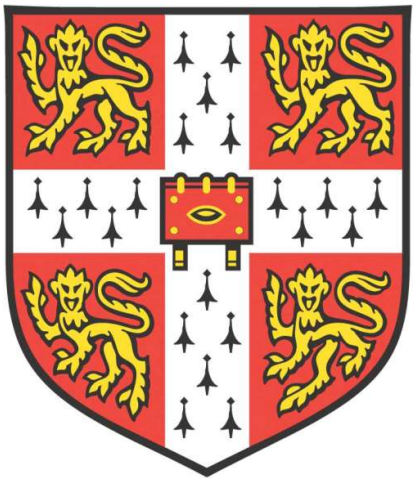
BUILDING AND EXPLORING DATABASES OF POROUS MATERIALS FOR ADSORPTION APPLICATIONS



Aurélia Li

Darwin College

Adsorption and Advanced Materials

Department of Chemical Engineering and Biotechnology

University of Cambridge

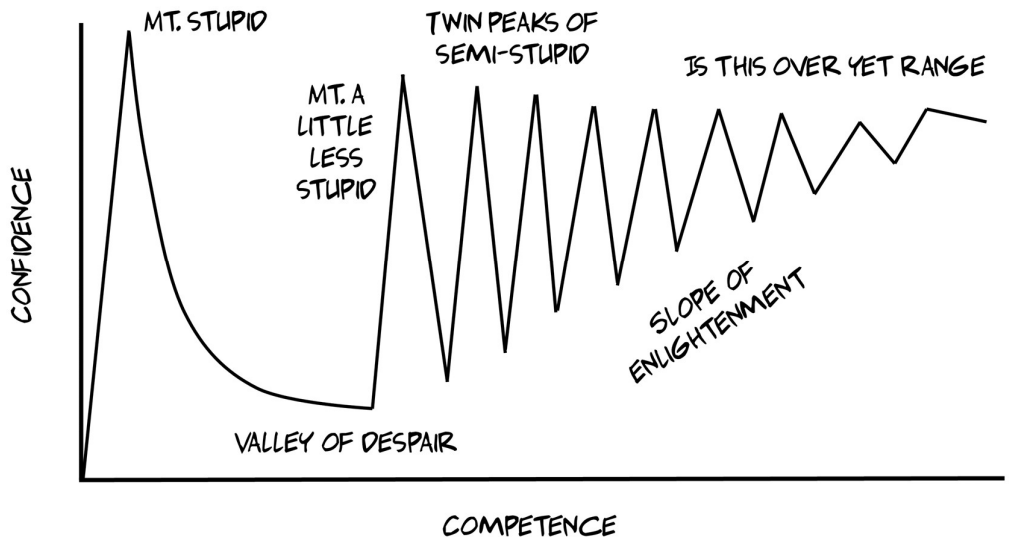
This dissertation is submitted for the degree of Doctor of Philosophy

April 2021

Building and Exploring Databases of Porous Materials for Adsorption Applications

Aurélia Li – April 2021

To the wolf pack



Modified model of the Dunning-Kruger knowledge graph.

Adapted from Joshua Render*

*Joshua Render, The Problem With Extremely Confident People,

agile-mercurial.com/2019/07/15/the-problem-with-extremely-confident-people/, 2019

DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

In accordance with the Department of Engineering guidelines, this thesis does not exceed 65,000 words, and it contains less than 150 figures.

Parts of this dissertation are presented in the following publications:

1. **Aurélia Li**, Rocio Bueno-Perez, Seth Wiggin, David Fairen-Jimenez, CrystEngComm, 2020, 22, 7152-7161.
2. Peyman Z. Moghadam*, **Aurélia Li***, Xiao-Wei Liu, Rocio Bueno-Perez, Shu-Dong Wang, Seth B. Wiggin, Peter A. Wood, David Fairen-Jimenez, Chem. Sci., 2020, 11, 8373-8387.
3. Pu Zhao, Hong Fang, Sanghamitra Mukhopadhyay, **Aurélia Li**, Svemir Rudic, Ian J. McPherson, Chiu C. Tang, David Fairen-Jimenez, S. C. Edman Tsang, Simon A. T. Redfern, Nat. Commun., 2019, 10, 999.
4. Claudia Orellana-Tavra, Milan Köppen, **Aurélia Li**, Norbert Stock, David Fairen-Jimenez, ACS Appl. Mater. Interfaces, 2020, 12, 5.

*Authors contributed equally to this work.

Signed: [Signature redacted]

Date: 30 April 2021

Aurélia Li

Cambridge

SUMMARY

In recent years, the field of porous materials has witnessed increasing interest towards customisable structures. Among them are metal-organic frameworks (MOFs), covalent organic frameworks (COFs), metal-organic cages (MOCs) and organic cages (OCs). MOFs are coordination networks assembled from organic ligands and metal clusters, while COFs are their fully organic equivalents. MOCs are crystalline structures obtained from the packing of discrete cage-like organometallic molecules; OCs being in turn their organic equivalents. While all these materials are attracting increasing attention, MOFs remain the star. The building block approach to their relatively straightforward synthesis, combined with an easy crystallographic characterisation, has enabled scientists to synthesise an increasing number of structures and deposit their data into the Cambridge Structural Database (CSD). It is from the same database that I previously derived, in 2017, the world's first automatically updated MOF subset (the so-called CSD MOF subset). Three years after its creation, the number of structures increased from 70,000 to almost 100,000.

Building on my previous work, I developed a set of tools for experimental and computational scientists alike to explore the CSD MOF subset. I devised a series of methods for the targeted classification of MOFs into different groups: MOFs from different chosen families, with specific surface functionalisation, chirality, as well as channel and framework dimensionalities. The obtained information, along with their geometric characterisation was made accessible via an online interactive data visualisation platform, thereby mapping out the properties landscape of the CSD MOF subset. I then carried out a high-throughput screening (HTS) of a selected number of MOFs for their hydrogen storage performance at 298 K and 200, 500 and 900 bar. In addition to confirming one of the top-performers for this task, I uncovered interesting structure-property relationships by matching the adsorption data to the structural information obtained from this classification. I found that the best performing structures tend to be part of one of these families: CPO-27, Cu-Cu paddlewheel, IRMOFs or zirconium-oxide MOFs. Structures with three-dimensional porous channels and/or with halogen surface functionalisation also seemed to have greater performance.

I extended the developed methods to the identification of COFs in the CSD. However, the experimental difficulty in obtaining their crystallographic data means only a few structures were deposited in the CSD, therefore only a small portion of COFs reported in the literature were found.

Due to their inherent structural difference, MOCs and OCs could not be identified using the methods used for MOFs and COFs. I, therefore, designed a separate approach for their identification, using a combination of topological data analysis, supervised and unsupervised classification. After successfully obtaining two datasets – the largest OC dataset to the best of my knowledge and the only existing MOC dataset, I carried out a HTS on them for their separation performance on a 20/80 mixture of xenon/krypton at 298 K and 10 bar. I identified the top MOC and OC for this application, and confirmed the high performance of the CC3 family. The mapping of the classification of these two datasets to the adsorption data unveiled interesting insights into the range of performance from these CC3-type structures.

In summary, in addition to the previously built MOF subset, I successfully developed subsets of COFs, MOCs and OCs. I also built programmable and customisable tools for the exploration and visualisation of these subsets. The obtained structural landscapes proved useful in uncovering insightful structure-property relationships when mapped to adsorption data.

ACKNOWLEDGEMENTS

There are many people I would like to thank for making my Cambridge experience one of the best times of my life.

First of all, I would like to thank Prof. David Fairen-Jimenez for giving me the opportunity to carry out this PhD under his supervision. You have been an inspiration not only as a scientist and a researcher, but also as a group manager overall. I hope I can be as kind, understanding, open-minded and available as you have been these last few years. Thank you for also financially supporting me with the ERC grant, and for allowing me to participate in so many conferences and workshops.

Many thanks to Dr Peyman Z. Moghadam for showing me the way to the fascinating and aesthetically pleasing field of MOFs, a field that completely and unexpectedly changed the course of my career choices – for the better.

Thank you to Dr Sarah Rough for accepting me on the MPhil in Advanced Chemical Engineering program in the first place, giving me the chance to discover Cambridge and fall in love with its international atmosphere. I am sorry I did not follow your advice on this PhD; whatever the future holds, I know at least that I spent the most self-fulfilling years here.

Thank you to my CCDC supervisor Dr Seth Wiggin for your considerable patience, availability and support. Thank you to the CCDC for funding my PhD. Thanks to Dr Peter Wood, Suzanna Wood and Dr Andrew Maloney for making this collaboration between the CCDC and our research group possible.

Thank you to all my colleagues from the Adsorption and Advanced Materials group. In particular, thank you to Dr Rocio Bueno-Perez for your guidance, support and availability. I am incredibly grateful that you always found the time to help me despite your humongous workload. I still don't understand how you manage it. Infinite thanks to Marta Aragones Anglada, my colleague, deskmate and friend for these few years. Thank you for putting up with me. I am grateful for all the ups and downs we have gone through, and I am looking forward to more adventures with you. I love you ❤. Thank you to Selina Zhuang for being such a generous and supportive friend. Thank you to Elise Siouve for your positive and inexhaustible battery of energy, much like Charles actually! Thank you to Ceren Çamur for all the chocolate cake and the Turkish soup

when I needed them the most. Thank you to Francesca Melle for all the lockdown outings and tastings. Thank you to Elena Avila for being such an amazing #analogicdinosaur. Thank you to Amélie Albon for all the laughs and the knitting. Thank you to Raymond Chung for being the unique character that you are. Thank you to Johannes Osterrieth for tolerating me in Stockholm.

Thank you to all my collaborators for giving me the opportunity to work with you and publish such interesting works: Dr Sven Rogge, Dr Jelle Wieme, Dr Diego A. Gómez-Gualdrón, Prof. Veronique Van Speybroeck, Dr Ismael Matito-Martos, Dr Valentina Colombo, Prof. Jorge AR Navarro, Prof. Sofia Calero, Dr Pu Zhao, Prof. Simon AT Redfern, Prof. SC Edman Tsang, Dr Claudia Orellana-Tavra, Dr Milan Köppen, Prof. Norbert Stock, Dr Panagiota Markopoulou, Dr Nikolaos Panagiotou, Dr David Madden, Dr Sarah Buchanan, Prof. Paul G. Shiels and Prof. Ross Forgan.

Thank you to all my friends at CEB: Cloé, Jana, Qi, Dabwan, Ru, Harry A., Harry Z., Ray, (An)Drew, Apoorv, Dushanth, Joseph, Anthie, Maria, Walter, Eduardo, Sam N., Sam H., Leander.

Thank you to Darwin college for providing me with such a unique environment to study in. Thank you to my college tutor Dr Julia Davies for being so supportive. Thank you to my fellow Darwinians: Lotti, Melanie, Laura, James L., Alex, Jyothi, Mimy, Carolina, Jeff, Dan, Michael. Thank you to Chloé and Alice for pushing me the extra mile by joining the DCSA.

Thank you to all the friends I made over the years across societies and colleges. Special thanks to the members of the Cambridge University Gospel Choir, singing with you was an important ritual for me. Thanks to Carolina, Peerapat, Maria Luisa, Pauline for your eternal friendships.

Thank you to Kayla Friedman and Malcolm Morgan of the Centre for Sustainable Development, University of Cambridge, UK for producing the Microsoft Word thesis template used to produce this document.

Thank you to my partner Steven for your support during the last sprint of this marathon.

Last but not least, thank you to my parents for everything. I am the luckiest and I am eternally grateful for being your daughter.

CONTENTS

1 INTRODUCTION 1

1.1 POROUS MATERIALS FOR GAS ADSORPTION APPLICATIONS 2

1.2 METAL-ORGANIC FRAMEWORKS 4

1.2.1 What is a MOF? 4

1.2.2 A haystack of MOFs 4

1.3 COMPUTER-AIDED STUDY AND DISCOVERY OF NEW MATERIALS 8

1.3.1 A toolbox of molecular simulation techniques 8

1.3.2 Computational high-throughput screening 14

1.4 TOO MUCH DATA? 20

1.4.1 Too many databases? 20

1.4.2 Towards easier data exploration 22

1.5 AIMS AND GOALS OF THIS WORK 23

1.5.1 Exploring the CSD MOF subset 23

1.5.2 Building other porous structures databases 24

1.6 THESIS OUTLINE 24

2 METHODS FOR EXPLORING POROUS MATERIALS IN THE CAMBRIDGE STRUCTURAL DATABASE 25

2.1 DIGGING INTO THE CSD FOR MOFS WITH CONQUEST 26

2.2 USING CONQUEST TO ACCESS THE MOF SUBSET 27

2.3 NOTES ON CONQUEST QUERIES 29

2.4 REMOVING SOLVENTS WITH THE CSD PYTHON API 30

2.5 ADDING MISSING HYDROGENS 31

2.6 CASE STUDY: APPLICATION TO COFS 32

2.6.1 Introduction 32

2.6.2 Methods 33

2.6.3 Results 35

2.7 CONCLUSION 36

3 METHODS FOR MOLECULAR SIMULATIONS 37

3.1 INTRODUCTION 38

3.2 MONTE CARLO SIMULATIONS 38

3.2.1 Force fields 40

3.2.2 Modelling the framework and periodic boundary conditions 43

3.2.3 Modelling the adsorbates *44*
3.2.4 RASPA *45*
3.3 GEOMETRIC CHARACTERISATION 46
3.3.1 LCD, PLD and percolation *46*
3.3.2 Surface area *48*
3.3.3 Pore volume *48*
3.3.4 Pore size distribution *48*
3.4 CASE STUDIES 49
3.4.1 Case study 1: using GCMC to understand the CO_2 migration-induced flexibility in ZIF-7 ..... *49*
3.4.2 Case study 2: assessing the uptakes of α-CHC and DCA in CAU-7 ..... *56*
3.4.3 Case study 3: a computational study of the chiral selectivity of CMOM-1S, -2S and -3S on a series of molecules *60*
3.5 CONCLUSION 71
4 TARGETED CLASSIFICATION OF METAL-ORGANIC FRAMEWORKS IN THE CSD 73
4.1 CONTRIBUTIONS 74
4.2 INTRODUCTION 74
4.3 A CSD-INTEGRATED TOOLBOX FOR THE EXPLORATION OF THE CSD MOF SUBSET 75
4.4 TEXTURAL PROPERTIES OF MOFS AND THEIR EVOLUTION 76
4.4.1 Identification of missing hydrogens and occupancy errors *78*
4.4.2 Geometrical characterisation of the obtained subset *78*
4.5 IDENTIFICATION OF TARGET MOF FAMILIES 79
4.6 IDENTIFICATION OF SURFACE FUNCTIONALITIES IN MOFS 85
4.7 IDENTIFICATION OF CHIRAL MOFS 87
4.8 POROUS NETWORK CONNECTIVITY AND FRAMEWORK DIMENSIONALITY 88
4.9 AN INSIGHT INTO THE MOFS' CRYSTAL DATA QUALITY 90
4.10 MOF EXPLORER FOR 5D EXPLORATION OF STRUCTURAL PROPERTIES 93
4.11 CONCLUSION 94
5 HIGH-THROUGHPUT SCREENING OF THE CLASSIFIED CSD MOF SUBSET FOR HYDROGEN STORAGE 95
5.1 CONTRIBUTIONS 96
5.2 INTRODUCTION 96

5.3 STRUCTURES PREPARATION FOR HIGH-THROUGHPUT HYDROGEN UPTAKE SIMULATIONS 97
5.4 GRAND CANONICAL MONTE CARLO SIMULATIONS 97
5.5 RESULTS 98
5.6 CONCLUSION 105
6 IDENTIFICATION OF METAL-ORGANIC CAGES AND ORGANIC CAGES IN THE CSD USING TOPOLOGICAL DATA ANALYSIS 107
6.1 CONTRIBUTIONS 108
6.2 INTRODUCTION 108
6.3 PERSISTENT HOMOLOGY 111
6.4 CAGE IDENTIFICATION 116
6.4.1 Metal-organic cages 118
6.4.2 Organic cages 126
6.5 MAPPING THE CAGES' SHAPES TO THEIR XENON/KRYPTON SEPARATION PERFORMANCE 128
6.5.1 Methods 129
6.5.2 Results 130
6.6 CONCLUSION 137
7 CONCLUSION AND FUTURE WORK 139
8 REFERENCES 149
9 APPENDICES 163

LIST OF TABLES

TABLE 1 COMPARISON OF ZEOLITES, MOFS, COFS AND POROUS MOLECULAR SOLIDS. 3

TABLE 2 EXAMPLES OF MOFS IDENTIFIED VIA HTS AND VALIDATED EXPERIMENTALLY. 18

TABLE 3 FORCE FIELD PARAMETERS FOR THE UFF+ 51

TABLE 4 PORE VOLUME AND SURFACE AREA OF ZIF-7 54

TABLE A1 FORCE FIELD PARAMETERS USED IN THE GCMC SIMULATIONS. 166

TABLE A2 BOND, BEND AND TORSION DEFINITIONS OF DCA. 168

TABLE A3 BOND, BEND AND TORSION DEFINITIONS OF A-CHC. 170

TABLE B1 FORCE FIELD PARAMETERS FOR THE CMOMS AND THE ANIONS 175

TABLE B2 FORCE FIELD PARAMETERS FOR 1P1P. 176

TABLE B3 BOND, BEND AND TORSION DEFINITIONS FOR 1P1P 177

TABLE B4 FORCE FIELD PARAMETERS FOR 1P2P. 182

TABLE B5 BOND, BEND AND TORSION DEFINITIONS FOR 1P2P 183

TABLE B6 FORCE FIELD PARAMETERS FOR 2P1P. 188

TABLE B7 BOND, BEND AND TORSION DEFINITIONS FOR 2P1P 189

TABLE B8 FORCE FIELD PARAMETERS FOR 1P1PEN. 194

TABLE B9 BOND, BEND AND TORSION DEFINITIONS FOR 1P1PEN 196

TABLE B10 FORCE FIELD PARAMETERS FOR BM. 202

TABLE B11 BOND, BEND AND TORSION DEFINITIONS FOR BM. 203

TABLE B12 FORCE FIELD PARAMETERS FOR EBM. 208

TABLE B13 BOND, BEND AND TORSION DEFINITIONS FOR EBM. 209

TABLE B14 FORCE FIELD PARAMETERS FOR PE, 3PE AND 4PE. 213

TABLE B15 BOND, BEND AND TORSION DEFINITIONS FOR PE, 3PE AND 4PE. 214

TABLE B16 FORCE FIELD PARAMETERS FOR CPBA. 218

TABLE B17 BOND, BEND AND TORSION DEFINITIONS FOR CPBA. 219

TABLE B18 FORCE FIELD PARAMETERS FOR VBA. 223

TABLE B19 BOND, BEND AND TORSION DEFINITIONS FOR VBA. 224

TABLE B20 FORCE FIELD PARAMETERS FOR BUTANOL, PENTANOL AND HEXANOL. 228

TABLE B21 BOND, BEND AND TORSION DEFINITIONS FOR 2-BUTANOL. 229

TABLE B22 BOND, BEND AND TORSION DEFINITIONS FOR 2-PENTANOL. 234

TABLE B23 BOND, BEND AND TORSION DEFINITIONS FOR 2-HEXANOL. 240

TABLE D1 NUMBER OF STRUCTURES OBTAINED FOR EACH HALOGEN GROUP. 256

TABLE D2 NUMBER OF HITS OBTAINED FOR EACH OF THE POLAR GROUPS. 259

TABLE D3 NUMBER OF HITS FOR EACH ALKOXY GROUP. 260

TABLE D4 NUMBER OF HITS FOR EACH ALKYL GROUP. 262

TABLE H1 FORCE FIELD PARAMETERS USED IN THE GCMC SIMULATIONS. 275

TABLE J1 ADDITIONAL QUERIES FOR ORGANIC CAGES. 294

LIST OF FIGURES

FIGURE 1 MOFS ARE SELF-ASSEMBLED FROM METAL CLUSTERS AND ORGANIC LINKERS. 4

FIGURE 2 EVOLUTION OF THE NUMBER OF CSD ENTRIES AND ESTIMATED NUMBER OF DEPOSITED MOF STRUCTURES UP TO 2016. 7

FIGURE 3 CRITERIA USED TO EXTRACT MOF AND MOF-LIKE STRUCTURES FROM THE CSD. 7

FIGURE 4 SCHEMATIC 2D ILLUSTRATION OF **A.** LARGEST CAVITY DIAMETER (LCD) AND PORE LIMITING DIAMETER (PLD), **B.** GEOMETRIC PORE VOLUME, **C.** ACCESSIBLE SURFACE AREA AND **D.** PERCOLATION. 10

FIGURE 5 NITROGEN ISOTHERM ON ZIF-8 AT 77 K. 12

FIGURE 6 GENERAL WORKFLOW FOR THE COMPUTATIONAL HTS OF EXPERIMENTAL MOFS DERIVED FROM THE CSD. 15

FIGURE 7 A. ORGANISATION OF THE MOF DATA IN THE CSD. **B.** DIFFERENCES AND OVERLAPS BETWEEN THE *NON-DISORDERED MOF SUBSET* AND A "NO-DISORDER" FILTERED SEARCH IN THE CSD MOF SUBSET. 28

FIGURE 8 EXAMPLE OF TWO DIFFERENT WAYS OF REPRESENTING CARBOXYLATE LIGANDS 30

FIGURE 9 EXAMPLE OF A STRUCTURE FROM THE *NON-DISORDERED MOF* SUBSET WITH SITELESS HYDROGEN ATOMS 32

FIGURE 10 CONQUEST QUERIES USED TO LOOK FOR SPECIFIC COF LINKAGES. 34

FIGURE 11 COFS DATASET OBTAINED FROM THE CSD. 36

FIGURE 12 ILLUSTRATION OF THE BOND STRETCHING, ANGLE BENDING AND TORSIONS OF BONDED ATOMS. 41

FIGURE 13 SHAPE OF THE LENNARD-JONES POTENTIAL ENERGY BETWEEN TWO ATOMS. 42

FIGURE 14 SCHEMATIC REPRESENTATION OF THE PERIODIC BOUNDARY CONDITIONS. 44

FIGURE 15 CALCULATION OF THE LARGEST CAVITY DIAMETER (LCD) AND PORE-LIMITING DIAMETER (PLD) IN POREBLAZER AND IN ZEO++ 47

FIGURE 16 ZIF-7 STRUCTURE AND ITS CO_2 ADSORPTION BEHAVIOUR. 50

FIGURE 17 CO_2 ADSORPTION ISOTHERMS FOR ZIF-7 52

FIGURE 18 SNAPSHOTS AND DENSITY DISTRIBUTIONS OF CO_2 IN ZIF-7. 53

FIGURE 19 A. SIMULATED ADSORPTION ISOTHERMS FOR N_2 AT 77 K AND **B.** PSD OF ZIF-7. 54

FIGURE 20 HEAT OF ADSORPTION OF CO_2 IN ZIF-7. 55

FIGURE 21 INDEXING OF THE ATOMS IN **A.** DCA AND **B.** A-CHC. 57

FIGURE 22 A. PSD OF CAU-7, **B.** SNAPSHOT OF A SUPERCELL OF EMPTY CAU-7. 58

FIGURE 23 SNAPSHOTS OF DCA (**A.**–**C.**) AND A-CHC (**D.**–**F.**) IN CAU-7 AT DIFFERENT LOADINGS 58

FIGURE 24 DENSITIES OF DCA (**A.**–**C.**) AND A-CHC (**D.**–**F.**) AT DIFFERENT LOADINGS IN CAU-7 59

FIGURE 25 A. A 3X1X3 CELL OF CMOM-3S FRAMEWORK, WITH THE RIGID CMOM FRAMEWORK, THE MOVING TRIFLATES AND 1P1P MOLECULES. **B.** ENANTIOSELECTIVITY OBTAINED BY ZHANG ET AL. IN VARIOUS CMOMS. 61

FIGURE 26 GRAVIMETRIC LOADING OF (*S*)-1P1P (ORANGE) AND (*R*)-1P1P (PURPLE) IN CMOM-3S 64

FIGURE 27 GRAVIMETRIC LOADING OF (*S*)-1P2P (ORANGE) AND (*R*)-1P2P (PURPLE) IN CMOM-3S 66

FIGURE 28 GRAVIMETRIC LOADING OF (*S*)-2P1P (ORANGE) AND (*R*)-2P1P (PURPLE) IN CMOM-3S 67

FIGURE 29 GRAVIMETRIC LOADING OF **A.** 1P1P, **B.** 1P2P AND **C.** 2P1P IN CMOM-1S, AND **D.** 1P1P, **E.** 1P2P AND **F.** 2P1P IN CMOM-2S IN SYSTEM II). 68

FIGURE 30 LIST OF MODELLED CHIRAL MOLECULES. 69

Figure 31 Gravimetric loading in CMOM-3S of **a.** 2-butanol, **b.** 2-pentanol, **c.** 2-hexanol, **d.** 1P1Pen, **e.** BM, **f.** EBM, **g.** PE, **h.** 3PE, **i.** 4PE, **j.** CPBA and **k.** VB. 70

Figure 32 Flowchart outlining the CSD MOF subsets structures preparation prior to geometrical and structural calculations. 77

Figure 33 a. An example structure in the CSD MOF subset with occupancy issues (refcode: CIYER), **b.** atoms with occupancy issues in the CIYER CIF are highlighted. 78

Figure 34 Histograms comparing geometric properties for all the porous MOFs in the CSD MOF subset from 1995 to 2015. 79

Figure 35 Criteria developed for the identification of MOF families in the CSD MOF subset 81

Figure 36 a. to **d.** Criteria developed to look for structures containing Cu-Cu paddlewheels. **e.** to **h.** Example structures found using the criterion on the left 83

Figure 37 a. to **d.** Criteria used to eliminate undesired structures and the number of structures eliminated at each step. **e.** to **h.** Examples of eliminated structures corresponding to the criteria on the left. 84

Figure 38 Histograms showing the geometric properties for each MOF family identified in the CSD MOF subset. 85

Figure 39 Criteria developed to identify MOFs with common functionalities in the CSD MOF subset. 86

Figure 40 Histograms of the geometric properties of 1,911 chiral structures with non-zero gravimetric surface area in the CSD MOF subset. 88

Figure 41 Histograms of framework and channel/pore dimensionalities characterised for the 52,787 structures. 89

Figure 42 Non-cumulative evolution of R-factors of the MOF subset from 1960 to 2015 91

FIGURE 43 DISTRIBUTION OF R-FACTORS AND DENSITY ACROSS DIFFERENT MOF FAMILIES AND CRYSTAL SYSTEMS OF LOW OR HIGH SYMMETRY. 92

FIGURE 44 CHARACTERISATION OF THE 3D MOFS SCREENED FOR HYDROGEN STORAGE. ABSOLUTE VOLUMETRIC UPTAKE VS. ABSOLUTE GRAVIMETRIC UPTAKE WT.% H_2 AT ROOM TEMPERATURE AT 200, 500 AND 900 BAR. 100

FIGURE 45 VOLUMETRIC UPTAKE VERSUS GRAVIMETRIC UPTAKE IN WT.% H_2 FOR THE SCREENED STRUCTURES FOR HYDROGEN STORAGE AT **A.** 200 BAR, **B.** 500 BAR AND **C.** 900 BAR. 101

FIGURE 46 QUANTITATIVE CHARACTERISATION OF THE 3D MOFS SCREENED FOR HYDROGEN STORAGE. 103

FIGURE 47 **A.** EVOLUTION OF THE STRUCTURE WITH THE HIGHEST HYDROGEN VOLUMETRIC UPTAKE AT ROOM TEMPERATURE AND AT 500 BAR IN THE CSD OVER THE YEARS. **B.** SNAPSHOT OF A SUPERCELL OF BOMCUB. **C.** SNAPSHOT OF A SUPERCELL OF JARMEU. 104

FIGURE 48 EXAMPLES OF **A.** A CAGE, **B.** A RING-LIKE STRUCTURE, **C.** A BOWL-SHAPED STRUCTURE. 111

FIGURE 49 DETERMINING THE PERSISTENT HOMOLOGY OF A SET OF POINTS UNIFORMLY DISTRIBUTED AS A CIRCLE. **A.** GROWING BEADS ON THE POINT CLOUD, **B.** THE CORRESPONDING PERSISTENCE BARCODE, **C.** THE CORRESPONDING PERSISTENCE DIAGRAM. 113

FIGURE 50 **A.** A SIMPLICIAL COMPLEX BUILT ON A POINT CLOUD. **C.** EXAMPLE OF A CAGE, CSD REFCODE QUFYIB. PERSISTENCE **B.** BARCODE AND **D.** DIAGRAM OBTAINED FOR **C.** 114

FIGURE 51 FROM DATA POINTS TO PERSISTENCE DIAGRAM TO PERSISTENCE LANDSCAPE. 115

FIGURE 52 WORKFLOW FOR THE IDENTIFICATION AND CLASSIFICATION OF MOCS AND OCS IN THE CSD. 117

FIGURE 53 QUICK "MUST-HAVE" CRITERIA DRAWN IN CONQUEST FOR SOME COMMON CAGES. 118

FIGURE 54 EXAMPLES OF STRUCTURES OBTAINED WITH THE IMIDAZOLE-BASED QUERY. 119

FIGURE 55 EXAMPLES OF STRUCTURES TARGETED WITH THE PYRIDINE-BASED CAGES QUERY. 119

FIGURE 56 EXAMPLES OF STRUCTURES TARGETED BY THE BANANA-SHAPED QUERY. .. 120

FIGURE 57 EXAMPLES OF STRUCTURES OBTAINED WITH THE BIS(IMINO)PYRIDYL QUERY. 120

FIGURE 58 EXAMPLES OF STRUCTURES TARGETED BY THE DIOXOLANE/DIOXANE-BASED QUERY. 121

FIGURE 59 EXAMPLE OF CAGES OBTAINED WITH THE CYCLOTRIVERATRYLENE-DERIVED QUERY. 121

FIGURE 60 REMOVING NOISY STRUCTURES 123

FIGURE 61 TRUNCATED DENDROGRAMS OF THE HIERARCHICAL CLUSTERINGS USED TO IDENTIFY MOCS. 125

FIGURE 62 QUICK "MUST-HAVE" CRITERIA DRAWN IN CONQUEST FOR SOME COMMON 3D ORGANIC CAGES. 126

FIGURE 63 QUICK "MUST-HAVE" CRITERIA DRAWN IN CONQUEST FOR CUCURBITURILS, CYCLODEXTRINS AND CRYPTOPHANES. 127

FIGURE 64 EXAMPLES OF **A.** CYCLODEXTRINS, **B.** CUCURBITURILS AND **C.** CRYPTOPHANES. 127

FIGURE 65 XENON/KRYPTON SEPARATION PERFORMANCE OF METAL-ORGANIC CAGES AND ORGANIC CAGES..... 132

FIGURE 66 **A.** CRYSTAL SYSTEMS OF TETRAHEDRAL CAGES AND THEIR XE/KR SELECTIVITY. **B.** EXAMPLE OF ORGANIC TETRAHEDRAL CAGE. **C.** EXAMPLE OF METAL-ORGANIC TETRAHEDRAL CAGE. 134

FIGURE 67 COMPARISON OF TWO M_6L_4 STRUCTURES WITH WIDELY DIFFERENT XE/KR SELECTIVITY VALUES 136

FIGURE B1 INDEXING OF THE ATOMS IN 1P1P. 176

FIGURE B2 INDEXING OF THE ATOMS IN 1P2P. 182

FIGURE B3 INDEXING OF THE ATOMS IN 2P1P. 188

FIGURE B4 INDEXING OF THE ATOMS IN 1P1PEN. 194

FIGURE B5 INDEXING OF THE ATOMS IN BM. 202

FIGURE B6 INDEXING OF THE ATOMS IN EBM. 208

FIGURE B7 INDEXING OF THE ATOMS IN **A.** PE, **B.** 3PE AND **C.** 4PE. 213

FIGURE B8 INDEXING OF THE ATOMS IN CPBA. 218

FIGURE B9 INDEXING OF THE ATOMS IN VBA. 223

FIGURE B10 INDEXING OF THE ATOMS IN 2-BUTANOL. 229

FIGURE B11 INDEXING OF THE ATOMS IN 2-PENTANOL. 234

FIGURE B12 INDEXING OF THE ATOMS IN 2-HEXANOL. 240

FIGURE C1 AN EXAMPLE ZR-OXIDE BASED MOF 246

FIGURE C2 CRITERIA USED TO LOOK FOR ZR-OXIDE BASED MOFS. 247

FIGURE C3 AN EXAMPLE STRUCTURE ELIMINATED BY THE USE OF "MUST NOT HAVE" CRITERION SHOWN IN **FIGURE C2**. 247

FIGURE C4 **A.** THE STRUCTURE OF IRMOF-1 (MOF-5). **B.** THE STRUCTURE OF AN EXAMPLE MOF WITH ZN-OXIDE SBUS. 248

FIGURE C5 CRITERIA USED TO LOOK FOR ZN-OXIDE BASED MOFS. 248

FIGURE C6 DERIVATION OF CRITERIA FOR IRMOF-LIKE STRUCTURES FROM THE PREVIOUS ZN-OXIDE-BASED STRUCTURES CRITERIA. 249

FIGURE C7 AN EXAMPLE MOF-74/CPO-27 STRUCTURE. 250

FIGURE C8 **A.** CRITERION DEVELOPED TO LOOK FOR MOF-74/CPO-27-TYPE MOFS. **B.** EXAMPLE OF AN UNDESIRED STRUCTURE FOUND IF THE CRITERION IS NOT

RESTRICTIVE ENOUGH, I.E. IF ONLY THE LEFT PART OF THE CRITERION IS REPRESENTED. 251

FIGURE C9 A. TO **C.** CRITERIA DEVELOPED TO ELIMINATE UNDESIRED STRUCTURES. **D.** TO **F.** EXAMPLES OF STRUCTURES ELIMINATED WITH THE CORRESPONDING CRITERION ON THE LEFT..... 251

FIGURE C10 "MUST HAVE" CRITERION USED TO LOOK FOR ZIF-TYPE STRUCTURES. 252

FIGURE C11 A. TO **F.** "MUST NOT HAVE" CRITERIA USED TO ELIMINATE UNDESIRED STRUCTURES. **G.** TO **L.** EXAMPLE STRUCTURES CORRESPONDING TO THE CRITERION DESCRIBED IN THE LEFT. 253

FIGURE C12 A. CRITERION USED TO LOOK FOR ZIF STRUCTURES WITH METAL COORDINATION OF 6 OR 8. **B.** AN EXAMPLE CORRESPONDING STRUCTURE. 254

FIGURE D1 CRITERION USED TO LOOK FOR HALOGEN GROUPS. 255

FIGURE D2 EXAMPLE STRUCTURES OBTAINED FOR THE -F GROUP IF THE F ATOM IS LINKED TO A P ATOM. 255

FIGURE D3 EXAMPLE STRUCTURE OBTAINED FOR THE -F GROUPS IF THE SECOND NEIGHBORS OF THE F ATOM ARE ALSO F ATOMS. 256

FIGURE D4 CRITERION USED TO LOOK FOR FMOFS. 256

FIGURE D5 CRITERION USED TO LOOK FOR POLAR FUNCTIONAL GROUPS. 257

FIGURE D6 CRITERIA USED TO LOOK FOR UNDESIRED DICYANIDE GROUPS AND CORRESPONDING EXAMPLES FOR EACH CRITERION. 258

FIGURE D7 CRITERION USED TO LOOK FOR ALKOXY GROUPS. 260

FIGURE D8 A. TO **C.** CRITERIA USED TO LOOK FOR STRUCTURES WITH ALKYL FUNCTIONAL GROUPS. **D.** TO **F.** EXAMPLE STRUCTURES TARGETED BY EACH CRITERION. 261

FIGURE D9 CRITERIA USED TO LOOK FOR ALKYL GROUPS OF MORE THAN 4 CARBON ATOMS. 262

FIGURE D10 CRITERIA USED TO LOOK FOR STRUCTURES WITH PERFLUOROALKANE CHAINS. 263

FIGURE E1 HISTOGRAMS SHOWING THE NUMBER OF HITS FOR MOFS WITH DIFFERENT ALKYL GROUPS..... 265

FIGURE E2 HISTOGRAMS SHOWING THE NUMBER OF HITS IN MOFS WITH DIFFERENT ALKOXY GROUPS. 266

FIGURE E3 HISTOGRAMS SHOWING THE NUMBER OF HITS IN MOFS WITH DIFFERENT POLAR GROUPS. 266

FIGURE E4 HISTOGRAMS SHOWING THE NUMBER OF HITS IN MOFS WITH DIFFERENT HALOGEN GROUPS. 267

FIGURE F1 HISTOGRAMS OF FRAMEWORK AND CHANNEL/PORE DIMENSIONALITIES FOR 8,253 POROUS MOFS. 271

FIGURE G1 HISTOGRAMS OF **A.** DENSITY (G CM^{-3}), **B.** LARGEST CAVITY DIAMETER (LCD) AND **C.** VOID FRACTION AGAINST R-FACTORS FOR STRUCTURES WITH NON-ZERO GRAVIMETRIC SURFACE AREA VALUES IN THE CSD MOF SUBSET. 272

FIGURE G2 **A.** HISTOGRAMS OF R-FACTORS FOR THE DIFFERENT MOF FAMILIES. **B.** HISTOGRAMS OF R-FACTORS FOR THE DIFFERENT MOF FAMILIES WITH NON-ZERO SURFACE AREA. 272

FIGURE G3 BOXPLOTS OF R-FACTORS VS. CRYSTAL SYSTEMS FOR EACH MOF FAMILY. 273

FIGURE G4 BOXPLOTS OF R-FACTORS VS. DEGREE OF SYMMETRY FOR OF THEIR CRYSTAL SYSTEMS FOR EACH MOF FAMILY..... 273

FIGURE G5 BOXPLOTS OF R-FACTORS VS. **A.** CRYSTAL SYSTEMS AND **B.** DEGREE OF SYMMETRY FOR ALL STRUCTURES IN THE CSD MOF SUBSET..... 274

FIGURE G6 BOXPLOTS OF R-FACTORS FOR **A.** DIFFERENT MOF FAMILIES AND **B.** ALL STRUCTURES IN THE CSD MOF SUBSET. 274

FIGURE H1 VOLUMETRIC UPTAKE VERSUS GRAVIMETRIC UPTAKE IN WT.% H_2 FOR THE SCREENED STRUCTURES FOR HYDROGEN STORAGE AT **A.** 200 BAR, **B.** 500 BAR AND **C.** 900 BAR. 280

FIGURE H2 VOLUMETRIC UPTAKE VERSUS GRAVIMETRIC UPTAKE IN WT.% H_2 FOR THE SCREENED STRUCTURES FOR HYDROGEN STORAGE AT **A.** 200 BAR, **B.** 500 BAR AND **C.** 900 BAR. 281

FIGURE H3 QUANTITATIVE CHARACTERISATION OF THE 3D MOFS SCREENED FOR HYDROGEN STORAGE. 282

FIGURE I1 VIETORIS-RIPS FILTRATION: SET OF POINTS WHERE THE DISTANCE BETWEEN TWO POINTS IS LESS OR EQUAL THAN ALPHA. 283

FIGURE I2 EXAMPLES OF TARGETED CARBON-BASED CAGES. 283

FIGURE I3 EXAMPLES OF TARGETED IMINE-BASED CAGES. 284

FIGURE I4 EXAMPLES OF TARGETED BORONATE-BASED CAGES 284

FIGURE I5 EXAMPLES OF TARGETED OXYGEN-BASED CAGES. 284

FIGURE I6 DIFFERENCES IN PACKING BETWEEN CC3-TYPE STRUCTURES AND M_6L_4-TYPE STRUCTURES. 293

List of Abbreviations and Acronyms

1P1P – 1-phenyl-1-propanol

1P1Pen – 1-phenyl-1-pentanol

1P2P – 1-phenyl-2-propanol

2P1P – 2-phenyl-1-propanol

3PE – 1-(3-chlorophenyl)ethanol

4PE – 1-(4-chlorophenyl)ethanol

α-CHC – α-cyano-4-hydroxycinnamic acid

AiiDA – Automated Interactive Infrastructure and Database for Computational Science

AMBER – Assisted Model Building with Energy Refinement

API – application programming interface

BET – Brunauer-Emmett-Teller (surface area)

BM – benzenemethanamine

CBMC – configurational bias Monte Carlo

CC3 – Covalent Cage 3

CCDC – Cambridge Crystallographic Data Centre

CDB – Cage Database

CIF – Crystallographic Information File

CMOM – chiral metal-organic material

COF – covalent-organic frameworks

CoRE MOF – Computation-ready, Experimental MOF (database). The corresponding MOF structures are referred to as CoRE MOFs.

CoRE COF – Computation-ready, Experimental COF (database). The corresponding COF structures are referred to as CoRE COFs.

CURATED COF – Clean, Uniform, and Refined with Automatic Tracking from Experimental Database of COF (database). The corresponding COF structures are referred to as CURATED COFs.

CPBA – α-cyclopropylbenzyl alcohol

CPO – coordination polymer of Oslo

CSD – Cambridge Structural Database

CSP – chiral stationary phase

DCA – sodium dichloroacetate

DDEC – density derived electrostatic and chemical (charges)

DDS – drug delivery system

DES – diethylsulfide

DFF – DREIDING force field

DFT – density functional theory

EBM – α-ethyl benzenemethanamine

EQeq – extended charge equilibration (method)

GA – genetic algorithm

GAFF – Generalized AMBER Force Field

GCMC – grand canonical Monte Carlo

HKUST – Hong Kong University of Science and Technology

hMOF – hypothetical MOF (database). The corresponding MOF structures are referred to as hMOFs.

HTS – high-throughput screening

IUPAC – International Union of Pure and Applied Chemistry

IRMOF – isoreticular MOF

LCD – largest cavity diameter

MC – Monte Carlo

MCMC – Markov chain Monte Carlo

MD – molecular dynamics

MFM – Manchester Framework Material (formerly NOTT)

MFU – metal-organic framework Ulm University

MOC – metal-organic cage

MOF – metal-organic framework

MOM – metal-organic material

MS – microstates

NOTT – University of Nottingham (now MFM)

NU – Northwestern University

OC – organic cage

OPLS - Optimized Potentials for Liquid Simulations

oPMC – organic porous molecular crystals (database)

PCN – porous coordination network

PE – α-phenylethyl alcohol

PLD – pore limiting diameter

PSD – pore size distribution

RAC – rectified autocorrelation (function)

RCSR - Reticular Chemistry Structure Resource

RU – repeating unit

SNU – Seoul National University

STP – standard temperature and pressure

TDA – topological data analysis

ToBaCCo - Topology-Based Crystal Constructor

TPS – temperature and pressure swing

TraPPE – Transferable Potentials for Phase Equilibria

UFF – Universal Force Field

UiO – Universitetet i Oslo (University of Oslo)

UMCM – University of Michigan Crystalline Material

VBA – α-vinylbenzyl alcohol

ZIF – zeolitic imidazolate framework

List of Appendices

Appendix A DCA and alpha-CHC molecule definition parameters and GCMC force field parameters 165

Appendix B Molecule parameters for the study of CMOMs for chiral separations 174

Appendix C MOF families classification: description of the criteria developed 246

Appendix D MOFs' functional groups: description of the criteria developed 255

Appendix E Calculations of the CSD MOFs' physical and geometrical properties 265

Appendix F Calculation of framework dimensionalities 268

Appendix G Quality assessment of the data in the CSD MOF subset using R factors 272

Appendix H HTS for H2 GCMC simulations parameters and additional results 275

Appendix I Identification of metal-organic cages and organic cages with topological data analysis – further details 283

Appendix J Additional ConQuest queries used for reducing the search space of organic cages in the CSD 294

1 INTRODUCTION

Parts of the following content are published in:

- *Enabling efficient exploration of metal-organic frameworks in the Cambridge Structural Database*, **Aurélia Li,** Rocio Bueno-Perez, Seth Wiggin, David Fairen-Jimenez, CrystEngComm, 2020, 22, 7152-7161.
- *Targeted classification of metal–organic frameworks in the Cambridge structural database (CSD)*, Peyman Z. Moghadam*, **Aurélia Li***, Xiao-Wei Liu, Rocio Bueno-Perez, Shu-Dong Wang, Seth B. Wiggin, Peter A. Wood, David Fairen-Jimenez, Chem. Sci., 2020, 11, 8373-8387.

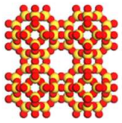
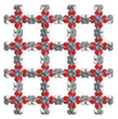
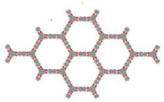
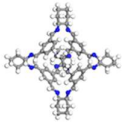
*Authors contributed equally to this work.

Abstract. Designable porous materials such as metal-organic frameworks (MOFs) have attracted much attention in the last two decades. The number of reported synthesised structures in the Cambridge Structural Database has reached almost 100,000. The increasing number of MOFs being synthesised, combined with the even larger number of imaginable hypothetical structures, paved the way to an entire computational research field based on high-throughput screenings (HTS) in order to i) find the best structure for a given application, ii) uncover interesting structure–property trends. This chapter introduces the various materials studied in this work and provides an overview of the role computational HTS have played so far in the rational design of MOFs for adsorption applications.

1.1 Porous materials for gas adsorption applications

Porous materials – such as the well-established zeolites and activated carbons – have been widely studied for their high surface areas (around 1,000 $m^2 g^{-1}$ and up to 3,000 $m^2 g^{-1}$ respectively) and are now commercially used in the fields of gas adsorption, separation and catalysis.[1] More recent porous materials, such as metal-organic frameworks (MOFs), covalent organic frameworks (COFs), metal-organic cages (MOCs) and organic cages (OCs) have been attracting growing interest, mostly due to their tunability and therefore the possibility for customisable pore shapes and sizes.[1-4] While MOFs and COFs are extended crystalline structures, MOCs and OCs are discrete molecules packed into crystalline or amorphous porous molecular solids.[5-9] Along with zeolites, MOFs, COFs, MOCs and OCs are all designable materials: the crystal engineering approaches developed for zeolites have paved the way to a paradigm of functional molecular design, where scientists can design and tune a structure for a target application. This was made possible by the advances in synthesis methods, computational simulations and physical measurements, all of which were essential to the better understanding of structure-property relationships.[1] However, conversely to zeolites, these four types of materials are at a lesser developed stage, and have not reached large-scale production yet. Although all offer a range of porosities, they differ in ease of modularity, variety range, stability and processing technologies. **Table 1** presents a comparison of these materials and zeolites and highlights the unique selling points of each**.** Zeolites remain the cheapest and most thermally stable designable porous materials and are processed into films, composites and pellets.[1] MOFs demonstrate the highest modularity of the showcased materials, albeit with varying stability, and growing research is directed towards processing MOFs into industry-friendly shapes (films, composites, pellets and most recently monoliths).[10-15] COFs distinguish themselves with their extended π-conjugated structures, both in-plane and in the stacking direction, thereby favouring electronic delocalisation and making them interesting candidates for photocatalysis[16, 17] and electronic applications.[18] Porous cages have the advantage of being soluble and increasing research is being carried out into shaping them into porous liquids.[19, 20] I will focus in this chapter on the case of MOFs; more details on COFs and cages will be given in the relevant chapters.

Table 1 Comparison of zeolites, MOFs, COFs and porous molecular solids. Largely adapted from Slater et al.[1] The International Union of Pure and Applied Chemistry (IUPAC) recommends the term "micropores" for pore sizes smaller than 2 nm and "mesopores" for pore sizes between 2 and 50 nm.[21] Modularity here refers to the degree to which the materials can be decomposed into building components. Designability refers to the availability of design frameworks to build these materials.

	 **Zeolites**	 **MOFs**	 **COFs**	 **Porous molecular solids (incl. MOCs and OCs)**
Porosity	Microporous to mesoporous	Microporous to mesoporous	Microporous to mesoporous	Can be mesoporous but rare so far
Modularity	High, less variety of building units but many possible topologies	Very high, a variety of building blocks	High, a (smaller) variety of building blocks	High, a variety of possible packings
Designability	High, but design templates sometimes challenging (e.g. organic templates)	Very high, well developed design strategies such as the isoreticular principle	High, but not as developed.	High, but not as developed
Stability	Good thermal stability	Low to high	Boronate-based: low; imine-based: high	Generally low
Processing	Insoluble. Processed in films, composites, pellets.	Insoluble. Processed in composites, films or as monoliths.	Insoluble	Soluble
Advantage	Stable, commercially in use, costs ranging from low to high	High modularity for a range of imaginable materials	Electronic properties	Solution processing, double porosity (intrinsic and extrinsic)
Development stage	Well-established, but still growing. Widely used commercially.	Established, but no large-scale applications yet.	Less developed, but promising	Less developed, but promising

1.2 Metal-organic frameworks

1.2.1 What is a MOF?

MOFs are defined by the International Union of Pure and Applied Chemistry (IUPAC) as "coordination network[s] with organic ligands containing potential voids".[22] This definition is purposely kept wide enough to account for the various ways different researchers across different disciplines conceptualise MOFs. Although MOFs are not required to be crystalline by this definition,[22] I will in this thesis describe MOFs as a class of crystalline materials assembled from metal atoms or clusters (secondary building units or SBUs) and organic ligands in a building block approach (see **Figure 1**).[23] The relatively straightforward synthesis of MOFs and the diverse possible combinations of SBUs and ligands have led to the design of ever more customised structures. Their pores span a range of sizes, geometries, internal surface areas (as high as 8,000 $m^2 g^{-1}$)[24] and pore volumes. These properties have encouraged researchers to consider MOFs for a wide variety of applications, ranging from gas storage,[25-29] gas separation,[3, 30-33] and catalysis[34-36] to drug delivery[37-41] and bio-imaging.[38, 39, 42]

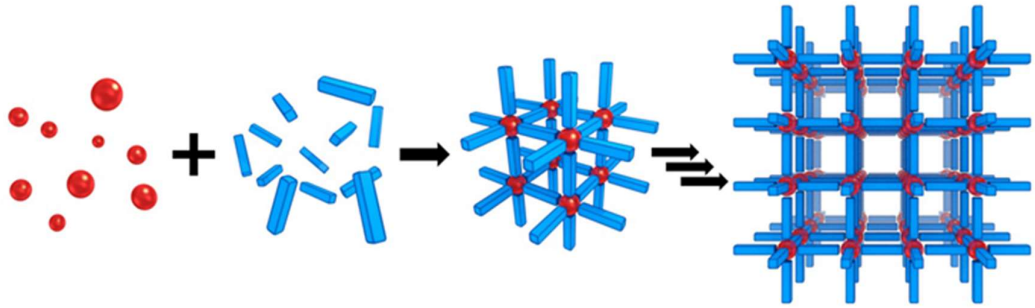


Figure 1 MOFs are self-assembled from metal clusters and organic linkers.[43]

1.2.2 A haystack of MOFs

Given the variety of metal clusters and organic ligands available or readily synthesisable, one can imagine the large space of possible combinations leading to MOFs. A few natural questions that arise from this are: what are possible MOFs? How many MOFs have been synthesised? What can we learn from all these data?

1.2.2.1 What are possible MOFs?

The building-block approach to the synthesis of MOFs has inspired researchers to computationally create hypothetical structures. Several groups have developed their own datasets, using different methods and different building blocks. For instance, Wilmer et al. built the hypothetical MOF (hMOF) database from a "bottom-up" – or "tinkertoy" – approach, where each structure is generated from the recombination of 102 SBUs and organic linkers from available crystallographic data of existing MOFs.[44] This method yielded 137,953 possible structures. However, this dataset was naturally biased towards six MOF topologies.[25] Topology here refers to the periodic nets in which metal clusters and organic ligands are connected as vertices and edges, respectively.[45] Therefore, topology is a useful tool for the classification of MOFs based on their connectivity. As a comparison, the Reticular Chemistry Structure Resource (RCSR), which is the topology system recommended by IUPAC for the case of MOFs, has identified 2918 unique 3-periodic, 200 2-periodic and 8 1-periodic nets.[46] To remedy the biased representation of MOFs in the hMOF database, Gómez-Gualdrón et al. used a "top-down" (or reverse topological) approach to focus on the diversity of possible MOF topologies.[25] The starting point was 41 predefined topologies, in which 78 building blocks were combined and added. This method led to the generation of 13,512 hypothetical MOFs gathered in the ToBaCCo (Topology-Based Crystal Constructor) database. A similar approach from Boyd et al. with 46 topologies led to ca. 300,000 structures.[47, 48] Although larger and more comprehensive, it is likely that the latter databases still do not cover the entire possible MOF space, as the number of topologies considered is only a fraction of possible topologies identified by IUPAC, and only a limited variety of building blocks was used. In addition, structures predicted by these methods are only *theoretical*. How many of these – apart from the well-known cases – have been synthesised, and how the remnant structures can be – if they can be at all – synthesised requires significant additional research.

1.2.2.2 How many MOFs have been synthesised?

The crystallographic data of most of the synthesised crystal structures accompanying a publication are deposited in the Cambridge Structural Database (CSD) curated by the Cambridge Crystallographic Data Centre (CCDC).[49] It contains data of experimentally-obtained organic and metal-organic crystal structures in the format of Crystallographic Information Files (CIFs) resulting from X-ray, neutron and electron diffraction

analyses.[50] Each structure has a unique CSD refcode composed of six letters, and if necessary, two digits. The number of submitted structures has greatly increased over the last 44 years, reaching the milestone of one million in 2019.[51] Among these data are those of MOFs. Several research groups published their methods for the extraction of a dataset of experimental MOFs.[52-54] Since most research groups focused on gas adsorption applications, the selected structures were porous 3D MOFs; the porosity being defined differently by each research group. All these studies followed a similar procedure: the MOF structures were i) extracted from the CSD, ii) filtered, iii) processed and cleaned in order to prepare them for simulations. Among the most common data processing steps are bound and/or unbound solvent removal, the addition of missing hydrogens and the elimination or repair of disordered structures. Watanabe et al. extracted 30,000 MOFs from the CSD, although no details were given regarding the selection of MOFs from the CSD.[52] Later on, Goldsmith et al. used a set of labelled MOFs and an algorithm to determine the features that indicate if a structure is a MOF.[53] Based on these features, they extracted 38,800 structures using unspecified CSD tools. Chung et al. developed the Computation-ready, Experimental (CoRE) MOF database, the first publicly available database of MOFs; in 2019, the CoRE MOF database was updated to contain over 14,000 curated structures.[54, 55]

However, these databases were mainly focused on gas adsorption applications. In addition, these databases required frequent manual updates, and provided little version control of the cleaning process the structures underwent. In order to offer a more global and free-to-process dataset, I, previously as an MPhil student in the Adsorption and Advanced Materials group, Department of Chemical Engineering and Biotechnology, University of Cambridge, UK, worked with the CCDC[43] to build the world's first CSD-integrated, automatically-updated MOF dataset: the CSD MOF subset. This new subset contains a wide variety of porous and non-porous, 1D, 2D and 3D MOF-like structures, which added up to a total of 70,000 structures in 2016 (see **Figure 2**) and is nearly 100,000 today.[43] To extract the data, I used ConQuest, search software implemented by the CCDC for the CSD (see Chapter 2 for more details), to develop seven criteria (see **Figure 3**) that filter the CSD database and return MOF-like structures. This subset is now available in the software package developed by the CCDC. The release of this dataset was accompanied by the publication of Python scripts for users to remove unbound and/or bound solvents on the appropriate structures.

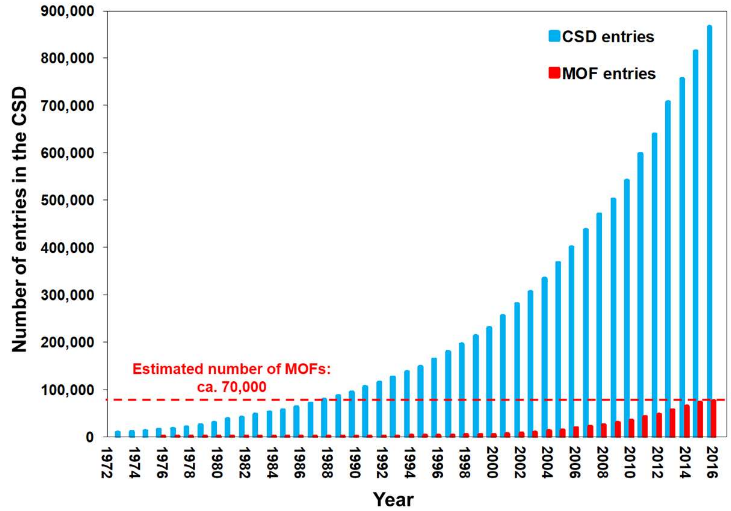


Figure 2 Evolution of the number of CSD entries and estimated number of deposited MOF structures up to 2016.[43]

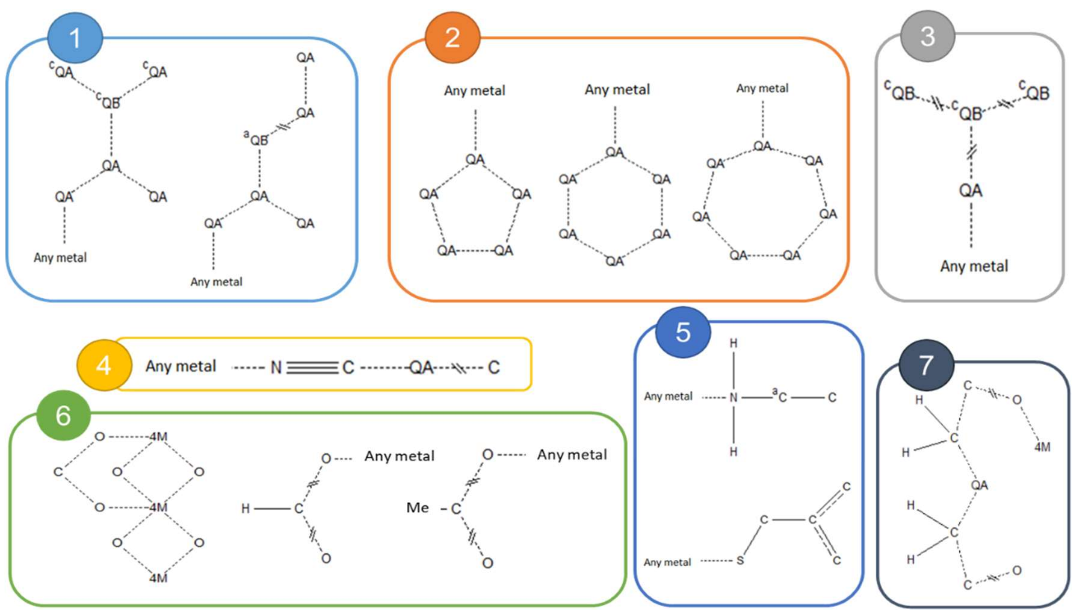


Figure 3 Criteria used to extract MOF and MOF-like structures from the CSD. QA = O, N, P, C, B, S. QB = N, P, B, S, C and superscripts "c" and "a" impose the corresponding atoms to be "cyclic" or "acyclic", respectively. "Cyclic" here means the atom is part of a path (formed by the atoms and bonds) that leads back to the given atom. "Acyclic" refers to the case where no path leads back to the given atom. 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.[43]

Avci et al. recently illustrated the importance of such regularly updated databases by investigating the newly added structures in the CSD MOF subset. The authors analysed the structures for carbon capture and hydrogen storage in the CSD MOF subset in 2018 (3,857 structures) and in 2020 (10,221 structures). The results showed that a larger number of MOFs among the updated dataset exceed the target adsorption performance; many MOFs even outperform benchmark zeolites on specific adsorption metrics.[56] The authors consequently encouraged further data production for the discovery of new materials.

While the databases presented above are a good starting point for mining MOFs, a range of computational tools are needed for the analysis of their structure and performance. This is all the more necessary as the number of synthesised MOFs available is already large and – as shown with the previous example – expected to continue its growth.

1.3 Computer-aided study and discovery of new materials

The large amount of data illustrated above, combined with the growing toolbox of computational chemistry[57] and computational power available today have paved the way to computer-aided studies and discoveries of new materials. The toolbox of molecular simulations allows researchers to easily characterise the materials and assess their performance for a given application. The computational power has significantly decreased the time needed to compute these properties. The huge amount of available data has introduced the use of computational high-throughput screenings (HTS), where the simulations are performed on a large number of structures. I herein review elements of molecular simulations and HTS relevant to this work.

1.3.1 A toolbox of molecular simulation techniques

Molecular simulations are a set of techniques used in a range of fields – biology, chemistry, physics, materials science – for either elucidating phenomena that are difficult to access experimentally, or complementing and confirming experimental observations. For the study of crystalline structures such as the porous materials presented in this thesis, the structural inputs are given as CIFs, which contain the coordinates of the structures' atoms. Then, depending on the techniques used, and with a correct definition of the system and the variables involved, information at scales ranging from a structure's pores (mesoscale) to its atoms (nanoscale) and its electrons (quantum-scale) can be calculated. Molecular simulations thus bring invaluable insights

into the understanding of a structure's behaviour, either as a stand-alone technique or as a complement to experimental results.

1.3.1.1 Geometrical characterisation

To understand the porous structure of MOFs (mesoscale), it is necessary to carry out a geometrical characterisation. Some of the properties that geometrically describe MOFs are density, largest cavity diameter (LCD), pore limiting diameter (PLD), pore volume, surface area and percolation. **Figure 4** presents these properties. The bulk material is represented in grey and the porous area in white. The LCD is the size of the largest sphere that can fit in a cavity, the PLD is the smallest opening of the channel that a molecule can diffuse through (**Figure 4a**). There are several definitions of pore volume and surface area, of which **Figure 4** presents the concepts used in this thesis. The yellow area in **Figure 4b** corresponds to the geometric pore volume of the structure. **Figure 4c** presents the notion of accessible surface area, obtained by the centre of a probe (represented in blue) rolling over the atoms. It is represented by the solid orange line in the cross-sectional view. The orange volume enclosed is the accessible pore volume. Conversely to the geometrical pore volume presented in **Figure 4b**, the accessible pore volume is obtained with a probe of non-zero size. A common probe used is nitrogen, as it allows comparisons with surface area measurements obtained experimentally, such as the BET (Brunauer-Emmett-Teller) area, where the molecules are assumed to form layers on the surface. Numerous studies have assessed and compared the different models behind the calculation and experimental measurements of surface area.[58, 59] In this thesis, I will use the accessible surface area as defined in this paragraph. **Figure 4d** presents the concept of accessibility and channel dimensionality. An isolated pore is shown in the top right corner. This pore is non-accessible; while its size, surface area and pore volume can be calculated, it cannot be occupied by any guest molecule. The rest of the schematic presents two porous channels crossing each other, thus forming a system of 2D channels – or 2D *percolation*. If another channel perpendicular to this page was connected to these two channels, the system would be 3D-percolated. The channel dimensionality is not to be confused with the overall structure dimensionality.

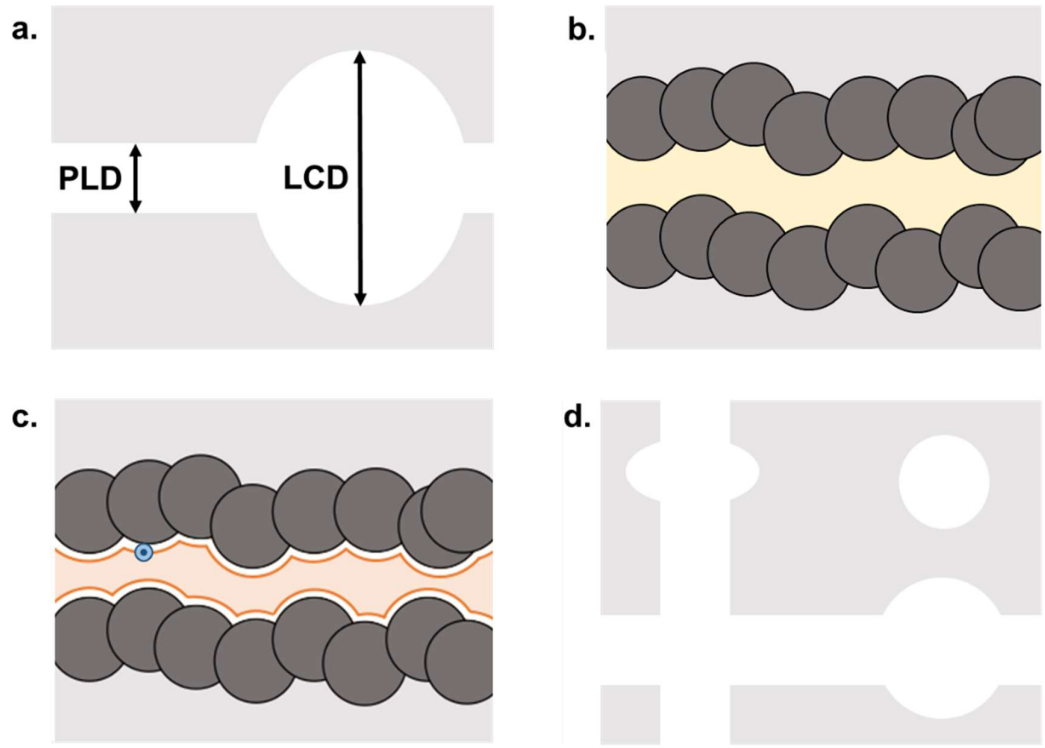


Figure 4 Schematic 2D illustration of **a.** largest cavity diameter (LCD) and pore limiting diameter (PLD), **b.** geometric pore volume, **c.** accessible surface area and **d.** percolation. The area coloured in grey represents the bulk of the structure. The porous areas are represented in white. The dark-grey circles in **b.** and **c.** represent the atoms at the frontier between the material and the pores. The blue circle in **c.** represents the probe used to obtain the accessible surface area. The yellow area in **b.** correponds to the geometric pore volume. The accessible surface area in **c.** is delimited by the orange line.

MOFs can be extended in one, two or three dimensions, forming MOF-rods, MOF-sheets or 3D-MOFs. While MOF-rods can only have 0D (isolated pores) or 1D channels, MOF-sheets can have 0D, 1D and/or 2D channels and 3D MOFs can have any type of channels. 2D and 3D MOFs can also have a combination of the various types of channels. The overall MOF dimensionality is another geometric property that can be calculated, for example, using open-source packages such as Zeo++.[60]

1.3.1.2 Classical simulations

With classical simulations, we enter the realm of atoms and molecules. For adsorption applications, it is important to understand the interactions between adsorbates and adsorbents, adsorbates and adsorbates, and the consequences of these interactions on a macroscopic level. This gap between the macroscopic and the atomic behaviour is bridged by methods based on stochastic processes, such as Monte Carlo (MC) simulations. MC methods use random sampling to compute a numerical approximation of a complex mathematical expression. In the case of adsorption, MC methods would typically sample from a system's microstates to compute the desired macroscopic properties. Each simulation relies on a model of the structure, a model of the adsorbates, a force field that describes the van der Waals and coulombic interactions involved and random moves (e.g. insertions, deletions, translations and rotations of adsorbates) that are accepted or rejected according the Boltzmann distribution. In many cases, the adsorbent structure is modelled as rigid, and it proves to be a good enough approximation for most MOFs.[61] Chapter 3 provides more details on the inner workings and the specifics of the MC methods used in this work. Amongst the information relevant to adsorption that can be obtained from MC methods are adsorption isotherms, Henry's coefficients and heats of adsorption.

Adsorption isotherms are used to describe the adsorption process of gases in a structure. They correspond to the plot of the amount of guest molecules adsorbed in a given structure at increasing pressures, at a fixed temperature. IUPAC identified six types of isotherms, the shapes of which reveal the types of porosities in the studied structures.[62] Using an MC ensemble of fixed temperature, pressure and chemical potential (called grand canonical Monte Carlo or GCMC, see Chapter 3), it is possible to obtain a full pure component or mixture isotherm by computing the amount of guest molecules adsorbed in a given structure at increasing pressure points. Comparisons between computational and experimental isotherms can either confirm a structure's behaviour, or reveal unusual properties – such as flexibility.[63, 64] As an example, **Figure 5** shows the experimental and computational nitrogen isotherm of ZIF-8 (zeolitic imidazolate framework) at 77 K obtained by Fairen-Jimenez et al.[64] The experimental isotherm represented with black circles shows two distinct steps. The computational isotherms obtained – assuming a rigid structure – are represented by the triangles. The isotherm with closed triangles was obtained with the conventional ZIF-8 observed at ambient pressures (referred to as ZIF-AP). The isotherm with open triangles was obtained with

an open phase of ZIF-8 observed at high pressures (referred to as ZIF-8HP). Neither computational isotherm matches the experimental results over the entire range of pressures. The ZIF-8AP isotherm presents a very different shape – a typical IUPAC Type I isotherm, with no steps, indicative of a microporous solid with relatively small external surfaces.[62] However, it perfectly replicates the isotherm at low pressures. The ZIF-8HP replicates the two-step shape well, matches the experimental isotherm at high pressures, but deviates from the experimental results at low pressures. The comparison of these three isotherms showed that the rigid models were not capable of explaining the structure's behaviour, and thus indicated a structural change around a pressure point of 2×10^{-4} P/P_0. This change was later on identified as a swing effect in ZIF-8's imidazolate linkers. The work presented in this thesis relies heavily on the computation of the amount of gas adsorbed in a structure – whether across a range of pressures or at single pressure points.

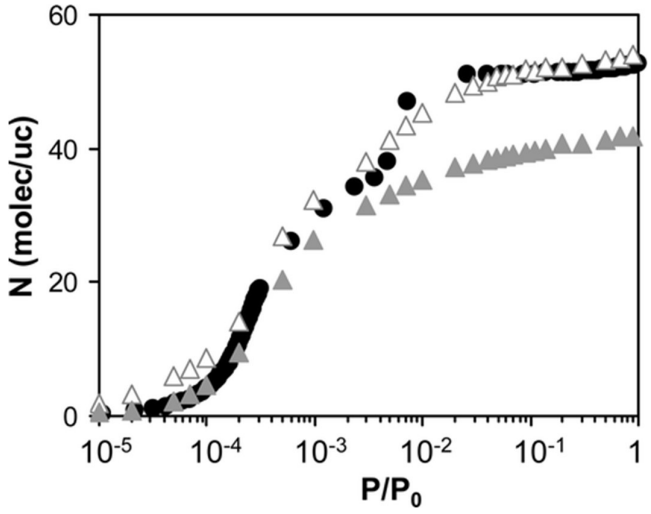


Figure 5 Nitrogen isotherm on ZIF-8 at 77 K. Black circles: experimental isotherm. Triangles: computational isotherm on the closed structure (closed triangles) obtained at ambient temperature, ZIF-8, and the open structure (open triangles) obtained at high pressures, ZIF-8HP.[64]

In the low-surface coverage regime – corresponding to Henry's regime, it is possible to obtain the Henry's coefficient via the slope of the curve. The Henry's coefficient is an indicator of the strength of the interactions between the adsorbates and adsorbent. The larger the slope is, the stronger are the interactions. In the previous example, the ZIF-8AP isotherm was able to replicate the Henry's coefficient well. However, it is also possible to compute it using a Widom test particle insertion[65] in the canonical ensemble

(fixed temperature, volume and number of molecules). This method inserts a ghost molecule into the system, thereby computing the molecule's energy as well as its averaged chain extension called Rosenbluth factor[66] (see Chapter 3) without affecting the system.[65] The Henry's coefficient is directly linked to the density of the framework and the obtained Rosenbluth factor and can be easily computed from there.[67] This method is particularly useful for calculating Henry's coefficients, as it quickly evaluates the guest-host interaction without including any guest-guest interactions.

The heat of adsorption (denoted Q_{st}) is another measure of the adsorbate-adsorbent interaction. Q_{st} is positive and corresponds to the opposite of the enthalpy of adsorption $\Delta H_{st} = -Q_{st}$, where ΔH_{st} is the heat released when a molecule is adsorbed to the surface. A higher magnitude of the enthalpy – or a higher heat of adsorption – indicates a stronger interaction. From the Clausius-Clapeyron equation, it is possible to obtain a value of Q_{st} by running simulations at different temperatures and differentiating with regards to $1/T$.[68] However, a large number of individual simulations need to be repeated to obtain an accurate value. Molecular simulations present two main alternatives. With GCMC simulations, the change in energy when one guest molecule is adsorbed can be calculated using the energy/particle fluctuations.[69] The final heat of adsorption can be obtained after a sufficient number of particle insertions and deletions are performed to obtain averages of energies and number of particles. Using the Widom test particle insertions, the heat of adsorption at zero coverage can be obtained directly via the average energies sampled in the system.[70]

The properties described in this paragraph are only a fraction of results that can be computed on MOFs. Classical molecular dynamics (MD) are a deterministic method based on Newton's law that describes the evolution of a system over time and can be used to compute a fluid's diffusion properties in a structure.[71] Quantum methods such as density functional theory are used to model crystals, calculate mechanical properties[57] or assign partial charges to MOFs.[72] Although MDs and quantum methods can unveil interesting behaviours, they remain too costly at this stage to perform HTS, the goal of which is to quickly and cheaply sift a large amount of data.

1.3.2 Computational high-throughput screening

The aim of HTS studies is usually two-fold: i) to identify the best performing structure for a given application and ii) to uncover interesting structure-property relationships that can guide researchers towards more rational designs of MOFs in the future. From the small dataset of 14 manually collected MOF data in 2009 for the study of carbon capture[73] to half a million structures screened for hydrogen storage in 2019,[74] a booming number of HTS studies have been published. **Figure 6** presents the general workflow adopted by computational MOF scientists for HTS studies. It focuses on experimental data derived from the CSD, but the approach is the same with another experimental or hypothetical database. The first part of the workflow – data gathering and processing – corresponds to what was described previously regarding databases. Starting from the CSD, search queries were made to extract potential MOF structures, thus forming the CSD MOF subset. Users have then the possibility to focus on non-disordered structures, remove certain solvents and add missing hydrogen atoms. Once a clean set of structures is obtained, users can proceed to the geometrical characterisation of the dataset. The PLD is an important property to calculate, as the number of structures to run simulations on can be reduced to MOFs with porous channels that are big enough for the studied adsorbate. Finally, the molecular simulations can serve the two goals of HTS: select a few structures of interest and map out the structure-property trends of the given application with the given structures. The HTS examples and studies presented throughout this thesis all follow roughly this workflow.

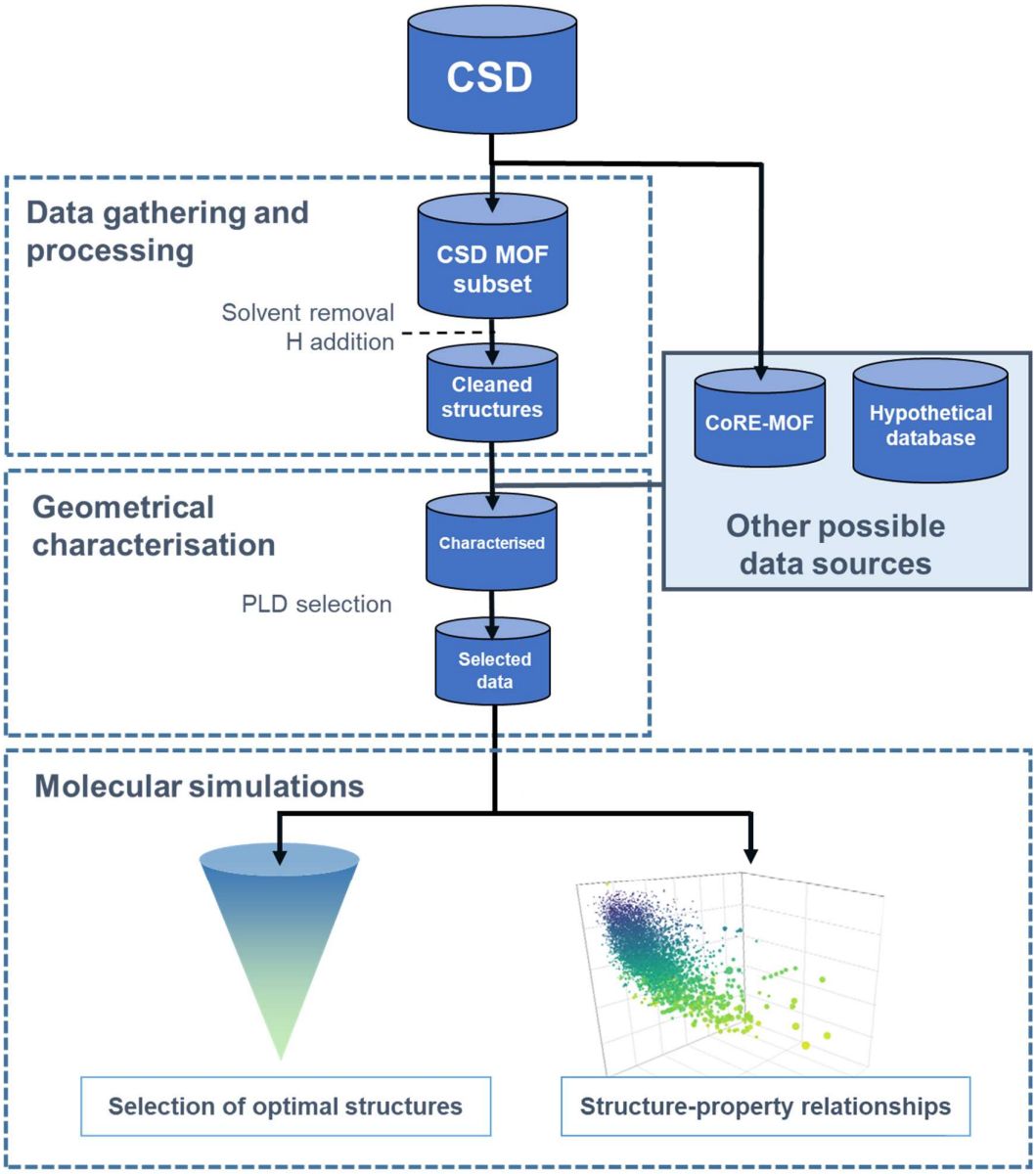


Figure 6 General workflow for the computational HTS of experimental MOFs derived from the CSD.

1.3.2.1 Identifying the best structure for a given application

While the attempts at identifying the best structures for a specific task have been numerous,[75] only a minority of these studies were experimentally evaluated. **Table 2** presents a few examples of these cases, which span a range of applications: methane storage, carbon capture, hydrogen storage, oxygen storage and chemical warfare agents capture.

Wilmer et al. carried out one of the earliest of these studies for the adsorption of methane at 298 K and 35 bar.[44] The storage of methane at room temperature and high pressures is extremely useful for natural gas-powered vehicles, the main challenge being the ability to store enough methane for given driving distances. MOFs could potentially lead to cheap, high-density tanks that meet the US Department of Energy target of 263 cm^3 cm^{-3} at standard temperature and pressure (STP). For this screening, the authors used their in-house hMOF database presented in the previous section and used several rounds of GCMC simulations with increased number of cycles on a smaller amount of data (the best-performing data after each round). Among the 300 top structures that performed better than the then world-record (230 cm^3 (STP) cm^{-3}), the existing – but unbeknownst to the authors – NOTT-107 (previously named after the University of Nottingham)/MFM-107 (later on named as Manchester Framework Material) was synthesised. However, the measured uptake was 8% lower than the predicted value, and lower than the record. The authors explained the disparity with the possible incomplete pore activation of the synthesised MOF.

The assignment of density-derived electrostatic and chemical (DDEC)[72, 76] charges to structures from the CoRE MOF database opened up more possibilities in the study of adsorptions where electrostatic interactions play a major role. The high quality of the charges and their availability in the CIFs themselves make the DDEC fully ready for HTS. Using the DDEC database, Moghadam et al. performed GCMC simulations and found the best existing candidate for oxygen storage at 298 K and a pressure swing of 5 – 140 bar, UMCM-152 (University of Michigan Crystalline Material).[77] Oxygen storage is a relatively less explored gas adsorption application with MOFs. Its promising uses include oxygen tanks in the healthcare industry as first aiders, in the military and aerospace industries.[78] The identified structure was then synthesised and its uptake experimentally confirmed to be 22.5% higher than the previously best-performing structure reported in the literature. Using the same database, Matitos-Martos et al. found an ideal structure for the capture of diethylsulfide (DES) in moist environments. DES is a simulant of mustard gas, used as a chemical warfare agent. After a first round of selection using the Henry's coefficients of water to estimate the structures' hydrophobicity, GCMC was performed on the top-performing structures. The Henry's constants were obtained using Widom test particle insertion methods[65] and were deemed a good indication of the adsorbent-adsorbate interactions. The identified

structure, of CSD refcode UTEWOG,[79] was synthesised according to the existing protocol and its performance validated.

Another approach is to combine hypothetical and experimental data. This is examplified by Chung et al. who combined GCMC with a genetic algorithm (GA) on hypothetical structures, before making precombustion carbon capture predictions on the CoRE MOF database at 313 K.[80] Carbon capture and storage represents an interesting transitional solution while fossil fuels are still in use. For recent power plants, carbon can be captured via a precombustion carbon technology, where natural gas is first reformed into a mixture of CO and H_2, before going through a water-gas shift reaction which produces a high pressure steam of CO_2 and H_2. GAs are a class of optimisation method inspired by the theory of natural selection. The algorithm starts with an initial population of structures and a definition of fitness function. The genetically fittest structures then evolve to give birth to the subsequent generations. In this case, the GA looked for structures with high carbon uptakes and high carbon/hydrogen selectivity. The ethoxy-functionalised NOTT-101/MFM-101 (NOTT-101/Oet or MFM-101/Oet) was found to be the best performing hypothetical MOF. After applying the trained GA to CoRE MOF, the structure with the CSD refcode VEXTUO[81] was found to be another promising structure. Both structures were synthesised and NOTT-101/Oet was confirmed as the new record for this application.

More recently, Bucior et al. combined GCMC and supervised learning based on the structures' potential energy histograms to screen a dataset of more than 50,000 structures composed of a mix of different available experimental databases for hydrogen adsorption under temperature and pressure swing (TPS) conditions (77 K, 100 bar – 160 K, 5 bar).[82] As a promising clean vehicular fuel, hydrogen is by far the most computationally studied gas for adsorption application in MOFs.[25, 53, 83-93] In this study, the authors found MFU-4l (metal-organic framework Ulm University) as one of the top-performing materials with an experimental deliverable capacity of 47 g L^{-1}, thus ranking among other previously identified structures.[93] Ahmed at al. soon after screened ca. 500,000 structures composed of a mix of all available hypothetical and experimental data for hydrogen storage at the cryogenic pressure swing conditions of 5 – 100 bar. After a first selection of structures using the semi-empirical Chahine rule, GCMC was applied to ca. 44,000 structures. Three candidates were identified: SNU-70 (Seoul National University), UMCM-9, PCN-610 (porous coordination network)/NU-100

(Northwestern University), all of which were synthesised and shown to perform better than MFU-4l at the same previous TPS conditions. PCN-610/NU-100 and UMCM-9 were existing MOFs whereas SNU-70 was a hypothetical one.

Table 2 Examples of MOFs identified via HTS and validated experimentally.

