 Crystal Structure Modifications

The initial stage of AR visualization begins by selecting a MOF, opening the corresponding CIF/PDB file, and expanding the representation to its unit cell or supercell. In this process, we select structures from CSD ConQuest and use Mercury [22] as a key tool to implement corrections on the crystal structures (e.g., addition of missing hydrogens or removal of bound/unbound solvents, if required), followed by the repair of broken or unusual chemical bonds. Often, PDB outputs have certain configured bonding patterns, and the CSD software suite is ideal for fast and easy corrections to these abnormalities. It is essential that all models are corrected in this preliminary stage so that errors are not carried forward into the AR representations, and it is recommended to manually check all atoms in the structure even after autocorrection to ensure there are no “floating” atoms remaining or undesired solvents still present in the pores. It is also possible to “snip-off” any over branching linkers that exceed the dimensions of the unit cell at this stage to ensure uniformity of the crystal structure.

6.4.2 Conversion Processes from CIF/PDB to FBX

Once all bonding information has been determined and corrected, and the structure representation has been chosen (e.g., ball and stick, ellipsoidal, wireframe, etc.) it is possible to export the file as either a PDB or CIF. Once the file is saved as a PDB or CIF file, it can be opened in Jmol [23], an open-source file conversion and visualization software. From Jmol, detailed instructions in the SI explain how to save and export this file as an object (OBJ) file, but note that after this point, it becomes much more difficult to make any chemical modifications to the structure: bonding can still be corrected in Jmol as detailed in the software documentation. The OBJ file can then be imported into the 3D modeling freeware, Blender [24]. (An alternative method which, requires the installation of the Atomic Blender plugin, can skip the Jmol step as it is possible to directly import PDB files into Blender: through various trials we determined that the OBJ technique produces equally, if not better, representations in the final stages.) Once the OBJ file is added to Blender, the remaining stages involve the removal of the light and camera layers, followed by a check of structure face count. If the face count exceeds 750,000, the image will not render correctly into AR so the "decimate" function can be used to reduce the number of faces. For best results, keeping the number of faces as close as possible to 750,000 is recommended. The reduced file can then be manipulated so that the orientation is presented as desired, before it is exported as an FBX file.

6.4.3 Publication on p3d.in and QR Code Generation

The publication stage is very simple but requires a free (or paid) subscription to p3d.in, an online 3D model hosting platform with built-in AR functionality [25]. The FBX file can be uploaded, and the structure's final orientation and color scheme can be selected. It is also possible to customize the online viewer to display the structure with different background colors, select mouse/keyboard configurations for controllability, and determine the scale of AR representation. Each file is then given a unique URL which can be assigned a static QR code using any, freely available QR code generator (we used `https://www.the-qrcode-generator.com/`) to create a QR link to the AR enabled structure. Figure 6.2 shows a graphical summary of the workflow we developed to create complex AR representations of MOFs and their gas adsorption snapshots using the molecular simulations software package, RASPA [26].

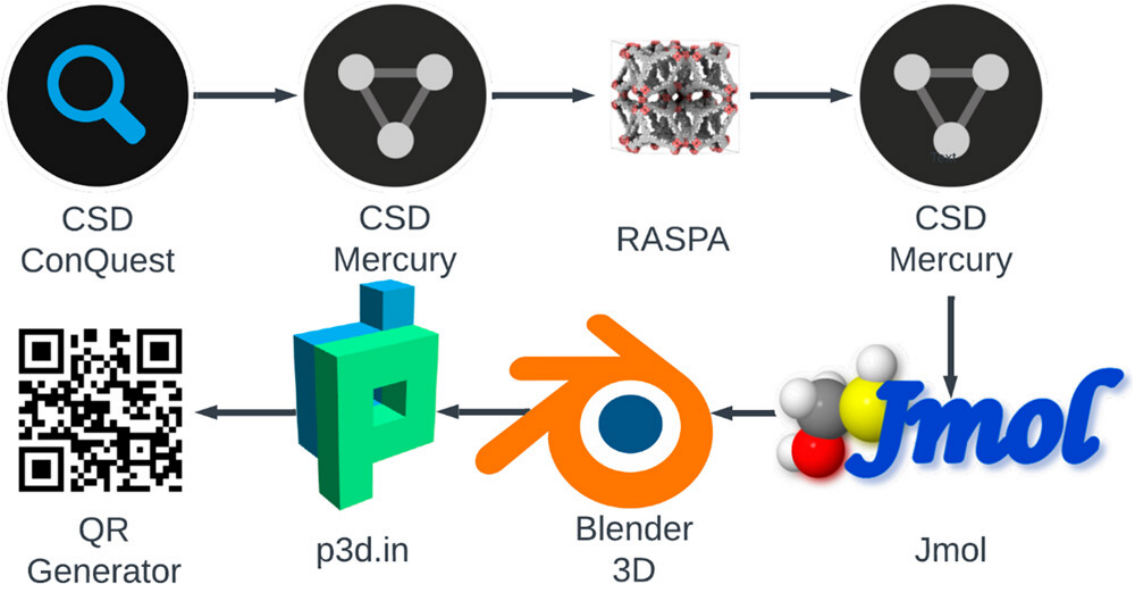


Figure 6.2: *Graphical representation of the AR visualization workflow for MOF adsorbents. We begin from initial structure selection in CSD ConQuest followed by exportation of the unit cell for use in RASPA via Mercury, structure cleanup, file format conversions in Jmol, modeling and export in Blender, and finally upload to the p3d.in platform and generation of a custom QR code.*

6.5 Applications of AR in Representing MOFs

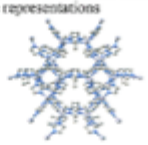
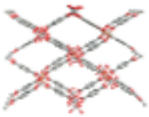
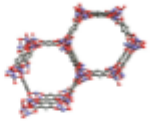
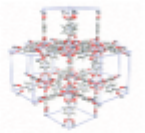
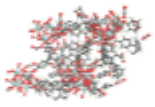
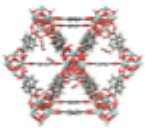
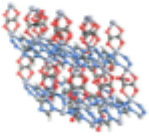
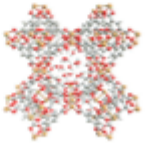
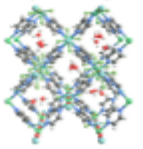
AR has seen increasing use in the field of 3D chemical and molecular structure modeling in recent years, originating from a very limited number of publications to an increasingly popular and more widespread audience [20]. We previously introduced a step-by-step guide to use AR in the field of MOFs in 2022 [27]; however, this process was tedious and complex requiring the development and publication of an app. The current approach follows an easier, app-free, workflow. Although it would be straightforward to produce AR representations for basic MOF structures, it is possible to make these tools more useful. For example, the representation of molecular interactions between gases adsorbed in MOFs are mostly represented in 2D which can make it difficult to fully appreciate the complex interplay between structural network, pore shapes and sizes, surface chemistry, and preferential adsorption sites. Clearly, a 3D AR technology to assist in the detection of adsorption sites or topological features of MOFs provides better understanding of the adsorption phenomena and the pore environment.

6.5.1 Crystal Representation

For this article, we created several interesting and diverse AR representations of MOFs under various conditions to demonstrate the relevance and reach of AR in the field, from basic MOF visualizations to more complex representations of their topologies and gas adsorption. Table 6.1 contains a selection of materials created using a combination of structure preparation approaches, demonstrating the range of uses for AR visualization. The initial representations are taken directly from the CSD 3D MOF subset with only minor modifications. Rows 1-3 in Table 6.1 show how AR can be used to visualize MOF crystal structures (e.g. ZIF-8, MIL-54, and CPO-27) without the need for complex modifications to CIF files from the CSD. The choice of crystal structures represented

here took several factors into consideration. For better aesthetics and clarity, we avoided materials with large unit cells because of computational expense when converting large unit cell structures to OBJ files, the file size can exceed several hundred MBs which makes rendering more difficult. The structures shown in Table 6.1 are illustrative examples, we mainly picked materials that are well-known in the MOF community for gas adsorption applications with CIFs that are easily accessible.

Table 6.1: *Selection of MOFs Visualized in AR under Different Conditions*

Name (CSD Refcode)	Condition	Diagram (3D)	QR CODE
Crystal structure representations			
ZIF-8 (FAWCEN)[31]	Original CIF with no H atoms		
MIL-53 (ASOHUL)[32]	Original CIF		
CPO-27-Co (MOHGEY)[33]	Original CIF		
Topology representations			
IRMOF-1 (SAHYIK)[34]	**pcu**		
UiO-67 (WIZMAV)[35]	**fcu**		
Gas adsorption site representations			
MOF-812 (BOHWOM)[36]	CO_2 adsorption at 298K and 0.15 bar		
UiO-66 (RUBTAK)[37]	CO_2 adsorption at 298 K and 0.15 bar		
CALF-20 (TASYAR)[38]	CO_2 adsorption at 323 K and 0.15 bar		
HKUST-1 (FIQCEN)[39]	H_2O adsorption at 323 K and 6 mbar		
KAUST-7 (REDQUQ)[40]	H_2O adsorption at 293 K and 0.5 bar		

6.5.2 Topology Representation

Another interesting use of AR is for the comparison between the complex crystal structure of MOFs and their underlying topologies. The creation of "MOF plus topology" AR representations involves the use of software such as ToposPro [28] or CrystalNets [29] to generate the underlying net for a given CIF. It is essential that topological nets correctly match the dimensions of the CIF so they can be overlaid together for AR representation. Once the topology is determined, the next stage is to convert both the CIF and its corresponding net into separate OBJ files where they can then be layered together in Blender for export into a single FBX file. Topology nets are typically created as CGD or MOL2 files that can be imported into CSD Mercury, before being exported as PDB files for use in Jmol. One disadvantage of this method is that the underlying net representation cannot later be switched on/off in the free version of p3d.in, but it remains an interesting tool for demonstrating the underlying connectivity of MOFs. Table 6.1, rows 4 and 5, show IRMOF-1 and UiO-67 as illustrative examples of MOFs with their overlaid topologies.

6.5.3 Gas Adsorption Representation

To obtain more information from AR modeling, one can run adsorption simulations and generate "snapshots" of gas adsorption sites for visualization in AR at desirable operating temperatures and pressures. Here, we use RASPA [26] for running Monte Carlo simulations of adsorption in MOFs. Once the simulation at a specific temperature and pressure in the isotherm is equilibrated, RASPA provides CIF and PDB output files for the framework and the adsorbate molecules, and to visualize gas adsorption snapshots, these two files should be merged into a single file followed by the AR development process, as explained in Section 6.4. Figure 6.3 demonstrates CO_2 adsorption isotherm simulated in UiO-67 at room temperature. As can be seen, we can use AR to visualize gas adsorption snapshots as pressure is increased from 0.15 to 20 bar. These snapshots are all produced in RASPA forming the basis of AR model creation for gas adsorption visualization.

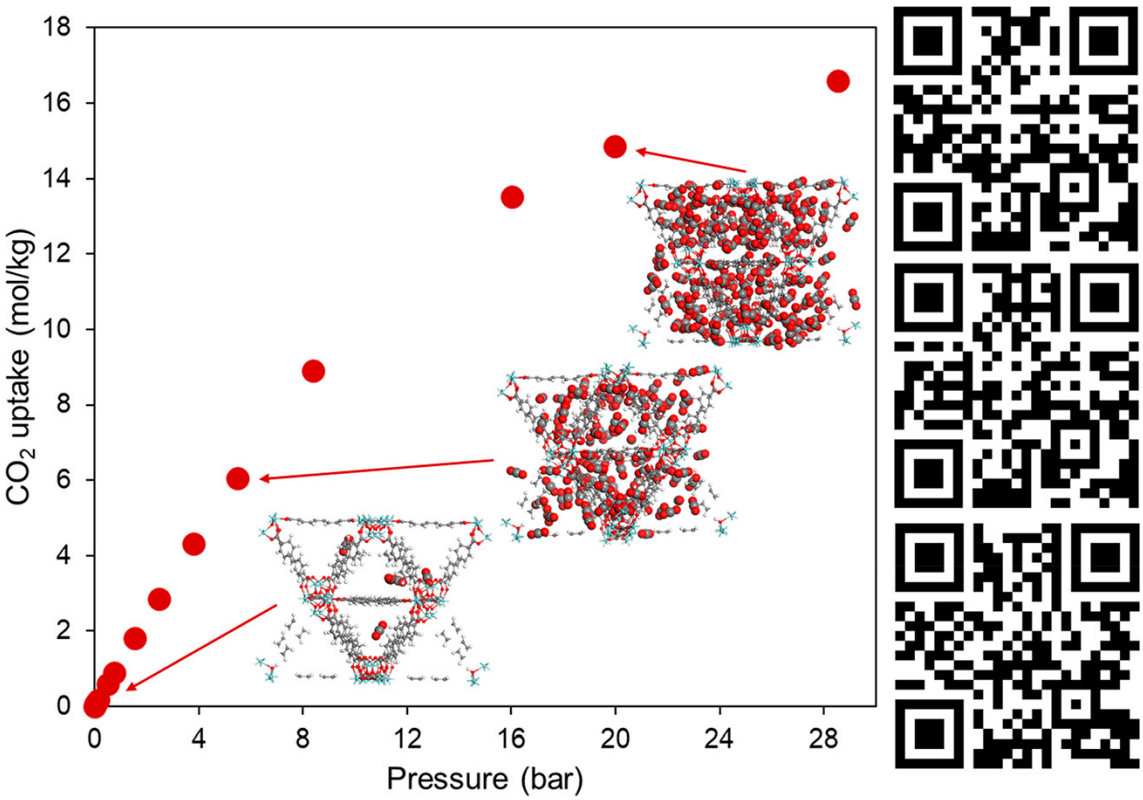


Figure 6.3: *CO2 adsorption isotherm simulated in UiO-67 at 298 K. Adsorption snapshots are highlighted at 0.15, 5.5, and 20 bar. QR codes for AR visualization are located adjacent to each snapshot.*

In Table 6.1, rows 6–10, we demonstrate AR representation of CO_2 and H_2O adsorption in selected MOFs. By simply scanning the relevant QR code, one can create an immersive experience, investigating how the pore environment and surface chemistry affect the adsorption of CO_2 or water in MOF-812, CALF-20, UiO-66, HKUST-1, and KAUST-7. In contrast to typical 2D visualization of simulation snapshots, AR can thoroughly capture the clustering of water molecules and the formation of hydrogen bonds, typically seen in hydrophilic MOFs such as in Cu-BTC. In CALF-20, it can be seen that CO_2 molecules sit tightly within the channels of the oxalic acid linkers where they strongly interact with the metal nodes. In UiO-66, we can use this experience to easily visualize that CO_2 molecules occupy the tetrahedral cages first at low pressure conditions.

6.5.4 Reception

We demonstrated some of the MOF AR visualizations via QR codes at the first Mediterranean Porous Materials Conference in May 2023 in Crete, Greece, and again at the sixth Annual UK Porous Materials Conference in June 2023 in Sheffield, United Kingdom, receiving approximately 350 views combined. We received many positive comments regarding the quality and clarity of the modeling, and we shared a number of additional QR codes to demonstrate the flexibility of these methods with various structures and gases. We also received constructive feedback for AR demonstration of, e.g., bond vibrations, structural flexibility, and increasing interactivity by implementing measuring tools.

