# Supplementary Information for "FAIR-MOFs:Structure-centred synthesis inference from three-dimensional structures of metal-organic frameworks"†

A.D. Dinga Wonanke,*[a,b,c,e] Antonio Longa,[f] Asha Pankajakshan,[h] Joseph O. Ogar, [c] Lauri Himanen,[b] Alvin N. Ladines,[b] José A. Márquez,[b] Matthew A. Addicoat, [c] Deborah Crittenden,[d] Markus Scheidgen,[b] Pietro Lio,[g] Stefanie Dehnen,[h] Christof Wöll *[e] and Thomas Heine*[a]

[a] *Chair of Theoretical Chemistry, Faculty of Chemistry and Food Chemistry, Technical University of Dresden, Bergstraße 66c, 01069 Dresden, Germany*

[b] *Consortium FAIRmat, Humboldt-University of Berlin, 12489, Berlin, Germany*

[c] *School of Science and Technology, Nottingham Trent University Nottingham, NG11 8NS, Nottingham, UK*

[d] *School of Physical and Chemical Sciences, University of Canterbury, 8140, Christchurch, New Zealand*

[e] *Institute of Functional Interfaces (IFG), Karlsruhe Institute of Technology (KIT), Eggenstein-Leopoldshafen, Germany*

[f] *Department of Information Engineering and Computer Science, University of Trento, Trento, Italy*

[g] *Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom*

[h] *Institute of Nanotechnology (INT), Karlsruhe Institute of Technology, Karlsruhe (KIT), Eggenstein-Leopoldshafen, Germany*

## Conventional synthetic approach

The synthesis of any new MOF still follows the conventional synthetic approach that begins with a thorough literature survey, risk assessment, calculations of reagent starting quantities and a series of trial-and-error syntheses that are followed by purification and characterisation. This approach is inherently cost-ineffective, environmentally unfriendly and does not provide any heuristics for effective syntheses. The adoption of electronic laboratory notebooks have contributed significantly in speeding up several intermediary processes like enabling proper understanding of risk assessment and determining of quantities of starting reagents and in some cases assisting with characterisation procedures. [1, 2] However, a proper understanding in mitigating the trial-and-error synthesis for optimising the synthesis of novel MOFs is still elusive.

Recently, newer synthetic methods have been developed in an effort to reduce the environmentally unfriendly nature of the existing synthetic approach as well as to intelligently improve the morphology of the newly synthesised MOF. [3] The synthetic methods have gradually moved from traditional solvo(hydro) thermal synthesis to greener methods such as microwave assisted synthesis, mechanochemical synthesis, electrochemical synthesis and sol-gel method. However, although these methods have proven to be resourceful in scalability, which is vital for industrial scale application of MOF, they still do not provide an effective solution to the trial-and-error synthesis of novel MOFs.

Hence, in this study we take a step into providing a solution to this unreliable approach by implementing a framework consisting of a FAIR dataset of MOFs that maps carefully curated crystal structures to experimental synthetic conditions and models for predicting the synthesis conditions directly from 3D structures so as to make experimental synthesis to be more cost-effective and less time consuming.

# S-1 Data Mining

## S-1.1 Extract MOF refcodes from the CSD

The refcode is a unique identifier given to every crystal structure in the CSD. Since there is no generalised convention on how to systematically name MOFs, the CSD refcodes have become the main tool for querying MOFs structures in different databases such as in the CoRE MOF and QMOF. A list of refcodes of the MOF subset in the CSD can be extracted using Conquest as illustrated in Fig. S1. Once conquest is opened, simply click on the icon **view Databases**. From the dropdown menu click on the **Subset in CSD version ...** then click on the **CSD MOF subset** and finally select the subset you desire from the list of subsets. Once this is done the refcodes can be saved in a **.gcd** text file.