Authors	Year	Application	Data	Identified and synthesised MOF
Wilmer et al.[44]	2012	Methane storage, 298 K, 35 bar	137,953 hMOFs	NOTT-107/MFM-107
Chung et al.[80]	2016	Carbon capture, 313 K, up to 16 bar	Genetic algorithm on 55,163 hMOFs and then on 5,169 CoRE MOFs	NOTT-101/Oet or MFM-101/Oet, VEXTUO
Matito-Martos et al.[94]	2018	Diethylsulfide (mustard simulant) over water selectivity	2,932 DDEC	UTEWOG
Moghadam et al.[77]	2018	Oxygen storage, 298 K, 5 – 140 bar	2,932 DDEC	UMCM-152
Bucior et al.[82]	2019	Hydrogen storage, 77K, 100 bar – 160 K, 5bar	A mix of > 50,000 including CSD subset	MFU-4l
Ahmed et al.[74]	2019	Hydrogen storage, 77 K, 5 – 100 bar	A mix of 493,458 including CoRE MOFs and the CSD	SNU-70, UMCM-9, PCN-610/NU-100

1.3.2.2 Uncovering structure-property relationships

In addition to identifying the best candidate for a given application, HTS studies can unveil interesting structure-property relationships, thereby providing the community with rational design rules. In this section, I present some of the trends observed in the studies mentioned above.

In their HTS of MOFs for methane storage, Wilmer et al. found a trade-off between maximing the structures' gravimetric surface area and their storage capability, with an optimum point at 2,500 – 3,000 m^2 g^{-1}. A pore volume that is too large also has a negative impact on the uptake. In fact, it was found the ideal pore size corresponds to either exactly one or two methane molecules. In addition, methyl-functionalised MOFs, such as the identified NOTT-107/MFM-107, usually performed better.[44] Similar trade-offs were observed in the work carried out by Moghadam et al. on oxygen storage, where a ceiling of 250 cm^3 (STP) cm^{-3} was reached. Structures with cavities larger than 10 Å and void fractions higher than 0.8 do not improve this volumetric uptake.[77] Matito-Martos et al. found that the strongest chemical warfare agents-MOFs interactions (i.e. with Q_{st} values between 100 and 200 kJ mol^{-1}) took place in structures with rather high surface area (up to 2000 m^2 g^{-1}) with an optimum Henry's contant for LCDs between 5 and 6 Å.[80] Bucior et al. found that a relatively weak adsorbate-MOF interaction is ideal for hydrogen storage at cryogenic conditions,[82] while Ahmed et al. estimated a volumetric ceiling of 40 g L^{-1}, where the volume refers to the volume of MOF.[74] The latter also suggests that total capacities and usable capacities might not follow the same structure-property relationships.

The cases presented here are only a fraction of the results published by the computational MOF community. The diversity and amount of crystallographic data generated in the last two decades, and the added calculated properties have moved the field into the new paradigm of big data. While each computational research group has adapted themselves slowly to the age of big data analysis and machine learning, the field still needs signficant improvements – maybe a holistic framework? – in order for different researchers to share their work efficiently.

1.4 Too much data?

At the moment, the computational MOF field is booming, but few standards exist to unify the various studies. This lack of structure means resources are often wasted as published results are either not seen, not accessible or not reproducible. It is even more difficult for new scientists joining the MOF world, as a wider, more varied and complex body of skills is required. In addition, the lack or the poor communication with experimental researchers means the majority of the simulated results remain theoretical and are not put into practice. In this section, I highlight in particular two major issues: the variety of databases available and the difficult visualisation of the data produced.

1.4.1 Too many databases?

As demonstrated earlier, there is a range of MOF databases one can choose from to carry out HTS. Evidently, the final calculated outcome depends heavily on the quality of the data at the source. Moosavi et al. very recently compared the diversity of MOFs in several databases – the CoRE MOF database being the only one based on experiments.[48] The authors used revised autocorrelation functions (RACs) to compute the correlations between heuristic atomic properties on a molecular graph, thus extracting information such as linker chemistry, metal chemistry and functional groups.[95] Combined with the analysis of simple geometric descriptors, it was found that the databases considered covered a significantly different chemical space. In particular, the one experimental database (CoRE MOF) was composed mostly of structures with small pores, while the hypothetical databases covered each a different pore size region. This is understandbly a natural consequence of how the databases were built, as they were intended to cover different topological spaces. Another result of these artificially built structures is a more thorough coverage of pore geometry, linker chemistry and functional groups spaces. However, the metal chemistry in these hypothetical databases is significantly less varied and present. This is because metal clusters in MOFs are only known once they have been synthesised, as opposed to readily available lists of possible organic linkers and functional groups. Taking their analysis even further, Moosavi et al. applied machine learning to these databases and found that the resulting most important variables differed from one database to another. Metal chemistry is likely to be considered not important for certain applications if studied on hypothetical databases for instance. Finally, the authors show that the biases in the different databases mean that machine learning might not always be transferable, the best case scenario being an algorithm trained on a more diverse set of structures and tested on a biased dataset.

Through their diversity study, Moosavi et al. highlighted the potential differences between CoRE MOFs and a range of hypothetical MOFs, and the dangerous general conclusions one can draw when relying on biased datasets. In addition, while hypothetical databases aim to map out the unexplored regions of the MOF space, the existing ones are very restricted and heavily depend on the knowledge drawn from real structures. Unfortunately, the CSD MOF subset was not considered in this study. How different is it from CoRE MOF?

To solve this question, Altintas et al. reported the comparison of the two experimental databases (5,109 structures from CoRE MOF version 2014 and 19,123 non-disordered from the CSD MOF subset version 5.37 May 2016, where the authors also removed solvents using the provided Python script) in the case of 3D MOFs, for methane and hydrogen adsorption.[96] Among the 3,490 structures in common in the two datasets, 387 differed significantly in the final gas uptakes. These differences stem from the different modifications the original data underwent during the cleaning process in both databases. The authors went on to compare the uptake obtained in both cases to the actually measured uptake from the original papers. According to this study, neither database is in perfect agreement with reported experimental results. However, the exact changes applied to the original CSD CIFs in both cases were also not made clear for the comparison of these two datasets. Although the nature of the modifications made on the original CSD data to obtain CoRE MOF is known, it is difficult to understand what exactly has been modified from the final CIF and the impact of these changes on the simulated uptake. Similarly, although the provided Python script for the removal of bound and unbound solvent was used on the CSD MOF subset, it is unclear on which subsets of structures it was run for this comparison. As I and others highlight and insist, researchers should be extremely careful when removing bound solvents, as the MOFs' structural integrity might be impacted, resulting in unrealistic simulated uptakes and selectivities.[43, 97]

To overcome the issue of reproducibility, a trackable workflow such as the Automated Interactive Infrastructure and Database for Computational Science (AiiDA) could be helpful.[98] Ongari et al. recently demonstrated the use of AiiDA on a database of CURATED (Clean, Uniform, and Refined with Automatic Tracking from Experimental Database) covalent-organic frameworks (COFs).[99] The workflow and results obtained at each stage are available online on the Materials Cloud platform.[98] While this

infrastructure is targeted towards computational researchers, the same authors suggested to take this idea one step further by creating a platform which could match experimentally obtained materials and applications, thereby bridging the gap between experimental and computational scientists.[100] Though promising, implementing such a workflow will need significant effort and time from the computational MOF community. In the meantime, more efforts are being made towards allowing easier exploration of published data, for both computational and experimental scientists alike.

1.4.2 Towards easier data exploration

With a large amount of data comes the following questions: what to visualise and how to best visualise it? While the topic of data visualisation might seem trivial, clear, flexible, informative, biased or un-biased plots are crucial for i) conveying the desired message to the entire community – computational and experimental, and ii) carrying out extensive exploratory data analyses prior to applying the plethora of now ubiquitous machine learning algorithms. And it seems that scientists are not the best at creating visualisations just yet,[101] so much so that *Nature Methods* published a set of guidelines from picking the right plot for the right data, to using colourblind-friendly colours, from avoiding rainbow gradients for continuous data to choosing the right fonts.[102-106] Some remarkable improvements have however been made recently in the field. Along with the (re)discovery of UMCM-152, Moghadam et al. published an online interactive data explorer where users can plot all the available textural and adsorption properties in order to spot interesting structure-property trends and potential structures of interest.[77] Over 1,000 plots can be easily obtained, by choosing different axes, colours and sizes for the available variables.[77] Users can also follow the evolution of the properties of a structure as the pressure point changes. In addition, each structure has a link to the corresponding CSD entry web page. Plots can be zoomed in and out, the corresponding data filtered a priori or a posteriori and snapshots can be extracted directly. Such data visualisation tools were then adopted by Matito-Martos et al. for the publication of the data obtained for the capture of chemical warfare agents.[94] Similar visualisations are now also incorporated in the Materials Cloud platform.[99]

There is still, however, a gap between the users being able to visualise other people's data and to plot their own. In a *Nature* toolbox section, Perkel called for more accessible data visualisation tools.[107] In particular, the ability for researchers to easily plot interactive figures could not only drive story-telling, but also reproducibility. Following

up on this, Balzer et al. recently developed Wiz, a free web app for the codeless, interactive visualisation of any large datasets.[108] This tool, born from the MOF field, is announced to extend its functionalities to data analysis. Going one step further, Sarkisov et al. published a program for the computation and visualisation of principal component analysis for MOFs with pre-tabulated data.[109] This tool is planned to accommodate any kind of data, thus paving the way to lowering the entry barrier to big data analysis and visualisation.

1.5 Aims and goals of this work

The previous sections introduced the background of the computational MOF field in terms of HTS. In particular, I highlighted the various MOF databases, the importance of the data quality at the source and the differences between CoRE MOF and the CSD MOF subset; the main advantages of the latter being that 1) users have access to the original experimental data, and thus can process them according to their own needs and standards, ii) it is curated and automatically updated quarterly. To map out more clearly the landscape of the CSD MOF subset and to provide users with even more control over the data they want to study, the aim of this thesis is to develop easy-to-use tools to simplify data selection and exploration. Drawing on my experience from building and exploring the CSD MOF subset, I then aim to extend the developed methods to COFs, MOCs and OCs.

1.5.1 Exploring the CSD MOF subset

Although MOF databases in conjunction with HTS have proven to be extremely useful for the study of structure-property relationships and the screening of MOFs to find optimal materials, little information exists on how MOFs can be classified based on their important chemical and structural anatomy. Such a breakdown could be very useful towards a better understanding of the materials' behaviour and consequently for researchers wishing to focus on certain chemistries deemed more relevant to their area of study. Unbeknownst to me, and in parallel to this work, Moosavi et al. presented a method using RACs to extract MOF chemistry information. I developed simpler methods using CCDC tools for the targeted classification of MOFs in the CSD MOF subset. The chosen classifications include: structures containing specific chemical functionalities, specific metal-cluster, framework dimensionality or chirality. In collaboration with Dr Marcus Fantham, Laser Analytics group, Department of Chemical Engineering and Biotechnology, University of Cambridge, UK, I also present an online interactive platform for the exploration of the CSD MOF subset.

1.5.2 Building other porous structures databases

As presented in the first section of this introduction, other rising and promising porous materials include COFs and porous molecular solids composed of MOCs and OCs. While a few COFs databases do exist (see Chapter 3), I demonstrate the use of the developed methods for their extraction from the CSD. Moving on from extended structures, I then dive into the world of cages, for which the inherent shape differences require a completely different approach to the extraction of their data (Chapter 6).

1.6 **Thesis outline**

As a result of the previous section, the remant of this thesis is organised in three parts. Chapters 2 and 3 introduce the various methods used throughout this work:

- Chapter 2 explains the CSD tools used to develop the CSD MOF subset and to obtain the results presented in Chapter 4. Chapter 2 is a modified version of a published tutorial review. A case study demonstrates the use of these tools for the derivation of a CSD COF subset.
- Chapter 3 gives an overview of the methods used for the molecular simulations carried out in this work. It is complemented by three case studies, two of which are published as part of two experimental-computational collaborations. The last one introduces the concept of reverse HTS, where a given MOF is screened with a library of adsorbate molecules.

Chapters 4 to 6 present the exploration of the CSD MOF subset and the derivation of the cages dataset:

- Chapter 4 presents the tools and results obtained for the targeted classification of MOFs in the CSD MOF subset, using the methods from Chapter 2.
- Chapter 5 presents the results of a HTS carried out on the CSD MOF subset with the methods from Chapter 3 and the tools developed in Chapter 4 for the study of hydrogen storage at room temperature and high pressures.
- Chapter 6 leaves the domain of framework structures and presents the use of topological data analysis as a method for the identification of organic and metal-organic cages in the CSD.

Chapter 7 concludes the thesis and provides an outlook on future steps.

2 METHODS FOR EXPLORING POROUS MATERIALS IN THE CAMBRIDGE STRUCTURAL DATABASE

Parts of the following content are published in *Enabling efficient exploration of metal-organic frameworks in the Cambridge Structural Database*, **Aurélia Li**, Rocio Bueno-Perez, Seth Wiggin, David Fairen-Jimenez, CrystEngComm, 2020, 22**,** 7152-7161.

Abstract. As the Cambridge Structural Database (CSD) reaches a record number of one million deposited structures in 2019, the metal-organic framework (MOF) community sees its own pool of synthesised structures continue to grow to almost 100 000 entries. The CSD has thus become the most complete treasure trove in which computational researchers are trying to find the most relevant data. This chapter, originally published as a tutorial review, presents the most useful tools for the efficient exploration of the CSD for MOFs applications and the future possible developments to further enhance the discovery of MOFs. A case study concludes the chapter by digging for COFs in the CSD. The results presented in this case study are my own work.

The Cambridge Crystallographic Data Centre (CCDC) has developed a range of software for the exploration and analysis of crystalline data gathered in the Cambridge Structural Database (CSD). I will focus here on ConQuest and the CSD application programming interface (API), for high-throughput search and analyses. I will also sometimes refer to Mercury, the CCDC software for the visualisation and analysis of molecules.[110]

2.1 Digging into the CSD for MOFs with ConQuest

ConQuest is the primary structure search software developed by the CCDC for exploring the data in the CSD.[111] It offers a wide range of search possibilities, from drawing a fragment of a targeted structure, to specifying its space group, and from combining different search queries to combining different search results. It can be used to find one specific structure, but also a subset of structures – such as metal-organic frameworks (MOFs). As early as in 2004, Ockwig et al. endeavoured to classify MOFs in the CSD according to their respective topologies.[112] An older version of the software, Quest, was then used to carry out a string search, which returned 1,127 three-periodic MOFs. However, no further details on the exact strings were given.

As previously explained in Chapter 1, several research groups have built their own databases from data extracted from the CSD. In particular, CoRE MOF was the first publicly available accessible database and contains today over 14,000 structures.[54, 55] Chung et al. selected structures from the CSD using ConQuest and their own definition of a MOF: "structures with more than one bond between metals and the elements O, N, B, P, S, and C [… and] any kind of bond from these six elements to C, N, P, or S atoms."[54] Users should be aware that, although a significant number of structures from the CoRE MOF database still have their original CSD refcode, they have been modified to be simulation-ready. Most recently, I developed the CSD MOF subset using criteria drawn in ConQuest to capture MOFs and MOF-like structures.[43] Conversely to the previous datasets, the CSD MOF subset was designed to be of use to researchers working on applications that are not restricted to gas adsorption. The subset thus contains 1D, 2D and 3D MOFs that do not necessarily have any apparent porosity. These criteria are implemented in the CSD as a filter so that the dataset is automatically updated, along with the quarterly CSD updates. These criteria are also very flexible and can be easily tailored to better fit the evolution of the definition of MOFs.[22, 43] The

structures in the CSD MOF subset are the original data as-deposited and curated by the CCDC.

2.2 Using ConQuest to access the MOF subset

One of the most useful search tools within ConQuest is *Draw*. With this function, users can draw entire molecules or fragments that should be contained in the targeted structure. The criteria developed for the CSD MOF subset heavily used *Draw* (see **Figure 3**), and was combined with an additional keyword search that captured the extended nature of MOFs. Indeed, as extended structures, only a repeating unit (RU) is used to represent MOFs in the CSD. Therefore, several expansions of this RU are necessary to form the final framework. These expansions are encoded in the CSD within the concept of "polymeric" bonds which connect these RUs. These bonds are represented in a zig-zag shape in a 2D diagram. Conveniently, polymeric structures containing metals are tagged in the CSD with the word "catena", which constituted the additional keyword search.[43]

The CSD MOF subset is accessible in ConQuest via *View Databases > Lists in CSD version X [this will depend on the software version] > MOF subset.* The *Non-disordered MOF subset* is also available and can be used as a starting point for high-throughput screenings (HTS). However, the term "non-disordered" here can be misleading and deserves further explanation. The crystallographic disorder is flagged differently in the CSD depending on the exact nature of the disorder. A structure is normally classified as disordered if there is any non-hydrogen disorder present in the whole structure, that is the framework and any other unmodelled molecule present – i.e. solvent and guest molecules commonly seen in MOF-like compounds where the disordered solvent is treated using the Platon/Squeeze[113] or Olex2/Mask[114] tools. In other words, a non-disordered structure in the CSD might still have missing or disordered hydrogen atoms. However, the *Non-disordered MOF subset* is intended to contain structures with no disorder within the framework, including no hydrogen disorder or missing hydrogens, but there might still be disorder in the unmodelled molecules. The algorithm developed by the CSD for identifying these "non-disordered" MOFs works as follows: i) look for disordered atoms (i.e. cases of multi-site disorder); ii) search for the nearest neighbouring non-disordered atom; iii) if this non-disordered atom is part of the framework, the structure is considered as disordered, if not, (i.e. near a solvent molecule), it is considered as non-disordered. **Figure 7a** summarises the

organisation of the MOF subsets in the CSD. **Figure 7b** gives a summary of the differences and overlaps between the *Non-disordered MOF subset* and the structures obtained with a "non-disordered" filtered search in the main *MOF subset*. In the latter case, the "non-disordered" filter applied to the CSD MOF subset excludes entries with the disorder in the unmodelled molecules but keeps frameworks with hydrogen disorder. Errors might still exist in the database and users are encouraged to report them to the CCDC. Once the desired subset is loaded, users can export the list of structures to a GCD list, a text file containing the list of refcodes, or to CIF or PDB files, amongst other formats. The GCD file can then be read directly from the CSD Python API.

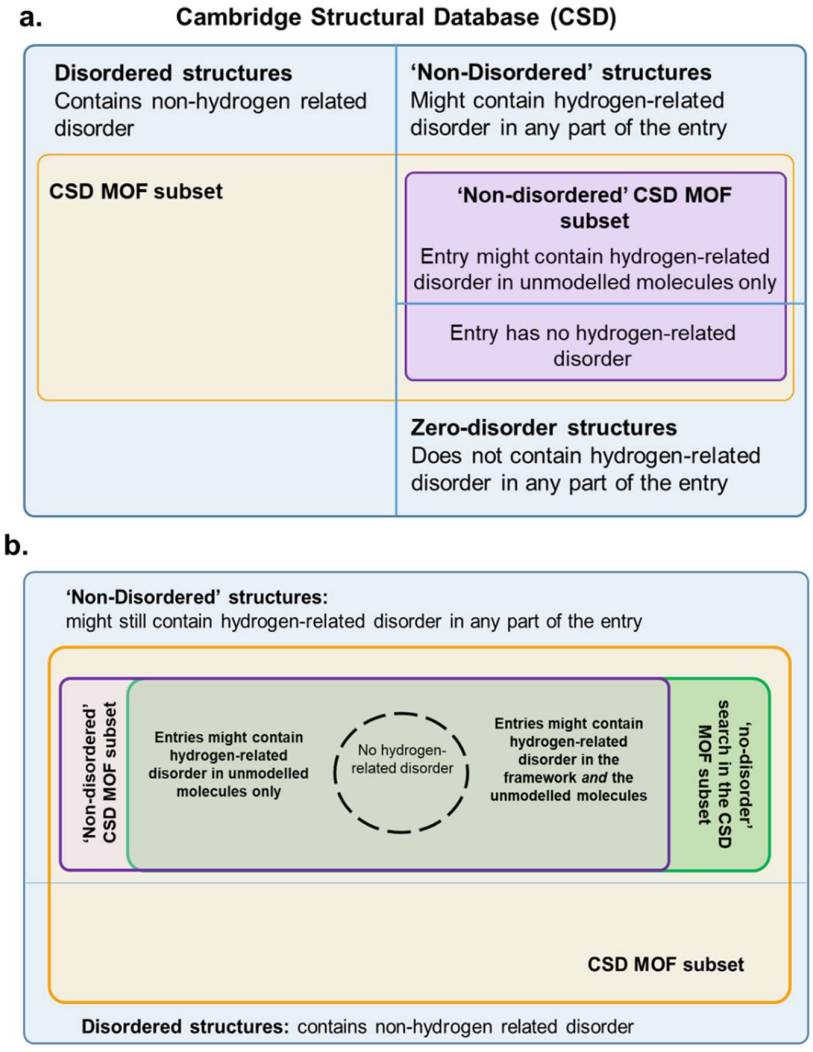


Figure 7 a. Organisation of the MOF data in the CSD. **b.** Differences and overlaps between the *Non-disordered MOF subset* and a "no-disorder" filtered search in the CSD MOF subset. About 80% of the *Non-disordered MOF subset* is included in the "no-disorder" filtered search.

2.3 Notes on ConQuest queries

Although a MOF subset is provided, users are encouraged to perform their own searches within either the CSD or the MOF subsets and, most importantly, to report the exact queries used when publishing. Indeed, queries are rarely reported or, at best, only translated into words. These can be highly misinterpreted and have very low reproducibility. I also noticed that, often, only one *Draw* query is mentioned, when most of the time, a combination of different queries is necessary to cover the whole spectrum of possible results. Since the basics of how to use ConQuest for queries are covered in the user guide (available in *Help* > *Help Index*) or online,[115] I will only give a few comments on how to improve MOF searches.

Draw is a deceiving function, as it appears simple and easy to use when experience shows that proper usage really is an art that requires many trials and errors. When a *Draw* query is made, ConQuest will look through the 2D diagrams of the CSD structures and find an exact match to these 2D diagrams. There are different ways of representing the same structures with a 2D diagram, and this is especially the case for extended structures which are only partially represented with "polymeric" bonds.[116] Thinking about where and how the polymeric bond can be defined in a RU is tricky. **Figure 8** shows two structures containing carboxylate ligands connected to zinc atoms – one of the simplest configurations. As highlighted by the blue circles, there are at least two ways of representing such a linkage. A single query describing the circled linkage in **Figure 8a** seems intuitive but is too specific and will miss structures such as the one in **Figure 8b**. Therefore, researchers are highly encouraged to examine the resulting diagrams after each search. These are displayed in the *View Results* panel by default, where the matching substructure is highlighted in red. This will guide users towards a better-tuned query, or to combining different queries instead. Users can either create different queries and combine the resulting hitlists in *Manage Hitlists* or combine queries directly in the *Combine Queries* panel. When a set of queries is satisfactory, it can be saved, exported and shared (e.g. as supplementary information) to be reused by other users.

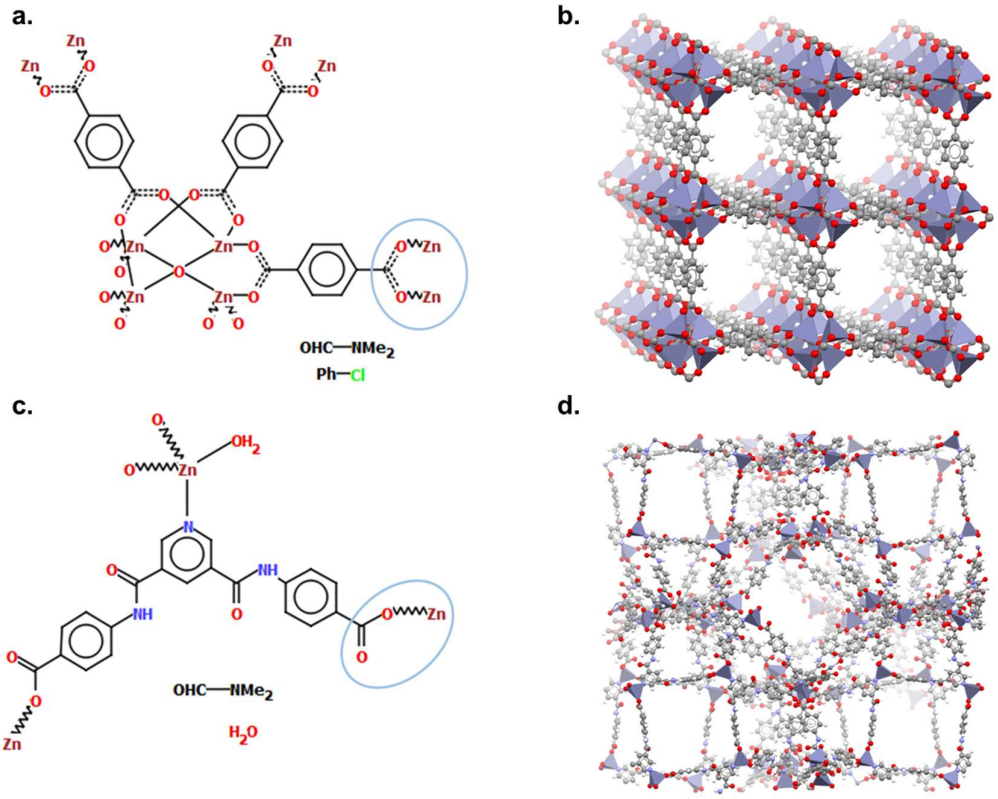


Figure 8 Example of two different ways of representing carboxylate ligands: **a.** chemical diagram of SAHYIK, **b.** 3D representation of SAHYIK, **c.** chemical diagram of ADUROI and **d.** 3D representation of ADUROI. On the chemical diagrams, the blue circles highlight two different representations of the same linkage. The 3D representations are obtained with Mercury, where oxygen atoms are represented in red, carbon atoms in grey and hydrogen atoms in white. Blue-grey tetrahedrals highlight the metal clusters.

2.4 Removing solvents with the CSD Python API

Once the desired MOF subset is obtained from ConQuest and saved as a GCD file, users can explore the data with the CSD Python API[49] for more efficient data mining and structures modification. For the removal of bound and unbound solvents, a Python script was previously published. [43] It takes a GCD list of structures, a solvent list when necessary and outputs the desired CIFs. The algorithm looks for metal atoms present in the framework, removes all bonds around them, and compares the removed fragments to a list of solvent. When no list is provided by the user, the algorithm uses the default CCDC most common solvent list. It is important to note here that the provided script will remove both bound and unbound solvents. However, it is recommended to remove

bound solvents on specific cases only (such as structures containing Cu-Cu paddlewheels or similar to CPO-27/MOF-74). To remove unbound solvent only with this script, use an empty solvent file. Another simple way of obtaining the same result with the API would be to look for the heaviest weight component (`heaviest_component`) of an entry and return only this part of the entry as a CIF. The `heaviest_component` corresponds to the component in the CSD entry with the highest molecular weight, a component being a group of atoms linked with bonds and thus forming a distinct unit. The `heaviest_component` is – in general – the framework. However, exceptions exist and it is wise to check that it is indeed polymeric (i.e. by using the `is.polymeric` attribute). If it is not, one of the substructures must be polymeric by definition of the subset, and that substructure should be kept as the framework.

2.5 Adding missing hydrogens

As explained earlier, the *Non-disordered MOF subset* should not include frameworks with missing hydrogens. However, "missing hydrogens" is another misleading expression that requires clarification. In a CSD entry, only the atoms modelled from the original data have coordinates and can be visualised in Mercury. Hydrogen atoms are sometimes not found in the original data, therefore not modelled, but are still accounted for in the CSD so the overall structure makes chemical sense. These hydrogen atoms are referred to as "siteless hydrogens" in the CSD. They do not appear in the original structure's CIF but are taken into account in the 2D diagrams and in search queries. **Figure 9** shows the example of RUBTAK01, one of many entry versions of UiO-66 (Universitetet i Oslo) in the CSD. RUBTAK01 is part of the *Non-disordered MOF subset* and has siteless hydrogens. To obtain the coordinates of these siteless hydrogens, users are recommended to apply the `add_hydrogen` function available in the API. The added hydrogen atoms will then appear in the CIF.

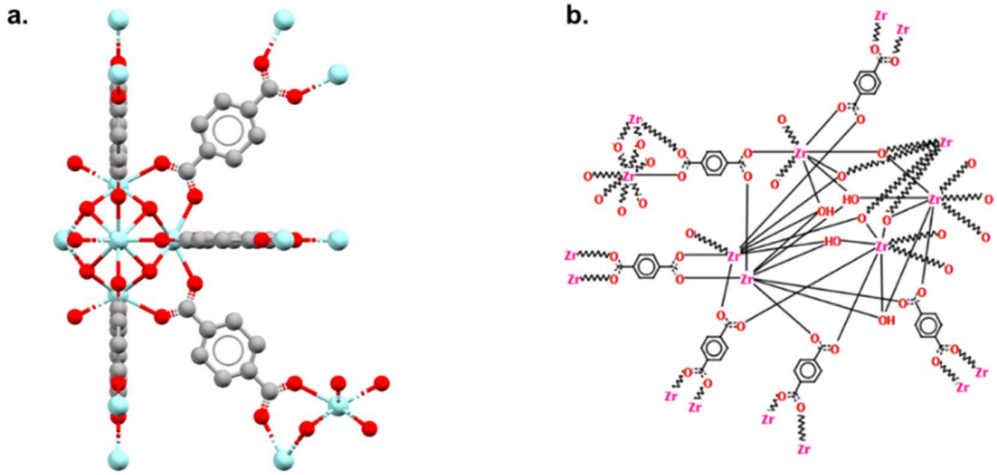


Figure 9 Example of a structure from the *Non-disordered MOF* subset with siteless hydrogen atoms: UiO-66 (CSD refcode: RUBTAK01). **a.** 3D visualisation of a repeating unit in Mercury. Oxygen atoms are represented in red, carbon atoms in grey and copper atoms in light blue. Hydrogen atoms are normally represented in white, and should be present on the aromatic rings, but are missing and not represented. **b.** 2D (hydrogen-depleted) diagram available in Mercury.

2.6 Case study: application to COFs

This case study is entirely the result of my own work.

2.6.1 Introduction

Covalent organic frameworks (COFs) are the organic equivalents to MOFs: they are crystalline structures assembled from – covalently bonded – organic building blocks. Similarly to MOFs, their reticular nature endows them with high tunability and a large design space for the community's imagination. However, COFs were only first conceptualised in 2005 by Yaghi and co-workers,[117] and only a few structures can be found in the CSD, as this case study will show. This is because it is experimentally more difficult to resolve COFs directly, as most of them present poor long range crystallinity necessary for the techniques required by the CCDC (X-ray, neutron or electron diffraction studies and powder studies using constrained refinement).[118]

There are two known attempts at building COF databases. Tong et al. developed CoRE COF (Computation-ready, Experimental COF) in 2017 as an equivalent to CoRE MOF, although the data gathering method is different.[119] While CoRE MOF derived its data from the CSD, the CoRE COF authors collected experimental studies published in the

literature and computationally reconstructed the CIFs from the experimental information available. The 187 1D, 2D and 3D obtained structures are thus solvent- and disorder-free. The later updates saw this number increase to 280 in 2018[120] and 309 in 2019.[121] In the latter year, Ongari et al. built their own Clean, Uniform, and Refined with Automatic Tracking from Experimental Database of COFs (CURATED COF database) from the 2018 version of CoRE COF and additional structures parsed from the textual literature to obtain 324 structures. The main differences with CoRE COF are: i) the building of the CURATED COF is entirely tracked in the AiiDA (Automated Interactive Infrastructure and Database for Computational Science) workflow management platform, allowing users to understand where their data come from; ii) the unit cells are optimised using density functional theory (DFT) and iii) DFT-obtained partial charges are also available. Whichever database, it seems the current number of COFs published in the literature should be around 300.

2.6.2 Methods

As extended structures, COFs share the essential "polymeric" feature with MOFs, meaning all the guidance given in this chapter are relevant for the search of COFs in the CSD. To frame the search, the table of six typical COF linkages published by Diercks et al. was used.[122] These six linkages are: C-N, C=N, C=N(Ar), B-O, B=N and C=C bonds. ConQuest was used to draw the target linkages. The searches were performed in the 5.41 version of CSD (with updates up to November 2019). The filter "only organics" was applied on all searches. **Figure 10** shows the queries used for the four linkage types that returned structures (C=C and B=N did not return any structure). Each coloured box corresponds to a type of linkage. The reaction involving the two reactive parts of the organic building blocks is presented in skeletal form in each linkage box. For each general type of linkage, there exists sub-types, the list of which is given by Diercks et al. and presented in **Figure 11b**.[2] More variety of linkages exists, as reported by Huang et al., but these have not returned any results in the CSD.[123] The developed ConQuest queries look for these specific sub-types and are represented in the dotted boxes in their ConQuest form. Only the sub-types that returned structures from the CSD are shown. A single query was used for each sub-type, each query consisting of two parts: i) a linkage-descriptive part, with most bonds being of "any type of bond" (represented in ConQuest with dotted lines), ii) a polymeric part, represented by two "QA" atoms linked by a polymeric bond (represented in ConQuest with zigzag lines). Note that this latter part is the equivalent of the text search for "catena" in the

structures' name when building the CSD MOF subset, since "catena" refers to any polymeric structure containing metal atoms.[43] In this case, QA was restricted to any atom among B, C, N, O or H, which are the constitutive atoms of COFs.

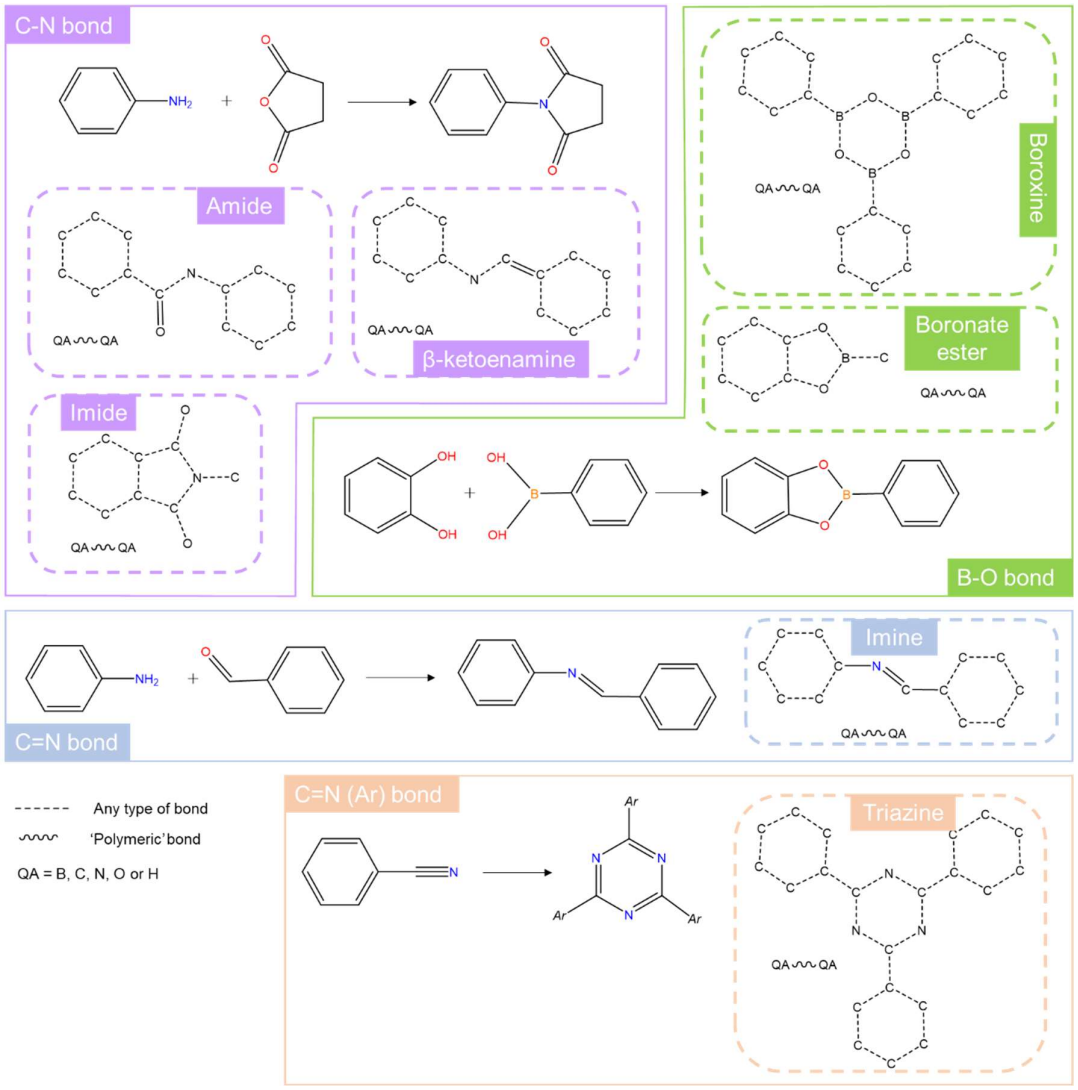


Figure 10 ConQuest queries used to look for specific COF linkages. QA = B, C, N, O, H. *Ar* denotes aromatic groups. The dotted line refers to any of the bond types stored in the CSD (single, double, triple, quadruple, aromatic, polymeric, delocalised, and pi). The zigzagging lines refer to polymeric bonds.

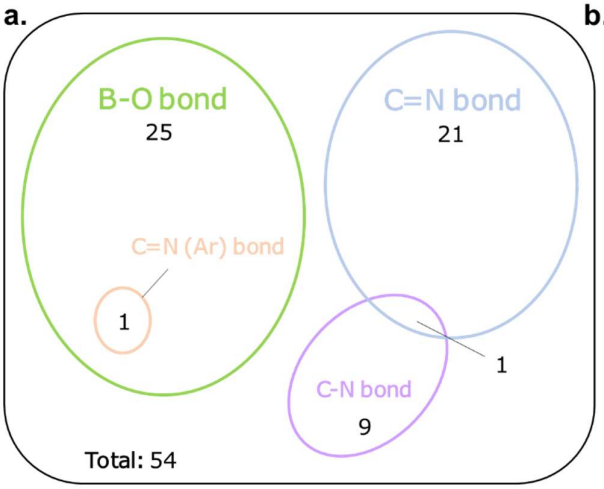
2.6.3 Results

A first quick search of any structure containing B, C, N, O or H atoms and at least a polymeric bond leads to a total of 204 hits, which constitutes the upper limit of potential COF structures in the CSD. **Figure 11b** presents the result obtained by looking at the classified linkages specifically and summarises the number of hits obtained for each query. As expected, the number of COFs structures found is staggeringly small: 54, that is $1/6^{th}$ of the number of structures reported in the literature. **Figure 11a** gives a visual representation of the COF dataset obtained. More structures might be present in the CSD, but given the experimental difficulty to obtain CIFs, the total number is likely to be in the same order of magnitude.

It is interesting to note several structural advantages of COFs over MOFs that make it relatively easier to mine COFs in the CSD:

- There are at least two reviews classifying the different linkages of COFs in the literature. The field of COFs is younger and smaller than that of MOFs, and the building blocks concept was mapped from the MOF field to COFs early on. Important MOF linkages have been reported, but the rapidly increasing number of structures means these reviews only captured a fraction of what can be considered as MOFs.
- The COF linkages looked for in this case study are rather small substructures involving two to three aromatic rings. Aromatic rings being usually represented in full, this gives little room for speculating over the location of the polymeric bonds. Therefore, it is sufficient to indicate the presence of a polymeric bond between any of the selected organic atoms. This is not the case for MOFs, where the polymeric bonds are most often assigned within the metal cluster. As previously explained in this chapter, there are different possible locations for the polymeric bonds to represent the same structure. Thus, the choice and location of the polymeric bonds directly affects the way the linkage of interest is represented in ConQuest, which is why designing queries for MOFs is trickier than for COFs.

a.



b.

COF linkage	Number of structures
Boronate	19
Boroxine	6
Borosilicate	0
Spiroborate	0
Imine	21
Hydrazone	0
Azine	0
Squaraine	0
Oxazole	0
Triazine	1 (part of boronate)
Phenazine	0
Alkene	0
Imide	2 (both part of amide, 1 part of imine)
Beta-ketoenamine	5
Amide	4 (1 part of imine)
Borazine	0

Figure 11 COFs dataset obtained from the CSD. **a.** Schematic representation of the overlaps of the different queries. **b.** Number of structures found with each query.

2.7 Conclusion

I reviewed in this chapter the most common CSD methods for mining MOFs. The manipulations presented here are enough for simulations where electrostatic interactions between the framework and the adsorbed molecules are neglected. I then illustrated the use of these methods for the extraction of COFs data in the CSD. The number of COFs found in the CSD is staggeringly low, as could be expected. These methods will be further used in Chapter 4 for the targeted search of specific types of MOFs. However, these methods cannot be easily directly applied to cages, as these are not extended structures. Chapter 6 will present a different approach to digging the CSD for cages, combining methods presented in this chapter with data analysis tools.

As described in this chapter, designing the best search query is not straightforward and there is currently no easy checks to determine how many structures can be missed; that is, how accurate a certain query is. The CCDC has already started assigning to some MOFs their common names (e.g. HKUST-1 from the Hong Kong University of Science and Technology, MOF-5, etc.). This information can be searched for as a string. I believe this effort will greatly facilitate the search for specific types of well-known MOFs. I also propose to flag any future deposition as “MOF” vs “non-MOF” and to double-check with the automatically updated CSD MOF subset, to ensure the accuracy of the criteria defined.

3 METHODS FOR MOLECULAR SIMULATIONS

The first two case studies presented in this chapter are published in:

- *Structural dynamics of a metal–organic framework induced by CO_2 migration in its non-uniform porous structure*, Pu Zhao, Hong Fang, Sanghamitra Mukhopadhyay, **Aurélia Li**, Svemir Rudic, Ian J. McPherson, Chiu C. Tang, David Fairen-Jimenez, S. C. Edman Tsang, Simon A. T. Redfern, Nat. Commun., 2019, 10, 999.
- *Biocompatible, crystalline, and amorphous bismuth-based metal-organic frameworks for drug delivery*, Claudia Orellana-Tavra, Milan Köppen, **Aurélia Li**, Norbert Stock, David Fairen-Jimenez, ACS Appl. Mater. Interfaces, 2020, 12, 5.

Abstract. This chapter introduces the various computational methods and tools used for the calculation of structural and adsorption properties of the nanoporous materials of interest in this thesis. Three case studies – including two that are now published – are presented as applications of these tools. These examples are the result of experimental-computational collaborative work, where I, unless otherwise specified, performed the computational part. The methods presented here are further on used in Chapters 4, 5 and 6.

3.1 Introduction

The previous chapter introduced the methods used in this work for mining the Cambridge Structural Database (CSD). Once a satisfying subset of structures is obtained, it is possible to further analyse the data with computational methods. The computational toolbox available to scientists today is extremely large and spans ranges of length and time scales: from quantum approaches focused on studying the electrons in the atoms to process simulations where a bulk of materials is considered for industrial scale applications.[57] In the context of the adsorption applications presented in this thesis, I aim to determine i) the structures' adsorption performance using classical simulations – where the focus is on the atoms or groups of atoms, and ii) their geometrical – mesoscale – characterisation. In particular, Monte Carlo (MC) methods will be used for the first goal – and for the second goal to some extent. The MC simulations presented here use statistical thermodynamics methods to obtain macroscopic properties of a system by averaging the underlying microstates' properties. Note that although these methods can be applied to a system in non-equilibrium, all the systems in this thesis are in equilibrium. This chapter first introduces the methods for the MC simulations and geometrical characterisations, before applying them to three case studies. The methods presented in Chapters 2 and 3 will then be used together in Chapters 4, 5 and 6.

3.2 Monte Carlo simulations

MC methods are a class of statistics algorithms based on the use of repeated random sampling to obtain numerical results.[124] In statistical physics in particular, they are useful in the numerical estimation of multivariate integrals that are otherwise too difficult to evaluate exactly. A typical example is the Hamiltonian of a given system that follows the Boltzmann statistics at a certain temperature. In this case, the mean value of a macroscopic property A corresponds to the integral sampled over all the system's microstates (MS):

$$< A > = \int_{MS} A_r \frac{e^{-\beta E_r}}{Z} dr \qquad \textbf{3-1}$$

where $E_r = E(r)$ is the energy of the system in the state r; $\beta = \frac{1}{k_B T}$, where k_B is the Boltzmann constant; and Z is the partition function of the system:

$$Z = \int_{MS} e^{-\beta E_r} dr \qquad \textbf{3-2}$$

A first estimation of this integral can be obtained with the MC integration, where the mean value of A is approximated as:

$$< A > \cong \frac{1}{N}\sum_{i=1}^{N}\frac{A_{r_i}e^{-\beta E_{r_i}}}{Z} \qquad \textbf{3-3}$$

with N the number of sampled points and the states r_i obtained uniformly across all the microstates. This sum can be further simplified with the importance sampling technique, where only the most significant states are sampled – these states being the ones with values of $e^{-\beta E_{r_i}}$ that are sufficiently high compared to the rest of the states. In our case, the most important microstates are those that maximise the overall Boltzmann distribution, which is, therefore, the chosen probability distribution for the importance sampling. To sample from this distribution, a class of algorithms called Markov chain Monte Carlo (MCMC) methods is used. The Metropolis scheme[125] in particular is widely adopted for its simplicity and generalisability.

Starting from a current (or *old*) state denoted (o) from the Boltzmann distribution, the MCMC method generates random trial moves to move the system to a new state denoted (n), which can be accepted or rejected. Let $P_B(o)$ and $P_B(n)$ be the probability of finding the system in state (o) and state (n) respectively, and $\alpha(o \to n)$ and $\alpha(n \to o)$ the conditional probability of performing the trial move between the two states. The Metropolis scheme assumes that a system composed of an arbitrary initial distribution of microstates eventually reaches equilibrium, in which case detailed balance must be satisfied: the probability of leaving state (o) to (n) by accepting the trial move o$\to n$, $P_{acc}(o \to n)$, is equal to the probability of leaving all other states (n) to (o) with the corresponding trial move $n \to o$, $P_{acc}(n \to o)$, such that:

$$P_B(o)\alpha(o \to n)P_{acc}(o \to n) = P_B(n)\alpha(n \to o)P_{acc}(n \to o) \qquad \textbf{3-4}$$

In addition, the Metropolis scheme assumes α to be a symmetric matrix, such that $\alpha(o \to n) = \alpha(n \to o)$ and the move acceptance is:

$$P_{acc}(o \to n) = \min\left(1, \frac{P_B(n)}{P_B(o)}\right) = \min(1, e^{-\beta \Delta E}) \qquad \textbf{3-5}$$

where $\Delta E = E_{(n)} - E_{(o)}$ is the energy difference between state (n) and state (o). Therefore, the move is always accepted if the energy of (n) is lower than that of (o). Otherwise, the move is accepted or rejected with probability $e^{-\beta \Delta E}$.

The types of moves used throughout this thesis include:[124]

- **Insertion:** a molecule is inserted at a random position,
- **Deletion:** a random molecule is chosen and removed from the system,
- **Rotation:** a random rotation is performed on a randomly selected molecule,
- **Translation:** a random displacement is applied to a randomly selected molecule,
- **Identity change:** in case of molecules of different nature, one type of molecule is chosen randomly and has its identity swapped with another type, which is grown where the previous molecule was.

Depending on the properties of interest to be computed, the MC methods should be applied in different ensembles.[126] In this thesis, two ensembles will be used:

- **Canonical ensemble (NVT):** the number of molecules (N), the volume (V) and the temperature (T) of the system are kept fixed. This ensemble will be used later on to determine the occupation density of molecules in the framework.
- **Grand canonical ensemble (μVT):** the chemical potential (μ), the volume (V) and the temperature (T) are kept fixed. Grand canonical Monte Carlo (GCMC) simulations are used to obtain adsorption isotherms.

3.2.1 Force fields

To accurately describe the interactions in a system of interest with classical simulations, we need a set of functions and parameters called force fields. These define the types of atoms present in the system based on their atomic number and their environment, as well as their interactions with other atoms. The functions enable the calculation of the total energy of the system U_{total} :

$$U_{total} = U_{bonded} + U_{non-bonded} \qquad \textbf{3-6}$$

The bonded atoms interact through their bonds, bends and torsions with their neighboring atoms (**Figure 12**):

$$U_{bonded} = U_{bonds} + U_{bends} + U_{torsions} \quad \textbf{3-7}$$

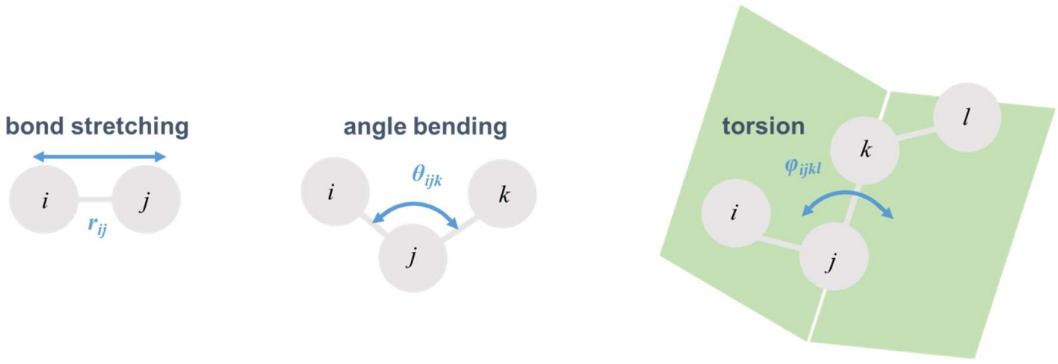


Figure 12 Illustration of the bond stretching, angle bending and torsions of bonded atoms.

U_{bonds} and U_{bends} are usually defined with a harmonic potential with respect to the interatomic distance and the bend angle, respectively. Given two atoms i and j at a distance r_{ij}, their bond energy is in the form of:

$$U_{bonds}(r_{ij}) = \frac{1}{2}k(r_{ij} - r_{eq})^2 \quad \textbf{3-8}$$

where k is the force constant and r_{eq} the bond length at equilibrium.

Given three atoms i, j and k with $i-j$ and $j-k$ bonds and a bend angle of θ_{ijk}, the bend energy is in the form of:

$$U_{bends}(\theta_{ijk}) = \frac{1}{2}k'(\theta_{ijk} - \theta_{eq})^2 \quad \textbf{3-9}$$

where k' is the force constant and θ_{eq} the bend angle at equilibrium.

Torsions are defined for four consecutively bonded atoms (i, j, k, l) with respect to their dihedral φ_{ijkl}. As the form of the torsion energy varies depending on the implementation, it will be provided when necesssary.

The energy for non-bonded atoms include interactions of atoms from different molecules and atoms from a single molecule separated by more than three bonds. It is the sum of the energy resulting from the van der Waals interactions and from the electrostatic interactions:

$$U_{non-bonded} = U_{VdW} + U_e \quad \textbf{3-10}$$

The van der Waals term corresponds to the short-range interactions described by the Lennard-Jones (LJ) 12-6 potential (**Figure 13**):

$$U_{VdW}(r_{ij}) = U_{LJ}(r_{ij}) = 4\varepsilon[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}] \quad \textbf{3-11}$$

Where r_{ij} is the distance between two atoms i and j, σ is the interatomic distance such that the attractive and repulsive forces are balanced out, and ε is the depth of the minimum energy well. These two parameters are defined for two atoms of the same type. For atoms of different types α and β, the Lorentz-Berthelot rules are used to obtain the cross terms:

$$\sigma_{\alpha\beta} = \frac{\sigma_{\alpha\alpha} + \sigma_{\alpha\beta}}{2} \quad \textbf{3-12}$$

$$\varepsilon_{\alpha\beta} = \sqrt{\varepsilon_{\alpha\alpha}\varepsilon_{\alpha\beta}} \quad \textbf{3-13}$$

Note that the Lorentz-Berthelot rules are analytically only correct for hard spheres. In this thesis, I assume all atoms to be hard spheres and apply this set of rules throughout.

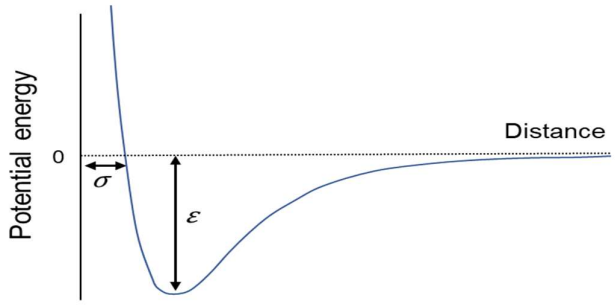


Figure 13 Shape of the Lennard-Jones potential energy between two atoms.

The electrostatic term corresponds to the long-range interactions in the system and is obtained with Coulomb's law:

$$U_e(r_{ij}) = U_{Coulomb}(r_{ij}) = \frac{q_i q_j}{4\pi\varepsilon_o\varepsilon_r r_{ij}} \quad \textbf{3-14}$$

where r_{ij} is the distance between atoms i and j, q_i and q_j are their partial charges, ε_0 is the electric constant of the vacuum and ε_r the relative dielectric constant of the medium.

The parameters necessary to compute these functions are obtained either experimentally or theoretically through quantum calculations. Many different sets of functions and parameters – i.e. force fields – exist. The most common ones in the metal-organic framework (MOF) simulation literature are the Universal Force Field (UFF)[127] and the DREIDING force field (DFF), both widely used for their universality and transferability.[128] Other force fields are more specifically tailored to a certain type of molecules: Assisted Model Building with Energy Refinement (AMBER) and its generalised version GAFF (Generalized AMBER Force Field)[129] are more suited for proteins and functional groups of amino-acids, Optimized Potentials for Liquid Simulations (OPLS)[130] is a better fit for proteins and organic molecules, and the Transferable Potentials for Phase Equilibria (TraPPE)[131] force field is optimised for the calculation of vapour-liquid coexistence curves for small organic molecules.

3.2.2 Modelling the framework and periodic boundary conditions

Throughout this thesis, the nanoporous materials are considered rigid, unless otherwise stated. The atoms' positions are obtained experimentally and modelled with the UFF and the DFF. This is enough when modelling the adsorption of apolar compounds in MOFs. However, when modelling polar compounds, it becomes necessary to assign partial charges to the framework. Several methods exist,[132] with varying accuracy and computational cost. In this work, the extended charge equilibration (EQeq)[133] was used when necessary, as it quickly provides satisfying point charges.[72, 132]

As we are only interested in the bulk properties of the framework, we need periodic boundaries on our simulation box to make sure we are infinitely far from any interface. To compute the total energy of the system, we need to sum over all the contributions, which include the very long-range non-bonded interaction terms. For the final electrostatic energy, the Ewald summation is used.[134] This summation method has become the standard method for computing long-range electrostatic interations in periodic systems, as it converges rapidly and ensures high accuracy.[135] As for the LJ terms, the contribution of pairs with high values of r are neglected as it is close to zero. A cut-off radius r_C is used for this purpose to truncate the additional LJ terms. The choice of r_C has then a direct consequence on the simulation box lengths, as these should be at least $2r_C$ to satisfy the nearest image convention, as illustrated in **Figure**

14.[124] Typically, a cut-off value is chosen such that it is a multiple of the largest LJ σ parameter in the simulation.[136] The lowest recommended value is 2.5σ. The higher the value, the more computational time is needed. Therefore, a compromise between computational accuracy and efficiency has to be made when choosing r_C. Throughout this thesis, a cut-off value of 12.8 Å is used for the LJ terms. This value has become a standard choice following the work of Düren et al. on methane storage in isoreticular MOFs (IRMOFs), for which 12.8 Å corresponds to 3.4σ and is just under half the length of the unit cell of IRMOF-1.[137]

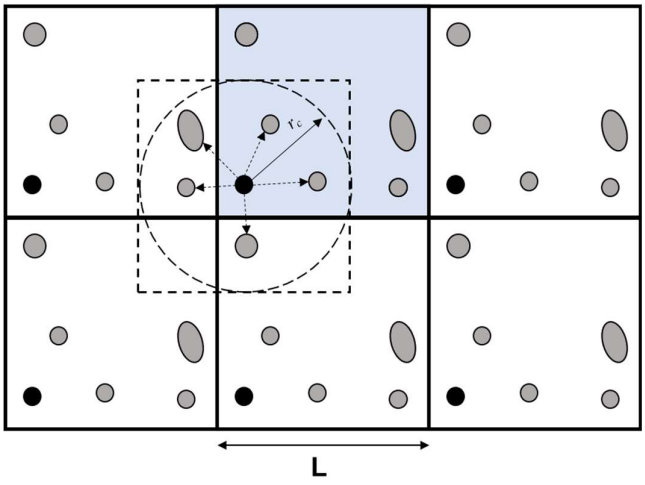


Figure 14 Schematic representation of the periodic boundary conditions. The simulation box is coloured in blue. It is surrounded by periodically repeated images of itself. For simplicity, the box is considered a square of length L. Particles are represented with rounded shapes. The black dot represents a particle of interest. The dotted arrows indicate the interactions with the neighbouring atoms, bound within a circle of radius r_c, i.e. the cut-off value. For easier comparison, a dotted box of length L surrounds the circle.

3.2.3 Modelling the adsorbates

The hydrogen used in the high-throughput screening (HTS) of Chapter 5 is modelled with TraPPE. In case studies 2 and 3 of this chapter, however, I present the modelling of unusual and longer molecules using OPLS, and more specifically the all-atom version of the force field, OPLS-AA. An all-atom force field takes into account the interactions of all the atoms of the molecules, as opposed to a united-atom force field where several atoms can be considered as a group. The choice of an all-atom versus united-atom force field depends on the desired application. A united-atom force field is usually a good

approximation in cases where the intermolecular interactions are much more important than the intramolecular interactions; it is a fast and accurate option for the simulation of methyl groups for instance.

The modelling of these adsorbates considers parts of the molecule as rigid and others as flexible. The rigid parts – in this thesis – correspond to aromatic rings and are defined by the relative coordinates of the atoms. The rigid parts are inserted or deleted directly in the simulation box. The flexible parts, however, have their bonds, bends and torsions parameters defined and are built atom by atom in the simulation box using configurational-bias Monte Carlo (CBMC).[138, 139] Indeed, the conventional MC methods are not efficient when the molecules to be inserted are bulky, as the odds for inserting large molecules are usually low, causing the overall efficiency of the algorithm to be low as well. CBMC improves the conformational sampling of a long molecule by growing the entire molecule bead by bead. In particular, it generates for each bead k trial orientations depending on the internal energy of the molecule. The most favorable orientation is then chosen depending on the external energy of each trial for a given bead, thus biasing the growth of the molecule. For each growth step at a given bead, the Rosenbluth factor is defined as the ratio of available growth directions in this step over the total number of growth directions.[66] When the entire molecule is grown, its acceptance or rejection is based on its Rosenbluth factor, which is the product of all the Rosenbluth weights accumulated throughout the process.

Other methods exist to further improve the insertion of molecules, such as the continuous fractional component Monte Carlo,[140] but these will not be presented in this thesis.

3.2.4 RASPA

To carry out the aforementioned simulations, the open-source multipurpose state-of-the-art package RASPA was used.[141] This software includes all the algorithms necessary for MC and molecular dynamics simulations of adsorption in nanoporous materials. After having defined a given force field and models of the adsorbent and adsorbates, users can work in different ensembles and define summation methods, cut-off values and types of moves, among other parameters. RASPA also outputs snapshots – positions of the atoms or groups of atoms – of the simulated adsorbates in the adsorbent. Unless otherwise stated, the snapshots presented in this work were obtained with RASPA and further processed with Materials Studio.[142]

3.3 Geometric characterisation

I introduced in Chapter 1 a few geometrical concepts relevant to the studied extended materials: largest cavity diameter (LCD), pore limiting diameter (PLD), surface area, pore volume and percolation. Several open-source programs allow the computation of these properties. In this work, Zeo++[60] was used for the characterisation of the CSD MOF subset in Chapter 4, and PoreBlazer[143] was used for the calculation of channel dimensionalities in Chapter 4 and HTS of the same subset for hydrogen storage in Chapter 5. The change of software was due to a homogenisation of the tools used in our research group (Adsorption and Advanced Materials group, Department of Chemical Engineering and Biotechnology, University of Cambridge, UK). As demonstrated by Sarkisov et al., however, the results obtained from Zeo++ and PoreBlazer are consistent to a large extent.[109] The discrepancies occur with properties based on network-accessibility, and especially in low-porosity structures. I herein describe the algorithms behind the computation of these values in Zeo++ and PoreBlazer.

3.3.1 LCD, PLD and percolation

Zeo++ is based on the computation of the Voronoi network – or tessellation – of a 3D structure. In this case, the Voronoi network is the partition of space into regions, or Voronoi cells, such that all the points in a cell are closer to a given atom than any other atoms. **Figure 15d** shows an example of Voronoi network obtained for a zeolite. The spheres represent the atoms of the structure, the unit cell is represented in blue and the network in white. The Voronoi network thus corresponds to the void space of the material. From the obtained tessellation, it is possible to obtain the largest included sphere (or LCD) and the largest free sphere (or PLD). Here, the largest included sphere corresponds to the largest distance between a node from the network and a neighbouring atom. The largest free sphere is the largest included sphere that can travel within the void space. The largest included sphere is obtained by iterating over all the Voronoi nodes in a periodic unit cell to find the node with the largest distance to a neighbouring atom. Then, the travel path with the largest opening is obtained with the Dijkstra's optimisation algorithm, thus computing largest free spheres.[144] The final PLD is obtained by retaining only the largest of the largest free spheres.

The percolation is obtained by first assessing the accessibility of the nodes. If a porous channel exists in a given structure, it should be periodic and run through an infinite number of cells. Therefore a node that is connected to its image constitutes a channel, so

is a node connected to a channel. Following this identification scheme, the nodes of a Voronoi network are classified as part of a channel or of an isolated pocket. During this process, the directions of the channels are saved, from which the dimensionality of the channels – or percolation – can be obtained.

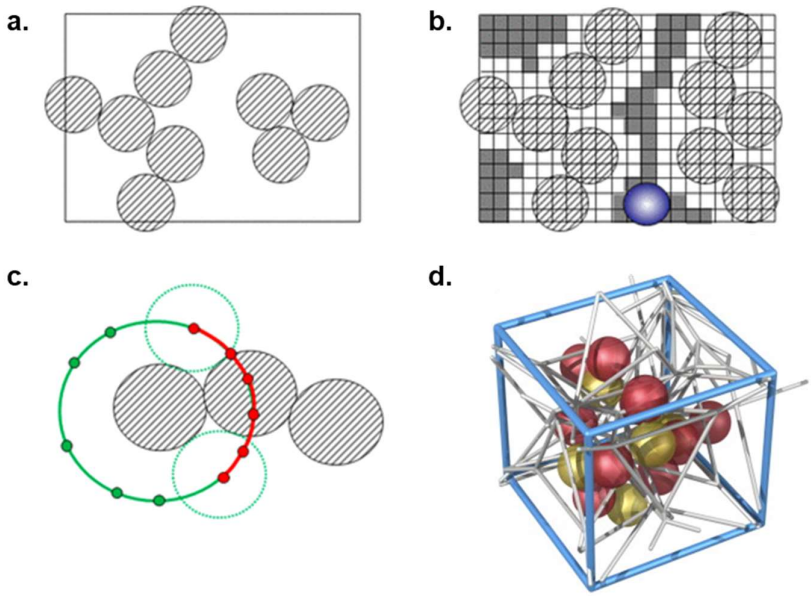


Figure 15 Calculation of the largest cavity diameter (LCD) and pore-limiting diameter (PLD) in PoreBlazer and in Zeo++: **a.** example of a system with modelled atoms (striped grey circles), **b.** a lattice of cubelets is formed on system **a.** (the striped circles remain the atoms, the dark grey cubes correspond to a percolated path for the blue spherical probe, the representation is not up to scale), **c.** illustration of the MC algorithm behind the calculation of surface area, **d.** example of Voronoi network obtained for a zeolite (the red and yellow spheres represent oxygen and silicon atoms respectively, the blue frame corresponds to the limits of the unit cell and the white structure to the Voronoi network).[60, 109]

PoreBlazer uses a drastically different approach, which divides the system (**Figure 15a**) into a lattice of cubelets (**Figure 15b**). Using the Hoshen-Kopelman algorithm for finding and labelling clusters, it is then possible to determine the percolation with various probe sizes.[145] The dark grey cubelets in **Figure 15b** correspond to those that do not overlap with the atoms (striped in grey) and thus to a cluster of lattices forming a path accessible to the blue probe. The PLD corresponds to the largest probe for which such a percolated path exists. In the process, PoreBlazer keeps track of the dimensions of the obtained percolation. The LCD is obtained by retrieving the largest distance

amongst the stored distances between the centres of the cubelets and the surfaces of the atoms.

3.3.2 Surface area

For the calculation of accessible surface area, Zeo++ and PoreBlazer both use MC integration methods, but on different systems. **Figure 15c** illustrated the MC algorithm applied in PoreBlazer.[109] The smaller, green dashed circles represent the probe. The green circle corresponds to the accessible surface area and is defined as cocentric to the given atom and of radius $k(r_i + r_p)$, where r_i is the radius of atom i and r_p that of the probe. In PoreBlazer, k is a constant chosen such that this radius corresponds to the location of the LJ well, and is equal to 1.22. In Zeo++, it corresponds to the distance at which the LJ potential is equal to 0 and $k = 1$. For each atom, the MC algorithm counts the proportions of sampled points f falling on this surface area, without overlapping with the atoms (striped grey atoms). The surface area accessible to the given probe on atom i is thus given by $a = f(4\pi r^2)$. Sampled points that collide with atoms are in the red arc of **Figure 15c**. In PoreBlazer's lattice system, identifying accessible surface area only requires a check-up of the cubelets labels. Zeo++ uses an additional step based on the convexity of the Voronoi cells to assess the viability of a sampled point.[60] I refer later on in this thesis to gravimetric and volumetric surface areas – these are obtained by dividing the previous surface area by the material's mass and volume, respectively.

3.3.3 Pore volume

In Zeo++, the accessible – probe centered – pore volume defined in Chapter 1 is obtained using the same method as for accessible surface area, by extending the sampling area to the entire unit cell. The distance between each sampled point and any other atomic surface needs to be larger than the radius of the probe. In PoreBlazer, the accessible pore volume is enclosed in the accessible surface area and corresponds to the volume of the dark grey cubelets in **Figure 15b**. In this thesis, the notion of void fraction will be used instead of the pore volumes. The void fraction corresponds to the ratio of pore volume over the framework volume.

3.3.4 Pore size distribution

The pore size distributions (PSD) presented in this work are obtained with RASPA,[141] in which the PSD is obtained following the method developed by Gelb and Gubbins.[146] To compute the PSD, the volume $V_{pore}(r)$ of void space coverable by spheres of radius r or lower is first calculated. A point a is in $V_{pore}(r)$ if there exists a sphere of radius r

that contains a without overlapping any framework atoms. The PSD is defined as the opposite of the derivative of this volume $-\frac{dV_{pore}(r)}{dr}$, which is the fraction of volume coverable by spheres of radius r but not $r + dr$.

3.4 Case studies

The case studies presented here are the result of collaborative work with experimental researchers: case study 1 presents the use of GCMC to elucidate the flexibility of ZIF-7. Case study 2 shows the illustrative power of GCMC to visualise where drug molecules adsorb in MOF pores. Case study 3 looks at the possibility of using GCMC to predict the chiral selectivity of a set of MOFs. The contribution of my collaborators are specified at the start of each case study.

3.4.1 Case study 1: using GCMC to understand the CO_2 migration-induced flexibility in ZIF-7

This case study is the result of a collaborative work led by Dr Pu Zhao, Department of Chemistry, University of Oxford, UK. The synthesis and characterisation of ZIF-7 (zeolitic-imidazolate framework) and the experimental isotherms of CO_2 in ZIF-7 referred to in this section were performed by Dr Pu Zhao. The computational contribution is entirely my work.

3.4.1.1 Introduction

This case study looks at ZIF-7, a flexible MOF composed of zinc atoms connected with benzimidazolate linkers. Conversely to most other flexible structures studied, ZIF-7 does not have uniform pores, but four different types: two six-ring windows (called window A and window B, window A being larger than window B), one four-ring window and the sodalite cavity. **Figure 16a** shows the location and shape of these windows and cavities. **Figure 16b** shows the experimental stepped CO_2 isotherms obtained at 195 K and 298 K. In this collaborative work, it is shown experimentally that the structure of ZIF-7 transitions between two phases (ZIF-7-I at high pressures of CO_2 and ZIF-7-II at low pressures of CO_2 as shown in **Figure 16**) during CO_2 adsorption, and that this flexibility is caused by the migration of CO_2 molecules between windows A and B.[147] More specifically, it is shown that CO_2 first binds weakly to window B when window A is inaccessible, until the accumulation of CO_2 triggers the linker between windows A and B to rotate, opening up access to window A and thus increasing the uptake of CO_2 at higher pressures. The goal of this case study is to corroborate the experimental finding by using GCMC simulations on the ZIF-7-I and

ZIF-7-II phases. I, therefore, simulated the adsorption isotherms of CO_2 at 195 and 298 K and N_2 at 77 K.

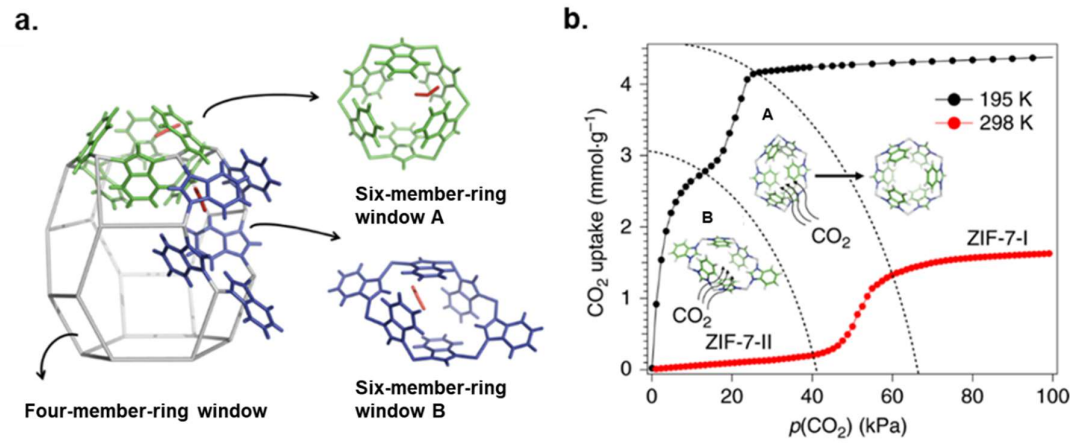


Figure 16 ZIF-7 structure and its CO_2 adsorption behaviour. **a.** The building unit, sodalite cage, of ZIF-7 with two types of six-member-ring windows (A, B) and one type of four-member-ring window on its walls. Part of the framework is simplified by replacing Zn−bIm−Zn with Zn−Zn (bIm = benzimidazolate, $C_7N_2H_5$, Zn: grey). The symmetry of CO_2 has been disregarded for clarity. **b.** CO_2 adsorption isotherms of ZIF-7 at 195 and 298 K, p_{CO2} = 1–100 kPa, illustrated by the structural behaviours of ZIF-7.

3.4.1.2 Simulation setup

I used an atomistic model of of the ZIF-7-I and ZIF-7-II phases for which the framework atoms were kept fixed at the crystallographic positions. I used 2x2x2 cells in each case to account for the periodic boundary conditions. In order to describe correctly the absence of CO_2 molecules in the windows A of ZIF-7 before the ZIF-7-II >> ZIF-7-I phase transition, all the windows A were blocked in ZIF-7-II during the simulation. The parameters for the framework atoms were derived from the UFF[127] and previously developed for ZIF-8,[148] whereas CO_2 and N_2 were modelled using the TraPPE potential with charges placed on each atom and at the centre of mass (**Table 3**).[149] 0 molecules were added at the start of the simulations.

Table 3 Force field parameters for the UFF+, derived from the UFF and used to model the framework atoms of ZIF-7-I and ZIF-7-II and TraPPE parameters used to model CO_2 and N_2.

	UFF+	
	σ (Å)	**ε/k_B (K)**
C	3.431	31.270
N	3.261	20.549
H	2.571	13.103
Zn	2.462	36.928
	TraPPE	
	σ (Å)	**ε/k_B (K)**
O_CO_2	3.05	79.0
C_CO_2	2.80	27.0
N_N_2	3.31	36.0

Up to 50,000 MC cycles were performed, the first 50% of which were used for equilibration, and the remaining steps were used to calculate the ensemble averages. The Peng-Robinson equation of state was used to calculate the gas-phase fugacity.[150] Surface areas were calculated by rolling a 3.681 Å diameter sphere, which corresponds to N_2, across the surface of the material.

3.4.1.3 Results

In order to describe the flexibility of the adsorbents using rigid models in the GCMC, I performed the simulations on both ZIF-7-I and ZIF-7-II phases separately. A similar strategy was followed in the past on ZIF-8.[64] **Figure 17** shows the experimental and simulated CO_2 isotherms of ZIF-7; the solid symbols in the background correspond to the experimental isotherm. At 298 K, the simulated isotherms match the experimental data, with a slight overprediction at lower pressures for ZIF-7-II. When decreasing the

temperature to 195 K, the simulated isotherm for ZIF-7-II – with blocked windows A – matches the experimental isotherm very well. Interestingly, the uptake of ZIF-7-I after the phase transition is largely underpredicted by the GCMC simulations. This underprediction of ca. 1.2 mol kg^{-1} is translated to ca. 7 molecules per unit cell. **Figure 18** shows the snapshots and density distributions of the adsorption of CO_2 in ZIF-I and ZIF-II at 10 and 30 kPa, 195 K.

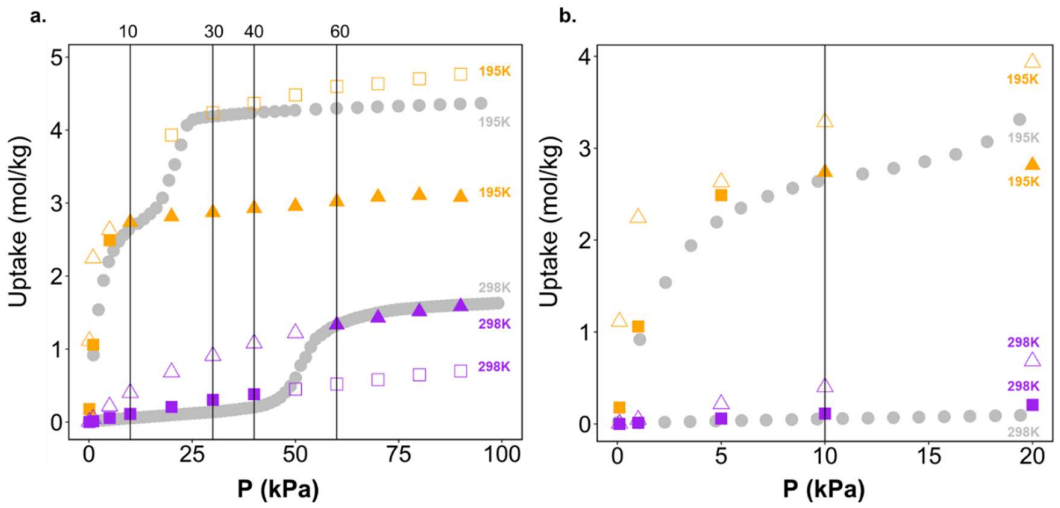


Figure 17 CO_2 adsorption isotherms for ZIF-7. **a.** Comparison of experimental (grey circles) and simulated adsorption isotherms, using ZIF-7-I (triangles) and ZIF-7-II (squares) phases at 195 and 298 K. **b.** Detail of the comparison at low pressure. Solid triangles and squares correspond to the regions of the isotherms that correspond to the phases observed in the experimental isotherm. The vertical lines highlight the range of pressures at which the transitions occur: 10 kPa – 30 kPa at 195 K and 40 kPa – 60 kPa at 298 K.

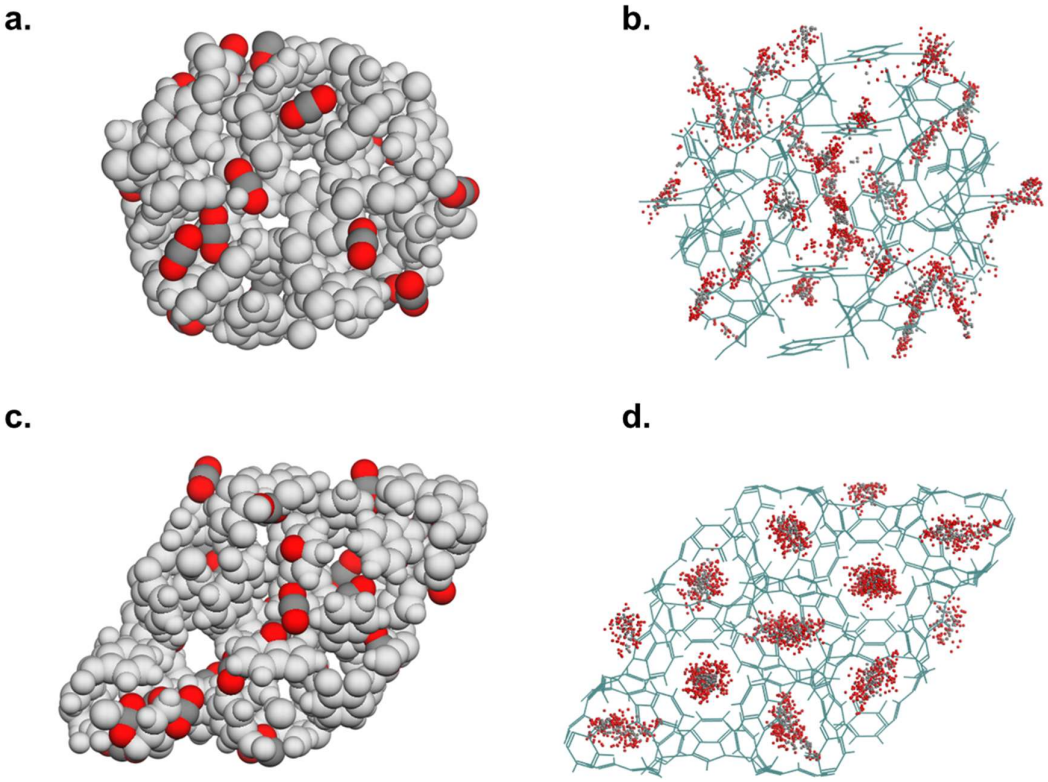


Figure 18 Snapshots and density distributions of CO_2 in ZIF-7. **a.** Snapshot and **b.** density distributions of CO_2 molecules in ZIF-7-II at 10 kPa at 195 K (18 molecules per unit cell). **c.** Snapshot and **d.** density distributions of CO_2 molecules in ZIF-7-I at 30 kPa at 195 K (16 molecules per unit cell).

In order to understand the underprediction observed by the GCMC simulation, I investigated the porosity of ZIF-7-I and ZIF-7-II. **Table 4** shows the values of the surface area and pore volumes obtained with N_2 as a probe; **Figure 19** shows the simulated adsorption isotherms and PSD for ZIF-7-I and ZIF-7-II – note that ZIF-7 is non-porous to N_2 at 77 K experimentally due to the kinetic effects induced by the low temperature, whereas a GCMC simulation will be able to insert N_2 molecules even in closed porosity. Although the experimental data from the adsorption isotherms of CO_2 showed a higher uptake, suggesting that ZIF-7-I had a larger pore volume and therefore larger surface area, the simulations found that ZIF-7-II has higher pore volumes and surface area than ZIF-7-I. This is also confirmed by the GCMC simulated isotherm of N_2. These discrepancies between experiments and simulations could be potentially related to inaccuracies in the force field employed during the simulation and the existence of very narrow porosity at ca. 4.5 Å of diameter.

Table 4 Pore volume and surface area of ZIF-7.

	ZIF-7-I	ZIF-7-II
Pore Volume ($cm^3.g^{-1}$)	0.69	0.76
Surface Area ($cm^2.g^{-1}$)	349	605

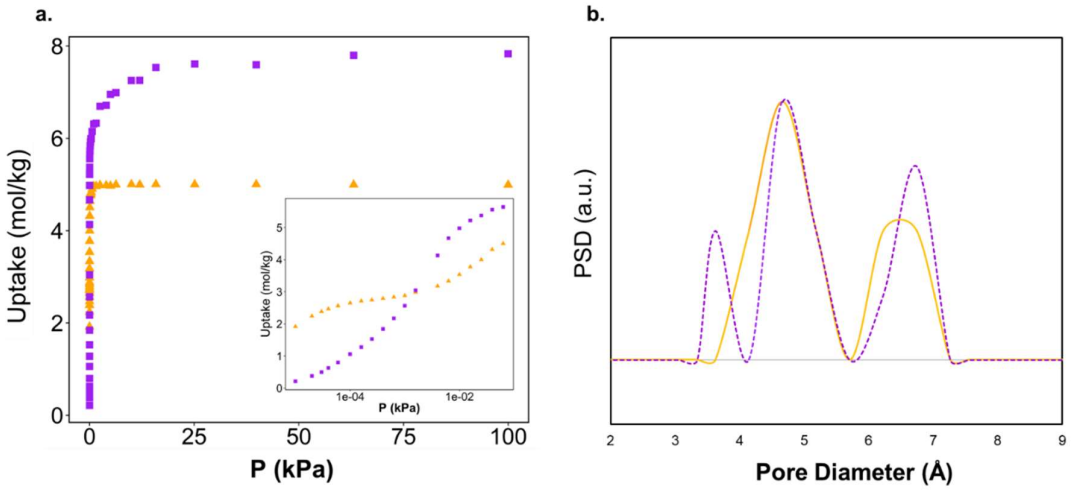


Figure 19 a. Simulated adsorption isotherms for N_2 at 77 K and **b.** PSD of ZIF-7. The data corresponding to ZIF-7-I are represented with orange triangles and the orange solid line. Those corresponding to ZIF-7-II are represented with purple squares and the purple dashed line. The inset in **a.** represents the semi-logarithmic isotherm at low pressures.

Finally, **Figure 20** shows the heat of adsorption (Q_{st}) as a function of the CO_2 uptake. The heat of adsorption of ZIF-7-I (i.e. Q_{st}-I) measured at 195 and 298 K starts at around 28.3 kJ mol^{-1}, and remains constant until 1.8 mol kg^{-1} of uptake when it starts decreasing abruptly. On the other hand, the Q_{st} of ZIF-7-II (Q_{st}-II) measured at both temperatures remains constant at around 24 kJ mol^{-1} until 1 mol kg^{-1} of uptake, when it starts increasing. Interestingly, Q_{st}-I and Q_{st}-II intersect at 26 kJ mol^{-1} and an uptake of 2.8 mol kg^{-1}, i.e. the loading where the phase transition takes place for low temperature. While I do not think this is a coincidence, I am not capable of giving a satisfying physical explanation to his phenomenon at this stage.