6.6 Conclusion

This article showcases the capability of AR modeling for visualization of MOFs and other porous materials for a variety of applications. We used AR for representing MOFs crystalline structure, their underlying topologies and favorable gas adsorption regions without the need for additional downloads, as the models can be viewed on an Android or iOS smartphone app-free. The technique outlined in this paper and the SI allows anyone to create attractive AR models that can be shared globally by simply distributing a QR code. This freely available, no-cost method is ideal for augmenting MOF posters at conferences, adsorption workshops, and crystal structure presentations as an engaging and interactive experience. Furthermore, the ability to modify the size of AR representations to over 5 m in diameter once placed in a room, establishes the use of AR as an educational tool in the field for furthering understanding of gas adsorption and topological complexities of these intriguing materials. Additional applications of AR include use in research and design for visualization of reactants, intermediates and products in catalysis, crystal engineering and visualization of defects, solvents and irregularities, as well as facilitating collaboration and communication between research groups via the Internet, and even artistic experiences of crystal structure representations.

Data Availability

All of the structures featured here can be downloaded directly from the Cambridge Structural Database (CSD), which can be obtained from https://www.ccdc.cam.ac.uk/support-and-resources/download-the-csd. RASPA can be freely obtained from https://iraspa.org/raspa/. Blender is available for free at https://www.blender.org/download/. The Jmol freeware can be found at http://jmol.sourceforge.net/download/. Online 3D model hosting platform at https://p3d.in/. These links to the various software packages used are also presented in the SI. We also include a ZIP file of gas adsorption files from the RASPA outputs (where applicable), as well as the FBX files used to create these AR representations.

References

[1] Omar K. Farha, A. Özgür Yazaydın, Ibrahim Eryazici, Christos D. Malliakas, Brad G. Hauser, Mercouri G. Kanatzidis, SonBinh T. Nguyen, Randall Q. Snurr, and Joseph T. Hupp. De novo synthesis of a metal-organic framework material featuring ultrahigh surface area and gas storage capacities. *Nat Chem*, 2(11):944–948, November 2010.

[2] Shengqian Ma and Hong-Cai Zhou. Gas storage in porous metal–organic frameworks for clean energy applications. *Chem. Commun.*, 46(1):44–53, January 2010. Publisher: The Royal Society of Chemistry.

[3] Bin Li, Hui-Min Wen, Wei Zhou, Jeff Q. Xu, and Banglin Chen. Porous Metal-Organic Frameworks: Promising Materials for Methane Storage. *Chem*, 1(4):557–580, October 2016.

[4] Jian-Rong Li, Julian Sculley, and Hong-Cai Zhou. Metal–Organic Frameworks for Separations. *Chem. Rev.*, 112(2):869–932, February 2012. Publisher: American Chemical Society.

[5] Qihui Qian, Patrick A. Asinger, Moon Joo Lee, Gang Han, Katherine Mizrahi Rodriguez, Sharon Lin, Francesco M. Benedetti, Albert X. Wu, Won Seok Chi, and Zachary P. Smith. MOF-Based Membranes for Gas Separations. *Chem. Rev.*, 120(16):8161–8266, August 2020. Publisher: American Chemical Society.

[6] Rui-Biao Lin, Shengchang Xiang, Wei Zhou, and Banglin Chen. Microporous Metal-Organic Framework Materials for Gas Separation. *Chem*, 6(2):337–363, February 2020.

[7] Sophie E. Miller, Michelle H. Teplensky, Peyman Z. Moghadam, and David Fairen-Jimenez. Metal-organic frameworks as biosensors for luminescence-based detection and imaging. *Interface Focus*, 6(4):20160027, August 2016. Publisher: Royal Society.

[8] Jack Gonzalez, Krishnendu Mukherjee, and Yamil J. Colón. Understanding Structure–Property Relationships of MOFs for Gas Sensing through Henry's Constants. *J. Chem. Eng. Data*, 68(1):291–302, January 2023. Publisher: American Chemical Society.

[9] Vlad Pascanu, Greco González Miera, A. Ken Inge, and Belén Martín-Matute. Metal–Organic Frameworks as Catalysts for Organic Synthesis: A Critical Perspective. *J. Am. Chem. Soc.*, 141(18):7223–7234, May 2019. Publisher: American Chemical Society.

[10] Yu Shen, Ting Pan, Liu Wang, Zhen Ren, Weina Zhang, and Fengwei Huo. Programmable Logic in Metal–Organic Frameworks for Catalysis. *Advanced Materials*, 33(46):2007442, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adma.202007442.

[11] Peyman Z. Moghadam, Aurelia Li, Seth B. Wiggin, Andi Tao, Andrew G. P. Maloney, Peter A. Wood, Suzanna C. Ward, and David Fairen-Jimenez. Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future. *Chem. Mater.*, 29(7):2618–2625, April 2017. Publisher: American Chemical Society.

[12] Christopher E. Wilmer, Omar K. Farha, Youn-Sang Bae, Joseph T. Hupp, and Randall Q. Snurr. Structure–property relationships of porous materials for carbon dioxide separation and capture. *Energy Environ. Sci.*, 5(12):9849–9856, November 2012. Publisher: The Royal Society of Chemistry.

[13] Majedeh Gheytanzadeh, Alireza Baghban, Sajjad Habibzadeh, Amin Esmaeili, Otman Abida, Ahmad Mohaddespour, and Muhammad Tajammal Munir. Towards estimation of CO2 adsorption on highly porous MOF-based adsorbents using gaussian process regression approach. *Sci Rep*, 11(1):15710, August 2021. Number: 1 Publisher: Nature Publishing Group.

[14] Min Xu, Zhangli Liu, Xiulan Huai, Lanting Lou, and Jiangfeng Guo. Screening of metal–organic frameworks for water adsorption heat transformation using structure–property relationships. *RSC Advances*, 10(57):34621–34631, 2020. Publisher: Royal Society of Chemistry.

[15] Kristina Eriksen, Bjarne E. Nielsen, and Michael Pittelkow. Visualizing 3D Molecular Structures Using an Augmented Reality App. *J. Chem. Educ.*, 97(5):1487–1490, May 2020. Publisher: American Chemical Society.

[16] Hootan Roshandel, Matthew Shammami, Shiyun Lin, Yin-Pok Wong, and Paula L. Diaconescu. App-Free Method for Visualization of Polymers in 3D and Augmented Reality. *J. Chem. Educ.*, 100(5):2039–2044, May 2023. Publisher: American Chemical Society.

[17] Fabio Cortés Rodríguez, Matteo Dal Peraro, and Luciano A. Abriata. Online tools to easily build virtual molecular models for display in augmented and virtual reality on the web. *Journal of Molecular Graphics and Modelling*, 114:108164, July 2022.

[18] Jonah Kailer Aw, Kevin Christopher Boellaard, Teck Kiang Tan, John Yap, Yi Ping Loh, Benoît Colasson, Étienne Blanc, Yulin Lam, and Fun Man Fung. Interacting with Three-Dimensional Molecular Structures Using an Augmented Reality Mobile App. *J. Chem. Educ.*, 97(10):3877–3881, October 2020. Publisher: American Chemical Society.

[19] Michael Ovens, Megan Ellyard, Jacob Hawkins, and Dino Spagnoli. Developing an Augmented Reality Application in an Undergraduate DNA Precipitation Experiment to Link Macroscopic and Submicroscopic Levels of Chemistry. *J. Chem. Educ.*, 97(10):3882–3886, October 2020. Publisher: American Chemical Society.

[20] Alba Fombona-Pascual, Javier Fombona, and Rubén Vicente. Augmented Reality, a Review of a Way to Represent and Manipulate 3D Chemical Structures. *J. Chem. Inf. Model.*, 62(8):1863–1872, April 2022. Publisher: American Chemical Society.

[21] P. Z. Moghadam, A. Li, X.-W. Liu, R. Bueno-Perez, S.-D. Wang, S. B. Wiggin, P. A. Wood, and D. Fairen-Jimenez. Targeted classification of metal–organic frameworks in the Cambridge structural database (CSD). *Chemical Science*, 11(32):8373–8387, 2020. Publisher: Royal Society of Chemistry RSC.

[22] C. F. Macrae, I. Sovago, S. J. Cottrell, P. T. A. Galek, P. McCabe, E. Pidcock, M. Platings, G. P. Shields, J. S. Stevens, M. Towler, and P. A. Wood. Mercury 4.0: from visualization to analysis, design and prediction. *J Appl Cryst*, 53(1):226–235, February 2020. Number: 1 Publisher: International Union of Crystallography.

[23] R. M. Hanson. Jmol – a paradigm shift in crystallographic visualization. *J Appl Cryst*, 43(5):1250–1260, October 2010. Number: ARRAY(0xb1ec5c8) Publisher: International Union of Crystallography.

[24] Blender Development Team. Blender - a 3D modelling and rendering package, 2022.

[25] p3d.in, July 2023.

[26] David Dubbeldam, Sofía Calero, Donald E. Ellis, and Randall Q. Snurr. RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Molecular Simulation*, 42(2):81–101, January 2016. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/08927022.2015.1010082.

[27] Rama Oktavian, Raymond Schireman, Lawson T. Glasby, Guanming Huang, Federica Zanca, David Fairen-Jimenez, Michael T. Ruggiero, and Peyman Z. Moghadam. Computational Characterization of Zr-Oxide MOFs for Adsorption Applications. *ACS Appl. Mater. Interfaces*, 14(51):56938–56947, December 2022. Publisher: American Chemical Society.

[28] Vladislav A. Blatov, Alexander P. Shevchenko, and Davide M. Proserpio. Applied Topological Analysis of Crystal Structures with the Program Package ToposPro. *Crystal Growth & Design*, 14(7):3576–3586, July 2014. Publisher: American Chemical Society.

[29] Lionel Zoubritzky and François-Xavier Coudert. CrystalNets.jl: Identification of Crystal Topologies. *SciPost Chem.*, 1(2):005, June 2022.

Chapter 7

Conclusions and Future Work

7.1 Conclusion

This project, and the work conducted within it, has resulted in the creation of the publications within this document in addition to the stimulation of further ongoing works within this research group and beyond. In this thesis, we began with a goal of improving the topological characterisation of MOFs and MOF-like structures within the CSD, as well as creating new computational tools to aid in this process with the potential to implement some forms of machine learning to achieve this outcome. Computational studies such as this are playing an increasingly important role in today's digital society by driving research into a more efficient and streamlined process for design and discovery of new functional materials.

Efficient and accurate topological characterisation is one area which still has scope for improvement. MOF topology is crucial for understanding structure-property relationships that can have significant performance impacts on materials depending on their chosen applications. By conducting this analysis of connectivity between the linkers and nodes it is possible to predict porosity, surface area, stability, and other key properties without the need for experimental synthesis to take place - this can also aid in MOF design with the ability to tailor functionality for a range of applications including gas storage, separation, catalysis, and sensing. Lastly, topological analysis enables classification and comparison between different MOFs based solely on their structure allowing for grouping in terms of common structural motifs whilst helping to guide rational design of hypothetical materials.

In Chapter 2 we investigated exactly what topology means in the context of crystalline materials with a deep perspective look into the requirements for accurate topological assignment and the tools that are currently available in the MOF domain. We also considered several datasets to find an ideal target database for the inclusion of topological information. This chapter had a specific focus on ensuring that the form of periodic networks were understood, and we talked about the knowledge required to select the best assignment software for the problems a researcher might be facing and where to find further information when investigating complex crystal structure analysis. We also touched on the IUPAC guidelines and what we expect to see in publications that report MOF topology, and suggested that the creation of a freely available MOF topology database would be of benefit to the community.

In Chapter 3 we introduced a new Python based tool for use in conjunction with the

CCDC software suite that can be used at the click of a button. The crystal structure characterisation report gives the user insight into the spacegroup, cell volume, void fractions, packing, and topology (if applicable) to any molecule or MOF that has been deposited into the CSD.

Chapter 4 introduced the concept of machine learning and digital manufacturing for solid-state materials development. ML is revolutionising digital manufacturing by offering unprecedented opportunity for efficiency, optimisation, and innovation across a number of industrial sectors. By learning from experience, recognising patterns, and making decisions without needing explicit programming ML plays a pivotal role in transforming traditional processes into smart, data-driven systems. Novel solid state materials are urgently required for energy applications that include carbon capture and storage, and the synthesis of these candidates has typically been a manual process that is inefficient for exploration of new materials. These emerging data science and digital technologies are developing as highly sought after techniques in academia and industry.

Following on from the previous introduction, in Chapter 5 we developed a text-mining approach using a modified version of ChemDataExtractor to parse 43,281 MOF publications and extract 52,680 associated MOF properties, such as synthesis route, metal precursor, solvent, and organic linker with an overall precision of 77%. Then, to supplement this data we calculated pore limiting diameters, largest cavity diameters, and independently assessed topology for all 3D MOFs in the CSD MOF subset. Further, we mined the organic linker names from the CSD for each of the corresponding structures, normalised them, and performed cost analysis. This study resulted in the creation of the DigiMOF database which contains a wealth of information, publicly available to researchers for use in training ML models, or to aid them in searching for specific structure properties.

Chapter 6 is a showcase of an additional new feature, augmented reality modeling for enhanced visualisation. As previously discussed, MOFs are ideal candidates for gas storage and adsorption applications, and further their topology is a key feature. The use of AR for visualisations allows researchers to bring atomic scale representations of MOFs into the room enabling them to gain insight into the adsorption sites at various conditions, or further understand the complex underlying connectivity of novel structures. These models are a great tool for understanding and educating with an almost limitless combination of available representations - if it can be rendered on a computer then it likely can be rendered in AR.

7.2 Future Work

Whilst this study has provided some valuable insight into topological analysis of MOFs, the use of natural language processing and data mining for synthesis information gathering, and the introduction of some new computational tools, several avenues for future research have emerged that warrant exploration. In this short section, I highlight a few key areas where additional investigation could improve our understanding and contribute to the advancement of the MOF and crystal chemistry field.

Firstly, large language models such as ChatGPT, are a hot topic at present. These models are trained on vast amounts of text data and excel in natural language understanding and generation, and in fact we have already seen the adoption of these tools begin

within the field in recent publications. In the context of MOFs, language models have several potential uses including data analysis, materials design, data curation, knowledge synthesis, and they could even be used for hypothesis generation.

Language models can analyse research papers and patents to identify trends, summarise findings, and parse important parameters as demonstrated by DigiMOF. The continual development of these tools increases their abilities and can lead to further MOF synthesis property extraction. By obtaining this information it allows researchers to gain more comprehensive understanding of MOFs to guide further research and innovation. Knowing these patters and correlations in data can also prompt the next steps for experimental or computational study - possibly through suggestions of optimal design parameters to tailor-make materials for specific applications. Lastly, this curated data can serve as a valuable resource for future analysis and modelling studies and form a basis for machine learning training for further development.