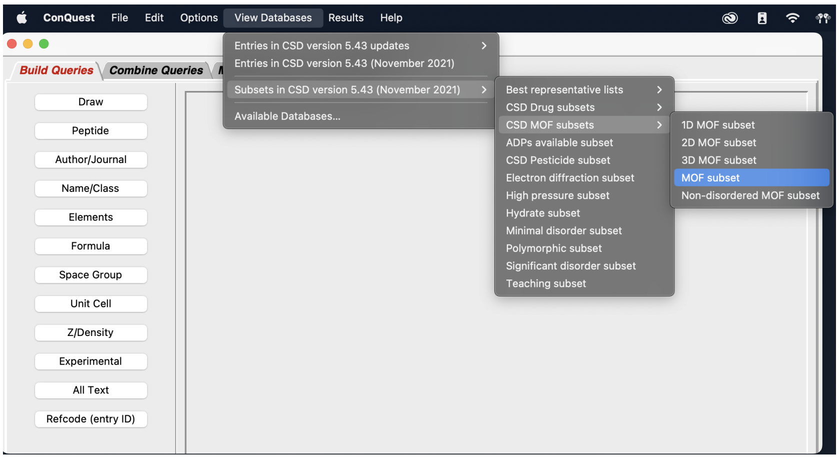


**Fig. S1** Extracting the refcodes of the MOF-subset in CSD using Conquest

## S-1.2 Downloading crystal structures from the CSD

The crystal structures can be downloaded from the CSD in an automated manner using the CSD python API. To facilitate this process, we implemented a python module (csd_data_extraction) that can be used to manipulate structures using the CSD python library. With this module the MOFs structures can be downloaded into a cif format using the following command, where in refcodes is a list of csd refcodes extracted from a file or pandas dataframe.

```
from csd_data_extraction import download_structures
download_structures.main_downloader(refcodes, format='cif')
```

Note that the above script will work only when csd python api is installed and the environment has been properly set up. However, this can only be done if you own a CCDC license. You can always contact us if you require any assistance with setting this up.

### S-1.3 Common errors in the CSD

Despite the requirement for rigorous structural refinement, the MOF subset from the CSD are known to possess several errors in their crystal structures. Lillerud and co-workers showed that while 56 % of all crystal structures in the CSD were of high quality (with an R-factor $< 5$ %) only 20 % of the crystal structures of MOFs had an R-factor $< 5\%$. Meanwhile more than 22 % of the MOFs structures were of significantly poor quality possessing an R-factor $> 10$ %. [4] The R-factor is a numerical value that indicates how well the modelled powder X-ray diffraction (PXRD) pattern matches the experimentally determine PXRD. Generally systems with R-factor $< 5$ % are considered to be of good quality meanwhile those with R-factor $> 10$ % are of extremely low quality with the tendency of systematic errors corresponding to wrong assignment of Laue class. [5]

The three most common errors found in the MOF subsets are **missing hydrogens**, **presence of unbound guest molecules** and **multiple atoms occupying the same symmetrical position** as illustrated in Fig. S2a, Fig. S2b and Fig. S2c respectively.