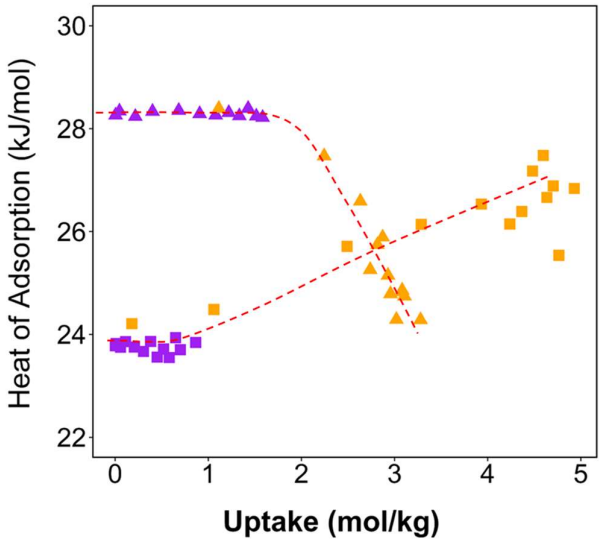


Figure 20 Heat of adsorption of CO_2 in ZIF-7. Triangles: ZIF-7-I, squares: ZIF-7-II, orange: 195 K, purple: 298 K. Red dotted lines are included as eye-guides.

3.4.1.4 Conclusion

The computational results in this case study confirm the initial uptake of CO_2 molecules in windows B in ZIF-7-II. Although the underprediction of CO_2 uptake in ZIF-7-1 at 195 K is still not well understood, it matches the results from previous simulations, where it was suggested that minor linker arrangements might enable more efficient packing of the molecules under these conditions, thus explaining the higher experimental uptake.[151] The fact that only the computed isotherms at 195 K do not match with the experimental results also suggests that the temperature might play a role in this structural rearrangement. This case study showed that it is possible to identify the flexibility of a material by assuming its phases as rigid. However, the observed discrepancies suggest that the method could be improved: the force field could be tuned more specifically to reproduce the isotherms more accurately, or a force field suitable for flexible structures could be used instead.

3.4.2 Case study 2: assessing the uptakes of α-CHC and DCA in CAU-7

This case study is the result of a collaborative work led by Dr Claudia Orellana-Tavra, then in the Adsorption and Advanced Materials group, Department of Chemical Engineering and Biotechnology, University of Cambridge, UK. Dr Milan Köppen, Institut für Anorganische Chemie, Christian-Albrechts-University, Germany, synthesised CAU-7. Dr Claudia Orellana-Tavra performed the MOF characterisation and the drug encapsulation referred to later on in this section. The computational contribution is entirely my work.

3.4.2.1 Introduction

Conventional therapeutics that exist today for cancer treatment suffer from a number of drawbacks, including limited drug solubility, low selectivity, poor distribution and early degradation before reaching the target organ.[152-154] Often, high concentrations of a drug need to circulate first in the blood stream before arriving at the desired organ, by which time healthy tissues have already been damaged. Drug delivery systems (DDS) are a possible solution for a more controlled release by protecting the drug from degradation and targeting delivery to the specific organ in question.

In this collaborative work, we studied the use of the biocompatible MOF CAU-7 (Bi^{3+} ions connected with 1,3,5-benzenetrisbenzoate) as a DDS for two cancer drugs: sodium dichloroacetate (DCA) and α-cyano-4-hydroxycinnamic acid (α-CHC). These two drugs are known for their ability to modify cancer metabolic pathways. **Figure 22b** shows a snapshot of the structure and **Figure 21** shows the two drugs. The goal of this case study is to use MC simulations to compare the theoretical possible uptakes against the experimental loadings in CAU-7. It is also an example of how MC simulations can be used to study drug uptakes.

3.4.2.1 Simulation setup

I first performed GCMC simulations of DCA and α-CHC in CAU-7 at body temperature (310 K) to obtain snapshots of the saturated state of the drugs in CAU-7. I then performed MC simulations in the NVT ensemble to obtain snapshots of the drugs in CAU-7 at low and medium loadings. I used an atomistic model of CAU-7 for which the framework atoms were kept fixed at the crystallographic positions in 1x1x7 cells to account for the periodic boundary conditions. The parameters for the framework atoms were derived from the UFF[127] and the DFF,[151] whereas DCA and α-CHC were modelled

with GAFF4RASPA using the general AMBER potential (**Table 5**).[129] 0 molecules were added at the start of the GCMC, while three loadings of DCA (low: 121 mg g^{-1}, medium: 307 mg g^{-1} and saturation: 500 mg g^{-1}) and α-CHC (low: 42 mg g^{-1}, medium: 126 mg g^{-1} and saturation: 248 mg g^{-1}) were added and kept fixed in the NVT ensemble. Up to 500,000 MC cycles were performed, the first 50% of which were used for equilibration, and the remaining steps were used to calculate the ensemble averages.

The molecule definition files and the force field definition files for DCA and α-CHC were obtained using GAFF4RASPA, a set of tools which calculates the LJ parameters and molecule definitions based on the general AMBER force field. Both molecules were modelled as flexible, except for the ring in α-CHC which was kept rigid. The LJ parameters, bonds, bends and torsion definitions used in RASPA for these simulations are described in **Table A1**Error! Reference source not found. and **Table A2** available in Appendix A.

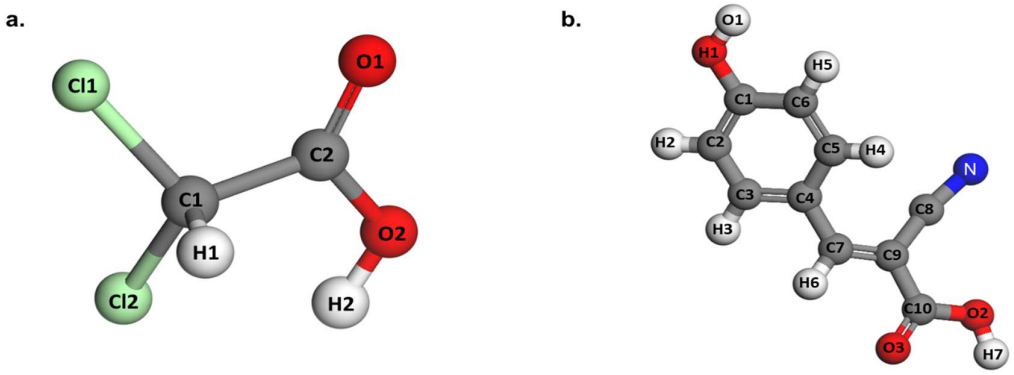


Figure 21 Indexing of the atoms in **a.** DCA and **b.** α-CHC.

3.4.2.2 Results

Figure 22a shows the PSD of CAU-7 and highlights the presence of one type of cavity of about 10 Å. The snapshot in **Figure 22b** gives a clearer view of this cavity.

Using MC, the maximum capacities were predicted to be 33.3 wt.% (10.1 mol of drug per mol of CAU-7) and 19.9 wt.% (13.7 mol of drug per mol of CAU-7) for DCA and α-CHC, respectively, whereas the experimental maximum loadings obtained were of 33.7 wt.% (9.8 mol of DCA per mol of CAU-7) and 9.3 wt.% (33.2 mol of α-CHC per mol of CAU-7). Looking at the sizes of both drugs (ca. 4.1 Å in length for DCA and 9.6 Å for α-CHC) and their uptakes, it is clear that DCA is able to fit in the pores, whereas α-CHC showed some issues getting loaded. However, the computed uptake of α-CHC

suggests that the experimental uptake could be improved. **Figure 23** shows snapshots of the adsorption process of these two drugs by CAU-7 at low, medium, and saturated loadings; **Figure 24** shows the density distributions during the adsorption process, highlighting the areas where the molecules get adsorbed. DCA and α-CHC are first adsorbed on the walls of the MOF structure at low loadings, before filling up the whole cavity at higher loadings.

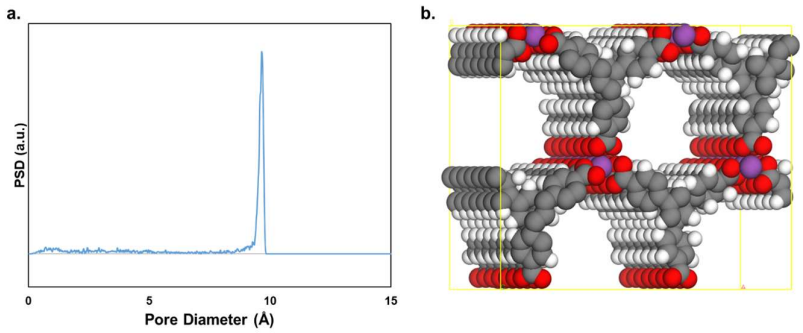


Figure 22 a. PSD of CAU-7, **b.** Snapshot of a supercell of empty CAU-7.

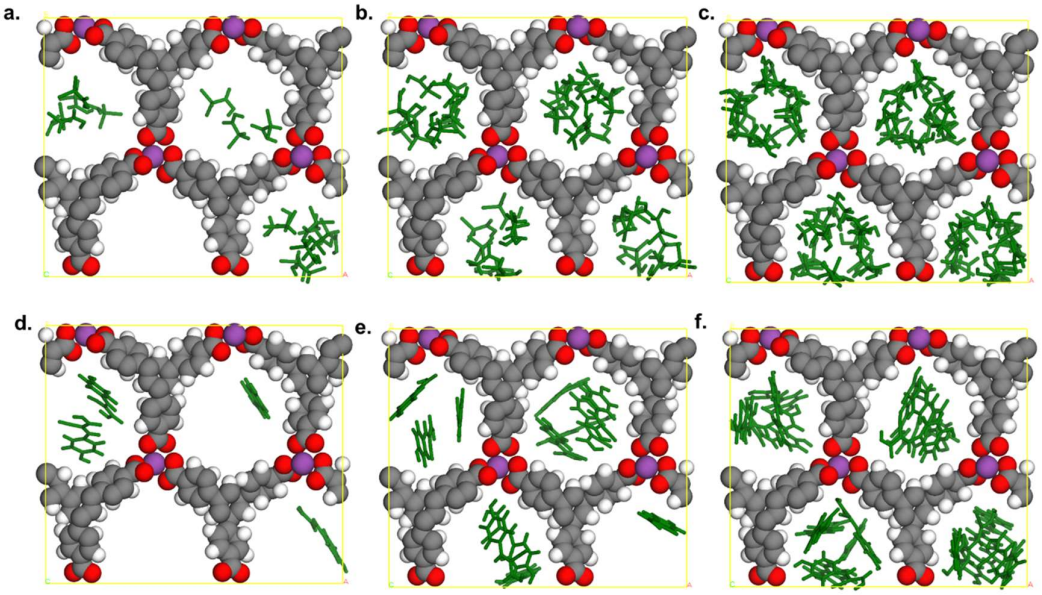


Figure 23 Snapshots of DCA (**a.**–**c.**) and α-CHC (**d.**–**f.**) in CAU-7 at different loadings: **a.** 121 mg g^{-1} or 10.8 wt.%, **b.** 307 mg g^{-1} or 23.5 wt.%, **c.** saturation with 500 mg g^{-1} or 33.3 wt.%, **d.** 42 mg g^{-1} or 4.0 wt.%, **e.** 126 mg g^{-1} or 11.2 wt.%, and **f.** saturation with 248 mg g^{-1} or 19.9 wt.%. The drug molecules are represented in the green-stick mode. The experimental loading of α-CHC in CAU-7 is 9.3 wt.%, which correspond to a situation between **Figures 23a** and **23b.**

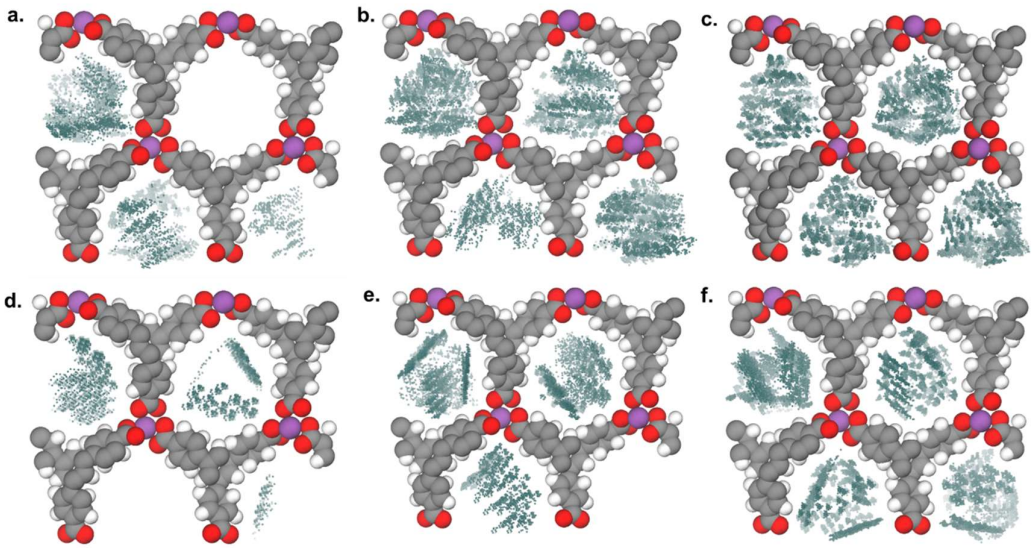


Figure 24 Densities of DCA (**a.**–**c.**) and α-CHC (**d.**–**f.**) at different loadings in CAU-7 (same loadings as in **Figure 23**): **a.**121 mg g^{-1}or 10.8 wt.%, **b.** 307 mg g^{-1}or 23.5 wt.%, **c.** saturation with 500 mg g^{-1}or 33.3 wt.%, **d.** 42 mg g^{-1} or 4.0 wt.%, **e.** 126 mg g^{-1} or 11.2 wt.% and **f.** saturation with 248 mg g^{-1} or 19.9 wt.%. The atoms of the drug molecules are coloured in grey, with a gradient showing the depth of the cell. The darker dots are closer to the reader, whereas the lighter ones are further from the reader.

3.4.2.3 Conclusion

This case study illustrated the use of GCMC simulations as a comparative guide to assess the maximum loading capacity of two cancer drugs in CAU-7. It also helped evaluate the possibility of fitting large drugs such as α-CHC inside the MOF. The MC simulations performed in the NVT ensemble also provided useful visualisation of the filling process. This case study is an example of how simple MC simulations can also help drug studies.

3.4.3 Case study 3: a computational study of the chiral selectivity of CMOM-1S, -2S and -3S on a series of molecules

This case study presents the first simulation results from an ongoing project led by Dr Shi-Yuan Zhang in collaboration with Dr Rocio Bueno-Perez, both from the Adsorption and Advanced Materials group, Department of Chemical Engineering and Biotechnology, University of Cambridge, UK. Dr Shi-Yuan Zhang synthesised the CMOM structures and provided their CIFs. Dr Rocio Bueno-Perez modelled the anions present in the frameworks and three of the considered adsorbates (2-butanol, 2-pentanol and 2-hexanol).

3.4.3.1 Introduction

The resolution of chiral compounds is of critical importance in the pharmaceutical, food, fragrance and agricultural fields. A compound is chiral when it cannot be superposed with its mirror image, in the same way that our left hand cannot be superposed to our right hand. Such a compound and its mirror image are called enantiomers. The existence of these two forms of a molecule is often problematic: one form of a drug can be effective, while the other is inactive or even potentially harmful.[155] One way of separating the two enantionmers is by means of chiral chromatography on a chiral stationary phase (CSP), which interacts selectively with one of two forms. Among potential CSP material candidates are MOFs. Homochiral MOFs, that is, MOFs with a single sense of chirality throughout the framework – as opposed to heterochiral MOFs,[156] are of particular interest. As an example, Zhang et al. recently synthesised a series of such homochiral MOFs, called chiral metal-organic materials (CMOMs).[157-159] From a computational perspective, the study of CMOMs is particularly interesting. Conversely to most homochiral MOFs modelled computationally, CMOMs have rigid frameworks but adaptable pores. In particular, the benzene rings in the linkers are prone to rotations, and the anions to self-rearragement.[157, 158] Previous studies of fully rigid homochiral MOFs have shown that the enantioselectivity of a MOF is very sensitive to slight changes of the adsorption site.[160-162] Indeed, even the smallest shift in the adsorption space can lead to more or less stereospecific environments, and therefore to more or less efficient separations. This was confirmed by the study of the impact of framework flexibility on the enantioselectivity of a zeolite.[163] Therefore, the modelling of CMOMs is expected to be particularly challenging.

In this study, I looked specifically at three parent CMOMs derived from CMOM-1S $[Co_2(man)_2(bpy)_3](NO_3)_2]$ (man = S-mandelate, bpy = 4,4'-bipyridine). CMOM-2S and -3S are obtained by substituting the anions in CMOM-1S with BF_4^- and $CF_3SO_3^-$ respectively. These structures were reported to have various separation properties with regard to three structural isomers of phenylpropanol: 1-phenyl-1-propanol (1P1P), 1-phenyl-2-propanol (1P2P) and 2-phenyl-1-propanol (2P1P) (see **Figure 25b**).[157]

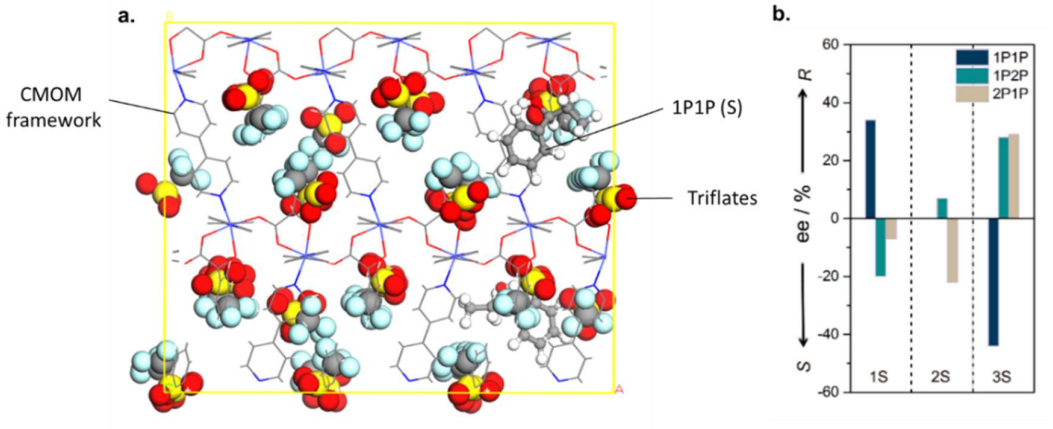


Figure 25 a. A 3x1x3 cell of CMOM-3S framework, with the rigid CMOM framework, the moving triflates and 1P1P molecules. **b.** Enantioselectivity obtained by Zhang et al. in various CMOMs.[157] There are four anions in each unit cell of framework presented in **a.**

The goal of this study is two-fold: i) develop a method to computationally confirm the enantioselectivities of CMOMs observed experimentally, ii) guide experimentalists in choosing potential chiral molecules to be separated with the considered CMOMs. To model the CMOMs, a series of MC simulations broke down the structures into a rigid framework (the benzene rings are kept fixed in this study for the sake of simplicity), the anions and the chiral molecules. More details about this method are given below. This method was first tested on CMOM-3S, before being applied to CMOM-1S and CMOM-2S.

3.4.3.1 Simulation setup

I carried out MC simulations in the systems composed of the three components below and shown in **Figure 25a**:

1. The CMOM's common framework

I used an atomistic model for the common part of the three CMOM's frameworks, where the atoms were kept fixed at their crystallographic positions. 3x1x3 cells of frameworks were used each time to account for the periodic boundary conditions.

2. Anions

The anions were modelled by Dr Rocio Bueno-Perez with the all-atom 0.8-scaled OPLS-2009IL force field.[164]

3. Chiral molecules

The regular all-atom force field OPLS was used for the modelling of the selected molecules. [130, 165-167] An all-atom force field was chosen to better capture the role of each atom in the chirality of the molecule. The aromatic rings were kept rigid. Details of the molecules definitions are provided in Appendix B.

MC simulations in the NVT ensemble were first performed to obtain an equilibrated (framework + anions) system, with a total of 36 anions in the total 3x1x3 cells of framework. I then performed GCMC simulations to simulate the uptake of chiral molecules in three different systems:

i) the previously obtained (framework + anions) system where the anions are fixed,
ii) the previously obtained (framework + anions) system where the anions are allowed to move,
iii) the framework and the anions kept fixed at their final positions obtained experimentally by Zhang et al.[157, 159] These systems represent the state in which the structures are experimentally proven to be selective. This scenario requires the availability of the corresponding experimental data, and were therefore only studied for CMOM-3S.

The number of anions were kept fixed at 36 in the GCMC simulations, and 0 chiral molecules were added at the start. The parameters for the framework were derived from a mix of the UFF[127] and the DFF.[128] Up to 1,000,000 MC cycles were performed, the first 10% of which were used for initialisation, and the remaining steps were used to calculate the ensemble averages.

3.4.3.2 Results

3.4.3.2.1 Phenylpropanols in CMOM-1S, -2S and -3S

Figure 26 summarises the results obtained computationally for 1P1P in CMOM-3S. **Figure 26a** shows the uptakes obtained in system iii). **Figure 26b** shows the predicted uptakes in system i), where the anions are fixed at their previously equilibrated positions**. Figure 26c** corresponds to system ii), where the anions are allowed to move, after being previously equilibrated. **Figure 26d** is a repeated independent simulation of system ii), with a different random seed. **Figure 26e** and **Figure 26f** show snapshots of system iii) (corresponding to **Figure 26a** and **Figure 26c**). The two figures are positioned for easier comparison. The comparable regions are further delimitated by the black lines.

The total uptake of the two enantiomers is the same in all three scenarios (0.84 mol kg^{-1} or 18 molecules in the simulation box composed of 3x1x3 unit cells of framework), but the selectivities are different. The obtained selectivity for system iii) in **Figure 26a** (*S* over *R*) matches the experimental results (**Figure 25b**). However, this selectivity could not be computationally confirmed in either system ii) or i). Indeed, although the predicted uptakes for each enantiomer are slightly different in **Figure 26c** and **Figure 26d** (corresponding to system ii) where the anions are moving), both plots show the inability to discriminate the 1P1P enantiomers. **Figure 26b**, for which the anions are fixed, also shows no significant difference in predicted uptakes compared to **Figure 26c** and **Figure 26d**. Examination of the movies obtained from the simulations of system ii) confirms that the anions indeed vibrate in the same positions obtained in system i). **Figure 26e** and **Figure 26f** show that, apart from a few molecules, the general location of the anions are similar. The black circle in **Figure 26f** indicates an example of an anion positioned at a significant different location than its peers. The orientations of the anions are however more varied. This is particularly the case along the z-axis, as the positions of the anions in system iii) are obtained experimentally for a unit cell and are repeated by symmetry and kept fixed.

Although cheaper, system i) is thereafter no longer used, as I believe the anions' movement is a determining feature of the CMOMs studied.

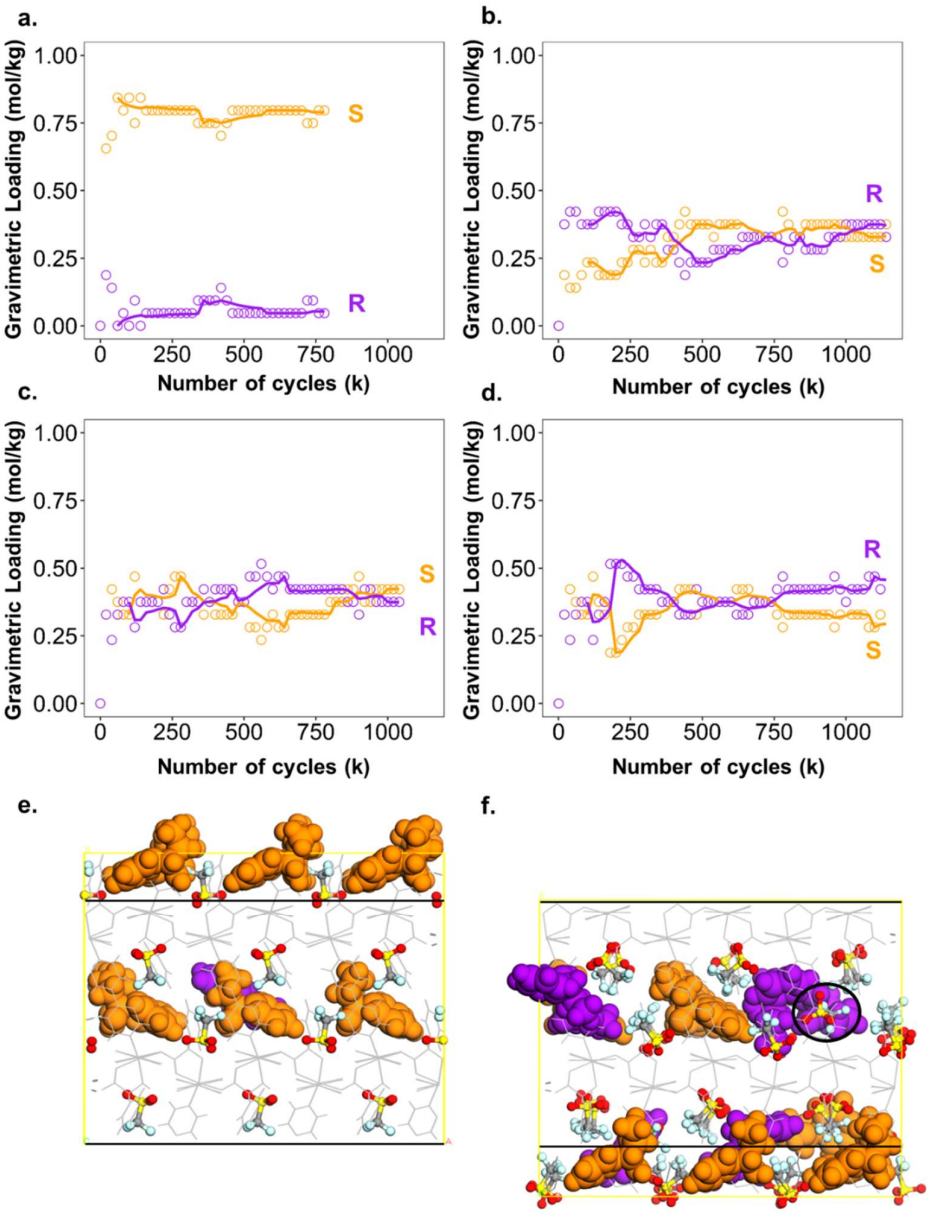


Figure 26 Gravimetric loading of (*S*)-1P1P (orange) and (*R*)-1P1P (purple) in CMOM-3S when **a.** the anions are at their experimentally obtained positions (system iii) and kept fixed, **b.** the anions are kept fixed at the positions obtained after computational equilibration of the anions in the framework (system ii), **c.** and **d.** when the anions are allowed to move (system i), repeated independent simulations. The circles correspond to the absolute uptakes, and the lines correspond to the cumulative average uptakes after equilibration. Snapshots of the simulation box (yellow lines) of **e.** system iii) and **f.** system ii) corresponding to **c.** The framework is presented in greyed lines, the anions in sticks and balls, the (*S*)-1P1P molecules in orange (CPK) and the (*R*)-1P1P molecules in purple (CPK). The black lines delimitate the comparable regions in the two snapshots.

Figure 27 shows the uptakes of the 1P2P enantiomers obtained computationally. **Figure 27a** corresponds to system iii) and **Figure 27b** to system ii). **Figure 27c** and **Figure 27d** present the corresponding snapshots. While the computational and experimental results (see **Figure 25b**) matched for 1P1P in scenario iii), it is not the case for 1P2P. The simulation in this case predicted the opposite selectivity of the experimental results, although it did predict the ability of the structure to separate the 1P2P enantiomers. Similarly to the 1P1P case, however, this discriminatory power was not reproducible in system ii). The snapshots show a similar situation to that of 1P1P: with a few exceptions – such as highlighted by the black circle in **Figure 27d** – the anions are generally located in the same places in both cases, but with much higher orientation variance in system ii).

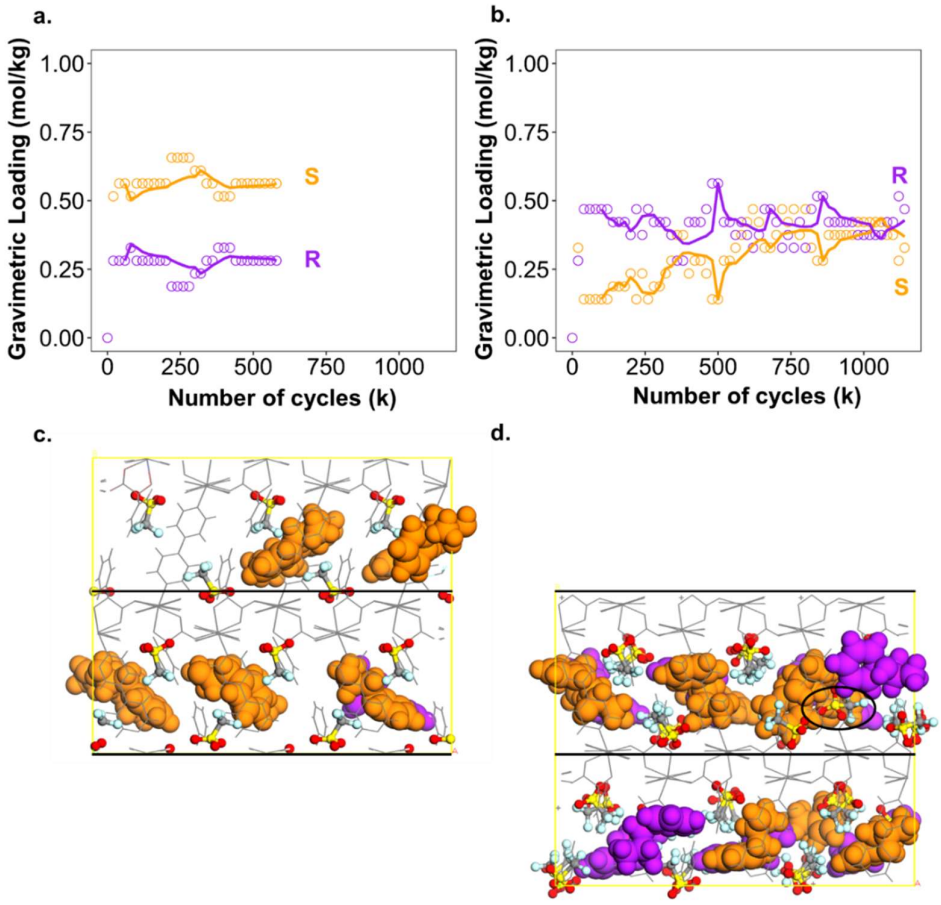


Figure 27 Gravimetric loading of (*S*)-1P2P (orange) and (*R*)-1P2P (purple) in CMOM-3S when **a.** the anions are at their experimentally obtained positions (system iii) and kept fixed and **b.** when the anions are allowed to move. The circles correspond to the absolute uptakes, and the lines correspond to the cumulative average uptakes after equilibration. Snapshots of the simulation box (yellow lines) of **c.** system iii) and **d.** system ii) corresponding to **a.** and **b.** respectively. The framework is presented in greyed lines, the anions in sticks and ball, the (*S*)-1P2P molecules in orange (CPK) and the (*R*)-1P2P molecules in purple (CPK). The black lines highlights a comparable region in the two snapshots.

Figure 28a shows the predicted uptakes of 2P1P in system iii) and **Figure 28b** the corresponding uptakes in system ii). Similarly to 1P2P, **Figure 28a** indicates a discriminatory ability that is opposite to the experimentally observed one. **Figure 28b** shows once again the inability of system ii) to identify any selectivity.

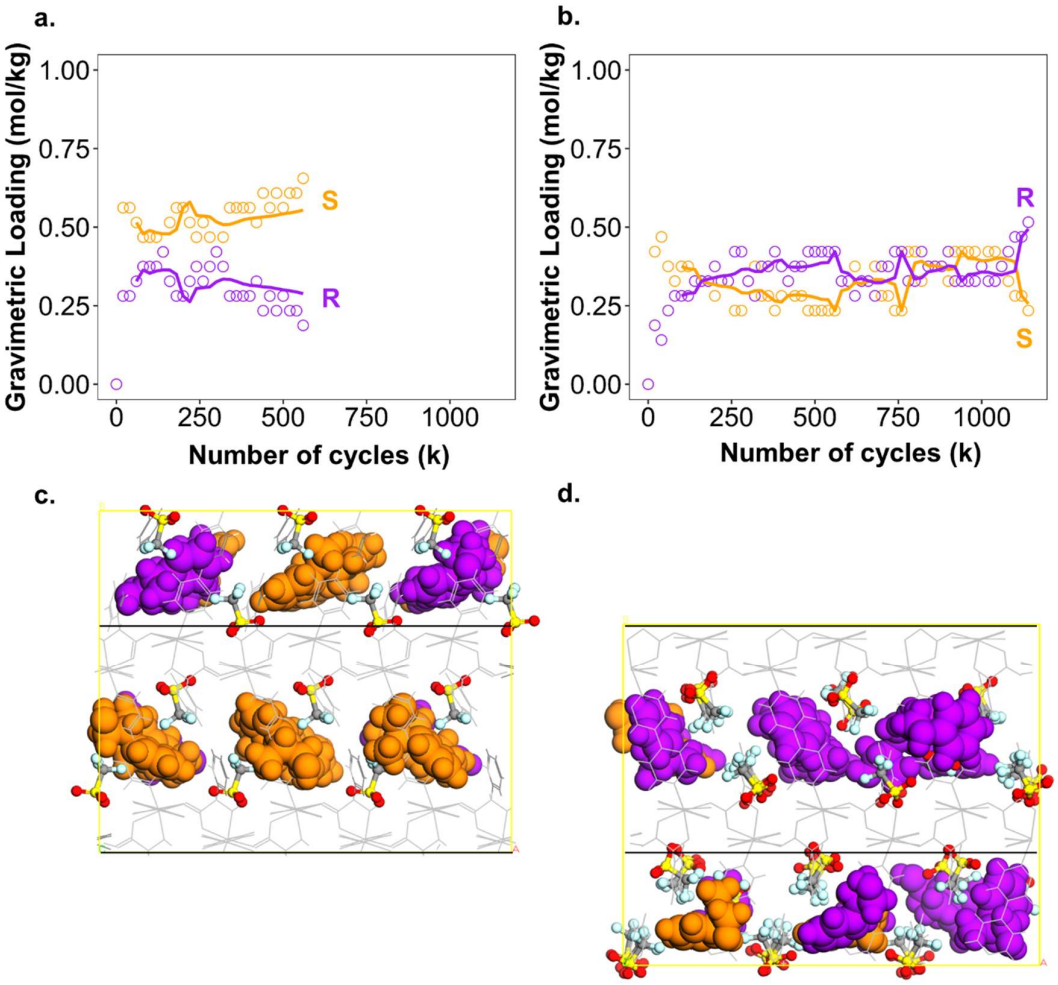


Figure 28 Gravimetric loading of (*S*)-2P1P (orange) and (*R*)-2P1P (purple) in CMOM-3S when **a.** the anions are at their experimentally obtained positions (system iii) and kept fixed and **b.** when the anions are allowed to move. The circles correspond to the absolute uptakes, and the lines correspond to the cumulative average uptakes after equilibration. Snapshots of the simulation box (yellow lines) of **c.** system iii) and **d.** system ii) corresponding to **a.** and **b.** respectively. The framework is presented in greyed lines, the anions in sticks and ball, the (*S*)-2P1P molecules in orange (CPK) and the (*R*)-2P1P molecules in purple (CPK). The black lines highlights a comparable region in the two snapshots.

Figure 29 presents the uptakes of 1P1P, 1P2P and 2P1P in CMOM-1S and CMOM-2S in system ii). These plots do not show any chiral discrimination ability from the structures.

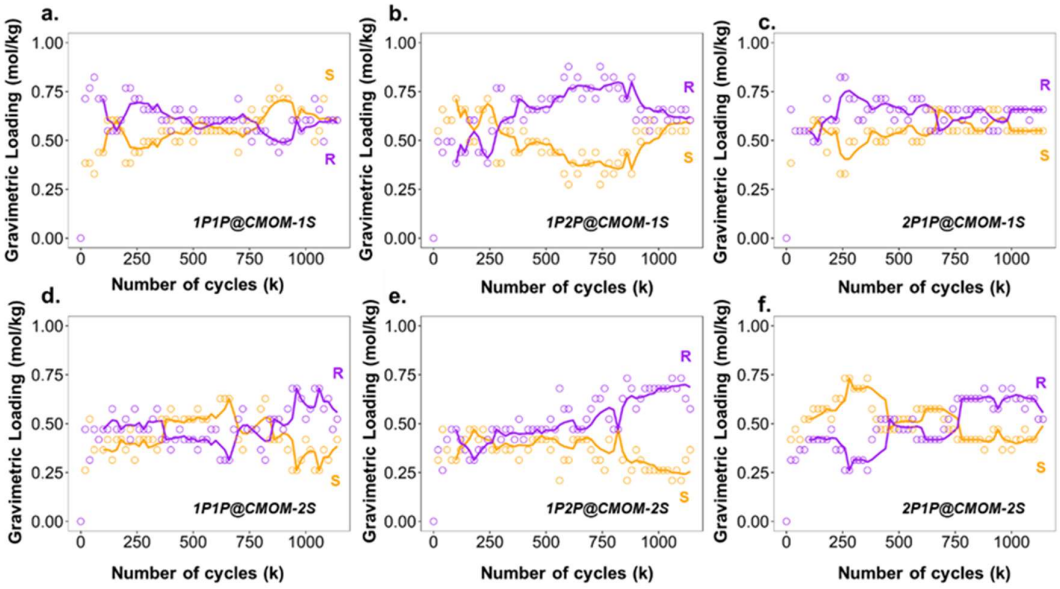


Figure 29 Gravimetric loading of **a.** 1P1P, **b.** 1P2P and **c.** 2P1P in CMOM-1S, and **d.** 1P1P, **e.** 1P2P and **f.** 2P1P in CMOM-2S in system ii). The circles correspond to the absolute uptakes, and the lines correspond to the cumulative average uptakes after equilibration. Orange: *S*-form, purple: *R*-form.

3.4.3.2.2 Reverse high-throughput screening of chiral molecules in CMOM-3S

Most computational HTS in the field of MOFs screen a large number of structures with one given adsorbate. Different approaches have been adopted with homochiral MOFs, where a number of MOFs are screened with a number of potential chiral molecules.[168] I present here a similar HTS where a library of chiral molecules are screened in one given MOF: CMOM-3S in system ii). The goal of this HTS is to quickly identify enantiomers that could be further studied experimentally, thereby enlarging the usability scope of CMOM-3S. **Figure 30** presents the library of molecules considered for this HTS. 2-butanol, 2-pentanol and 2-hexanol were modelled by Dr Rocio Bueno Perez while I modelled the rest.

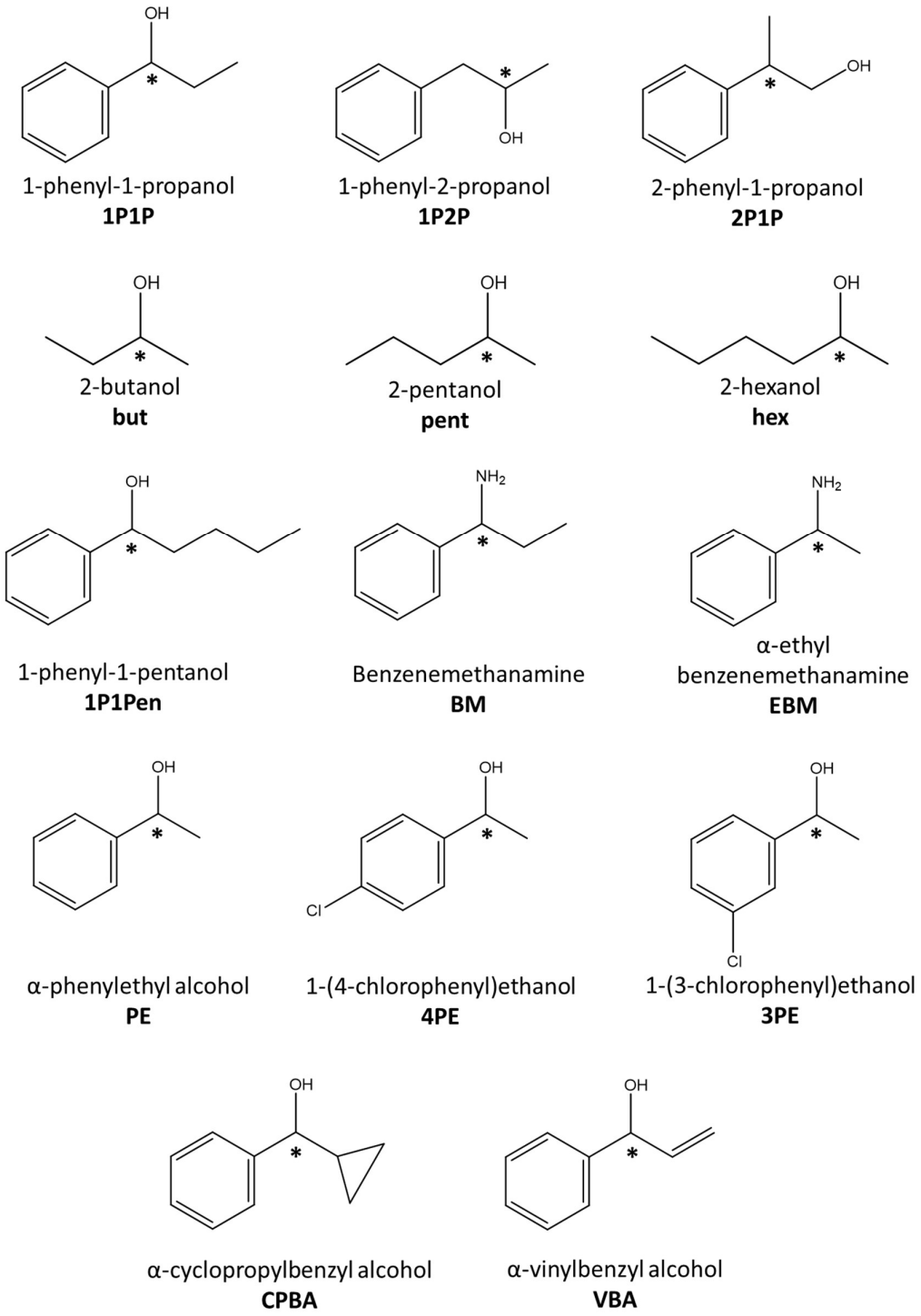


Figure 30 List of modelled chiral molecules. The asymmetric carbon is indicated with *. The full name of the molecules and the corresponding abbreviations used in this study are given below each structure.

Figure 31 presents the uptakes predicted for the library of molecules presented in **Figure 30**. All the simulations were carried out in system ii). While most plots show no significant chiral discriminatory power, four molecules stand out: 3PE, 4PE, CPBA and VB. 3PE and 4PE both have a chlorine atom, therefore it is likely that one of the enantiomers' configuration creates more favorable electrostatic interactions with the (framework + anions) system. The three-atom cycle in CPBA and the double bond in VB are likely to create comparatively bulkier steric environment, for which one of the enantiomers fits more comfortably in the given (framework + anions) system. In addition, the bulkiness and stronger electrostatic interactions could also reduce the mobility of the anions, thereby in turn impacting the structure.

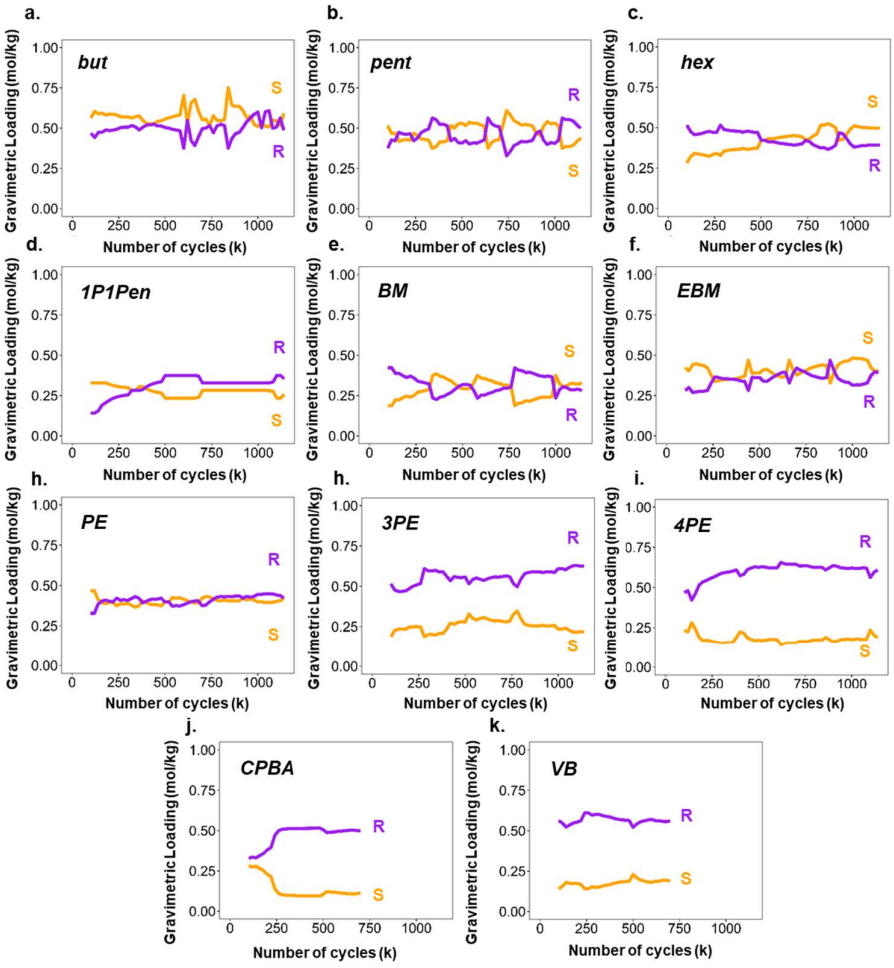


Figure 31 Gravimetric loading in CMOM-3S of **a.** 2-butanol, **b.** 2-pentanol, **c.** 2-hexanol, **d.** 1P1Pen, **e.** BM, **f.** EBM, **g.** PE, **h.** 3PE, **i.** 4PE, **j.** CPBA and **k.** VB. The lines correspond to the cumulative average uptakes after equilibration. Orange: *S*-form, purple: *R*-form.

3.4.3.3 Conclusion

In this study, I examined a MC method to simulate the chiral discriminatory power of a chiral MOM. The close look at 1P1P, 1P2P and 2P1P in CMOM-3S indicates that given the right (framework + anions) system, it is possible to predict if CMOM-3S has chiral separation abilities. However, the predicted selectivity may not match with experimental results. In addition, the method used to equilibrate the (framework + anions) system does not give reliable results. It could be good enough, however, for a quick screening of chiral adsorbates in which some molecules have strong defining features, such as favourable electrostatic and steric environments. Among the library of eleven molecules specifically modelled in this study, 3PE, 4PE, CPBA and VB are promising candidates for further experimental study for chiral separation in CMOM-3S. This case study also showed the importance of considering all the molecules involved in a framework when looking at selective adsorption. One main issue found in this case study is the difficulty for the system to be stable. A possible way to solve this in the future is to constrain the movement of the anions to a certain degree.

3.5 Conclusion

I reviewed in this chapter the molecular simulation methods used in this work. In particular, I illustrated with three examples the use of MC to i) uncover unusual phase change behaviour in ZIF-7, ii) complement experimental drug studies with CAU-7, and iii) investigate the chiral separation power of a series of CMOMs. In Chapters 5 and 6, the same methods will be used in a high-throughput manner.

4 TARGETED CLASSIFICATION OF METAL-ORGANIC FRAMEWORKS IN THE CSD

The following content was published in *Targeted classification of metal–organic frameworks in the Cambridge structural database (CSD)*, Peyman Z. Moghadam*, **Aurélia Li***, Xiao-Wei Liu, Rocio Bueno-Perez, Shu-Dong Wang, Seth B. Wiggin, Peter A. Wood, David Fairen-Jimenez, Chem. Sci., 2020, 11, 8373-8387.

*Authors contributed equally to this work.

Abstract. I present in this chapter the development of algorithms to break down the overarching family of metal-organic frameworks (MOFs) into a number of subgroups according to some of their key chemical and physical features: metal-cluster, network and pore dimensionality, surface chemistry (i.e. functional groups) and chirality. These tools are backed-up by an interactive web-based data explorer containing all the data obtained. This toolbox, integrated in the Cambridge Crystallographic Data Centre software, will guide future exploration of MOFs and similar materials, as well as their design and development for an ever-increasing range of potential applications.

4.1 Contributions

I designed and performed all the classifications presented in this chapter.

The data selection, cleaning, and the geometrical characterisation of the selected MOFs presented in this chapter are performed by Dr Xiao-Wei Liu, previously in the Adsorption and Advanced Materials group, Department of Chemical Engineering and Biotechnology, University of Cambridge, UK, and currently in the Advanced Membranes and Porous Materials Center, King Abdullah University of Science and Technology, Saudi Arabia.

Dr Marcus Fantham, previously in the Laser Analytics group, Department of Chemical Engineering and Biotechnology, University of Cambridge, UK, created the first version of the online data explorer. With his help, I modified and updated this tool to fit the purpose of the work presented in this chapter.

Dr Seth Wiggin, CCDC, UK, and I worked jointly on the framework dimensionality script. We conceptualised the algorithm together, and fine-tuned the iterative versions of the script together. Dr Seth Wiggin wrote the final CCDC-approved version of the script, and I performed all the calculations on the CSD MOF subset.

4.2 Introduction

Metal-organic framework (MOF) databases in conjunction with molecular simulations have proven to be extremely useful for the exploration of structure-property landscapes and screening of MOFs to find optimal materials. This can be exemplified by the efforts of the United States Materials Genome Initiative, aiming to accelerate the way materials are developed and deployed to market.[169] In spite of the enormous advances implemented in high throughput screenings (HTS) and data mining, no standard convention exists on how MOFs can be classified based on their important chemical and structural anatomy. Indeed, previous studies focused on the computational geometric analysis of structures such as surface area, pore size and void fraction. This is clearly useful for performing brute-force HTS for gas adsorption and/or separation in the entire structural phase space, giving a birds-eye point of view on property-performance relationships. Despite being of huge interest for experimentalists, large-scale targeted exploration of MOFs with specific characteristics such as a given chemical functionality, or a family of specific metal-cluster, has not been widely explored so far. A MOF identification scheme was recently developed to enable rapid data searches amongst the

existing databases.[170] The open software decomposes the structure and topology of a given MOF using standard cheminformatics formats to assign a unique identifier to the MOF. In this process, interesting information can be extracted from MOF databases, such as the most common linkers, polymorphs and topologies. The necessity for such capabilities results from the MOF community's growing knowledge on the advantages and challenges of MOFs, which has enabled them to focus their research interests on certain chemistries deemed relevant to their practice – an excellent example is the recognition of the outstanding stability of Zr-MOFs. By breaking down the big family of MOFs into smaller hierarchical categories of materials that exhibit similar features, researchers would benefit from a clearer evaluation on how the MOF landscape is structured in terms of what materials have already been synthesised. Precise identification of different classes of materials, as opposed to brute-force screening, can also significantly improve the way they are studied for different applications.

As part of the Cambridge Crystallographic Data Centre's (CCDC) efforts to categorise crystalline materials, I present here the classification of MOFs according to some of their key features and their evolution over time since they were first synthesised. Although the methods presented here do not represent a standardised approach to the classification of MOFs, these simple tools can help MOF researchers navigate through the data available and highlight the necessity to establish such standards. For easier data exploration, all the obtained information is available in an interactive data visualisation website at aam.ceb.cam.ac.uk/mof-explorer/CSD_MOF_subset.

4.3 A CSD-integrated toolbox for the exploration of the CSD MOF subset

As explained in Chapters 1 and 2, a set of scripts for the removal of bound and unbound solvents was previously released. This is useful for processing the structural data before further calculations. To enable easy data exploration of the Cambridge Structural Database (CSD) MOF subset, I present here two additions to the CSD toolbox consisting of: i) ConQuest and CSD Python API search queries and methods for specific types of MOFs and ii) a new script for the determination of framework dimensionality. This toolbox uses the CCDC software package and can therefore be applied to the CSD MOF subset directly. First, the MOFs are categorised into some of the most well-known secondary building units (SBUs) and functional groups, providing the possibility of looking for specific families of MOFs within the CSD using a combination of the CSD Python API and the *Draw* function in ConQuest. Second, the dimensionality of MOF

networks is investigated using an in-house python script. These new features – all integrated in the CSD – will allow users to have access to some of the most widely studied classes of MOFs in a single resource and offer a unique platform to boost the applicability of MOFs for a wide range of uses from gas storage/separation to asymmetric catalysis and enantiomer separation. Researchers can use the algorithms developed here to exploit the most recent MOF subset in the CSD release and maintained by the CCDC every quarter.[43] The principles outlined here are also customisable if need be; therefore, users can develop similar algorithms for new families of MOFs according to their interests, where the structures can be downloaded for computational studies.

4.4 Textural properties of MOFs and their evolution

Before presenting the aforementioned classifications, a geometric characterisation of the structures considered as porous was carried out by Dr Xiao-Wei Liu, previously in the Adsorption and Advanced Materials group, Department of Chemical Engineering and Biotechnology, University of Cambridge, UK, and currently in the Advanced Membranes and Porous Materials Center, King Abdullah University of Science and Technology, Saudi Arabia. This characterisation required prior data selection and cleaning, mostly performed by Dr Xiao-Wei Liu – apart from the identification of structures from which bound solvent should be removed, which I performed. This section briefly (4.4) presents the methods used by Dr Xiao-Wei Liu, as well as the results obtained, as these are thereafter combined with the targeted classification of MOFs.

The structural characterisation discussed here is focused on the porous MOFs from the CSD MOF subset version 5.37.[43] From a total of 55,547 non-disordered structures in the *Non-disordered MOF subset*, a number of MOFs were excluded from the structural analysis due to the presence of partial occupancy issues (583 MOFs) and those containing missing framework hydrogens (2,177 MOFs), leaving 52,787 structures (see details in the next paragraph). 8,253 materials were found to be porous according to previously described criteria, i.e. a nitrogen probe sized molecule with a radius of 1.86 Å can access the pores for surface area calculations.[43] The unbound solvents were removed for all structures – using previously developed Python scripts[43] – prior to the calculations. The bound solvents were removed for a total of 739 previously identified materials containing Cu-Cu paddle-wheels as well as CPO-27 (coordination polymer of

Oslo)/MOF-74-like structures.[43] **Figure 32** presents the workflow used for the selection of MOFs.

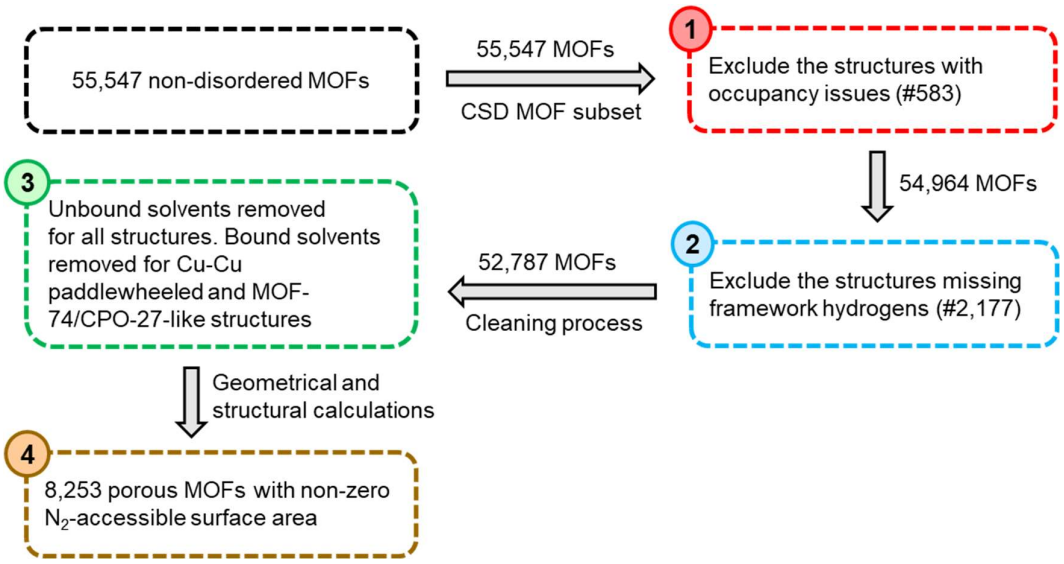


Figure 32 Flowchart outlining the CSD MOF subsets structures preparation prior to geometrical and structural calculations.

4.4.1 Identification of missing hydrogens and occupancy errors

As highlighted in Chapter 3, missing hydrogens are a common issue in the crystallographic data obtained experimentally. Occupancy issues related to non-hydrogen atoms have been observed as well. **Figure 33a** shows an example of such occupancy issues. The corresponding lines in the CIF are highlighted in **Figure 33b**. The atom label in these cases are followed by a question mark, which Dr Xiao-Wei Liu identified with bash scripts. Another script was developed for the identification of structures with missing hydrogens.[171] At the time of this work, the `add_hydrogen` function presented in Chapter 2 was not fully functional yet. Therefore, structures with missing hydrogens were directly discarded.

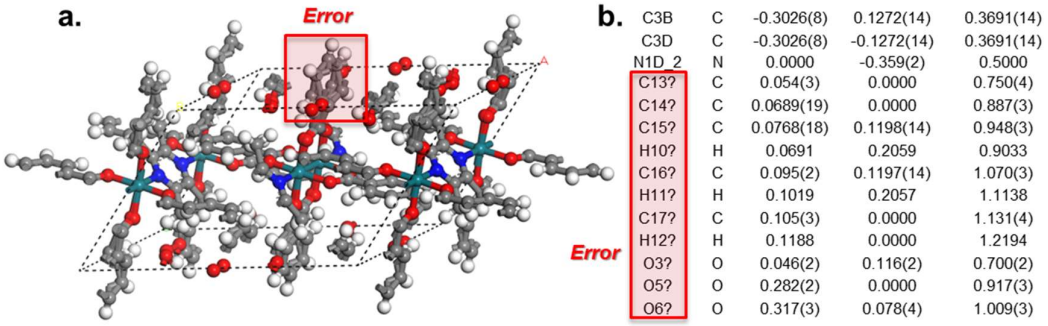


Figure 33 a. An example structure in the CSD MOF subset with occupancy issues (refcode: CIYER), **b.** atoms with occupancy issues in the CIYER CIF are highlighted.

4.4.2 Geometrical characterisation of the obtained subset

Figure 34 shows distributions of the geometric properties of MOFs and their evolution from 1995 to 2015. While very few MOFs were known until the early 21st century, the dramatic increase in the number of structures from 2000 to 2015 is evidence of how the remarkable characteristics of MOFs enable the exploration of a wide range of physical properties in porous materials. Most MOFs are concentrated in regions with pore sizes smaller than 10 Å and surface areas above 2000 m^2 g^{-1}, possibly due to the use of relatively inexpensive and commercially available short linkers such as terephthalic acid and the fact that this range of pore size is optimal for many gas storage and separation applications. As new synthesis methods of MOFs are designed every day, the introduction of longer linkers, more sophisticated SBUs and new topologies have continued to increase during the past decade.[172]

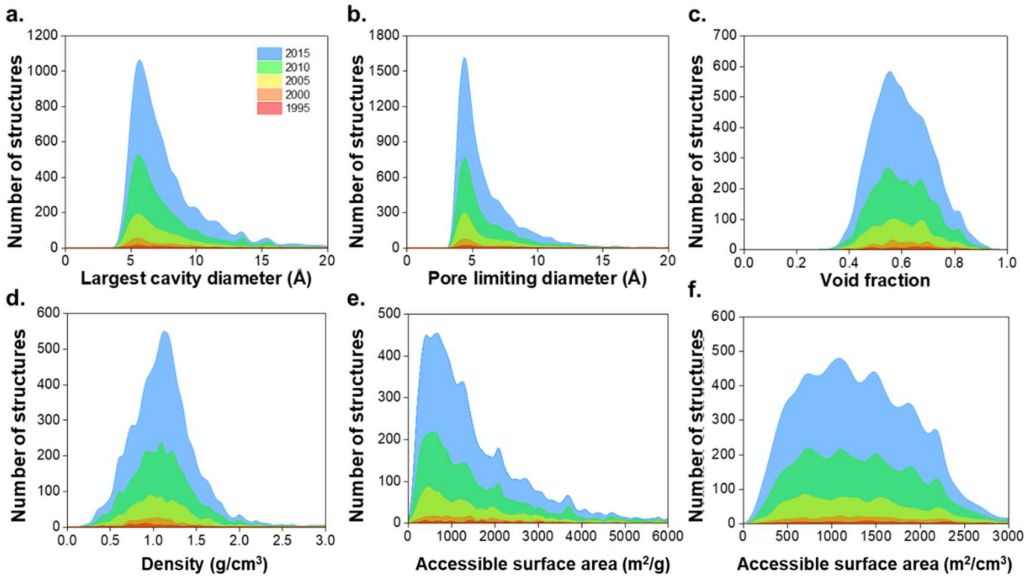


Figure 34 Histograms comparing geometric properties for all the porous MOFs in the CSD MOF subset from 1995 to 2015. **a.** Largest cavity diameter (LCD), **b.** pore limiting diameter (PLD), **c.** void fraction, **d.** density, **e.** gravimetric accessible surface area, **f.** volumetric accessible surface area. All family-property relationships of the 8,253 porous MOFs presented in this work can be found online at aam.ceb.cam.ac.uk/mof-explorer/CSD_MOF_subset.

4.5 Identification of target MOF families

I used ConQuest in the CSD MOF subset to identify MOFs with the desired SBUs; ConQuest offers the user a wide range of flexible search options based on the metal centers, organic linkers or SBUs. I developed search criteria for six prototypical MOF families well studied in the literature: Zr-oxide nodes (e.g. UiO-66), Cu-Cu paddlewheels (e.g. HKUST-1), ZIF (zeolitic imidazolate framework)-like, Zn-oxide nodes, IRMOF (isoreticular MOF)-like, and MOF-74/CPO-27-like materials. I also devised search criteria to identify MOFs containing common functional groups such as alkyls, alkoxys, halogens as well as polar functionalities, allowing to discriminate on the surface chemistry and therefore on the hydrophilic/phobic nature of the MOFs. These criteria introduce guidelines for MOF researchers to perform quickly targeted MOF searches, not only for the above classes of MOFs and surface chemistry but also for additional ones; the criteria can be customised in ConQuest, as explained below, to look for new MOF chemistries.

Intuitively, the initial approach to look for specific MOF families was to fully draw and search for each SBU in ConQuest. However, this approach resulted in fewer than expected MOF hits in each category. As explained in Chapter 2, when dealing with infinite polymeric structures, ConQuest carries out its searches on the smallest repeating unit based on the crystallographic symmetry, which may be different from the desired SBU, and therefore MOFs where the full metal cluster is not represented can be missed out. In other words, complete metal cluster information is only "assembled" in full when the unit cell is requested. To overcome this challenge regarding cluster representation, a series of criteria were developed to ensure that even MOFs with partially represented secondary building units are included in the search. **Figure 35** summarises the criteria developed for the identification of each MOF family. I used a step-by-step approach, where I started from the simplest search for a MOF family and then gradually tuned the search criteria by including or excluding certain bonds and connections in the metal cluster. At each step, the resulting materials were visually inspected until all unwanted structures were removed and the targeted MOFs were identified. The green and red diagrams included in **Figure 35** represent search queries in ConQuest that are respectively labeled as "must-have" and "must not have" queries. A criterion for a target MOF family is either one single "must-have" query, such as IRMOF-like structures, or a combination of "must-have" and "must not have" queries. When several "must-have" queries are represented separately, they correspond to an OR statement, and therefore only one of the green diagrams is required to be present in each search hit (see for example the Zr-oxide family in **Figure 35**). When several "must-have" queries are represented in the same dotted box, they correspond to an "AND" statement, and therefore each search hit should contain all the green diagrams (see MOF-74/CPO-27-type in **Figure 35**).

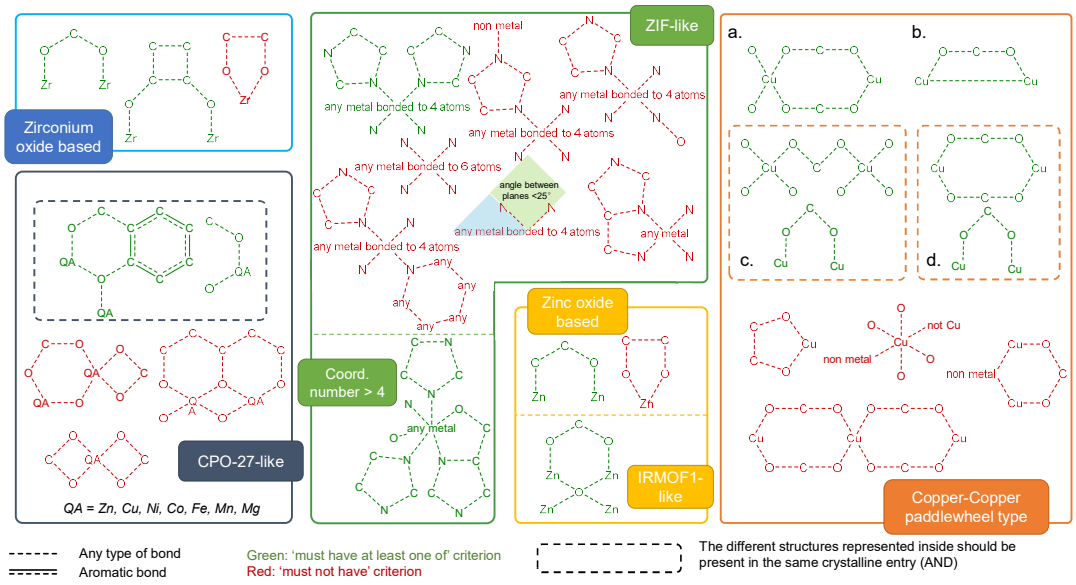


Figure 35 Criteria developed for the identification of MOF families in the CSD MOF subset based on specific SBUs and their connection to the organic linkers. The target MOF families are Zr-oxide, MOF-74/CPO-27-like, ZIF-like, Zn-oxide and IRMOF-like, as well as Cu-Cu paddle-wheeled materials. **a.-d.** diagrams used to look for structures containing Cu-Cu paddlewheels. The dotted box for **c.** and **d.** means the structures inside should be considered as one single query. The red diagrams are queries used to eliminate undesired structures. See Appendix C for more details on each MOF family.

I present here the derivation of the four search criteria for the family of Cu-Cu paddlewheel MOFs, which are a good example because they are usually not fully represented in ConQuest; the derivation of criteria for other MOF families is presented in Appendix C. **Figure 35a** represents the diagram of one complete paddlewheel and its connection to the linker via the two oxygen atoms. However, there are multiple cases where only half of the paddlewheel is represented. These structures are found using **Figure 35b** diagram, which contains only a section of the paddlewheel. The oxygen atoms were omitted from the linker, as I found that keeping these atoms returns fewer target structures. In this case, the two copper atoms are now bonded, corresponding to the rotational axis of the paddlewheel. More structures were found using the search criterion shown in **Figure 35c** diagram, which is in turn comprised of two parts. The upper part brings in structures in which the represented paddlewheel is "broken". However, other Cu-based structures with linear linkers are also included; this is avoided by adding the lower part, which represents the connection between the metal atoms and the linkers. The upper part of **Figure 35d** diagram is similar to the diagram in **Figure**

35a, without the oxygen atoms from the linkers bonded to the Cu atoms. Together with the lower part of the search criterion, the diagram from **Figure 35d** captures structures where the paddlewheel and the metal-linker connections are represented separately in ConQuest. **Figure 36** shows the structure hits. All in all, the four "must-have" queries result in 1,426 structures, some of which are not of the target type. To filter out these unwanted structures, I included another set of "must not have" criteria according to specific undesired structures (**Figure 37**). The combination of the "must-have" and the "must not have" criteria leads to a total of 1,015 MOFs containing Cu-Cu paddlewheel building blocks.

Combined together, Zn-oxide and IRMOF-like materials account for 3,187 structures, followed by 1,015 for Cu-Cu paddlewheels, 274 for ZIFs, 108 for CPO-27-like structures and 77 for Zr-oxide structures in the CSD 5.37 version from May 2016. **Figure 38** presents histograms that map the geometric properties of each MOF category. Zn-oxide MOFs being the largest family, the corresponding structures cover the widest range of LCD, PLD, void fraction, density and surface area. Despite this wide coverage, a few modes can be observed: LCDs of 5, 9, 10 and 16 Å, PLDs of 5, 7 and 10 Å, void fractions around 0.6 and 0.85, densities around 0.5 and 1.0 g cm^{-3}, volumetric surface areas of 1000, 1500 and 2200 m^2 cm^{-3} and gravimetric surface areas of 1000, 1500, 1800 and 4000 m^2 g^{-1}. In particular, the IRMOF-like materials fit perfectly one mode in each of these structural properties: LCD of 16 Å, PLDs of 7 Å, void fractions of 0.85, densities around 0.5 g cm^{-3}, volumetric surface area of 2200 m^2 cm^{-3} and gravimetric surface area of 4000 m^2 g^{-1}. MOFs containing Cu-Cu paddlewheels also show distinct modes despite covering a range of values: LCDs around 7, 9, 12, 14 Å, PLDs around 5 and 7 Å, void fractions around 0.5 and 0.7, densities around 0.9 g cm^{-3}, volumetric surface areas around 1000, 1400 and 2000 m^2 cm^{-3} and gravimetric surface areas around 1000 and 2000 m^2 g^{-1} mostly. CPO-27-like structures present one distinct mode for each property: LCDs of 12 Å, PLDs of 11 Å, void fractions of 0.65, densities of 1.2 g cm^{-3}, volumetric surface areas of 1500 m^2 cm^{-3} and gravimetric surface areas of 1100 m^2 g^{-1}. ZIF-like structures and Zr-oxide MOFs are in significant smaller numbers with values covering a wide spectrum for most properties. However, ZIF-like structures do show a distinct peak of void fraction at 0.65, volumetric surface area of 1200 m^2 cm^{-3} and gravimetric surface area of 1100 m^2 g^{-1}. As for Zr-oxide MOFs, most structures have void fractions around 0.7, three modes of densities (0.6, 0.7 and 1.0 g cm^{-3}), volumetric

surface areas around 2200 m^2 cm^{-3}. These histograms thus show that the chosen classification was able to capture MOFs with a range of differing behaviours.

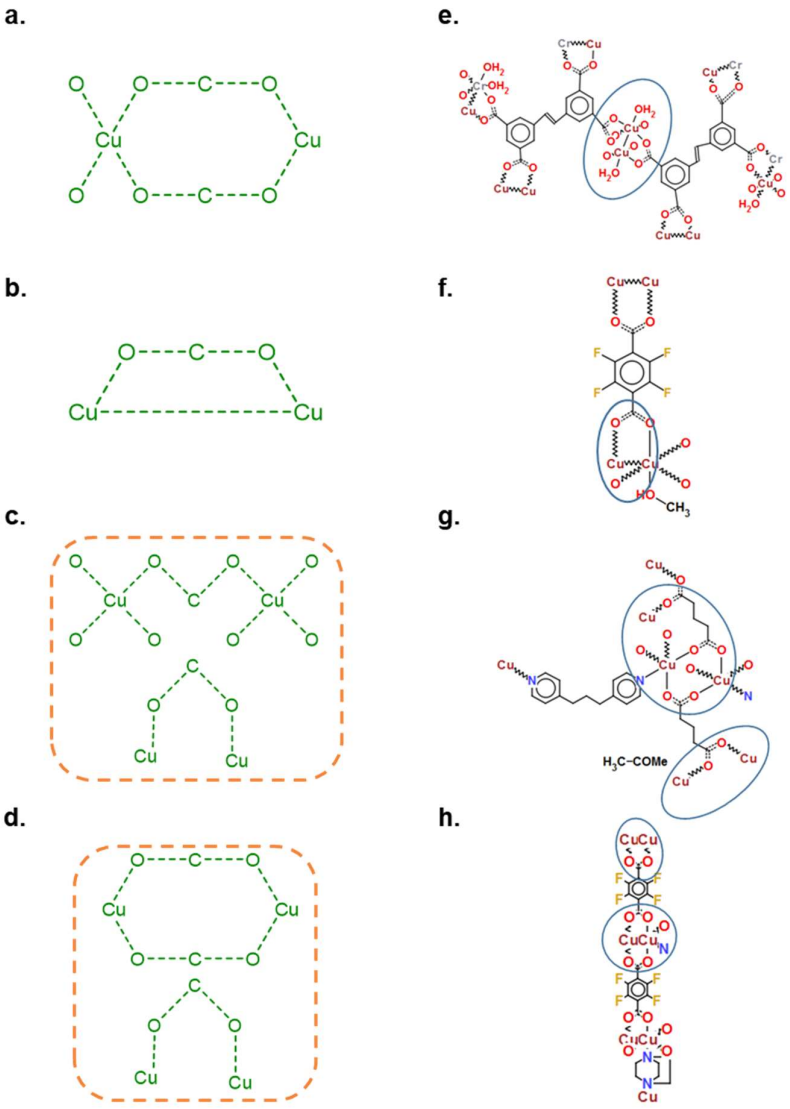


Figure 36 a. to **d.** Criteria developed to look for structures containing Cu-Cu paddlewheels. **e.** to **h.** Example structures found using the criterion on the left. **a.** returns 988 hits, **b.** returns 611 hits, adds 178 to the list, **c.** returns 716 hits, adds 248 to the list, **d.** returns 647 hits, adds 12 to the list. For **c.** and **d.**, the dotted box means the structures inside should be considered as one single query. The blue circled areas show the parts that have been searched for in ConQuest. CSD refcodes: **e.** ACUJOZ, **f.** ACASUT, **g.** ACAJOF, **h.** ACATAA.

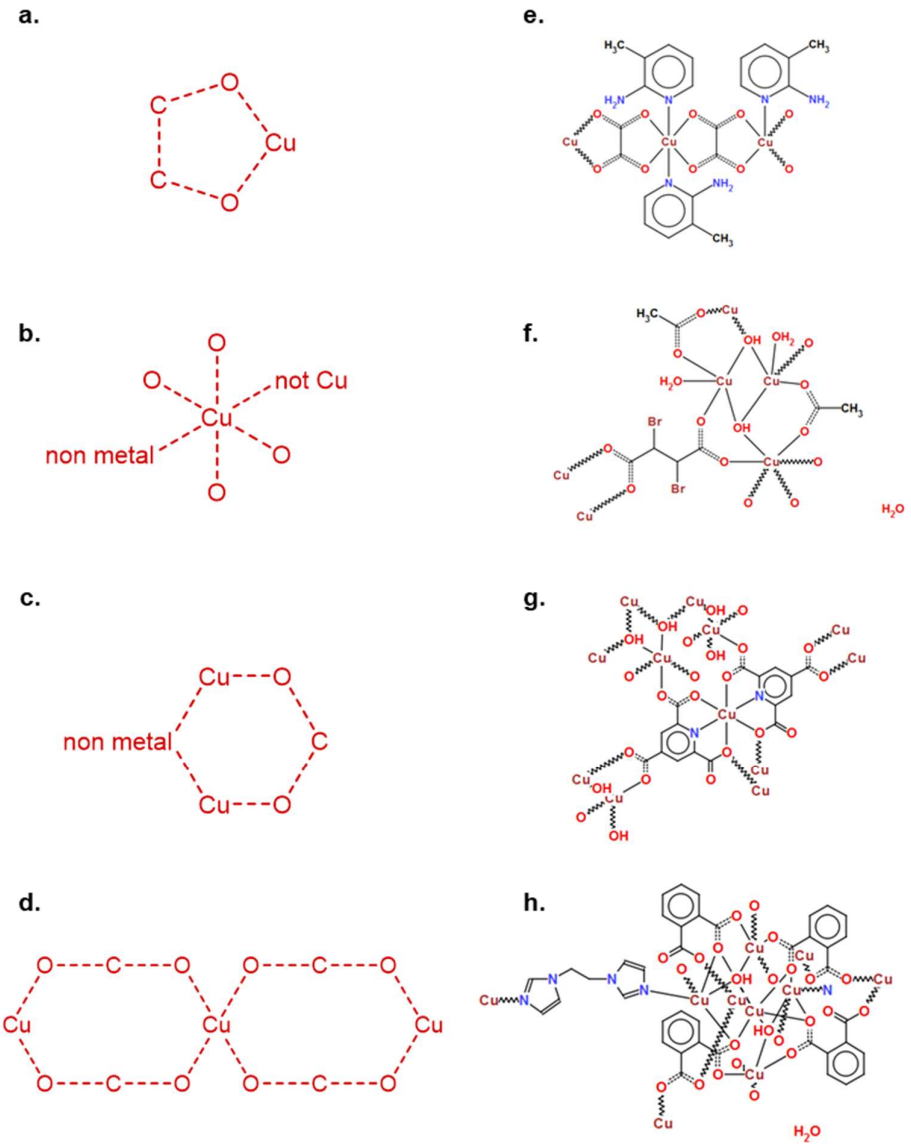


Figure 37 a. to **d.** Criteria used to eliminate undesired structures and the number of structures eliminated at each step. **e.** to **h.** Examples of eliminated structures corresponding to the criteria on the left. **a.** eliminates 128 hits, **b.** eliminates 190 hits, **c.** eliminates 88 hits, **d.** eliminates 5 hits. CSD refcodes: **e.** ABOCUP, **f.** AHEGIF, **g.** AGUMAR, **h.** ASEWEB.

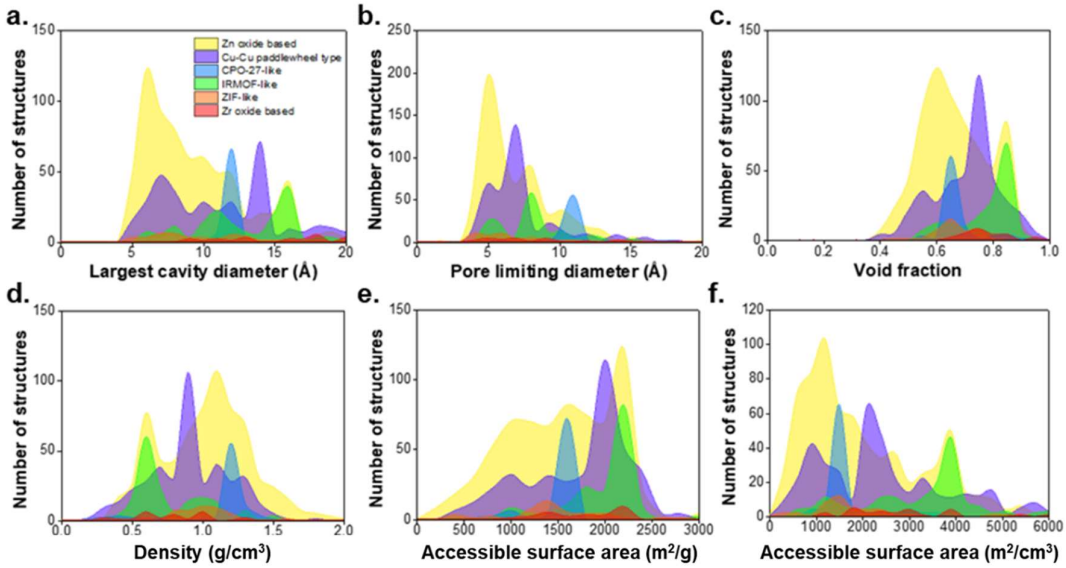


Figure 38 Histograms showing the geometric properties for each MOF family identified in the CSD MOF subset. **a.** largest cavity diameter (LCD), **b.** pore limiting diameter (PLD), **c.** void fraction, **d.** density, **e.** gravimetric accessible surface area, **f.** volumetric accessible surface area.

4.6 Identification of surface functionalities in MOFs

Functionalisation plays a crucial role in fine-tuning the chemical and physical properties in MOFs. Rational incorporation of chemical functionalities has been extensively employed using various pre- or post-synthetic engineering techniques as well as in computer models of MOFs for a breadth of applications including carbon capture,[173, 174] gas separation and sensing,[175-177] catalysis,[178, 179] light harvesting[180] and optical luminescence.[181] I present in this section the targeted identification of a number of distinct functional groups categories such as polar functional groups (-NH_2, -NO_2, -CN, -COOH, -OH), alkoxys (methoxy, ethoxy, propyloxy), alkyls (methyl, ethyl, propyl and alkyls containing more than 4 carbon atoms) and halogens (-F, -Cl, -Br). **Figure 39** shows the combination of ConQuest queries used to target these functionalised MOFs. These queries should be combined with a CSD Python API script that ensures the search fragments are only present in the main framework and not part of a solvent. This script is provided in Appendix D. Users can directly reproduce the queries drawn in **Figure 39**, or download them from the publication corresponding to this work.[171] **Figures E1** to **E4** in Appendix E show the frequency of occurrence as well the geometric properties for all MOFs with the functional groups described above.

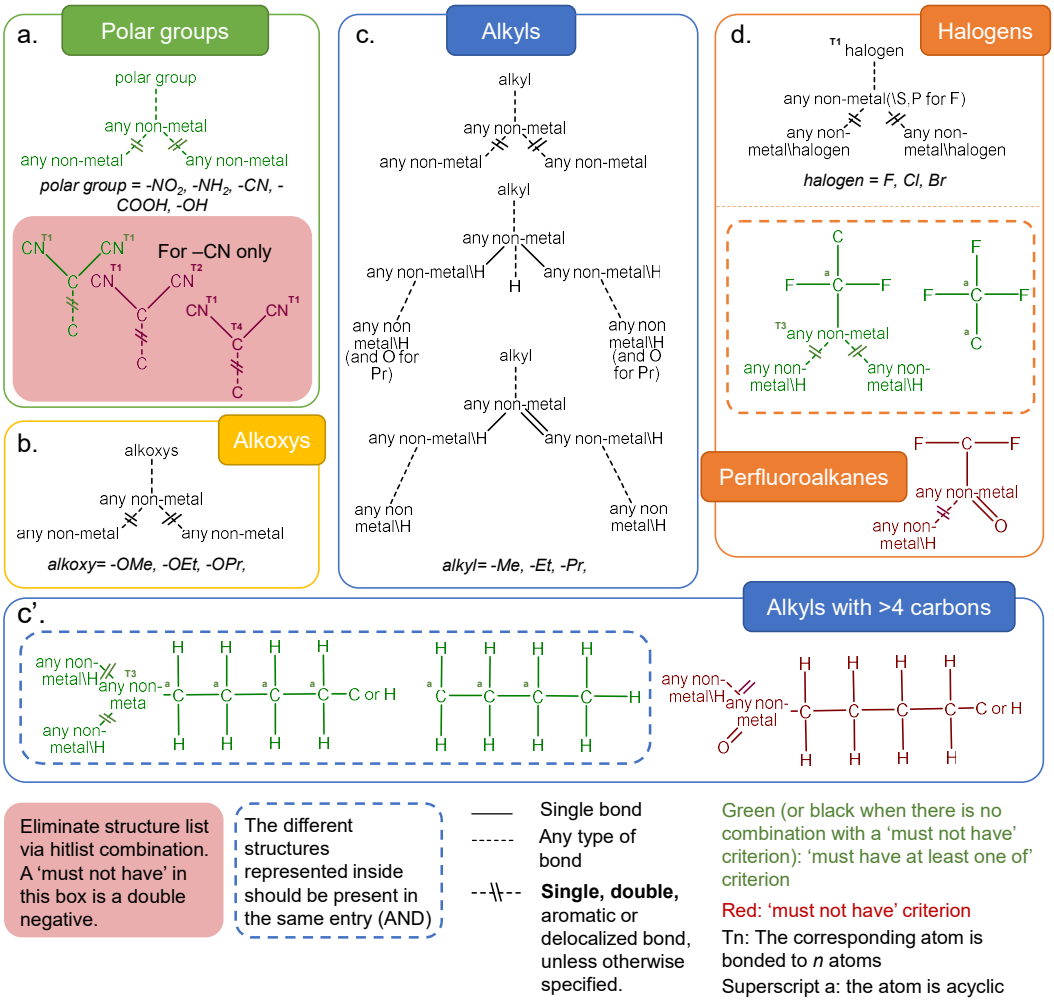


Figure 39 Criteria developed to identify MOFs with common functionalities in the CSD MOF subset. **a.** polar groups (-NH_2, -NO_2, -CN, -COOH and -OH). For the -CN case, the red box represents queries which target dicyanides that are chosen to be eliminated. This dicyanide search is obtained via a combination of one "must-have" query and two "must not have" queries. The green diagram is thus an overall negative and the red diagrams are double negatives. **b.** alkoxys (methoxy, ethoxy, propyloxy); **c.** alkyls (methyl, ethyl, propyl); **c'.** alkyls with more than 4 carbon atoms and **d.** halogens (-F, -Cl, -Br), and structures with perfluoroalkane groups. The variable bonds are all the same type for queries within the grey dotted box: single, double, aromatic or delocalised. For the three queries outside of the grey dotted box, the variable bonds are either aromatic or delocalised. See Appendix D for more details on each functional group.