Future work should also include the continual development and refinement of the CCDC's powerful software suite. In recent times we have seen the inclusion of PoreBlazer within CSD Mercury as a welcome addition to crystal structure analysis, and in this thesis we also developed a processes through which CrystalNets can be used to assign topology to any structure within the database. I would recommend the inclusion of new tools and the refinement of existing tools should continue as is evidenced by the quarterly updates released by the CCDC and their support of researcher-created Python tools hosted on their GitHub. The 2024.1 update has strengthened the hydrogen addition capabilities, and also the packing similarity functionality for structures with internal symmetries.

Appendices

Appendix A

Supporting Information for DigiMOF

A.1 Article Retrieval

Article retrieval is achieved by using DOIs to automatically download articles from journal websites. Two methods may be used to retrieve article DOIs when assembling a corpus of MOF articles using CDE. For the first method, a web scraping script developed in the most recent version of CDE can be used to send a search query to Elsevier and the Royal Society of Chemistry to extract the DOIs which are then used to download the article in the form of a HTML file. We also developed a second method which involves retrieving MOF reference codes and their associated DOIs from the Cambridge Structural Database (CSD) using the CSD Python API. Both methods produce a CSV file which stores the DOIs of the articles to be downloaded. Here, we used the CSD Python API as utilising search queries was found to significantly increase the time required for web scraping. We wrote a Python script which calls the Selenium webdriver to navigate to the article webpage and the PyAutoGui library to save the articles as HTML files. After running the web scraping script, to get access to the publications, a window appears where the user must sign into their DOI account via their institution's website. The scraper will then automatically copy and paste article DOIs from the CSV file where they are stored, prior to downloading them. The web scraping script can download approximately three articles per minute. We recommend researchers use high-performance computing clusters to assemble corpuses that contain thousands of articles to avoid a bottleneck in the pipeline.

A.2 Database Overview and Performance

As the data mined for the DigiMOF database consisted of text-text relationships, in contrast to the text-numerical records from previously conducted text mining studies, there were considerably more linguistic and syntactical variations in the reporting of the properties of interest compared to previous projects.

Table A.1: *Parsing elements used to create the rule-based grammars to identify MOF names and corresponding topology, solvent, synthesis route, organic linker, and/or metal precursor [1].*

Element	Description	Element	Description
R(regex)	Match text with regular expression	T(tag)	Match tags
W(word)	Match case-insensitive token text	I(iword)	Match case-insensitive token text
Any	Match any single token	H(hide)	Ignore the matched tokens
Not	Match only if not followed by some text	Followed By	Match only if followed by some text
ZeroOrMore	Match zero or more of the expressions	OneOrMore	Match one or more of the expressions
Optional	Match if it exists	SkipTo	Skips to the next occurrence of text

Table A.2 contains examples of compound records which were extracted in this project and in previous chemistry text mining projects. Previous projects enlisting CDE mined text-numerical data, whereas this project mined qualitative text-text relationships. Note that for the example for this database a real record was used and in this instance no topology or solvent was associated with the MOF compound by the parser. This is typical as it is rare for all 5 properties to be found in one compound record. It has been attempted to represent records from previous projects as faithfully as possible but the exact format of the scraped records is not always available in the source material.

Table A.2: *Examples of compound records from previous versions of CDE, the previous attempt to text mine MOF data and this work.*

Source	Project	Compound	Properties				
This Work	*CDE MOF Synthesis Routes*	**Names:** MOF-5	**Linker:** H_2BDC	**Metal_precursor:** Zinc(II) terephthalate	**Synthesis_route:** solvothermal		
Huang and Cole[1]	*CDE Battery Materials Database*	**Names:** Li1.5Co02	**Current_value:** 16	**Current_units:** mAg-1	**Cycle_value:** 25	**Cycle_units:** cycles	
Court and Cole[2]	*CDE Neel and Curie Temperatures*	**Names:** BiFeO3	**Temp_type:** Curie	**Temp_value:** 647	**Temp_units:** K		
Park et al.[3]	*MOF Surface Area and Pore volume*	**Names:** MOF-210	**Surface_type:** BET	**Surface_value:** 6240	**Surface_units:** m^2/g	**Pore_volume_value:** 3.6	**Pore_volume_units:** cm^3/g

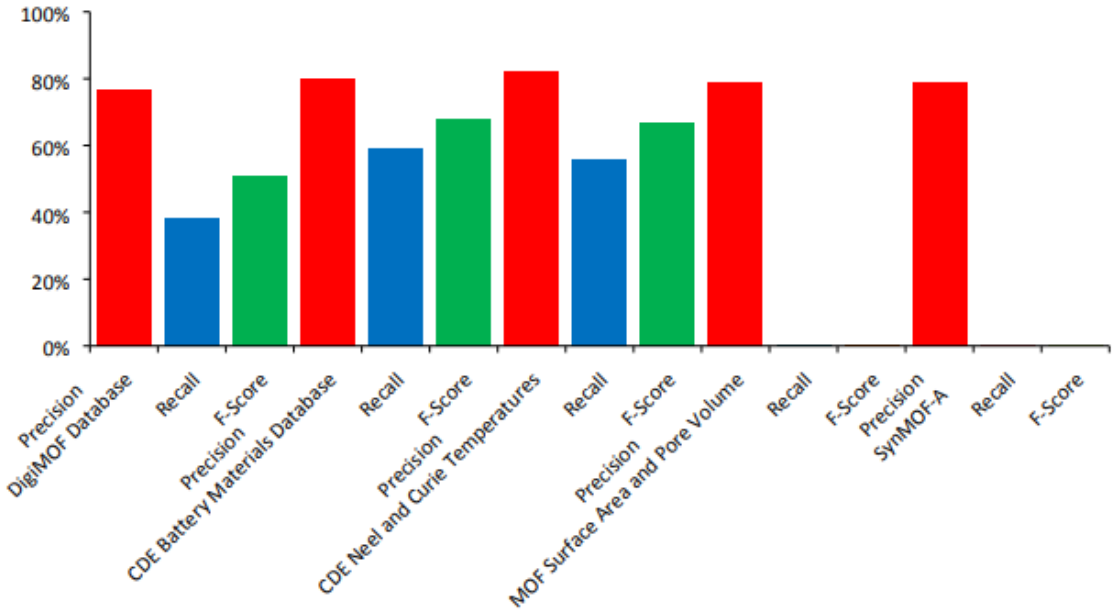


Figure A.1: *MOF CDE parser performance compared with previous versions of CDE [1, 2] and the work from Park et al. [3] MOF text mining tool. Performance of individual parsers and detailed methodology for calculation of these metrics is available in the supporting information, Table A.3.*

The machine-learning assisted version of CDE enlisted in the Neel and Curie Temperature database achieved a precision of 82% on its test-set, but this is expected to converge to 66% over time as the algorithm is trained on broader datasets [2]. Park et al.[3] reported their accuracy to be 79% but recall and F-score were not reported in their work. Luo et al. [4] also had an accuracy report of 78.9% which was referred to as consistency, this value was obtained by matching the manually extracted records in the SynMOF-M database with the automatically extracted records in the SynMOF-A database. It's also important to note that when sentences contained multiple compound names associated with other properties, our parsers could only identify the properties correctly if the MOF compound name was preceded or followed by a property without another MOF compound name separating the two. Some sentences however have multiple MOF names listed first with their corresponding properties listed second and this resulted in the erroneous association of the last MOF compound name with the first property name. Finally, a filter for MOF names was created using a regular expression to ensure that only MOF compound names were extracted into the database which further limited the entries, increasing precision.

The performance of each individual parser was also manually assessed on 50 random journal articles. For each property (synthesis routes, topologies, solvents, linkers, and metal precursors) in the database, both the precision and recall was calculated as shown in Table A.3 below.

Table A.3: *Summary of the performance of each individual parser.*

Property	Precision	Recall	F-score
Synthesis Route	100%	37.7%	54.7%
Topologies	70%	40%	50.9%
Linkers	62.9%	35.9%	45.7%
Metal Precursors	89.4%	40.3%	55.6%

For each property, the precision was calculated by manually extracting each property from all 50 papers. Following this, the values extracted by the parsers were given the value of "1" if the match was correct and a value of "0" if the match was incorrect. The total of correct extractions was then divided by the total number of identified properties (incorrect are false positives) to obtain the precision. To calculate the recall, the total of correct values was divided by the all the correct possible values (false negatives) that the parsers could have extracted from the papers.

Table A.4 contains a list of regular expression samples which were used to parse the respective properties. These expressions were modified and refined to improve on the aforementioned recall and precision metrics of the parser techniques for the DigiMOF database.

Table A.4: *Simplified MOF CDE Regular Expression (Regex) examples.*

Variable		CDE Regex	Evaluation	TP	FP	FN
MOF names	*1.*	I(^MOF[a-zA-Z0-9\-]*$') ***Too lenient***	Rejected	"MOF-5"	"MOF" "Moffat"	"Zr-MOF-808"
	2.	R('^MOF-[a-zA-Z0-9]+$') ***Only matches MOF and a suffix***	Accepted	"MOF-5"	-	"Zr-MOF-808"
	3.	R('^[aA-zZ]*\-*MOF-[a-zA-Z0-9]+$') **Matches MOF and a suffix with an optional prefix**	Accepted	"MOF-5" "Zr-MOF-808"	-	-
Common Linkers	*4.*	R('H(2|3)[A-Z]+') **Too lenient**	Rejected	"H_2-BDC" "H_2DABCO"	"H_2O"	"$h_2$bdc"
	5.	R('H(2|3)[BDCTANHFPVIMEP]+') **Only matches capitalised linker abbreviations**	Accepted	"H_2BDC"	-	"H_2DABCO" "$h_2$bdc"
	6.	I('H(2|3)[BDCTANHFPVIMEP]+') **Case-insensitive version of 5**	Accepted	"H_2BDC" "$h_2$bdc"	-	"H_2DABCO"
	7.	I('h2dabco') **Case insensitive match to only these characters**	Accepted	"H_2DABCO"	-	"H_2BDC" "$h_2$bdc"
Metal Precursors	*8.*	R('Fe[0-9\+\-A-Za-z]*') **Too lenient**	Rejected	"Fe" "$FeSO_4$"	"$Fe_3OH(H_2O)_2O[(BDC)]_3$" "Feral" "Ferrous"	-
	9.	R('Fe[0-9\+\-A-Z]+') **Matches both MOF and metal salt names. Exclusion list can be used to exclude MOF names.**	Accepted only with exclusion list item	"$FeSO_4$"	"$Fe_3OH(H_2O)_2O[(BDC)]_3$"	"Fe"
	10.	R('Fe\W') **Use of a word boundary only matches with the metal element formula**	Accepted	"Fe"	-	-

Table A.5 below demonstrates the development of regular expressions to eliminate compounds which were frequently misidentified as linkers in exclusion lists. This is not shown for metal precursors and MOF names as the principle was similar and misidentifications were much less common for these variables as their definitions included less ambiguous regular expressions. As with the regular expressions for the variable definitions, strings may be added to the exclusion list if they have false negatives but should be avoided if they have false positives as this will prevent the identification of compounds

such as linkers (which is why expression 5 may be preferred to expression 2). For abbreviations or compound names which convey ambiguity or overlap between variable types, it can be advisable to use more tailored and/or case-sensitive regular expression which corresponds to a limited number of strings or to a unique string of characters (as with 10 expressions 5 and 6), rather than attempting to accommodate or exclusion list many strings using more general rules. These examples are simplified and the actual exclusion list regular expressions can be found in the MOF CDE on GitHub.

Table A.5: *Simplified examples of organic linker exclusion list item regular expression development. Note that compound types (such as MOF names and metal precursors) were also added to this exclusion list.*

	CDE Style Exclusion List Regular Expression	Evaluation	TP	FP	FN
1.	R('[A-Za-z0-9]*OH')	Rejected	"CH_3OH"	"C_6H_5COOH"	"methanol"
	Too lenient, exclusion lists common carboxylic acid type linker formulae		"OH^-"		"2,4-dimethyl-3-pentanol"
			"OHIO"		"DMF"
2.	R('A-Za-z0-9\(\)\-\,]*\W[Aa]cid')	Rejected	"Formic acid"	"Dicarboxylic acid"	"DMF"
	Too lenient, exclusion lists common carboxylic acid type linker names		"Acetic acid"		
3.	R('^[CH0-9]*OH$')	Accepted	"CH_3OH"	-	"methanol"
	Exclusion lists solvent and adsorbate alcohol formulae				"2,4-dimethyl-3-pentanol"
					"DMF"
4.	R('[A-Za-z0-9\(\)\-\,]*[Oo][Ll]')	Accepted	"methanol"	-	"DMF"
	Exclusion lists solvent and adsorbate alcohol names		"2,4-dimethyl-3-pentanol"		
5.	I('acetic\Wacid')	Accepted	"Acetic acid"	-	"methanol"
	Exclusion lists only one commonly-used MOF synthesis modulator				"2,4-dimethyl-3-pentanol"
					"DMF"
6.	I('dmf')	Accepted	"DMF"	-	"methanol"
	Exclusion lists only one commonly-used MOF synthesis solvent abbreviation				"2,4-dimethyl-3-pentanol"

A.3 Synthesis Proportionality

Figure A.2 is a pie chart representation of the proportionality of synthesis techniques extracted over the previous 25 years of MOF synthesis, spanning the period of 1995 to

2020. The results show a significant preference for hydrothermal techniques over the next leading method. The third most commonly reported technique has a share of only 1.44% of all reported techniques, although we must note that it is anticipated this is an even smaller proportion due to the non–specific reporting of the most common techniques which is noted in many MOF synthesis papers. Despite the increasing prevalence of novel techniques, they at present likely make up significantly less than 4% of all MOF synthesis pathways.

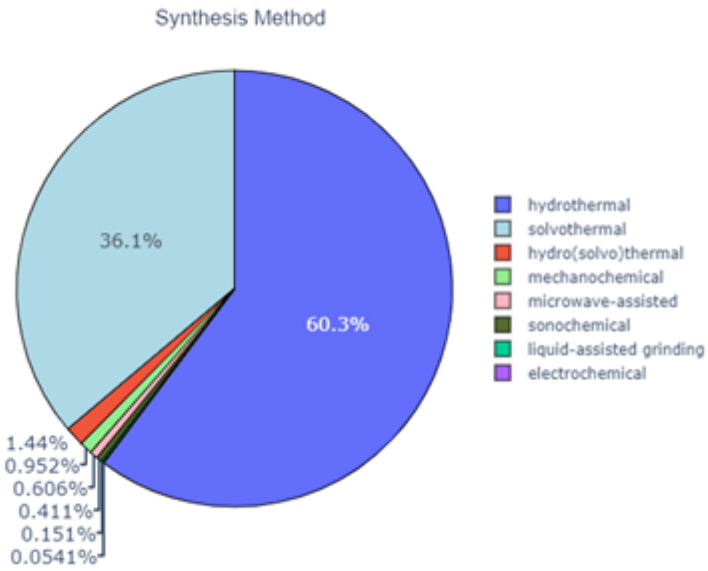


Figure A.2: *Proportion of synthesis methods present in the MOF Database.*

A.4 Data Transformation and Visualization

Following the data extraction process, the data was converted from a JSON format to a Microsoft Excel (.csv) file. While the filter was able to exclude many non-MOF names, it did miss some such as "[Co2]", "Cd-", and "Cu()" which were therefore removed using Excel's find and replace function. Additionally, the transfer of the data to Excel format led to the addition of special characters such as "Â", "â€, and "âˆž" and these were also removed. Furthermore, data that were obviously not linkers or metal precursors such as "KOH" and "NbO" were also deleted from the database with notes made of frequent misidentifications to be added to exclusion lists. During this transformation process, synonyms were also combined such as "DMF", "N,N-dimethylformamide", and "dimethylformamide" to ensure that data entries were only counted once. After the data was transformed, it was combined with the data extracted from the CSD using Excel's Power Query which combined the data based on the article download number which corresponded to the row number in the CSD thus matching the two separate data records. Figure A.3 shows the most extracted metal precursor results which were deemed not suitable to reported in the metal precursor histogram. This chart can be compared with Figure 5.4 e) within the main manuscript to reflect on the impact of conducting data transformation and augmentation following extraction.