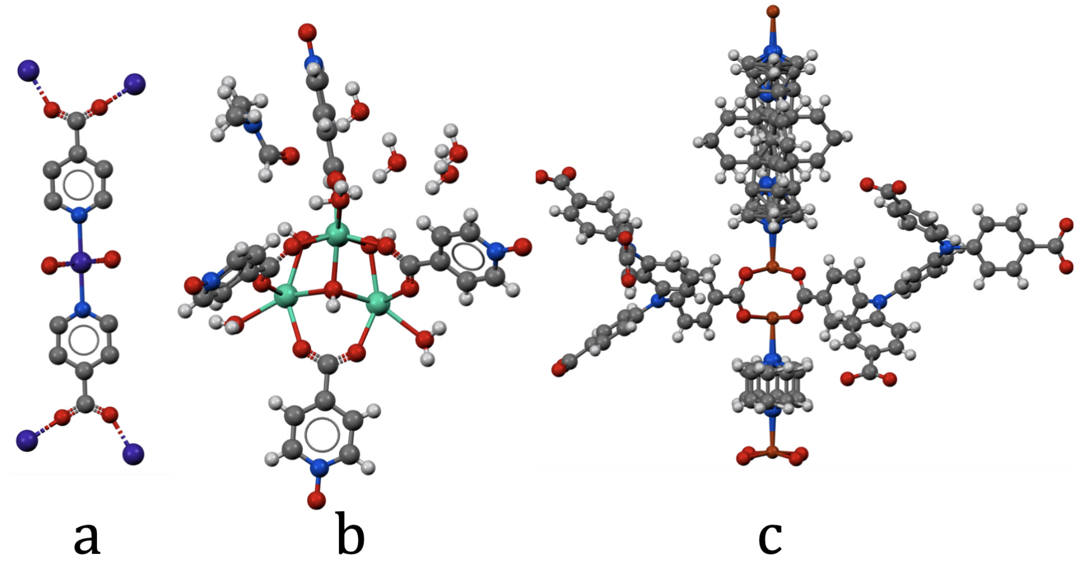


**Fig. S2** Illustration of some of the most common errors found in the CSD. **a** represents an example of a system with one or more missing hydrogen atoms (refcode:ABAVUV). **b** is an example of a system with one or more unbound guest molecules (refcode:CISVEH) and **c** is an example of a system in which multiple atoms occupy the same symmetrical positions in space (refcode:XUDPUJ)

### S-1.3-1 Missing hydrogen

The occurrence of missing hydrogen is expected from crystal structures analysed using X-ray diffraction. This is because X-ray diffraction patterns do not provide information on the position of light atoms like hydrogen and these positions are rather approximated. Hence to add missing hydrogen, we used the CSD Python API to normalise all the bonds from which we added missing hydrogen atoms. This approach is by far the most rigorous bond typing method because it performs bond typing by using a machine learning model that has been trained on millions of structures found in the database. [6] The algorithm begins by dividing a system into small fragments and then calculates the frequencies of the fragments in the CSD. Finally, geometry tests are performed to obtain conditional probability from which Baye's model is applied to assign bond types.

### S-1.3-2 Removal of unbound guest

The presence of unbound guest molecules are ubiquitous in porous systems, which can result from the adsorption of gases in the atmosphere, unreacted reagents and solvents. Fairen-Jimenez and co-workers reported that more than 52 % of MOFs in the CSD have unbound guest molecules in their crystal structures. [7] For this reason, we decided to implement a robust guest removal method, given that several errors have been reported from the existing approaches, which in some cases removes the MOF instead of the guest.

Our guest removal method heavily relies on the Atom Simulation Environment neighbour (ASE) list, in which bonded atoms in a system correspond to atoms whose bond lengths are smaller or equal to the sum of their covalent radii plus an empirical skin value of 0.3. [8, 9] Our approach begins by creating a graph of the system that correspond to a Python dictionary in which each key correspond to an atom and the value correspond to a list of neighbouring atoms. From this graph, we compute a list of connected components using a Depth-first search graph algorithm. This list of connected components correspond to a list of all unconnected entities in the system. The entity that is periodic is then selected as the main system and considered every other entities as unbound guest. We define periodic here as an entity whose cluster expansion still lead to a single entity in the list of connected components. Our approach is very robust because it can effectively be used to remove guest molecules from all common porous systems such as Zeolites, Covalent-organic frameworks (COFs) and MOFs.

### S-1.3-3 Systems with overlapping atoms

Errors occurring from systems containing overlapping atoms are tougher to fix. We decided to perform this by manually editing structures using the CSD Mercury graphical user interface. [10] We found that in two months, we were able to successfully edit only 600 MOFs, which was not a cost-effective strategy. For this reason, we decided to filter out systems containing overlapping atoms by performing two basic chemistry assessments as illustrated in Fig. S3.

We started with a valence test, which ensured that no atom in the system should have more connected neighbours than the maximum possible valency. After the valence test, we proceeded with an inter-atomic test, which ensured that the distance between every pair of atoms should be greater than 0.9 Å except if one of the atom is a hydrogen atom. Using this filter on MOFs systems that had previously been curated for missing hydrogen and unbound guest molecules, we were able to compile a new curated database containing 47,800 MOFs.