4.7 Identification of chiral MOFs

The previously targeted subsets of MOFs were closely related to adsorption applications, which guided the choice and design of the criteria presented earlier. For instance, the list of 55,547 structures in the CSD MOF subset was narrowed down to 8,253 porous MOFs. Similarly, considering other applications from a wider range of areas, these queries can be tuned according to a new set of criteria and a different subset suitable for these purposes can be designed. As an example, precise knowledge of existing chiral MOFs and their structural properties facilitates the identification and engineering of MOF chirality for niche catalytic and enantio-separation applications.[30, 182-184] Given the flexibility provided by CSD Python API scripts, I also included the chirality of MOFs. Here, a chiral MOF is defined as presenting either chiral atoms in the structure or a chiral crystal packing. The atom attribute `is_chiral` from the CSD Python API corresponds to the first case, whereas the crystal attribute `is_sohncke` to the latter. 4,504 structures were found to contain *S*/*R*-chiral atoms and 6,859 structures in Sohncke-chiral space groups; combinatorial searches of chiral-ligand MOFs in chiral space groups gave 2,010 structures. It should be noted that only *R*/*S* chirality was taken into account and, therefore, structures with e.g. metal lambda/delta or axially-chiral structures were not included. **Figure 40** shows the physical and geometric properties of 1,911 chiral structures with non-zero surface area values. The group of chiral porous MOFs is included in the 8,253 porous MOF subset and represents around 23% of the latter. The distribution of geometrical properties is similar, and the majority of chiral structures synthesised so far contain small pores of less than 10 Å and surface area values of more than 2,000 m^2 g^{-1}. However, non-porous structures make up only 5% of the chiral MOFs, which suggests that researchers were actively looking for porous chiral structures. This could be related to the fact that more than 90% of chiral MOFs were synthesised after the 2000s, when MOFs were increasingly explored for their potential in catalytic applications and enantiomeric resolution.

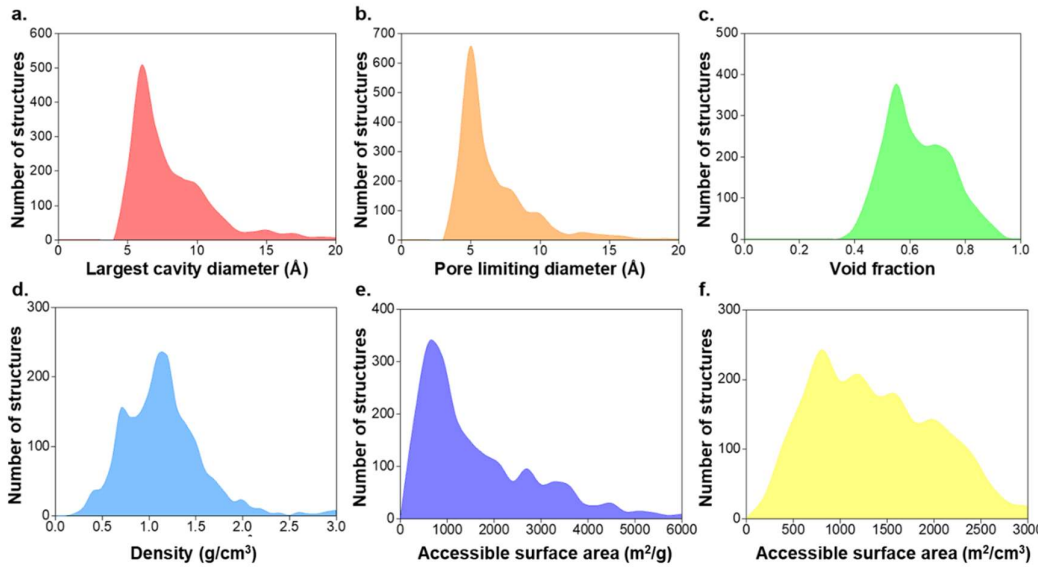


Figure 40 Histograms of the geometric properties of 1,911 chiral structures with non-zero gravimetric surface area in the CSD MOF subset.**a.** largest cavity diameter (LCD), **b.** pore limiting diameter (PLD), **c.** void fraction, **d.** density, **e.** gravimetric accessible surface area, **f.** volumetric accessible surface area.

4.8 Porous network connectivity and framework dimensionality

Knowing the porous network connectivity or dimensionality (also referred to as percolation) is important in determining MOFs applicability in certain adsorption applications. For example, 1D-channeled MOFs have shown to be highly selective in the separation of hydrocarbons due to favorable thermodynamic or kinetic origins towards one component, depending on channel size and shape.[185-187] The diverse nature of building units' linkage in MOFs results in variations of porous networks, where the connectivity of a porous network is determined by a geometric analysis of connecting pathways of porous components, resulting in 1D channels and 2D or 3D networks. Porous networks are normally sampled using mesh/grid-based propagation techniques that map the void space into connected components.[143, 188-190] To investigate the pore system accessibility and dimensionality, I used Poreblazer[109, 143] to determine the geometrical parameters of the pore networks for all 8,253 porous structures in the MOF subset. **Figure 41** shows the analysis, resulting in 86% 1D, 9% 2D and 4% 3D pore connectivity for these porous structures.

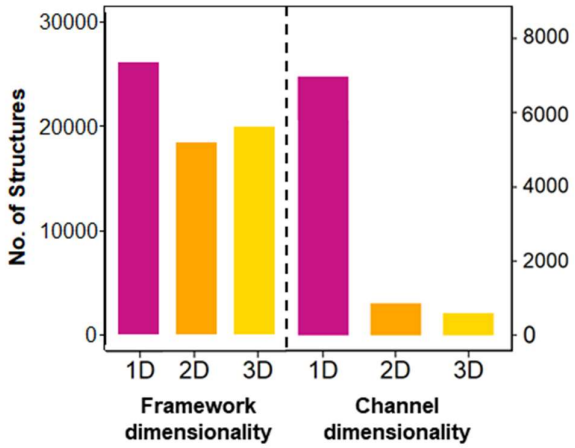


Figure 41 Histograms of framework and channel/pore dimensionalities characterised for the 52,787 structures. The framework dimensionality refers to the 1D (rod), 2D (sheet) or 3D shape of the structure, whereas the channel dimensionality refers to the 1D, 2D or 3D extension of porous network within the structure.

In addition to the pore network, framework dimensionality is also critical for selecting an optimal MOF for a given application. As defined in Chapter 1, the channel dimensionality characterises the extension of the porous network within the structure, whereas framework dimensionality describes the shape of the structure. The dimensionality of the structure is crucial in deciding which material is more practical for a given application. The algorithm co-developed with Dr Seth Wiggin, CCDC, UK (available in Appendix F) generates the smallest box containing the smallest repeating unit of each structure. The latter is then expanded and a new smallest-containing box is created. The dimensions of the initial box and the last box are then compared to determine in which directions the structure has expanded. The script was tested on 1/5th of the 52,787 structures (i.e. 11,515). The structures were randomly chosen and visually checked in Mercury, with the help of the "Polymer Expansion" functionality, which extends all the polymeric bonds in a given structure.[110] The results were compared to those obtained with Zeo++[60], open-source software that is able to determine framework dimensionality based on atom connectivity. 30% (i.e. 3,663) of the results disagreed, which led to the visual inspection of 2,157 of these structures. Our in-house script was found to be correct in 93% of the cases where there was a disagreement. Based on these comparisons and checks, our predictions are estimated to be overall 97% accurate. The results for all 52,787 porous and non-porous MOFs are included in **Figure 41**, where 40% of the structures are 1D, 29% are 2D and 31% are 3D.

4.9 An insight into the MOFs' crystal data quality

When dealing with such a high amount of experimental data, it is useful and interesting to have a better idea of the data quality. A simple way of assessing the quality of crystal structures is to analyse their crystallographic R-factors, available in the CSD and extractable via the CSD Python API. The R-factor is a measure of the discrepancy between the observed structure factor F_{obs} and the calculated structure factore F_{calc} upon crystal determination:

$$R = \frac{\sum ||F_{obs}| - |F_{calc}||}{\sum |F_{obs}|} \qquad \textbf{4-1}$$

where the sum extends over all the X-ray reflections measured and calculated. In other words, it measures the agreement between the theoretical crystallographic model and the experimental measurement. An R-factor of 0 corresponds to a perfect agreement between the experimental measurements and the predicted structure factors. High R-factors, typically above 10%, reflect refinement models that may contain systematic errors.[191] **Figure 42** shows the evolution of the R-factors of the MOF materials from 1960 to 2015; **Figure G1** and **Figure G2** in Appendix G show the characterisation of the physical and geometric properties for all MOFs and the corresponding families vs. R-factors. Although the field of MOFs is generally considered to have started in the late 1990s[192, 193] – as reflected by the increasing number of structures in **Figure 42**, scientists have been working on coordination polymers since the late 1950s, and even before. However, since the definition of MOFs is still debated today,[22, 43, 194] it is not straightforward to tell which structure truly is the first MOF. The oldest structure in the CSD MOF subset dates back to 1940 and consists of a sodium formate (NAFORM).[195] The general opinion would hardly consider this a MOF nowadays, although it still marginally fits the criteria required for being part of the CSD MOF subset. The most "MOF-like" 3D coordination polymer structure from the early days must be ADINCU by Saito and coworkers from 1959,[196] which is widely recognised by the community. This work was followed by Hoskins and Robson (JARMEU) and then by the groups of Yaghi and Kitagawa. I have, therefore, started the timeline in 1960. Despite the fact that the number of structures with R-factors higher than 10% has increased over the last decade, reaching 0.7% of the MOF subset in 2013, the mean and the median R-factor values have remained fixed at around 5%, and 99% of the structures have R-factors lower than 12%. To understand the evolution, it is worth noting the technological

advances in crystal structure determination between the 1960s and today. Until the 1970s, the mean values for most structures are above 10%, while in the 1980s, the R-factors significantly dropped to below 10% despite the increase in more complex and large structures being synthesised.[197]

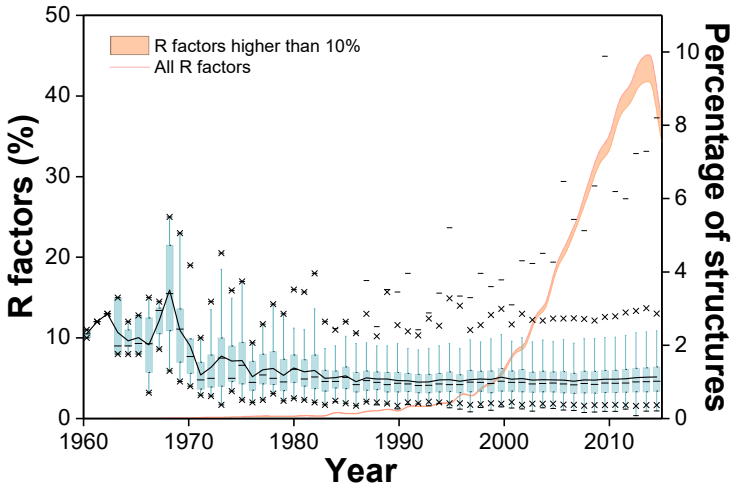


Figure 42 Non-cumulative evolution of R-factors of the MOF subset from 1960 to 2015. Blue: boxplots of R-factors per year. Percentiles used: 1% (lower dash symbol), 25% (lower cross symbol), 50% (dash in the box), 75% (upper cross symbol), 99% (upper dash symbol). A black line connects the means across all the boxes. The corresponding values are given by the y-axis on the left. The orange curve shows the percentage of structures added to the database per year. The corresponding values are given on the y-axis on the right. The orange area under the orange curve highlights the number of structures with an R-factor higher than 10%.

The development of MOF families such as the ones introduced above enables data analyses that provide an overview of the properties dedicated to these smaller subsets. As an example, **Figure 43** explores the quality of MOF structures – via their R-factors – by looking at their family (e.g. IRMOF-like, ZIF, etc.), crystal system, symmetry and density. For each family, structures are divided into their crystal systems and a boxplot shows the distribution of their R-factors. The crystal systems are arranged in decreasing order of symmetry: cubic, hexagonal, trigonal and tetragonal systems considered as "high symmetry", and orthorhombic, monoclinic and triclinic considered as "low symmetry". Each point representing a structure is then coloured according to its density. The property-landscape provided here shows for example that some families crystallise

in specific crystal systems (see CPO-27/MOF-74 and Zr-oxide MOFs), whereas others crystallise in all crystal systems, with different distributions. For instance, IRMOF-like structures tend to crystallise mainly in cubic or hexagonal systems and show higher R-factors in these systems. In general, the data presented here suggest that for all the families, low-density MOFs tend to form high symmetry structures – in accordance with the analysis of Øien-Ødegaard and co-workers.[191] From the general overview given in **Figure 43**, it is possible to focus on more specific aspects of R-factors for each family. For example, the boxplots in **Figure G3** show the distribution of R-factors among each crystal system for each family; those in **Figure G5** show the distribution of R-factors among high and low symmetry structures for each family.

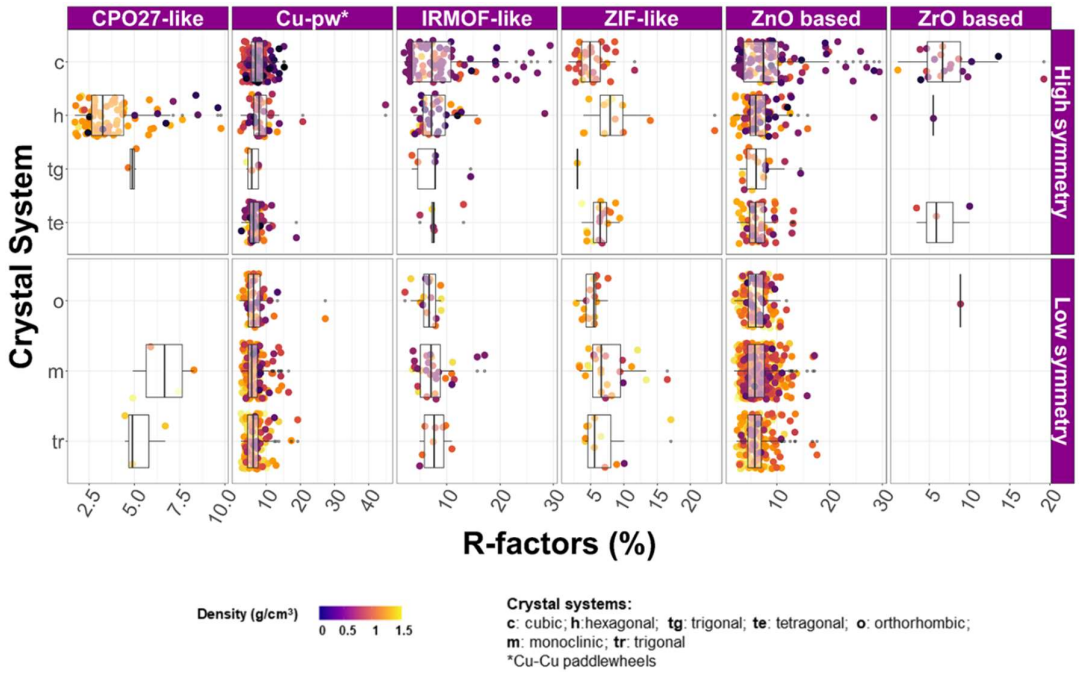


Figure 43 Distribution of R-factors and density across different MOF families and crystal systems of low or high symmetry. Each jittered point corresponds to a structure, and is categorised according to its crystal system (of high or low symmetry, as indicated on the right side) and its MOF family. Each colour corresponds to the density. The boxplot overlayed on top of the jittered points indicate the minimum, first quartile, median, third quartile, and maximum values of the corresponding points' R-factors.

An artificial way of "correcting" the experimental values obtained from X-ray diffraction patterns is to mask the solvent. To explore the effect of solvent masking on the quality of the crystal structure data, I finally compared the role of the structure refinement software SQUEEZE[113] in the distribution of R-factors. SQUEEZE enables users to identify and include the contribution of disordered solvent in the calculated structure factors upon determination of the crystal structure. **Figure G6** shows boxplot representation of the R-factors for the different MOF families, comparing the values on structures that have had their solvent masked through SQUEEZE and those that have not gone through this process. Although it might seem simple to assume that the use of SQUEEZE will lead to lower R-factors, there is not a clear trend to support this statement. One of the major difficulties when considering solvent masking and R-factors is how to determine what will produce the best structure for your purposes; a slightly lower R-factor structure that has had SQUEEZE applied, or a higher R-factor structure with an attempt to model all the disorder positions of the framework and/or guests.

It should be remembered that, although the R-factor is a convenient single metric to assess the quality of crystal structures, it simply measures the agreement between the refined model and the experimental data. The R-factor does not take into account how chemically and physically meaningful the resulting structure is, whether any use of solvent masking is appropriate or whether there are large residual electron density peaks. However, despite its simplicity, the R-factor is the only quality assessment metric publicly available in the CSD.

4.10 MOF explorer for 5D exploration of structural properties

All family-property relationships of the 8,253 porous MOFs presented in this work can be found online at aam.ceb.cam.ac.uk/mof-explorer/CSD_MOF_subset. Users can explore the structural landscape of the selected porous MOFs interactively with any one of up to 18 variables plotted in 5 dimensions. MOFs can be searched for and filtered by name, or by selecting them from the graph, allowing the user to track particular MOFs' characteristics. The different subsets developed in this Chapter can be colour-coded for easier visualisation. The corresponding data points can then be directly selected on the plots or from the data table for further exploration.

4.11 Conclusion

The coordination geometry of inorganic units and the diverse nature of MOF linkers have given rise to the emergence of thousands of diverse MOF materials with currently ca. 100,000 structures present in the CSD MOF subset. Here, I developed a customised set of criteria to identify specific families of MOFs as a powerful tool to classify them and speed up the way MOFs are being investigated for different applications. The computational tools and the interactive online data explorers provided in this work will allow MOF researchers to browse and look for targeted MOF categories based on secondary building units, chirality, surface chemistry as well as geometrical properties including pore and framework dimensionality. Through CCDC's structure search program ConQuest, the principles supplied here are customisable to enable the search and identification of new MOF families and functionalities based on any of the diverse pool of MOF building blocks.

5 HIGH-THROUGHPUT SCREENING OF THE CLASSIFIED CSD MOF SUBSET FOR HYDROGEN STORAGE

Parts of the following content are published in *Targeted classification of metal–organic frameworks in the Cambridge structural database (CSD)*, Peyman Z. Moghadam*, **Aurélia Li***, Xiao-Wei Liu, Rocio Bueno-Perez, Shu-Dong Wang, Seth B. Wiggin, Peter A. Wood, David Fairen-Jimenez, Chem. Sci., 2020, 11, 8373-8387.

*Authors contributed equally to this work.

Abstract. I demonstrate in this chapter the usefulness of the tools developed in Chapter 4 with a high-throughput screening (HTS) for hydrogen storage at room temperature. The results of this study are two-fold: i) mapping the previously obtained data on HTS results offer interesting behavioural insights regarding metal-organic frameworks (MOFs), ii) using MOFs for hydrogen storage at room temperature and high pressures (500 to 900 bar) is not industrially interesting and other temperature and pressure conditions should be sought.

5.1 Contributions

I performed the data selection, cleaning and all the simulations presented in this chapter.

Dr Marcus Fantham, previously in the Laser Analytics group, Department of Chemical Engineering and Biotechnology, University of Cambridge, UK, created the first version of the online data explorer. With his help, I modified and updated this tool to fit the purpose of the work presented in this chapter.

5.2 Introduction

To demonstrate the usefulness of the methods and analysis presented in Chapter 4, I included their application into hydrogen storage, using a high-throughput screening (HTS) based on grand canonical Monte Carlo (GCMC) simulations. Cost-effective and high capacity hydrogen storage remains a challenge for the widespread use of fuel cell applications. Although hydrogen has a higher gravimetric energy density than most other fuels, its volumetric energy density is one of the lowest.[198] The main challenge is thus to store enough hydrogen in a compact space. The US Department of Energy had set a target of 30 g L^{-1} of volumetric capacity by 2020 in order to first reach 40 g L^{-1} in 2025 and ultimately 50 g L^{-1}.[199] Among the possible storage solutions being currently researched, adsorption in porous materials is a promising one. As current on-board containers operate at high pressures (700 bar for Toyota fuel cell vehicles) and room temperature,[11] I predicted the adsorption uptake at 298 K over a range of low to high pressures of 200, 500 and 900 bar. Although HTS has been widely performed on metal-organic frameworks (MOFs) for hydrogen storage, very little work published results at these conditions.[74] In addition, the classification presented in this paper enables interesting visualisations regarding the performance of different classes of MOFs, thereby either further confirming previous observations with the amount of data available in the Cambridge Structural Database (CSD) MOF subset or presenting new ones. Using the methods described above, readers can also create their own classification and map it to their screening results.

5.3 Structures preparation for high-throughput hydrogen uptake simulations

3D structures were selected from the CSD version 5.37 using the Python API script described in Chapter 4. All structures had their unbound solvent removed using the CSD Python API scripts published previously. Structures containing Cu-Cu paddlewheels and CPO-27/MOF-74-like structures had their bound solvent removed using the same scripts. Missing hydrogens were added using the `add_hydrogen` function in the CSD Python API. Any additional hydrogen-related disorder was removed by using the "non-disordered" filter in ConQuest, following the protocol described in Chapter 2 to differentiate between the "non-disordered" filter and the *Non-disordered MOF subset*.[51] A pore limiting diameter (PLD) of 2.8 Å, corresponding to the lowest σ of the hydrogen atom across different force fields, was used to eliminate structures with lower PLDs.

5.4 Grand canonical Monte Carlo simulations

The GCMC simulations were performed in the multi-purpose code RASPA.[141] I used an atomistic model of each structure where the framework atoms were kept fixed at their crystallographic positions. The standard Lennard-Jones (LJ) 12-6 potential was used to model the interactions between the framework and fluid atoms. In addition, a Coulomb potential was used for fluid-fluid interactions. The parameters for the framework atoms were obtained from the DREIDING force field[128] and, when not available, from the Universal Force Field[127], whereas the hydrogen molecule was modelled by placing a single LJ sphere at the center of mass.[200] The Lorentz-Berthelot mixing rules were employed to calculate fluid-solid LJ parameters, and LJ interactions beyond the cut-off value of 12.8 Å were neglected. The simulation box for each structure is defined so that the cell lengths are larger than twice the cut-off distance. 30,000 Monte Carlo (MC) cycles were performed, the first third of which were used for equilibration and the remaining steps were used to calculate the ensemble averages. MC moves consisted of insertions, deletions and displacements. In a cycle, N MC moves are attempted, where N is defined as the maximum of 20 or the number of adsorbates in the simulation box. To calculate the gas-phase fugacity I used the Peng-Robinson equation of state.[150]

5.5 Results

Figures 44a-c show the absolute volumetric uptake (mass of hydrogen over volume of framework) versus the absolute gravimetric uptake (mass of hydrogen over total system mass) of these structures at the three considered pressures. Each circle represents a MOF. The colours highlight the six different families of MOFs chosen in this work, as described above, whereas grey circles represent the structures that do not fit in this classification; **Figures 44d-f** and **Figures 44g-i** highlight the pore dimensionality and surface chemistry, respectively, of the structures. The size of each circle represents the LCD of the corresponding structure. The corresponding gravimetric uptake in an empty tank is represented with a dashed line. A dynamic representation of the simulations can be found at aam.ceb.cam.ac.uk/mof-explorer/H2_HTS. Similarly to Chapter 4,[77, 177, 201] this allows the visualisation of the absolute hydrogen gravimetric and volumetric uptakes with respect to different structural properties such as void fraction, largest cavity diameter (LCD), PLD, isosteric heat of adsorption, and surface area to better understand their role. More importantly, it allows the multidimensional visualisation of the generated data in an interactive way, where, each data point (i.e. each MOF) can be individually identified and tracked into the CSD and the Cambridge Crystallographic Data Centre website.

The empty tank reference shows that, for pressures higher than 200 bar and at room temperature, the MOFs do not provide any improvement in terms of volumetric uptake. Room temperature and high pressure are therefore not the way forward for efficient hydrogen storage in porous materials unless new radical ideas are implemented. Nevertheless, the trends obtained still unveil valuable insights; I will henceforth focus on the information gained from mapping the classification previously obtained to the screening results.

Figures 44a-c show that the highest uptakes, especially gravimetric, are obtained for Cu-Cu paddlewheel, CPO-27/MOF-74-like and IRMOFs structures, whereas other Zn-oxide-type structures tend to have lower performance. Zr-MOFs, known to have large chemical stability among MOFs, show moderate gravimetric uptakes but competitive volumetric values. When looking at the pore connectivity, the trends reproduce those from the MOF families found here (**Figures 44d-f**). In particular, Cu-Cu paddlewheel MOFs form 3D-pore networks whereas CPO-27/MOF-74 form 1D channels and therefore the highest uptakes are for 3D and 1D MOFs. **Figures 44g-i** show that alkyl,

alkoxy and polar groups are often present in high uptakes, whereas structures containing alkyl groups have a slightly lower volumetric uptake. **Figure H1** in Appendix H shows in more detail the nature of the functional group in these cases: $-CH_3$, -OH and $-OCH_3$ are the functional groups present in the best-performing structures. However, this seems to be due to their larger numbers and wider spread of values. **Figure H2** in Appendix H and **Figure 45** provide similar information with regard to the structures' crystal systems and the metal atoms they contain. **Figure 45** is particularly interesting when combined with **Figures 44a-c**, as they suggest the best-performing CPO-27/MOF-74-type structures – which are among the overall best-performing ones – are frameworks containing magnesium atoms due to its lighter character. This is in agreement with studies on the role of magnesium in better hydrogen adsorption in MOFs.[198] All in all, the structure with the best volumetric and absolute uptake is a Cu-Cu paddlewheel, 3D-channeled unfunctionalised MOF, BAZGAM (**Figures 44a-c**), which has been identified previously in the literature for its exceptional performance at 77 K and 100 bar (reported simulated values of 34.3 g L^{-1} and 19.3 wt.% H_2).[74] At room temperature and 900 bar, its uptake values calculated in this work are 42.7 g L^{-1} and 25.1 wt.% H_2.

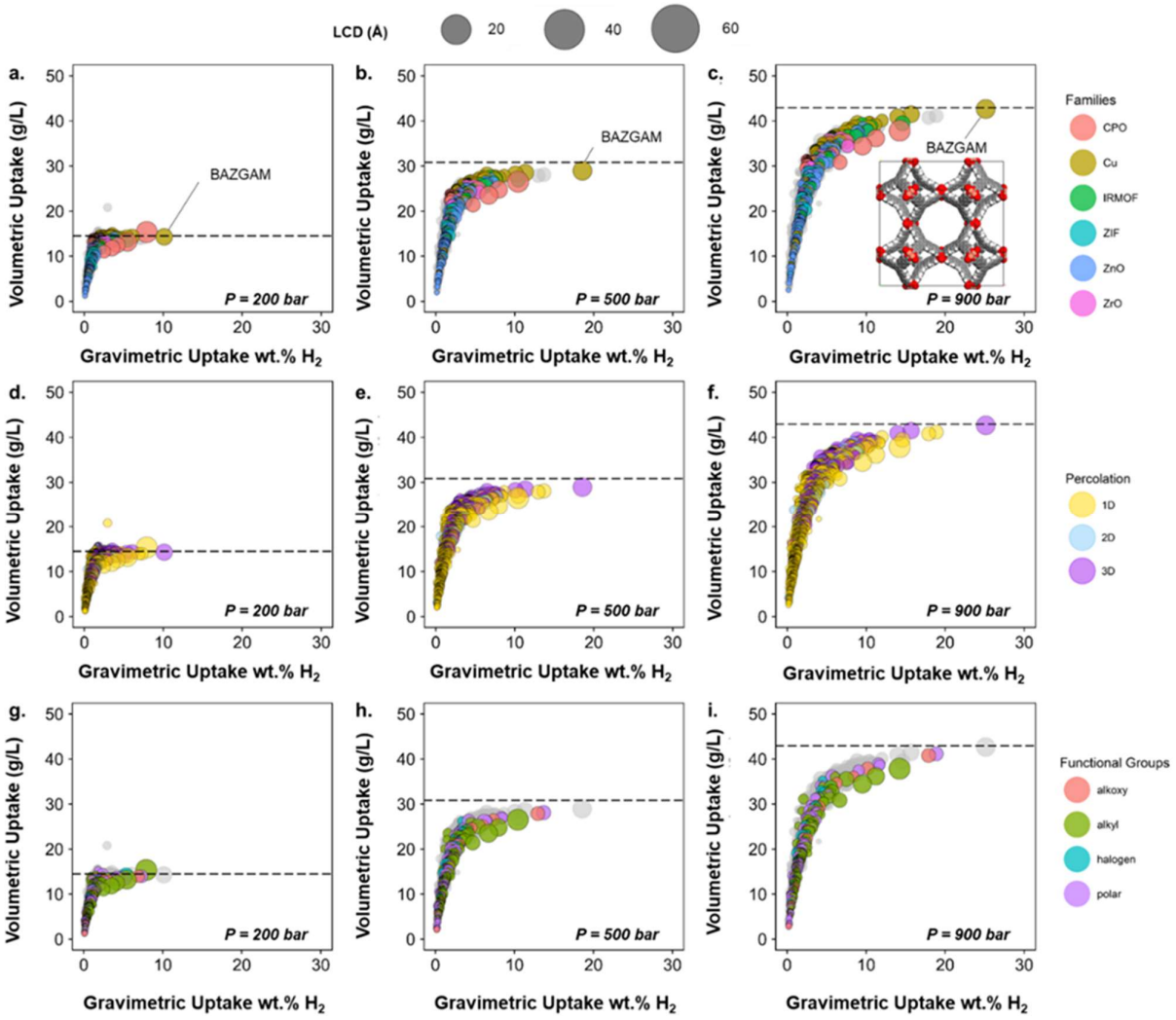


Figure 44 Characterisation of the 3D MOFs screened for hydrogen storage. Absolute volumetric uptake vs. absolute gravimetric uptake wt.% H_2 at room temperature at 200, 500 and 900 bar. Each circle represents a MOF structure. The sizes of the circles represent the LCD in all plots. The dashed line corresponds to the volumetric uptake obtained in an empty tank. **a.-c.** Families of the screened structures; structures that have not been assigned a family are coloured in grey in the background. The highlighted structure BAZGAM is shown in the inset at 900 bar. **d.-f.** Percolation of the screened structures. Structures containing 1D, 2D and 3D pore channels are respectively represented in yellow, blue and purple. **g.-i.** Functional groups identified in the screened structures. Structures that have no particular functional groups identified are coloured in grey in the background. Full hydrogen adsorption data can be found online at aam.ceb.cam.ac.uk/mof-explorer/H2_HTS.

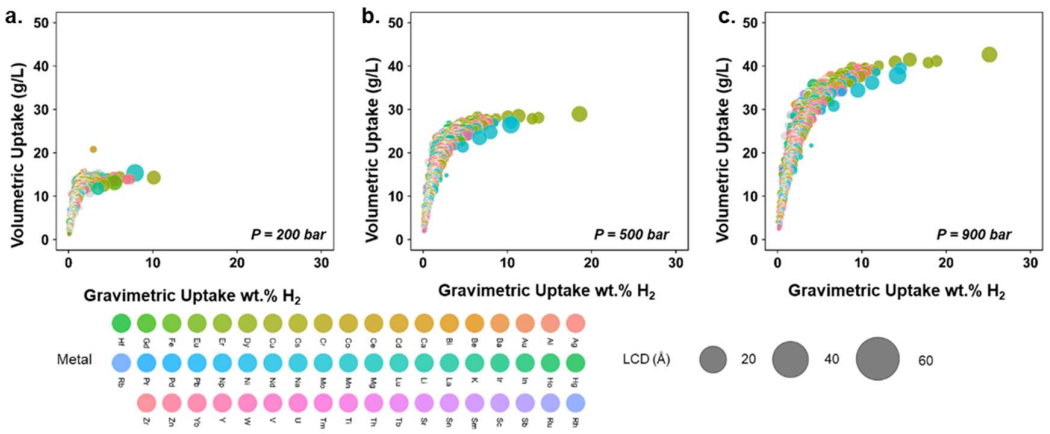


Figure 45 Volumetric uptake versus gravimetric uptake in wt.% H_2 for the screened structures for hydrogen storage at **a.** 200 bar, **b.** 500 bar and **c.** 900 bar. Each circle corresponds to a structure. The colours highlight a metal present in each structure, and the size of the circles indicate the largest cavity diameter (LCD).

While **Figure 44** highlighted the characteristics of the best-performing structures, **Figure 46** gives more quantitative insights, through statistical analyses, of these observations; **Figure H3** in Appendix H provides similar boxplots in terms of gravimetric uptake. **Figures 46a-c, d-f** and **g-i** show boxplots representations of the volumetric uptake for each of the MOF families, the percolation and the type of surface chemistry present, respectively. **Figures 46a-c** show that CPO-27/MOF-74-like, Cu-Cu paddlewheels, IRMOFs and Zr-oxide MOFs perform better at all three different pressures. In addition, they adsorb hydrogen more easily as the pressure increases: the amount of hydrogen adsorbed in ZIFs and Zn-oxide-type structures quadruples from ca. 5 to 20 g L^{-1} as the storage pressure increases from 200 to 900 bar, whereas the amount adsorbed in CPO-27/MOF-74-like, Cu-Cu paddlewheels, Zr-oxide and IRMOFs structures increases from ca. 7 to 30 g L^{-1}, reaching 32 g L^{-1} in IRMOFs, over the same range of pressures. Interestingly, **Figures 46d-f** show that, as could be expected, 3D-channeled structures have, on average, higher volumetric uptake than 2D-channeled structures, which in turn have higher volumetric uptake than 1D-channeled structures. In addition, the difference in performance increases as the storage pressure increases: 3D-channeled structures have in average a 40, 48 and 53% higher uptake at 200, 500 and 900 bar, respectively, than 1D-channeled structures. **Figures 46g-i** show that structures containing halogen groups perform better overall, and the spread of volumetric uptake of structures containing alkyl groups is wider as the pressure

increases. **Figure H1** provides a breakdown of each functional group, showing that structures containing -Br, -F and -OCH_2CH_3 groups stand out as having the highest volumetric uptakes.

Previous similar work that screened MOFs for hydrogen storage focused on the relationship between their geometrical properties (such as pore volume[93] or void fraction[198]) and performance. In this case, I mapped out the behaviour of the different classes of MOFs outlined in Chapter 4, thus providing a clearer picture of the CSD MOF subset landscape. In particular, I identified the volumetric and gravimetric storage limits for different families of MOFs, thus offering more insights into which MOF space is more promising or unexplored.

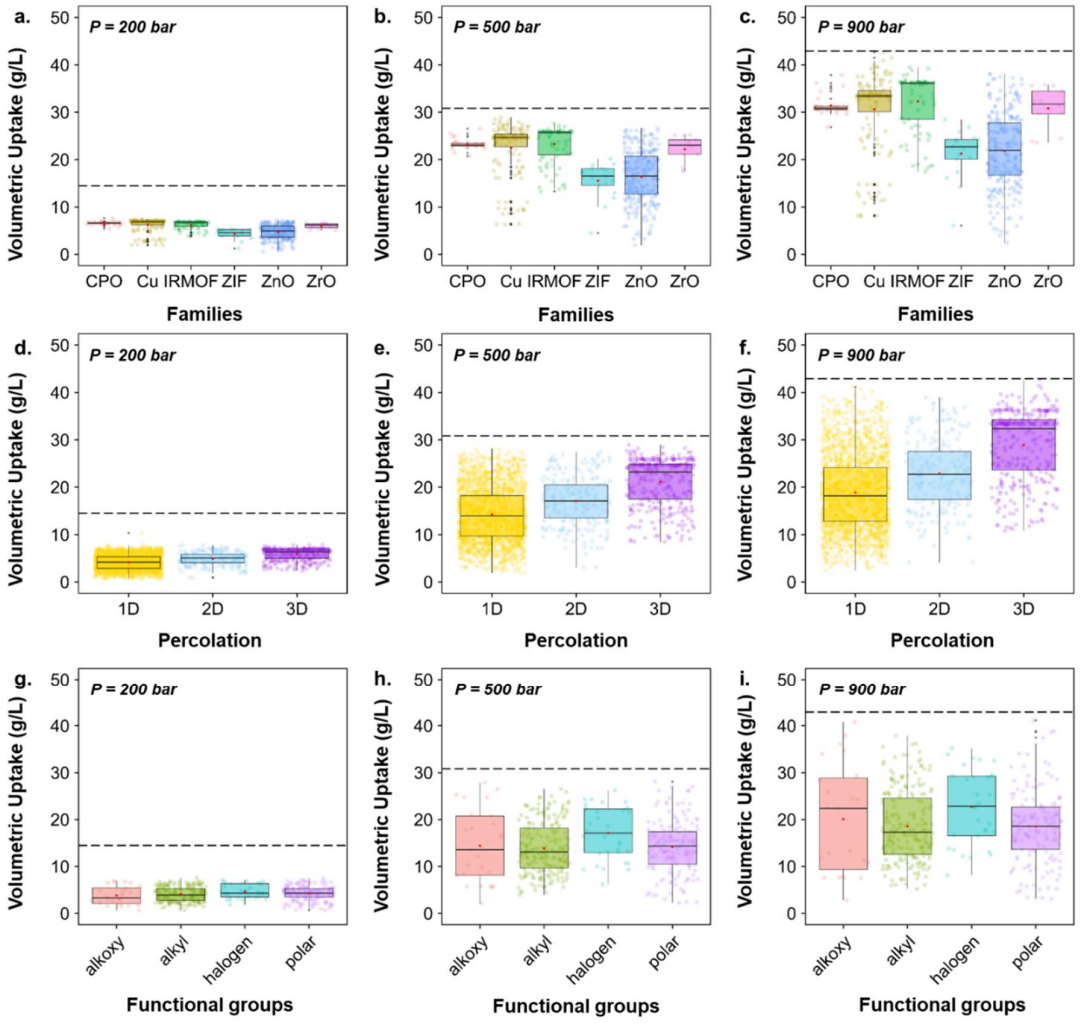


Figure 46 Quantitative characterisation of the 3D MOFs screened for hydrogen storage. Boxplots of volumetric uptake of H_2 at room temperature at 200, 500 and 900 bar versus **a.-c.** families of the screened structures, **d.-f.** percolation of the screened structures and **g.-i.** functional groups identified in the screened structures. The jittered points in the background give an idea on the number of structures considered for each boxplot. The markers represent the minimum, first quartile, median, third quartile, and maximum values, respectively. Outliers are represented by black data points. The dashed line corresponds to the volumetric uptake obtained in an empty tank.

In addition to the structure-property relationships that can be uncovered from combining simulation data and the structural data available via the CSD and the developed subsets, the tools developed here allows a better understanding of the evolution of the MOF field. **Figure 47a** shows the evolution of the hydrogen volumetric uptakes at room temperature and 500 bar for the 3D MOFs included in the CSD over the years. Each

circle represents a MOF; their size corresponds to their LCD and the colours indicate their R-factors. The yellow line traces the best-performing structure throughout time. Interestingly, the biggest jumps in terms of volumetric uptake – reaching 9.7 and 12.6 g L^{-1} – happened in 1983 and 1989, with structures BOMCUB[202] and JARMEU[203], respectively, when only a few fairly good quality structures were submitted. **Figures 47b-c** show the snapshots of these two structures: BOMCUB being an oxalate complex synthesised by Siftar and coworkers; and JARMEU being an infinite polymeric framework consisting of three dimensionally-linked rod-like segments synthesised by Hoskins and Robson. The number of structures then significantly increased in the late 1990s, with slightly higher R-factors and higher LCDs. Starting from the 2000s, the R-factors and LCDs become more varied and the highest volumetric uptake reaches a maximum of 28.8 g L^{-1}.

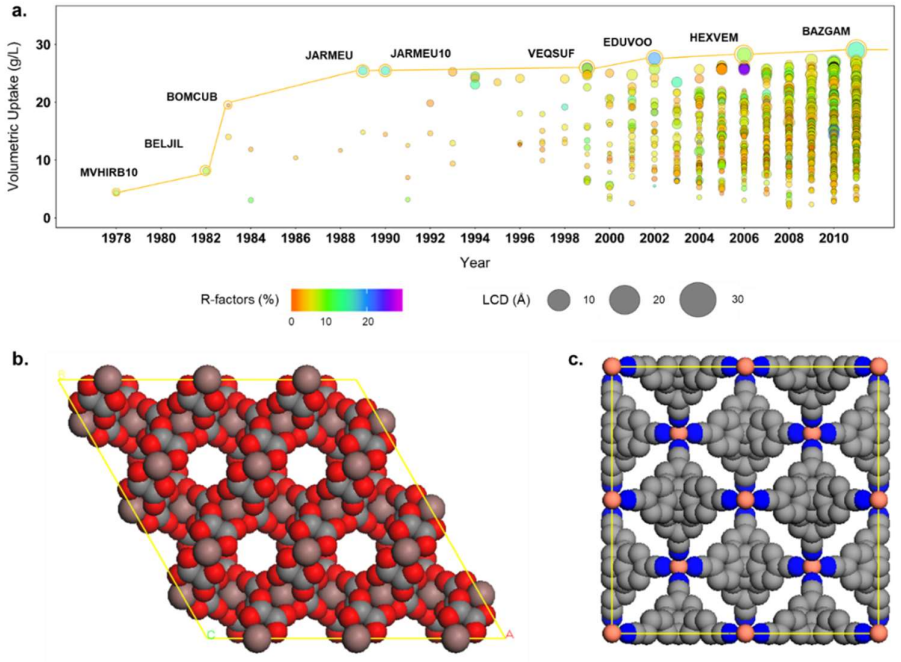


Figure 47 a. Evolution of the structure with the highest hydrogen volumetric uptake at room temperature and at 500 bar in the CSD over the years. **b.** Snapshot of a supercell of BOMCUB. **c.** Snapshot of a supercell of JARMEU. In **a.**, each circle represents a structure. The size indicates the LCD, the colour the corresponding R-factor. Each new best performing structure is highlighted with a yellow circle and the yellow line tracks the best performing structure over the years.The counter-ions and water molecules were removed from the snapshots for clarity. In **b.** and **c.**, carbon atoms are represented in grey, oxygen in red, nitrogen in blue, indium in brown in **b.** and copper in pink in **c.**

5.6 Conclusion

I demonstrated in this chapter the usefulness of the toolbox previously developed in Chapter 2. To the best of my knowledge, this is the only computational study of hydrogen storage at room temperature and high pressures, conditions for which a solution could greatly benefit the industry. Whilst the additional information obtained bring interesting insights into the behaviour of MOFs, the empty tank comparison suggests the experimental structures available in the CSD MOF subset are not good candidates for hydrogen storage. The strikingly absent publicly available research in these conditions may be explained by the lack of encouragement to publish negative results. I believe these should nevertheless be shared, so as to direct researchers towards other directions, and I hope this example will encourage other scientists to share all results, positive or negative, to the wider community.

6 IDENTIFICATION OF METAL-ORGANIC CAGES AND ORGANIC CAGES IN THE CSD USING TOPOLOGICAL DATA ANALYSIS

Abstract. Metal-organic cages (MOCs) and organic cages (OCs) hold a special place in the landscape of microporous materials, as they are discrete molecules non-covalently packed into solids. As rationally designable materials, the number of existing MOCs and OCs has increased in the last few years, albeit not as quickly or to the same extent as metal-organic frameworks (MOFs). I showed in the previous chapters the methods for identifying extended crystalline structures such as MOFs, and the potential of the Cambridge Structural Database MOF subset in identifying the best-performing structure for a given application and in uncovering structure-property trends. In this chapter, I demonstrate how topological data analysis, combined with supervised and unsupervised classification, is a useful tool for the identification and classification of cages. I then illustrate a possible use of such a dataset with a high-throughput screening of MOCs and OCs for xenon/krypton separation.

6.1 Contributions

The work presented in this chapter is entirely my work.

Dr Andrew Tarzia, Jelfs Computational Materials Group, Department of Chemistry, Imperial College London, UK, provided me with a list of labelled cages, as explained later in this chapter.

6.2 Introduction

Amongst the burgeoning field of microporous materials, metal-organic cages (MOCs) and organic cages (OCs) are of particular interest. Conversely to metal-organic frameworks (MOFs) and covalent organic frameworks (COFs), which are extended crystalline structures constructed from strongly bonded building blocks, MOCs and OCs are discrete individual molecules with a cage-like shape. The cage's internal cavity defines the material's *intrinsic* porosity. When packed, MOCs and OCs assemble through non-covalent interactions into a porous structure, where the packing gives rise to *extrinsic* porosity. The combined porosity of these materials justify the growing research for their applications in molecular[204, 205] or gas separations,[5, 6, 204, 206, 207] encapsulation,[208] catalysis,[209, 210] molecular sensing,[9, 211, 212] and as porous liquids.[19]

Similarly to MOFs and COFs, the modularity of MOCs and OCs can lead to a large space of possible structures and has thus attracted the attention of computational researchers. The two types of porosities – intrinsic and extrinsic – add yet another customisable dimension. Evans et al. estimated that using only small organic molecules as building blocks for cage-based porous molecular materials, there could be 10^{60} potential candidates.[213] Inspired by the success of high-throughput screenings (HTS) on extended porous materials, several research groups started to apply the same data mining methods to organic molecular materials, whether their porosity is intrinsic and/or extrinsic. McKeown et al. carried out a targeted structure search in the Cambridge Structural Database (CSD) to identify promising organic microporous crystals for nitrogen and hydrogen adsorption.[214] In particular, the authors were looking for structures that might possess enhanced microporosity compared to existing examples of microporous crystals. Therefore, to narrow their search in the database, they looked for structures 1) with densities lower than 0.9 g cm^{-3}, as the lowest density of any known microporous organic crystal was 0.96 g cm^{-3} for p-tert-butylcalix[4]dihydroquinone after water removal, 2) containing mostly aromatic rings

as these play an essential role in the structures' stability and 3) with pore diameters smaller than 10 Å, as it was shown it ensures strong gas adsorption in this particular case. Following this data sieving and after the elimination of additional structures with questionable data quality, 23 organic and metal-organic structures were retained. Among them, 3,3′,4,4′-tetra(trimethylsilylethynyl)biphenyl (CSD refcode: BALNIM[215]) was synthesised experimentally and demonstrated a BET (Brunauer-Emmett-Teller) surface area of 278 m^2 g^{-1} and the highest amount of nitrogen adsorbed at 77 K and at saturation for an organic, crystalline compound with such low molecular mass.[215] Later on, Mastalerz et al. used similar criteria to rationally build an extrinsically porous molecular crystal with flat ordered sheets self-assembled with hydrogen bonding.[216] They found that benzimidazolones were promising subunits for extrinsic porous crystalline structures with one-dimensional channels. The synthesised structure (a trisbenzimidazolone, CSD refcode: DEBXIT) showed an exceptional BET surface area of 2,796 m^2 g^{-1}. Moving on from single searches to larger datasets, Evans et al. derived the first organic porous molecular crystals database (oPMC).[217] From the CSD (version 5.35, including updates up to March 2014), the authors used ConQuest to look for 1) organic structures with 2) densities lower than 2 g cm^{-3}, containing either 3) only one residue (therefore removing co-crystals) or 3') more than one symmetry-independent molecule. To this search, they excluded 1) disordered structures, 2) structure data solved from powder diffraction methods, 3) a list of structure types including organic polymers, amino-acids, peptides and complexes. From the initial obtained dataset of 160,000 structures, entries without explicit hydrogen atoms were removed, leading to 156,333 candidates. Among these, 16,000 were found to be porous to helium. However, a significant number of structures presented unphysically large pores that could potentially lead to mechanically unstable structures. To remove these, molecular mechanics simulations were performed to optimise the geometry of the crystals. 481 final organic porous molecular crystals were eventually retained to form the oPMC. These include well-studied structures as well as previously unknown ones. The authors also demonstrated the possible structure-properties trends identifiable with such a dataset. In particular, they applied support vector machines on oPMC to show that descriptors related to molecular size – such as van der Waals surface – are the most important factors in predicting the structures propensity to form structural voids. While the previous studies did not specifically focus on intrinsically porous materials, Miklitz et al. built the Cage Database (CDB), which contains organic cages, cucurbiturils,

cyclodextrins and cryptophanes, for xenon/krypton separation purposes specifically.[218] Starting from about 120 structures identified through a literature review, 41 structures were first retained after visual inspection. 26 of them were then found to have pore sizes suitably close to the diameters of xenon and krypton, after which only the structures with the highest host-guest binding energies and the relative xenon/krypton binding energies were kept, leading to 12 potential candidates. Conversely to the previous studies, this screening focused on the analysis of single host molecules, rather than the solid state structure of the material. The experimental adsorption isotherms of xenon and krypton were then measured for the selected materials at 1 bar and 298 K. It was found that the cage molecule Covalent Cage 3 (CC3) remained the best performing structure for this task, theoretically and experimentally, as previously reported.[204]

While the computational organic molecular materials field has significantly grown, the metal-organic equivalent seems currently non-existent. Yet, the field is growing, as demonstrated by the increasing number of reviews tackling a significantly diverse range of MOCs.[8, 219-228] Similarly to the previously studied MOFs and COFs, armed with the CSD tools, I would like to answer the question: how many MOCs and OCs are there?

Conversely to MOFs and COFs, however, cages are discrete molecules. Extended structures were identified in the CSD using a combination of i) substructure search based on the most common ligands and clusters linkages and ii) a search for keywords tagging polymeric structures specifically. The latter captured the essence of extended structures and significantly reduced the search space for i). However, simply changing the second criterion to non-polymeric structures significantly enlarges the search area without getting us any closer to cage-shaped molecules. Indeed, a search for non-polymeric organo-metallic structures with 3D coordinates determined leads to 447,336 hits, and the same search for organic structures returns 442,503 candidates (CSD version 5.41 with updates up to November 2019). While the specific linkage between the organic and metal subunits can still be described with ConQuest, there is no simple specific keyword to capture the shape of cages. In addition, the lack of clear definition of cages makes their automatic identification even more difficult. The International Union of Pure and Applied Chemistry defines cage compounds as ”polycyclic compound[s] having the shape of a cage”,[229] which, similarly to the MOFs’ case,[43] translate into yet another tautology. While it seems widely assumed and accepted that cages should contain cavities, it remains unclear when a cage should no longer be

considered a cage: how closed or open can these cavities be, in order to be labelled as a cage? While certain structures are undoubtedly cages (3D) and others rings (2D), there is a wide range of structures in between these two extremes. **Figure 48** shows an example of each.

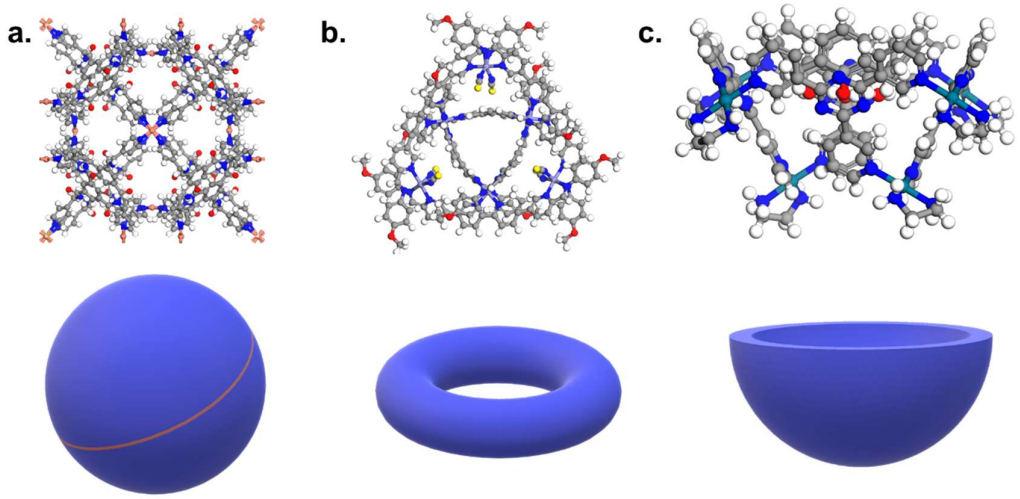


Figure 48 Examples of **a.** a cage, **b.** a ring-like structure, **c.** a bowl-shaped structure. The following structures are represented on top (CSD refcodes): **a.** CIYWOX, **b.** AVELIY, **c.** BOGYUT. Schematic of the corresponding shapes are provided at the bottom.

What is certain is that cages, from a structural point of view, whether 3D or 2D, should contain some kind of hole in which another molecule can fit, at least partially. To capture the presence of this hole – cavity in 3D or window in 2D, I chose to use a well-established data analysis tool called topological data analysis (TDA).[230] TDA is a mathematical method that studies the shape – or topology, in mathematical terms – of big data. In particular, the persistent homology theory[231] enables us to identify holes and clusters of data points.

6.3 Persistent homology

Persistent homology identifies the features that are the most spatially representative in a point cloud. For this, a nested family of *simplicial complex* – e.g. a set of points, line segments and triangles forming a graph where the segments and triangles represent the relationships between the points – is first obtained once an appropriate metric distance is chosen. This family is called *filtration*. **Figure 50a** shows an example of a filtration

built on a point cloud. To compute this family, a bead of a given radius is placed on each of the data points and the starting time $t = 0$ is recorded as their birth time. Each bead represents a feature, e.g. each individual data point at $t = 0$. The Betti number of a feature refers to its dimension, so the individual points have a Betti number equal to 0. From there, the beads are grown stepwise. When two beads merge, a new feature – the merging of these two beads – is created, while the previous individual beads are destroyed. The creation of this new feature marks another birth, accompanied by the death of the previous beads. Notice that, if these two beads were of Betti number 0, merging them means connecting them with a line, thus creating a feature of Betti number 1. The beads continue to be grown until the chosen stop time of the computation is reached. If the algorithm gave enough time for the beads to grow, the last remaining features are the most persistent ones. **Figure 49a** shows the process of growing beads for points distributed as a circle, in four time steps. At t_0, the algorithm has not started yet, and the beads have not started growing. At t_1, some of the beads have already merged. At t_2, most of the beads have merged and finally at t_3, the beads have filled the circle – or hole in the 2D space, referred to as windows in the case of porous materials. The same procedure in 3D enables the identification of cavities, these correspond to Betti number 2.

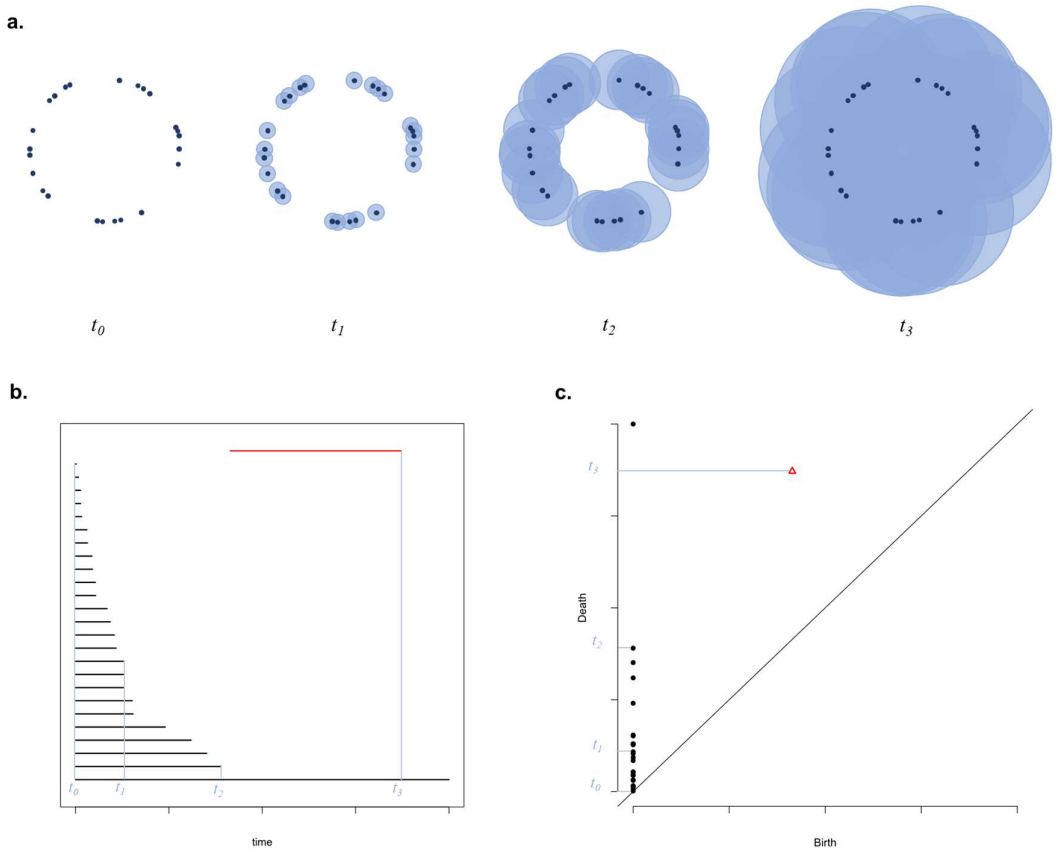


Figure 49 Determining the persistent homology of a set of points distributed as a circle. **a.** Growing beads on the point cloud, **b.** the corresponding persistence barcode, **c.** the corresponding persistence diagram. Four time steps are shown: t_0, t_1, t_2 and t_3. t_0 corresponds to the situation at $t = 0$, when the algorithm has not started yet. t_3 corresponds to the moment when the circle is fully filled with the beads. The barcode records the evolution of the features as the beads gros. Each horizontal line on the y-axis represents the lifetime of a feature. The birth and death dates are recorded on the diagram on the x- and y- axis, respectively. The lifetimes and (birth, death) points are colour-coded according to the corresponding Betti numbers: Betti 0 in black and Betti 1 in red. The four time steps indicated in blue.

From the recorded births and deaths, it is then possible to represent the persistence homology with the help of persistence diagrams and persistence barcodes. In the latter, each individual feature at any given time is stacked on the y axis, in order of successive births. Their lifetime is represented on the x axis by joining the features' birth time and death time. The features corresponding to different Betti numbers are represented in different colours. The number of most persistent features at the final time gives the

number of independent features. In the persistence diagram, the births are recorded on the x axis and the deaths on the y axis. Each feature is then represented by a single point, above the diagonal line. **Figures 49b** and **c** show the persistence barcode and diagram corresponding to the identification of the circle in **Figure 49a**. The black lines and (birth, death) points correspond to Betti 0 points. The long red line and the red triangle correspond to the circle (Betti 1). The four time steps highlighted in **Figure 49a** are indicated in the barcode and in the diagram. **Figures 50b** and **d** give an example of the persistence diagram and barcodes obtained for a cage (**Figure 50c**). A few red lines (Betti 1) are significantly longer than other red lines – they correspond to the red points located at death times equal to about 8, high above the diagonal. They signal the presence of large windows. The blue lines and blue points correspond to Betti 2 features. Importantly, the most persistent feature is indicated by the only long blue line left at the end of the calculation. This blue line is translated into one distinct blue point at a death time of about 10, and well above the diagonal. This point corresponds to the cavity.

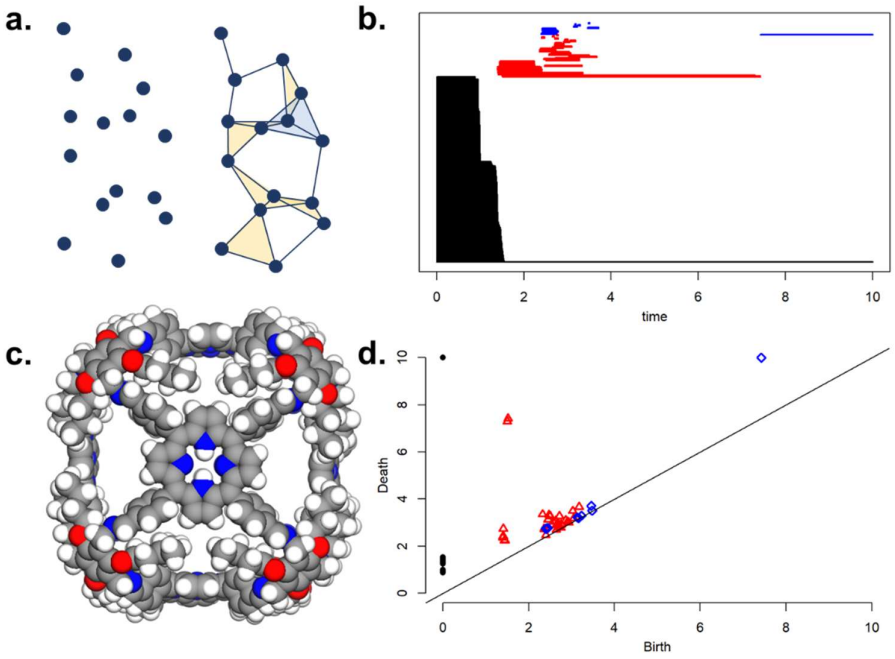


Figure 50 a. A simplicial complex built on a point cloud. **c.** Example of a cage, CSD refcode QUFYIB. Persistence **b.** barcode and **d.** diagram obtained for **c.** Each horizontal line on the y-axis of the barcode represents the lifetime of a feature. Its birth and death dates are recorded on the diagram. Their corresponding Betti number is indicated by their colour: Betti 0 (black), Betti 1 (red) and Betti 2 (blue).

While persistence diagrams and barcodes are the most intuitive representations of persistence homology, they are often not readily useable for further comparative data analysis, as each structure has a different number of (birth, death) points. This is when persistence landscapes come in handy.[232] A persistence landscape takes as input the previously obtained persistence diagram, and turns it into a set of functions as illustrated in **Figure 51**. It is then possible to choose a fixed number of points from this set of functions to represent a given set of structures. Putting the chosen points into vectors of same lengths allows us to then apply machine learning to all the structures, as I illustrate later on with the unsupervised and supervised classification of cages. Persistent homology has been applied in the field of nanoporous materials on several occasions. Lee et al. used TDA to analyse the pore shapes of zeolites and their impact on the adsorption performances.[233] Later on, TDA-based descriptors were used to predict with machine learning the performance of structures of similar shapes.[234] Moosavi et al. used persistent homology to define the geometry landscapes of porous molecular crystals from the crystal structure prediction datasets.[235] Three molecules were chosen and the shapes of their various packings studied. The different types of packings identified were mapped to their corresponding lattice energies, thereby revealing the best-performing structures. Machine learning was then applied using these geometric landscapes as descriptors, and performed remarkably well for the prediction of methane storage.

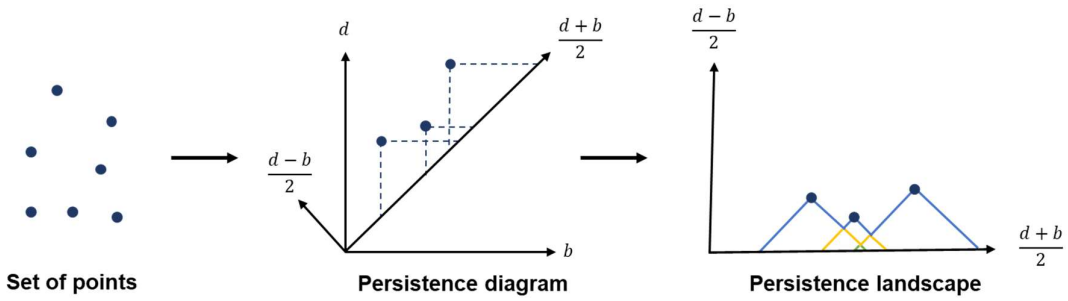


Figure 51 From data points to persistence diagram to persistence landscape.

6.4 Cage identification

I aim to apply machine learning to the persistent homology fingerprints obtained to predict whether the candidate structures are cages or not. Similarly to the CSD MOF subset, I chose to identify both 2D (rings) and 3D (cages) structures, in order to keep the dataset useful to the widest audience possible. It is important to note that I am focused on identifying the presence of a single molecule with a cavity or windows, and not on the periodic structure obtained from their packing. Although interesting and essential to understanding their adsorption behaviour, the extrinsic porosity of MOCs and OCs is beyond the scope of this study.

While the problem for MOCs and OCs is the same – identifying their defining shapes, the starting point for these two types of structures is different. As previously explained, there are no known datasets of experimental MOCs in the literature, whereas some experimental OCs have been extracted already.[218] In particular, I used in this chapter a list of known experimental OCs kindly provided by the Jelfs Computational Materials Group, Department of Chemistry, Imperial College London, UK. This existing dataset of OCs was obtained by looking for known author names in the CSD and consists of 929 2D and 3D structures. Therefore, two distinct workflows were used for MOCs and OCs, as highlighted in **Figure 52**. While MOCs underwent unsupervised classification, OCs were determined with supervised classification. In addition, the large amount of discrete organic or metal-organic structures in the CSD encouraged me to reduce the search space and computational time by first carrying out ConQuest searches for potential MOCs and OCs candidates. The CSD version 5.41 with updates up to November 2019 was used. Once a list of potential candidates was obtained, the structures were further processed with the CSD Python API:

- Many entries have either additional solvent molecules or multiple identical cages. Both cases add unnecessary noise to the TDA analysis, and only the heaviest weight component corresponding to the cage of interest was kept.
- Although rare, some entries are fully linear. Therefore, an additional check made sure that at least one atom is part of a cycle. Note that “cycle” here includes any closed path from a given atom to the same atom.

The fractional coordinates of the cleaned structures were then extracted. TDA was performed on each structure using the Python GUDHI module.[236] The Vietoris-Rips

complex was used to build the simplicial complex. This complex is a set of points built such that the distance between two points is less or equal to a given alpha (see **Figure I1**in Appendix I. The maximum value of alpha (`max_edge_length`, see full code in the Appendix I) is provided by the user. To choose `max_edge_length`, several values ranging from 0.2 to 1.6 were attempted on randomly selected structures. The resulting persistences were compared. A large maximum value such as 1.6 considerably slows down the computation of the complex, while a low value such as 0.2 shows a more significant difference from the persistences obtained with higher values. Any value between these two extremes returned identical results and 0.8 proved to be a good middle-ground. Since I am interested in the Betti 1 (windows) and Betti 2 (cavities) features, I obtained landscapes for each of these two dimensions for each structure. 2,500 points were then regularly sampled from each of the Betti 1 and Betti 2 landscapes, totalling 5,000 fingerprint points. More details specific to each type of structures are given below. In the case of MOCs, an additional noise removal step was added to the workflow. As I explain later in this Chapter, some noisy structures were found and their presence hampered the clustering algorithm. The Python scripts I wrote are provided in the Appendix I.

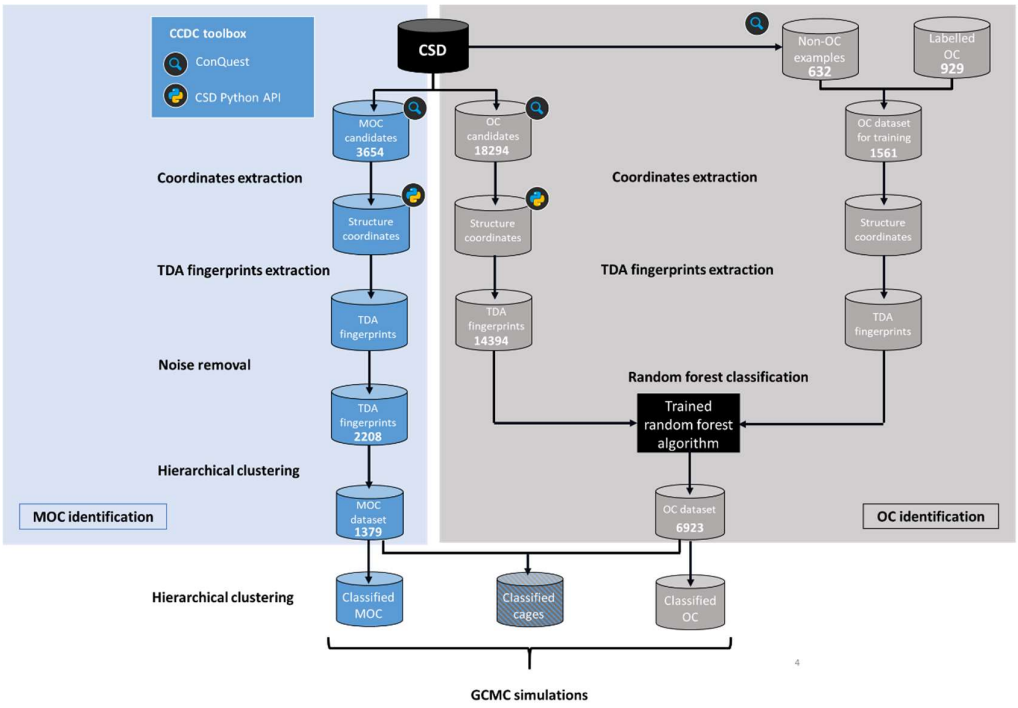


Figure 52 Workflow for the identification and classification of MOCs and OCs in the CSD.

6.4.1 Metal-organic cages

6.4.1.1 Data preparation

I identified the most common types of cages synthesised by the largest MOC groups (Ward, Hardie, Clever, Nitschke, Raymond and Fujita)[228, 237-242] and built simple "must-have" criteria describing the linkage between their organic and metal parts with ConQuest. Six main groups were identified. As the goal of these criteria is only to reduce the search space, they were not fine-tuned to match specific cages. **Figure 53** below gives a summary of the different linkages, criteria and hits obtained. In addition to these criteria, the filters "3D coordinates determined", "not polymeric" and "only organometallics" were used. Note that, conversely to MOFs, the absence of polymeric bonds means there are no multiple ways to represent the same substructure. The following paragraphs detail each of these criteria.

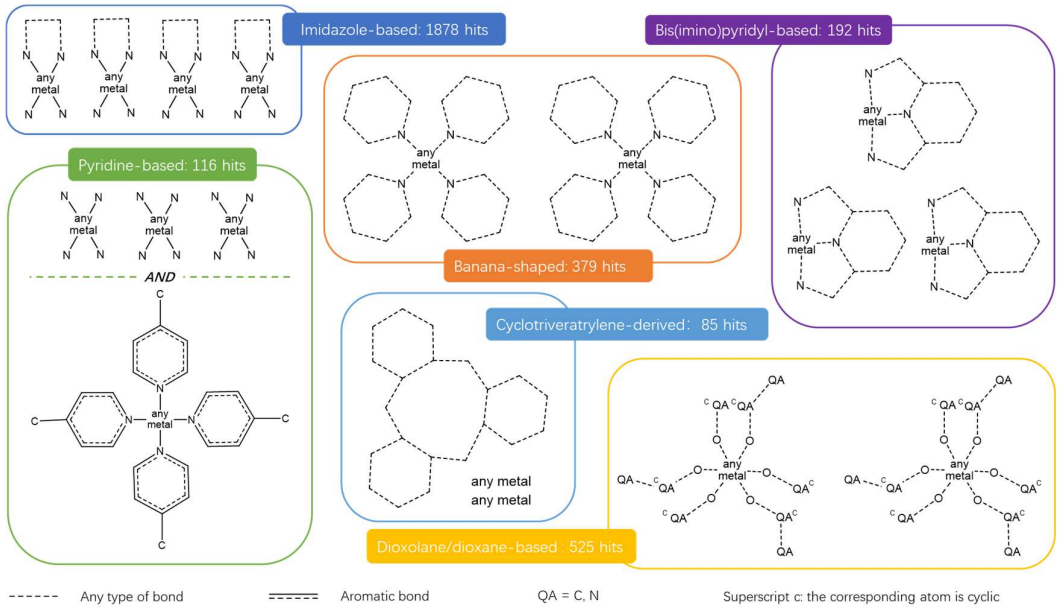


Figure 53 Quick "must-have" criteria drawn in ConQuest for some common cages. The dotted lines refer to "Any" type of bonds. QA = C or N. Superscript c: the corresponding atom should be cyclic. When atoms are not explicitly indicated, they correspond to C atoms.

6.4.1.1.1 Imidazole-based cages

The imidazole-based cages describe structures where the metal atoms are connected to the organic ligands via at least four nitrogen atoms, two of which should be part of an imidazole. As most targeted cages have at least four metal atoms, four identical units of such atoms connected to an imidazole are repeated. **Figure 54** gives three examples of structures obtained with this query. Note the variety of shape: EHIHIN is a tetrahedral cage, while LAVMOM has the shape of a funnel and ZULJAT is a helicate. 1,878 hits were obtained from this search.

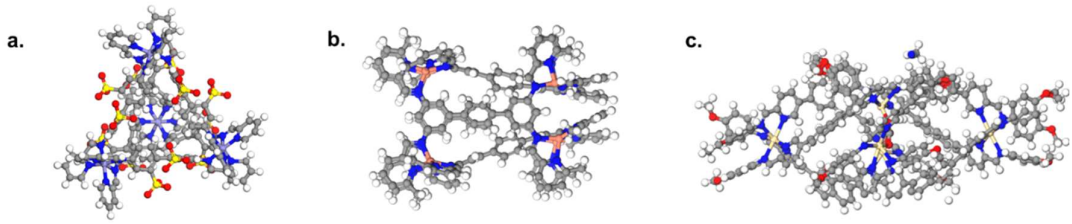


Figure 54 Examples of structures obtained with the imidazole-based query. CSD refcodes: **a.** EHIHIN, **b.** LAVMOM and **c.** ZULJAT.

6.4.1.1.2 Pyridine-based cages

The pyridine-based query describes structures where the metal atoms are connected to four pyridine compounds, each of which are then connected to a carbon atom. In addition, each entry should have at least three metal atoms. The green dashed line in **Figure 53** separates the queries above and below, meaning only one of these pyridine units is necessary. These two queries should be combined in ConQuest as an AND statement. **Figure 55** shows two examples of structures targeted with this query. 116 hits were obtained from this search.

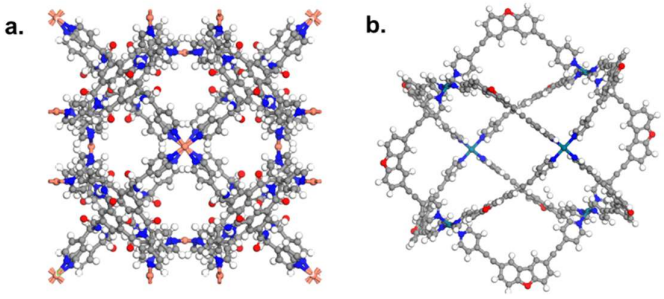


Figure 55 Examples of structures targeted with the pyridine-based cages query. CSD refcodes: **a.** CIYWOX and **b.** COWBIA.

6.4.1.1.3 Banana-shaped cages

The term "banana" was coined by Han et al. to describe the shape of the ligands, and not actually the overall cage. For the sake of simplicity, I will refer to these structures as banana-shaped. An example of such a cage is shown in **Figure 56a**. Other non-banana-shaped cages can also be found with this query; an example of a spherical cage is given in **Figure 56b**. 379 hits were obtained with this search.

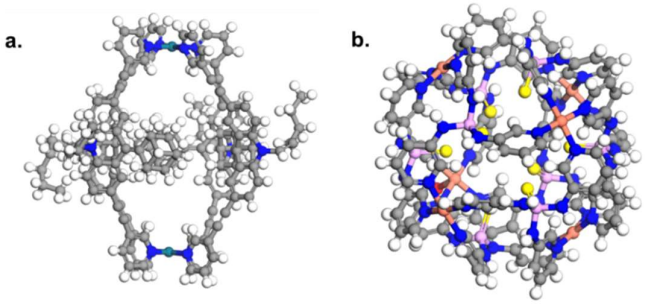


Figure 56 Examples of structures targeted by the banana-shaped query. CSD refcodes: **a.** ALEPEO, **b.** AGEMAD.

6.4.1.1.4 Bis(imino)pyridyl-based cages

The bis(imino)pyridyl-derived query describes structures where the metal atoms are part of a group containing two imidazole units that share the metal-nitrogen bond and a pyridine unit that shares a bond and a nitrogen atom with each of the imidazole units. **Figure 57** gives two examples of cages obtained with this query. Note that ZOKDEL in **Figure 57b** is referred to by the original authors as a macrocycle. The presence of a hollow in the macrocycle means it qualifies as a cage, in the case of our definition. 192 hits were obtained with this search.

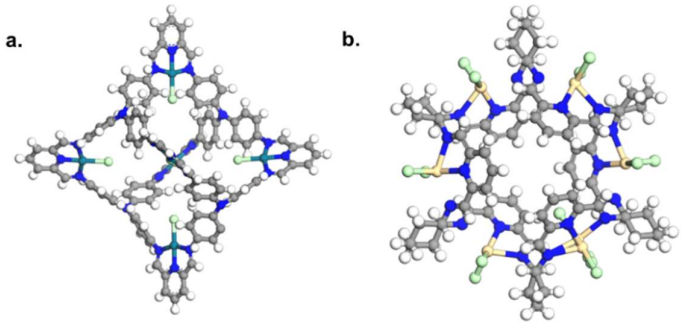


Figure 57 Examples of structures obtained with the bis(imino)pyridyl query. CSD refcodes: **a.** VOMGOW, **b.** ZOKDEL.

6.4.1.1.5 Dioxolane/dioxane-based cages

The dioxolane/dioxane-based query addresses the case of structures where the metal atoms are connected to either a 1,3-dioxolane or a 1,3-dioxane, as well as variants of these heterocycles where certain carbon atoms can be replaced with nitrogen atoms. **Figure 58** shows three examples of cages of different shapes – cylinder (ADODUS), helicate (ANITAT) and tetrahedral (BOBZUP) – obtained with this query. 525 hits were obtained with this search.

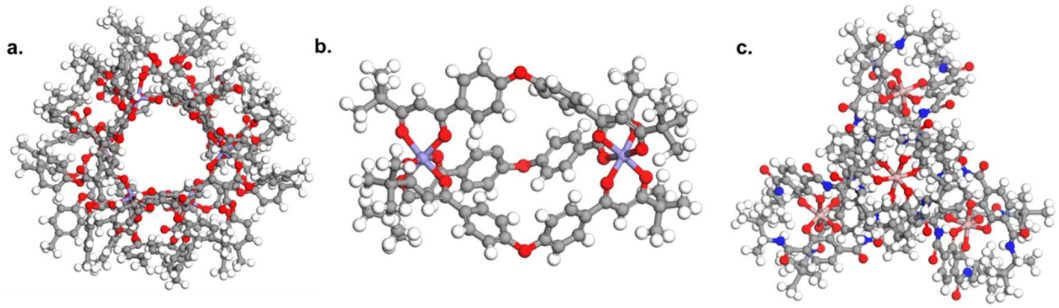


Figure 58 Examples of structures targeted by the dioxolane/dioxane-based query. CSD recodes: **a.** ADODUS, **b.** ANITAT and **c.** BOBZUP.

6.4.1.1.6 Cyclotriveratrylene-derived cages

This query tackles specifically the emerging field of cyclotriveratrylene-derived coordination cages. As the essence of these cages lies in their organic ligand, the query consists in the description of the cylotriveratrylene ligand, accompanied by the presence of at least two metal atoms. **Figure 59** gives two examples of such cages with different shapes. These cages are prone to structures with multiple cavities. **Figure 59c** gives an example of such a structure, where two cages, each with two distinct pores, are linked via an organic ligand. 85 hits were obtained with this search.

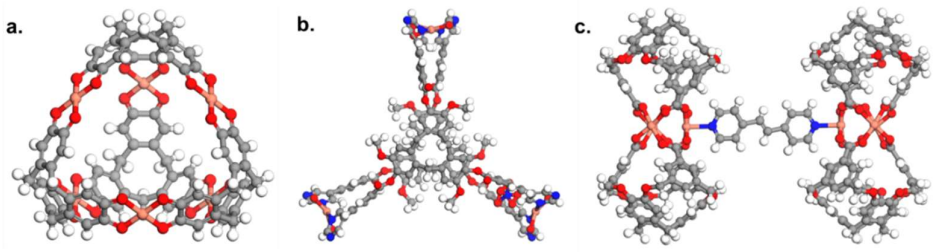


Figure 59 Example of cages obtained with the cyclotriveratrylene-derived query. CSD refcodes: **a.** ATOXIR, **b.** UTADOJ and **c.** EHEJAD.

6.4.1.1.7 Large cages

Some cages are too large and do not have an assigned 2D chemical diagram, which means a substructure search in ConQuest will miss them. However, these structures have the word "exceeded" in their textual description. This search returned 612 hits.

The combination of the presented queries led to a total of 3,654 structures. Visual checks revealed a large number of questionable structures such as AHABOA and BOYJOP (see **Figures 60b-c**). These structures have the shape of single and quadruple grids, respectively, in addition to being in large numbers. Structures in the shape of AGAPAA (**Figure 60d**) are also in large numbers. These structures are usually not labelled as cages in the literature but qualify structurally – mathematically – as cages. I obtained their persistence landscapes and after a preliminary unsupervised classification of the candidates including these unusual structures, I found the latter confused the algorithm and reduced the overall classification performance. I, therefore, proceeded to remove these structures and here on refer to them as noise.

6.4.1.2 Noise removal

Given two persistence diagrams, it is possible to compute their similarity. Several different measures of similarity exist. In this work, I used the standard bottleneck distance, defined as the shortest distance d for which there exists a perfect matching between all the points from the two persistent diagrams, such that any couple of matched points are at a distance of at most d. The matching is performed on the diagonal. **Figure 60a** illustrates the calculation of the bottleneck distance: points from two persistence diagrams are shown in two colours. The corresponding matching is represented as red edges on the diagonal. The bottleneck distance corresponds to the length of the longest of these edges. I removed structures similar by 95% to the three identified noisy structures using the GUDHI module. The distance metric used in GUDHI for the bottleneck distance is the sup norm in $\mathbb{R}^2$. It was also observed that a number of structures did not contain an organic part in their main component. These structures were discarded using ConQuest, which left me with 2,194 structures.