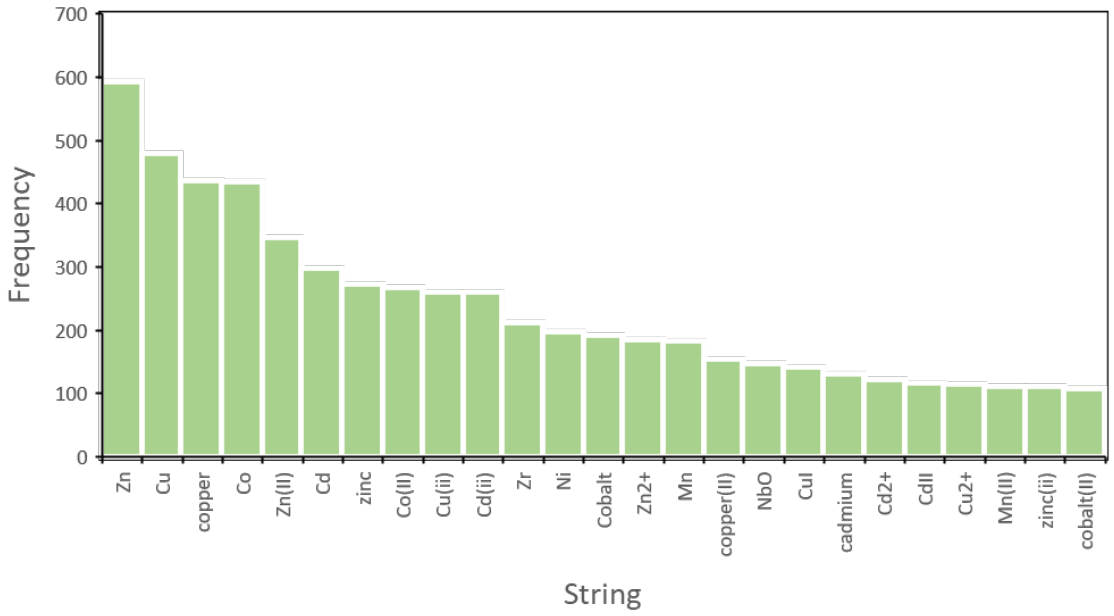


Figure A.3: *A histogram displaying the 25 most extracted strings marked up as metal precursors.*

A.5 Building Blocks and Topology

Further analysis was performed to compare the most common MOF building blocks and available topologies. The structures which most commonly reported topology and metal cluster in the experimental manuscripts were all metal nitrates, and primarily hydrated nitrates of transition metals. As for the linker types, these are primarily organic compounds which bond to metal clusters at each end of a straight chain. In Figure A.4, 'bipy' refers exclusively to 2,2'-bipyridine.

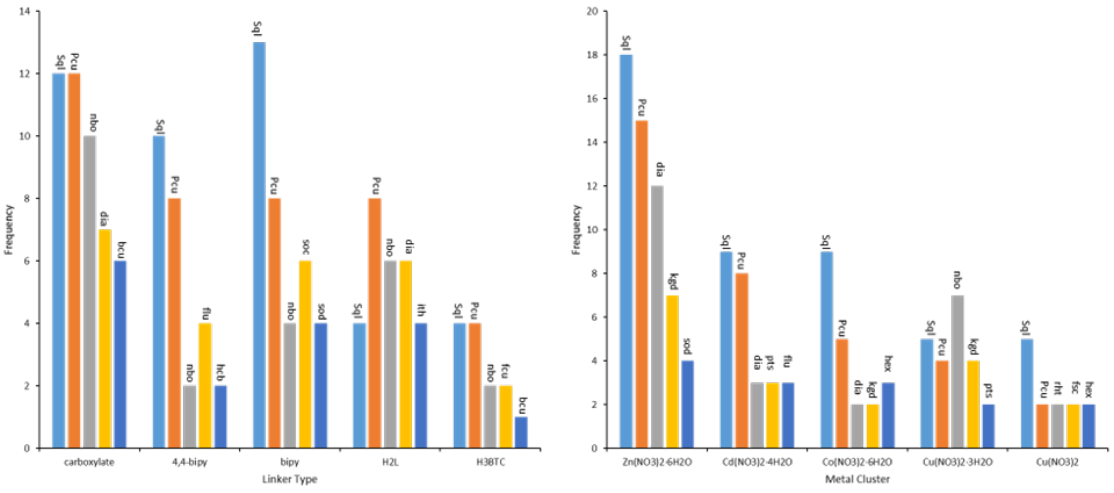


Figure A.4: *Clustered columns reflecting the top five topological allocations to a. the top five linker types, and b. the top five metal clusters.*

We also investigated the ratio of LCD/PLD of the building blocks with respect to the

topology. Figure A.5 shows the top 20 topologies for porous MOFs in the 3D MOF subset against the LCD/PLD ratio across the whole range of linker types. Here we note that the diversity of the LCD/PLD ratio suggests different pore accessibility, noting that the median ratio for each topology is different. For LCD/PLD ratio close to 1, we expect to see channel type pores, whereas for larger ratios we expect larger pores and small PLD values.

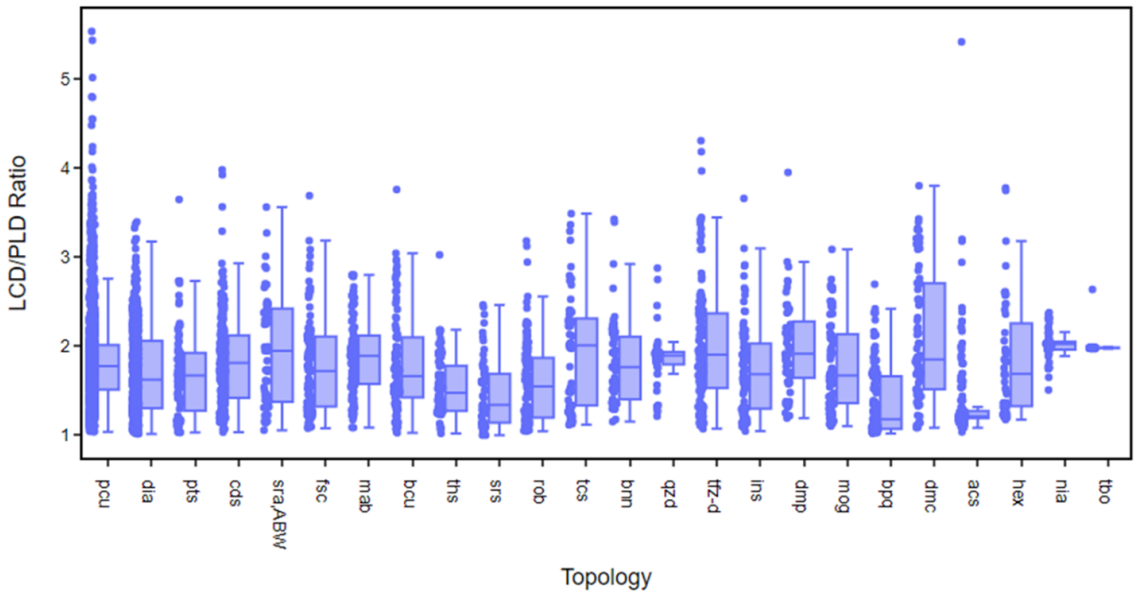


Figure A.5: *Top 20 topologies versus the LCD/PLD ratio in descending order of frequency for structures with PLD > 0.55 Å.*

Lastly, Figure A.6. shows a comparison of the linker length against the LCD/PLD ratio for the **pcu** topology. At shorter lengths, the ratio can be seen to reach a higher maximum value of 5.5, as well as a higher median value. This pattern of ratio decrease continues as the linker length increases. Given that the topology is the same across all of these structures, it would be unlikely that this change in ratio could be attributed to a structure being restricted to a single pore shape as opposed to a variety of mesopores and micropores. Whilst the pore sizes may vary, we would expect to see more uniform pore shapes for matching topologies, however given that pcu is one of the most basic nets, and that the topological assignment was performed using the Single Node algorithm, it is possible that these structures do form different structures and therefore do display some variety in micro and mesopores.

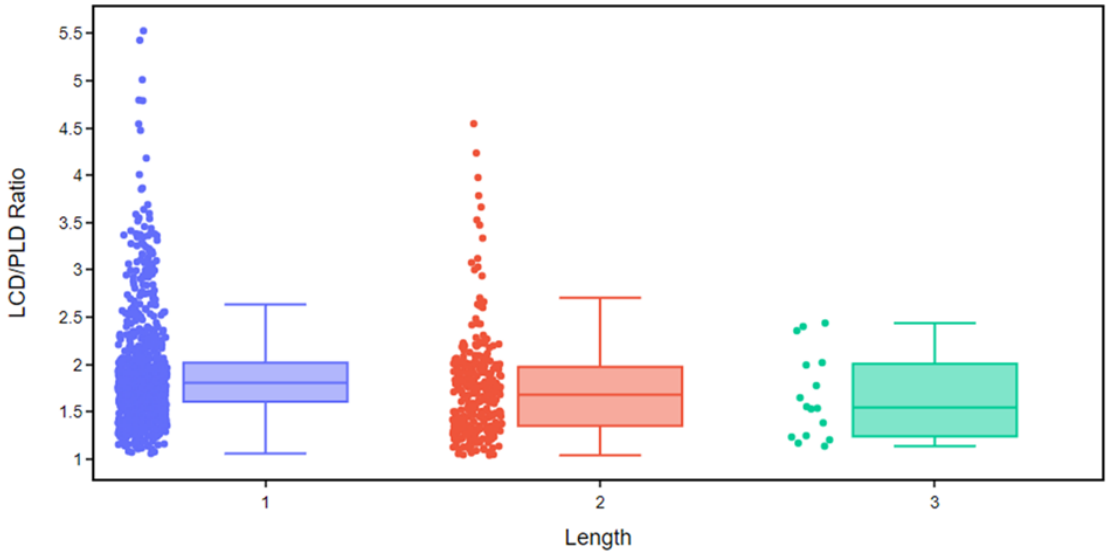


Figure A.6: *A box and whisker plot of linker length, as categorised by the number of aromatic rings, against LCD/PLD ratio for all porous MOFs which were assigned* ***pcu*** *topology.*

A.6 Text Mining Overview

An overview of the outcome of text mining can be found in Figure A.7 where the five main parameters are listed with their properties listed in order of recurrence frequency, alongside the CSD temperature values. All data presented here were obtained using the modified ChemDataExtractor found in the associated GitHub.

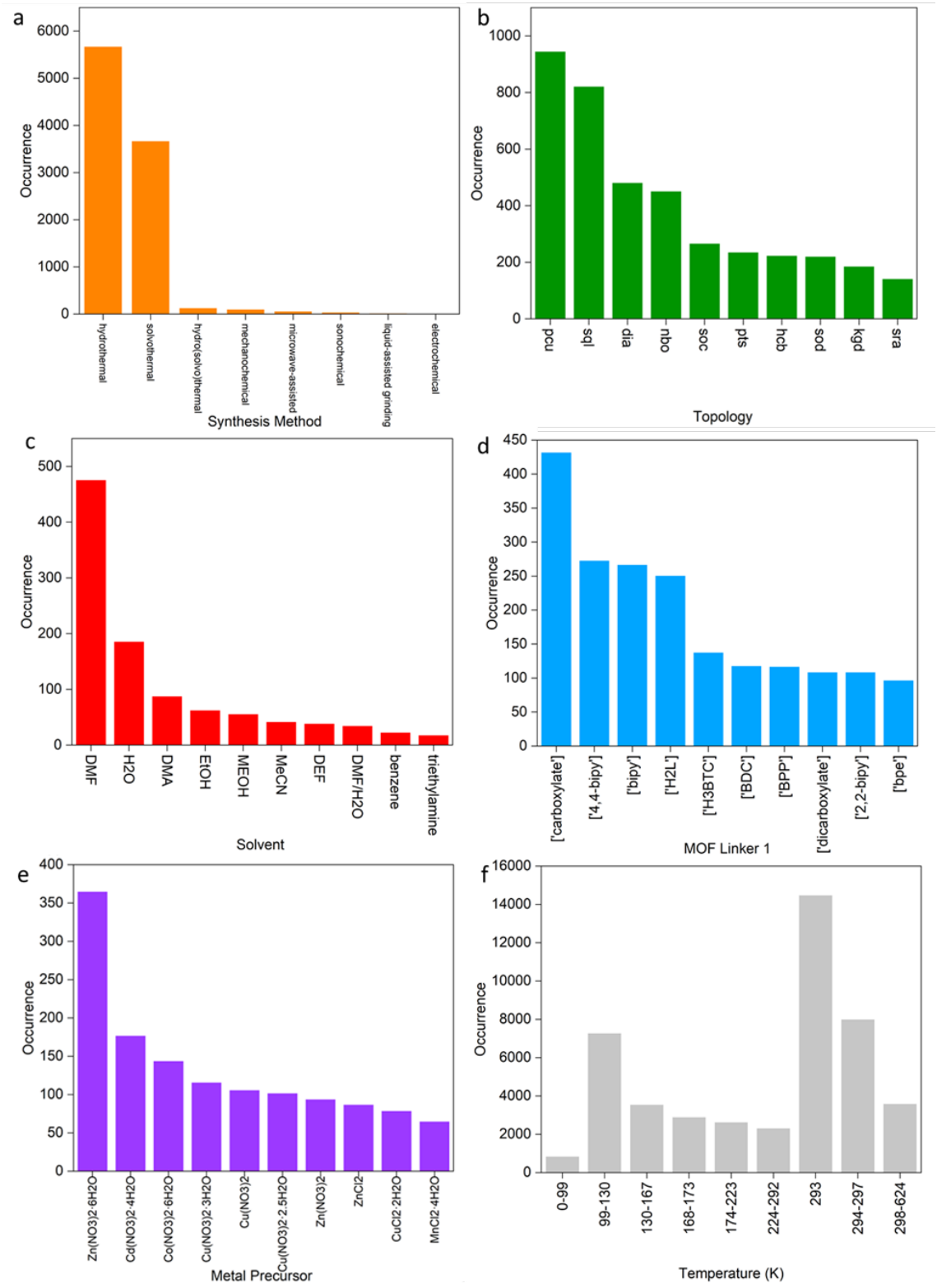


Figure A.7: *Histograms showing the most common MOF properties extracted in the DigiMOF database. a. synthesis methods, b. topologies, c. solvents, d. organic linkers, e. metal precursors, and f. temperature.*

A.7 Parsed Articles

The test set of 50 journal articles can be found below.