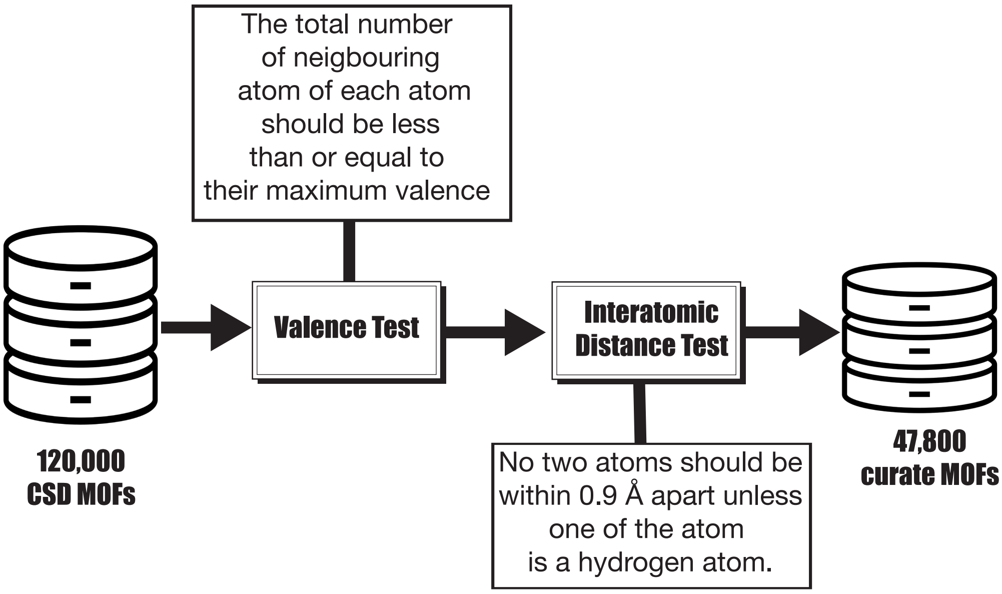

The total number of neigbouring atom of each atom should be less than or equal to their maximum valence
Valence Test
Interatomic Distance Test
No two atoms should be within 0.9 Å apart unless one of the atom is a hydrogen atom.
120,000 CSD MOFs
47,800 curate MOFs

**Fig. S3** Workflow to filter out systems containing errors due to multiple atoms occupying the same symmetrical position in space.The workflow uses two fundamental chemistry assessments. A valence test and an inter-atomic distance test.

### S-1.4 Some errors in the IUPAC names of MOFs in the CSD

While extracting organic ligands from the IUPAC names found in the CSD, we identify a couple of errors with the names. Some of these errors include typos, which can render name extraction difficult. Some of the common typos on the name of the metals can be seen in table Table S1.

**Table S1** Common errors on IUPAC names of MOFs found in the CSD

| **Typo** | **Correct Name** | **Typo** | **Correct Name** |
|---|---|---|---|
| Sodmium | Sodium | Potaasium | Potassium |
| Stontium | Strontium | Siliver | Silver |
| Bairium | Barium | Zirconim | Zirconium |
| abrium | Aabrium | Bismiuth | Bismuth |
| Stannane | Stannane | Zirconocene | Zirconocene |
| Zirconiun | Zirconium | Stannate | Stannate |
| Zirconinum | Zirconium | Ytterbum | Ytterbium |
| Neodmium | Neodymium | Zirconiuum | Zirconium |
| Siler | Silver | Lathanum | Lanthanum |
| Cadmiu | Cadmium | Dypsrosium | Dysprosium |
| Laed | Lead | Stontium | Strontium |
| Germanium | Germanium | Magneium | Magnesium |
| Terbium | Terbium | Praseodimium | Praseodymium |

## S-2 Deconstruction of MOFs

To obtain the unique building units, topology and prospective machine learning descriptors, we implemented a robust MOF deconstruction module called **mof-structure**. The module uses two main procedures to deconstruct MOFs into their building units as illustrated in Fig. S4.