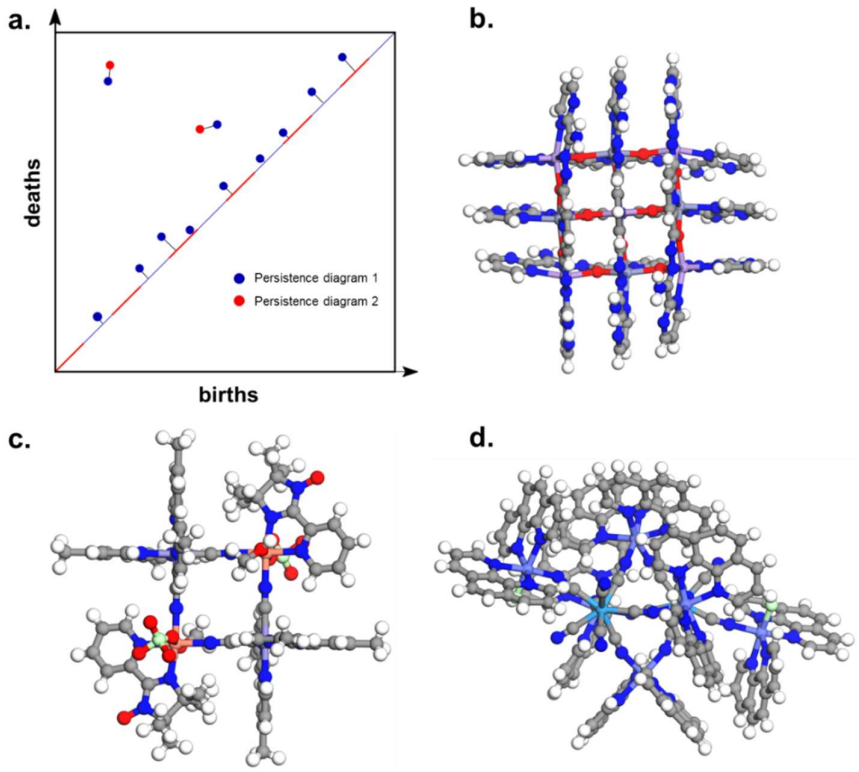


Figure 60 Removing noisy structures: **a.** illustration of the calculation of the bottleneck distance between two persistence diagrams. CSD refcodes of example noisy structures: **b.** BOYJOP, **c.** AHABOA and **d.** AGAPAA.

6.4.1.3 MOC identification

I then applied hierarchical clustering on the persistence landscapes of the filtered structures. Hierarchical clustering is a type of unsupervised classification algorithm.[243] More precisely, it is a type of agglomerative clustering, i.e. the initial clusters correspond to each different data point. The initial clusters are then merged together successively according to a specific merge strategy. The process resembles the building of a nested tree, where each branch corresponds to two merged clusters. That is why the final hierarchy of the clusters are represented as a tree – or dendrogram, where the root of the tree corresponds to the overall cluster containing all the structures. The obtained dendrogram is a useful way of visualising the similarity and relationships between the clusters, and is the reason why this algorithm was chosen for this classification. I used the hierarchy, cluster and distance modules from the open-source Python library SciPy[244] to compute the dendrograms on the 5,000 fingerprint points regularly sampled from the persistence landscapes. The chosen merge strategy was the standard Ward

linkage, which defines the distance between two clusters as the variance between them and attempts to minimise it. In this case, the Ward linkage compares the regularly sampled points to determine how different two persistence landscapes are. **Figure 61a** presents the first dendrogram obtained. The x-axis shows the different clusters obtained, and the number of structures in each cluster. For readability reasons, the drendrograms are truncated, therefore showing only a small number of possible clusters. The y-axis represents the distance between each cluster merge, as calculated by the Ward method. The horizontal black line indicates interesting cut-off distances and guides users in choosing the number of desired clusters. Above this line, the distance at which two clusters merge is indicated in blue near the merging point. This first dendrogram is composed of three main branches. Visual checks of the classification show that the algorithm was able to clearly classify 2D and 3D cages (two right branches), with very few cases of non-cages. The first branch, however, is a mix of cages and non-cages. The corresponding data were isolated and hierarchical clustering was applied again. The resulting dendrogram is presented in **Figure 61b**. There are again three main branches, the middle one being composed of 2D and 3D cages. The outer branches are however still composed of a mix of cages and non-cages. These clusters were then further classified. 48 clusters were extracted from this clustering for each of the two branches. The corresponding dendrograms are presented in **Figures 61c-d** and truncated at 24 clusters for readability. Each of the smaller clusters was then visually checked. For each cluster, if the number of cages was higher than 60%, it was considered as composed of cages. However, if less than 40% consist of cages, the whole cluster was considered as composed of non-cages. Using this decision method, I estimateed the accuracy of the overall method to be 94.7% for the 1,377 structures labelled as cages. 112 structures were not classified, as they belonged to clusters with similar numbers of cages and non-cages. Among the structures labelled as non-cages, 26 were false negatives.

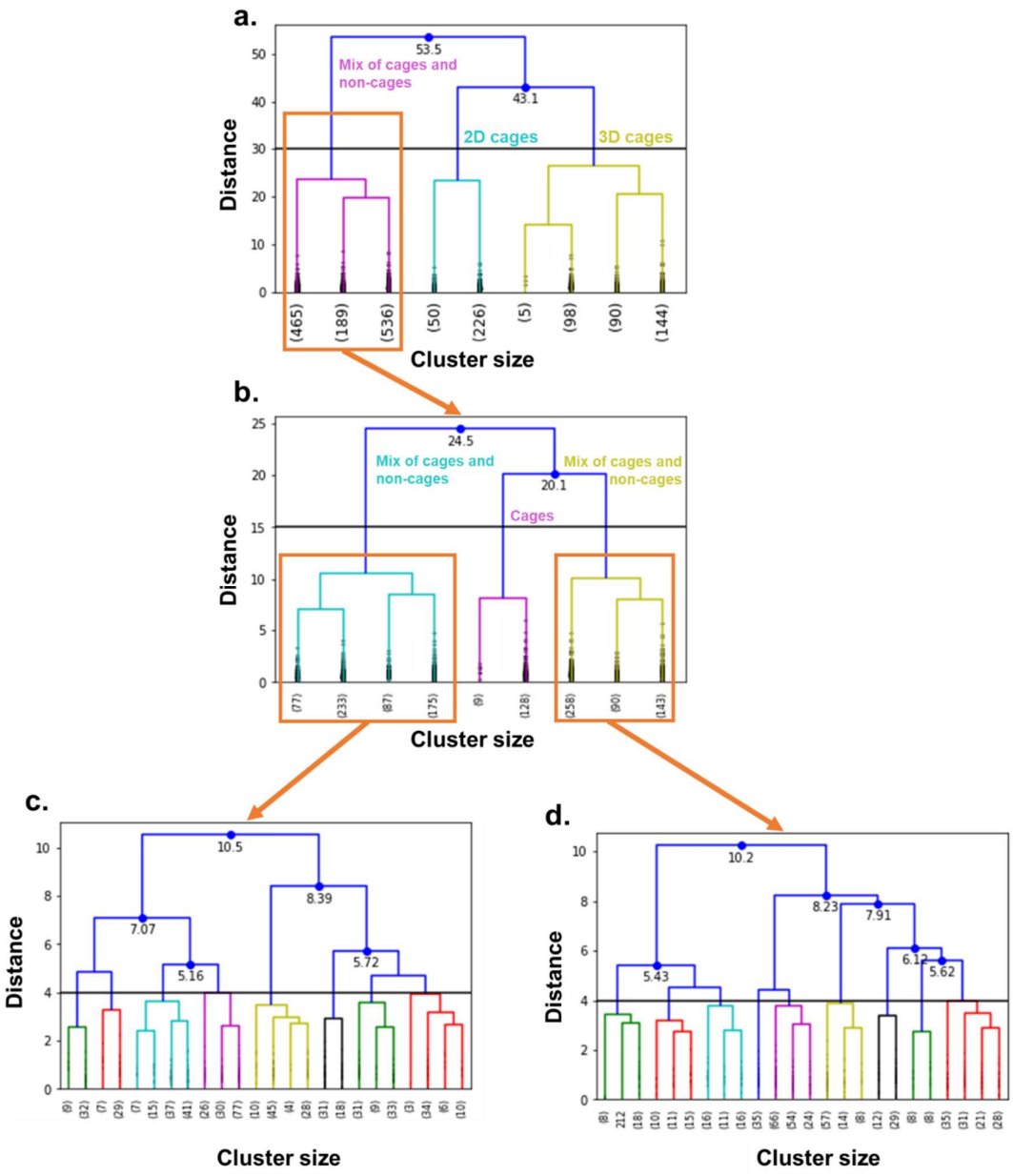


Figure 61 Truncated dendrograms of the hierarchical clusterings used to identify MOCs. Each structure is represented by 5,000 regularly-spaced fingerprint points sampled from their persistence landscape. The distance between these sampled points is then computed using the Ward linkage. The black horizontal line indicates interesting cut-off distance values.

6.4.2 Organic cages

6.4.2.1 Data preparation

Similarly to MOCs, I gathered potential OC candidates using ConQuest. A few main families of 2D and 3D cages were first identified based on literature reviews.[6, 9] Quick general queries were then designed to capture most of them, without any fine-tuning. The filters "3D coordinates determined", "not polymeric" and "only organics" were used. In addition, I eliminated any structure with any metal atom. **Figure 62** summarises the main groups of organic cages and their respective number of hits. Examples of each type of OCs are given in Appendix I.

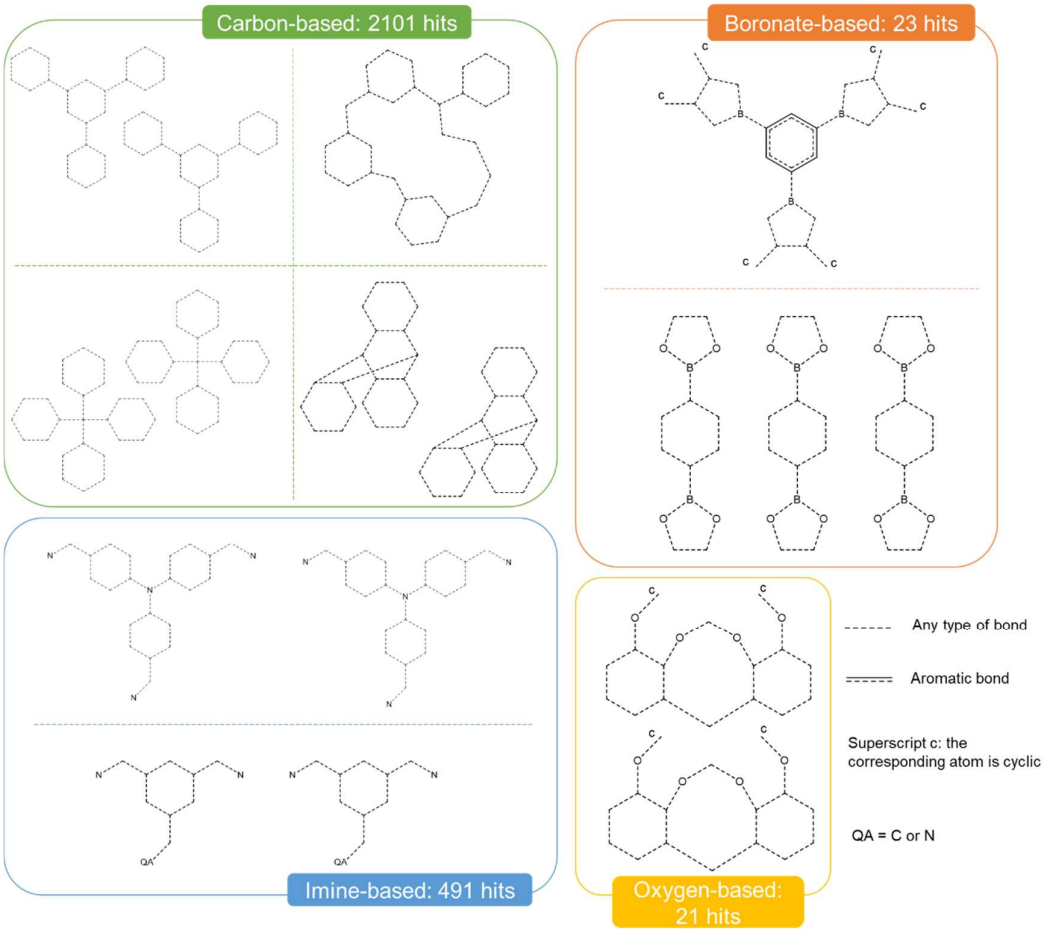


Figure 62 Quick "must-have" criteria drawn in ConQuest for some common 3D organic cages. The dotted lines refer to "Any" type of bonds. QA = C or N. Superscript c: the corresponding atom should be cyclic. When atoms are not explicitly indicated, they correspond to C atoms.

Figure 63 presents the criteria used for cucurbiturils, cyclodextrins and cryptophanes, and the corresponding number of hits returned.

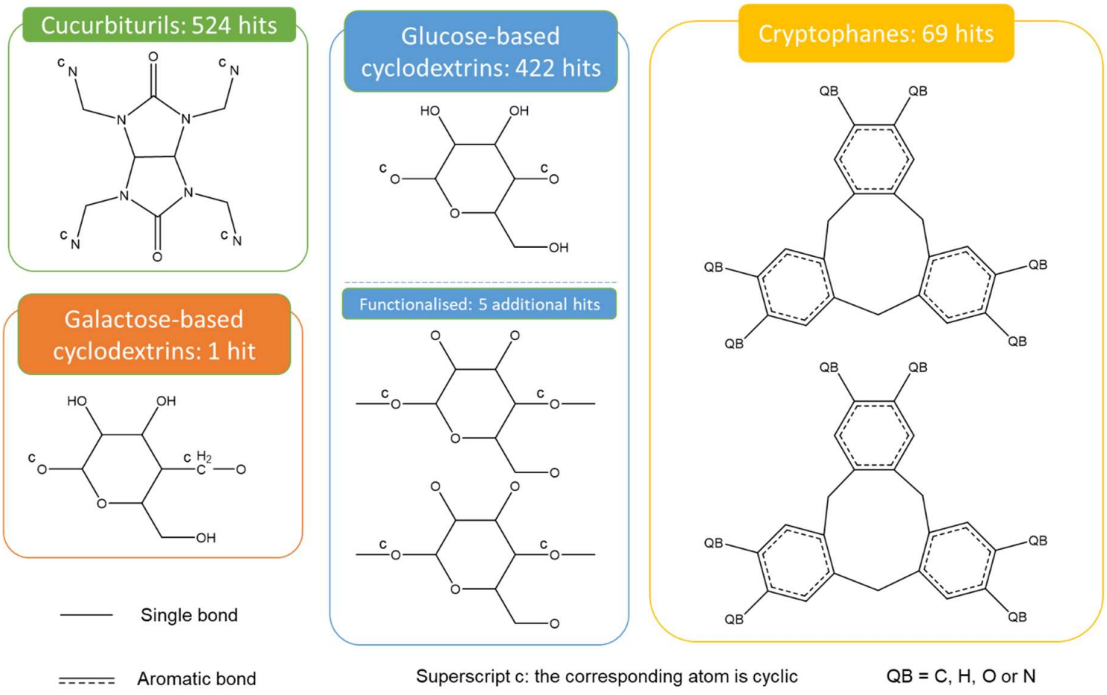


Figure 63 Quick "must-have" criteria drawn in ConQuest for cucurbiturils, cyclodextrins and cryptophanes. The dotted lines refer to "Any" type of bonds. QA = C or N. Superscript c: the corresponding atom should be cyclic. When atoms are not explicitly indicated, they correspond to C atoms.

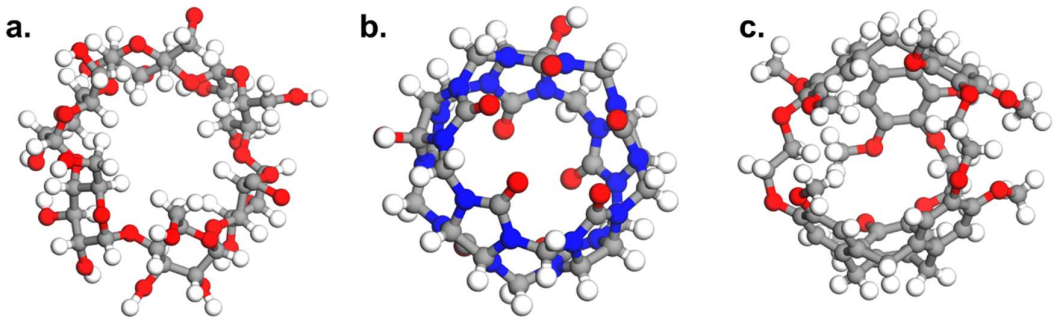


Figure 64 Examples of **a.** cyclodextrins, **b.** cucurbiturils and **c.** cryptophanes. CSD refcodes: **a.** ACDHBA, **b.** AHUPOK, **c.** XIHQAI.

The combination of the above queries led to a total of 3,746 structures. When compared with the list of 929 labelled cages, it was found that 462 structures were not included.

Although in minority, visual inspection showed these missing structures represented a wide variety of cages, not corresponding to any of the previously identified big families of OCs. For each type of missing structure, an additional query was created, until all missing structures were found. These queries are provided in Appendix J. A total number of 12,310 candidates was obtained, all of which had their persistence calculated.

6.4.2.2 Supervised classification

To prepare the training dataset, I added to the list of structures labelled as cages 633 random non-cage structures from the CSD. These structures were visually checked to be indeed non-cages. I used the RandomForestClassifier module from the Python library Scikit-learn[245] for the supervised classification of the OCs. The base random forest model with 100 estimators was found to be good enough for the task. 85% of the data was used for training and 15% for testing. The trained algorithm had an accuracy of 96%. Among the 12,310 potential candidates, 6,923 structures were labelled as cages, which is 7 times the size of the initial 929 labelled OCs.

6.5 Mapping the cages' shapes to their xenon/krypton separation performance

I have built two subsets of cages and showcased in particular the use of hierarchical clustering as an unsupervised classification method. This algorithm can be applied again on these two subsets – independently or jointly – to classify the different types of cages. I now demonstrate the usefulness of such methods when mapped together with adsorption data. To this end, I carried out a HTS of a 20/80 xenon/krypton mixture on the two datasets obtained at 298 K and 10 bar.

Xenon/krypton separation is of great industrial interest. As rare gases, they both exist in low concentrations in nature. Xenon is found at 0.087 parts per million by volume (ppmv) in the atmosphere, and krypton at 1.14 ppmv.[246] Yet both play important roles in applications ranging from medical imaging[247] to anaesthetics,[247, 248] from lighting,[249] lasers[250] to double-glazing[250] and satellite propellant.[251] Currently, a 20/80 mixture of xenon/krypton is first obtained as a byproduct of cryogenic distillations for the separation of oxygen and nitrogen in the air.[252, 253] Additional cryogenic technologies are then required to obtain pure xenon and krypton. The low concentrations mean the price of high-purity xenon is currently as high as 5,000 USD per kilogram.[249] Selective adsorption in porous materials could be a potentially cheaper alternative.

Several computational studies have already looked at the use of porous materials for xenon/krypton separation at 298 K.[204, 254-258] In particular, Simon et al. screened the Nanoporous Materials Genome, composed of over 670,000 hypothetical and experimental zeolites, MOFs, COFs and other extended structures and found SBMOF-1[259] to be a top MOF performer at 1 bar.[252] Banerjee et al. later on screened 125,000 hypothetical and experimental MOFs and identified the very same SBMOF-1 in the same conditions.[256] On the side of the discrete molecules, the tetrahedral organic cage CC3 was identified twice as the best performer at 298 K and 1 bar. These studies highlight, in particular, the importance of pore size and morphology with the selectivity of the material.[204, 252, 258] Importantly, xenon's van der Waals radius is 1.985 Å, which is larger than that of krypton (1.83 Å).[252] Combined with a deeper potential well for xenon, most structures are expected to be selective towards xenon. While the current record-holder is well-established (SBMOF-1 has a predicted selectivity of 82 versus 13.8 for CC3),[252] this is the first study that classifies cages according to their shapes and maps these onto their separation performance. In addition, previous HTS were performed at 298 K and 1 bar. In this study, I explore the cages' performance at the higher pressure of 10 bar.

6.5.1 Methods

6.5.1.1 Data preparation

I used the CIFs of the 1,377 MOCs and 6,923 OCs identified previously. Removing disordered structures using the "non-disordered" filter in ConQuest would only leave a minority of structures. As most of the disorder encountered in these structures are located in the solvent molecules, the disorder is usually removed when removing these molecules. Visual checking revealed very few structures presented missing hydrogens. Since xenon/krypton separation is based on the size of the atoms, I hypothesised structures with missing hydrogens would present higher uptakes. I, therefore, kept all the structures for the screening and checked for disorders *a posteriori* among well-ranked structures. The processing used in TDA kept only one single cage in cases where multiple cages were present in one asymmetric unit. I, therefore, also checked *a posteriori* that the well-ranked structures indeed only contained a single cage in one asymetric unit.

6.5.1.2 GCMC simulations

I used the multi-purpose code RASPA to perform grand canonical Monte Carlo simulations of the said mixture in the selected MOCs and OCs.[141] I used an atomistic model of each previously cleaned, packed structure where the atoms were kept fixed at their crystallographic positions. The symmetry operations and space groups describing the packing of the molecules are included in the original CIF from the CSD. I used the standard Lennard-Jones (LJ) 12-6 potential to model the interactions between the framework and fluid atoms. The parameters for the framework atoms were obtained from the DREIDING force field[128] and, when not available, from the Universal Force Field.[127] The Lorentz-Berthelot mixing rules were employed to calculate fluid-solid LJ parameters, and LJ interactions beyond the cut-off value of 12.8 Å were neglected. The simulation box for each structure is defined so that the cell lengths are larger than twice the cut-off distance. 20,000 Monte Carlo (MC) cycles were performed, the first third of which was used for equilibration and the remaining steps were used to calculate the ensemble averages. MC moves consisted of insertions, deletions and displacements. In a cycle, *N* MC moves were attempted, where *N* is defined as the maximum of 20 or the number of adsorbates in the simulation box. To calculate the gas-phase fugacity, I used the Peng-Robinson equation of state.[150]

6.5.1.3 Cage classification

I applied hierarchical clustering on the two datasets. I computed the dendrograms, visually identified interesting cut-off distances and chose the corresponding number of clusters or classes accordingly. I chose 16 clusters for the MOC dataset and 11 clusters for the OC dataset.

6.5.2 Results

Figure 65 presents the results obtained by combining the GCMC results with the cages classification. The figures in the left column correspond to the MOC dataset, and the figures in the right to the OC dataset. **Figures 65a** and **b** present the xenon uptake versus the selectivity of xenon over krypton, here defined as:

$$S_{\frac{Xe}{Kr}} = \frac{(\frac{x_{Xe}}{x_{Kr}})}{(\frac{y_{Xe}}{y_{Kr}})} \qquad \textbf{6-1}$$

where x_{Xe} and x_{Kr} are the molar fractions of xenon and krypton in the adsorbed phase, and y_{Xe} and y_{Kr} are their molar fractions in the bulk gas phase, here 0.2 and 0.8. Each point corresponds to a structure. The majority of the data points have selectivities between 1 and 10. As I am interested in the outstanding structures, and for clearer visibility, I masked this range of selectivites with a pink band. The size and y-positions of the structure points on the right of this band (i.e. very xenon-selective) correspond to their xenon uptakes. On the left side of the band, where structures are krypton-selective, the points' size and y-positions correspond to their krypton uptakes. The colours correspond to the data points' cluster, indicated in **Figures 65e** and **f**. For such separations, the ideal structure should be highly selective towards xenon whilst showing high xenon uptake. The highlighted area in red boxes is zoomed-in in **Figures 65c** and **d**. The images of some of the best performing structures are also presented. Interesting structures are enclosed in orange boxes. For both MOCs and OCs, the cluster colours in **Figures 65c** and **d** reveal families of structures with similar xenon uptakes and a range of selectivities (class 13 for MOCs and class 4 for OCs). Visual inspection of these structures reveals they are all CC3-type of structures. **Figures 65e** and **f** present the boxplots of the different clusters, for selectivities over 10. The jittered points behind the boxplots indicate the number of data points involved. The markers represent the minimum, first quartile, median, third quartile, and maximum values, respectively, while the red dots indicate the average value in each boxplot. Outliers are shown with additional grey dots. For structures in this range of selectivities, MOC class 13 and OC class 4 indeed stand out as families with high values of selectivity. The structure of a CC3-type cage is given in **Figure 65e**.

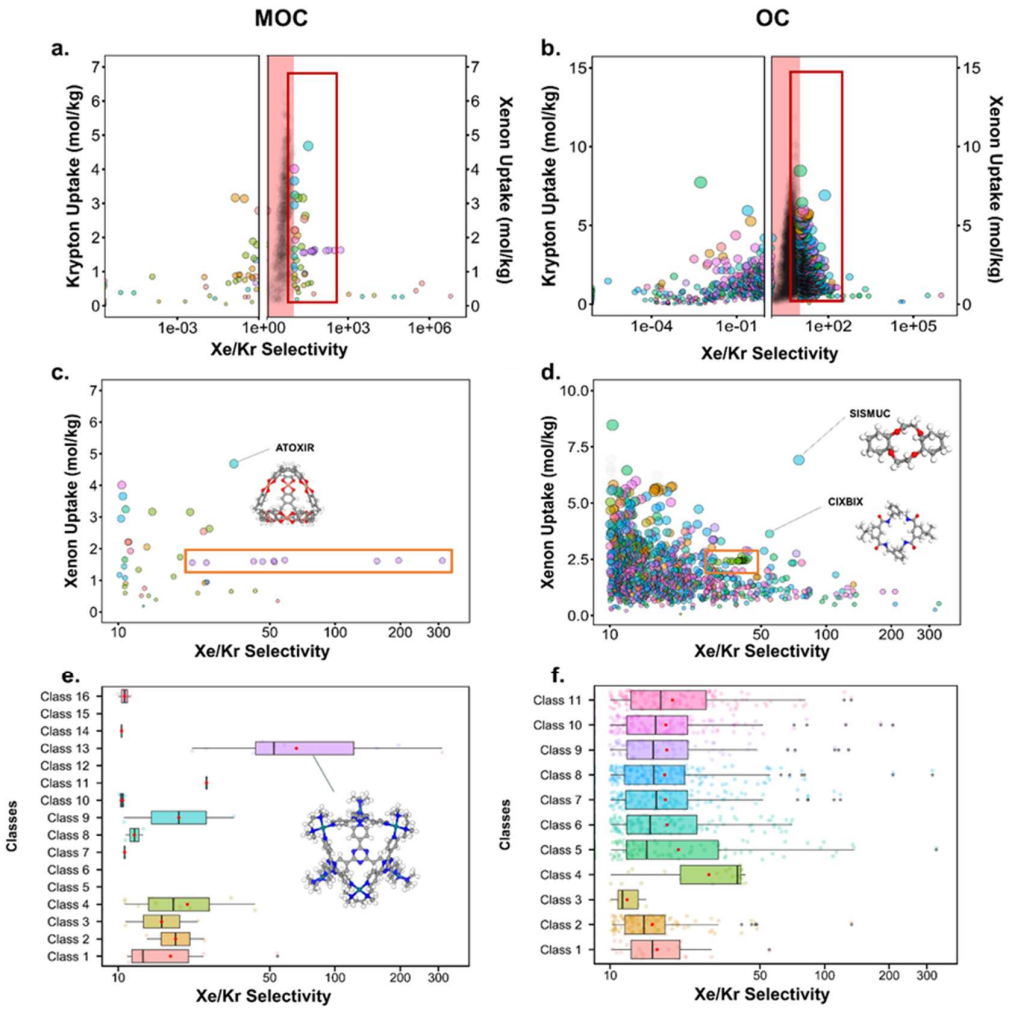


Figure 65 Xenon/krypton separation performance of metal-organic cages and organic cages. **a.** and **b.** Xenon and krypton uptakes versus Xe/Kr selectivity. Each point represents a structure. The points' colour corresponds to their cluster or class, indicated with the same colours in **e.** and **f.** The pink band hides structures with selectivities between 1 and 10. On the right side of the pink band, the y-axis and the size of the points correspond to the xenon uptake. On the left side of the band, the y-axis and size of the points correspond to the krypton uptake. The red boxes highlight areas of interest, zoomed-in in **c.** and **d.** The orange boxes indicate structures with a CC3-type shape. The best-performing structures are also provided. **e.** and **f.** show the boxplots of the different classes of materials identified for xenon/krypton selectivities of over 10. The jittered points in the background give an idea of the number of structures considered for each boxplot. The markers represent the minimum, first quartile, median, third quartile, and maximum values, respectively. The red dot indicates the mean. Outliers are represented by black data points.

Figures 65c and **d** revealed some of the best-performing structures, such as SISMUC and CIXBIX, both rings. **Figures 65e** and **f** however show the statistical behaviour of the different classes of cages. While CC3 was not predicted to have the highest xenon uptakes, its family of structures spans a range of selectivity (from ca. 22 to over 300 vs. previously reported computational value of 20.4[204] and experimental value of 14[218] at 1 bar) for a similar xenon uptake (1.6 mol kg^{-1}). This latter value is lower than the previouly calculated value of 2.69 mol kg^{-1} [204] and the measured value of 2.43 mol kg^{-1} at 1 bar.[218] These discrepancies between the simulated values are likely due to the modified DFF used to better mach the experimental results at 1 bar.[204] However, the CC3 selectivities remain of the highest, thus maintaining these cages amongst the most promising structures. The variability of the selectivity could be an artefact due to the very low uptake of krypton (close to 0), thus causing large variance. However, it is also important to note that the classification here was only applied to the cage itself, and does not reflect the extrinsic pore shapes. Similar cages can pack differently, causing more or less efficiency in their selectivity. I, therefore, looked at the crystal systems in which these structures crystallise. I gathered all tetrahedral cages and extracted their crystal system and space group information from the CSD. The results are shown in **Figure 66a**. Each point corresponds to a structure. Its shape indicates whether it is a MOC or an OC and its colour its space group. The structures are organised into rows, each of which corresponds to their crystal system. The x-axis gives the xenon/krypton selectivity. The red line indicates the previously chosen threshold of selectivity equal to 10. Structures on the left side of the red line tend to have lower selectivity and higher xenon uptakes, whereas structures on the right side have higher selectivities and lower xenon uptakes. Interestingly, structures with higher selectivities that crystallise in cubic systems are also organic, while structures with higher selectivities that crystallise in tetragonal systems are also metal-organic. These two types of structures have distinct features: most of the organic structures with higher selectivities gather around selectivity values of about 40 and uptakes of around 2.5 mol kg^{-1}. These structures – highlighted in orange – correspond to different versions of CC3 obtained under different conditions from the Cooper group.[204, 260-262] A typical CC3 structure is shown in **Figure 66b**. The metal-organic structures, however, span a range of selectivities at lower xenon uptakes of around 1.6 mol kg^{-1} and correspond to M_6L_4 structures (6 metal nodes and 4 ligands) synthesised under different conditions by the Fujita group.[263-270] These structures are highlighted in green and **Figure 66c** shows their typical structure.

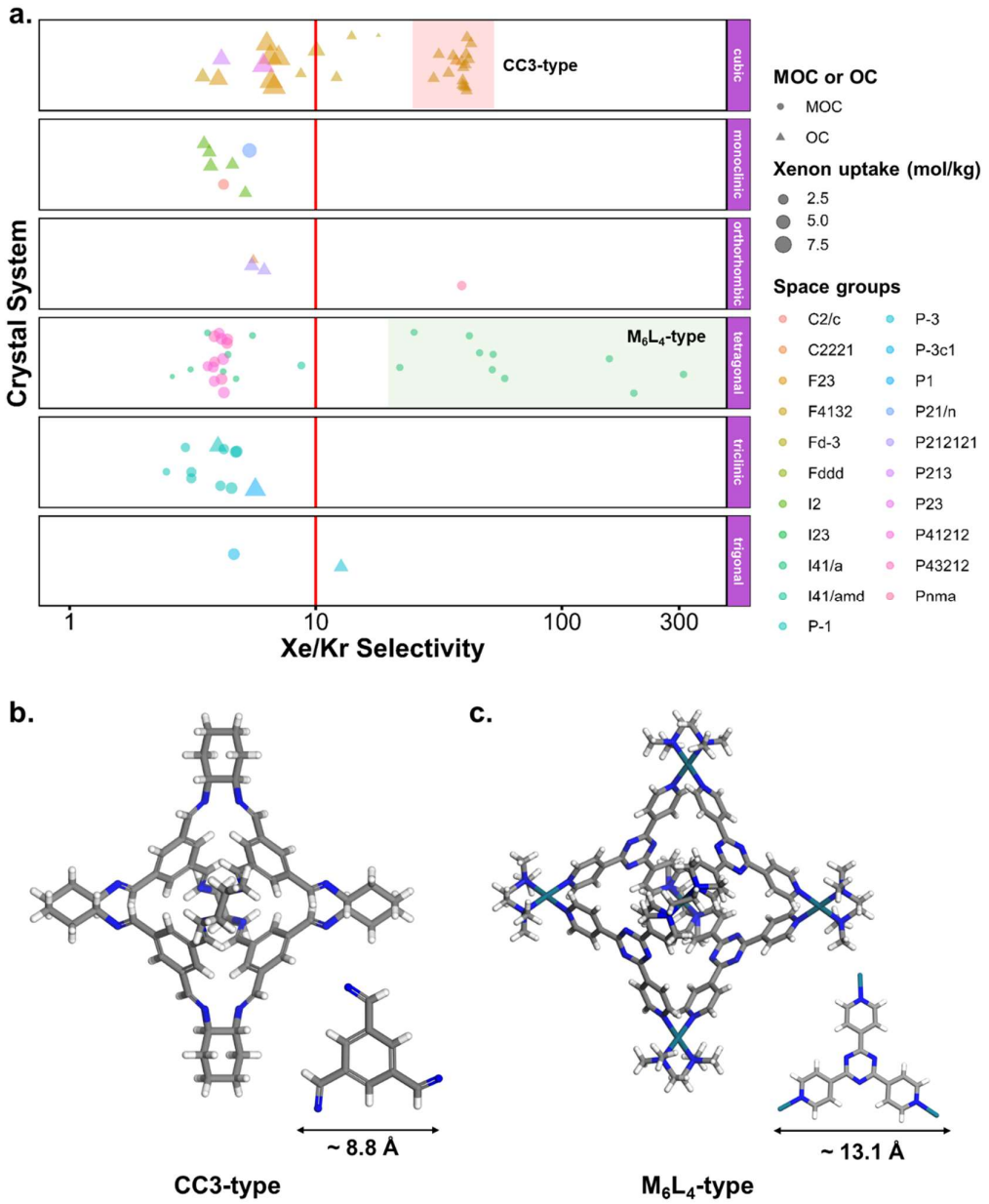


Figure 66 a. Crystal systems of tetrahedral cages and their Xe/Kr selectivity. **b.** Example of organic tetrahedral cage. **c.** Example of metal-organic tetrahedral cage. In **a.,** each point corresponds to a structure. Its color corresponds to its space group, its shape to its classification as MOC or OC and its size to its xenon uptake. The points are organised into different rows according to their crystal systems. The points are jittered in the y-axis for easier visualisation. The vertical red line indicates a selectivity of 10. The CC3-type structures are highlighted in orange. The M_6L_4-type structures are highlighted in green. One additional M_6L_4-type structure with a selectivity of 537 is not shown for clearer visualisation (CSD refcode: AJENIO).

Since the structures within each of the two highlighted groups were obtained under different conditions, why are the metal-organic cages more prone to selectivity variance? To understand this, I visually compared two M_6L_4 structures: one at the relatively lower selectivity of 25 (CSD refcode: COPPAA) and the structure with the highest selectivity (CSD refcode: AJENIO). The two structures are presented in **Figure 67**. Although the individual cages share the same ligands, metal nodes and space groups, the size of the cells and the void fraction differ. Using the Cambridge Crystallographic Data Centre software Mercury of structure visualisation and analysis,[110] I computed the surface surrounding the porous areas in both structures. The result in **Figure 67** shows that the two surfaces differ significantly in shape. While a continuous channel runs through COPPAA from left to right, this channel is cut short in AJENIO. By comparing the two structures, I found that this difference in channel morphology is due to the difference in the bending of the organic ligands. To go from **Figure 67a** to **Figure 67b**, one can imagine pulling on the ligands at their centre in their perpendicular direction. This movement is indicated in **Figure 67a** by the yellow arrows. This difference in ligand bending possibly caused the observed differences in cell lengths, leading to an overall larger cell in the case of AJENIO, as well as larger pore volumes. These structural differences seem to have a large impact on the observed selectivities: a difference of 1 to 3% in cell lengths is related to a 33% difference in void fraction and one selectivity that is 21 times higher than the other. While the exact mechanism behind the difference in selectivity could be further investigated, the main take-away from this example is that different synthesis conditions can lead to slight differences in ligand bending, which then leads to differences in the pores morphology that can have a significant impact on the calculated performance of the structures.

Such high-impact structural variations were however not observed in the CC3-type structures. There are two possible reasons for this:

1. As shown in **Figures 66b** and **c**, CC3 structures have shorter ligands which are therefore harder to bend.
2. CC3 structures crystallise in cubic systems, which provide more efficient packing and less leeway for structural variations. **Figure I6** shows the differences in packing in the two systems. This results in cages that are structurally extremely close, despite having been obtained under different

conditions. The low structural variance in turns explains the observed low selectivity variance.

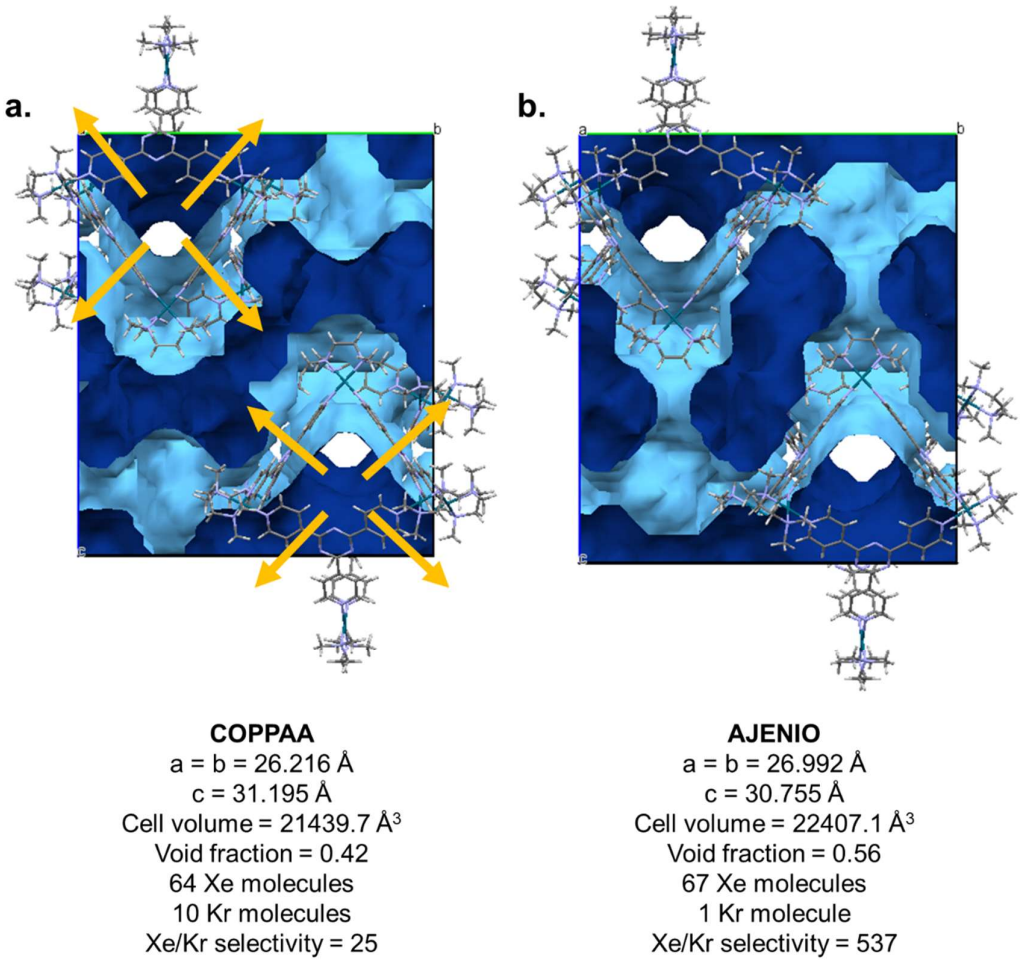


Figure 67 Comparison of two M_6L_4 structures with widely different Xe/Kr selectivity values: **a.** COPPAA and **b.** AJENIO. The blue surface maps out the porous areas, obtained in Mercury with a probe of radius 1.83 Å, corresponding to krypton's van der Waals radius. The light blue corresponds to the outer surface and the dark blue to the inner surface. The yellow arrows in **a.** indicate the bending direction of the ligands to reach the cage morphology of AJENIO.

While I was able to shed some light on the spread of selectivity values observed for M_6L_4 cages, this case study revealed how sensitive simulations can be to slight structural differences among similar or identical structures obtained under different conditions. Similar sensitivities with MOFs were shown in the third example of Chapter 3. These cases are not single point observations, and I expect such sensitivity to increase

with the potential flexibility of a structure. They show the limit of assuming the host structure as rigid in molecular simulations.

6.6 Conclusion

I have presented in this chapter the use of topological data analysis for the identification of cages in the CSD. In addition, I have demonstrated the usefulness of hierarchical clustering in the unsupervised classification of cages, as well as in visualising the structures' similarity. Using these methods, I successfully obtained the first MOC dataset and an OC dataset which expands the OC space previously known by seven-fold. Whilst the presented procedure is more complex to integrate into the CSD for automatic updates, I suggest applying random forest on persistent homology landscapes to determine whether a new structure is a cage. I illustrated the information obtained with a xenon/kryption separation simulation and confirmed the high performance at 10 bar of the CC3 cages, previously identified for their high selectivity at 1 bar. More interestingly, I found the metal-organic equivalents to CC3 (M_6L_4) and compared their respective selectivities. The sensitivity of molecular simulations to slight structural variations in the case of M_6L_4 cages showed yet again the limits of the rigid host assumption.

While the computational field of organic porous cages is growing fast, this is – to the best of my knowledge – the first extensive search of OCs in the CSD. A significant amount of work on the classification and prediction of OCs have already been produced, albeit relying on a cage-specific topology nomenclature – different from the mathematical concept of topology used in this work. Of the predicted 20 most common topologies defined by Santolini et al.,[271] 12 have been experimentally reported. Greenaway et al. took a step further by creating a hybrid computational-experimental high-throughput workflow where conventional virtual HTS was combined with robotic synthesis to discover new cages.[272] Of the 78 precursor combinations chosen, 33 cages were eventually synthesised and one previously unknown topology was discovered. Mapping the OCs dataset obtained in this work to the predicted possible cage space and the cage topologies would bring additional insight into the regions that have been explored. Extending the same cage-specific topology definitions and mappings to MOCs could not only accelerate the computational discovery of MOCs, but also provide a clear research framework early on in the development of the field.

7 CONCLUSION AND FUTURE WORK

The work presented in this dissertation aimed at building and exploring databases of porous materials for adsorption applications. Four types of materials were studied: metal-organic frameworks (MOFs), covalent organic frameworks (COFs), metal-organic cages (MOCs) and organic cages (OCs). The conclusions regarding these datasets are the following:

1. While the Cambridge Structural Database (CSD) MOF subset contained 69,999 structures at the time of its creation in 2017, it now has reached almost 100,000 in the latest update. The use of ConQuest searches integrated in the CSD proved to be extremely useful in updating the database automatically, thus maintaining its status of the most complete dataset of MOFs.
2. A toolbox for the exploration of the CSD MOF subset is now available: I extracted structural information (pore limiting diameter or PLD, largest cavity diameter or LCD, surface area, pore volume, percolation), built a new in-house algorithm for the determination of framework dimensionality, and created CSD-integrated methods for the classification of MOFs according to their family and surface functionalities. All the obtained information is available on an online multi-dimensional web tool for exploration. The ease of use of such data

visualisation can benefit both computational and experimental MOF scientist alike.

3. The framework dimensionality analysis revealed the diversity of MOFs in the subset, with 40% 1D, 29% 2D and 31% 3D structures.
4. The high-throughput screening (HTS) of the CSD MOF subset for hydrogen adsorption at the industrially sought-after conditions of 298 K and 200, 500 and 900 bar shows MOFs are not ideal materials for hydrogen storage. The lack of publications regarding these conditions indicate a reluctance to publish negative results in the scientific community. I did however identify the best-performing structure, BAZGAM, which had been previously found as one of the highest-performing structures for hydrogen storage at lower pressures.
5. By combining information extracted using the toolbox with simulated information obtained from molecular simulations, I mapped out the performance landscape of non-disordered 3D porous MOFs from the CSD MOF subset for hydrogen storage at the previous conditions. I showed that the best-performing structures for this task are either one of or a combination of the following: COP-27-like, Cu-Cu paddlewheel, IRMOFs and Zr-oxide structures, 3D-channeled structures or structures with halogen functional groups. While these structures were shown not to be ideal for hydrogen storage, the observed trends are likely to remain the same at published conditions, e.g. lower pressures.
6. The building of a COF dataset showed the lack of available experimental data in the CSD: 54 structures, i.e. $1/6^{th}$ of the reported synthesised structures in the literature. This is likely due to the difficulty in obtaining suitable single crystals for analysis.
7. I successfully built two CSD subsets of cages (MOCs and OCs) by using a combination of topological data analysis, hierarchical clustering and random forest. I then further classified these datasets with hierarchical clustering. These subsets are the biggest experimental datasets of cages to my knowledge.
8. I performed a HTS of a 20/80 xenon/krypton mixture separation on the obtained datasets of cages and identified the best-performing structures: ATOXIR for MOCs and SISMUC for OCs.
9. By mapping the obtained unsupervised classification and the results from the molecular simulations, I was able to identify a well-performing class of material,

from the same family as CC3, previously reported as the best-performing structure for this task.

In addition to these datasets, I demonstrated through several smaller case studies the following:

1. Molecular simulations can map out the adsorption process of drug molecules in MOFs. I looked specifically at the case of DCA and α-CHC in CAU-7, where the density distribution showed the drugs adsorbed initially on the walls of the MOFs, before filling up the pores.
2. Simple molecular simulations such as grand canonical Monte Carlo (GCMC) can elucidate complex MOF phenomena such as flexibility. I showed how combining adsorption isotherms at different conditions can explain the phase transitions of ZIF-7-II to ZIF-7-I, two phases of a flexible MOF.
3. I showed one possible method to investigate three chiral MOFs – CMOMs – using GCMC for chiral separations. I showed that, given the correct crystallographic data, it is possible to predict the chiral selectivity of these MOFs.
4. I also introduced the concept of reverse HTS, where a library of molecules is built to be screened inside a few given MOFs. This approach is useful to expand the use of a given structure. I thus identified four chiral molecules – 3PE, 4PE, CPBA and VB – that can potentially be separated in CMOM-3S.

The points above highlight the successful building of datasets for the four types of materials considered and the exploration of three of them for adsorption applications. I showed the power of GCMC for the analysis of adsorption phenomena, from gas adsorption to drug delivery. Through the two large-scale adsorption studies (hydrogen storage and xenon/krypton separation), the ability of HTS to identify top-performing structures was confirmed. Taking the power of HTS further, I showed the wealth of data extractable from the CSD and the interesting structure-property relationships it revealed. I believe making our results openly available through online data visualisation tools will help accelerate and simplify the search for the most relevant structures for a given application. Most importantly, these tools are useful communication means between computational and experimental MOF scientists.

While the aims and objectives of this thesis were achieved, further improvement needs to be implemented in the field for each research group's findings to be beneficial for the wider community:

1. Amongst recently discussed issues in the HTS for MOF community is that of identical and similar structures found in the CSD. Indeed, the same MOF is often represented several times, corresponding to syntheses performed in different research groups or measurements obtained under different conditions. HKUST-1 alone is present at least 50 times.[54] Whether or not users should discard similar structures depends on the exact research. The presence of numerous identical structures can indeed skew a data analysis, but the study of a variety of similar structures may also reveal interesting structural behaviours. The study of similar cages in Chapter 6 showed the variety of results one can obtain from visually similar structures. A possible solution to identifying these MOFs is to flag them in the CSD, provided a concept of similarity can be agreed on. Currently, similar structures in the CSD are assembled under the same refcode family. Entries from the same family share the same six-letter code but have different ending digits. RUBTAK and RUBTAK01 are for instance two similar entries corresponding to UiO-66. CSD editors use different techniques – including molecules overlay and powder patterns analyses – and the chemistry described by the authors to assess the similarity of two molecules. However, this classification is not straightforward for MOFs, as two identical frameworks with different or unknown guest molecules will not be considered as part of the same family. Barthel et al. proposed to exploit the structures' bond networks to determine whether two structures should be considered as duplicates.[273] In their case, after analysing 502 CoRE MOFs with assigned partial charges, 15.5% were found redundant. Bucior et al. recently developed systematic identifiers that assign to each unique MOF a MOFid and a MOFkey using automated cheminformatics algorithms.[170] These methods could not only identify duplicates but also initiate a more standardised way of naming MOFs. The latest update of CoRE MOF includes a similarity check performed with a Python script that compares the CIFs directly.[55] The StructureMatcher algorithm from Pymatgen, an open-source Python library for materials analysis, uses a similar method.[55, 274] It would be beneficial for the community, in addition of the CSD tagging, to integrate a user-friendly version of this software in the CSD.

2. As the field of gas adsorption is growing wider, and more researchers carry out similar simulations using the CSD data, it can be power- and time-saving to gather the obtained results and link them to the CSD. Several research groups have started to exploit the power of machine learning to rapidly predict gas uptakes by training their algorithms on data calculated from hypothetical and experimental databases.[275-279] Bucior et al. for instance used a combination of machine learning and a reduced number of GCMC simulations, which used less than 10% of the computational resources of a brute-force screening method.[82] The genetic algorithm used by Chung et al. was estimated to reduce the total computational time by two orders of magnitude.[80] The Adsorption and Advanced Materials group, Department of Chemical Engineering and Biotechnology, University of Cambridge, UK, is currently working with the Cambridge Crystallographic Data Centre to include i) the automatic calculation of geometrical properties, ii) calculated uptakes of a set of gases at given conditions to the MOF structures' information and iii) reasonably accurate machine learning models in order to ultimately automatically predict the uptakes of these gases for any new MOF. The gases considered at this stage are simulated at room temperature: hydrogen (5 and 200 bar), xenon (1 and 10 bar), krypton (1 and 10 bar), 2:8 xenon/krypton (1 and 10 bar), methane (0.1, 1, 5, 10 and 65 bar), ethane (0.1, 1, 10 and 20 bar), ethene (0.1, 1, 10 and 20 bar), propane (0.1, 1, 10 and 20 bar), propene (0.1, 1, 10 and 20 bar), 1:1 propane/propene (0.1 and 1 bar) oxygen, nitrogen (5 and 25 bar), carbon dioxide (1 and 35 bar), 15:85 carbon dioxide/nitrogen (1 and 20 bar), 1:9 carbon dioxide/methane (1 and 18 bar).
3. Partial charges are essential for the modelling of adsorption phenomena where electronic interactions are not negligible. The Adsorption and Advanced Materials group, Department of Chemical Engineering and Biotechnology, University of Cambridge, UK, is currently using density functional theory to compute partial charges for 3D porous frameworks in the subset. In addition, the group is developing graph neural networks for the prediction of partial charges on any new submitted structure in the database.
4. While it is important to make the CSD MOF subset data more accessible, it is also useful and interesting to compare the MOFs space covered by that dataset. In Chapter 1, I summed up the results from Moosavi et al., in which the diversity

of the MOFs ecosystem was explored.[48] This study did not include the CSD MOF subset. A thorough, quantitative comparison of CoRE MOFs and the CSD subset would be beneficial for all computational researchers.

These research topics are immediately related to my thesis. If we take a step back and look at the MOF field in general, there are many more exciting challenges and opportunities. I presented in Chapter 1 examples of HTS studies that led to experimental confirmation of the selected materials' performance. Validating the labscale feasibility of MOFs found *in silico* is only the first of many steps to bring the material to an industrially usable stage. And yet, only a minority of published HTS studies have led to experimental testing. There are many possible reasons for such few experiments-backed HTS studies, such as the lack of human resources or laboratory equipment, expensive reagents or difficult – or even unreproducible – synthesis protocols. I give here three directions that I believe the MOF field should work on:

1. *Fostering computational-experimental collaboration.* In Chapter 1, I discussed the importance of bridging the communication gap between computational and experimental researchers by improving data visualisation and lowering the barrier to using data analytics tool. I also mentioned how, within the computational community, tools such as the Automated Interactive Infrastructure and Database (AiiDA) could help scientists track their workflow and share their data processing and analysis in a more transparent – and thus reproducible – way. While such initiatives, and the much needed efforts to create open databases that follow the FAIR principle (findable, accessible, interoperable, reusable) are widely acknowledged in the computational field,[280] these efforts could make a more significant impact on the MOF field if combined with experimental knowledge. I think the next important step, from a MOF data perspective, is to connect the simulated data with the experimental data, i.e. all measured data pertaining to the characterisation of the materials. This can be done either with literature scraping, or by enriching the existing data with new measured results. I believe continuously collating computational data with experimental data will facilitate communication and collaboration between computational and experimental experts. In the long term, training machine learning on this data could also potentially point scientists directly to the most promising materials. I talk about this last point a little more later on.

2. *Towards a holistic HTS approach.* Most HTS studies focus on identifying top-performing structures based on only a few metrics, such as volumetric and gravimetric uptakes, selectivities and geometrical properties. However, to be useful at an industrial scale, MOFs need to be integrated in broader systems that have their own constraints. These bring a new – large – set of conditions that the materials need to satisfy. Rampal et al. screened a subset of 183 Cu-Cu paddlewheeled structures for the separation of CO/N_2. This is one rare example where the GCMC simulations were combined with process modelling. The latter consisted of the simulation of a simplified 3-step pressure-swing adsorption (PSA) simulation at 298 K and 1 – 40 bar, a simplified 3-step temperature-swing adsorption simulation (TSA) at 1 bar and 200 – 298 K, and a simplified 3-step TSA at 1 bar, 298 – 398 K. The analysis of the uptakes obtained from the GCMC simulations and the added metrics of purity, recovery and amount of product generated per unit of mass adsorbent calculated from the process simulations, led to the selection of four candidates, of which the monolithic form of HKUST-1 showed the best performance. This is an example where the selected material is an experimentally well-known and relatively inexpensive MOF. Unfortunately, this is not the case for many HTS, and cost and feasibility are also important aspects to take into account. While it might be difficult to accurately estimate the economics of a final MOF-system, some additional data can be included early on in the HTS, such as reagents costs, equipment needed (at a labscale first) and associated costs, estimated overall synthesis time needed and estimated humain time needed. These indicators, either included as standalone measures or combined into a new feasibility metric, can help discard any structure that would be too costly or difficult to produce. All this data is already available, albeit scattered across the web. For structures that have been synthesised, the original papers contain the procedures, and, therefore, the reagents needed and synthesis steps. In fact, Park et al. very recently extracted synthesis protocols by applying natural language processing on 47,187 papers from the CSD.[281] The mined information include the precursors, solvents and various synthesis conditions. The next step would be to connect the reagents to their costs, either by connecting the relevant databases or by scraping the web. Adding feasibility metrics to a comprehensive database – such as a

computational-experimental knowledge base, would be very useful to the MOF community.

3. *Lab digitalisation for better reproducibility.* One major issue, when it comes to synthesising a structure following a procedure written by another scientist, is its reproducibility. From one lab to another, many things can change and affect the synthesis: lab equipment, reagents providers and product batches, to name a few, but also human intervention. This means that even if a structure is synthesised, it might behave differently from the original report. One way to remedy this is automation. Not only can machines remove human biases in the steps where they are introduced, but they also save scientists from time-consuming repetitive tasks. This is all the more true when it comes to optimising an experimental procedure, where only one variable is changed at a time. Robots are particularly helpful in these situations, as they can be programmed to explore chemical spaces that would take human scientists an incomparable longer time to achieve. Even more time and resources can be saved in the long run if the robots are equipped with an active learning brain, where it chooses the next condition to test based on learned data, thus closing the loop of scientific discovery. The combination of automated high-throughput experiments and artifical intelligence in the lab is not new. In fact, King et al. introduced the concept of 'Robot Scientist' in 2009, with their robot 'Adam' who autonomously tested its own hypotheses.[282] However, most of the robots developed since then remained static and could not cater for the complexity and variety of experiments required in a chemistry lab. In addition, setting up such a robot took significant time and effort ('Adam' was born after a 7-year long process, for instance).[283] The development in 2020 of a *mobile* robot chemist by Burger et al. changed the game.[284] This time, the modularity introduced means the same robot can be more easily tailored to another lab space with different operations and equipment, and the set-up time significantly reduced. While it took Burger et al. two years to set up theirs, it is estimated that transfering the same robot using the pre-developed protocols and software should take less time.[284] Still, adapting the robots' brains to a completely different experimental goal is not completely straightforward. To help tune a robotic platform, Steiner et al. developed 'chempiler' – a program that translates experimental procedures into instructions for the robot.[285] Importantly, the synthetic protocols are codified with a chemical

> programming language based on a universal and interopable standard, meaning any procedure can be converted to a shareable code, and thus guaranteed to be reproducible. Although such significant digitalisation is not within every lab's reach, small improvements can still be made, such as switching to electronic lab notebooks to track experimental procedures or sharing "failed" syntheses – e.g. in a computational-experimental knowledge base. "Negative" results not only prevent other chemists from wasting time and resources, but also provides computational scientists with valuable data on which to train machine learning algorithms for the prediction of synthesis conditions.[286]

Improving reproducibility within the computational community and communication with experimental experts, including more metrics into standard HTS and creating constant, reproducible experimental feedback loops to the simulations – these are directions the MOF field could focus on. These suggestions are by no means straightforward to implement, nor are they the only solutions. But I think most of the building blocks are out there, and putting them together could take MOFs out of the computer and bring them a few steps closer to being studied for industry-friendly systems.

8 REFERENCES

1. A. G. Slater and A. I. Cooper, *Science*, 2015, **348**.
2. C. S. Diercks and O. M. Yaghi, *Science*, 2017, **355**.
3. H. Furukawa, K. E. Cordova, M. O'Keeffe and O. M. Yaghi, *Science*, 2013, **341**.
4. S. L. James, *Chemical Society Reviews*, 2003, **32**, 276-288.
5. M. A. Little and A. I. Cooper, *Advanced Functional Materials*, 2020, **30**, 1909842.
6. T. Hasell and A. I. Cooper, *Nature Reviews Materials*, 2016, **1**, 16053.
7. Ben S. Pilgrim and Jonathan R. Nitschke, *Chem*, 2016, **1**, 19-21.
8. M. M. J. Smulders, I. A. Riddell, C. Browne and J. R. Nitschke, *Chemical Society Reviews*, 2013, **42**, 1728-1754.
9. J. D. Evans, C. J. Sumby and C. J. Doonan, *Chemistry Letters*, 2015, **44**, 582-588.
10. M. I. Nandasiri, S. R. Jambovane, B. P. McGrail, H. T. Schaef and S. K. Nune, *Coordination Chemistry Reviews*, 2016, **311**, 38-52.
11. B. M. Connolly, D. G. Madden, A. E. H. Wheatley and D. Fairen-Jimenez, *Journal of the American Chemical Society*, 2020.
12. J. Dhainaut, C. Avci-Camur, J. Troyano, A. Legrand, J. Canivet, I. Imaz, D. Maspoch, H. Reinsch and D. Farrusseng, *CrystEngComm*, 2017, **19**, 4211-4218.
13. T. Tian, Z. Zeng, D. Vulpe, M. E. Casco, G. Divitini, P. A. Midgley, J. Silvestre-Albero, J.-C. Tan, P. Z. Moghadam and D. Fairen-Jimenez, *Nature Materials*, 2018, **17**, 174-179.
14. T. Tian, J. Velazquez-Garcia, T. D. Bennett and D. Fairen-Jimenez, *Journal of Materials Chemistry A*, 2015, **3**, 2999-3005.
15. B. M. Connolly, M. Aragones-Anglada, J. Gandara-Loe, N. A. Danaf, D. C. Lamb, J. P. Mehta, D. Vulpe, S. Wuttke, J. Silvestre-Albero, P. Z. Moghadam, A. E. H. Wheatley and D. Fairen-Jimenez, *Nature Communications*, 2019, **10**, 2345.

16. G.-B. Wang, S. Li, C.-X. Yan, F.-C. Zhu, Q.-Q. Lin, K.-H. Xie, Y. Geng and Y.-B. Dong, *Journal of Materials Chemistry A*, 2020, **8**, 6957-6983.
17. Q. Yang, M. Luo, K. Liu, H. Cao and H. Yan, *Applied Catalysis B: Environmental*, 2020, **276**, 119174.
18. L. Ma, S. Wang, X. Feng and B. Wang, *Chinese Chemical Letters*, 2016, **27**, 1383-1394.
19. N. Giri, M. G. Del Pópolo, G. Melaugh, R. L. Greenaway, K. Rätzke, T. Koschine, L. Pison, M. F. C. Gomes, A. I. Cooper and S. L. James, *Nature*, 2015, **527**, 216-220.
20. L. Ma, C. J. E. Haynes, A. B. Grommet, A. Walczak, C. C. Parkins, C. M. Doherty, L. Longley, A. Tron, A. R. Stefankiewicz, T. D. Bennett and J. R. Nitschke, *Nature Chemistry*, 2020, **12**, 270-275.
21. J. Rouquerol, D. Avnir, C. W. Fairbridge, D. H. Everett, J. M. Haynes, N. Pernicone, J. D. F. Ramsay, K. S. W. Sing and K. K. Unger, *Pure and Applied Chemistry*, 1994, **66**, 1739-1758.
22. R. Batten Stuart, R. Champness Neil, X.-M. Chen, J. Garcia-Martinez, S. Kitagawa, L. Öhrström, M. O'Keeffe, M. Paik Suh and J. Reedijk, in *Pure and Applied Chemistry*, 2013, vol. 85, p. 1715.
23. M. Eddaoudi, D. B. Moler, H. Li, B. Chen, T. M. Reineke, M. O'Keeffe and O. M. Yaghi, *Accounts of Chemical Research*, 2001, **34**, 319-330.
24. O. K. Farha, I. Eryazici, N. C. Jeong, B. G. Hauser, C. E. Wilmer, A. A. Sarjeant, R. Q. Snurr, S. T. Nguyen, A. Ö. Yazaydın and J. T. Hupp, *Journal of the American Chemical Society*, 2012, **134**, 15016-15021.
25. D. A. Gomez-Gualdron, Y. J. Colon, X. Zhang, T. C. Wang, Y.-S. Chen, J. T. Hupp, T. Yildirim, O. K. Farha, J. Zhang and R. Q. Snurr, *Energy & Environmental Science*, 2016, **9**, 3279-3289.
26. L. J. Murray, M. Dinca and J. R. Long, *Chemical Society Reviews*, 2009, **38**, 1294-1314.
27. R. B. Getman, Y.-S. Bae, C. E. Wilmer and R. Q. Snurr, *Chemical Reviews*, 2012, **112**, 703-723.
28. Y. He, W. Zhou, G. Qian and B. Chen, *Chemical Society Reviews*, 2014, **43**, 5657-5678.
29. J. A. Mason, J. Oktawiec, M. K. Taylor, M. R. Hudson, J. Rodriguez, J. E. Bachman, M. I. Gonzalez, A. Cervellino, A. Guagliardi, C. M. Brown, P. L. Llewellyn, N. Masciocchi and J. R. Long, *Nature*, 2015, **527**, 357-361.
30. B. Van de Voorde, B. Bueken, J. Denayer and D. De Vos, *Chemical Society Reviews*, 2014, **43**, 5766-5788.
31. J.-R. Li, J. Sculley and H.-C. Zhou, *Chemical Reviews*, 2012, **112**, 869-932.
32. P. Z. Moghadam, J. F. Ivy, R. K. Arvapally, A. M. dos Santos, J. C. Pearson, L. Zhang, E. Tylianakis, P. Ghosh, I. W. H. Oswald, U. Kaipa, X. Wang, A. K. Wilson, R. Q. Snurr and M. A. Omary, *Chemical Science*, 2017, **8**, 3989-4000.
33. N. S. Bobbitt, M. L. Mendonca, A. J. Howarth, T. Islamoglu, J. T. Hupp, O. K. Farha and R. Q. Snurr, *Chemical Society Reviews*, 2017, **46**, 3357-3385.
34. J. Lee, O. K. Farha, J. Roberts, K. A. Scheidt, S. T. Nguyen and J. T. Hupp, *Chemical Society Reviews*, 2009, **38**, 1450-1459.
35. T. Zhang and W. Lin, *Chemical Society Reviews*, 2014, **43**, 5982-5993.
36. S. M. J. Rogge, A. Bavykina, J. Hajek, H. Garcia, A. I. Olivos-Suarez, A. Sepulveda-Escribano, A. Vimont, G. Clet, P. Bazin, F. Kapteijn, M. Daturi, E. V. Ramos-Fernandez, F. X. Llabres i Xamena, V. Van Speybroeck and J. Gascon, *Chemical Society Reviews*, 2017, **46**, 3134-3184.

37. M. H. Teplensky, M. Fantham, P. Li, T. C. Wang, J. P. Mehta, L. J. Young, P. Z. Moghadam, J. T. Hupp, O. K. Farha, C. F. Kaminski and D. Fairen-Jimenez, *Journal of the American Chemical Society*, 2017, **139**, 7522-7532.
38. P. Horcajada, T. Chalati, C. Serre, B. Gillet, C. Sebrie, T. Baati, J. F. Eubank, D. Heurtaux, P. Clayette, C. Kreuz, J.-S. Chang, Y. K. Hwang, V. Marsaud, P.-N. Bories, L. Cynober, S. Gil, G. Férey, P. Couvreur and R. Gref, *Nature Materials*, 2010, **9**, 172-178.
39. J. Della Rocca, D. Liu and W. Lin, *Accounts of Chemical Research*, 2011, **44**, 957-968.
40. P. Li, Justin A. Modica, Ashlee J. Howarth, E. Vargas L, Peyman Z. Moghadam, Randall Q. Snurr, M. Mrksich, Joseph T. Hupp and Omar K. Farha, *Chem*, 2016, **1**, 154-169.
41. I. Abánades Lázaro, S. Haddad, S. Sacca, C. Orellana-Tavra, D. Fairen-Jimenez and R. S. Forgan, *Chem*, 2017, **2**, 561-578.
42. S. E. Miller, M. H. Teplensky, P. Z. Moghadam and D. Fairen-Jimenez, *Interface Focus*, 2016, **6**.
43. P. Z. Moghadam, A. Li, S. B. Wiggin, A. Tao, A. G. P. Maloney, P. A. Wood, S. C. Ward and D. Fairen-Jimenez, *Chemistry of Materials*, 2017, **29**, 2618-2625.
44. C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp and R. Q. Snurr, *Nature Chemistry*, 2011, **4**, 83.
45. M. O'Keeffe, M. A. Peskov, S. J. Ramsden and O. M. Yaghi, *Accounts of Chemical Research*, 2008, **41**, 1782-1789.
46. C. Bonneau, M. O'Keeffe, D. M. Proserpio, V. A. Blatov, S. R. Batten, S. A. Bourne, M. S. Lah, J.-G. Eon, S. T. Hyde, S. B. Wiggin and L. Öhrström, *Crystal Growth & Design*, 2018.
47. P. G. Boyd and T. K. Woo, *CrystEngComm*, 2016, **18**, 3777-3792.
48. S. M. Moosavi, A. Nandy, K. M. Jablonka, D. Ongari, J. P. Janet, P. G. Boyd, Y. Lee, B. Smit and H. J. Kulik, *Nature Communications*, 2020, **11**, 4068.
49. C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Crystallographica Section B*, 2016, **72**, 171-179.
50. I. D. Brown and B. McMahon, *Acta Crystallographica Section B*, 2002, **58**, 317-324.
51. A. Li, R. Bueno-Perez, S. Wiggin and D. Fairen-Jimenez, *CrystEngComm*, 2020.
52. T. Watanabe and D. S. Sholl, *Langmuir*, 2012, **28**, 14114-14128.
53. J. Goldsmith, A. G. Wong-Foy, M. J. Cafarella and D. J. Siegel, *Chemistry of Materials*, 2013, **25**, 3373-3382.
54. Y. G. Chung, J. Camp, M. Haranczyk, B. J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O. K. Farha, D. S. Sholl and R. Q. Snurr, *Chemistry of Materials*, 2014, **26**, 6185-6192.
55. Y. G. Chung, E. Haldoupis, B. J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling, J. S. Camp, B. Slater, J. I. Siepmann, D. S. Sholl and R. Q. Snurr, *Journal of Chemical & Engineering Data*, 2019.
56. G. Avci, I. Erucar and S. Keskin, *ACS Applied Materials & Interfaces*, 2020, **12**, 41567-41579.
57. J. D. Evans, G. Fraux, R. Gaillac, D. Kohen, F. Trousselet, J.-M. Vanson and F.-X. Coudert, *Chemistry of Materials*, 2017, **29**, 199-212.
58. T. Düren, F. Millange, G. Férey, K. S. Walton and R. Q. Snurr, *The Journal of Physical Chemistry C*, 2007, **111**, 15350-15356.
59. K. S. Walton and R. Q. Snurr, *Journal of the American Chemical Society*, 2007, **129**, 8552-8556.

60. T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza and M. Haranczyk, *Microporous and Mesoporous Materials*, 2012, **149**, 134-141.
61. T. Düren, Y.-S. Bae and R. Q. Snurr, *Chemical Society Reviews*, 2009, **38**, 1237-1247.
62. K. S. W. Sing, *Pure and Applied Chemistry*, 1985, **57**, 603-619.
63. N. A. Ramsahye, G. Maurin, S. Bourrelly, P. L. Llewellyn, T. Loiseau, C. Serre and G. Férey, *Chemical Communications*, 2007, 3261-3263.
64. D. Fairen-Jimenez, S. A. Moggach, M. T. Wharmby, P. A. Wright, S. Parsons and T. Düren, *Journal of the American Chemical Society*, 2011, **133**, 8900-8902.
65. B. Widom, *The Journal of Chemical Physics*, 1963, **39**, 2808-2812.
66. M. N. Rosenbluth and A. W. Rosenbluth, *The Journal of Chemical Physics*, 1955, **23**, 356-359.
67. D. Frenkel and B. Smit, in *Understanding Molecular Simulation (Second Edition)*, eds. D. Frenkel and B. Smit, Academic Press, San Diego, 2002, pp. 269-287.
68. A. Nuhnen and C. Janiak, *Dalton Transactions*, 2020, **49**, 10295-10307.
69. T. Vuong and P. A. Monson, *Langmuir*, 1996, **12**, 5425-5432.
70. T. J. H. Vlugt, E. García-Pérez, D. Dubbeldam, S. Ban and S. Calero, *Journal of Chemical Theory and Computation*, 2008, **4**, 1107-1118.
71. D. Frenkel and B. Smit, in *Understanding Molecular Simulation (Second Edition)*, eds. D. Frenkel and B. Smit, Academic Press, San Diego, 2002, pp. 63-107.
72. D. Nazarian, J. S. Camp and D. S. Sholl, *Chemistry of Materials*, 2016, **28**, 785-793.
73. A. Ö. Yazaydın, R. Q. Snurr, T.-H. Park, K. Koh, J. Liu, M. D. LeVan, A. I. Benin, P. Jakubczak, M. Lanuza, D. B. Galloway, J. J. Low and R. R. Willis, *Journal of the American Chemical Society*, 2009, **131**, 18198-18199.
74. A. Ahmed, S. Seth, J. Purewal, A. G. Wong-Foy, M. Veenstra, A. J. Matzger and D. J. Siegel, *Nature Communications*, 2019, **10**, 1568.
75. Y. J. Colón and R. Q. Snurr, *Chemical Society Reviews*, 2014, **43**, 5735-5749.
76. T. A. Manz and D. S. Sholl, *Journal of Chemical Theory and Computation*, 2010, **6**, 2455-2468.
77. P. Z. Moghadam, T. Islamoglu, S. Goswami, J. Exley, M. Fantham, C. F. Kaminski, R. Q. Snurr, O. K. Farha and D. Fairen-Jimenez, *Nature Communications*, 2018, **9**, 1378.
78. J. B. DeCoste, M. H. Weston, P. E. Fuller, T. M. Tovar, G. W. Peterson, M. D. LeVan and O. K. Farha, *Angewandte Chemie International Edition*, 2014, **53**, 14092-14095.
79. V. Colombo, S. Galli, H. J. Choi, G. D. Han, A. Maspero, G. Palmisano, N. Masciocchi and J. R. Long, *Chemical Science*, 2011, **2**, 1311-1319.
80. Y. G. Chung, D. A. Gómez-Gualdrón, P. Li, K. T. Leperi, P. Deria, H. Zhang, N. A. Vermeulen, J. F. Stoddart, F. You, J. T. Hupp, O. K. Farha and R. Q. Snurr, *Science Advances*, 2016, **2**, e1600909.
81. J. Jia, X. Lin, C. Wilson, A. J. Blake, N. R. Champness, P. Hubberstey, G. Walker, E. J. Cussen and M. Schröder, *Chemical Communications*, 2007, 840-842.
82. B. J. Bucior, N. S. Bobbitt, T. Islamoglu, S. Goswami, A. Gopalan, T. Yildirim, O. K. Farha, N. Bagheri and R. Q. Snurr, *Molecular Systems Design & Engineering*, 2019, **4**, 162-174.

83. A. Ahmed, Y. Liu, J. Purewal, L. D. Tran, A. G. Wong-Foy, M. Veenstra, A. J. Matzger and D. J. Siegel, *Energy & Environmental Science*, 2017, **10**, 2459-2471.
84. A. G. Wong-Foy, A. J. Matzger and O. M. Yaghi, *Journal of the American Chemical Society*, 2006, **128**, 3494-3495.
85. M. Witman, S. Ling, A. Gladysiak, K. C. Stylianou, B. Smit, B. Slater and M. Haranczyk, *The Journal of Physical Chemistry C*, 2017, **121**, 1171-1181.
86. J. L. C. Rowsell and O. M. Yaghi, *Angewandte Chemie International Edition*, 2005, **44**, 4670-4679.
87. A. W. Thornton, C. M. Simon, J. Kim, O. Kwon, K. S. Deeg, K. Konstas, S. J. Pas, M. R. Hill, D. A. Winkler, M. Haranczyk and B. Smit, *Chemistry of Materials*, 2017, **29**, 2844-2854.
88. D. P. Broom, C. J. Webb, K. E. Hurst, P. A. Parilla, T. Gennett, C. M. Brown, R. Zacharia, E. Tylianakis, E. Klontzas, G. E. Froudakis, T. A. Steriotis, P. N. Trikalitis, D. L. Anton, B. Hardy, D. Tamburello, C. Corgnale, B. A. van Hassel, D. Cossement, R. Chahine and M. Hirscher, *Applied Physics A*, 2016, **122**, 151.
89. D. A. Gómez-Gualdrón, T. C. Wang, P. García-Holley, R. M. Sawelewa, E. Argueta, R. Q. Snurr, J. T. Hupp, T. Yildirim and O. K. Farha, *ACS Applied Materials & Interfaces*, 2017, **9**, 33419-33428.
90. D. J. Collins and H.-C. Zhou, *Journal of Materials Chemistry*, 2007, **17**, 3154-3160.
91. N. L. Rosi, J. Eckert, M. Eddaoudi, D. T. Vodak, J. Kim, M. Keeffe and O. M. Yaghi, *Science*, 2003, **300**, 1127.
92. N. S. Bobbitt, J. Chen and R. Q. Snurr, *The Journal of Physical Chemistry C*, 2016, **120**, 27328-27341.
93. P. García-Holley, B. Schweitzer, T. Islamoglu, Y. Liu, L. Lin, S. Rodriguez, M. H. Weston, J. T. Hupp, D. A. Gómez-Gualdrón, T. Yildirim and O. K. Farha, *ACS Energy Letters*, 2018, **3**, 748-754.
94. I. Matito-Martos, P. Z. Moghadam, A. Li, V. Colombo, J. A. R. Navarro, S. Calero and D. Fairen-Jimenez, *Chemistry of Materials*, 2018, **30**, 4571-4579.
95. J. P. Janet and H. J. Kulik, *The Journal of Physical Chemistry A*, 2017, **121**, 8939-8954.
96. C. Altintas, G. Avci, H. Daglar, A. Nemati Vesali Azar, I. Erucar, S. Velioglu and S. Keskin, *Journal of Materials Chemistry A*, 2019, **7**, 9593-9608.
97. P. Zarabadi-Poor and R. Marek, *ACS Applied Materials & Interfaces*, 2019, **11**, 16261-16265.
98. G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari and B. Kozinsky, *Computational Materials Science*, 2016, **111**, 218-230.
99. D. Ongari, A. V. Yakutovich, L. Talirz and B. Smit, *ACS Central Science*, 2019, **5**, 1663-1675.
100. D. Ongari, L. Talirz and B. Smit, *ACS Central Science*, 2020, **6**, 1890-1900.
101. B. Mason, *Why scientists need to be better at data visualization*, https://knowablemagazine.org/article/mind/2019/science-data-visualization, 2020.
102. B. Wong, *Nature Methods*, 2011, **8**, 101-101.
103. B. Wong, *Nature Methods*, 2011, **8**, 189-189.
104. B. Wong, *Nature Methods*, 2011, **8**, 277-277.
105. B. Wong, *Nature Methods*, 2011, **8**, 365-365.
106. B. Wong, *Nature Methods*, 2011, **8**, 441-441.
107. J. M. Perkel, *Nature*, 2018, 133-134.

108. C. Balzer, R. Oktavian, M. Zandi, D. Fairen-Jimenez and P. Z. Moghadam, *Patterns*, 2020, **1**, 100107.
109. L. Sarkisov, R. Bueno-Perez, M. Sutharson and D. Fairen-Jimenez, *Chemistry of Materials*, 2020, **32**, 9849-9867.
110. 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, *Journal of Applied Crystallography*, 2020, **53**, 226-235.
111. I. J. Bruno, J. C. Cole, P. R. Edgington, M. Kessler, C. F. Macrae, P. McCabe, J. Pearson and R. Taylor, *Acta Crystallographica Section B*, 2002, **58**, 389-397.
112. N. W. Ockwig, O. Delgado-Friedrichs, M. O'Keeffe and O. M. Yaghi, *Accounts of Chemical Research*, 2005, **38**, 176-182.
113. A. Spek, *Acta Crystallographica Section C*, 2015, **71**, 9-18.
114. O. V. Dolomanov, L. J. Bourhis, R. J. Gildea, J. A. K. Howard and H. Puschmann, *Journal of Applied Crystallography*, 2009, **42**, 339-341.
115. CCDC, *ConQuest User Guide and Tutorials*, www.ccdc.cam.ac.uk/support-and-resources/ccdcresources/ConQuest-UserGuide.pdf, 2019.
116. CCDC, *I find searching for polymeric structures difficult and I get results I don't fully understand. Can you explain how polymeric structures are defined so I can tailor my searches to be more effective?*, www.ccdc.cam.ac.uk/support-and-resources/support/case/?caseid=f75281ce-d3fe-472e-8438-7b6cdea5accb, 2019.
117. A. P. Côté, A. I. Benin, N. W. Ockwig, M. Keeffe, A. J. Matzger and O. M. Yaghi, *Science*, 2005, **310**, 1166.
118. CCDC, *Deposit a Structure*, www.ccdc.cam.ac.uk/Community/depositastructure/, 2020.
119. M. Tong, Y. Lan, Q. Yang and C. Zhong, *Chemical Engineering Science*, 2017, **168**, 456-464.
120. M. Tong, Y. Lan, Z. Qin and C. Zhong, *The Journal of Physical Chemistry C*, 2018, **122**, 13009-13016.
121. T. Yan, Y. Lan, M. Tong and C. Zhong, *ACS Sustainable Chemistry & Engineering*, 2019, **7**, 1220-1227.
122. C. S. Diercks and O. M. Yaghi, *Science*, 2017, **355**, eaal1585.
123. N. Huang, P. Wang and D. Jiang, *Nature Reviews Materials*, 2016, **1**, 16068.
124. D. Frenkel and B. Smit, in *Understanding Molecular Simulation (Second Edition)*, eds. D. Frenkel and B. Smit, Academic Press, San Diego, 2002, pp. 23-61.
125. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, *The Journal of Chemical Physics*, 1953, **21**, 1087-1092.
126. D. Frenkel and B. Smit, in *Understanding Molecular Simulation (Second Edition)*, eds. D. Frenkel and B. Smit, Academic Press, San Diego, 2002, pp. 111-137.
127. A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard III and W. M. Skiff, *Journal of American Chemical Society*, 1992, **114**, 10035-10046.
128. S. L. Mayo, B. D. Olafson and W. A. Goddard III, *J. Phys. Chem.*, 1990, **94**, 8897-8909.
129. J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *Journal of Computational Chemistry*, 2004, **25**, 1157-1174.
130. W. L. Jorgensen, D. S. Maxwell and J. Tirado-Rives, *Journal of the American Chemical Society*, 1996, **118**, 11225-11236.
131. M. G. Martin and J. I. Siepmann, *The Journal of Physical Chemistry B*, 1998, **102**, 2569-2577.