(1) Hu, X.-L.; Qin, C.; Wang, X.-L.; Shao, K.-Z.; Su, Z.-M. A Luminescent Dye@MOF as a Dual-Emitting Platform for Sensing Explosives. Chem. Commun. 2015, 51 (99), 17521–17524. https://doi.org/10.1039/C5CC07004J.

(2) Song, C.; Hu, J.; Ling, Y.; Feng, Y.; Krishna, R.; Chen, D.; He, Y. The Accessibility of Nitrogen Sites Makes a Difference in Selective CO2 Adsorption of a Family of Isostructural Metal–Organic Frameworks. J. Mater. Chem. A 2015, 3 (38), 19417–19426. https://doi.org/10.1039/C5TA05481H.

(3) Yi, F.-Y.; Jiang, H.-L.; Sun, Z.-M. Linearly Bridging CO2 in a Metal–Organic Framework. Chem. Commun. 2015, 51 (40), 8446–8449.
https://doi.org/10.1039/C5CC01244A.

(4) Ahrenholtz, S. R.; Landaverde-Alvarado, C.; Whiting, M.; Lin, S.; Slebodnick, C.; Marand, E.; Morris, A. J. Thermodynamic Study of CO2 Sorption by Polymorphic Microporous MOFs with Open Zn(II) Coordination Sites. Inorg Chem 2015, 54 (9), 4328–4336. https://doi.org/10.1021/ic503047y.

(5) Zhang, S.-Y.; Zhang, X.; Li, H.; Niu, Z.; Shi, W.; Cheng, P. Dual-Functionalized MetalOrganic Frameworks Constructed from Hexatopic Ligand for Selective CO2 Adsorption. Inorg Chem 2015, 54 (5), 2310–2314. https://doi.org/10.1021/ic502921j.

(6) Hu, Z.; Huang, G.; Lustig, W. P.; Wang, F.; Wang, H.; Teat, S. J.; Banerjee, D.; Zhang, D.; Li, J. Achieving Exceptionally High Luminescence Quantum Efficiency by Immobilizing an AIE Molecular Chromophore into a Metal–Organic Framework. Chem. Commun. 2015, 51 (15), 3045–3048. https://doi.org/10.1039/C4CC07642G.

(7) Paraschiv, C.; Cucos, A.; Shova, S.; Madalan, A.; Maxim, C.; Visinescu, D.; Cojocaru, B.; Parvulescu, V.; Andruh, M. New Zn(II) Coordination Polymers Constructed from AminoAlcohols and Aromatic Dicarboxylic Acids: Synthesis, Structure, Photocatalytic Properties, and Solid-State Conversion to ZnO. Crystal Growth & Design 2015, 15, 799–811. https://doi.org/10.1021/cg501604c.

(8) Yang, F.; Zheng, Q.; Chen, Z.; Ling, Y.; Liu, X.; Weng, L.; Zhou, Y. A Three-Dimensional Structure Built of Paddle-Wheel and Triazolate-Dinuclear Metal Clusters: Synthesis, Deformation and Reformation of Paddle-Wheel Unit in the Single-Crystal-to-Single-Crystal Transformation. CrystEngComm 2013, 15 (35), 7031–7037.
https://doi.org/10.1039/C3CE40855H.

(9) Yang, Y.-Y.; Lin, Z.-J.; Liu, T.-T.; Liang, J.; Cao, R. Synthesis, Structures and Physical Properties of Mixed-Ligand Coordination Polymers Based on a V-Shaped Dicarboxylic Ligand. CrystEngComm 2015, 17 (6), 1381–1388.
https://doi.org/10.1039/C4CE02163K.

(10) Tu, B.; Pang, Q.; Wu, D.; Song, Y.; Weng, L.-H.; Li, Q. Ordered Vacancies and Their Chemistry in Metal Organic Frameworks. Journal of the American Chemical Society 2014, 136. https://doi.org/10.1021/ja5063423.

(11) Kapelewski, M. T.; Geier, S. J.; Hudson, M. R.; Stück, D.; Mason, J. A.; Nelson,

J. N.; Xiao, D. J.; Hulvey, Z.; Gilmour, E.; FitzGerald, S. A.; Head-Gordon, M.; Brown, C. M.; Long, J.R. M2(m-Dobdc) (M = Mg, Mn, Fe, Co, Ni) Metal-Organic Frameworks Exhibiting Increased Charge Density and Enhanced H2 Binding at the Open Metal Sites. J Am Chem Soc 2014, 136 (34), 12119–12129. https://doi.org/10.1021/ja506230r.

(12) Zhao, N.; Sun, F. X.; He, H.; Jia, J.; Zhu, G. Solvent-Induced Single Crystal To Single Crystal Transformation and Complete Metal Exchange of a Pyrene-Based Metal–Organic Framework. Crystal Growth & Design 2014, 14, 1738–1743. https://doi.org/10.1021/cg401887b.

(13) Patel, D. G. (Dan); Walton, I. M.; Cox, J. M.; Gleason, C. J.; Butzer, D. R.; Benedict, J. B. Photoresponsive Porous Materials: The Design and Synthesis of Photochromic DiaryletheneBased Linkers and a Metal–Organic Framework. Chem. Commun. 2014, 50 (20), 2653–2656. https://doi.org/10.1039/C3CC49666J.

(14) Crane, A. K.; Wong, E. Y. L.; MacLachlan, M. J. Metal–Organic Frameworks from Novel Flexible Triptycene- and Pentiptycene-Based Ligands. CrystEngComm 2013, 15 (45), 9811–9819. https://doi.org/10.1039/C3CE41459K.

(15) Makal, T. A.; Zhuang, W.; Zhou, H.-C. Realization of Both High Hydrogen Selectivity and Capacity in a Guest Responsive Metal–Organic Framework. J. Mater. Chem. A 2013, 1 (43), 13502–13509. https://doi.org/10.1039/C3TA12761C.

(16) Schoedel, A.; Boyette, W.; Wojtas, L.; Eddaoudi, M.; Zaworotko, M. J. A Family of Porous Lonsdaleite-e Networks Obtained through Pillaring of Decorated Kagomé Lattice Sheets. J Am Chem Soc 2013, 135 (38), 14016–14019. https://doi.org/10.1021/ja406030p.

(17) Han, L.; Xu, L.-P.; Qin, L.; Zhao, W.-N.; Yan, X.-Z.; Yu, L. Syntheses, Crystal Structures, and Physical Properties of Two Noninterpenetrated Pillar-Layered Metal Organic Frameworks Based on N,N-Di(4-Pyridyl)-1,4,5,8-Naphthalenetetracarboxydiimide Pillar. Crystal Growth & Design 2013, 13, 4260–4267. https://doi.org/10.1021/cg400454c.

(18) Li, J.-R.; Yu, J.; Lu, W.; Sun, L.-B.; Sculley, J.; Balbuena, P. B.; Zhou, H.-C. Porous Materials with Pre-Designed Single-Molecule Traps for CO2 Selective Adsorption. Nat Commun 2013, 4 (1), 1538. https://doi.org/10.1038/ncomms2552.

(19) Dau, P. V.; Polanco, L. R.; Cohen, S. M. Dioxole Functionalized Metal–Organic Frameworks. Dalton Trans. 2013, 42 (11), 4013–4018. https://doi.org/10.1039/C3DT32588A.

(20) Qin, Y.; Feng, X.; Luo, F.; Sun, G.; Song, Y.; Tian, X.; Huang, H.; Zhu, Y.; Yuan, Z.; Luo, M.; Liu, S.; Xu, W. A Microporous Metal–Organic Framework Containing an Exceptional FourConnecting 4264 Topology and a Combined Effect for Highly Selective Adsorption of CO2 over N2. Dalton Trans. 2012, 42 (1), 50–53. https://doi.org/10.1039/C2DT31905E.

(21) Zhang, J.-P.; Zhu, A.-X.; Chen, X.-M. Single-Crystal X-Ray Diffraction and Raman Spectroscopy Studies of Isobaric N2 Adsorption in SOD-Type Metal–Organic Zeolites. Chem. Commun. 2012, 48 (93), 11395–11397. https://doi.org/10.1039/C2CC35544B.

(22) Chen, Z.; Zhang, C.; Liu, X.; Zhang, Z.; Liang, F.; Chen, Z.; Zhang, C.; Liu,

X.; Zhang, Z.; Liang, F. Synthesis, Structure, and Properties of a Chiral Zinc(II) Metal-Organic Framework Featuring Linear Trinuclear Secondary Building Blocks. Aust. J. Chem. 2012, 65 (12), 1662–1666. https://doi.org/10.1071/CH12270.

(23) Hou, C.; Liu, Q.; Fan, J.; Zhao, Y.; Wang, P.; Sun, W.-Y. Novel (3,4,6)-Connected MetalOrganic Framework with High Stability and Gas-Uptake Capability. Inorg Chem 2012, 51 (15), 8402–8408. https://doi.org/10.1021/ic300950h.

(24) He, J.-H.; Sun, D.-Z.; Xiao, D.-R.; Yan, S.-W.; Chen, H.-Y.; Wang, X.; Yang, J.; Wang, E.-B. Syntheses and Structures of Five 1D Coordination Polymers Based on Quinolone Antibacterial Agents and Aromatic Polycarboxylate Ligands. Polyhedron 2012, 42 (1), 24–29. https://doi.org/10.1016/j.poly.2012.04.022.

(25) Guo, M.; Sun, Z.-M. Solvents Control over the Degree of Interpenetration in MetalOrganic Frameworks and Their High Sensitivities for Detecting Nitrobenzene at Ppm Level. J. Mater. Chem. 2012, 22 (31), 15939–15946.
https://doi.org/10.1039/C2JM32066E.

(26) Shi, D.; Ren, Y.; Jiang, H.; Cai, B.; Lu, J. Synthesis, Structures, and Properties of Two Three-Dimensional Metal-Organic Frameworks, Based on Concurrent Ligand Extension. Inorg Chem 2012, 51 (12), 6498–6506. https://doi.org/10.1021/ic202624e.

(27) Kanoo, P.; Ghosh, A. C.; Cyriac, S. T.; Maji, T. K. A Metal–Organic Framework with Highly Polar Pore Surfaces: Selective CO2 Adsorption and Guest-Dependent On/Off Emission Properties. Chemistry – A European Journal 2012, 18 (1), 237–244. https://doi.org/10.1002/chem.201101183.

(28) Feng, D.; Gu, Z.-Y.; Li, J.-R.; Jiang, H.-L.; Wei, Z.; Zhou, H.-C. ZirconiumMetalloporphyrin PCN-222: Mesoporous Metal–Organic Frameworks with Ultrahigh Stability as Biomimetic Catalysts. Angewandte Chemie International Edition 2012, 51 (41), 10307–10310. https://doi.org/10.1002/anie.201204475.

(29) Wu, P.; Wang, J.; He, C.; Zhang, X.; Wang, Y.; Liu, T.; Duan, C. Luminescent MetalOrganic Frameworks for Selectively Sensing Nitric Oxide in an Aqueous Solution and in Living Cells. Advanced Functional Materials 2012, 22 (8), 1698–1703. https://doi.org/10.1002/adfm.201102157.

(30) Schaate, A.; Roy, P.; Preuße, T.; Lohmeier, S. J.; Godt, A.; Behrens, P. Porous Interpenetrated Zirconium–Organic Frameworks (PIZOFs): A Chemically Versatile Family of Metal–Organic Frameworks. Chemistry – A European Journal 2011, 17 (34), 9320–9325. https://doi.org/10.1002/chem.201101015.

(31) Liu, D.; Xie, Z.; Ma, L.; Lin, W. Three-Dimensional Metal-Organic Frameworks Based on Tetrahedral and Square-Planar Building Blocks: Hydrogen Sorption and Dye Uptake Studies. Inorg Chem 2010, 49 (20), 9107–9109. https://doi.org/10.1021/ic1009169.

(32) Zhuang, W.; Ma, S.; Wang, X.-S.; Yuan, D.; Li, J.-R.; Zhao, D.; Zhou, H.-C. Introduction of Cavities up to 4 Nm into a Hierarchically-Assembled Metal–Organic Framework Using an Angular, Tetratopic Ligand. Chem. Commun. 2010, 46 (29), 5223–5225. https://doi.org/10.1039/C0CC00779J.

(33) Barquín, M.; Cocera, N.; González Garmendia, M. J.; Larrínaga, L.; Pinilla, E.; Torres, M. R. Acetato and Formato Copper(II) Paddle-Wheel Complexes with Nitrogen

Ligands. Journal of Coordination Chemistry 2010, 63 (13), 2247–2260. https://doi.org/10.1080/00958972.2010.502227.

(34) Tian, Y.-Q.; Yao, S.-Y.; Gu, D.; Cui, K.-H.; Guo, D.-W.; Zhang, G.; Chen, Z.-X.; Zhao, D.-Y. Cadmium Imidazolate Frameworks with Polymorphism, High Thermal Stability, and a Large Surface Area. Chemistry – A European Journal 2010, 16 (4), 1137–1141. https://doi.org/10.1002/chem.200902729.

(35) Kishan, M. R.; Tian, J.; Thallapally, P. K.; Fernandez, C. A.; Dalgarno, S. J.; Warren, J. E.; McGrail, B. P.; Atwood, J. L. Flexible Metal–Organic Supramolecular Isomers for Gas Separation. Chem. Commun. 2010, 46 (4), 538–540. https://doi.org/10.1039/B913910A.

(36) Wang, X.-Z.; Zhu, D.; Xu, Y.; Yang, J.; Shen, X.; Zhou, J.; Fei, N.; Ke, X.; Peng, L. Three Novel Metal-Organic Frameworks with Different Topologies Based on 3,3-Dimethoxy-4,4-Biphenyldicarboxylic Acid: Syntheses, Structures, and Properties. 2010. https://doi.org/10.1021/CG9012262.

(37) Hong, S.; Oh, M.; Park, M.; Yoon, J. W.; Chang, J.-S.; Lah, M. S. Large H2 Storage Capacity of a New Polyhedron-Based Metal–Organic Framework with High Thermal and Hygroscopic Stability. Chem. Commun. 2009, No. 36, 5397–5399. https://doi.org/10.1039/B909250A.

(38) Xiang, S.; Zhou, W.; Gallegos, J. M.; Liu, Y.; Chen, B. Exceptionally High Acetylene Uptake in a Microporous Metal-Organic Framework with Open Metal Sites. J Am Chem Soc 2009, 131 (34), 12415–12419. https://doi.org/10.1021/ja904782h.

(39) Dai, F.; He, H.; Gao, D.; Ye, F.; Sun, D.; Pang, Z.; Zhang, L.; Dong, G.; Zhang, C. SelfAssembly of 2D Zinc Metal–Organic Frameworks Based on Mixed Organic Ligands. Inorganica Chimica Acta 2009, 362 (11), 3987–3992. https://doi.org/10.1016/j.ica.2009.05.038.