132. S. Hamad, S. R. G. Balestra, R. Bueno-Perez, S. Calero and A. R. Ruiz-Salvador, *Journal of Solid State Chemistry*, 2015, **223**, 144-151.
133. C. E. Wilmer, K. C. Kim and R. Q. Snurr, *The Journal of Physical Chemistry Letters*, 2012, **3**, 2506-2511.
134. D. Frenkel and B. Smit, in *Understanding Molecular Simulation (Second Edition)*, eds. D. Frenkel and B. Smit, Academic Press, San Diego, 2002, pp. 291-320.
135. B. A. Wells and A. L. Chaffee, *Journal of Chemical Theory and Computation*, 2015, **11**, 3684-3695.
136. M. J. Lennox, M. Bound, A. Henley and E. Besley, *Molecular Simulation*, 2017, **43**, 828-837.
137. T. Düren, L. Sarkisov, O. M. Yaghi and R. Q. Snurr, *Langmuir*, 2004, **20**, 2683-2689.
138. J. I. Siepmann and D. Frenkel, *Molecular Physics*, 1992, **75**, 59-70.
139. D. Frenkel and B. Smit, in *Understanding Molecular Simulation (Second Edition)*, eds. D. Frenkel and B. Smit, Academic Press, San Diego, 2002, pp. 321-387.
140. W. Shi and E. J. Maginn, *Journal of Chemical Theory and Computation*, 2007, **3**, 1451-1463.
141. D. Dubbeldam, S. Calero, D. E. Ellis and R. Q. Snurr, *Molecular Simulation*, 2016, **42**, 81-101.
142. A. Inc, *Accelrys Software Inc., San Diego*, 2012.
143. L. Sarkisov and A. Harrison, *Molecular Simulation*, 2011, **37**, 1248-1257.
144. E. W. Dijkstra, *Numerische Mathematik*, 1959, **1**, 269-271.
145. J. Hoshen and R. Kopelman, *Physical Review B*, 1976, **14**, 3438-3445.
146. L. D. Gelb and K. E. Gubbins, *Langmuir*, 1999, **15**, 305-308.
147. P. Zhao, H. Fang, S. Mukhopadhyay, A. Li, S. Rudić, I. J. McPherson, C. C. Tang, D. Fairen-Jimenez, S. C. E. Tsang and S. A. T. Redfern, *Nature Communications*, 2019, **10**, 999.
148. D. Fairen-Jimenez, R. Galvelis, A. Torrisi, A. D. Gellan, M. T. Wharmby, P. A. Wright, C. Mellot-Draznieks and T. Düren, *Dalton Transactions*, 2012, **41**, 10752-10762.
149. J. J. Potoff and J. I. Siepmann, *AIChE Journal*, 2004, **47**, 1676-1682.
150. R. C. Reid, J. M. Prausnitz and B. E. Poling, *The properties of gases and liquids*, McGraw Hill Book Co.,New York, NY, United States, 1987.
151. Y. Du, B. Wooler, M. Nines, P. Kortunov, C. S. Paur, J. Zengel, S. C. Weston and P. I. Ravikovitch, *Journal of the American Chemical Society*, 2015, **137**, 13603-13611.
152. R. Erttmann, N. Erb, A. Steinhoff and G. Landbeck, *Journal of Cancer Research and Clinical Oncology*, 1988, **114**, 509-513.
153. L. Meng, X. Zhang, Q. Lu, Z. Fei and P. J. Dyson, *Biomaterials*, 2012, **33**, 1689-1698.
154. K. Shan, A. M. Lincoff and J. B. Young, *Annals of Internal Medicine*, 1996, **125**, 47-58.
155. W. H. Porter, *Pure and Applied Chemistry*, 1991, **63**, 1119-1122.
156. L. A. Clark, S. Chempath and R. Q. Snurr, *Langmuir*, 2005, **21**, 2267-2272.
157. S.-Y. Zhang, D. Fairen-Jimenez and M. J. Zaworotko, *Angewandte Chemie International Edition*, 2020, **59**, 17600-17606.
158. S.-Y. Zhang, S. Jensen, K. Tan, L. Wojtas, M. Roveto, J. Cure, T. Thonhauser, Y. J. Chabal and M. J. Zaworotko, *Journal of the American Chemical Society*, 2018, **140**, 12545-12552.

159. S.-Y. Zhang, L. Wojtas and M. J. Zaworotko, *Journal of the American Chemical Society*, 2015, **137**, 12045-12049.
160. X. Bao, L. J. Broadbelt and R. Q. Snurr, *Physical Chemistry Chemical Physics*, 2010, **12**, 6466-6473.
161. X. Bao, R. Q. Snurr and L. J. Broadbelt, *Langmuir*, 2009, **25**, 10730-10736.
162. P. Z. Moghadam and T. Düren, *The Journal of Physical Chemistry C*, 2012, **116**, 20874-20881.
163. R. Bueno-Perez, S. R. G. Balestra, M. A. Camblor, J. G. Min, S. B. Hong, P. J. Merkling and S. Calero, *Chemistry – A European Journal*, 2018, **24**, 4121-4132.
164. B. Doherty, X. Zhong, S. Gathiaka, B. Li and O. Acevedo, *Journal of Chemical Theory and Computation*, 2017, **13**, 6131-6145.
165. S. W. I. Siu, K. Pluhackova and R. A. Böckmann, *Journal of Chemical Theory and Computation*, 2012, **8**, 1459-1470.
166. R. C. Rizzo and W. L. Jorgensen, *Journal of the American Chemical Society*, 1999, **121**, 4827-4836.
167. W. L. Jorgensen, *The Journal of Physical Chemistry*, 1986, **90**, 1276-1284.
168. X. Bao, L. J. Broadbelt and R. Q. Snurr, *Microporous and Mesoporous Materials*, 2012, **157**, 118-123.
169. MGI, *Materials Genome Initiative*, www.mgi.gov/, 2019.
170. B. J. Bucior, A. S. Rosen, M. Haranczyk, Z. Yao, M. E. Ziebel, O. K. Farha, J. T. Hupp, J. I. Siepmann, A. Aspuru-Guzik and R. Q. Snurr, *Crystal Growth & Design*, 2019, **19**, 6682-6697.
171. P. Z. Moghadam, A. Li, X.-W. Liu, R. Bueno-Perez, S.-D. Wang, S. B. Wiggin, P. A. Wood and D. Fairen-Jimenez, *Chemical Science*, 2020.
172. P. Li, N. A. Vermeulen, C. D. Malliakas, D. A. Gómez-Gualdrón, A. J. Howarth, B. L. Mehdi, A. Dohnalkova, N. D. Browning, M. O'Keeffe and O. K. Farha, *Science*, 2017, **356**, 624-627.
173. K. Sumida, D. L. Rogow, J. A. Mason, T. M. McDonald, E. D. Bloch, Z. R. Herm, T.-H. Bae and J. R. Long, *Chemical Reviews*, 2012, **112**, 724-781.
174. R. W. Flaig, T. M. Osborn Popp, A. M. Fracaroli, E. A. Kapustin, M. J. Kalmutzki, R. M. Altamimi, F. Fathieh, J. A. Reimer and O. M. Yaghi, *Journal of the American Chemical Society*, 2017, **139**, 12125-12128.
175. S. M. Cohen, *Chemical Reviews*, 2012, **112**, 970-1000.
176. P. Deria, J. E. Mondloch, O. Karagiaridi, W. Bury, J. T. Hupp and O. K. Farha, *Chemical Society Reviews*, 2014, **43**, 5896-5912.
177. P. Z. Moghadam, D. Fairen-Jimenez and R. Q. Snurr, *Journal of Materials Chemistry A*, 2016, **4**, 529-536.
178. V. Bernales, M. A. Ortuño, D. G. Truhlar, C. J. Cramer and L. Gagliardi, *ACS Central Science*, 2017.
179. L. Zhu, X.-Q. Liu, H.-L. Jiang and L.-B. Sun, *Chem. Rev.*, 2017, **117**, 8129-8176.
180. M. C. So, G. P. Wiederrecht, J. E. Mondloch, J. T. Hupp and O. K. Farha, *Chemical Communications*, 2015, **51**, 3501-3510.
181. Y. Cui, Y. Yue, G. Qian and B. Chen, *Chem. Rev.*, 2012, **112**, 1126-1162.
182. Y. Peng, T. Gong, K. Zhang, X. Lin, Y. Liu, J. Jiang and Y. Cui, *Nat. Commun.*, 2014, **5**, 4406.
183. J. Navarro-Sánchez, A. I. Argente-García, Y. Moliner-Martínez, D. Roca-Sanjuán, D. Antypov, P. Campíns-Falcó, M. J. Rosseinsky and C. Martí-Gastaldo, *J. Am. Chem. Soc.*, 2017, **139**, 4294-4297.
184. L. Ma, J. M. Falkowski, C. Abney and W. Lin, *Nat. Chem.*, 2010, **2**, 838.

185. Z. R. Herm, B. M. Wiers, J. A. Mason, J. M. van Baten, M. R. Hudson, P. Zajdel, C. M. Brown, N. Masciocchi, R. Krishna and J. R. Long, *Science*, 2013, **340**, 960-964.
186. A. Torres-Knoop, R. Krishna and D. Dubbeldam, *Angew. Chem. Int. Ed.*, 2014, **53**, 7774-7778.
187. J. M. Holcroft, K. J. Hartlieb, P. Z. Moghadam, J. G. Bell, G. Barin, D. P. Ferris, E. D. Bloch, M. M. Algaradah, M. S. Nassar, Y. Y. Botros, K. M. Thomas, J. R. Long, R. Q. Snurr and J. F. Stoddart, *J. Am. Chem. Soc.*, 2015, **137**, 5706-5719.
188. M. Haranczyk and J. A. Sethian, *Journal of Chemical Theory and Computation*, 2010, **6**, 3472-3480.
189. E. Haldoupis, S. Nair and D. S. Sholl, *Phys. Chem. Chem. Phys.*, 2011, **13**, 5053-5060.
190. F.-X. Coudert and A. H. Fuchs, *Coord. Chem. Rev.*, 2016, **307**, 211-236.
191. S. Oien-Odegaard, G. C. Shearer, D. S. Wragg and K. P. Lillerud, *Chemical Society Reviews*, 2017, **46**, 4867-4876.
192. H. Li, M. Eddaoudi, T. L. Groy and O. M. Yaghi, *Journal of the American Chemical Society*, 1998, **120**, 8571-8572.
193. M. Kondo, T. Yoshitomi, H. Matsuzaka, S. Kitagawa and K. Seki, *Angewandte Chemie International Edition in English*, 1997, **36**, 1725-1727.
194. S. R. Batten, N. R. Champness, X.-M. Chen, J. Garcia-Martinez, S. Kitagawa, L. Ohrstrom, M. O'Keeffe, M. P. Suh and J. Reedijk, *CrystEngComm*, 2012, **14**, 3001-3004.
195. W. H. Zachariasen, *Journal of the American Chemical Society*, 1940, **62**, 1011-1013.
196. Y. Kinoshita, I. Matsubara, T. Higuchi and Y. Saito, *Bulletin of the Chemical Society of Japan*, 1959, **32**, 1221-1226.
197. N. L. Strutt, D. Fairen-Jimenez, J. Iehl, M. B. Lalonde, R. Q. Snurr, O. K. Farha, J. T. Hupp and J. F. Stoddart, *Journal of the American Chemical Society*, 2012, **134**, 17436-17439.
198. Y. J. Colón, D. Fairen-Jimenez, C. E. Wilmer and R. Q. Snurr, *The Journal of Physical Chemistry C*, 2014, **118**, 5383-5389.
199. DoE, *Materials-Based Hydrogen Storage*, www.energy.gov/eere/fuelcells/materials-based-hydrogen-storage, 2019.
200. V. Buch, *The Journal of Chemical Physics*, 1994, **100**, 7610-7629.
201. P. Z. Moghadam, S. M. J. Rogge, A. Li, C.-M. Chow, J. Wieme, N. Moharrami, M. Aragones-Anglada, G. Conduit, D. A. Gomez-Gualdron, V. Van Speybroeck and D. Fairen-Jimenez, *Matter*, 2019, **1**, 219-234.
202. N. Bulc, L. Golic and J. Siftar, *Acta Crystallographica Section C*, 1983, **39**, 176-178.
203. B. F. Hoskins and R. Robson, *Journal of the American Chemical Society*, 1989, **111**, 5962-5964.
204. L. Chen, P. S. Reiss, S. Y. Chong, D. Holden, K. E. Jelfs, T. Hasell, M. A. Little, A. Kewley, M. E. Briggs, A. Stephenson, K. M. Thomas, J. A. Armstrong, J. Bell, J. Busto, R. Noel, J. Liu, D. M. Strachan, P. K. Thallapally and A. I. Cooper, *Nature Materials*, 2014, **13**, 954-960.
205. T. Mitra, K. E. Jelfs, M. Schmidtmann, A. Ahmed, S. Y. Chong, D. J. Adams and A. I. Cooper, *Nature Chemistry*, 2013, **5**, 276-281.
206. T. Hasell, M. Miklitz, A. Stephenson, M. A. Little, S. Y. Chong, R. Clowes, L. Chen, D. Holden, G. A. Tribello, K. E. Jelfs and A. I. Cooper, *Journal of the American Chemical Society*, 2016, **138**, 1653-1659.

207. A. Kewley, A. Stephenson, L. Chen, M. E. Briggs, T. Hasell and A. I. Cooper, *Chemistry of Materials*, 2015, **27**, 3207-3210.
208. T.-C. Lee, E. Kalenius, A. I. Lazar, K. I. Assaf, N. Kuhnert, C. H. Grün, J. Jänis, O. A. Scherman and W. M. Nau, *Nature Chemistry*, 2013, **5**, 376-382.
209. M. Yoshizawa, J. K. Klosterman and M. Fujita, *Angewandte Chemie International Edition*, 2009, **48**, 3418-3438.
210. M. D. Ward, C. A. Hunter and N. H. Williams, *Accounts of Chemical Research*, 2018, **51**, 2073-2082.
211. K. Acharyya and P. S. Mukherjee, *Chemical Communications*, 2014, **50**, 15788-15791.
212. M. Brutschy, M. W. Schneider, M. Mastalerz and S. R. Waldvogel, *Advanced Materials*, 2012, **24**, 6049-6052.
213. J. D. Evans, K. E. Jelfs, G. M. Day and C. J. Doonan, *Chemical Society Reviews*, 2017, **46**, 3286-3301.
214. K. J. Msayib, D. Book, P. M. Budd, N. Chaukura, K. D. M. Harris, M. Helliwell, S. Tedds, A. Walton, J. E. Warren, M. Xu and N. B. McKeown, *Angewandte Chemie International Edition*, 2009, **48**, 3273-3277.
215. K. P. U. Perera, M. Krawiec and D. W. Smith, *Tetrahedron*, 2002, **58**, 10197-10203.
216. M. Mastalerz and I. M. Oppel, *Angewandte Chemie International Edition*, 2012, **51**, 5252-5255.
217. J. D. Evans, D. M. Huang, M. Haranczyk, A. W. Thornton, C. J. Sumby and C. J. Doonan, *CrystEngComm*, 2016, **18**, 4133-4141.
218. M. Miklitz, S. Jiang, R. Clowes, M. E. Briggs, A. I. Cooper and K. E. Jelfs, *The Journal of Physical Chemistry C*, 2017, **121**, 15211-15222.
219. Z. Wu, K. Zhou, A. V. Ivanov, M. Yusobov and F. Verpoort, *Coordination Chemistry Reviews*, 2017, **353**, 180-200.
220. A. Schmidt, A. Casini and F. E. Kühn, *Coordination Chemistry Reviews*, 2014, **275**, 19-36.
221. R. Chakrabarty, P. S. Mukherjee and P. J. Stang, *Chemical Reviews*, 2011, **111**, 6810-6918.
222. N. Ahmad, A. H. Chughtai, H. A. Younus and F. Verpoort, *Coordination Chemistry Reviews*, 2014, **280**, 1-27.
223. R. Custelcean, *Chemical Society Reviews*, 2014, **43**, 1813-1824.
224. S. Mukherjee and P. S. Mukherjee, *Chemical Communications*, 2014, **50**, 2239-2248.
225. S. R. Seidel and P. J. Stang, *Accounts of Chemical Research*, 2002, **35**, 972-983.
226. L. Li, D. J. Fanna, N. D. Shepherd, L. F. Lindoy and F. Li, *Journal of Inclusion Phenomena and Macrocyclic Chemistry*, 2015, **82**, 3-12.
227. M. Yoshizawa and M. Yamashina, *Chemistry Letters*, 2016, **46**, 163-171.
228. M. D. Ward, *Chemical Communications*, 2009, 4487-4499.
229. P. Muller, *Pure and Applied Chemistry*, 1994, **66**, 1077-1184.
230. P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson and G. Carlsson, *Scientific Reports*, 2013, **3**, 1236.
231. G. Carlsson, *Bull. Amer. Math. Soc*, 2009, **46**, 255-308.
232. P. Bubenik, *The Journal of Machine Learning Research*, 2015, **16**, 77-102.
233. Y. Lee, S. D. Barthel, P. Dłotko, S. M. Moosavi, K. Hess and B. Smit, *Nature Communications*, 2017, **8**, 15396.
234. Y. Lee, S. D. Barthel, P. Dłotko, S. M. Moosavi, K. Hess and B. Smit, *Journal of Chemical Theory and Computation*, 2018, **14**, 4427-4437.

235. S. M. Moosavi, H. Xu, L. Chen, A. I. Cooper and B. Smit, *Chemical Science*, 2020, **11**, 5423-5433.
236. T. G. Project, *GUDHI User and Reference Manual*, 3.4.1 edn., GUDHI Editorial Board, 2021.
237. J. J. Henkelis and M. J. Hardie, *Chemical Communications*, 2015, **51**, 11929-11943.
238. M. Han, D. M. Engelhard and G. H. Clever, *Chemical Society Reviews*, 2014, **43**, 1848-1860.
239. D. Zhang, T. K. Ronson and J. R. Nitschke, *Accounts of Chemical Research*, 2018, **51**, 2423-2436.
240. D. L. Caulder and K. N. Raymond, *Journal of the Chemical Society, Dalton Transactions*, 1999, 1185-1200.
241. K. Harris, D. Fujita and M. Fujita, *Chemical Communications*, 2013, **49**, 6703-6712.
242. M. Fujita, K. Umemoto, M. Yoshizawa, N. Fujita, T. Kusukawa and K. Biradha, *Chemical Communications*, 2001, 509-518.
243. F. Nielsen, in *Introduction to HPC with MPI for Data Science*, ed. F. Nielsen, Springer International Publishing, Cham, 2016, pp. 195-211.
244. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G.-L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, Y. Vázquez-Baeza and C. SciPy, *Nature Methods*, 2020, **17**, 261-272.
245. G. V. Fabian Pedregosa, 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, Édouard Duchesnay, 2011, **12**, 2825-2830.
246. F. G. Kerry, *Industrial Gas Handbook: Gas Separation and Purification (1st ed.)*, CRC Press., 2007.
247. N. P. Franks, R. Dickinson, S. L. M. de Sousa, A. C. Hall and W. R. Lieb, *Nature*, 1998, **396**, 324-324.
248. S. C. Cullen and E. G. Gross, *Science*, 1951, **113**, 580.
249. Y. Serdar, D. Douglas, B. Rudi and H. Rick, Proc.SPIE, 2005.
250. P. W. Hoff, J. C. Swingle and C. K. Rhodes, *Applied Physics Letters*, 1973, **23**, 245-246.

251. J. R. Beattie, J. N. Matossian, R. L. Poeschel, W. P. Rogers and R. M. Martinelli, *Journal of Propulsion and Power*, 1989, **5**, 438-444.
252. C. M. Simon, R. Mercado, S. K. Schnell, B. Smit and M. Haranczyk, *Chemistry of Materials*, 2015, **27**, 4459-4475.
253. D. Banerjee, C. M. Simon, S. K. Elsaidi, M. Haranczyk and P. K. Thallapally, *Chem*, 2018, **4**, 466-494.
254. A. Soleimani Dorcheh, D. Denysenko, D. Volkmer, W. Donner and M. Hirscher, *Microporous and Mesoporous Materials*, 2012, **162**, 64-68.
255. K. V. Lawler, A. Sharma, B. Alagappan and P. M. Forster, *Microporous and Mesoporous Materials*, 2016, **222**, 104-112.
256. D. Banerjee, C. M. Simon, A. M. Plonka, R. K. Motkuri, J. Liu, X. Chen, B. Smit, J. B. Parise, M. Haranczyk and P. K. Thallapally, *Nature Communications*, 2016, **7**, ncomms11831.
257. D. Banerjee, A. J. Cairns, J. Liu, R. K. Motkuri, S. K. Nune, C. A. Fernandez, R. Krishna, D. M. Strachan and P. K. Thallapally, *Accounts of Chemical Research*, 2015, **48**, 211-219.
258. B. J. Sikora, C. E. Wilmer, M. L. Greenfield and R. Q. Snurr, *Chemical Science*, 2012, **3**, 2217-2223.
259. D. Banerjee, Z. Zhang, A. M. Plonka, J. Li and J. B. Parise, *Crystal Growth & Design*, 2012, **12**, 2162-2165.
260. T. Hasell, M. Schmidtmann and A. I. Cooper, *Journal of the American Chemical Society*, 2011, **133**, 14920-14923.
261. S. Tothadi, M. A. Little, T. Hasell, M. E. Briggs, S. Y. Chong, M. Liu and A. I. Cooper, *CrystEngComm*, 2017, **19**, 4933-4941.
262. A. G. Slater, P. S. Reiss, A. Pulido, M. A. Little, D. L. Holden, L. Chen, S. Y. Chong, B. M. Alston, R. Clowes, M. Haranczyk, M. E. Briggs, T. Hasell, G. M. Day and A. I. Cooper, *ACS Central Science*, 2017, **3**, 734-742.
263. H. Takezawa, T. Murase, G. Resnati, P. Metrangolo and M. Fujita, *Angewandte Chemie International Edition*, 2015, **54**, 8411-8414.
264. H. Takezawa, T. Murase, G. Resnati, P. Metrangolo and M. Fujita, *Journal of the American Chemical Society*, 2014, **136**, 1786-1788.
265. Y. Kohyama, T. Murase and M. Fujita, *Angewandte Chemie International Edition*, 2014, **53**, 11510-11513.
266. S. Horiuchi, T. Murase and M. Fujita, *Chemistry – An Asian Journal*, 2011, **6**, 1839-1847.
267. T. Murase, S. Horiuchi and M. Fujita, *Journal of the American Chemical Society*, 2010, **132**, 2866-2867.
268. S. Horiuchi, Y. Nishioka, T. Murase and M. Fujita, *Chemical Communications*, 2010, **46**, 3460-3462.
269. Y. Nishioka, T. Yamaguchi, M. Yoshizawa and M. Fujita, *Journal of the American Chemical Society*, 2007, **129**, 7000-7001.
270. T. Furusawa, M. Kawano and M. Fujita, *Angewandte Chemie International Edition*, 2007, **46**, 5717-5719.
271. V. Santolini, M. Miklitz, E. Berardo and K. E. Jelfs, *Nanoscale*, 2017, **9**, 5280-5298.
272. R. L. Greenaway, V. Santolini, M. J. Bennison, B. M. Alston, C. J. Pugh, M. A. Little, M. Miklitz, E. G. B. Eden-Rump, R. Clowes, A. Shakil, H. J. Cuthbertson, H. Armstrong, M. E. Briggs, K. E. Jelfs and A. I. Cooper, *Nature Communications*, 2018, **9**, 2849.
273. S. Barthel, E. V. Alexandrov, D. M. Proserpio and B. Smit, *Crystal Growth & Design*, 2018, **18**, 1738-1747.

274. S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, *Computational Materials Science*, 2013, **68**, 314-319.
275. M. Pardakhti, E. Moharreri, D. Wanik, S. L. Suib and R. Srivastava, *ACS Combinatorial Science*, 2017, **19**, 640-645.
276. M. Fernandez, T. K. Woo, C. E. Wilmer and R. Q. Snurr, *The Journal of Physical Chemistry C*, 2013, **117**, 7681-7689.
277. M. Fernandez, P. G. Boyd, T. D. Daff, M. Z. Aghaji and T. K. Woo, *The Journal of Physical Chemistry Letters*, 2014, **5**, 3056-3060.
278. R. Anderson, J. Rodgers, E. Argueta, A. Biong and D. A. Gómez-Gualdrón, *Chemistry of Materials*, 2018, **30**, 6325-6337.
279. G. Borboudakis, T. Stergiannakos, M. Frysali, E. Klontzas, I. Tsamardinos and G. E. Froudakis, *npj Computational Materials*, 2017, **3**, 40.
280. F.-X. Coudert, *Advanced Theory and Simulations*, 2019, **2**, 1900131.
281. H. K. Park, Yeonghun; Choe, Wonyoung; Kim, Jihan, *arXiv*, 2021, 2108.13590.
282. D. King Ross, J. Rowland, G. Oliver Stephen, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, N. Soldatova Larisa, A. Sparkes, E. Whelan Kenneth and A. Clare, *Science*, 2009, **324**, 85-89.
283. R. King, E4R Symposium, Youtube, 2020.
284. B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick and A. I. Cooper, *Nature*, 2020, **583**, 237-241.
285. S. Steiner, J. Wolf, S. Glatzel, A. Andreou, M. Granda Jarosław, G. Keenan, T. Hinkley, G. Aragon-Camarasa, J. Kitson Philip, D. Angelone and L. Cronin, *Science*, 2019, **363**, eaav2211.
286. S. M. Moosavi, A. Chidambaram, L. Talirz, M. Haranczyk, K. C. Stylianou and B. Smit, *Nature Communications*, 2019, **10**, 539.
287. J. M. Vicent-Luna, *SITES-ANALYZER*, https://github.com/jmviclun/SITES-ANALYZER, 2020.

9 APPENDICES

APPENDIX A DCA AND ALPHA-CHC MOLECULE DEFINITION PARAMETERS AND GCMC FORCE FIELD PARAMETERS 165

APPENDIX B MOLECULE PARAMETERS FOR THE STUDY OF CMOMS FOR CHIRAL SEPARATIONS 174

APPENDIX C MOF FAMILIES CLASSIFICATION: DESCRIPTION OF THE CRITERIA DEVELOPED 246

APPENDIX D MOFS' FUNCTIONAL GROUPS: DESCRIPTION OF THE CRITERIA DEVELOPED 255

APPENDIX E CALCULATIONS OF THE CSD MOFS' PHYSICAL AND GEOMETRICAL PROPERTIES 265

APPENDIX F CALCULATION OF FRAMEWORK DIMENSIONALITIES 268

APPENDIX G QUALITY ASSESSMENT OF THE DATA IN THE CSD MOF SUBSET USING R FACTORS 272

APPENDIX H HTS FOR H2 GCMC SIMULATIONS PARAMETERS AND ADDITIONAL RESULTS 275

APPENDIX I IDENTIFICATION OF METAL-ORGANIC CAGES AND ORGANIC CAGES WITH TOPOLOGICAL DATA ANALYSIS – FURTHER DETAILS 283

APPENDIX J ADDITIONAL CONQUEST QUERIES USED FOR REDUCING THE SEARCH SPACE OF ORGANIC CAGES IN THE CSD 294

APPENDIX A DCA AND ALPHA-CHC MOLECULE DEFINITION PARAMETERS AND GCMC FORCE FIELD PARAMETERS

Below are the equations implemented in RASPA:

- Lennard-Jones

$$U = 4\varepsilon[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}] \quad \textbf{A-1}$$

where $\frac{\varepsilon}{k_B}$ is in K and σ in Å.

- Harmonic bond

$$U = \frac{1}{2}\varepsilon(r - \sigma)^2 \quad \textbf{A-2}$$

where $\frac{\varepsilon}{k_B}$ is in K/Å^2 and σ in Å.

- Harmonic bend

$$U = \frac{1}{2}\varepsilon\left(\theta_{ijk} - \sigma\right)^2 \quad \textbf{A-3}$$

where $\frac{\varepsilon}{k_B}$ is in K/rad^2 and σ in degrees.

- Torsion

$$U = \varepsilon[1 + cos\ cos\left(\sigma\varphi_{ijkl} - p_2\right)] \quad \textbf{A-4}$$

where $\frac{\varepsilon}{k_B}$ is in K, σ is dimensionless and p_2 in degrees.

- Improper torsion

$$U = \varepsilon[1 + cos\ cos\left(\sigma\varphi_{ijkl} - p_2\right)] \quad \textbf{A-5}$$

where $\frac{\varepsilon}{k_B}$ is in K, σ is dimensionless and p_2 in degrees.

The parameters in the tables below can also be found at:

github.com/ayl23/DCA_aCHC_in_CAU7

Table A1 Force field parameters used in the GCMC simulations. See **Figure 21** for the indexing of the atoms in DCA and α-CHC. The parameters are as defined in equation **A-1**

CAU-7	**DFFUFF**	
	ε/k_B (K)	**σ (Å)**
C	47.8562	3.47299
H (MOF)	7.64893	2.84642
O	48.1581	3.03315
Bi	260.658	3.894
DCA	**AMBER**	
C1	55.053	3.400
C2	43.277	3.400
H1	7.901	2.293
H2	0.00	0.000
O1	105.677	2.960
O2	105.878	3.066
Cl1	133.354	3.471
Cl2	133.354	3.471
α-CHC		
C1	49.718	3.315

C2	49.718	3.315
C3	49.718	3.315
C4	49.718	3.315
C5	49.718	3.315
C6	49.718	3.315
C7	49.718	3.315
C8	80.314	3.479
C9	49.718	3.315
C10	49.718	3.315
O1	46.800	3.243
O2	46.800	3.243
O3	73.621	3.048
N	55.254	3.274
H1	2.365	0.538
H2	8.102	2.625
H3	8.102	2.625
H4	8.102	2.625
H5	8.102	2.625
H6	8.102	2.625
H7	2.365	0.538

Table A2 Bond, bend and torsion definitions of DCA. The parameters are as defined in equations **A-2** to **A-5**.

Bond stretch

Type of potential	Parameters		Atoms	
	$\frac{\varepsilon}{k_B}$ (K/Å^2)	σ (Å)		
Harmonic bond	333838.9	1.096	C1	H1
Harmonic bond	373793.8	0.973	O2	H2
Harmonic bond	315017.4	1.524	C1	C2
Harmonic bond	268016.4	1.804	C1	Cl1
Harmonic bond	268016.4	1.804	C1	Cl2
Harmonic bond	641810.3	1.218	C2	O1
Harmonic bond	402678.8	1.351	C2	O2

Bend stretch

Type of potential	Parameters		Atoms		
	$\frac{\varepsilon}{k_B}$ (K/rad^2)	σ (°)			
Harmonic bend	47001.00	109.69	C2	C1	H1
Harmonic bend	50221.63	106.55	C2	O2	H2
Harmonic bend	40257.82	106.99	Cl1	C1	H1
Harmonic bend	40257.82	106.99	Cl2	C1	H1
Harmonic bend	67834.42	123.2	C1	C2	O1
Harmonic bend	68840.87	112.73	C1	C2	O2

Harmonic bend	57870.61	110.41	C2	C1	Cl1
Harmonic bend	57870.61	110.41	C2	C1	Cl2
Harmonic bend	76389.21	122.1	O1	C2	O2
Harmonic bend	54448.70	109.33	Cl1	C1	Cl2

Torsion

Type of potential	**Parameters**			**Atoms**			
	$\frac{\varepsilon}{k_B}$ **(K)**	σ	p_2 **(°)**				
Torsion	1157.412	2.0	180.0	C1	C2	O2	H2
Torsion	0	2.0	180.0	H1	C1	C2	O1
Torsion	956.1232	1.0	0.0	O1	C2	O2	H2
Torsion	1157.412	2.0	180.0	O1	C2	O2	H2
Torsion	0	2.0	180.0	H1	C1	C2	O2
Torsion	0	2.0	180.0	Cl1	C1	C2	O1
Torsion	0	2.0	180.0	Cl2	C1	C2	O1
Torsion	0	2.0	180.0	Cl1	C1	C2	O2
Torsion	0	2.0	180.0	Cl2	C1	C2	O2

Improper torsion

Type of potential	**Parameters**			**Atoms**			
	$\frac{\varepsilon}{k_B}$ **(K)**	σ	p_2 **(°)**				
Improper torsion	553.545	2.0	180.0	C1	C2	O1	O2

Table A3 Bond, bend and torsion definitions of α-CHC. The parameters are as defined in equations **A-2** to **A-5**.

Bond stretch				
Type	**Parameters**		**Atoms**	
	$\frac{\varepsilon}{k_B}$ **(K/Å²)**	σ **(Å)**		
Rigid bond			C1	Sp
Rigid bond			C2	Sp
Rigid bond			C3	Sp
Rigid bond			C4	Sp
Rigid bond			C5	Sp
Rigid bond			C6	Sp
Rigid bond			C2	H2
Rigid bond			C3	H3
Rigid bond			C5	H4
Rigid bond			C6	H5
Harmonic bond	196981.5	1.088	C7	H6
Rigid bond			O1	H1
Harmonic bond	283571	0.973	O2	H7
Rigid bond			C1	O1
Harmonic bond	144178.3	1.476	C4	C7
Harmonic bond	227537.2	1.351	C7	C9
Harmonic bond	171679.5	1.427	C9	C8

Harmonic bond	141103.7	1.482	C9	C10	
Harmonic bond	442493.8	1.157	C8	N	
Harmonic bond	328388.1	1.218	C10	O3	
Harmonic bond	192799.7	1.351	C10	O2	

Bend stretch

Type	**Parameters**		**Atoms**		
	$\frac{\varepsilon}{k_B}$ **(K/rad²)**	σ **(°)**			
Harmonic bend	23874.90	115.1	C4	C7	H6
Harmonic bend	25338.27	118.2	C9	C7	H6
Harmonic bend	25974.85	106.6	C10	O2	H7
Harmonic bend	33506.58	120.8	C3	C4	C7
Harmonic bend	33080.86	127.5	C4	C7	C9
Harmonic bend	33506.58	120.8	C5	C4	C7
Harmonic bend	34345.96	123.1	C7	C9	C8
Harmonic bend	33140.24	126.41	C7	C9	C10
Harmonic bend	37688.36	177.97	C9	C8	N
Harmonic bend	43452.28	123.2	C9	C10	O3
Harmonic bend	43768.30	113.62	C9	C10	O2
Harmonic bend	33464.31	118.42	C8	C9	C10
Harmonic bend	58245.51	122.1	O3	C10	O2

Torsion

Type	**Parameters**			**Atoms**			
	$\frac{\varepsilon}{k_B}$ **(K)**	σ	p_2 **(°)**				
Torsion	3346.431	2.0	180.0	C3	C4	C7	H6
Torsion	3346.431	2.0	180.0	C5	C4	C7	H6
Torsion	1824.182	2.0	180.0	C7	C4	C3	H3
Torsion	1824.182	2.0	180.0	C7	C4	C5	H4
Torsion	1157.412	2.0	180.0	C9	C10	O2	H7
Torsion	3346.431	2.0	180.0	C8	C9	C7	H6
Torsion	3346.431	2.0	180.0	C10	C9	C7	H6
Torsion	956.1232	1.0	0.0	O3	C10	O2	H7
Torsion	1157.412	2.0	180.0	O3	C10	O2	H7
Torsion	1824.182	2.0	180.0	C2	C3	C4	C7
Torsion	3346.431	2.0	180.0	C3	C4	C7	C9
Torsion	3346.431	2.0	180.0	C4	C7	C9	C8
Torsion	3346.431	2.0	180.0	C4	C7	C9	C10
Torsion	3346.431	2.0	180.0	C5	C4	C7	C9
Torsion	1824.182	2.0	180.0	C6	C5	C4	C7
Torsion	0	2.0	180.0	C7	C9	C8	N
Torsion	1094.509	2.0	180.0	C7	C9	C10	O3
Torsion	1094.509	2.0	180.0	C7	C9	C10	O2

Torsion	1094.509	2.0	180.0	C8	C9	C10	O3
Torsion	1094.509	2.0	180.0	C8	C9	C10	O2
Torsion	0	2.0	180.0	C10	C9	C8	N

Improper torsion

Type	**Parameters**			**Atoms**			
	$\frac{\varepsilon}{k_B}$ **(K)**	σ	p_2 **(°)**				
Improper torsion	553.545	2.0	180.0	C4	C7	C9	H6
Improper torsion	553.545	2.0	180.0	C3	C4	C5	C7
Improper torsion	553.545	2.0	180.0	C10	C9	C7	C8
Improper torsion	5283.839	2.0	180.0	C9	C10	O3	O2

APPENDIX B MOLECULE PARAMETERS FOR THE STUDY OF CMOMS FOR CHIRAL SEPARATIONS

Below are the equations implemented in RASPA:

- Lennard-Jones

$$U = 4\varepsilon[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}] \quad \textbf{B-1}$$

where $\frac{\varepsilon}{k_B}$ is in K and σ in Å.

- Harmonic bond

$$U = \frac{1}{2}\varepsilon(r - \sigma)^2 \quad \textbf{B-2}$$

where $\frac{\varepsilon}{k_B}$ is in K/Å^2 and σ in Å.

- Harmonic bend

$$U = \frac{1}{2}\varepsilon\left(\theta_{ijk} - \sigma\right)^2 \quad \textbf{B-3}$$

where $\frac{\varepsilon}{k_B}$ is in K/rad^2 and σ in degrees.

- Torsion

$$U = \frac{1}{2}p_0(1 + \cos(\varphi)) + \frac{1}{2}p_1(1 - \cos(2\varphi)) + \frac{1}{2}p_2(1 + \cos(3\varphi)) \quad \textbf{B-4}$$

where p_0, p_1, p_2 are in kcal/mol.

- Improper torsion

$$U = p_0(1 + \cos(\varphi)) + p_1(1 + \cos(2\varphi)) + p_2(1 + \cos(3\varphi)) \quad \textbf{B-5}$$

where p_0, p_1, p_2 are in kcal/mol.

Table B1 Force field parameters for the CMOMs and the anions. The parameters are as defined in equation **B-1**.

	ε/k_B (K)	σ (Å)
Co_rig	7.045	2.559
N_rig	38.948	3.263
C_rig	47.854	3.474
C_piv	47.854	3.474
C_fle	33.23	3.5
O_rig	48.156	3.034
H_rig	7.649	2.847
H_fle	15.1	2.42
N_no3	85.59	3.15
O_no3	105.73	2.86
B_bf4	47.83	3.5814
F_bf4	30.21	3.1181
S_tof	125.876	3.55
C_tof	33.23	3.5
O_tof	105.736	2.96
F_tof	26.685	2.95

Molecule definitions

1P1P

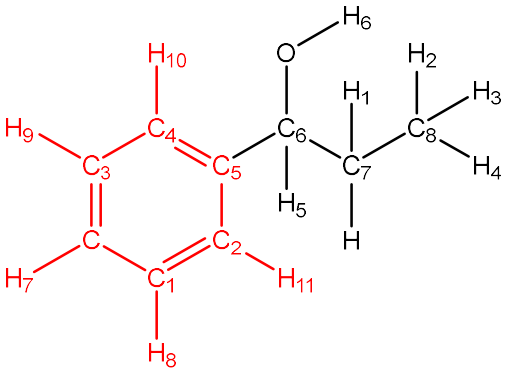


Figure B1 Indexing of the atoms in 1P1P. The rigid parts are coloured in red.

Table B2 Force field parameters for 1P1P. The parameters are as defined in equation **B-1**.

	ε/k_B (K)	**σ (Å)**
C	35.25061	3.55
C1	35.25061	3.55
C2	35.25061	3.55
C3	35.25061	3.55
C4	35.25061	3.55
C5	35.25061	3.55
C6	33.23628	3.5
C7	33.23628	3.5
C8	33.23628	3.5
H	15.1074	2.5
H1	15.1074	2.5

H2	15.1074	2.5
H3	15.1074	2.5
H4	15.1074	2.5
O	85.60861	3.07
H5	15.1074	2.5
H6	0	0
H7	15.1074	2.42
H8	15.1074	2.42
H9	15.1074	2.42
H10	15.1074	2.42
H11	15.1074	2.42

Table B3 Bond, bend and torsion definitions for 1P1P. The parameters are as defined in equations **B-2** to **B-5**.

Bond stretch				
Type of potential	**Parameters**		**Atoms**	
	$\frac{\varepsilon}{k_B}$ **(K/Å²)**	σ **(Å)**		
Harmonic bond	684373.0	1.090	C6	H5
Harmonic bond	684373.0	1.090	C7	H
Harmonic bond	684373.0	1.090	C7	H1
Harmonic bond	684373.0	1.090	C8	H2
Harmonic bond	684373.0	1.090	C8	H3

Harmonic bond	684373.0	1.090	C8	H4
Harmonic bond	1113119.5	0.945	O	H6
Harmonic bond	638092.3	1.510	C5	C6
Harmonic bond	539445.0	1.529	C6	C7
Harmonic bond	644130.0	1.410	C6	O
Harmonic bond	539445.0	1.529	C7	C8

Bend stretch

Type of potential	**Parameters**		**Atoms**		
	$\frac{\varepsilon}{k_B}$ **(K/rad²)**	σ **(°)**			
Harmonic bend	100643.8	109.500	C5	C6	H5
Harmonic bend	75482.9	110.700	C6	C7	C8
Harmonic bend	75482.85	110.700	C6	C7	H
Harmonic bend	23852.8	108.500	C7	C6	H5
Harmonic bend	75482.9	110.700	C7	C8	H2
Harmonic bend	75482.9	110.700	C7	C8	H3
Harmonic bend	75482.9	110.700	C7	C8	H4
Harmonic bend	75482.9	110.700	C8	C7	H
Harmonic bend	75482.9	110.700	C8	C7	H1
Harmonic bend	75482.9	110.700	H	C7	H1
Harmonic bend	66414.3	107.800	H4	C8	H3
Harmonic bend	66414.3	107.800	H4	C8	H2

Harmonic bend	66414.3	107.800	H3	C8	H2
Harmonic bend	70455.5	109.500	O	C6	H5
Harmonic bend	117457.9	112.700	C2	C5	C6
Harmonic bend	117457.9	112.700	C4	C5	C6
Harmonic bend	117457.9	112.700	C5	C6	C7
Harmonic bend	100643.8	109.500	C5	C6	O
Harmonic bend	117457.9	112.700	C6	C7	C8
Harmonic bend	100643.8	109.500	C7	C6	O

Torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
	(kcal/mol)						
Torsion	0	0	0	C2	C5	C6	H5
Torsion	0	0	0	C4	C5	C6	H5
Torsion	0	0	0.232	C5	C6	C7	H
Torsion	0	0	0.232	C5	C6	C7	H1
Torsion	-0.453	0	0	C5	C6	O	H6
Torsion	0	3.688	0	C6	C5	C2	H11
Torsion	0	3.688	0	C6	C5	C4	H10
Torsion	0	0	0.151	C6	C7	C8	H2
Torsion	0	0	0.151	C6	C7	C8	H3

Torsion	0	0	0.151	C6	C7	C8	H4
Torsion	-0.179	-0.088	0.248	C7	C6	O	H6
Torsion	0	0	0.151	C8	C7	C6	H5
Torsion	0	0	0.236	H	C7	C6	O
Torsion	0	0	0.151	H	C7	C6	H5
Torsion	0	0	0.151	H	C7	C8	H2
Torsion	0	0	0.151	H	C7	C8	H3
Torsion	0	0	0.151	H	C7	C8	H4
Torsion	0	0	0.236	H1	C7	C6	O
Torsion	0	0	0.151	H1	C7	C6	H5
Torsion	0	0	0.151	H1	C7	C8	H2
Torsion	0	0	0.151	H1	C7	C8	H3
Torsion	0	0	0.151	H1	C7	C8	H4
Torsion	0	0	0.226	H5	C6	O	H6
Torsion	0	3.648	0	C1	C2	C5	C6
Torsion	0	0	0	C2	C5	C6	C7
Torsion	0	0	0	C2	C5	C6	O
Torsion	0	3.648	0	C3	C4	C5	C6
Torsion	0	0	0	C4	C5	C6	C7
Torsion	0	0	0	C4	C5	C6	O
Torsion	0.654	-0.025	0.101	C5	C6	C7	C8

Torsion	0.861	-0.252	0.334	C8	C7	C6	O

Improper torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
	(kcal/mol)						
Improper torsion	0	1.107	0	C2	C5	C4	C6

1P2P

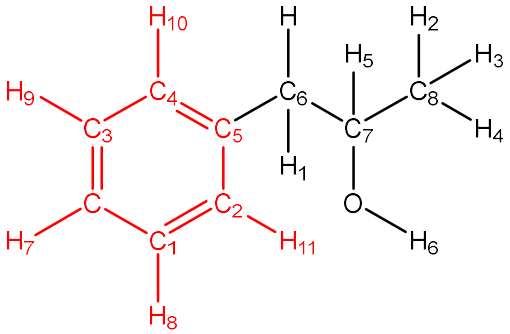


Figure B2 Indexing of the atoms in 1P2P. The rigid parts are coloured in red.

Table B4 Force field parameters for 1P2P. The parameters are as defined in equation **B-1**.

	ε/k_B (K)	**σ (Å)**
C	35.25061	3.55
C1	35.25061	3.55
C2	35.25061	3.55
C3	35.25061	3.55
C4	35.25061	3.55
C5	35.25061	3.55
C6	33.23628	3.5
C7	33.23628	3.5
C8	33.23628	3.5
H	15.1074	2.5
H1	15.1074	2.5
H2	15.1074	2.5

H3	15.1074	2.5
H4	15.1074	2.5
O	85.60861	3.07
H5	15.1074	2.5
H6	0	0
H7	15.1074	2.42
H8	15.1074	2.42
H9	15.1074	2.42
H10	15.1074	2.42
H11	15.1074	2.42

Table B5 Bond, bend and torsion definitions for 1P2P. The parameters are as defined in equations **B-2** to **B-5**.

Bond stretch				
Type of potential	**Parameters**		**Atoms**	
	$\frac{\varepsilon}{k_B}$ **(K/Å²)**	σ **(Å)**		
Harmonic bond	684373.0	1.090	C7	H5
Harmonic bond	684373.0	1.090	C6	H
Harmonic bond	684373.0	1.090	C6	H1
Harmonic bond	684373.0	1.090	C8	H2
Harmonic bond	684373.0	1.090	C8	H3
Harmonic bond	684373.0	1.090	C8	H4

Harmonic bond	1113119.5	0.945	O	H6
Harmonic bond	638092.3	1.510	C5	C6
Harmonic bond	539445.0	1.529	C6	C7
Harmonic bond	644130.0	1.410	C7	O
Harmonic bond	539445.0	1.529	C7	C8

Bend stretch

Type of potential	**Parameters**		**Atoms**		
	$\frac{\varepsilon}{k_B}$ **(K/rad²)**	σ **(°)**			
Harmonic bend	100643.8	109.500	C5	C6	H
Harmonic bend	100643.8	109.500	C5	C6	H1
Harmonic bend	117457.9	112.700	C6	C7	C8
Harmonic bend	75482.9	110.700	C7	C6	H
Harmonic bend	75482.9	110.700	C7	C6	H1
Harmonic bend	23852.8	108.500	C7	O	H6
Harmonic bend	75482.9	110.700	C6	C7	H5
Harmonic bend	75482.9	110.700	C7	C8	H2
Harmonic bend	75482.9	110.700	C7	C8	H3
Harmonic bend	75482.9	110.700	C7	C8	H4
Harmonic bend	75482.9	110.700	C8	C7	H5
Harmonic bend	75482.9	110.700	H	C6	H1
Harmonic bend	66414.3	107.800	H4	C8	H3

Harmonic bend	66414.3	107.800	H4	C8	H2
Harmonic bend	66414.3	107.800	H3	C8	H2
Harmonic bend	70455.5	109.500	O	C7	C8
Harmonic bend	117457.9	112.700	C2	C5	C6
Harmonic bend	117457.9	112.700	C4	C5	C6
Harmonic bend	117457.9	112.700	C5	C6	C7
Harmonic bend	100643.8	109.500	C6	C7	O
Harmonic bend	100643.8	109.500	H5	C7	O

Torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
	(kcal/mol)						
Torsion	0	0	0	C2	C5	C6	H
Torsion	0	0	0	C2	C5	C6	H1
Torsion	0	0	0	C2	C5	C6	C7
Torsion	0	3.648	0	C1	C2	C5	C6
Torsion	0	3.648	0	C3	C4	C5	C6
Torsion	0	0	0	C4	C5	C6	H
Torsion	0	0	0	C4	C5	C6	H1
Torsion	0	0	0	C4	C5	C6	C7
Torsion	0	0	0.232	C5	C6	C7	H5

Torsion	0.654	-0.025	0.101	C5	C6	C7	C8
Torsion	-0.861	-0.252	0.334	C5	C6	C7	O
Torsion	0	3.688	0	C6	C5	C2	H11
Torsion	0	3.688	0	C6	C5	C4	H10
Torsion	0	0	0.151	C8	C7	C6	H
Torsion	0	0	0.151	C8	C7	C6	H1
Torsion	-0.179	-0.088	0.248	C8	C7	O	H6
Torsion	0	0	0.236	H	C6	C7	O
Torsion	0	0	0.151	H	C6	C7	H5
Torsion	0	0	0.151	H5	C7	C8	H2
Torsion	0	0	0.151	H5	C7	C8	H3
Torsion	0	0	0.151	H5	C7	C8	H4
Torsion	0	0	0.236	H1	C6	C7	O
Torsion	0	0	0.151	H1	C6	C7	H5
Torsion	0	0	0.236	O	C7	C8	H2
Torsion	0	0	0.236	O	C7	C8	H3
Torsion	0	0	0.236	O	C7	C8	H4
Torsion	0	0	0.226	H5	C7	O	H6

Improper torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
	(kcal/mol)						
Improper torsion	0	1.107	0	C2	C5	C4	C6

2P1P

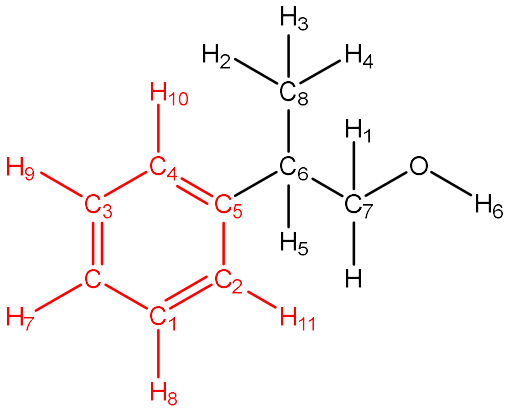


Figure B3 Indexing of the atoms in 2P1P. The rigid parts are coloured in red.

Table B6 Force field parameters for 2P1P. The parameters are as defined in equation **B-1**.

	ε/k_B (K)	**σ (Å)**
C	35.25061	3.55
C1	35.25061	3.55
C2	35.25061	3.55
C3	35.25061	3.55
C4	35.25061	3.55
C5	35.25061	3.55
C6	33.23628	3.5
C7	33.23628	3.5
C8	33.23628	3.5
H	15.1074	2.5
H1	15.1074	2.5

H2	15.1074	2.5
H3	15.1074	2.5
H4	15.1074	2.5
O	85.60861	3.07
H5	15.1074	2.5
H6	0	0
H7	15.1074	2.42
H8	15.1074	2.42
H9	15.1074	2.42
H10	15.1074	2.42
H11	15.1074	2.42

Table B7 Bond, bend and torsion definitions for 2P1P. The parameters are as defined in equations **B-2** to **B-5**.

Bond stretch				
Type of potential	**Parameters**		**Atoms**	
	$\frac{\varepsilon}{k_B}$ **(K/Å²)**	σ **(Å)**		
Harmonic bond	684373.0	1.090	C6	H5
Harmonic bond	684373.0	1.090	C7	H
Harmonic bond	684373.0	1.090	C7	H1
Harmonic bond	684373.0	1.090	C8	H2
Harmonic bond	684373.0	1.090	C8	H3

Harmonic bond	684373.0	1.090	C8	H4
Harmonic bond	1113119.5	0.945	O	H6
Harmonic bond	638092.3	1.510	C5	C6
Harmonic bond	539445.0	1.529	C6	C7
Harmonic bond	644130.0	1.410	C7	O
Harmonic bond	539445.0	1.529	C6	C8

Bend stretch

Type of potential	**Parameters**		**Atoms**		
	$\frac{\varepsilon}{k_B}$ **(K/rad²)**	σ **(°)**			
Harmonic bend	100643.8	109.500	C5	C6	H5
Harmonic bend	75482.9	110.700	C6	C7	H1
Harmonic bend	75482.85	110.700	C6	C7	H
Harmonic bend	23852.8	108.500	C7	O	H6
Harmonic bend	75482.9	110.700	C6	C8	H2
Harmonic bend	75482.9	110.700	C6	C8	H3
Harmonic bend	75482.9	110.700	C6	C8	H4
Harmonic bend	75482.9	110.700	C7	C6	H5
Harmonic bend	75482.9	110.700	C8	C6	H5
Harmonic bend	75482.9	110.700	H	C7	H1
Harmonic bend	66414.3	107.800	H4	C8	H3
Harmonic bend	66414.3	107.800	H4	C8	H2

Harmonic bend	66414.3	107.800	H3	C8	H2
Harmonic bend	70455.5	109.500	O	C7	H
Harmonic bend	70455.5	109.500	O	C7	H1
Harmonic bend	117457.9	112.700	C2	C5	C6
Harmonic bend	117457.9	112.700	C4	C5	C6
Harmonic bend	117457.9	112.700	C5	C6	C7
Harmonic bend	117457.9	112.700	C7	C6	C8
Harmonic bend	117457.9	112.700	C5	C6	C8
Harmonic bend	100643.8	109.500	C6	C7	O

Torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
	(kcal/mol)						
Torsion	0	0	0	C2	C5	C6	C8
Torsion	0	0	0	C4	C5	C6	C8
Torsion	0	0	0	C2	C5	C6	H5
Torsion	0	0	0	C4	C5	C6	H5
Torsion	0	0	0.232	C5	C6	C7	H1
Torsion	0	0	0.232	C5	C6	C7	H
Torsion	0	0	0.232	C5	C6	C8	H2
Torsion	0	0	0.232	C5	C6	C8	H3

Torsion	0	0	0.232	C5	C6	C8	H4
Torsion	0	3.688	0	C6	C5	C2	H11
Torsion	0	3.688	0	C6	C5	C4	H10
Torsion	0	0	0.151	C7	C6	C8	H2
Torsion	0	0	0.151	C6	C7	C8	H3
Torsion	0	0	0.151	C6	C7	C8	H4
Torsion	-0.179	-0.088	0.248	C6	C7	O	H6
Torsion	0	0	0.236	H5	C6	C7	O
Torsion	0	0	0.151	H5	C6	C8	H4
Torsion	0	0	0.151	H5	C6	C8	H2
Torsion	0	0	0.151	H5	C6	C8	H3
Torsion	0	0	0.236	H5	C6	C7	O
Torsion	0	0	0.151	C8	C6	C7	H
Torsion	0	0	0.151	C8	C6	C7	H1
Torsion	0	0	0.151	H	C7	C6	H5
Torsion	0	0	0.151	H1	C7	C6	H5
Torsion	0	0	0.226	H	C7	O	H6
Torsion	0	0	0.226	H1	C7	O	H6
Torsion	0	3.648	0	C1	C2	C5	C6
Torsion	0	3.648	0	C3	C4	C5	C6
Torsion	0	0	0	C2	C5	C6	C7

Torsion	0	0	0	C4	C5	C6	C7
Torsion	0.861	-0.025	0.334	C5	C6	C7	O
Torsion	0.861	-0.252	0.334	C8	C6	C7	O

Improper torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
	(kcal/mol)						
Improper torsion	0	1.107	0	C2	C5	C4	C6

1P1Pen

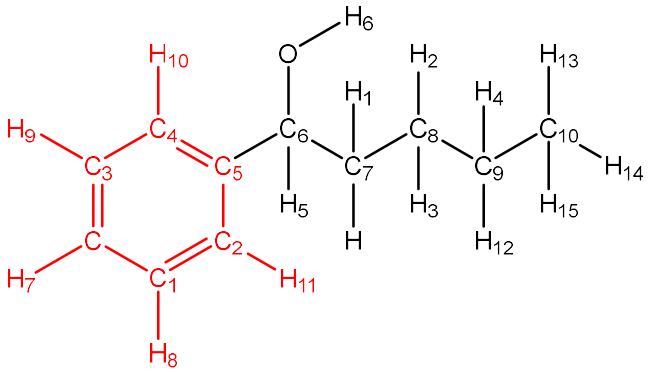


Figure B4 Indexing of the atoms in 1P1Pen. The rigid parts are coloured in red.

Table B8 Force field parameters for 1P1Pen. The parameters are as defined in equation **B-1**.

	ε/k_B (K)	**σ (Å)**
C	35.25061	3.55
C1	35.25061	3.55
C2	35.25061	3.55
C3	35.25061	3.55
C4	35.25061	3.55
C5	35.25061	3.55
C6	33.23628	3.5
C7	33.23628	3.5
C8	33.23628	3.5
C9	33.23628	3.5
C10	33.23628	3.5
H	15.1074	2.5

H1	15.1074	2.5
H2	15.1074	2.5
H3	15.1074	2.5
H4	15.1074	2.5
O	85.60861	3.07
H5	15.1074	2.5
H6	0	0
H7	15.1074	2.42
H8	15.1074	2.42
H9	15.1074	2.42
H10	15.1074	2.42
H11	15.1074	2.42
H12	15.1074	2.5
H13	15.1074	2.5
H14	15.1074	2.5
H15	15.1074	2.5

Table B9 Bond, bend and torsion definitions for 1P1Pen. The parameters are as defined in equations **B-2** to **B-5**.

Bond stretch

Type of potential	Parameters		Atoms	
	$\frac{\varepsilon}{k_B}$ (K/Å^2)	σ (Å)		
Harmonic bond	684373.0	1.090	C6	H5
Harmonic bond	684373.0	1.090	C7	H
Harmonic bond	684373.0	1.090	C7	H1
Harmonic bond	684373.0	1.090	C8	H2
Harmonic bond	684373.0	1.090	C8	H3
Harmonic bond	539445.0	1.529	C8	C9
Harmonic bond	1113119.5	0.945	O	H6
Harmonic bond	638092.3	1.510	C5	C6
Harmonic bond	539445.0	1.529	C6	C7
Harmonic bond	644130.0	1.410	C6	O
Harmonic bond	539445.0	1.529	C7	C8
Harmonic bond	684373.0	1.090	C9	H4
Harmonic bond	684373.0	1.090	C9	H12
Harmonic bond	539445.0	1.529	C10	C9
Harmonic bond	684373.0	1.090	C10	H13
Harmonic bond	684373.0	1.090	C10	H14
Harmonic bond	684373.0	1.090	C10	H15

Bend stretch

Type of potential	**Parameters**		**Atoms**		
	$\frac{\varepsilon}{k_B}$ **(K/rad²)**	σ **(°)**			
Harmonic bend	100643.8	109.500	C5	C6	H5
Harmonic bend	75482.9	110.700	C6	C7	H1
Harmonic bend	75482.9	110.700	C6	C7	H
Harmonic bend	23852.8	108.500	C6	O	H6
Harmonic bend	75482.9	110.700	C7	C6	H5
Harmonic bend	75482.9	110.700	C7	C8	H2
Harmonic bend	75482.9	110.700	C7	C8	H3
Harmonic bend	117457.9	112.700	C7	C8	C9
Harmonic bend	75482.9	110.700	C8	C7	H
Harmonic bend	75482.9	110.700	C8	C7	H1
Harmonic bend	75482.9	110.700	H	C7	H1
Harmonic bend	66414.3	107.800	H2	C8	H3
Harmonic bend	75482.9	110.700	H2	C8	C9
Harmonic bend	75482.9	110.700	H3	C8	C9
Harmonic bend	70455.5	109.500	O	C6	H5
Harmonic bend	117457.9	112.700	C2	C5	C6
Harmonic bend	117457.9	112.700	C4	C5	C6
Harmonic bend	117457.9	112.700	C5	C6	C7

Harmonic bend	100643.8	109.500	C5	C6	O
Harmonic bend	117457.9	112.700	C6	C7	C8
Harmonic bend	100643.8	109.500	C7	C6	O
Harmonic bend	75482.9	110.700	C8	C9	H4
Harmonic bend	75482.9	110.700	C8	C9	H12
Harmonic bend	75482.9	110.700	H4	C9	C10
Harmonic bend	75482.9	110.700	H12	C9	C10
Harmonic bend	75482.9	110.700	C9	C10	H13
Harmonic bend	75482.9	110.700	C9	C10	H14
Harmonic bend	75482.9	110.700	C9	C10	H15

Torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
	(kcal/mol)						
Torsion	0	0	0	C2	C5	C6	H5
Torsion	0	0	0	C4	C5	C6	H5
Torsion	0	0	0.232	C5	C6	C7	H
Torsion	0	0	0.232	C5	C6	C7	H1
Torsion	-0.453	0	0	C5	C6	O	H6
Torsion	0	3.688	0	C6	C5	C2	H11
Torsion	0	3.688	0	C6	C5	C4	H10

Torsion	0	0	0.151	C6	C7	C8	H2
Torsion	0	0	0.151	C6	C7	C8	H3
Torsion	0.654	-0.025	0.101	C6	C7	C8	C9
Torsion	-0.179	-0.088	0.248	C7	C6	O	H6
Torsion	0	0	0.151	C8	C7	C6	H5
Torsion	0	0	0.236	H	C7	C6	O
Torsion	0	0	0.151	H	C7	C6	H5
Torsion	0	0	0.151	H	C7	C8	H2
Torsion	0	0	0.151	H	C7	C8	H3
Torsion	0	0	0.151	H	C7	C8	H4
Torsion	0	0	0.236	H1	C7	C6	O
Torsion	0	0	0.151	H1	C7	C6	H5
Torsion	0	0	0.151	H1	C7	C8	H2
Torsion	0	0	0.151	H1	C7	C8	H3
Torsion	0	0	0.151	H1	C7	C8	H4
Torsion	0	0	0.226	H5	C6	O	H6
Torsion	0	3.648	0	C1	C2	C5	C6
Torsion	0	0	0	C2	C5	C6	C7
Torsion	0	0	0	C2	C5	C6	O
Torsion	0	3.648	0	C3	C4	C5	C6
Torsion	0	0	0	C4	C5	C6	C7

Torsion	0	0	0	C4	C5	C6	O
Torsion	0.654	-0.025	0.101	C5	C6	C7	C8
Torsion	0.861	-0.252	0.334	C8	C7	C6	O
Torsion	0	0	0.151	C7	C8	C9	H4
Torsion	0	0	0.151	C7	C8	C9	H12
Torsion	0	0	0.151	H2	C8	C9	H4
Torsion	0	0	0.151	H2	C8	C9	H12
Torsion	0	0	0.151	H3	C8	C9	H4
Torsion	0	0	0.151	H3	C8	C9	H12
Torsion	0	0	0.151	C7	C8	C9	H13
Torsion	0	0	0.151	C7	C8	C9	H14
Torsion	0	0	0.151	C7	C8	C9	H15
Torsion	0	0	0.151	H4	C8	C9	H13
Torsion	0	0	0.151	H4	C8	C9	H14
Torsion	0	0	0.151	H4	C8	C9	H15
Torsion	0	0	0.151	H12	C8	C9	H13
Torsion	0	0	0.151	H12	C8	C9	H14
Torsion	0	0	0.151	H12	C8	C9	H15
Torsion	0	0	0.151	H2	C7	C8	C9
Torsion	0	0	0.151	H3	C7	C8	C9

Improper torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
	(kcal/mol)						
Improper torsion	0	1.107	0	C2	C5	C4	C6

BM

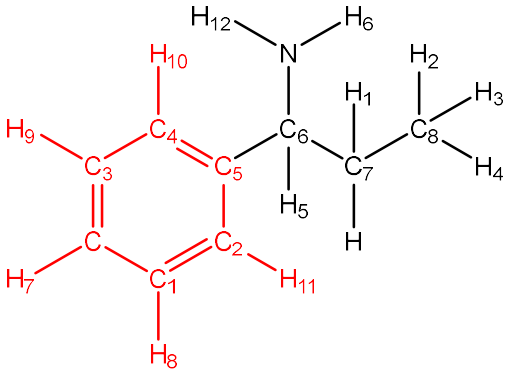


Figure B5 Indexing of the atoms in BM. The rigid parts are coloured in red.

Table B10 Force field parameters for BM. The parameters are as defined in equation **B-1**.

	ε/k_B (K)	**σ (Å)**
C	35.25061	3.55
C1	35.25061	3.55
C2	35.25061	3.55
C3	35.25061	3.55
C4	35.25061	3.55
C5	35.25061	3.55
C6	33.23628	3.5
C7	33.23628	3.5
C8	33.23628	3.5
H	15.1074	2.5
H1	15.1074	2.5
H2	15.1074	2.5

H3	15.1074	2.5
H4	15.1074	2.5
N	85.60861	3.3
H5	15.1074	2.5
H6	0	0
H7	15.1074	2.42
H8	15.1074	2.42
H9	15.1074	2.42
H10	15.1074	2.42
H11	15.1074	2.42
H12	0	0

Table B11 Bond, bend and torsion definitions for BM. The parameters are as defined in equations **B-2** to **B-5**.

Bond stretch				
Type of potential	**Parameters**		**Atoms**	
	$\frac{\varepsilon}{k_B}$ **(K/Å²)**	σ **(Å)**		
Harmonic bond	684373.0	1.090	C6	H5
Harmonic bond	684373.0	1.090	C7	H
Harmonic bond	684373.0	1.090	C7	H1
Harmonic bond	684373.0	1.090	C8	H2
Harmonic bond	684373.0	1.090	C8	H3

Harmonic bond	684373.0	1.090	C8	H4
Harmonic bond	873585.3	0.010	N	H6
Harmonic bond	873585.3	0.010	N	H12
Harmonic bond	638092.3	1.510	C5	C6
Harmonic bond	539445.0	1.529	C6	C7
Harmonic bond	986304.4	1.335	C6	N
Harmonic bond	539445.0	1.529	C7	C8

Bend stretch

Type of potential	**Parameters**		**Atoms**		
	$\frac{\varepsilon}{k_B}$ **(K/rad²)**	σ **(°)**			
Harmonic bend	100643.8	109.500	C5	C6	H5
Harmonic bend	75482.9	110.700	C6	C7	H1
Harmonic bend	75482.9	110.700	C6	C7	H
Harmonic bend	70455.5	119.800	C6	N	H6
Harmonic bend	70455.5	119.800	C6	N	H12
Harmonic bend	75482.9	110.700	C7	C6	H5
Harmonic bend	75482.9	110.700	C7	C8	H2
Harmonic bend	75482.9	110.700	C7	C8	H3
Harmonic bend	75482.9	110.700	C7	C8	H4
Harmonic bend	75482.9	110.700	C8	C7	H
Harmonic bend	75482.9	110.700	C8	C7	H1

Harmonic bend	75482.9	110.700	H	C7	H1
Harmonic bend	66414.3	107.800	H4	C8	H3
Harmonic bend	66414.3	107.800	H4	C8	H2
Harmonic bend	66414.3	107.800	H3	C8	H2
Harmonic bend	161030.1	109.700	C7	C6	N
Harmonic bend	117457.9	112.700	C2	C5	C6
Harmonic bend	117457.9	112.700	C4	C5	C6
Harmonic bend	70455.5	120.000	H6	N	H12
Harmonic bend	80515.0	114.000	N	C6	H5
Harmonic bend	117457.9	112.700	C5	C6	C7
Harmonic bend	117457.9	112.700	C6	C7	C8
Harmonic bend	161030.1	109.700	C5	C6	N

Torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
	(kcal/mol)						
Torsion	0	0	0	C2	C5	C6	H5
Torsion	0	0	0	C4	C5	C6	H5
Torsion	0	0	0.232	C5	C6	C7	H
Torsion	0	0	0.232	C5	C6	C7	H1
Torsion	0	3.688	0	C6	C5	C2	H11

Torsion	0	3.688	0	C6	C5	C4	H10
Torsion	0	0	0.151	C6	C7	C8	H2
Torsion	0	0	0.151	C6	C7	C8	H3
Torsion	0	0	0.151	C6	C7	C8	H4
Torsion	0	0	0.151	C8	C7	C6	H5
Torsion	0	0	0.236	H	C7	C6	N
Torsion	0	0	0.151	H	C7	C8	H4
Torsion	0	0	0.151	H	C7	C8	H2
Torsion	0	0	0.151	H	C7	C8	H3
Torsion	0	0	0.151	H1	C7	C6	N
Torsion	0	0	0.151	H1	C7	C6	H4
Torsion	0	0	0.151	H1	C7	C8	H2
Torsion	0	0	0.151	H1	C7	C8	H3
Torsion	0	0	0.234	H1	C7	C6	H5
Torsion	0	0	0.234	H	C7	C6	N
Torsion	0	2.466	0	C5	C6	N	H6
Torsion	0	2.466	0	C5	C6	N	H12
Torsion	0	2.466	0	C7	C6	N	H6
Torsion	0	2.466	0	C7	C6	N	H12
Torsion	0	2.466	0	H5	C6	N	H6
Torsion	0	2.466	0	H5	C6	N	H12

Torsion	0	3.648	0	C1	C2	C5	C6
Torsion	0	0	0	C2	C5	C6	C7
Torsion	0	0.554	0	C2	C5	C6	N
Torsion	0	3.648	0	C3	C4	C5	C6
Torsion	0	0	0	C4	C5	C6	C7
Torsion	0	0.554	0	C4	C5	C6	N
Torsion	0.654	-0.025	0.101	C5	C6	C7	C8
Torsion	0.989	0	0.332	C8	C7	C6	N

Improper torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
	(kcal/mol)						
Improper torsion	0	1.107	0	C2	C5	C4	C6

EBM

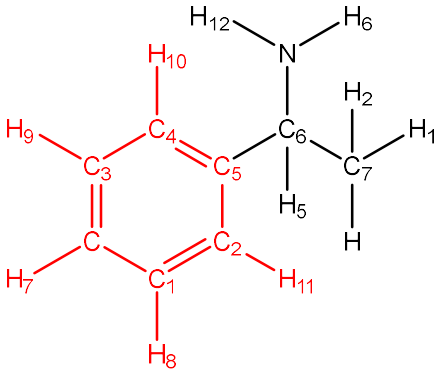


Figure B6 Indexing of the atoms in EBM. The rigid parts are coloured in red.

Table B12 Force field parameters for EBM. The parameters are as defined in equation **B-1**.

	ε/k_B (K)	σ (Å)
C	35.25061	3.55
C1	35.25061	3.55
C2	35.25061	3.55
C3	35.25061	3.55
C4	35.25061	3.55
C5	35.25061	3.55
C6	33.23628	3.5
C7	33.23628	3.5
C8	33.23628	3.5
H	15.1074	2.5
H1	15.1074	2.5
H2	15.1074	2.5

H3	15.1074	2.5
H4	15.1074	2.5
N	85.60861	3.3
H5	15.1074	2.5
H6	0	0
H7	15.1074	2.42
H8	15.1074	2.42
H10	15.1074	2.42
H11	15.1074	2.42
H9	15.1074	2.42
H12	0	0

Table B13 Bond, bend and torsion definitions for EBM. The parameters are as defined in equations **B-2** to **B-5**.

Bond stretch				
Type of potential	**Parameters**		**Atoms**	
	$\frac{\varepsilon}{k_B}$ **(K/Å²)**	σ **(Å)**		
Harmonic bond	684373.0	1.090	C6	H5
Harmonic bond	684373.0	1.090	C7	H
Harmonic bond	684373.0	1.090	C7	H1
Harmonic bond	684373.0	1.090	C7	H2
Harmonic bond	873585.3	1.010	N	H6

Harmonic bond	873585.3	1.010	N	H12	
Harmonic bond	638092.3	1.510	C5	C6	

Bend stretch

Type of potential	Parameters		Atoms		
	$\frac{\varepsilon}{k_B}$ (K/rad^2)	σ (°)			
Harmonic bend	100643.8	109.500	C5	C6	H5
Harmonic bend	75482.9	110.700	C6	C7	H
Harmonic bend	75482.9	110.700	C6	C7	H1
Harmonic bend	75482.9	110.700	C6	C7	H2
Harmonic bend	70455.5	119.800	C6	N	H6
Harmonic bend	70455.5	119.800	C6	N	H12
Harmonic bend	75482.9	110.700	C7	C6	H5
Harmonic bend	75482.9	110.700	H	C7	H1
Harmonic bend	75482.9	110.700	H	C7	H2
Harmonic bend	75482.9	110.700	H1	C7	H2
Harmonic bend	80515.4	114.000	N	C6	H5
Harmonic bend	161030.1	109.700	C7	C6	N
Harmonic bend	66414.3	107.800	H6	C6	H12
Harmonic bend	117457.9	112.700	C2	C5	C6
Harmonic bend	117457.9	112.700	C4	C5	C6
Harmonic bend	117457.9	112.700	C5	C6	C7

Harmonic bend	161030.8	109.700	C5	C6	N		

Torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
	(kcal/mol)						
Torsion	0	0	0	C2	C5	C6	H5
Torsion	0	0	0	C4	C5	C6	H5
Torsion	0	0	0.232	C5	C6	C7	H
Torsion	0	0	0.232	C5	C6	C7	H1
Torsion	0	0	0.232	C5	C6	C7	H2
Torsion	0	2.466	0	C5	C6	N	H6
Torsion	0	2.466	0	C5	C6	N	H12
Torsion	0	2.466	0	C7	C6	N	H6
Torsion	0	2.466	0	C7	C6	N	H12
Torsion	0	2.466	0	H5	C6	N	H6
Torsion	0	2.466	0	H5	C6	N	H12
Torsion	0	3.688	0	C6	C5	C4	H10
Torsion	0	0	0.151	H	C7	C6	H5
Torsion	0	0	0.151	H1	C7	C6	H5
Torsion	0	0	0.151	H2	C7	C6	H5
Torsion	0	0	0.236	H	C7	C6	N

Torsion	0	0	0.151	H1	C7	C6	N
Torsion	0	0	0.151	H2	C7	C8	N
Torsion	0	3.648	0	C1	C2	C5	C6
Torsion	0	0	0	C2	C5	C6	C7
Torsion	0	0	0	C2	C5	C6	N
Torsion	0	3.648	0	C3	C4	C5	C6
Torsion	0	0	0	C4	C5	C6	C7
Torsion	0	0	0	C4	C5	C6	N

Improper torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
	(kcal/mol)						
Improper torsion	0	1.107	0	C2	C5	C4	C6

PE, 3PE and 4PE

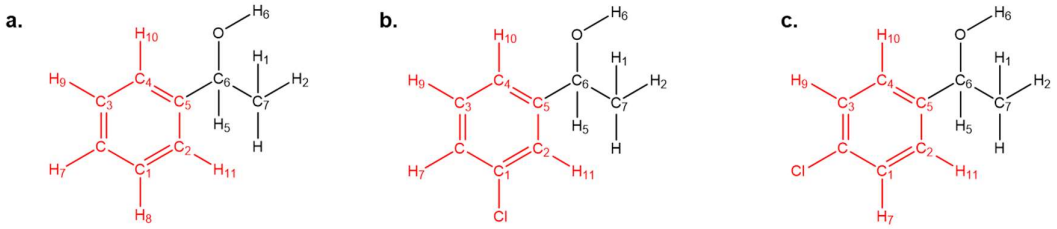


Figure B7 Indexing of the atoms in **a.** PE, **b.** 3PE and **c.** 4PE. The rigid parts are coloured in red.

Table B14 Force field parameters for PE, 3PE and 4PE. The parameters are as defined in equation **B-1**.

	ε/k_B (K)	σ (Å)
C	35.25061	3.55
C1	35.25061	3.55
C2	35.25061	3.55
C3	35.25061	3.55
C4	35.25061	3.55
C5	35.25061	3.55
C6	33.23628	3.5
C7	33.23628	3.5
H	15.1074	2.5
H1	15.1074	2.5
H2	15.1074	2.5
O	85.60861	3.07
H5	15.1074	2.5

H6	0	0
H7	15.1074	2.42
H9	15.1074	2.42
H10	15.1074	2.42
H11	15.1074	2.42
Cl	151.074	3.40
H8 (PE)	15.1074	2.42

Table B15 Bond, bend and torsion definitions for PE, 3PE and 4PE. The parameters are as defined in equations **B-2** to **B-5**.

Bond stretch

Type of potential	**Parameters**		**Atoms**	
	$\frac{\varepsilon}{k_B}$ **(K/Å²)**	σ **(Å)**		
Harmonic bond	539445.0	1.529	C6	H5
Harmonic bond	644130.0	1.410	C7	H
Harmonic bond	684373.0	1.090	C6	H5
Harmonic bond	684373.0	1.090	C7	H2
Harmonic bond	684373.0	1.090	C7	H
Harmonic bond	684373.0	1.090	C7	H1
Harmonic bond	1113119.5	0.945	O	H6
Harmonic bond	638092.3	1.510	C5	C6

Bend stretch

Type of potential	**Parameters**		**Atoms**		
	$\frac{\varepsilon}{k_B}$ **(K/rad²)**	σ **(°)**			
Harmonic bend	100643.8	109.500	C5	C6	H5
Harmonic bend	75482.9	110.700	C6	C7	H1
Harmonic bend	75482.85	110.700	C6	C7	H
Harmonic bend	75482.85	110.700	C6	C7	H2
Harmonic bend	23852.8	108.500	C6	O	H6
Harmonic bend	75482.9	110.700	C7	C6	H5
Harmonic bend	75482.9	110.700	H2	C7	H
Harmonic bend	75482.9	110.700	H2	C7	H1
Harmonic bend	75482.9	110.700	H	C7	H1
Harmonic bend	70455.5	109.500	O	C6	H5
Harmonic bend	117457.9	112.700	C2	C5	C6
Harmonic bend	117457.9	112.700	C4	C5	C6
Harmonic bend	117457.9	112.700	C5	C6	C7
Harmonic bend	100643.8	109.500	C5	C6	O
Harmonic bend	100643.8	109.500	C7	C6	O

Torsion

Type of potential	Parameters			Atoms			
	p_0	p_1	p_2				
	(kcal/mol)						
Torsion	0	0	0	C2	C5	C6	H5
Torsion	0	0	0	C4	C5	C6	H5
Torsion	0	0	0.232	C5	C6	C7	H
Torsion	0	0	0.232	C5	C6	C7	H1
Torsion	0	0	0.232	C5	C6	C7	H2
Torsion	-0.453	0	0	C5	C6	O	H6
Torsion	0	3.688	0	C6	C5	C2	H11
Torsion	0	3.688	0	C6	C5	C4	H10
Torsion	-0.179	-0.088	0.248	C7	C6	O	H6
Torsion	0	0	0.151	C8	C7	C6	H5
Torsion	0	0	0.236	H	C7	C6	O
Torsion	0	0	0.236	H1	C7	C6	O
Torsion	0	0	0.151	H	C7	C6	H5
Torsion	0	0	0.236	H2	C7	C6	O
Torsion	0	0	0.151	H1	C7	C6	H5
Torsion	0	0	0.226	H5	C6	O	H6
Torsion	0	3.648	0	C1	C2	C5	C6

Torsion	0	0	0	C2	C5	C6	C7
Torsion	0	0	0	C2	C5	C6	O
Torsion	0	3.648	0	C3	C4	C5	C6
Torsion	0	0	0	C4	C5	C6	C7
Torsion	0	0	0	C4	C5	C6	O

Improper torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
	(kcal/mol)						
Improper torsion	0	1.107	0	C2	C5	C4	C6

CPBA

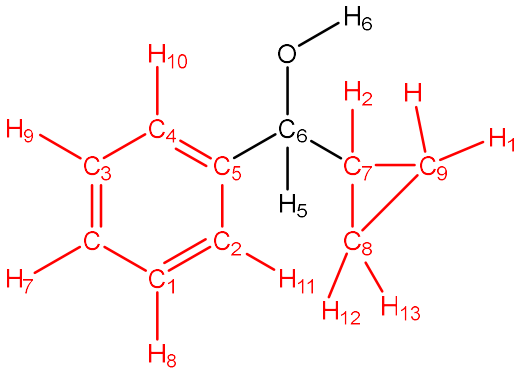


Figure B8 Indexing of the atoms in CPBA. The rigid parts are coloured in red.

Table B16 Force field parameters for CPBA. The parameters are as defined in equation **B-1**.

	ε/k_B (K)	**σ (Å)**
C	35.25061	3.55
C1	35.25061	3.55
C2	35.25061	3.55
C3	35.25061	3.55
C4	35.25061	3.55
C5	35.25061	3.55
C6	33.23628	3.5
C7	33.23628	3.5
C8	33.23628	3.5
C9	33.23628	3.5
H	15.1074	2.5
H1	15.1074	2.5

H2	15.1074	2.5
H3	15.1074	2.5
H4	15.1074	2.5
O	85.60861	3.07
H5	15.1074	2.5
H6	0	0
H7	15.1074	2.42
H8	15.1074	2.42
H9	15.1074	2.42
H10	15.1074	2.42

Table B17 Bond, bend and torsion definitions for CPBA. The parameters are as defined in equations **B-2** to **B-5**.

Bond stretch				
Type of potential	**Parameters**		**Atoms**	
	$\frac{\varepsilon}{k_B}$ **(K/Å²)**	σ **(Å)**		
Harmonic bond	684373.0	1.090	C6	H5
Harmonic bond	1113119.5	0.945	O	H6
Harmonic bond	638092.3	1.510	C5	C6
Harmonic bond	644130.0	1.410	C6	O
Harmonic bond	539445.0	1.529	C6	C7

Bend stretch

Type of potential	Parameters		Atoms		
	$\frac{\varepsilon}{k_B}$ (K/rad^2)	σ (°)			
Harmonic bend	100643.8	109.500	C5	C6	O
Harmonic bend	75482.9	110.700	C7	C6	H5
Harmonic bend	75482.85	110.700	C6	C7	H2
Harmonic bend	23852.8	108.500	C6	O	H6
Harmonic bend	70455.5	109.500	C7	C6	O
Harmonic bend	117457.9	112.700	C2	C5	C6
Harmonic bend	117457.9	112.700	C4	C5	C6
Harmonic bend	117457.9	112.700	C5	C6	C7
Harmonic bend	100643.8	109.500	C5	C6	H5
Harmonic bend	117457.9	112.700	C6	C7	C8
Harmonic bend	117457.9	112.700	C6	C7	C9
Harmonic bend	100643.8	109.500	H5	C6	O

Torsion

Type of potential	Parameters			Atoms			
	p_0	p_1	p_2				
	(kcal/mol)						
Torsion	0	0	0	C2	C5	C6	C7
Torsion	0	0	0	C2	C5	C6	H5

Torsion	0	0	0	C2	C5	C6	O
Torsion	0	0	0	C4	C5	C6	C7
Torsion	0	0	0	C4	C5	C6	H5
Torsion	0	0	0	C4	C5	C6	O
Torsion	0	0	0.236	O	C6	C7	H2
Torsion	0	0	0.232	C5	C6	C7	H2
Torsion	-0.453	0	0	C5	C6	O	H6
Torsion	0	3.688	0	C6	C5	C2	H11
Torsion	0	3.688	0	C6	C5	C4	H10
Torsion	-0.179	-0.088	0.248	C7	C6	O	H6
Torsion	0	0	0.226	H5	C6	O	H6
Torsion	0	0	0.151	H2	C7	C6	H5
Torsion	0	0	0.151	C6	C7	C8	H
Torsion	0	0	0.151	C6	C7	C8	H1
Torsion	0	0	0.151	C6	C7	C9	H12
Torsion	0	0	0.151	C6	C7	C9	H13
Torsion	0	0	0.226	H5	C6	O	H6
Torsion	0	0	0.151	H5	C6	C7	C8
Torsion	0	0	0.151	H5	C6	C7	C9
Torsion	0	3.648	0	C1	C2	C5	C6
Torsion	0	3.648	0	C3	C4	C5	C6

Torsion	0.654	-0.025	0.101	C8	C9	C7	C6
Torsion	0.654	-0.025	0.101	C6	C7	C8	C9
Torsion	0.654	-0.025	0.101	C5	C6	C7	C8
Torsion	0.654	-0.025	0.101	C5	C6	C7	C9
Torsion	0.861	-0.252	0.334	C8	C7	C6	O
Torsion	0.861	-0.252	0.334	C9	C7	C6	O

Improper torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
		(kcal/mol)					
Improper torsion	0	1.107	0	C2	C5	C4	C6

VBA

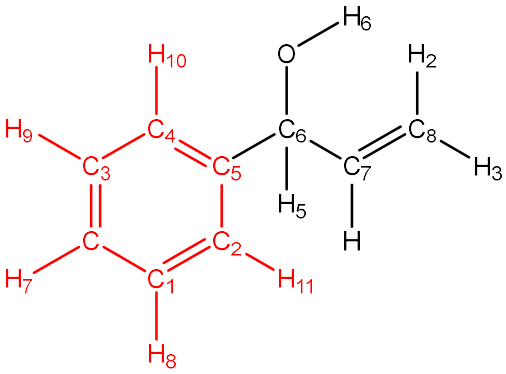


Figure B9 Indexing of the atoms in VBA. The rigid parts are coloured in red.

Table B18 Force field parameters for VBA. The parameters are as defined in equation **B-1**.

	ε/k_B (K)	**σ (Å)**
C	35.25061	3.55
C1	35.25061	3.55
C2	35.25061	3.55
C3	35.25061	3.55
C4	35.25061	3.55
C5	35.25061	3.55
C6	33.23628	3.5
C7	38.27209	3.55
C8	38.27209	3.55
H	15.1074	2.42
H2	15.1074	2.5
H3	15.1074	2.5

O	85.60861	3.07
H5	15.1074	2.5
H6	0	0
H7	15.1074	2.42
H8	15.1074	2.42
H9	15.1074	2.42
H10	15.1074	2.42
H11	15.1074	2.42

Table B19 Bond, bend and torsion definitions for VBA. The parameters are as defined in equations **B-2** to **B-5**.

Bond stretch

Type of potential	**Parameters**		**Atoms**	
	$\frac{\varepsilon}{k_B}$ **(K/Å²)**	σ **(Å)**		
Harmonic bond	684373.0	1.090	C6	H5
Harmonic bond	684373.0	1.080	C7	H
Harmonic bond	684373.0	1.080	C8	H2
Harmonic bond	684373.0	1.080	C8	H3
Harmonic bond	1113119.5	0.945	O	H6
Harmonic bond	638092.3	1.510	C5	C6
Harmonic bond	638081.7	1.510	C6	C7
Harmonic bond	644130.0	1.410	C6	O

Harmonic bond	1105061.2	1.340	C7	C8	

Bend stretch

Type of potential	**Parameters**		**Atoms**		
	$\frac{\varepsilon}{k_B}$ **(K/rad²)**	σ **(°)**			
Harmonic bend	100643.8	109.500	C5	C6	H5
Harmonic bend	70450.7	117.000	C6	C7	H
Harmonic bend	23852.8	108.500	C7	C6	H5
Harmonic bend	70455.5	125.700	C7	C8	H2
Harmonic bend	70455.5	125.700	C7	C8	H3
Harmonic bend	70455.5	125.700	C8	C7	H
Harmonic bend	70450.7	117.000	H2	C8	H3
Harmonic bend	80515.0	109.500	C5	C6	C7
Harmonic bend	140901.3	124.000	C6	C7	C8
Harmonic bend	70455.5	109.500	C7	C6	H5
Harmonic bend	70455.5	109.500	O	C6	H5
Harmonic bend	117457.9	112.700	C2	C5	C6
Harmonic bend	117457.9	112.700	C4	C5	C6
Harmonic bend	100643.8	109.500	C5	C6	O
Harmonic bend	100643.8	109.500	C7	C6	O

Torsion

Type of potential	Parameters			Atoms			
	p_0	p_1	p_2				
	(kcal/mol)						
Torsion	0	0	0	C2	C5	C6	H5
Torsion	0	0	0	C4	C5	C6	H5
Torsion	0.000	0	0.236	C5	C6	C7	H
Torsion	-0.453	0	0	C5	C6	O	H6
Torsion	0	3.688	0	C6	C5	C2	H11
Torsion	0	3.688	0	C6	C5	C4	H10
Torsion	0	7.045	0	C6	C7	C8	H2
Torsion	0	7.045	0	C6	C7	C8	H3
Torsion	-0.453	0	0	C7	C6	O	H6
Torsion	0.000	0	0.187	C8	C7	C6	H5
Torsion	0.000	0	0.236	H	C7	C6	O
Torsion	0.000	0	0.160	H	C7	C6	H5
Torsion	0	0	0.151	H	C7	C8	H3
Torsion	0	0	0.151	H	C7	C8	H2
Torsion	0	0	0.226	H5	C6	O	H6
Torsion	0	3.648	0	C1	C2	C5	C6
Torsion	0	0	0	C2	C5	C6	C7

Torsion	0	0	0	C2	C5	C6	O
Torsion	0	3.648	0	C3	C4	C5	C6
Torsion	0	0	0	C4	C5	C6	C7
Torsion	0	0	0	C4	C5	C6	O
Torsion	0.174	0.204	-0.455	C5	C6	C7	C8
Torsion	0.252	0	0	C8	C7	C6	O

Improper torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
	(kcal/mol)						
Improper torsion	0	1.107	0	C2	C5	C4	C6

Butanol, pentanol and hexanol

(Modelled by Dr Rocio Bueno-Perez)

Table B20 Force field parameters for butanol, pentanol and hexanol. The parameters are as defineds in equation **B-1**.

	ε/k_B (K)	**σ (Å)**
C_CH3	33.212	33.212
C_CH2	33.212	33.212
C_CH	33.212	33.212
C_CH2OH	33.212	33.212
C_CHOH	33.212	33.212
C_COH	33.212	33.212
O_OH	85.5472	85.5472
H_OH	-	-
H_C	15.09	15.09

Butanol

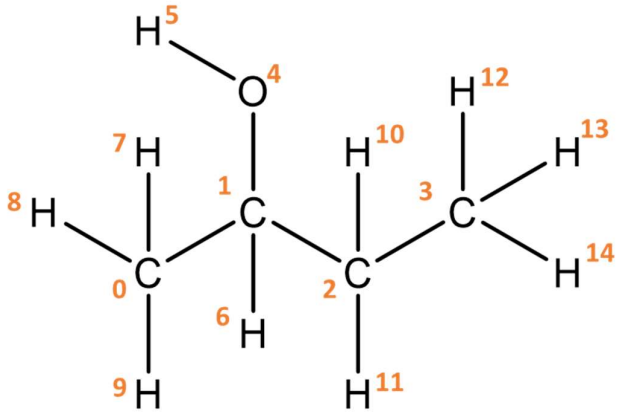


Figure B1068 Indexing of the atoms in 2-butanol.

Table B21 Bond, bend and torsion definitions for 2-butanol. The atom indices are indicated in orange in **Figure B10.** The parameters are as defined in equations **B-2** to **B-5**.

Bond stretch				
Type of potential	**Parameters**		**Atoms**	
	$\frac{\varepsilon}{k_B}$ **(K/Å²)**	σ **(Å)**		
Harmonic bond	269725	1.529	0	1
Harmonic bond	269725	1.529	1	2
Harmonic bond	269725	1.529	2	3
Harmonic bond	322060	1.410	1	4
Harmonic bond	556560	0.945	4	5
Harmonic bond	342188	1.090	0	7
Harmonic bond	342188	1.090	0	8
Harmonic bond	342188	1.090	0	9
Harmonic bond	342188	1.090	1	6

Harmonic bond	342188	1.090	2	10	
Harmonic bond	342188	1.090	2	11	
Harmonic bond	342188	1.090	3	12	
Harmonic bond	342188	1.090	3	13	
Harmonic bond	342188	1.090	3	14	

Bend stretch

Type of potential	**Parameters**		**Atoms**		
	$\frac{\varepsilon}{k_B}$ **(K/rad²)**	σ **(°)**			
Harmonic bend	58725	112.7	0	1	2
Harmonic bend	58725	112.7	1	2	3
Harmonic bend	55391	108.5	1	4	5
Harmonic bend	50355	109.5	0	1	4
Harmonic bend	50355	109.5	2	1	4
Harmonic bend	35248	109.5	6	1	4
Harmonic bend	37766	110.7	7	0	1
Harmonic bend	37766	110.7	8	0	1
Harmonic bend	37766	110.7	9	0	1
Harmonic bend	37766	110.7	6	1	0
Harmonic bend	37766	110.7	6	1	2
Harmonic bend	37766	110.7	11	2	1
Harmonic bend	37766	110.7	11	2	3

Harmonic bend	37766	110.7	10	2	1
Harmonic bend	37766	110.7	10	2	3
Harmonic bend	37766	110.7	12	3	2
Harmonic bend	37766	110.7	13	3	2
Harmonic bend	37766	110.7	14	3	2
Harmonic bend	33234	107.8	7	0	8
Harmonic bend	33234	107.8	7	0	9
Harmonic bend	33234	107.8	8	0	9
Harmonic bend	33234	107.8	10	2	11
Harmonic bend	33234	107.8	12	3	13
Harmonic bend	33234	107.8	12	3	14
Harmonic bend	33234	107.8	13	3	14

Torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
	(kcal/mol)						
Torsion	1.740	-0.157	0.279	0	1	2	3
Torsion	1.711	-0.5	0.663	3	2	1	4
Torsion	-0.356	-0.174	0.492	2	1	4	5
Torsion	-0.356	-0.174	0.492	0	1	4	5
Torsion	0.0	0.0	0.450	6	1	4	5

Torsion	0.0	0.0	0.468	7	0	1	4
Torsion	0.0	0.0	0.468	8	0	1	4
Torsion	0.0	0.0	0.468	9	0	1	4
Torsion	0.0	0.0	0.468	10	2	1	4
Torsion	0.0	0.0	0.468	11	2	1	4
Torsion	0.0	0.0	0.366	7	0	1	2
Torsion	0.0	0.0	0.366	8	0	1	2
Torsion	0.0	0.0	0.366	9	0	1	2
Torsion	0.0	0.0	0.366	10	2	1	0
Torsion	0.0	0.0	0.366	11	2	1	0
Torsion	0.0	0.0	0.366	12	3	2	1
Torsion	0.0	0.0	0.366	13	3	2	1
Torsion	0.0	0.0	0.366	14	3	2	1
Torsion	0.0	0.0	0.318	6	1	0	7
Torsion	0.0	0.0	0.318	6	1	0	8
Torsion	0.0	0.0	0.318	6	1	0	9
Torsion	0.0	0.0	.318	6	1	2	10
Torsion	0.0	0.0	0.318	6	1	2	11
Torsion	1.740	0.0	0.450	10	2	3	12
Torsion	1.711	0.0	0.468	10	2	3	13
Torsion	-0.356	0.0	0.468	10	2	3	14

Torsion	-0.356	0.0	0.468	11	2	3	12
Torsion	0.0	0.0	0.468	11	2	3	13
Torsion	0.0	0.0	0.468	11	2	3	14

Pentanol

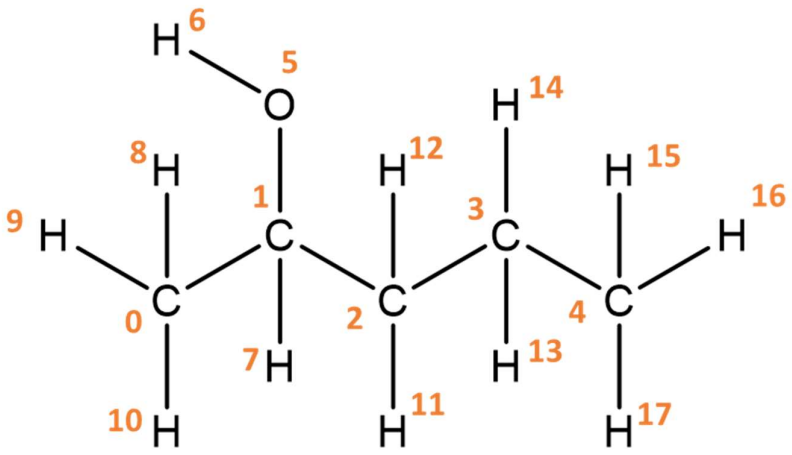


Figure B11 Indexing of the atoms in 2-pentanol.

Table B22 Bond, bend and torsion definitions for 2-pentanol. The atom indices are indicated in orange in **Figure B11**. The parameters are as defined in equations **B-2** to **B-5**.

Bond stretch				
Type of potential	**Parameters**		**Atoms**	
	$\frac{\varepsilon}{k_B}$ **(K/Å²)**	σ **(Å)**		
Harmonic bond	269725	1.529	0	1
Harmonic bond	269725	1.529	1	2
Harmonic bond	269725	1.529	2	3
Harmonic bond	322060	1.410	3	4
Harmonic bond	556560	0.945	1	5
Harmonic bond	342188	1.090	5	6
Harmonic bond	342188	1.090	1	7
Harmonic bond	342188	1.090	0	8

Harmonic bond	342188	1.090	0	9
Harmonic bond	342188	1.090	0	10
Harmonic bond	342188	1.090	2	11
Harmonic bond	342188	1.090	2	12
Harmonic bond	342188	1.090	3	13
Harmonic bond	342188	1.090	3	14
Harmonic bond	342188	1.090	4	15
Harmonic bond	342188	1.090	4	16
Harmonic bond	342188	1.090	4	17

Bend stretch

Type of potential	**Parameters**		**Atoms**		
	$\frac{\varepsilon}{k_B}$ **(K/rad²)**	σ **(°)**			
Harmonic bend	58725	112.7	0	1	2
Harmonic bend	58725	112.7	1	2	3
Harmonic bend	58725	112.7	2	3	4
Harmonic bend	37741	110.7	1	0	8
Harmonic bend	37741	110.7	1	0	9
Harmonic bend	37741	110.7	1	0	10
Harmonic bend	37741	110.7	1	2	11
Harmonic bend	37741	110.7	1	2	12
Harmonic bend	37741	110.7	3	2	11

Harmonic bend	37741	110.7	3	2	12
Harmonic bend	37741	110.7	2	3	13
Harmonic bend	37741	110.7	2	3	14
Harmonic bend	37741	110.7	4	3	13
Harmonic bend	37741	110.7	4	3	14
Harmonic bend	37741	110.7	3	4	15
Harmonic bend	37741	110.7	3	4	16
Harmonic bend	37741	110.7	3	4	17
Harmonic bend	37741	110.7	0	1	7
Harmonic bend	37741	110.7	2	1	7
Harmonic bend	50321	109.5	0	1	5
Harmonic bend	50321	109.5	2	1	5
Harmonic bend	55354	108.5	1	5	6
Harmonic bend	35225	109.5	7	1	5
Harmonic bend	33212	107.8	8	0	9
Harmonic bend	33212	107.8	8	0	10
Harmonic bend	33212	107.8	9	0	10
Harmonic bend	33212	107.8	11	2	12
Harmonic bend	33212	107.8	13	3	14
Harmonic bend	33212	107.8	15	4	16
Harmonic bend	33212	107.8	15	4	17

Harmonic bend	33212	107.8	16	4	17		

Torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
	(kcal/mol)						
Torsion	1.300	-0.050	0.200	0	1	2	3
Torsion	1.300	-0.050	0.200	1	2	3	4
Torsion	0.0	0.0	0.300	8	0	1	2
Torsion	0.0	0.0	0.300	9	0	1	2
Torsion	0.0	0.0	0.300	10	0	1	2
Torsion	0.0	0.0	0.300	7	1	2	3
Torsion	0.0	0.0	0.300	11	2	1	0
Torsion	0.0	0.0	0.300	12	2	1	0
Torsion	0.0	0.0	0.300	15	4	3	2
Torsion	0.0	0.0	0.300	16	4	3	2
Torsion	0.0	0.0	0.300	17	4	3	2
Torsion	0.0	0.0	0.300	11	2	3	4
Torsion	0.0	0.0	0.300	12	2	3	4
Torsion	0.0	0.0	.300	13	3	2	1
Torsion	0.0	0.0	0.300	14	3	2	1
Torsion	0.0	0.0	0.300	8	0	1	7

Torsion	0.0	0.0	0.300	9	0	1	7
Torsion	0.0	0.0	0.300	10	0	1	7
Torsion	0.0	0.0	0.300	7	1	2	11
Torsion	0.0	0.0	0.300	7	1	2	12
Torsion	0.0	0.0	0.300	11	2	3	13
Torsion	0.0	0.0	0.300	11	2	3	14
Torsion	0.0	0.0	0.300	12	2	3	13
Torsion	0.0	0.0	0.300	12	2	3	14
Torsion	0.0	0.0	0.300	13	3	4	15
Torsion	0.0	0.0	0.300	13	3	4	16
Torsion	0.0	0.0	0.300	13	3	4	17
Torsion	0.0	0.0	0.300	14	3	4	15
Torsion	0.0	0.0	0.300	14	3	4	16
Torsion	0.0	0.0	0.300	14	3	4	17
Torsion	-0.356	-0.174	0.492	0	1	5	6
Torsion	-0.356	-0.174	0.492	2	1	5	6
Torsion	0.0	0.0	0.352	7	1	5	6
Torsion	0.0	0.0	0.468	8	0	1	5
Torsion	0.0	0.0	0.468	9	0	1	5
Torsion	0.0	0.0	0.468	10	0	1	5
Torsion	0.0	0.0	0.468	11	2	1	5

Torsion	0.0	0.0	.468	12	2	1	5
Torsion	1.711	-0.500	0.663	3	2	1	5

Hexanol

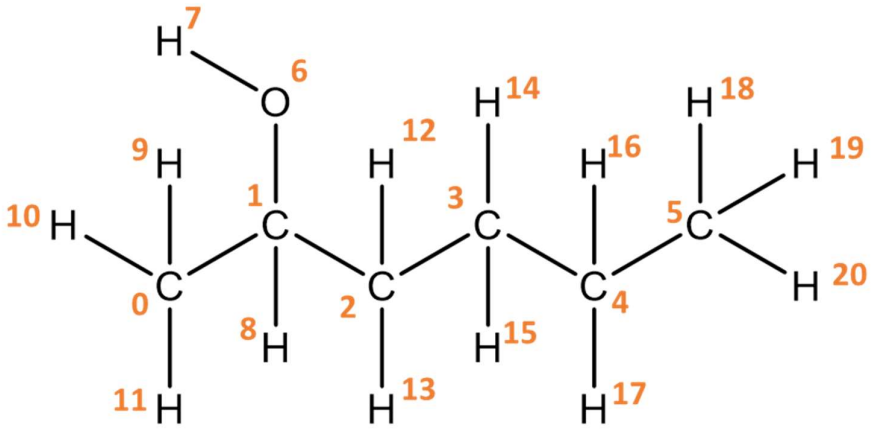


Figure B12 Indexing of the atoms in 2-hexanol.

Table B23 Bond, bend and torsion definitions for 2-hexanol. The atom indices are indicated in orange in **Figure B12**. The parameters are as defined in equations **B-2** to **B-5**.

Bond stretch

Type of potential	**Parameters**		**Atoms**	
	$\frac{\varepsilon}{k_B}$ **(K/Å²)**	σ **(Å)**		
Harmonic bond	269725	1.529	0	1
Harmonic bond	269725	1.529	1	2
Harmonic bond	269725	1.529	2	3
Harmonic bond	269725	1.529	3	4
Harmonic bond	269725	1.529	4	5
Harmonic bond	322060	1.410	1	6
Harmonic bond	556560	0.945	6	7
Harmonic bond	342188	1.090	1	8

Harmonic bond	342188	1.090	0	9	
Harmonic bond	342188	1.090	0	10	
Harmonic bond	342188	1.090	0	11	
Harmonic bond	342188	1.090	2	12	
Harmonic bond	342188	1.090	2	13	
Harmonic bond	342188	1.090	3	14	
Harmonic bond	342188	1.090	3	15	
Harmonic bond	342188	1.090	4	16	
Harmonic bond	342188	1.090	4	17	
Harmonic bond	342188	1.090	5	18	
Harmonic bond	342188	1.090	5	19	
Harmonic bond	342188	1.090	5	20	

Bend stretch

Type of potential	**Parameters**		**Atoms**		
	$\frac{\varepsilon}{k_B}$ **(K/rad²)**	σ **(°)**			
Harmonic bend	58725	112.7	0	1	2
Harmonic bend	58725	112.7	1	2	3
Harmonic bend	58725	112.7	2	3	4
Harmonic bend	58725	112.7	3	4	5
Harmonic bend	37741	110.7	1	0	9
Harmonic bend	37741	110.7	1	0	10

Harmonic bend	37741	110.7	1	0	11
Harmonic bend	37741	110.7	1	2	12
Harmonic bend	37741	110.7	1	2	13
Harmonic bend	37741	110.7	3	2	12
Harmonic bend	37741	110.7	3	2	13
Harmonic bend	37741	110.7	2	3	14
Harmonic bend	37741	110.7	2	3	15
Harmonic bend	37741	110.7	4	3	14
Harmonic bend	37741	110.7	4	3	15
Harmonic bend	37741	110.7	3	4	16
Harmonic bend	37741	110.7	3	4	17
Harmonic bend	37741	110.7	5	4	16
Harmonic bend	37741	110.7	5	4	17
Harmonic bend	37741	110.7	4	5	18
Harmonic bend	37741	110.7	4	5	19
Harmonic bend	37741	110.7	4	5	20
Harmonic bend	37741	110.7	0	1	8
Harmonic bend	37741	110.7	2	1	8
Harmonic bend	50321	109.5	0	1	6
Harmonic bend	50321	109.5	2	1	6
Harmonic bend	55354	108.5	1	6	7

Harmonic bend	35225	109.5	8	1	6
Harmonic bend	33212	107.8	9	0	10
Harmonic bend	33212	107.8	10	0	11
Harmonic bend	33212	107.8	9	0	11
Harmonic bend	33212	107.8	12	2	13
Harmonic bend	33212	107.8	14	3	15
Harmonic bend	33212	107.8	16	4	17
Harmonic bend	33212	107.8	18	5	19
Harmonic bend	33212	107.8	19	5	20
Harmonic bend	33212	107.8	18	5	20

Torsion

Type of potential	**Parameters**			**Atoms**			
	p_0	p_1	p_2				
	(kcal/mol)						
Torsion	1.300	-0.050	0.200	0	1	2	3
Torsion	1.300	-0.050	0.200	1	2	3	4
Torsion	1.300	-0.050	0.200	2	3	4	5
Torsion	0.0	0.0	0.300	9	0	1	2
Torsion	0.0	0.0	0.300	10	0	1	2
Torsion	0.0	0.0	0.300	11	0	1	2
Torsion	0.0	0.0	0.300	8	1	2	3

Torsion	0.0	0.0	0.300	12	2	1	0
Torsion	0.0	0.0	0.300	13	2	1	0
Torsion	0.0	0.0	0.300	12	2	3	4
Torsion	0.0	0.0	0.300	13	2	3	4
Torsion	0.0	0.0	0.300	14	3	2	1
Torsion	0.0	0.0	0.300	15	3	2	1
Torsion	0.0	0.0	0.300	14	3	4	5
Torsion	0.0	0.0	0.300	15	3	4	5
Torsion	0.0	0.0	0.300	16	4	3	2
Torsion	0.0	0.0	0.300	17	4	3	2
Torsion	0.0	0.0	0.300	18	5	4	3
Torsion	0.0	0.0	0.300	19	5	4	3
Torsion	0.0	0.0	0.300	20	5	4	3
Torsion	0.0	0.0	0.300	9	0	1	8
Torsion	0.0	0.0	0.300	10	0	1	8
Torsion	0.0	0.0	0.300	11	0	1	8
Torsion	0.0	0.0	0.300	8	1	2	12
Torsion	0.0	0.0	0.300	8	1	2	13
Torsion	0.0	0.0	0.300	12	2	3	14
Torsion	0.0	0.0	0.300	13	2	3	14
Torsion	0.0	0.0	0.300	12	2	3	15

Torsion	0.0	0.0	0.300	13	2	3	15
Torsion	0.0	0.0	0.300	14	3	4	16
Torsion	0.0	0.0	0.300	14	3	4	17
Torsion	0.0	0.0	0.300	15	3	4	16
Torsion	0.0	0.0	0.300	15	3	4	17
Torsion	0.0	0.0	0.300	16	4	5	18
Torsion	0.0	0.0	0.300	16	4	5	19
Torsion	0.0	0.0	0.300	16	4	5	20
Torsion	0.0	0.0	0.300	17	4	5	18
Torsion	0.0	0.0	0.300	17	4	5	19
Torsion	0.0	0.0	0.300	17	4	5	20
Torsion	-0.356	-0.174	0.492	0	1	6	7
Torsion	-0.356	-0.174	0.492	2	1	6	7
Torsion	0.0	0.0	0.352	8	1	6	7
Torsion	0.0	0.0	0.468	9	0	1	6
Torsion	0.0	0.0	0.468	10	0	1	6
Torsion	0.0	0.0	0.468	11	0	1	5
Torsion	0.0	0.0	0.468	12	2	1	6
Torsion	0.0	0.0	0.468	13	2	1	6
Torsion	1.711	-0.500	0.663	3	2	1	6

APPENDIX C MOF FAMILIES CLASSIFICATION: DESCRIPTION OF THE CRITERIA DEVELOPED

All searches were carried out in the MOF subset developed in my previous work based on CSD version 5.37 with May 2016 update.[43] The number of hits obtained for each criterion is given for all the MOF structures, ordered and disordered alike, in this document. Unless specified, all dotted bonds represented in the following queries are of "any" type. Green diagrams represent "must have" criteria, and red diagrams represent "must not have" criteria as explained in the paper.

Zr-oxide based MOFs

The Zr atoms in Zr-oxide-based MOFs such as UiO-66 (**Figure C1**) are bonded to the oxygen atoms of the carboxylic linkers. The circled area in **Figure C1** shows the connection between the metal cluster and the organic linker. This specificity is expressed in the first "must have" criterion in **Figure C2a**. **Figure C2b** refers to the special case of Zr-oxide structures containing squarates in their linkers, in which case the two oxygen atoms of the carboxylic linker are bonded to two carbon atoms that are part of a four-atom ring. Only one such structure was found in the MOF subset. The combination of these searches returns 85 hits, of which some are not the target structures (see **Figure C3**). The criteria shown in the red diagram eliminates all these undesired hits, leading to a total of 77 Zr-oxide-based MOF structures.

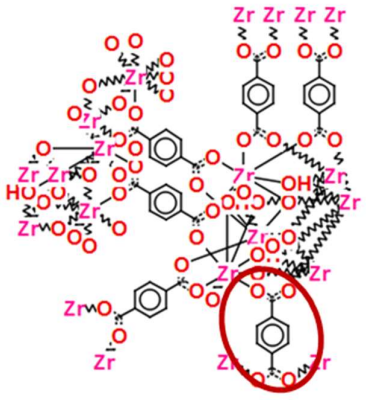


Figure C1 An example Zr-oxide based MOF; UiO-66, CSD refcode: RUBTAK02. The circle highlights the part of the MOF described by the criterion in **Figure C2a**.

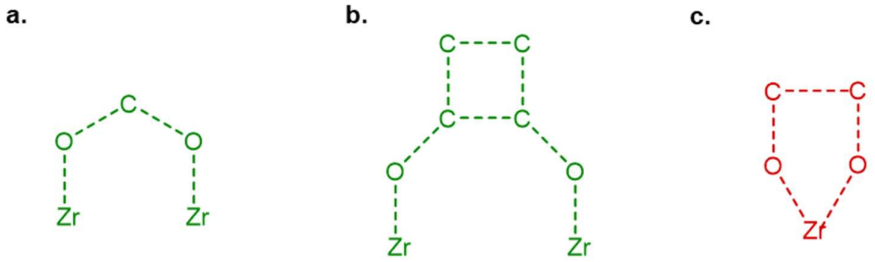


Figure C2 Criteria used to look for Zr-oxide based MOFs. **a.** and **b.** "must have" criteria, **c.** "must not have" criterion.

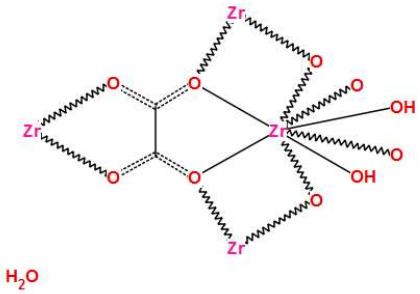


Figure C3 An example structure eliminated by the use of "must not have" criterion shown in **Figure C2**. CSD refcode: VIXGAM.

Zn-oxide based MOFs

The Zn-oxide based MOFs is a family of which IRMOF-1 (MOF-5) is a member (**Figure C4a**). The diagrams presented in **Figure C5** were developed using the same approach as for the Zr-oxide based MOFs. The combination of criteria results in 3,187 structures including IRMOF-1 materials. **Figure C4b** shows an example of a MOF with Zn-oxide SBUs (CSD refcode: ACOCUS).

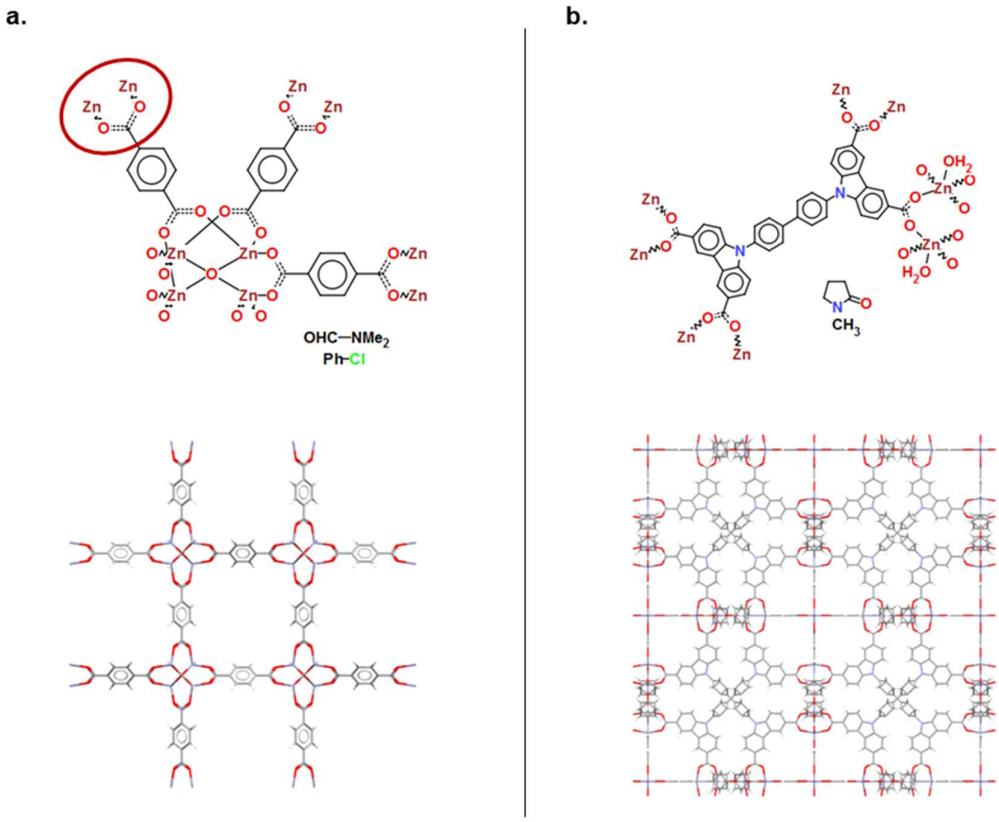


Figure C4 a. The structure of IRMOF-1 (MOF-5). **b.** the structure of an example MOF with Zn-oxide SBUs. CSD refcodes: **a.** SAHYIK, **b.** ACOCUS. The circle highlights the area captured by the criterion shown in **Figure C5a**.

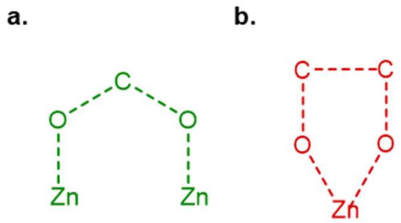


Figure C5 Criteria used to look for Zn-oxide based MOFs. **a.** "must have" criterion, **b.** "must not have" criterion.

To specifically look for IRMOF-like structures, another criterion was developed (**Figure C6**). Since the main difference between these MOFs and the more general Zn-oxide MOFs is the shape of the cluster, the criterion in **Figure C6** was obtained from further customisation of the metal node description in **Figure C5a**. 354 structures are returned for this criterion.

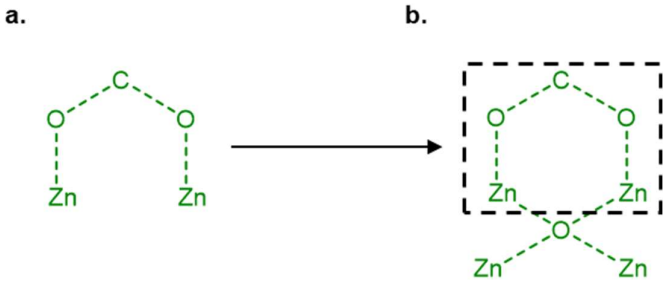


Figure C6 Derivation of criteria for IRMOF-like structures from the previous Zn-oxide-based structures criteria. Starting from **a.**, which is the "must have" criterion for the Zn-oxide based MOFs shown in **Figure C5a**, the metal cluster part is further described by including two additional Zn atoms and an oxygen atom, leading to criterion shown in **b.** The dotted box shows the same area in both criteria.

MOF-74/CPO-27-type MOFs

A typical example of CPO-27 structure is shown in **Figure C7**. A similar approach as that of Zr-oxide-based MOFs was used here to find the target structures: **Figure C7a** represents the connection between the metal cluster and the organic linker. The metal atoms QA could be any of Zn, Cu, Ni, Co, Fe, Mn or Mg, which are, by comparison with a query where QA would simply be "any metal", the most common metals in CPO-27-type of structures. This search leads to 147 hits. However, using only the criterion described in **Figure C8a** is not restrictive enough and a few undesired structures are found. An example is given in **Figure C9**. Adding a diagram which represents part of the ring to which the metal atoms belong to effectively eliminates 16 of these hits. Other untargeted structures are more difficult to remove; **Figure C9** shows three "must not have" criteria that were developed based on specific examples, some of which are shown in the right column of **Figure C9**. The combination of all these criteria returns 108 hits.

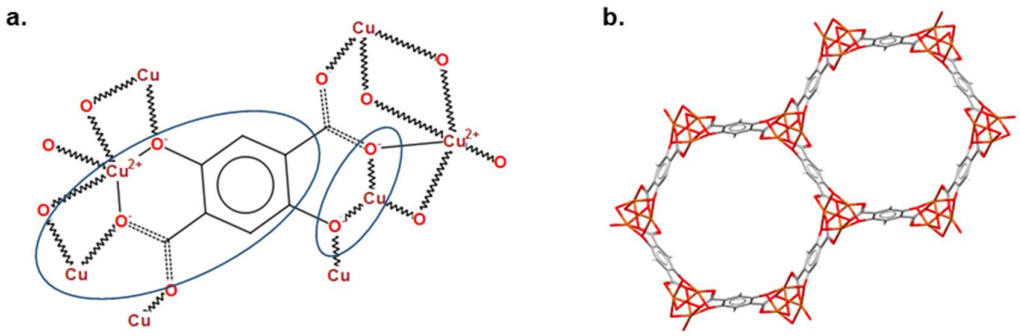


Figure C7 An example MOF-74/CPO-27 structure. **a.** Chemical diagram. The blue circled areas show the parts that are looked for by the search criterion in **Figure C9**. **b.** Spatial representation of the hexagonal channels formed in CPO-27. CSD refcode: COKNIB.

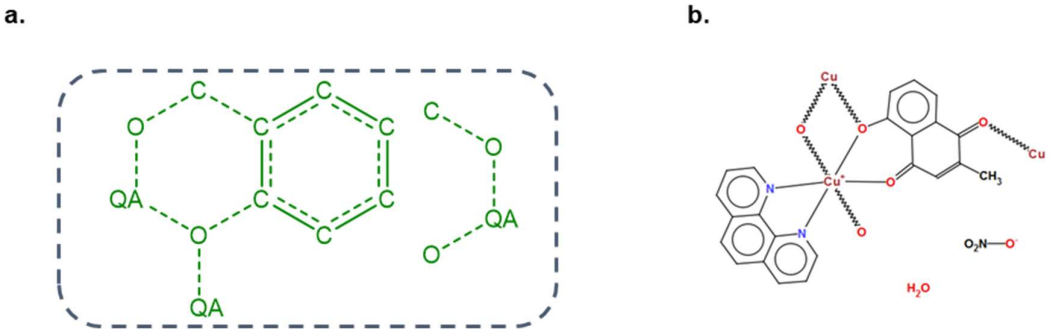


Figure C8 a. Criterion developed to look for MOF-74/CPO-27-type MOFs. **b.** Example of an undesired structure found if the criterion is not restrictive enough, i.e. if only the left part of the criterion is represented. Both parts should be considered as one single query. QA = Zn, Cu, Ni, Co, Fe, Mn, Mg. CSD refcode: XUVNUZ.

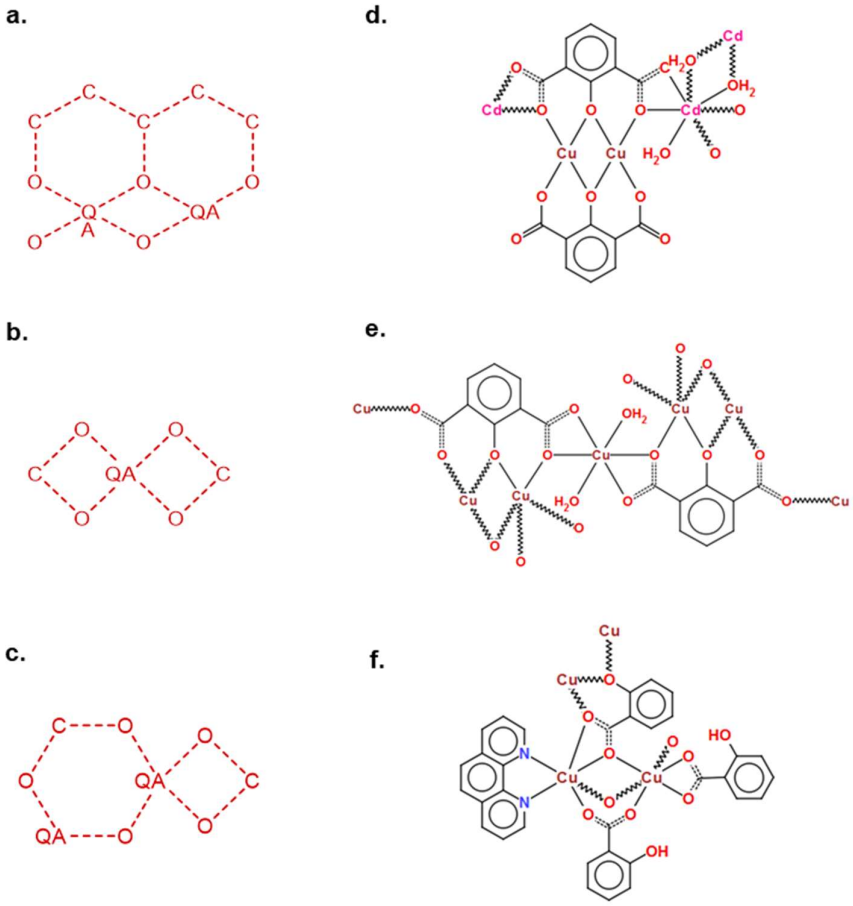


Figure C9 a. to **c.** Criteria developed to eliminate undesired structures. **d.** to **f.** Examples of structures eliminated with the corresponding criterion on the left. **a.** eliminates 20 hits, **b.** eliminates 1 hit and **c.** eliminates 2 hits. CSD refcodes: **d.** ADICIA, **e.** FODHIQ, **f.** WUYTUF.

ZIF-type MOFs

In ZIF-type of structures, the metal is tetrahedrally coordinated with four imidazolates. A “must have” criterion was first developed by describing the connection between the metal atoms and the organic linker. One metal atom is linked to four nitrogen atoms, two of which are part of an imidazolates. For symmetry reasons, two imidazolates are enough. It is specifically stressed that the metal atom should only be bonded to four atoms (**Figure C10**). This search leads to 331 hits, of which some structures need to be removed. **Figure C11** summarises the list of “must not have” criteria. Diagrams **a.** to **e.** were developed based on specific structures, some of which are shown in the corresponding examples. A 3D “must not have” criterion was also added, as the cluster in some structures does not correspond to a tetrahedron, but are almost planar. A constraint with respect to the angle between two planes, each defined by a N-metal-N chain was added. As the data in the CSD are experimental, only a few clusters are close to a perfect tetrahedron. Therefore, different values of angles were tested in order to keep most of the tetrahedra, as deformed as they may be, and filter out clusters that are not tetrahedra. At the lower end, eliminating structures with an angle below 5° is not restrictive enough, and most flat clusters are included. At the higher end, eliminating structures with an angle below 30° is too restrictive. Though all the non-tetrahedra clusters were excluded, some flat tetrahedra were also filtered out. 25° was found to be a good cut-off value. **Figure C11l** shows an example of structure where the metal and the nitrogen atoms almost belong to the same plane. An additional criterion (**Figure C12**) used on its own targets ZIFs with a metal coordination number of 6 or 8.

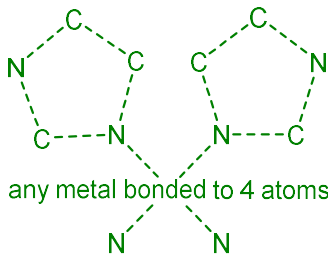


Figure C10 "Must have" criterion used to look for ZIF-type structures.

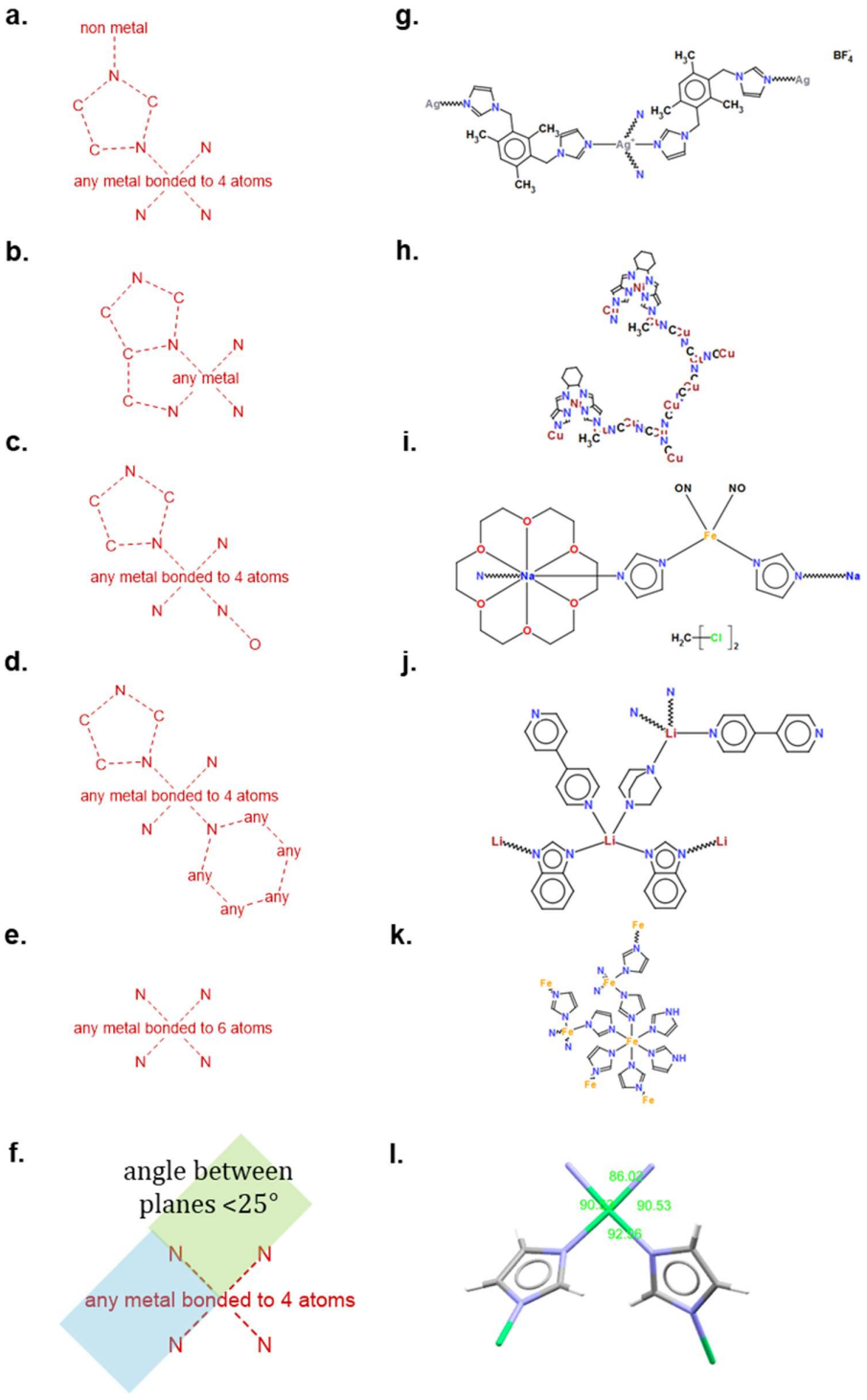


Figure C11 a. to **f.** "Must not have" criteria used to eliminate undesired structures. **g.** to **l.** Example structures corresponding to the criterion described in the left. **a.** eliminates 99 hits, **b.** eliminates 4 hits, **c.** eliminates 1 hit, **d.** eliminates 6 hits, **e.** eliminates 7 hits, **f.** eliminates 4. CSD refcodes: **g.** BUGKOF, **h.** KURPOE, **i.** VOBFIC, **j.** ALIHAF, **k.** IMIDFE, **l.** ALIDUU.

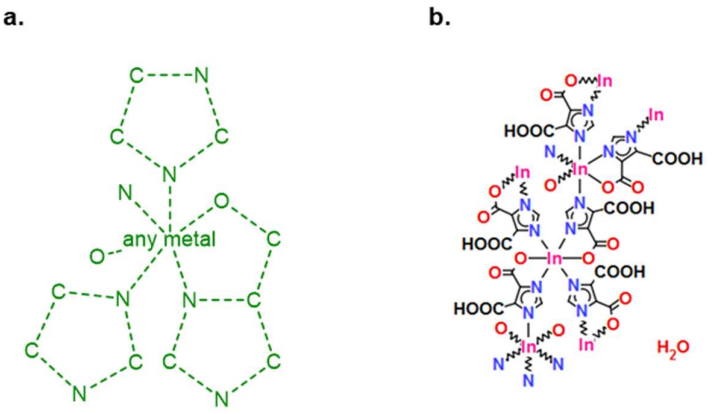


Figure C12 a. Criterion used to look for ZIF structures with metal coordination of 6 or 8. **b.** An example corresponding structure. CSD refcode: TEFWOR.

APPENDIX D MOFS' FUNCTIONAL GROUPS: DESCRIPTION OF THE CRITERIA DEVELOPED

Halogen groups

Figure D1 shows the criterion used to look for MOFs containing halogen groups. X represents any of F, Cl and Br and should be connected to only one other atom. X should not be part of the metal cluster, and should therefore not be bonded to a metal atom. In the particular case of F, the halogen atom should not be bonded to S or P. An example of undesired structure is given in **Figure D2**. The configuration of the bonds between the three non-metal atoms ensures that the functional group is attached to the organic linker, but is not part of the main chain of the linker. The variable bonds are either aromatic, delocalised, single or double, so that the functional groups that are looked for are not only those bonded to an aromatic structure, but also those that are linked to a linear organic chain. The second neighbors of the halogen atoms should not be halogens. An example of undesired structure for the -F group is given in **Figure D3**. A summary of the number of structures obtained for each case is given in **Table D1**.

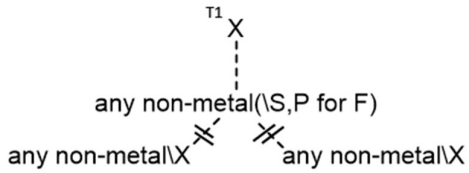


Figure D1 Criterion used to look for halogen groups. Replace X with F, Cl or Br. The variable bond is either single, double, aromatic or delocalised.

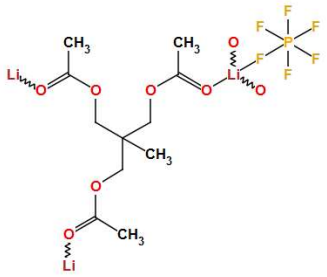


Figure D2 Example structures obtained for the -F group if the F atom is linked to a P atom. CSD refcode: WAHJOF.

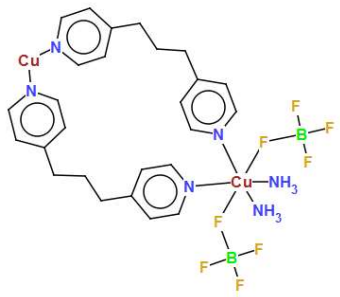


Figure D3 Example structure obtained for the -F groups if the second neighbors of the F atom are also F atoms. CSD refcode: ADOKOV.

Table D1 Number of structures obtained for each halogen group.

Funtional group	Number of hits
F	827
Cl	864
Br	503

The particular case of FMOFs

Figure D4 shows the criterion used to find FMOFs in the MOF subset. It consists in describing the organic linker of the MOF. This search led to 12 structures.

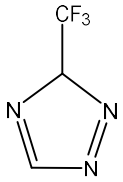


Figure D4 Criterion used to look for FMOFs.

Polar groups

The criterion used for the polar groups is very similar to that of the halogen groups. Apart from -CN, the criterion shown in **Figure D5** is enough to find structures containing -NH_2, NO_2, -COOH and -OH. It is also necessary to impose a number of bonded atoms to each atom of the polar groups. For instance, in -OH, O should be bonded to two atoms only, and H to only one.

For the particular case of -CN, an extra search was carried out in order to eliminate structures in which the cyanides are part of dicyanide functional groups. The criteria used to look for these dicyanides are shown in **Figure D6**. The list of structures obtained from this combination of searches is then eliminated from the main search presented in **Figure D5**.

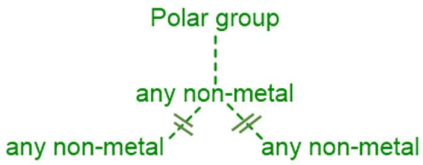


Figure D5 Criterion used to look for polar functional groups. The variable bond is either single, double, aromatic or delocalised.

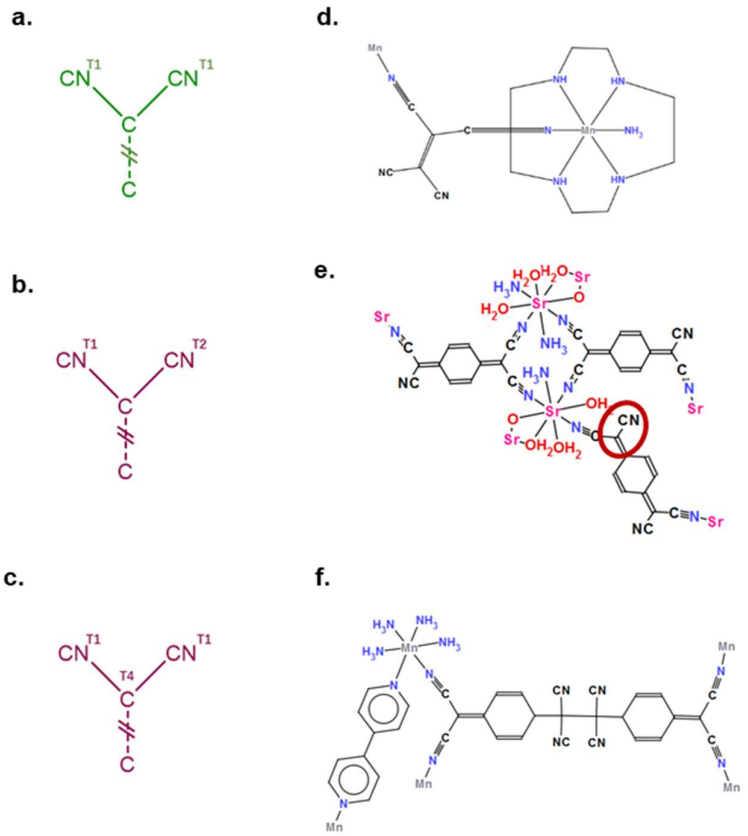


Figure D6 Criteria used to look for undesired dicyanide groups and corresponding examples for each criterion. The variable bonds are single or double. **a.** is the main criterion used to look for dicyanide functional groups. It is represented in green as it is a "must have" criterion for the search of dicyanides. **b.** is the criterion used to target dicyanide structures where one of the cyanide is part of the linker and the other cyanide can therefore be considered as a cyanide group. It is represented in red as it is a "must not have" criterion and returns structures that should be eliminated from the search for dicyanides. **c.** another "must not have" criterion used to look for structures where the carbon linked to the cyanide groups is bonded to more than three atoms. **d.** to **f.** Example structures for each case. CSD refcodes: **d.** AGAMUR, **e.** BENZOL, **f.** BUSQEM.

NB: the obtained list is to be eliminated from the main search corresponding to **Figure D5**, therefore the criteria in red are overall double negatives, i.e. positives.

Table D2 Number of hits obtained for each of the polar groups.

Functional group	Number of hits
NH_2	1996
NO_2	1198
COOH	1918
OH	1729
CN	520

Alkoxy groups

A similar approach to that of polar groups leads to the criterion in **Figure D7**, where "alkoxy" should be replaced by -OMe, -OEt and -OPr.

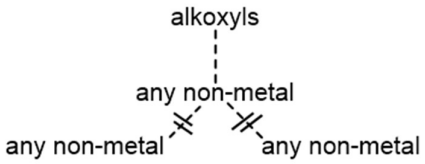


Figure D7 Criterion used to look for alkoxy groups. The variable bonds are either single, double, aromatic or delocalised.

Table D3 Number of hits for each alkoxy group.

Functional group	Number of hits
-OMe	707
-OEt	130
-OPr	31

Alkyl groups

Using the same criterion previously described for alkoxy groups returns circa 20,000 structures for alkyl groups, of which about two thirds are not the target type. This is because alkyl groups are ubiquitous and are very often part of another functional group. One alternative way of looking for alkyl groups is to break down the search into three: one search (**Figure D8a**) looks for alkyl groups attached to aromatic structures *only*, another one looks for alkyl groups attached to a linear chain, and more specifically via single bonds, the last one looks for groups attached a linear chain via a single bond and a double bond, or groups attached to an aromatic ring represented with single and double bonds.

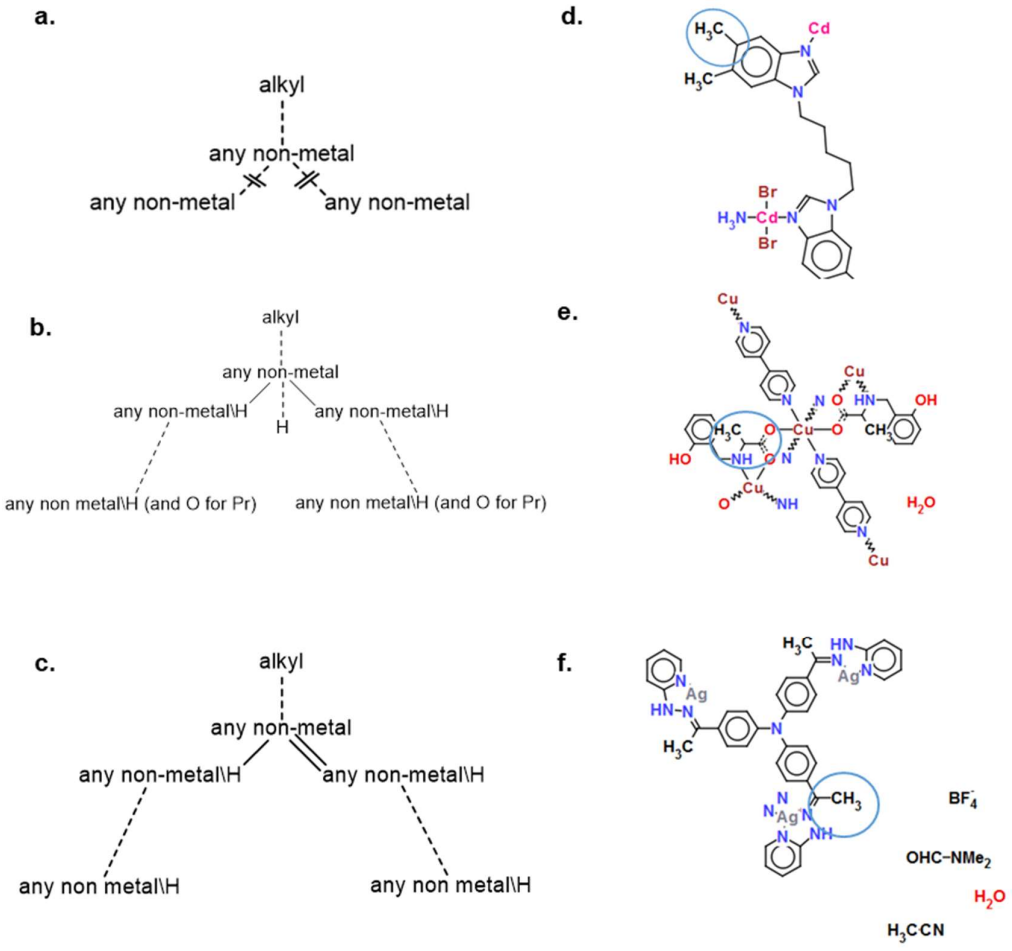


Figure D8 a. to **c.** Criteria used to look for structures with alkyl functional groups. **d.** to **f.** Example structures targeted by each criterion. The variable bonds in **a.** are aromatic or delocalised. The circles highlight the area captured by each criterion. CSD refcodes: **d.** ACELAX, **e.** ACABEM, **f.** ADAVEI.

Table D4 Number of hits for each alkyl group

Functional group	Number of hits
Me	7126
Et	437
Pr	126

Alkyl groups with more than 4 carbon atoms

For alkyl groups with more than 4 carbon atoms, the combination of criteria in **Figure D9** can be used. The two green "must have" criteria each describe one end of the alkane chain: it is bonded to the linker on one side and ends with -CH_3 on the other. They are grouped together to form an "AND" statement: each structure should meet both criteria. The resulting hitlist contained undesired structures which are eliminated using the red "must not have" criterion. The final hitlist contains 261 structures.

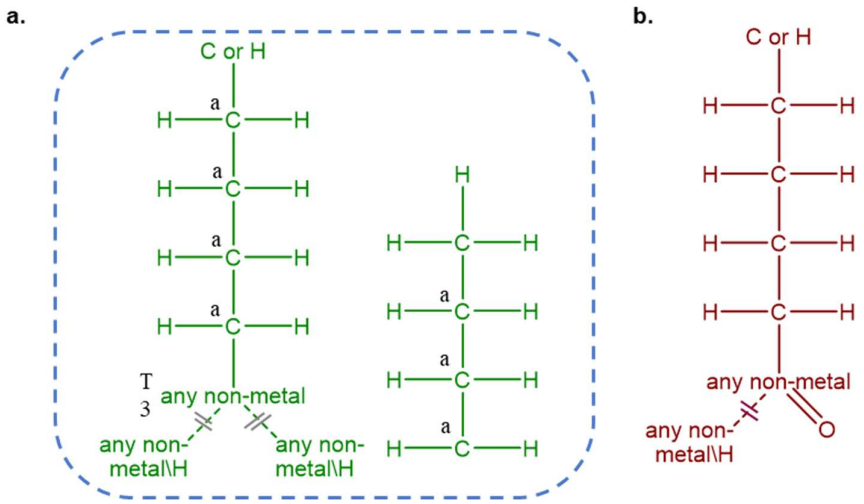


Figure D9 Criteria used to look for alkyl groups of more than 4 carbon atoms. **a.** constraints on both ends of the chain: one end has to be bonded to the linker but not be part of it and the other must be free and end with -Me. **b.** eliminates undesired structures. The variable bonds are either single, double, aromatic or delocalised. Upperscript a: the corresponding atom is acyclic. T3: the corresponding atom can only be bonded to three other atoms.

Perfluoroalkane groups

Another set of criteria is used to target structures with perfluoroalkane groups. It follows the same reasoning as above: each of the two green "must have" queries describe one end of the chain and are grouped together to form an "AND" statement, and the red "must not have" criterion eliminates undesired structures. The hitlist contains 64 structures.

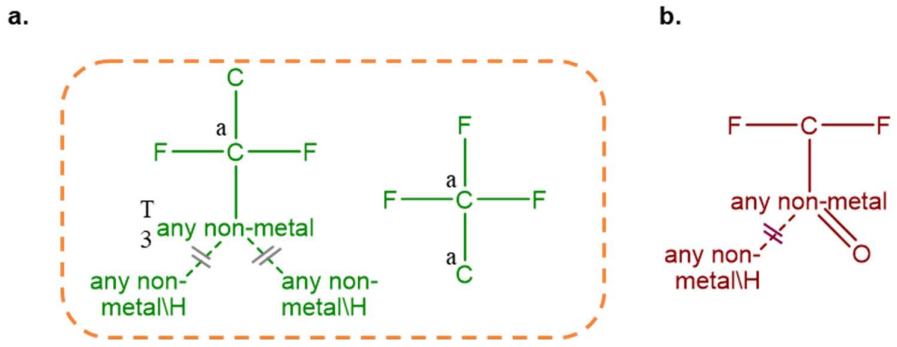


Figure D10 Criteria used to look for structures with perfluoroalkane chains. **a.** constraints on both ends of the perfluoroalkane group: one end is bonded to a linker but not part of it, the other must end with -CF3. **b.** eliminates undesired structures. The variable bonds are either single, double, aromatic or delocalised. Superscript a: the corresponding atom is acyclic.

Python script to be used with the ConQuest queries

To make sure only structures where the functional groups are located in the frameworks are returned, the ConQuest queries presented above should be saved as a .con file and used in combination with the CSD Python AP script below. A GCD list of the structures to be screened should also be provided. This script was written jointly with Dr Seth Wiggin, CCDC, UK. I conceptualised the algorithm and wrote the first version. Dr Seth Wiggin modified it to obtain the final version presented here.

```
"""
This script takes:
    -  a GCD list of structures to screen,
    -  a CON file containing the ConQuest queries for the functional
       group (here NO2)
And returns a GCD list of structures containing the functional groups
in the framework.
"""
import ccdc
import ccdc.search
from ccdc import io
from ccdc.io import EntryReader
from ccdc.io import MoleculeReader

MOF_entries=EntryReader(location_to_your_gcd_list)

def proper_identifier(m):
    '''Put the right identifier on a component.'''
    c = m.heaviest_component
    c.identifier = m.identifier
    return c

structure = 'NO2.con'

connser_substructure = ccdc.search.ConnserSubstructure(structure)
substructure_search = ccdc.search.SubstructureSearch()
sub_id = substructure_search.add_substructure(connser_substructure)

hits      =      substructure_search.search(location_to_your_gcd_list,
max_hits_per_structure=1)
hits2 = []

for old_h in hits:
    if proper_identifier(old_h.molecule).is_polymeric == True:
        for                              new_h                              in
substructure_search.search(proper_identifier(old_h.molecule)):
            hits2.append(new_h)

result_list = []

for hit in hits2:
    refcode = hit.identifier
    print refcode
    result_list.append(refcode)

f = open('clean_NO2.gcd', 'w')
f.write('\n'.join('%s' % x for x in result_list))
f.close()
```

APPENDIX E CALCULATIONS OF THE CSD MOFS' PHYSICAL AND GEOMETRICAL PROPERTIES

(work done by Dr Xiaowei Liu)

Zeo++[60] was used to characterise and calculate the geometric properties for all cleaned MOF structures (i.e. after removing solvent molecules). The calculations of accessible surface area and pore volume were performed by using a spherical N_2 probe with a radius of 1.86 Å, whilst geometrical volume fraction was calculated by setting the radius of the probe to zero.

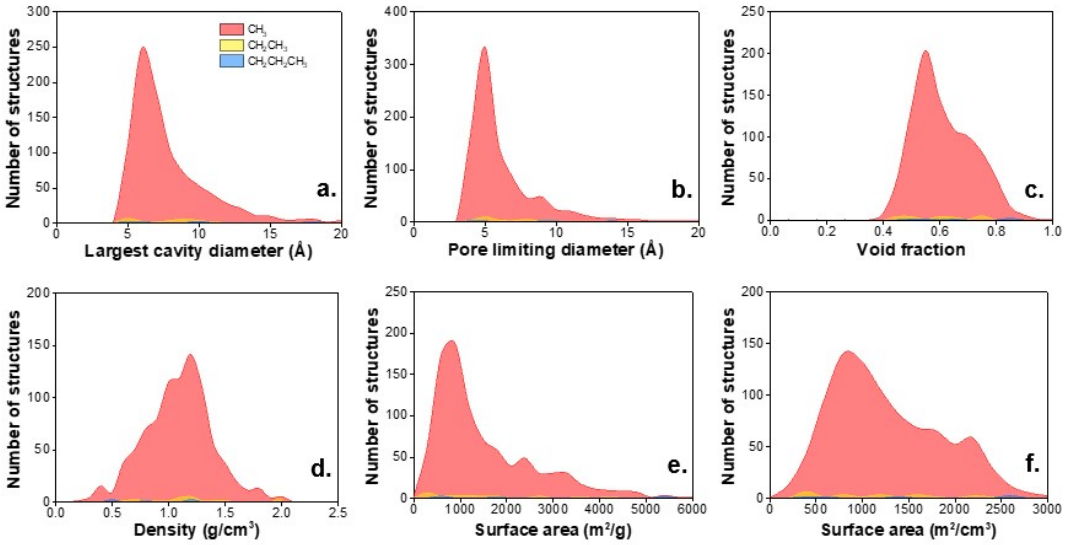


Figure E1 Histograms showing the number of hits for MOFs with different alkyl groups. **a.** largest cavity diameter (LCD), **b.** pore limiting diameter (PLD), **c.** void fraction, **d.** density, **e.** gravimetric surface area, **f.** volumetric surface area.

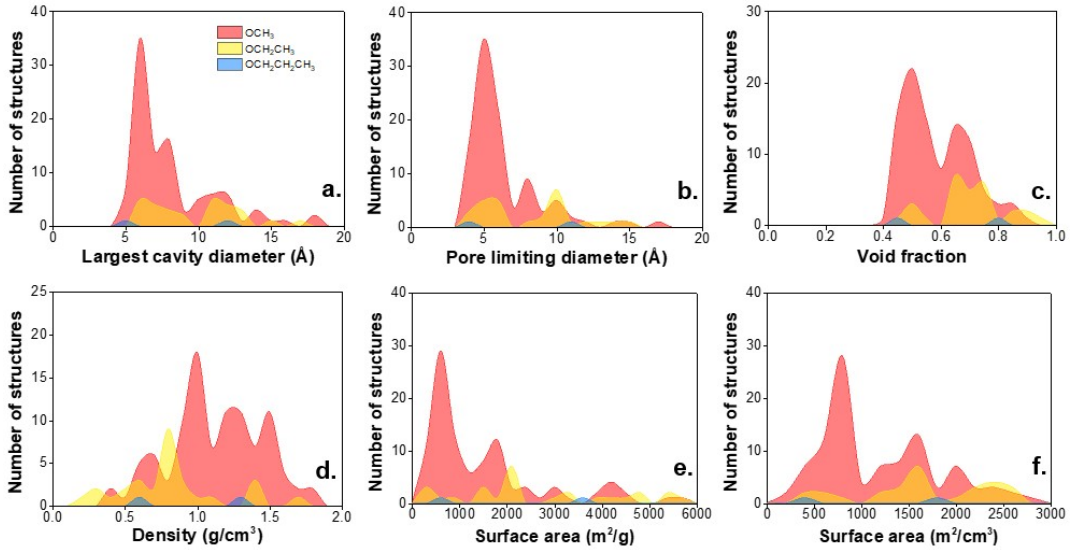


Figure E2 Histograms showing the number of hits in MOFs with different alkoxy groups. **a.** largest cavity diameter (LCD), **b.** pore limiting diameter (PLD), **c.** void fraction, **d.** density, **e.** gravimetric surface area, **f.** volumetric surface area.

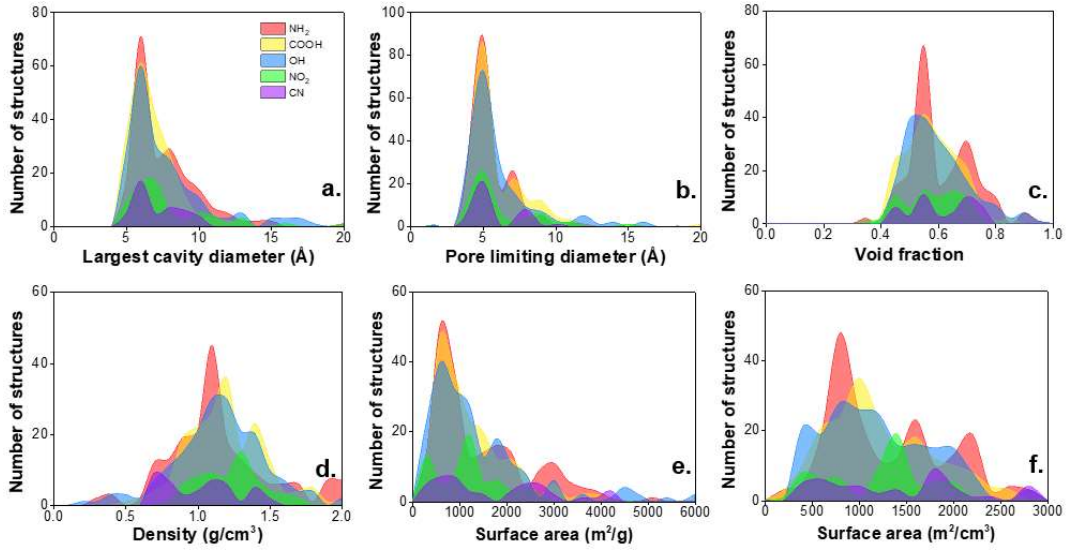


Figure E3 Histograms showing the number of hits in MOFs with different polar groups. **a.** largest cavity diameter (LCD), **b.** pore limiting diameter (PLD), **c.** void fraction, **d.** density, **e.** gravimetric surface area, **f.** volumetric surface area.

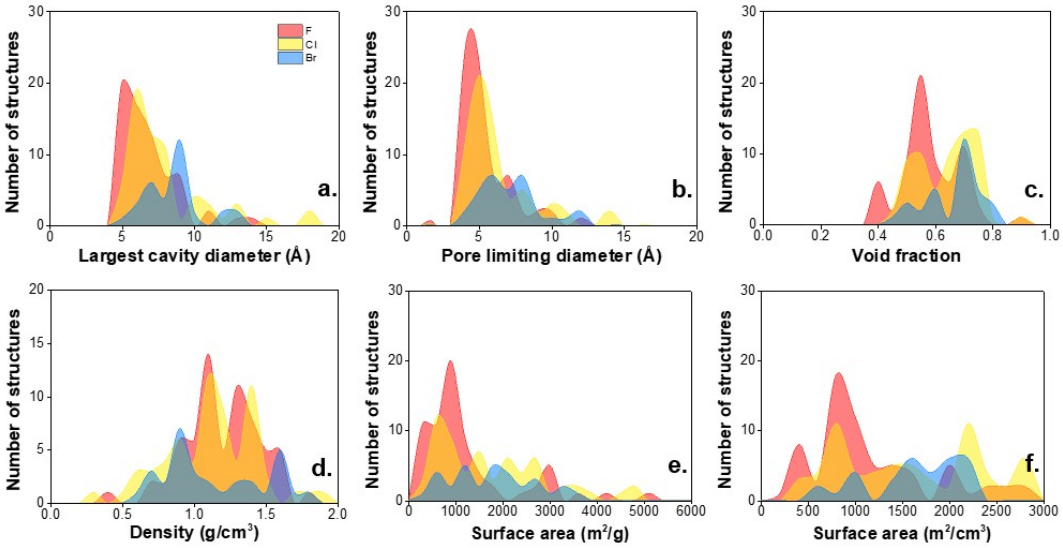


Figure E4 Histograms showing the number of hits in MOFs with different halogen groups. **a.** largest cavity diameter (LCD), **b.** pore limiting diameter (PLD), **c.** void fraction, **d.** density, **e.** gravimetric surface area, **f.** volumetric surface area.

APPENDIX F CALCULATION OF FRAMEWORK DIMENSIONALITIES

The calculation of the framework dimensionalities was performed with the CSD Python API script below, written by Dr Seth Wiggin and myself. Dr Seth Wiggin and I conceptualised the algorithm together, and I checked the different versions of the algorithms against test structures. Dr Seth Wiggin wrote the final CCDC-approved version of the script, and I tested it against the CSD MOF subset.

This script can also be found at:

github.com/ayl23/targeted_classification_CSD_MOF_subset/blob/main/framework_dimensionality.py.

The channel dimensionalities were obtained using PoreBlazer,[143] as explained in the manuscript.

```
# This script can be used for any purpose without limitation subject
to the
# conditions at
http://www.ccdc.cam.ac.uk/Community/Pages/Licences/v2.aspx
#
# This permission notice and the following statement of attribution
must be
# included in all copies or substantial portions of this script.
#
# 2020-01-21: created by S.B.Wiggin, the Cambridge Crystallographic
Data Centre,
# and Aurelia Li, Adsorption and Advanced Materials Group
(aam.ceb.cam.ac.uk),
# led by David Fairen-Jimenez from the Department of Chemical
Engineering and
# Biotechnology, University of Cambridge.

"""
    Classify_MOFs_dimensionality.py  - performs two expansions of a
polymeric network and produces minimum area bounding
    boxes. Comparison of the two bounding boxes gives the growth
dimensions of the framework
"""

from ccdc.io import EntryReader
import numpy as np
import argparse
import csv
import os.path
import time


def generate_bounding_box(atoms):
```

```
    all_pts = np.array([[c for c in atom.coordinates] for atom in
atoms])
    cov = np.cov(all_pts, rowvar=False)
    evals, evecs = np.linalg.eig(cov)

    lengths = np.sqrt(evals)
    lengths = lengths + 0.0001  # required when the box dimension is
exceedingly small and tiny changes effect the ratio
    lengths.sort()

    return lengths
def dimensionality(entry):
    """
    Calculates the dimensionality of the crystal.
    Grows 2 instances of the polymeric unit (four cycles and seven
cycles) and calculates the change in size
    The number of cycles needs to grow a representative part of the
framework, but more cycles will take longer
    """

    t0 = time.time()
    shell1 = entry.crystal.polymer_expansion(repetitions=4)
    shell2 = entry.crystal.polymer_expansion(repetitions=7)

    t1 = time.time()
    time_taken = str(t1-t0)
    print('total time elapsed for polymer expansion is % s' %
time_taken)

    lengths1 = generate_bounding_box(shell1.atoms)
    print('number of atoms in first expansion ' +
str(len(shell1.atoms)))
    s1, m1, l1 = lengths1

    lengths2 = generate_bounding_box(shell2.atoms)
    print('number of atoms in second expansion ' +
str(len(shell2.atoms)))
    s2, m2, l2 = lengths2

    ratios = [s2 / s1, m2 / m1, l2 / l1]
    print('ratio of box dimensions from first and second expansions: '
+ str(ratios))
    thresh = 1.15
    ndims = [r > thresh for r in ratios].count(True)
    return ndims

def analyse_structures(user_gcd_input, user_csv_output):

    if len(os.path.splitext(user_csv_output)[1]) == 0:
        user_csv_output += ".csv"

    with open(user_csv_output, 'w', newline='') as f:
        writer = csv.writer(f)
        writer.writerow(('Refcode', 'dimensionality', 'number in gcd
file'))

        csd_reader = EntryReader(user_gcd_input, 'CSD')

        t2 = time.time()
        n_structures = 0
        n_mof = 0
        n_non_mof = 0
```

```
        for entry in csd_reader:
            print('CSD entry: ' + str(entry.identifier))
            n_structures += 1  # quick counter
            count_polymers = 0
            for component in entry.molecule.components:
                if component.is_polymeric:
                    count_polymers += 1
            if count_polymers > 1:
                print('multiple polymer units present')
            if entry.molecule.heaviest_component.is_polymeric:
                n_mof += 1
                framework = entry.molecule.heaviest_component

                framework.remove_hydrogens()  # next steps fail if any
atoms in the unit do not have coordinates

                entry.crystal.molecule = framework

                fig = dimensionality(entry)

                if fig == 0:
                    dimension = '0D non-MOF'
                elif fig == 1:
                    dimension = '1D chain'
                elif fig == 2:
                    dimension = '2D sheet'
                elif fig == 3:
                    dimension = '3D framework'
            else:
                n_non_mof += 1
                dimension = 'no polymeric bonds detected'

            print('Framework dimensions for CSD entry % s: % s \n' %
(entry.identifier, dimension))
            writer.writerow((entry.identifier, dimension,
n_structures))
            f.flush()

        print('Total MOF subset size is: % d' % n_structures)
        print('Entries recognised as polyermic is: % d' % n_mof)
        print('Entries not recognised as polymeric (and ignored) is: %
d' % n_non_mof)

        t3 = time.time()
        overall_time_taken = str(t3 - t2)
        print('total time elapsed for script % s' % overall_time_taken)
        f.close()

def get_args():
    parser = argparse.ArgumentParser()

    parser.add_argument('-i', '--input', help='CSD refcode list (.gcd
file) filename')
    parser.add_argument('-o', '--output', help='Results CSV filename')

    return parser.parse_args()

def main():
    args = get_args()

    if args.input is None:
        args.input = input("Enter filepath for CSD refcode list (.gcd
file)" '\n')
```

```
    if args.output is None:
        args.output = input("Enter filename for results file" '\n')

    analyse_structures(args.input, args.output)

if __name__ == '__main__':
    main()
```

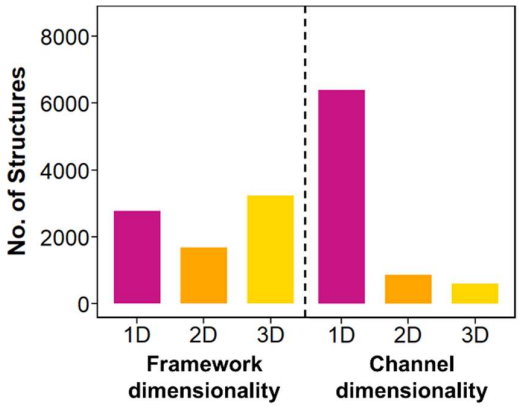


Figure F1 Histograms of framework and channel/pore dimensionalities for 8,253 porous MOFs. The channel dimensionalities of some structures could not be determined and are therefore not shown.

APPENDIX G QUALITY ASSESSMENT OF THE DATA IN THE CSD MOF SUBSET USING R FACTORS

The R-factors and crystal systems were extracted using the CSD Python API.[49] The geometric and physical properties were obtained using Zeo++.[60]

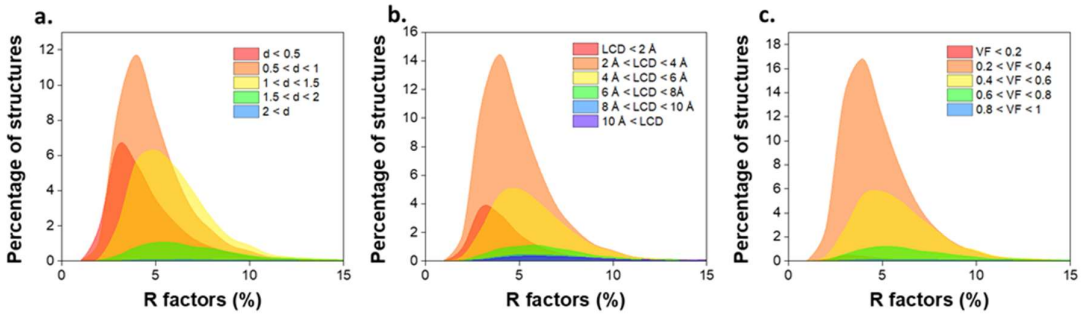


Figure G1 Histograms of **a.** density (g cm^{-3}), **b.** largest cavity diameter (LCD) and **c.** void fraction against R-factors for structures with non-zero gravimetric surface area values in the CSD MOF subset.

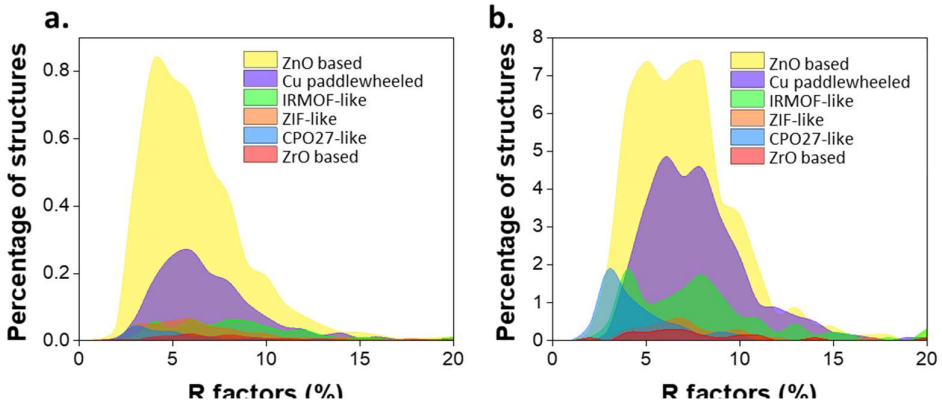


Figure G2 a. Histograms of R-factors for the different MOF families. **b.** Histograms of R-factors for the different MOF families with non-zero surface area.

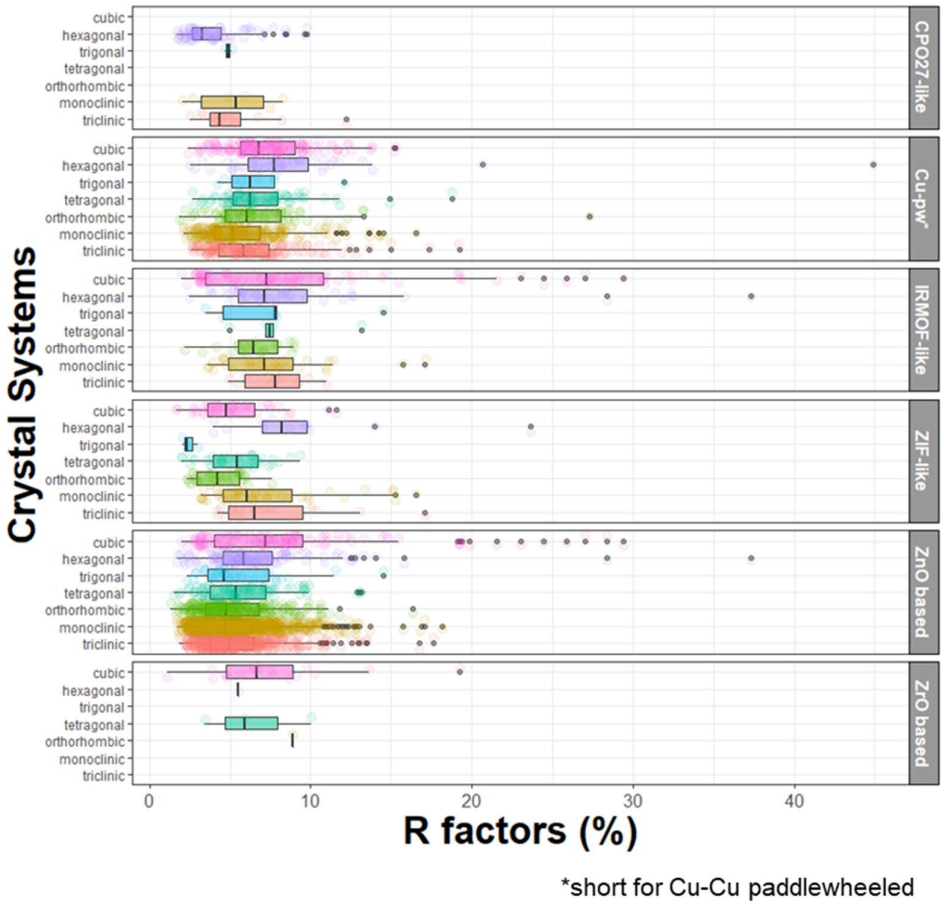


Figure G3 Boxplots of R-factors vs. crystal systems for each MOF family.