(40) Ma, S.; Wang, S.; Collier, C.; Manis, S.; Zhou, H.-C. Ultramicroporous Metal-Organic Framework Based on 9, 10-Anthracenedicarboxylate for Selective Gas Adsorption Inorg. Inorganic chemistry 2007, 46, 8499–8501. https://doi.org/10.1021/ic701507r.

(41) Ma, Y.; Han, Z.; He, Y.; Yang, L. A 3D Chiral Zn(II) Coordination Polymer with Triple Zn–Oba–Zn Helical Chains (Oba = 4,4-Oxybis(Benzoate)). Chem. Commun. 2007, No. 40, 4107–4109. https://doi.org/10.1039/B708479J.

(42) Sun, D.; Ke, Y.; Mattox, T. M.; Parkin, S.; Zhou, H.-C. Stability and Porosity Enhancement through Concurrent Ligand Extension and Secondary Building Unit Stabilization. Inorg Chem 2006, 45 (19), 7566–7568. https://doi.org/10.1021/ic0609002.

(43) Sun, D.; Ke, Y.; Collins, D.; Lorigan, G.; Zhou, H.-C. Construction of Robust Open Metal-Organic Frameworks with Chiral Channels and Permanent Porosity. Inorganic chemistry 2007, 46, 2725–2734. https://doi.org/10.1021/ic0624773.

(44) Clausen, H. F.; Poulsen, R. D.; Bond, A. D.; Chevallier, M.-A. S.; Iversen, B. B. Solvothermal Synthesis of New Metal Organic Framework Structures in the Zinc Terephthalic Acid Dimethyl Formamide System. Journal of Solid State Chemistry 2005, 178 (11), 3342–3351. https://doi.org/10.1016/j.jssc.2005.08.013.

(45) Chun, H.; Dybtsev, D. N.; Kim, H.; Kim, K. Synthesis, X-Ray Crystal Structures,

and Gas Sorption Properties of Pillared Square Grid Nets Based on Paddle-Wheel Motifs: Implications for Hydrogen Storage in Porous Materials. Chemistry – A European Journal 2005, 11 (12), 3521– 3529. https://doi.org/10.1002/chem.200401201.

(46) Adams, H.; Fenton, D. E.; McHugh, P. E. A Heteronuclear [Nickel(II)–Sodium] Infinite Chain Complex Derived from 3-[(2-Diethylamino-Ethyl)-Methyl-Amino]-Methyl-2-Hydroxy5-Methyl-Benzaldehyde. Inorganic Chemistry Communications 2004, 7 (7), 880–883. https://doi.org/10.1016/j.inoche.2004.04.025.

(47) Dong, Y.-B.; Ma, J.-P.; Smith, M. D.; Huang, R.-Q.; Tang, B.; Chen, D.; zur Loye, H.-C. New Coordination Polymers Generated from Oxadiazole-Containing Bidentate Ligands and Cu-Cu Dimetal Units. Solid State Sciences 2002, 4 (10), 1313–1320. https://doi.org/10.1016/S1293-2558(02)00014-6.

(48) Tao, J.; Yin, X.; Huang, R.; Zheng, L.; Weng Ng, S. Assembly of a Microporous MetalOrganic Framework [Zn(Bpdc)(DMSO)] (Bpdc=4,4-biphenyldicarboxylate) Based on PaddleWheel Units Affording Guest Inclusion. Inorganic Chemistry Communications 2002, 5 (11), 975–977. https://doi.org/10.1016/S1387-7003(02)00623-8.

(49) Papaefstathiou, G. S.; MacGillivray, L. R. An Inverted Metal-Organic Framework with Compartmentalized Cavities Constructed by Using an Organic Bridging Unit Derived from the Solid State. Angewandte Chemie International Edition 2002, 41 (12), 2070–2073.

(50) Medishetty R.; Jung D.; Song X.; Kim D.; Lee S. S.; Lah M. S.; Vittal J. J. Solvent-Induced Structural Dynamics in Noninterpenetrating Porous Coordination Polymeric Networks. Inorg. Chem. 2013, 52 (6), 2951–2957.
https://doi.org/10.1021/ic302334x.

References

[1] Shu Huang and Jacqueline M. Cole. A database of battery materials auto-generated using ChemDataExtractor. *Sci Data*, 7(1):260, August 2020. Number: 1 Publisher: Nature Publishing Group.

[2] Callum J. Court and Jacqueline M. Cole. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Sci Data*, 5(1):180111, June 2018. Publisher: Nature Publishing Group.

[3] Sanghoon Park, Baekjun Kim, Sihoon Choi, Peter G. Boyd, Berend Smit, and Jihan Kim. Text Mining Metal–Organic Framework Papers. *J. Chem. Inf. Model.*, 58(2):244–251, February 2018. Publisher: American Chemical Society.

[4] Yi Luo, Saientan Bag, Orysia Zaremba, Adrian Cierpka, Jacopo Andreo, Stefan Wuttke, Pascal Friederich, and Manuel Tsotsalas. MOF Synthesis Prediction Enabled by Automatic Data Mining and Machine Learning**. *Angewandte Chemie International Edition*, 61(19):e202200242, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.202200242.

Appendix B

Supporting Information for Augmented Reality

B.1 Visualising MOFs with Augmented Reality (AR)

A step-by-step guide to create AR models of MOFs hosted by p3d.in from the Cambridge Structural Database (CSD), with further instructions on how to manipulate RASPA movie files, or topology output files.

B.1.1 Requirements:

CSD Mercury - https://www.ccdc.cam.ac.uk/support-and-resources/download-the-csd/

Jmol - http://jmol.sourceforge.net/download/

Blender - https://www.blender.org/download/

An active p3d.in account - https://p3d.in

(Optional but required for gas adsorption representations)

RASPA - https://iraspa.org/raspa/

(Optional but required for topology representations - choice of either)

CrystalNets - https://github.com/coudertlab/CrystalNets.jl

ToposPro - https://topospro.com/software/topospro/download/

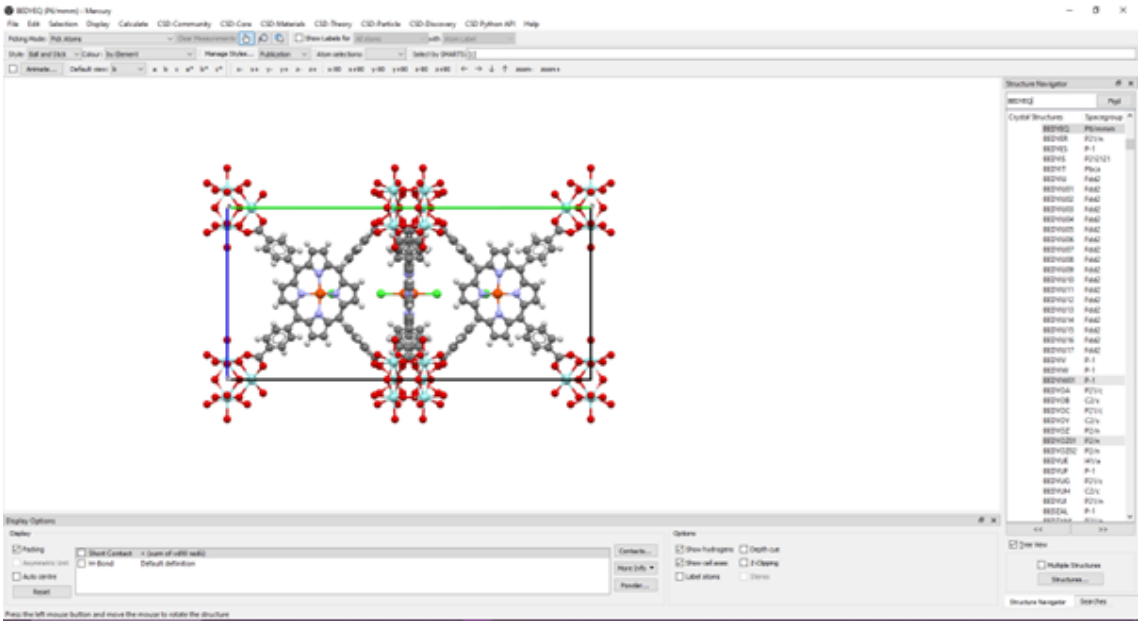
B.2 AR File Creation Method

B.2.1 Part A – Selecting and Modifying Files:

1. Install and licence CSD Mercury and ensure the database is the latest version.

2. Select a CSD refcode that you would like to develop an AR interaction for (or have a file available for the structure you are interested in, in PDB, MOL2, or CIF format from any other resource, including RASPA Movie outputs)

3. Open the structure in CSD Mercury by either searching for it by refcode or dragging the file to the viewing window (here we search for BEDYEQ). Ensure that Packing (bottom left) is selected to display the unit cell. This is important as what we see – is what we get.

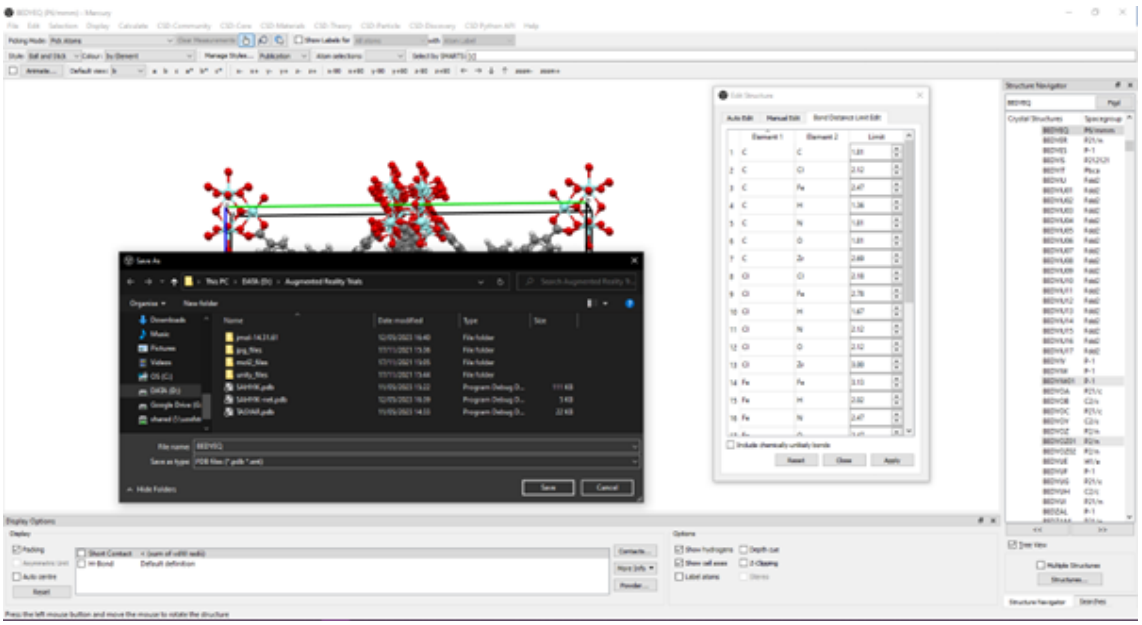


4. Once the file is loaded, make any changes to the structure that you see fit (e.g remove unbound solvents, trim the edges of the unit cell to make it more uniform etc.) You can do this by going to Edit -> Edit Structure. Find Remove at the bottom and click Atoms & Bonds, then click any atom or bond to remove it.

   Note: If you are loading a PDB file, the bonding information may be corrupted. To solve this, go to Edit -> Edit Bond Distance Limit and click Apply. This should reset the bonding, but it is important at this stage to check your structure and manually verify that the bonding is correct. If you find any abnormalities, then they can be corrected using the Edit Structure window as above to remove or change the bond types.

5. When you are content with the structure representation, you should save it as a new PDB file (eg. refcode_new.pdb). Ensure you have made all final structure changes before this step. It is also possible to save the file as a CIF at this stage, it should not make a difference.

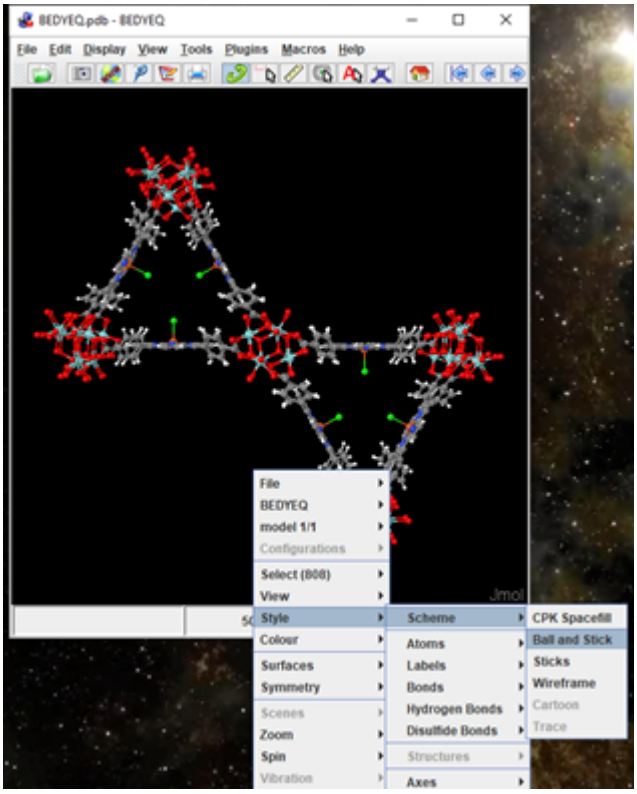
6. You can now close CSD Mercury.



B.2.2 Part B - Jmol:

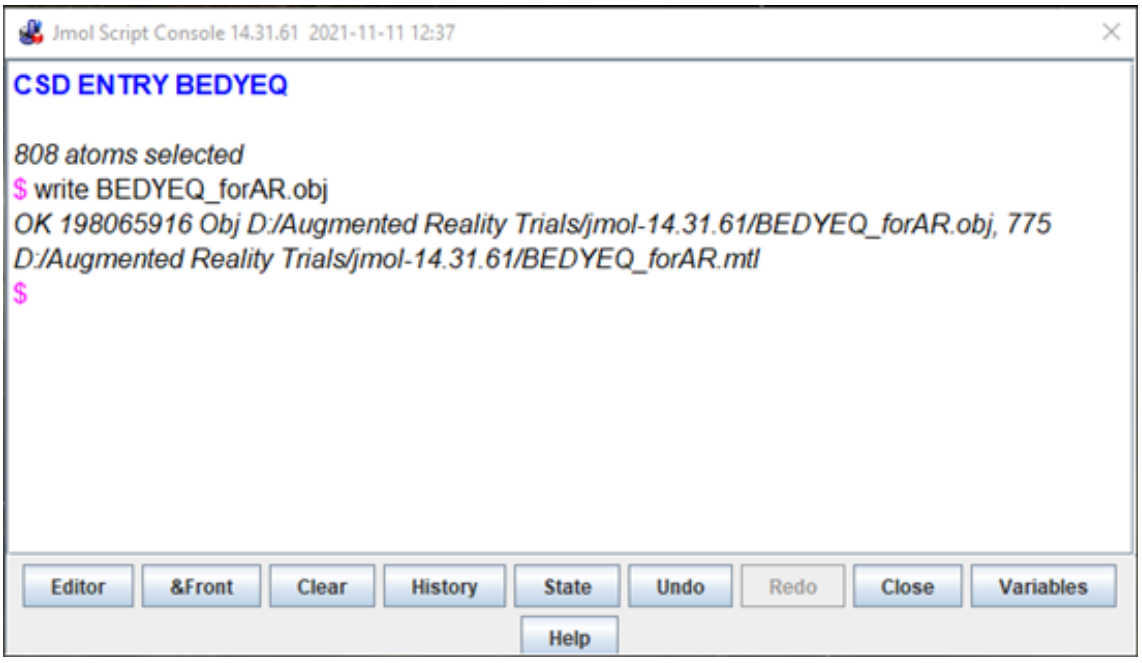
1. Open your new PDB file in Jmol.