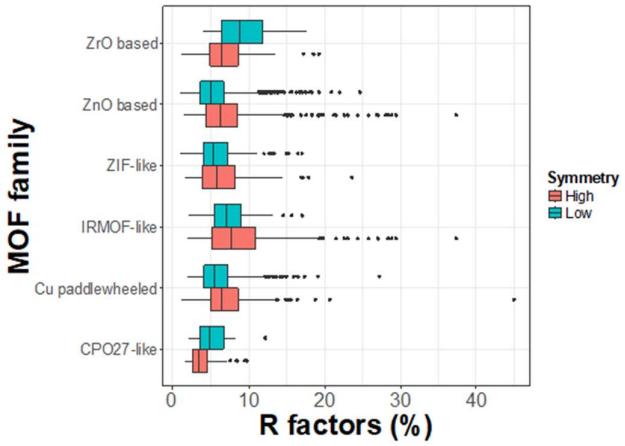


Figure G4 Boxplots of R-factors vs. degree of symmetry for of their crystal systems for each MOF family.

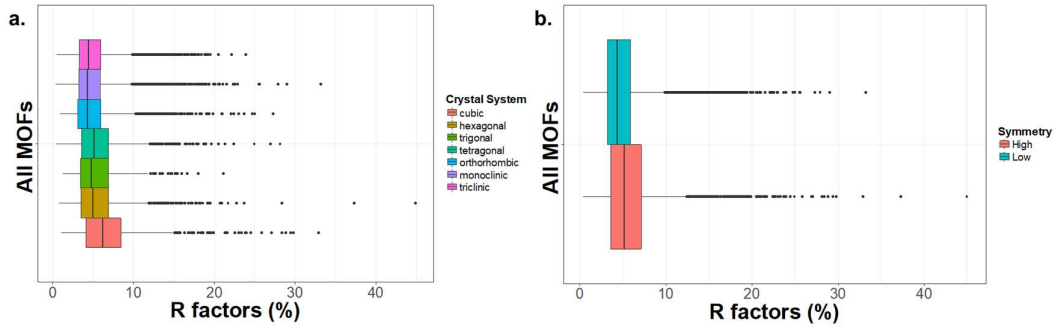


Figure G5 Boxplots of R-factors vs. **a.** crystal systems and **b.** degree of symmetry for all structures in the CSD MOF subset.

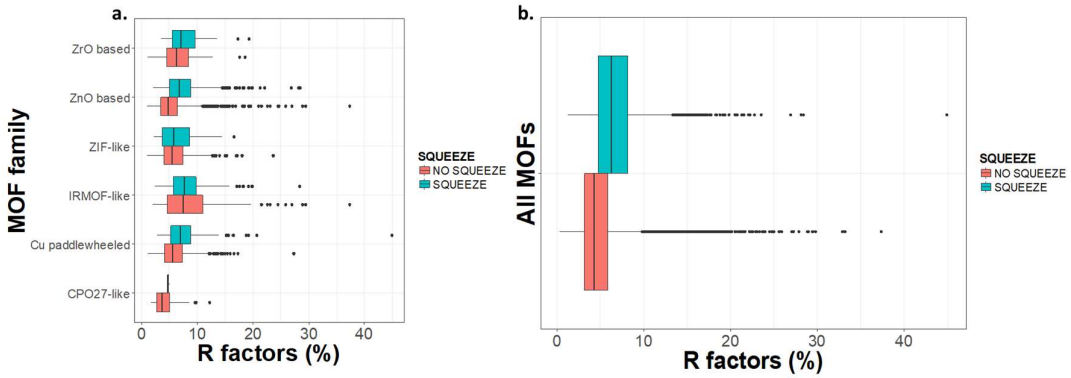


Figure G6 Boxplots of R-factors for **a.** different MOF families and **b.** all structures in the CSD MOF subset. Only 4,769 structures from the MOF subset have gone through the SQUEEZE process.

APPENDIX H HTS FOR H2 GCMC SIMULATIONS PARAMETERS AND ADDITIONAL RESULTS

The parameters presented below can also be found at:

github.com/ayl23/targeted_classification_CSD_MOF_subset/tree/main/HTS_H2

Table H1 Force field parameters used in the GCMC simulations.

	ε/k_B (K)	**σ (Å)**
B_	47.804	3.582
C_	47.854	3.474
H_	7.649	2.847
N_	38.948	3.263
O_	48.156	3.034
P_	161.024	3.698
S_	173.101	3.591
F_	36.482	3.094
I_	256.632	3.698
K_	17.612	3.397
V_	8.051	2.801
W_	33.714	2.735
Y_	36.23	2.981
U_	11.07	3.025
Li_	12.58	2.184
Be_	42.772	2.446

Al_	155.992	3.912
Si_	155.992	3.805
Cl_	142.557	3.52
Na_	251.6	2.801
Mg_	55.855	2.692
Ga_	201.28	3.912
Ge_	201.28	3.805
As_	206.312	3.698
Se_	216.376	3.591
Br_	186.184	3.52
Ca_	25.16	3.094
Sc_	9.561	2.936
Ti_	8.554	2.829
Cr_	7.548	2.694
Mn_	6.542	2.638
Fe_	27.676	4.045
Co_	7.045	2.559
Ni_	7.548	2.525
Cu_	2.516	3.114
Zn_	27.676	4.045
In_	276.76	4.09

Sn_	276.76	3.983
Sb_	276.76	3.876
Te_	286.824	3.769
Rb_	20.128	3.666
Sr_	118.252	3.244
Zr_	34.721	2.784
Nb_	29.689	2.82
Mo_	28.179	2.72
Tc_	24.154	2.671
Ru_	28.179	2.64
Rh_	26.67	2.61
Pd_	24.154	2.583
Ag_	18.115	2.805
Cd_	114.73	2.538
Tl_	342.176	3.873
Pb_	333.622	3.829
Bi_	260.658	3.894
Po_	163.54	4.196
At_	142.909	4.233
Rn_	124.794	4.246
Cs_	22.644	4.025

Ba_	183.165	3.3
Hf_	36.23	2.799
Ta_	40.759	2.825
Re_	33.211	2.632
Os_	18.618	2.78
Ir_	36.734	2.531
Pt_	40.256	2.454
Au_	19.625	2.934
Hg_	193.732	2.41
Fr_	25.16	4.366
Ra_	203.293	3.276
La_	8.554	3.138
Ce_	6.542	3.169
Pr_	5.032	3.213
Nd_	5.032	3.186
Pm_	4.529	3.161
Sm_	4.026	3.137
Eu_	4.026	3.112
Gd_	4.529	3.001
Tb_	3.522	3.075
Dy_	3.522	3.055

Ho_	3.522	3.038
Er_	3.522	3.022
Tm_	3.019	3.006
Yb_	114.73	2.99
Lu_	20.631	3.243
Ac_	16.606	3.099
Th_	13.083	3.026
Pa_	11.07	3.051
Np_	9.561	3.051
Pu_	8.051	3.051
Am_	7.045	3.013
Cm_	6.542	2.964
Bk_	6.542	2.975
Cf_	6.542	2.952
Es_	6.038	2.94
Fm_	6.038	2.928
Md_	5.535	2.917
No_	5.535	2.894
Lw_	5.535	2.883
H2	34.2	2.96

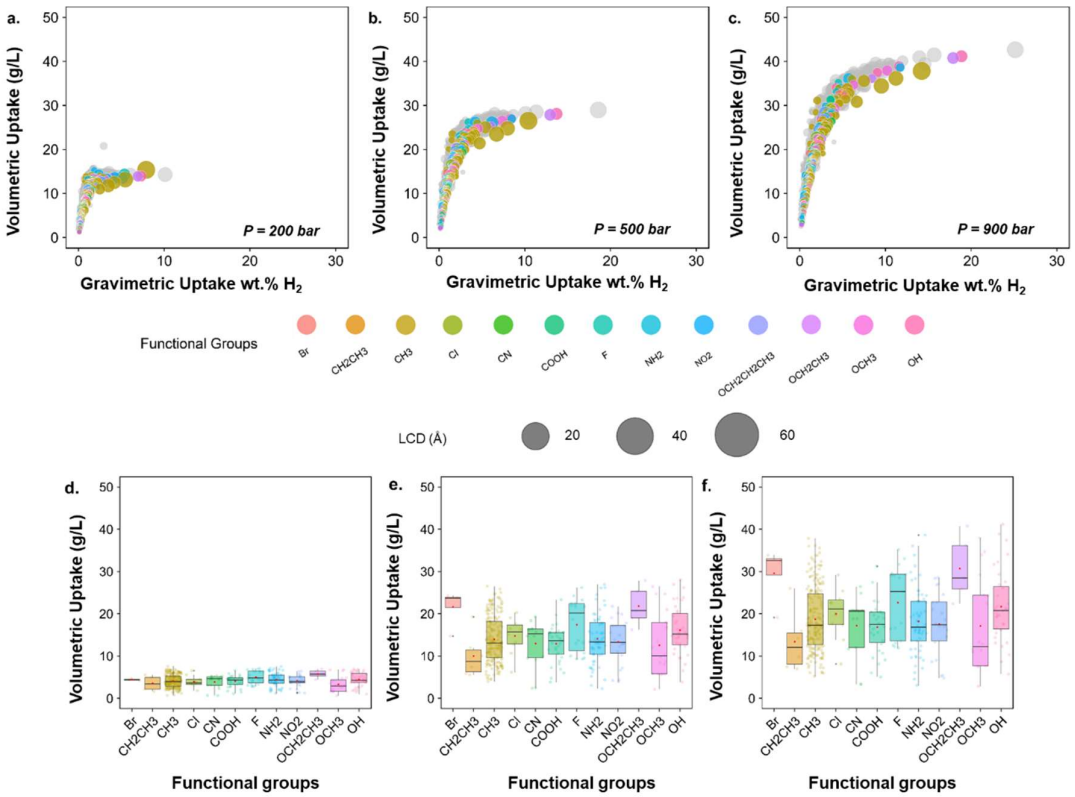


Figure H1 Volumetric uptake versus gravimetric uptake in wt.% H_2 for the screened structures for hydrogen storage at **a.** 200 bar, **b.** 500 bar and **c.** 900 bar. Each circle corresponds to a structure. The colours highlight the functional groups the structures contain, the size of the circles indicate the largest cavity diameter (LCD). **d.-f.** Boxplots of the volumetric uptake for each functional group. The markers represent the minimum, first quartile, median, third quartile, and maximum values, respectively. Outliers are represented by black data points.

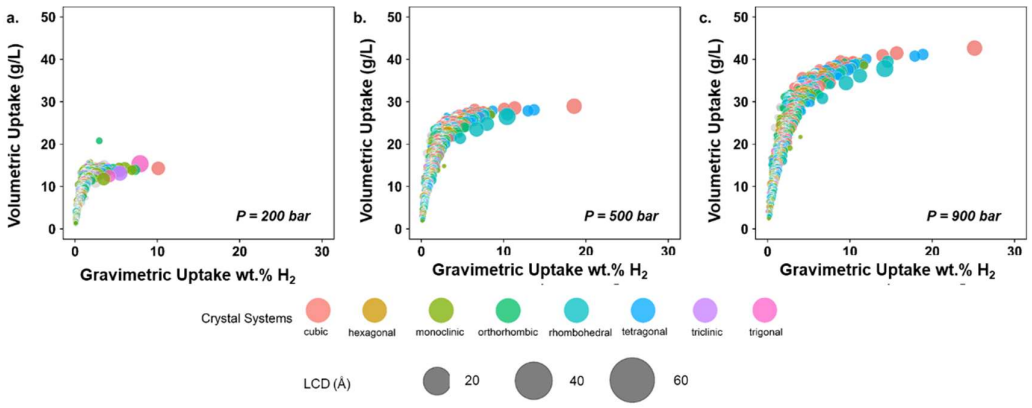


Figure H2 Volumetric uptake versus gravimetric uptake in wt.% H_2 for the screened structures for hydrogen storage at **a.** 200 bar, **b.** 500 bar and **c.** 900 bar. Each circle corresponds to a structure. The colours highlight the crystal systems of each structure, and the size of the circles indicate the largest cavity diameter (LCD).

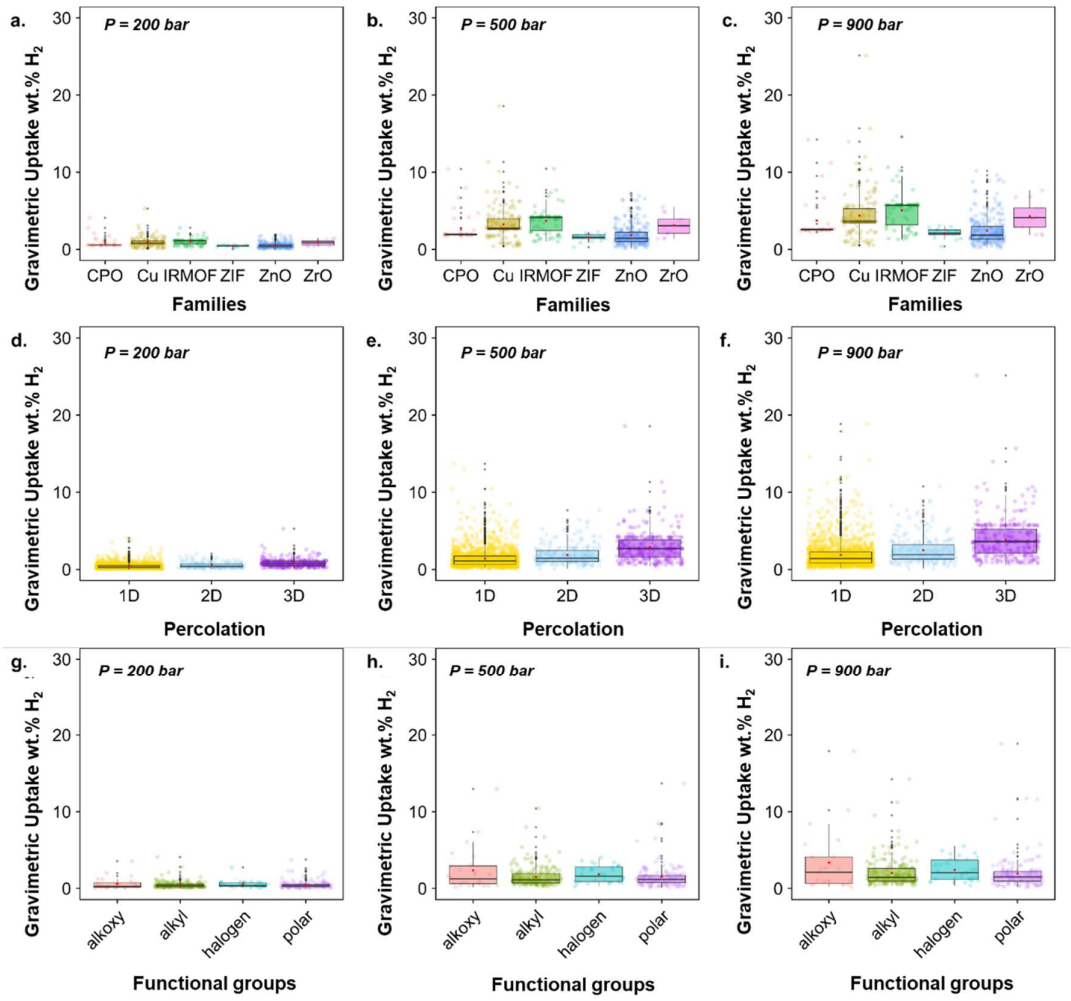


Figure H3 Quantitative characterisation of the 3D MOFs screened for hydrogen storage. Boxplots of gravimetric uptake of H_2 at room temperature at 200, 500 and 900 bar versus **a**.-**c.** families of the screened structures, **d.-f.** percolation of the screened structures and **g.-i.** functional groups identified in the screened structures. The jittered points in the background give an idea on the number of structures considered for each boxplot. The markers represent the minimum, first quartile, median, third quartile, and maximum values, respectively. Outliers are represented by black data points.

APPENDIX I IDENTIFICATION OF METAL-ORGANIC CAGES AND ORGANIC CAGES WITH TOPOLOGICAL DATA ANALYSIS – FURTHER DETAILS

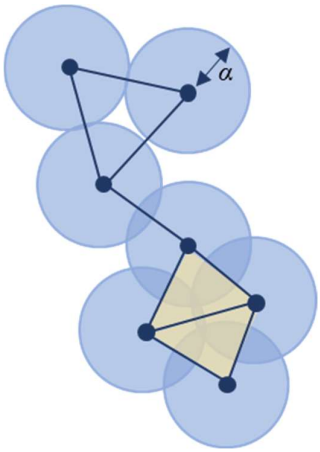
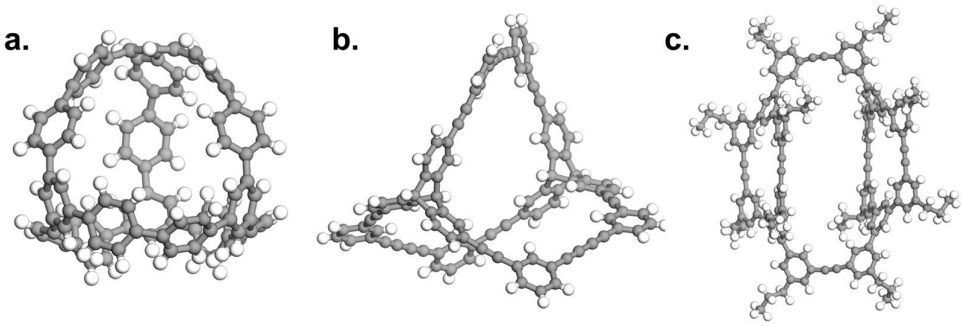


Figure I1 Vietoris-Rips filtration: set of points where the distance between two points is less or equal than alpha.

Types of OCs identified with the queries presented in Figure 62.

Carbon-based cages



E

Figure I2 Examples of targeted carbon-based cages. CSD refcodes: **a**. CIMCIM, **b.** LISTOX and **c.** YOHXOK.

Imine-based cages

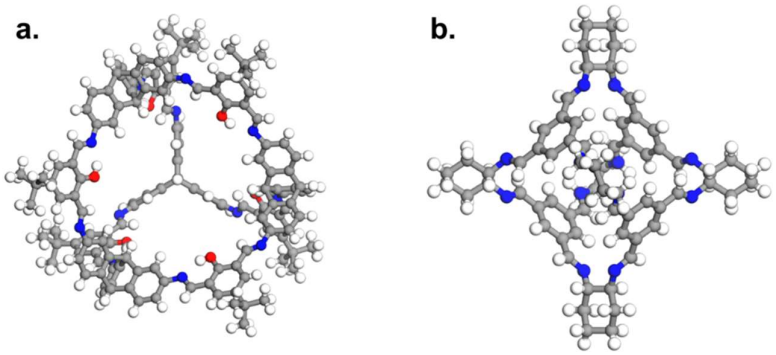


Figure I3 Examples of targeted imine-based cages. CSD refcodes: **a.** EKUKUR and **b.** FOXLAG.

Boronate-based cages

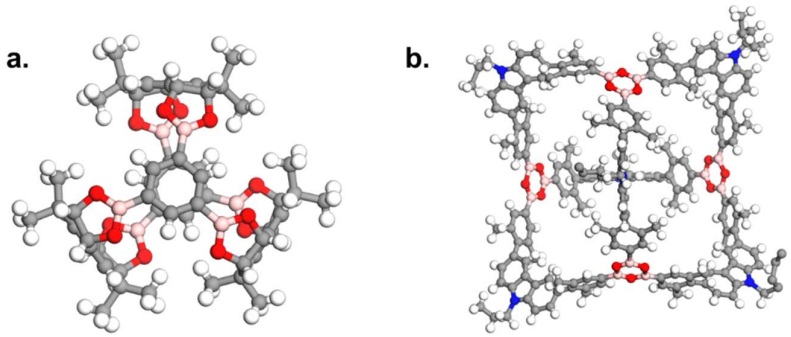


Figure I4 Examples of targeted boronate-based cages. CSD refcodes: **a.** AJOHUD and **b.** YUKHOD.

Oxygen-based cages

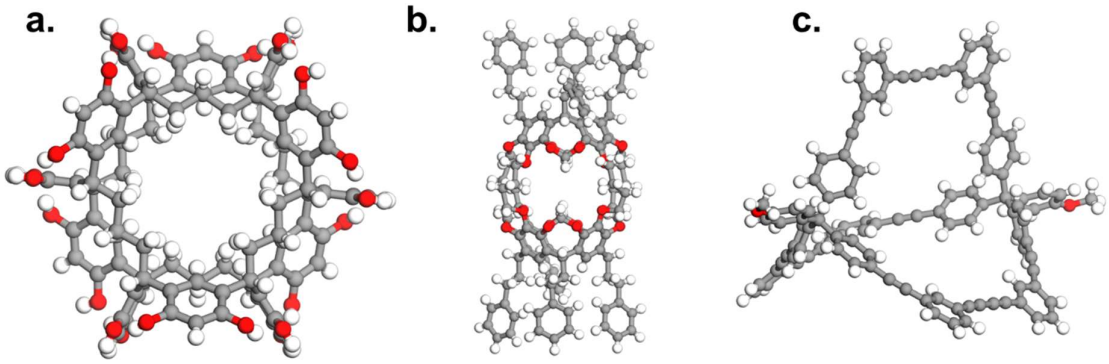


Figure I5 Examples of targeted oxygen-based cages. CSD refcodes: **a.** GUMCIB, **b.** PAQFES and **c.** REQXES.

Python scripts used for the TDA

I wrote the entirety of this script using the CSD Python API, numpy, pandas, gudhi, matplotlib, seaborn and scipy.

```
"""
PART I - DATA PREPARATION
"""

## Part Ia - check the structures identified with ConQuest and obtain
the clean CIFs
import ccdc.molecule
from ccdc.io import EntryReader, CrystalWriter
import csv

# Read list of structures identified with ConQuest
csd_reader = EntryReader('CSD')
MOF_list_entries = EntryReader('C:/Potential cages/all_queries.gcd')

potential_cage_list = []
total_structures = 0
potential_structures = 0

for entry in MOF_list_entries:
    atom_list=[]
    refcode = entry.identifier
    # Check that the heaviest weight component has organic parts, and
if so,
    # keep it. Otherwise, if it's not organic at all, look at the
other components
    # (there should be one other in general) and check that it is
organometallic.
    if 'Atom(C' in str(entry.molecule.heaviest_component.atoms):
        cage = entry.molecule.heaviest_component
    else:
        for idx, component in
enumerate(list(entry.molecule.components)):
            if entry.molecule.components[idx] !=
entry.molecule.heaviest_component:
                if 'Atom(C' in
str(entry.molecule.components[idx].atoms) and
entry.molecule.components[idx].is_organometallic:
                    cage  = entry.molecule.components[idx]
    # Check that at least one atom is part of a 'cycle'
    for atom in cage.atoms:
        if atom.is_cyclic is True:
            atom_list.append(atom)
    if len(atom_list) != 0:
        potential_cage_list.append(refcode)
        potential_structures += 1
        total_structures += 1
        print("This is the", potential_structures, "th potential
structure out of", total_structures)
        print("This structure is", refcode)
        entry.crystal.molecule = cage
        # Write out the corresponding cif
        with CrystalWriter(refcode+'.cif') as cryst_writer:
            cryst_writer.write(entry.crystal)
        # Write out the final list of structures
        with open('all_potential_cages_selected.gcd', 'a+') as f:
            f.write(refcode)
            f.write("\n")
```

```
            f.close()

## Part Ib - extract the fractional coordinates from the CIFs

# Specify the directory we want the csv files to be written to
directory = 'C:/Potential cages/'

# Function to remove parenthesis in cifs
def remove_parenthesis(list_of_strings):
    '''
    Takes a list of strings that contains numbers and something like
this:([0-9][0-9]).
    Remove the parentheses and what's between them
    Return a list of floats
    '''
    for i, element in enumerate(list_of_strings):
        if '(' in element:
            first = element.find('(')
            second = element.find(')')
            element = element[:-(second - first +1)]
            element = float(element)
            list_of_strings[i] = element
    return list_of_strings

# Write coordinates into CSV
for filename in os.listdir(directory):
    if filename.endswith(".cif"):
        cif_reader = io.EntryReader(filename)
        for cif in cif_reader:
            cif = cif_reader[0] # need to specify which datablock in
the CIF we're reading
            if cif.has_3d_structure:
                try:
                    atom_x_coordinate =
remove_parenthesis(cif.attributes['_atom_site_fract_x'])
                    atom_y_coordinate =
remove_parenthesis(cif.attributes['_atom_site_fract_y'])
                    atom_z_coordinate =
remove_parenthesis(cif.attributes['_atom_site_fract_z'])
                    rows = zip(atom_x_coordinate, atom_y_coordinate,
atom_z_coordinate)
                    with open(filename[:-4]+".csv", "w") as f:
                        writer = csv.writer(f)
                        for row in rows:
                            writer.writerow(row)
                except:
                    print('Error when extracting coordinates from',
filename)

"""
PART II - Obtaining the persistent landscapes
"""
import numpy as np
import gudhi as gd
import os
import time
import pandas as pd
import gudhi.representations

# Prepare dictionaries for easier browsing of diagrams (results for
diagrams,
# times for computation time and simplex_trees for simplex_trees)
results = {}
```

```
times = {}
simplex_trees = {}

# Read in the relevant list of structures
gcd_list = open("query_1.gcd", 'r').read()
refcodes = gcd_list.split("\n")

for refcode in refcodes:
    if os.path.exists(refcode+".csv"):
        print('Loading', refcode)
        # Extract the corresponding coordinates from CSV
        coordinates = np.genfromtxt(refcode+".csv", delimiter=",")

        print('Now trying Rips')
        # calculate persistence (I know some structures such as ABIGEY
don't work
        # hence the try) (this is actually because of memory issues)
        try:
            # Keep track of computation time
            start=time.time()
            # Perform TDA
            Rips_complex_sample = gd.RipsComplex(points = coordinates,
max_edge_length=0.8 )
            Rips_simplex_tree_sample =
Rips_complex_sample.create_simplex_tree(max_dimension=3)
            diag_Rips = Rips_simplex_tree_sample.persistence()
            stop=time.time()
            # store persistent diagrams in results
            results[refcode]=diag_Rips
            times[refcode]=stop-start
            simplex_trees[refcode] = Rips_simplex_tree_sample
            print('Persistence for', refcode, 'calculated')
        except:
            print('Something went wrong with', refcode)
            results[refcode]='Error'
        # Calculate landscapes, L1 for Betti 1 and L2 for Betti 2
        LS = gd.representations.Landscape(resolution=1000)
        print('LS calculated')
        L1 =
LS.fit_transform([Rips_simplex_tree_sample.persistence_intervals_in_di
mension(1)])
        print('L1 calculated')
        L2 =
LS.fit_transform([Rips_simplex_tree_sample.persistence_intervals_in_di
mension(2)])
        print('L2 calculated')
        L = np.concatenate((L1[0], L2[0]))
        print('L computed')
    else:
        results[refcode]='xyz missing'
        times[refcode]='NA'
        print('xyz file not found')
    w = csv.writer(open("results.csv", "a"))
    w.writerow([refcode, results[refcode]])
    t= csv.writer(open("times.csv", "a"))
    t.writerow([refcode, times[refcode]])
    # landscape is saved as an NPY
    with open(refcode+".npy", 'wb') as f:
        np.save(f,L)
    print("trees array saved")

"""
PART III - Noise removal
```

```
"""
# Process previously obtained dictionary results
results.columns=['refcode', 'persistence']
results = results.query("persistence != 'xyz missing'")
results = results.dropna()
results = dict(zip(list(results.refcode), list(results.persistence)))

# Check that all the entries are 'complete'.
for refcode in results.keys():
    if results[refcode][-1] != ']' :
        find_index = len(results[refcode])-1
        match = ', (0, (0'
        segment = results[refcode][-1]
        while match not in segment:
            find_index -= 1
            segment_new = results[refcode][find_index] + segment
            segment = segment_new
        results[refcode] = results[refcode][:find_index-1] + ')]'

# The persistence lists are read in as strings so we need to
# convert them to lists first
# But, the presence of 'inf' in the list make it hard for ast or jason
to
# convert them to lists so we need to replace the places where inf
appears first

for refcode in results.keys():
    try:
        string_to_convert = results[refcode]
        string_to_convert = string_to_convert.replace("(0, (0.0, inf)),
", "")
        results[refcode]=eval(string_to_convert)
    except:
        print('error with', refcode)

# look for structure wihout betti 2 numbers
def check_betti_2(results, refcode):
    '''
    This function takes a persistence diagram and checks if there are
results
    corresponding to a betti number of 2.
    '''
    if results[refcode][0][0] < 2:
        return False
    else:
        return True

# Gather these structures without betti 2s into a new dict and also
save the list in a csv file
no_betti_2 = {}
for refcode in results.keys():
    if check_betti_2(results, refcode) == False:
        w = csv.writer(open("no_betti_2.csv", "a"))
        w.writerow([refcode])
        no_betti_2[refcode]=results[refcode]

# We want to compare the persistence diagrams cooresponding to betti
numbers 1 and 2
# We need to extract the diagrams as a list of lists type
def persistence_to_compare(results, refcode, betti):
    '''
    results - dictionary of persistence diagram previously obtained in
the format of
```

```
            keys: refcodes, values: persistence diagram (list of
tuples (betti, (birth, death)))
    refcode - structure we are looking at
    betti - the number we want to look at, 1 or 2
    This function returns a persistence diagram of betti number for a
given structure by
    preparing the diagram in the format of a list of lists [[birth,
death], [birth, death]]
    '''
    persistence = []
    i = 0
    while results[refcode][i][0] != betti and i <=
len(results[refcode]):
        i += 1
    while results[refcode][i][0] == betti and i <
len(results[refcode]):
        persistence.append(list(list(results[refcode][i])[1]))
        i += 1
    return persistence

# Load results obtained previsouly
structures = results.keys()

# Prepare dataframes with all the bottleneck distances
heatmap_data_1 = pd.DataFrame(index = structures, columns = structures)
heatmap_data_2 = pd.DataFrame(index = structures, columns = structures)

for refcode in structures:
    for refcode_to_compare in structures:
        if refcode not in no_betti_2.keys() and refcode_to_compare not
in no_betti_2.keys():
            print(refcode, refcode_to_compare)
            print('Comparing', refcode, 'and', refcode_to_compare)
            heatmap_data_2.loc[refcode][refcode_to_compare] =
gd.bottleneck_distance(persistence_to_compare(results, refcode, 2),
persistence_to_compare(results, refcode_to_compare, 2))
            heatmap_data_1.loc[refcode][refcode_to_compare] =
gd.bottleneck_distance(persistence_to_compare(results, refcode, 1),
persistence_to_compare(results, refcode_to_compare, 1))

# Function to identify similar structures
def find_similar(heatmap, refcode, threshold):
    '''
    This function transforms the bottleneck distances stored in
heatmap into
    another similarity scale:
        1 - structure is very similar (corresponds to the structure
itself)
        0 - structure is completely dissimilar
    Then, the function looks at the second most similar structure
apart
    from itself, and looks for the closest values to that second most
similar
    structure
    Threshold corresponds to the % to which we want the structures to
be similar
    to that second most similar structure.
    Returns a dict of similar structures with name: bottleneck
distance
    '''
    similar = {}
    refcode_col = heatmap.loc[refcode]
```

```
    # normalising the scale of values and ranging similarity from 1
(very similar) to 0 (least similar) and take out the value
corresponding to refcode
    norm_values = refcode_col.sort_values().transform(lambda
x:(refcode_col.max()-x)/refcode_col.max())
    indices = norm_values.index.values.tolist()
    # after being sorted, the second element has the highest
similarity value (doesn't mean necessarily they are similar in
absolute terms)
    best_max = norm_values.loc[indices[1]]
    similar[indices[0]] = heatmap.loc[refcode, indices[0]]
    # take out the element corresponding to best_max
    norm_values = norm_values.iloc[1:]
    for index in indices[1:]:
        if  norm_values.loc[index] >= threshold*best_max:
            similar[index]=heatmap.loc[refcode][index]

    return similar

"""
PART IV - Classification on TDA landscapes
"""
## Part IVa - hierarchical clustering

import matplotlib.pyplot as plt
import seaborn as sns

# Read in list of structures
gcd_list = open("class7_1.gcd", 'r').read()
refcodes = gcd_list.split("\n")

# Create list of indices for pandas dataframe (not using the refcodes
list
# in case something wrong happens)
indices = []
shapes = {}

# Check the shapes of the structures landscapes
for refcode in refcodes:
    try:
        if os.path.exists(refcode+"_1000.npy"):
            with open(refcode+"_1000.npy", 'rb') as f:
                L = np.load(f)
                shapes[refcode] = L.shape
            indices.append(refcode)
    except:
        print("issue with", refcode)

# Create a master array where each row is a landscape
master = np.empty((1, 10000))
indices_1000 = []
for refcode in shapes.keys():
    if shapes[refcode][0] == 10000:
        with open(refcode+"_1000.npy", 'rb') as f:
            L = np.load(f)
            master = np.vstack([master, L])
            indices_1000.append(refcode)

# Remove first row of approx zeros generated by np.empty
master = np.delete(master, 0, 0)

# create list of columns for pandas dataframe
columns = []
```

```
for i in range(5000):
    columns.append("dimension 1 " + str(i))
for i in range(5000):
    columns.append("dimension 2 " + str(i))

# Convert to pandas dataframe and save
panda_master = pd.DataFrame(data=master, index=indices_1000,
columns=columns)
panda_master.to_csv('cages_final.csv')

# Computing dendrogram
from scipy.cluster.hierarchy import dendrogram, linkage, fcluster
from scipy.spatial.distance import squareform

# Compute linkage
Z = linkage(panda_master, 'ward')

# This function is used later on for plotting dendrogram:
def fancy_dendrogram(*args, **kwargs):
    max_d = kwargs.pop('max_d', None)
    if max_d and 'colour_threshold' not in kwargs:
        kwargs['colour_threshold'] = max_d
    annotate_above = kwargs.pop('annotate_above', 0)

    ddata = dendrogram(*args, **kwargs)

    if not kwargs.get('no_plot', False):
        plt.title('Hierarchical Clustering Dendrogram (truncated)')
        plt.xlabel('Cluster size')
        plt.ylabel('Distance')
        for i, d, c in zip(ddata['icoord'], ddata['dcoord'],
ddata['colour_list']):
            x = 0.5 * sum(i[1:3])
            y = d[1]
            if y > annotate_above:
                plt.plot(x, y, 'o', c=c)
                plt.annotate("%.3g" % y, (x, y), xytext=(0, -5),
                             textcoords='offset points',
                             va='top', ha='center')
        if max_d:
            plt.axhline(y=max_d, c='k')
    return ddata

# look at the dendrogram and decide on a max_d
max_d = 30

# Plot dendrogram with max_d
fancy_dendrogram(
    Z,
    truncate_mode='lastp',
    p=48,
    leaf_rotation=90.,
    leaf_font_size=12.,
    show_contracted=True,
    annotate_above=30,
    max_d=max_d,  # plot a horizontal cut-off line
)
plt.show()

# See number of clusters k
k=11
clusters = fcluster(Z, k, criterion='maxclust')
```

```
# Then assemble data into a dataframe with refcodes and the
corresponding predicted class
labeled = np.array([indices_1000, clusters])
labeled = labeled.T
labeled_pd = pd.DataFrame(data = labeled, columns =['refcode',
'class'])

# To see the data in a dictionary format
classes = {}
for i in range(1,19):
    classes[i] = labeled_pd.loc[labeled_pd['class'] == str(i)]

# Save results as GCD for easy visualisation
for i in range(1,19):
    for j in range(classes[i].shape[0]):
        w = csv.writer(open("class_11"+str(i)+".gcd", "a"))
        w.writerow([classes[i].iloc[j][0]])

## Part IVb - Random forest

from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn import metrics

# importing dataframe and selecting columns
dataset=pd.read_csv('OC_NOC.csv')
dataset.head()

X = dataset.iloc[:-1,1:-1]
Y = dataset.iloc[:-1,-1:]
Y = Y.values.ravel()

# setting an out-split to validate performance after cv
X_cv, X_out,Y_cv, Y_out = train_test_split(X,Y,test_size=0.15,
random_state=42)

#Base model
clf=RandomForestClassifier(n_estimators=100, random_state=75)
clf.fit(X_cv, Y_cv)
Y_pred=clf.predict(X_out)
all_pred=clf.predict(X)

print("Accuracy:",metrics.accuracy_score(Y_out, Y_pred))
print('Mean Absolute Error:', metrics.mean_absolute_error(Y_out,
Y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(Y_out, Y_pred))
print('Root Mean Squared Error:',
np.sqrt(metrics.mean_squared_error(Y_out, Y_pred)))
print('R Squared:', metrics.r2_score(Y_out, Y_pred))
```

CC3 vs M6L4

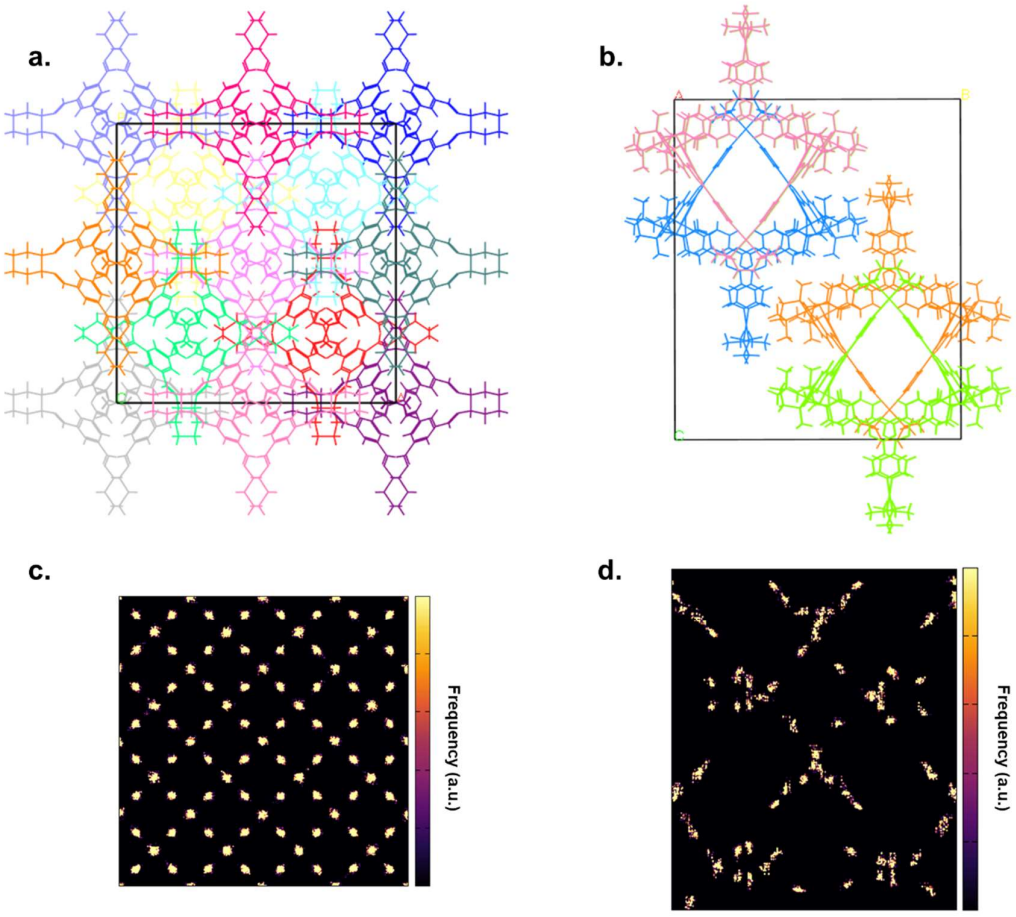
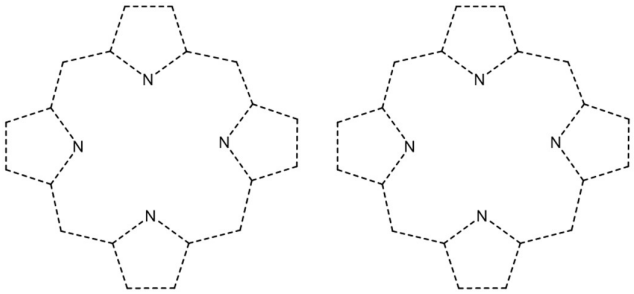
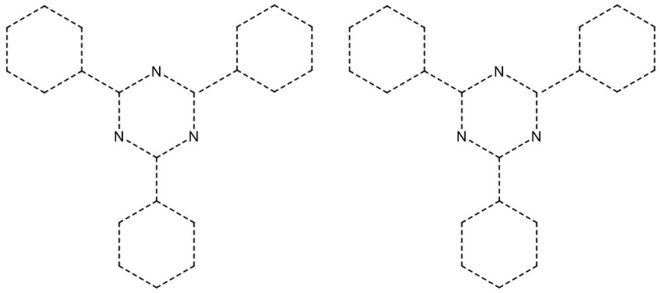
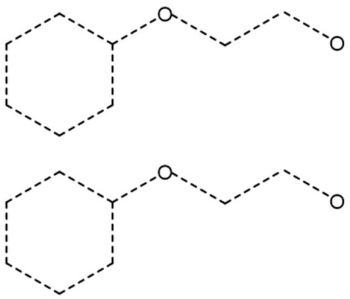


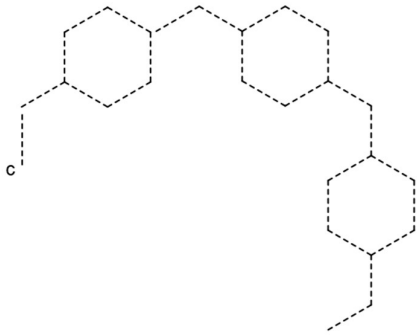
Figure I6 Differences in packing between CC3-type structures and M_6L_4-type structures. **a.** CC3 in its cubic system and **b.** COPPAA in its tetragonal system. The cages are coloured for easier visualisation. The corresponding adsorption sites (obtained with SITES ANALYZER)[287] are shown in **c.** for CC3 and **d.** for COPPAA.

APPENDIX J ADDITIONAL CONQUEST QUERIES USED FOR REDUCING THE SEARCH SPACE OF ORGANIC CAGES IN THE CSD

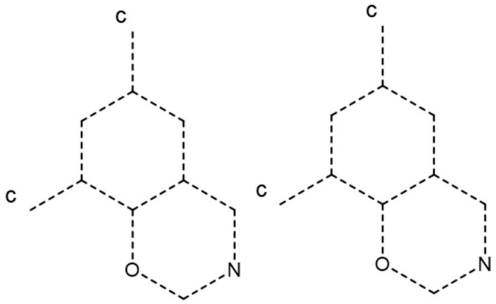
The following queries for organic cages and rings were added to the general queries presented in Chapter 6. Dotted lines correspond to "any" type of bond. Superscript c means the corresponding atom should be cylic. Superscript a means the corresponding atom should be acyclic. Sub-queries highlighted in a red box refer to "must not have" criteria. "TN" means the corresponding atom is attached to N other atoms only.

Table J1 Additional queries for organic cages.

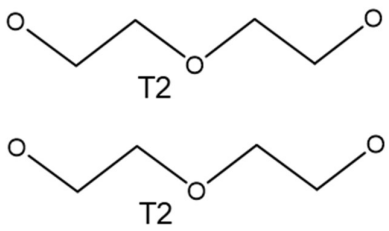
Query	Number of hits
	105
	22
	3670



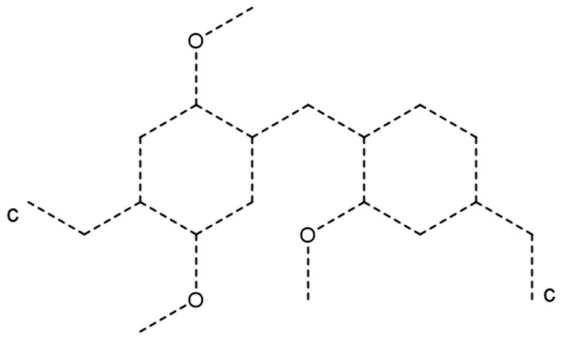
1567



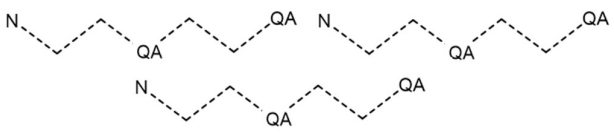
17



3793

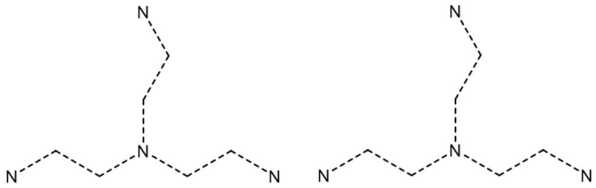


329

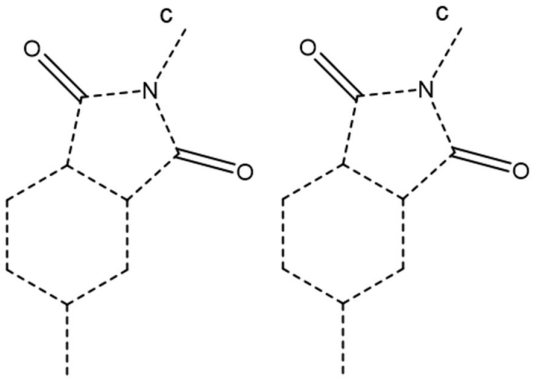


1867

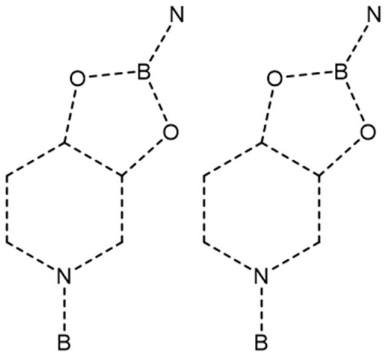
QA = O or N



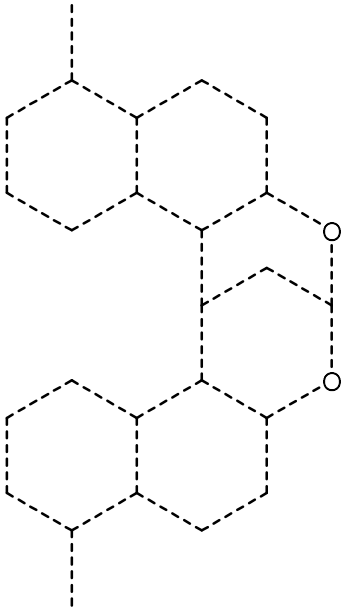
334



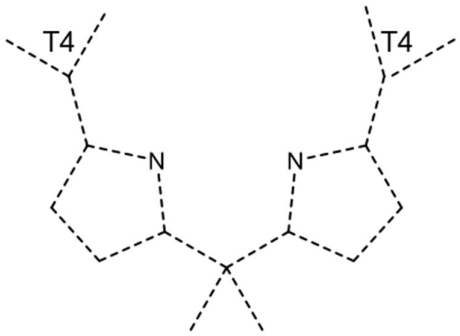
71



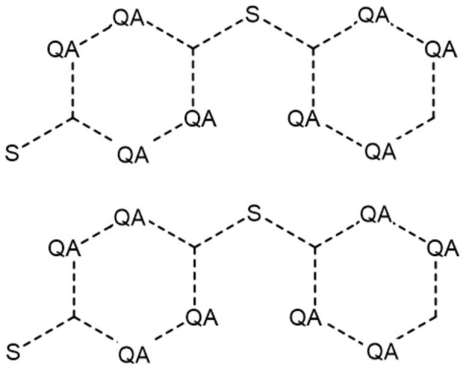
4



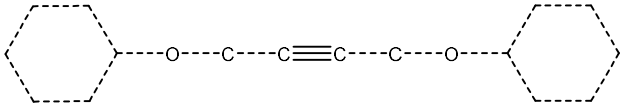
39



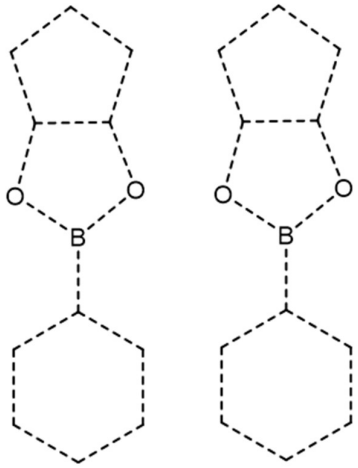
529



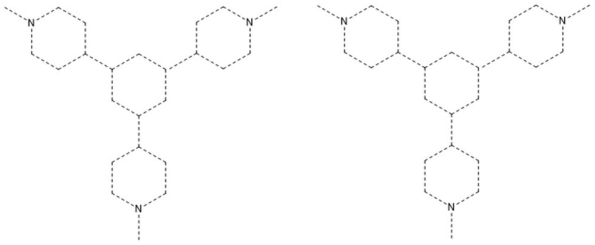
70



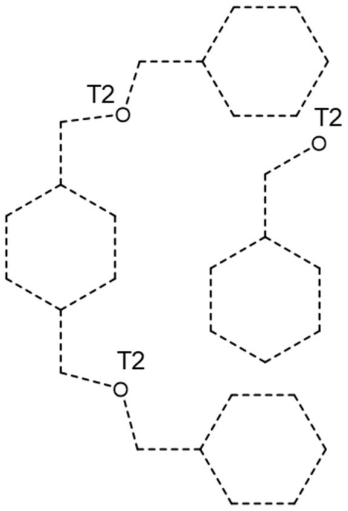
47



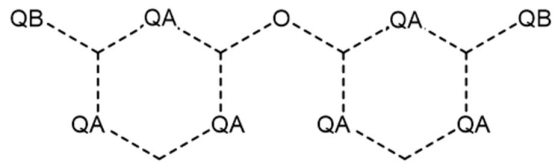
30

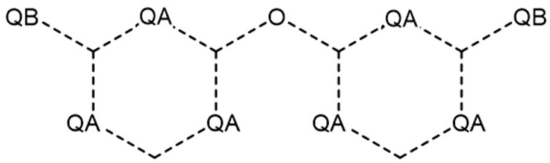


10



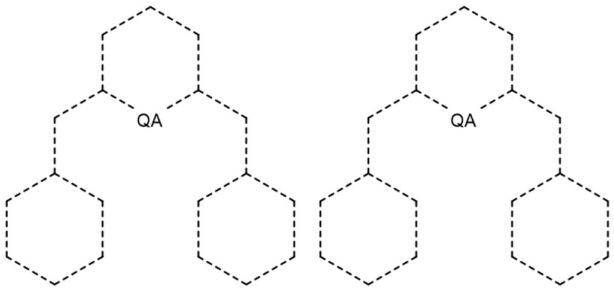
35





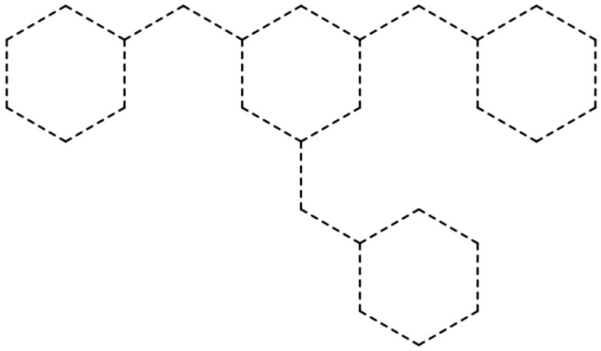
245

QA = C or N, QB = C or O

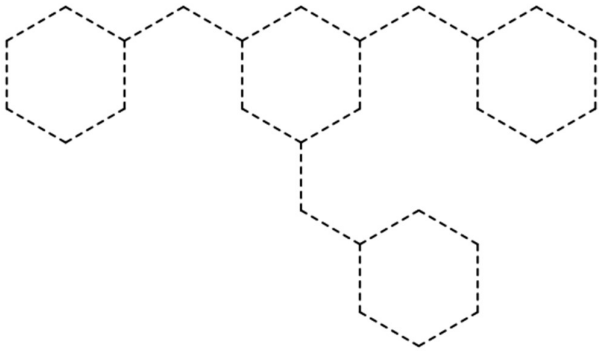


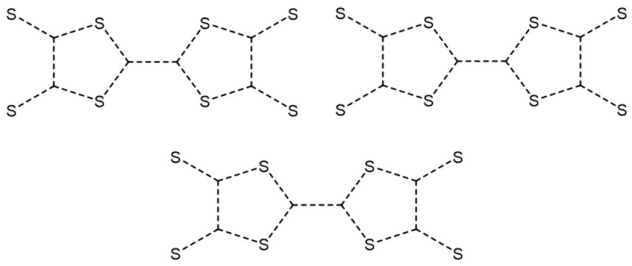
87

QA = C or N

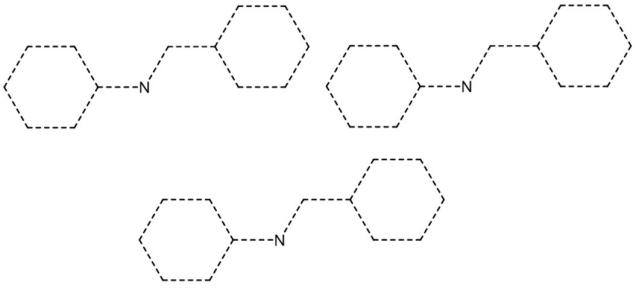


1494

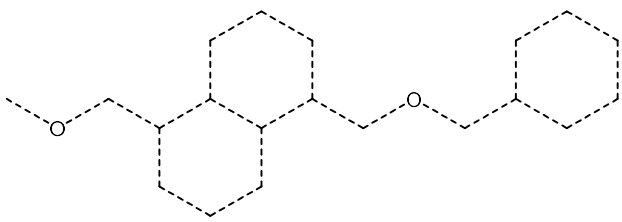




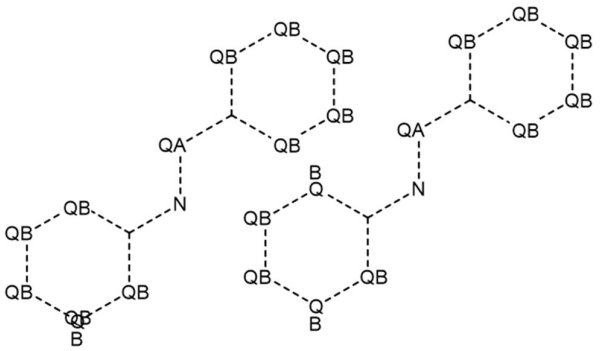
10



605

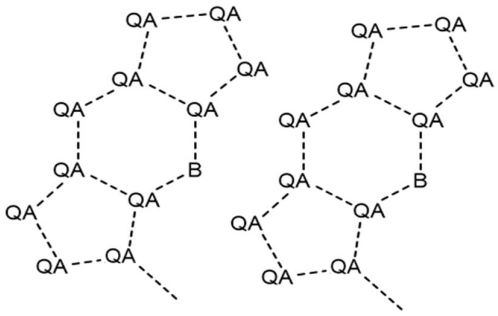


16



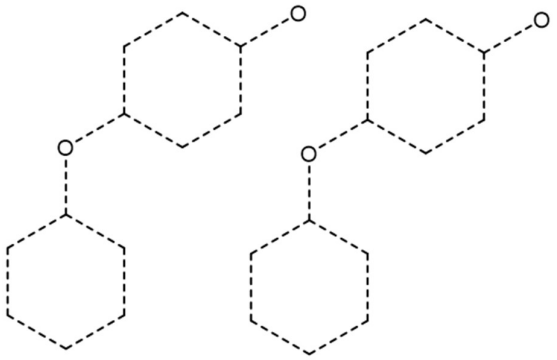
4036

QA = C or S, QB = C or N

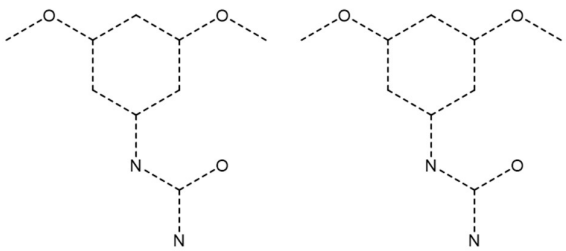


160

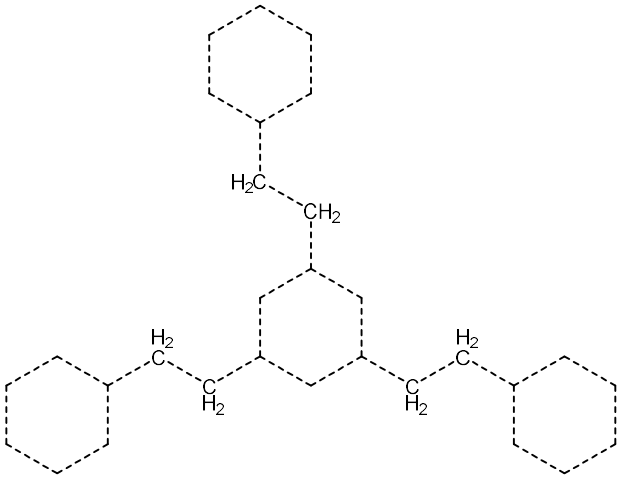
QA = C or N



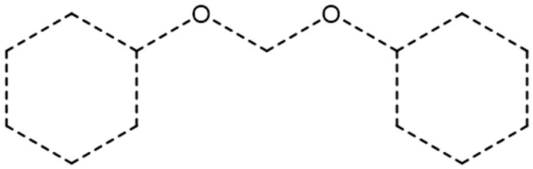
64

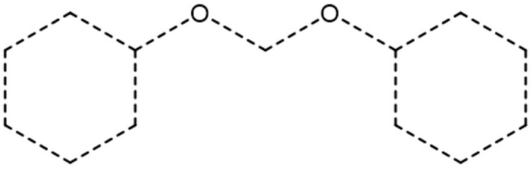


5

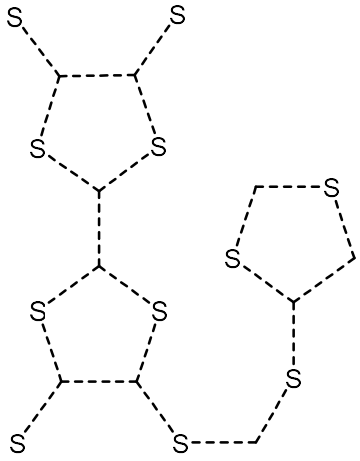


11

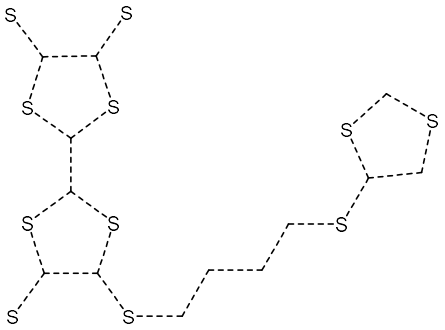




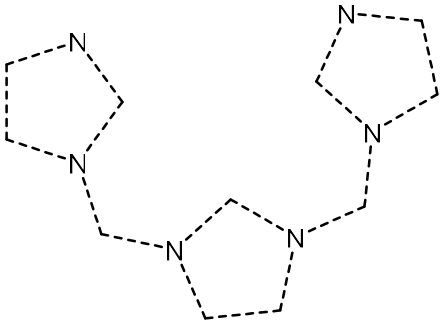
296



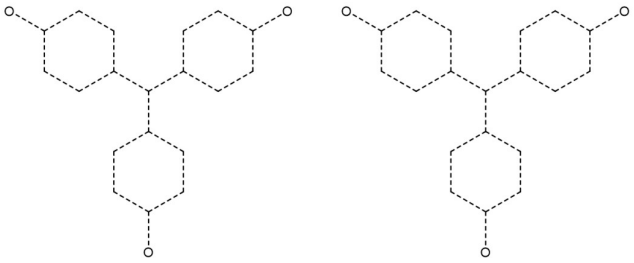
13



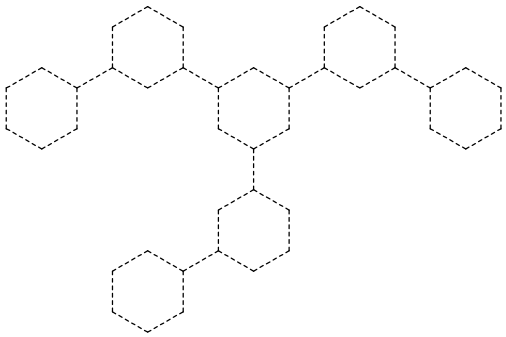
3



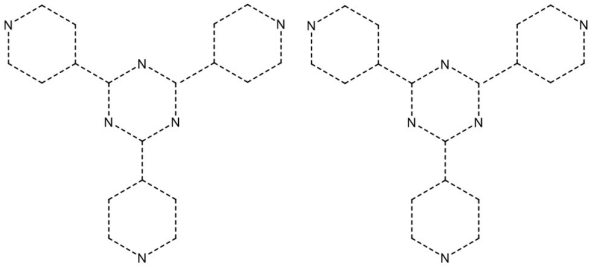
575



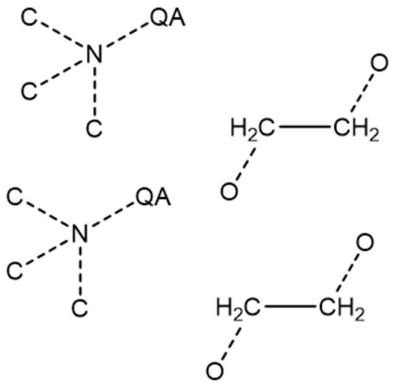
53



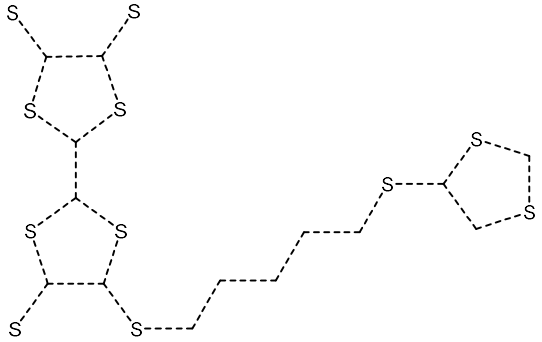
1561



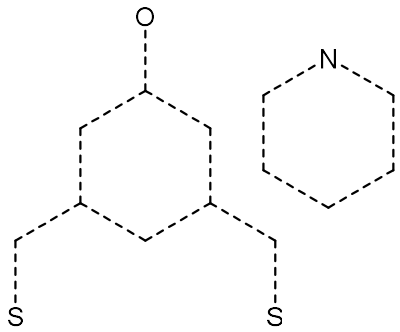
7


C
QA
N
C
C
O
H2C
CH2
O
C
QA
N
C
C
O
H2C
CH2
O

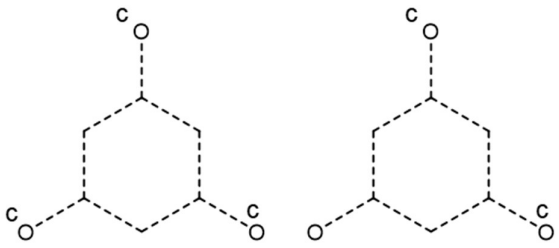
75


S
S
S
S
S
S
S
S
S
S
S
S

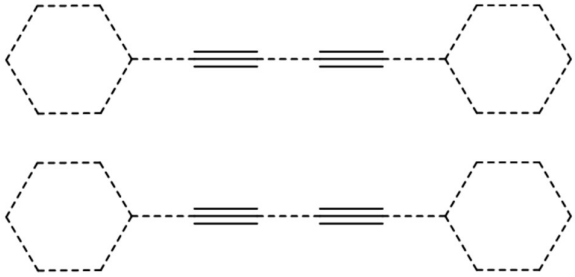
5


O
N
S
S

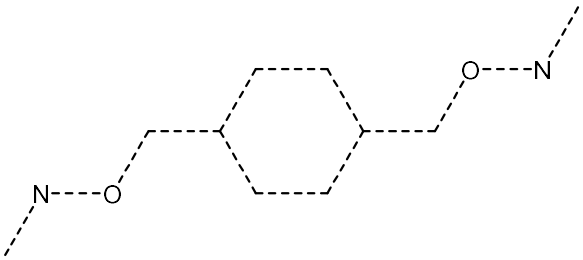
3


c
O
c
O
c
O
c
O
c
O

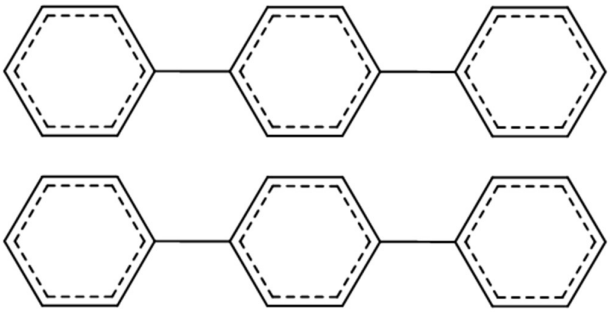
213



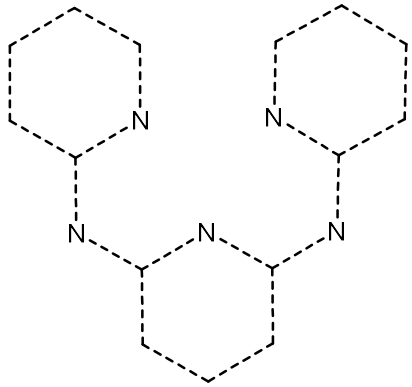
79



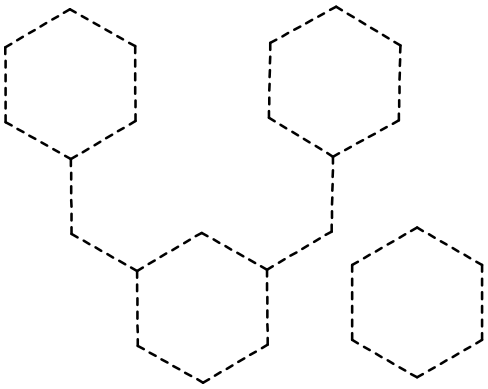
9



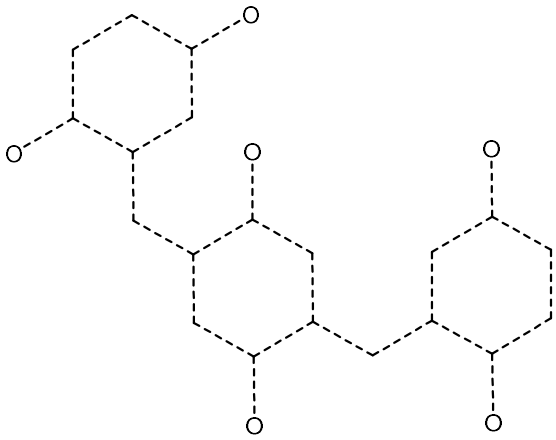
462



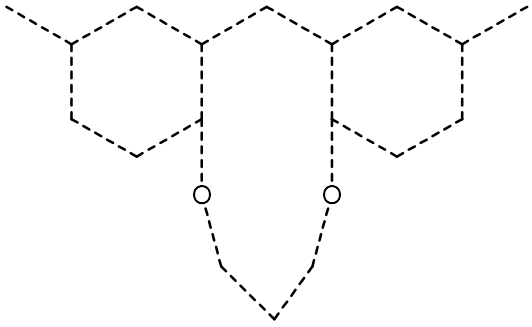
79



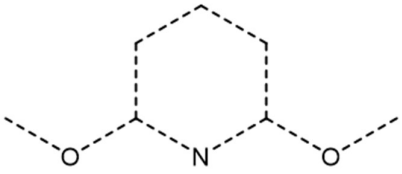
27

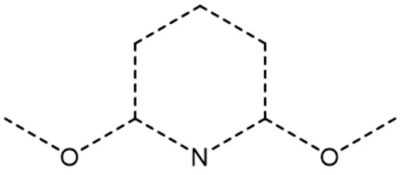


339

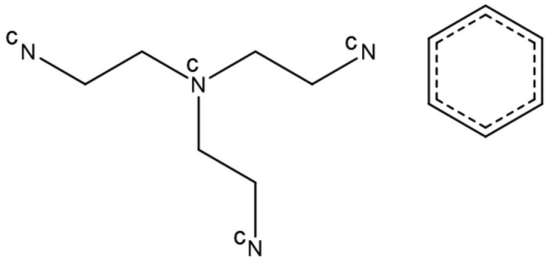


6

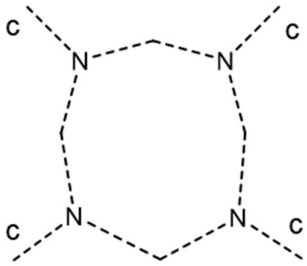




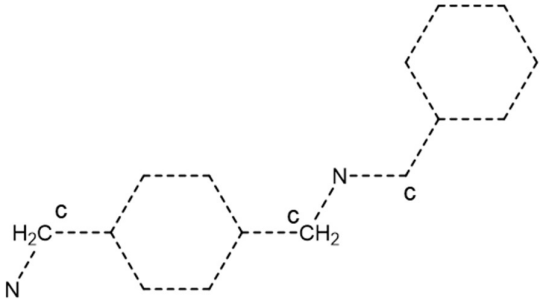
38



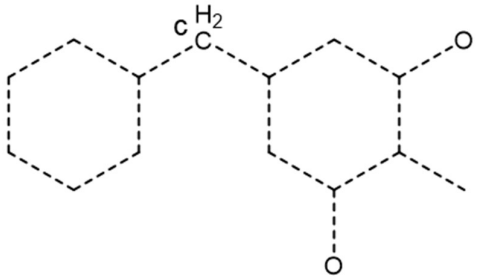
362



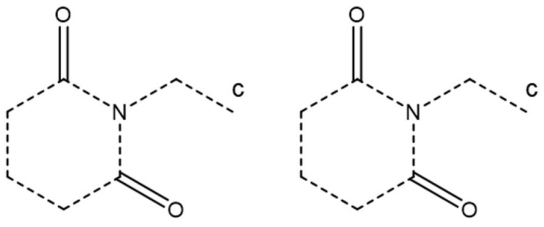
796



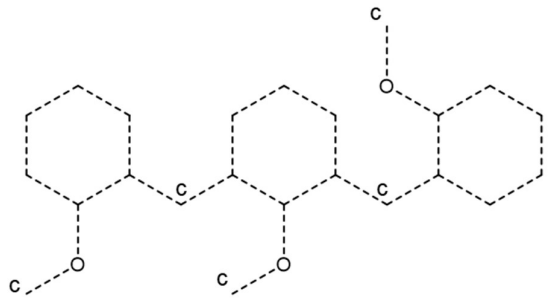
166



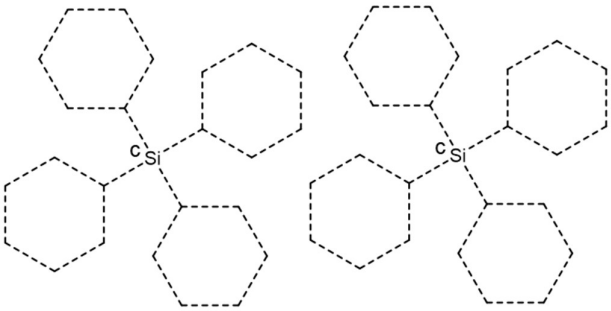
29



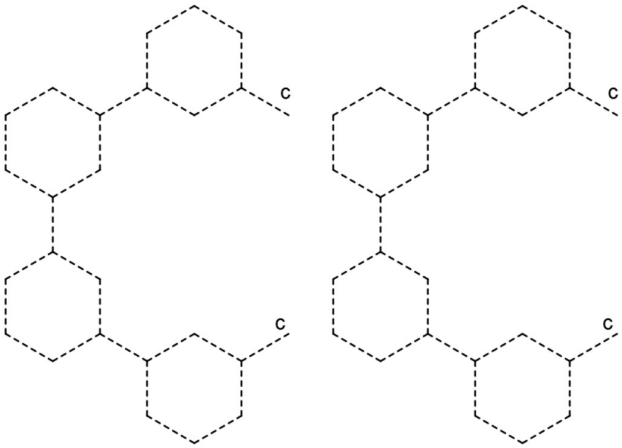
369



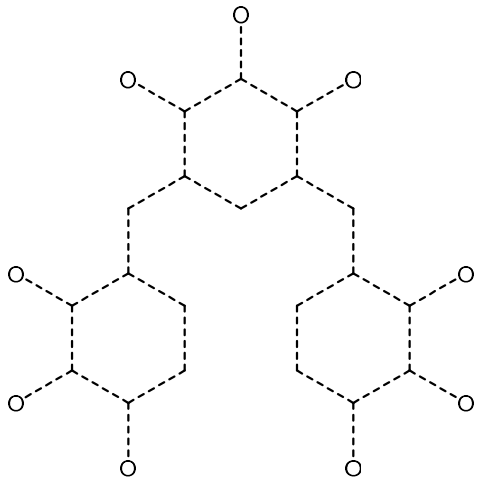
145



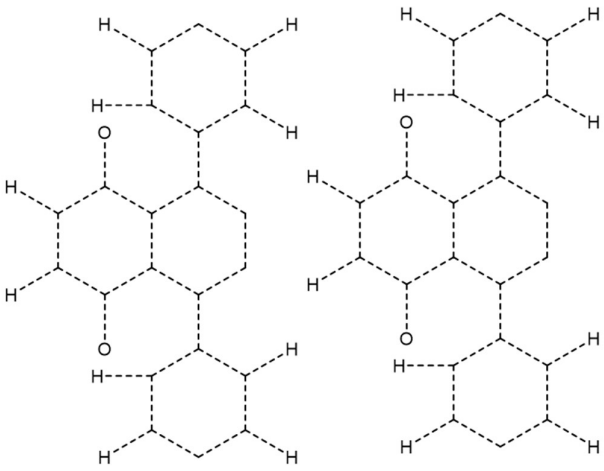
3



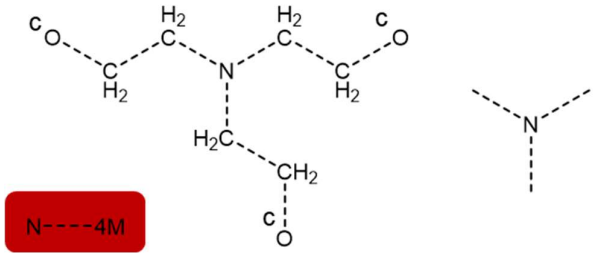
15



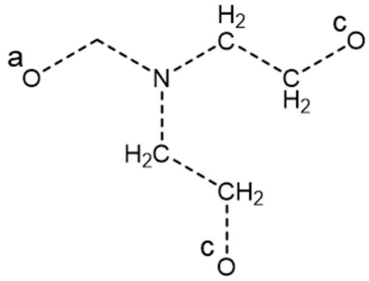
252



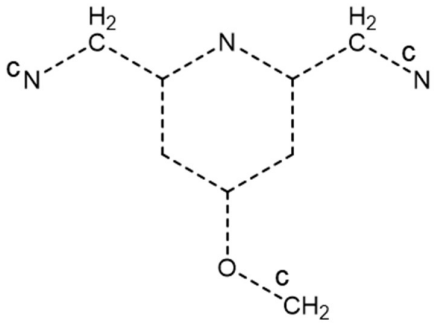
4



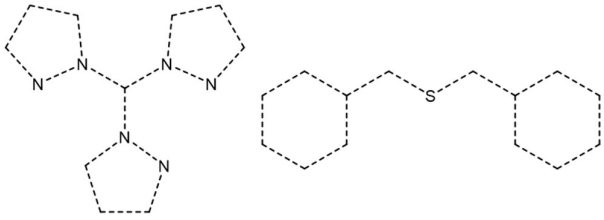
169



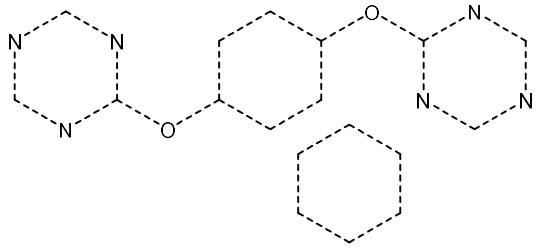
53



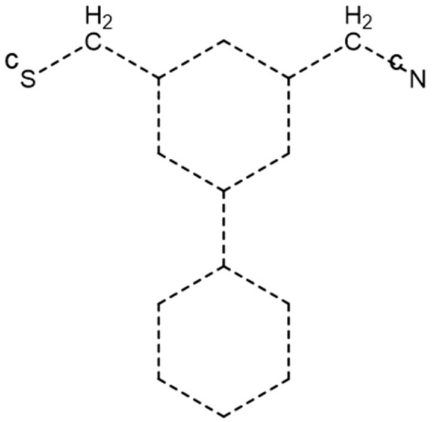
1



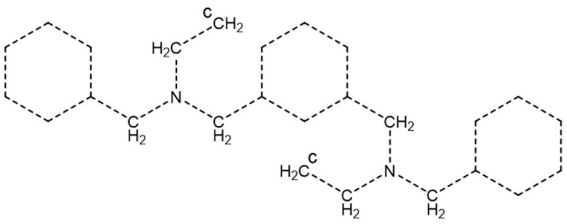
1



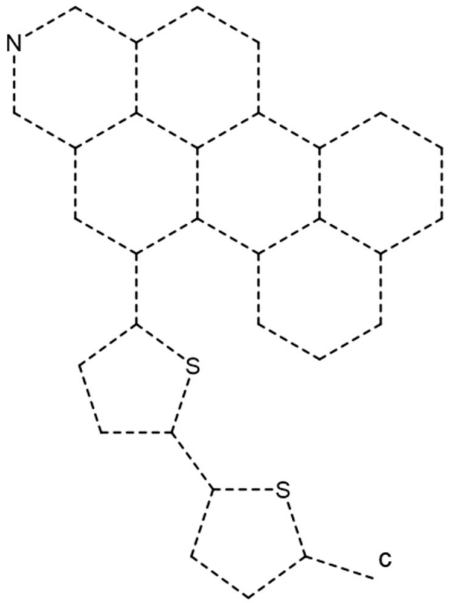
1



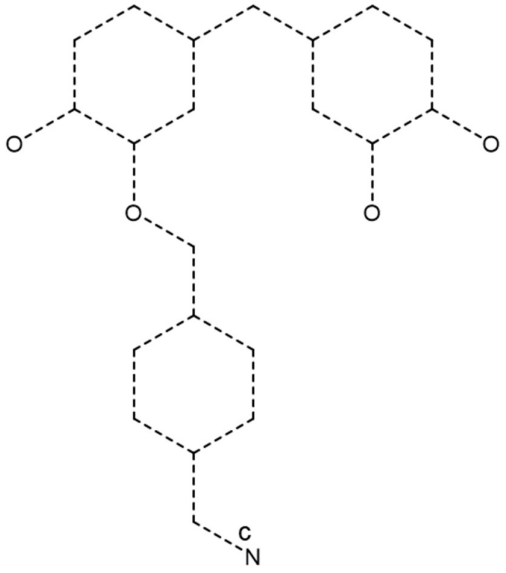
2



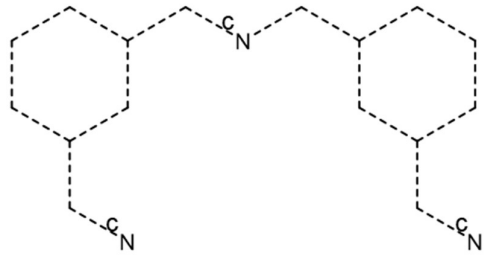
7



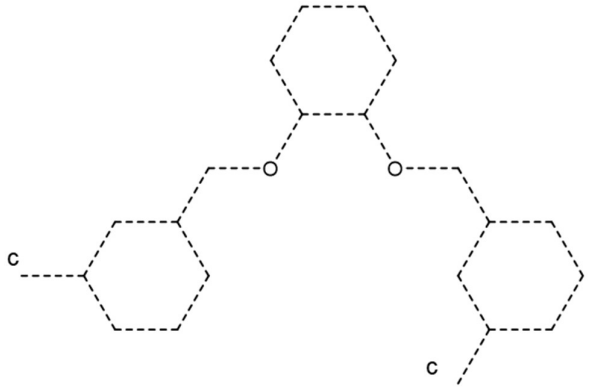
1



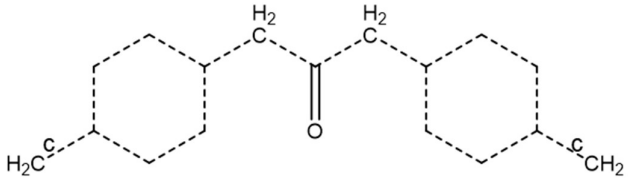
1



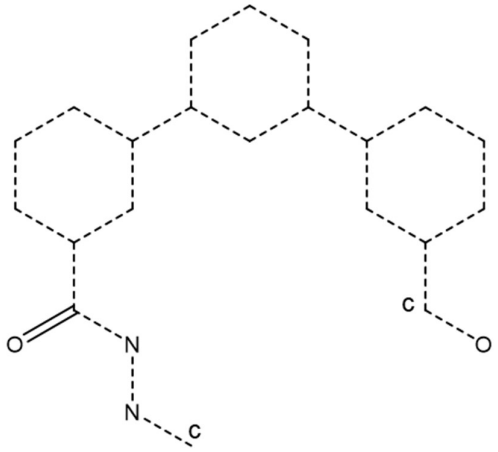
126



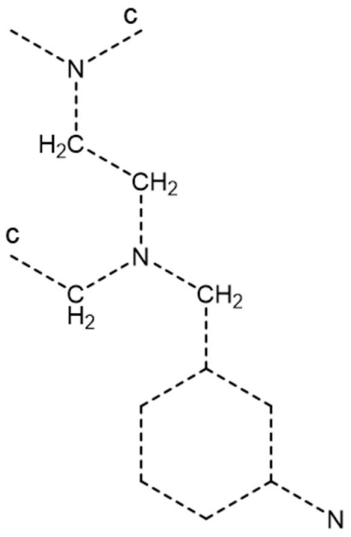
10



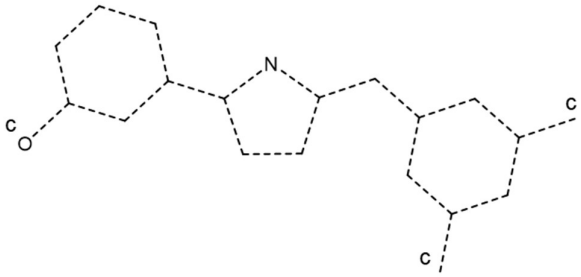
16



1



7



1