2. Right Click and go to Style -> Scheme -> Ball and Stick. Make sure the that the representation is in this format as once the file has been saved, it will stay in this representation. (Although you are free to choose whichever format you prefer.)



3. Go to File -> Console... This should open the console, where it will load the file name, e.g CSD ENTRY BEDYEQ. Next to the pink $ enter: write new_file_name.obj and press enter. This will print an object (OBJ) file into the default Jmol output folder with the name you have assigned it. If the command is incorrect, the text will turn red.

4. Wait for the console to give the OK, and the file path of your new file will be shown on screen.

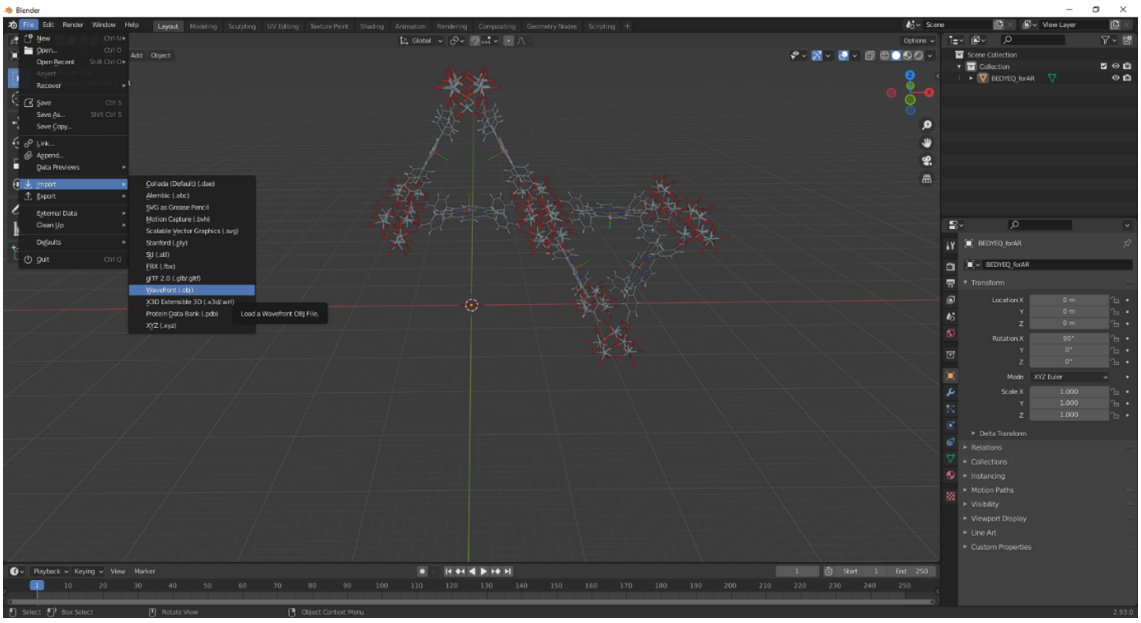


5. You can now close Jmol.

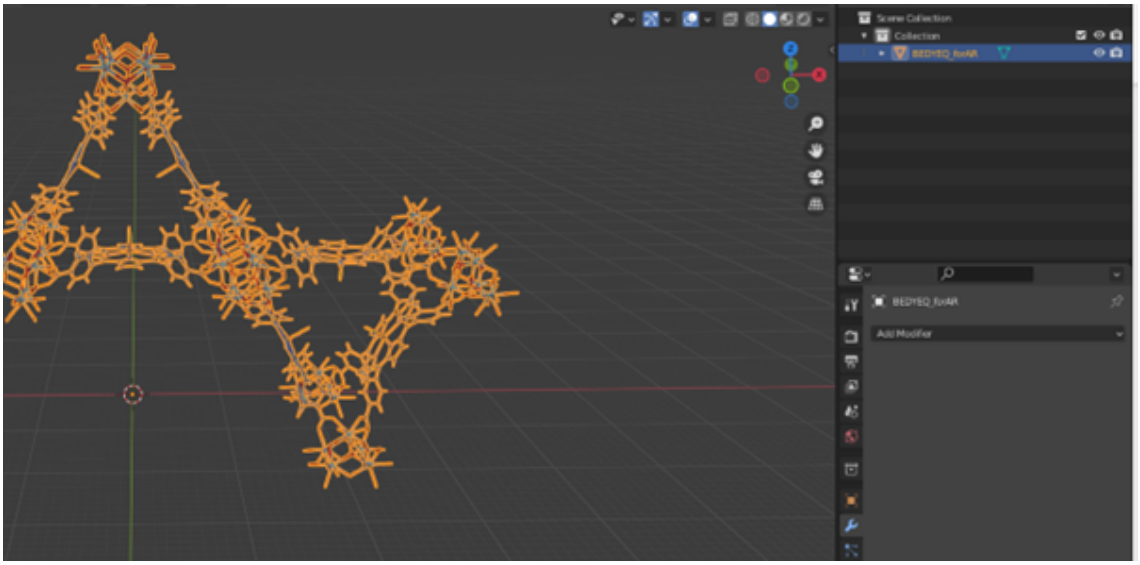
B.2.3 Part C - Blender:

1. Open Blender. Select a new general starting platform. Delete the default grey cube by selecting it and pressing delete. In the top right-hand corner, also delete the camera, and light layers by clicking on them and pressing delete. (This is required to keep file sizes low, so that the detail of the AR figure is kept at a maximum).

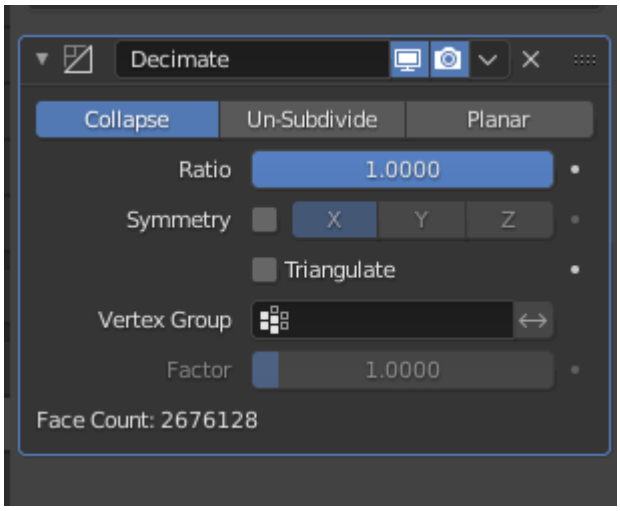
2. Go to File -> Import -> Wavefront (.obj) and import your object file from the output folder of Jmol. (You may need to wait for it to render, these files can be >250MB and may take some time to load.)



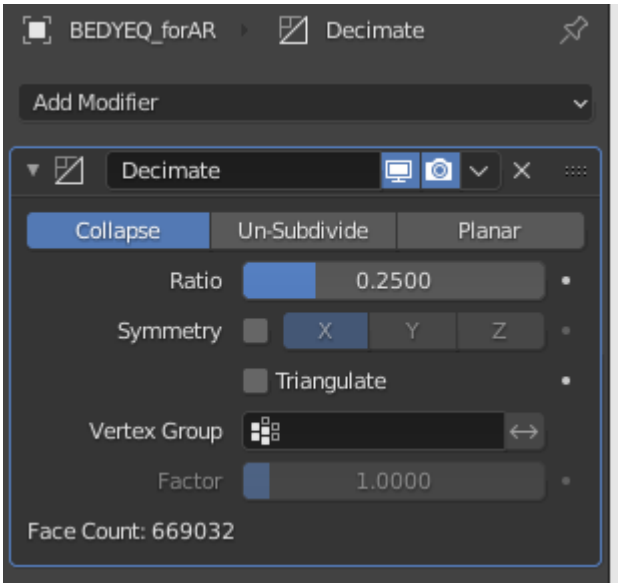
3. Select the object using left click and find the Spanner (Wrench) icon in the right-hand side panel. Here it will ask you to “Add Modifier”.



4. Select Add Modifier -> Decimate (in the Generate column). Check the Face Count (FC) of your object – for AR this must be <750,000, if it is higher, calculate the required ratio adjustment to bring it down to an acceptable level.



5. In this case, the FC is 3.57 times larger than acceptable. Here we have adjusted the ratio to 0.25 to bring it within an acceptable limit – try to get the FC to as close to 750,000 as possible as to not compromise on the quality of the representation.

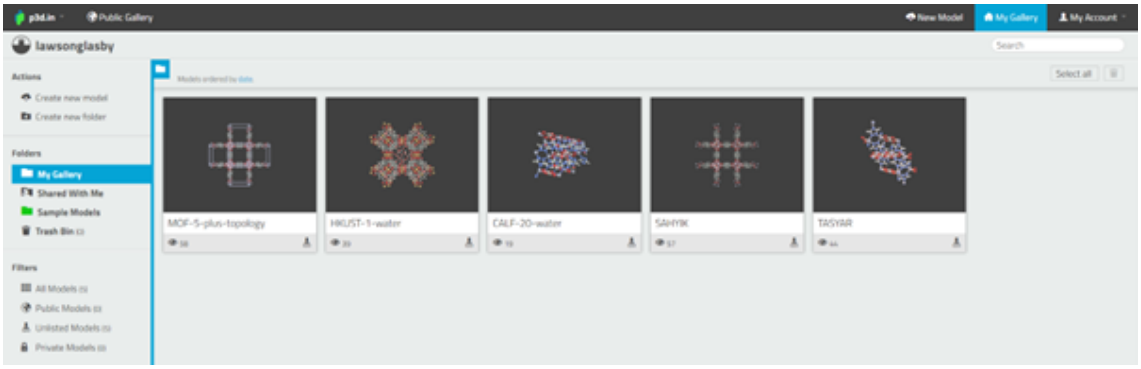


6. Now we need to export the file into a 3D rendering format. Go to File -> Export -> FBX (.fbx) and save the file as something appropriate, e.g refcode_AR.fbx.

7. You can now close Blender.

B.2.4 Part D - p3d.in:

1. Go to p3d.in and sign up as a Free User. Once you have logged in, you should see the dashboard.



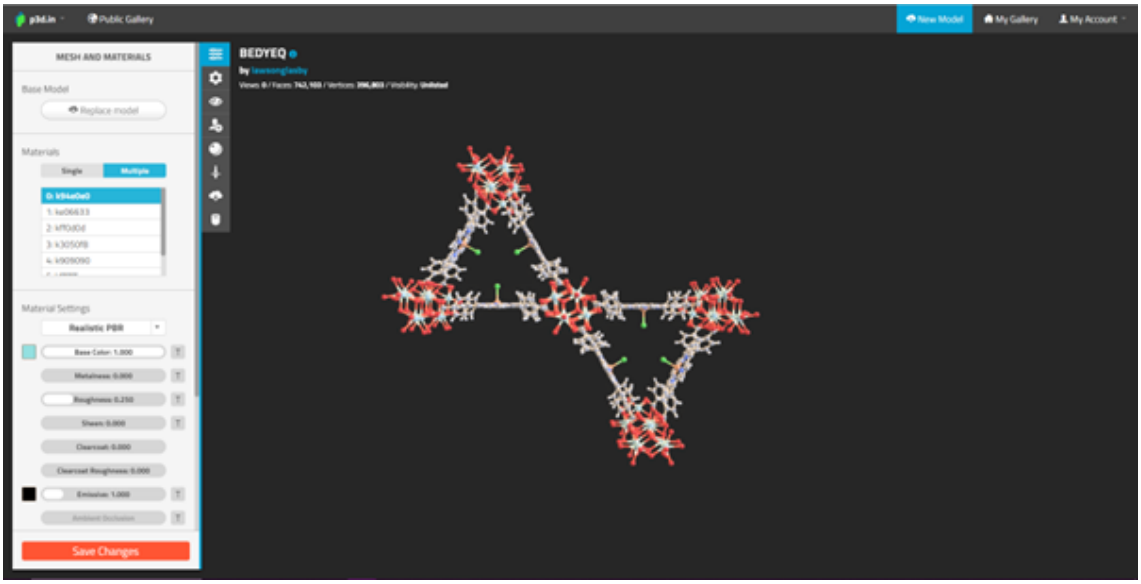
2. From here, select "Create new model" in the top left. Drag your FBX file into the box and wait for it to render. Alternatively, you can upload the file from the file explorer.

   Note: When using the free version your FBX file must be smaller than 50 MB. If you use the decimate tool, then it typically will condense the file to somewhere between 7-20 MB.

3. Once the model has rendered you can make changes to the colours of the object, and the background by following the settings available to you via the modification pane on the left-hand side of the website.

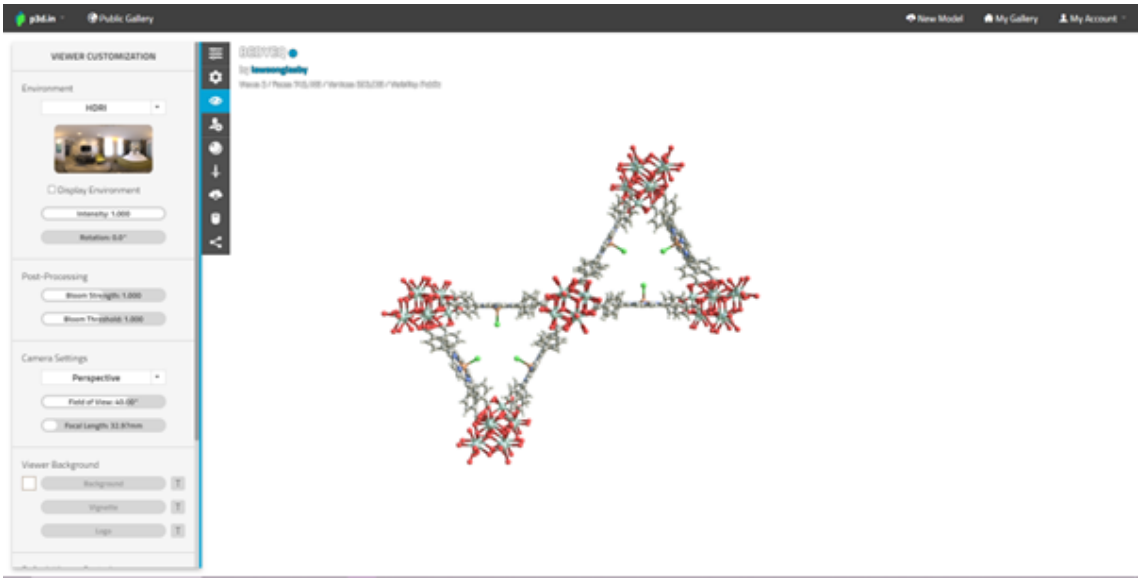
4. In Mesh and Materials, you will find a list of atoms – the colour of all C atoms will be modified by changing the colour of this entry in the list.

   We feel that changing the atom colours to "Classic" is a better representation than the default "Realistic PBR". These atoms colours can easily be swapped in a drop-down menu.



5. Background settings can be changed in Viewer Customisation. Find Viewer Background and set the background colour to white. Ensure in this section that Augmented Reality is set to "Enable". (This is per your preference but we find either black or white works best, although there are also options to select an environment such as hotel room.)

6. Save changes to exit and publish the structure.

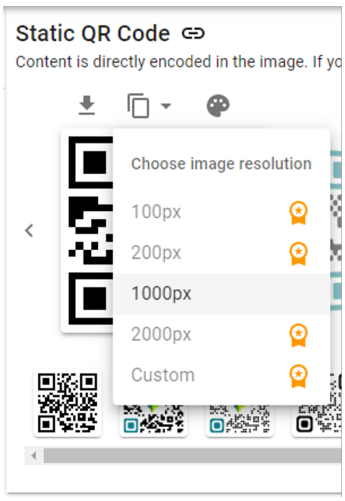


7. Copy the new unique URL from the published structure. You can choose to keep the structure private such that it can only be accessed via that URL (and corresponding QR) or you can publish it publicly, or hide it completely.

B.2.5 Part E - QR Code:

1. Take the unique URL published from the structure, in the format of p3d.in/xxxxx and visit the-qrcode-generator.com/. Select static QR generator, choose URL and paste the link in the Enter URL box. You can then customise the adjacent QR code.

2. Download the 1000px version of the QR code that is unique to your new structure and keep it safe.



3. Post your QR on research posters, websites, journal articles, and more!

4. If at any time you would like to modify your structure, you can edit the file corresponding to this QR code. Log back in to p3d.in and edit the relevant link, here you can upload new FBX files entirely and retain the same URL, you can also modify any of the settings at any time and they will be applied as soon as you save the entry again.

5.

B.3 RASPA AR file generation for gas adsorption visualisation

1. Download RASPA. (If you are not familiar with RASPA there are several online guides and workshops available on the iRASPA website linked at the beginning of this document).

2. Submit a CIF to RASPA and define your input parameters. Ensure that the "Movie" parameter is set to "yes" and specify the number of iterations between snapshots.

```
SimulationType MonteCarlo
NumberOfCycles 20000
NumberOfInitializationCycles 20000
PrintEvery 5000
Restartfile no

RemoveAtomNumberCodeFromLabel yes

Movies yes
WriteMoviesEvery 1000
ComputeRDF yes
WriteRDFEvery 1000
RDFHistogramSize 100
RDFRange 12.0

CutOffVDW 12.8
Forcefield MOF_DFF_UFF_rama

Framework 0
FrameworkName SIFSIX_unopt_charge
ChargeMethod Ewald
UseChargesFromCIFFile yes

UnitCells 4 4 4
HeliumVoidFraction 0.297
ExternalTemperature 298
ExternalPressure 100000

Component 0 MoleculeName              CO2
            MoleculeDefinition        TraPPE
            IdealGasRosenbluthWeight  1.0
            TranslationProbability    1.0
            RotationProbability       1.0
            ReinsertionProbability    1.o
            SwapProbability           1.0
            CreateNumberOfMolecules   0
```

3. Run RASPA. Once the simulations are completed, a Movie output folder will have been created containing the snapshots at each point of the isotherm. Select the one you would like to represent in AR and copy either the CIF or PDB file to a new folder where you will begin to create the AR representations.

4. Follow the steps in this guide from Part A.

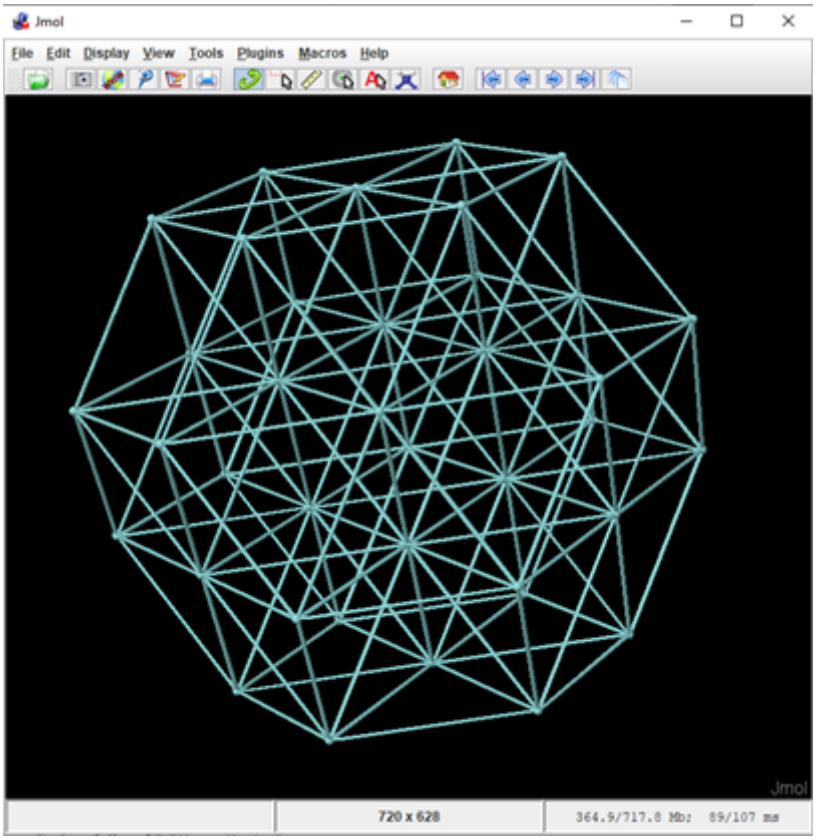
B.4 CrystalNets AR file generation to visualise topology

B.4.1 Part I: Obtaining OBJ format topological nets.

1. Download the relevant CIF file for your chosen structure.

2. Visit https://progs.coudert.name/topology and upload the CIF, cycle through the main options and select settings that are relevant to the chosen crystal. For a MOF, we selected the following settings: Structure Type: MOF, Bonding: Auto, Clusterings: SingleNodes, Exports: Trimmed, Subnets, Clusters.

3. The output file will be obtained in a VTF format and can be downloaded from the subnets section. Note: To view these nets in Mercury, it is then necessary to convert the output VTF file into a MOL2 format. VTF files cannot be opened in Merucry or Jmol, however MOL2 can be opened in Jmol and converted into an OBJ file, as in Part B of this guide.

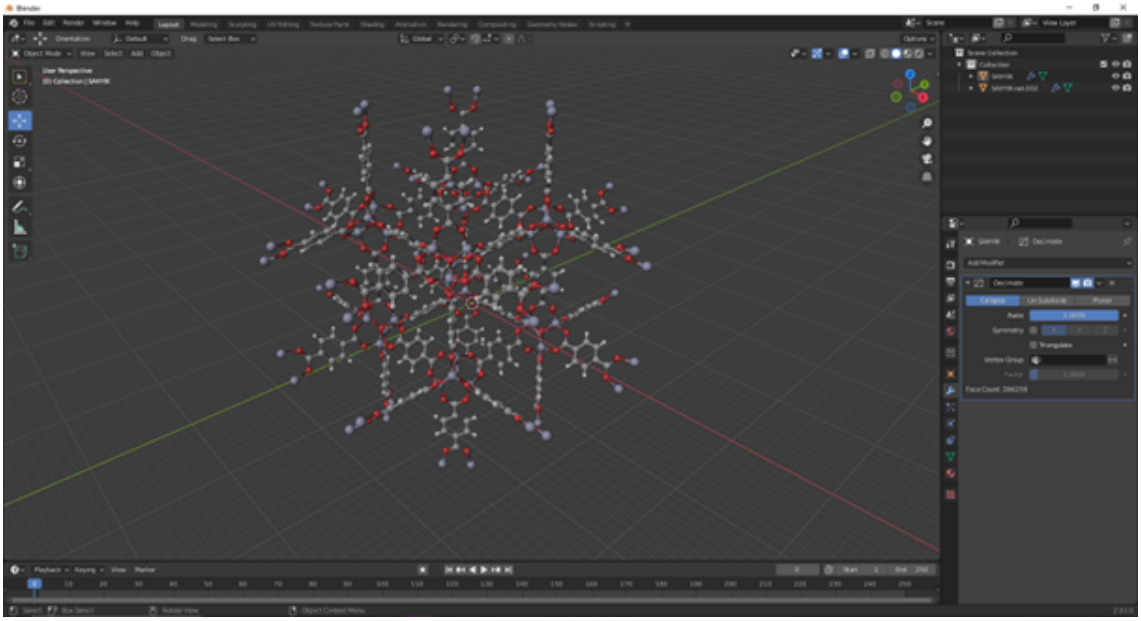


4. Open your newly created MOL2 file in Jmol and follow the steps in Part B using the instructions for the console in Jmol to obtain an OBJ file from the MOL2 input.

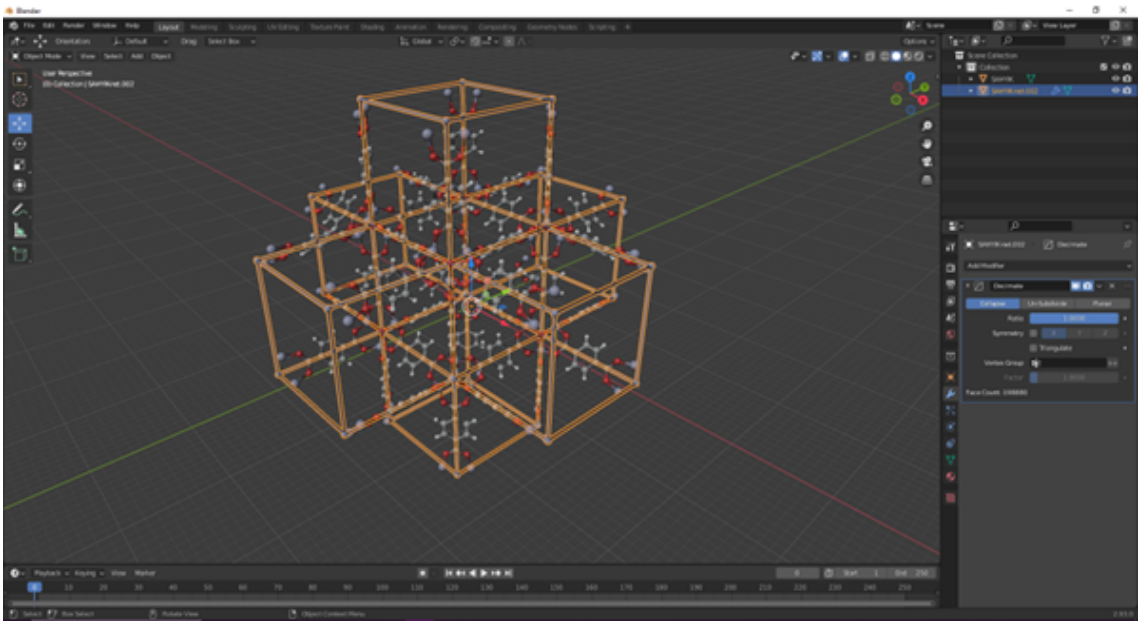
B.4.2 Part II: Combining OBJ nets and OBJ crystals in Blender.

1. Once the OBJ file of the net has been created, it is necessary to create the OBJ of the crystal structure itself, although these stages can be completed in either order. For the AR representation of the original crystal, follow Parts A and B of this guide and return here to combine the two files.

2. Begin by importing the crystal OBJ file. Go to File -> Import -> Wavefront (.obj) and import your crystal object file from the output folder of Jmol. (You may need to wait for it to render, these files can be >250MB and may take some time to load.)

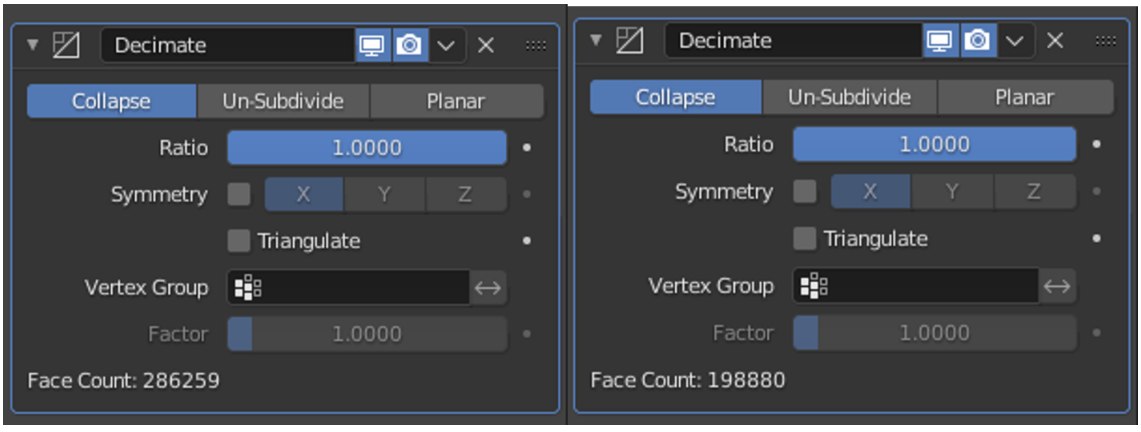


3. Now the OBJ files can be combined in Blender if the same original CIF was used to create both files. Some modification may be required in Blender to ensure the underlying net is configured in the correct position, although if the exact same CIFs are used this is unlikely. Import the nets OBJ file. Go to File -> Import -> Wavefront (.obj) and import your topology net object file.



Note: These stages draw many similarities with Part C, with a few additional steps. If the net requires some adjustment, it can easily be moved around and re-scaled in Blender.

4. Once the initial structures have been imported, to ensure that the combined files will render in Augmented Reality we need to add the total face counts for both OBJ files together and decimate them both so that the sum of faces does not exceed 750,000. Select the object using left click and find the Spanner (Wrench) icon in the right-hand side panel. Here it will ask you to "Add Modifier". Select Add Modifier -> Decimate (in the Generate column). Calculate the required ratio adjustment to bring it down to an acceptable level.



Note: We would recommend decimating the underlying net more than the crystal structure as the quality of the render for the net is less significant.

Ensure that the SUM of face counts does not exceed 750,000!

5. Once the face count is at an acceptable level, the combined structures can be ex-

ported as an FBX file. Go to File -> Export -> FBX (.fbx) and save the file as something appropriate, e.g refcode_topology_AR.fbx.

6. Close Blender and follow Part E to complete the upload to p3d for AR visualisation.