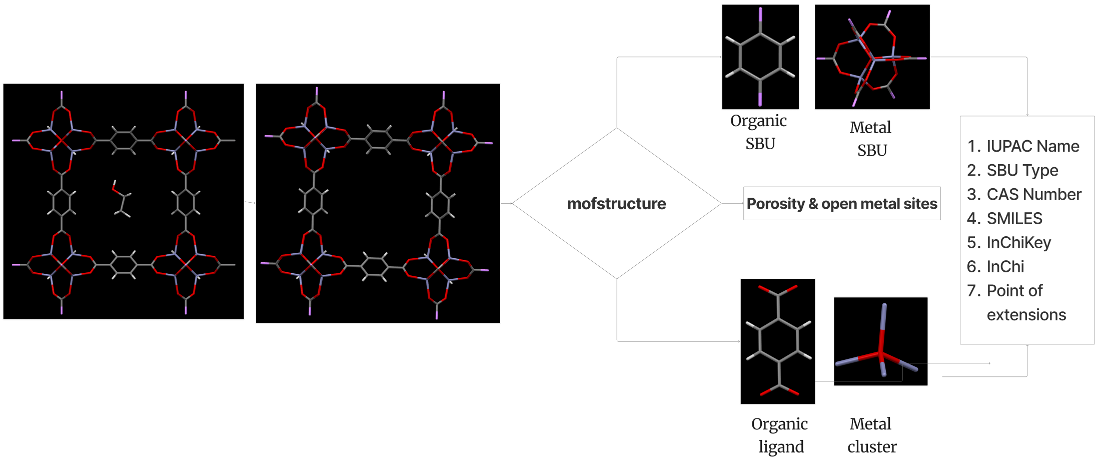
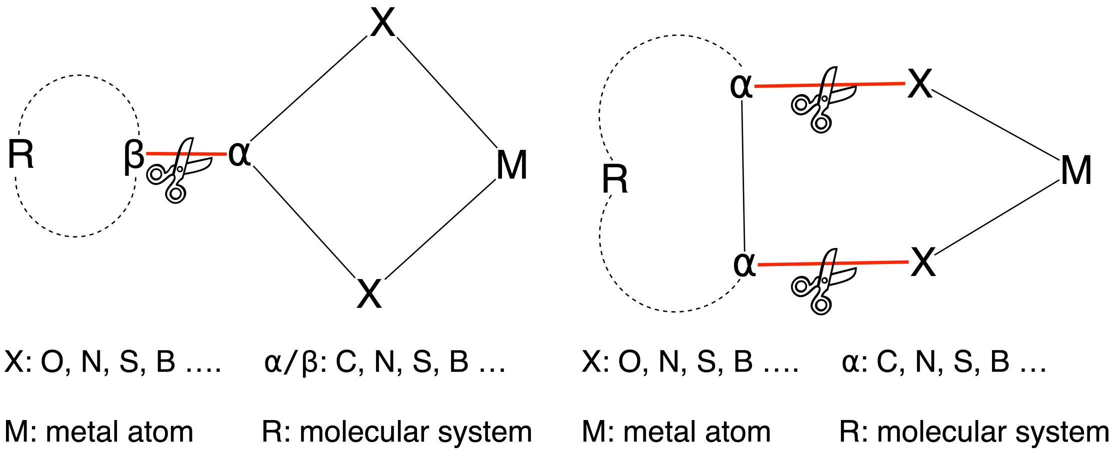


**Fig. S4** Illustration of the two MOF deconstruction procedures. The first procedure deconstruct MOFs into the metal and organic secondary building units, which encodes the topological information of the MOF. The second procedure deconstruct MOFs into metal clusters/ions and organic ligands, which encodes the chemical information of the MOF. Both procedures provide full atom mappings of each unique building units found in the MOF.

### S-2.1 Metal and linker secondary building unit

The metal and linker secondary building units corresponds to lego-like building units that can be used to represent MOFs as graphs or for building the structure of MOFs using packages such as AuToGraFS. [11] To effectively deconstruct MOFs into SBUs, we implemented a rule that recognises carboxylates, sulphates, phosphates, phosphides, sulfides and many other functional groups. Once each of these functional groups are identified in a MOF, we searched for all the atoms within these functional groups that are connected to the metal and break the bond between the $\alpha$ and the $\beta$-atoms related to the aforementioned atoms as illustrated in Fig. S5a. When these functional groups are not identified, we use other rigorous rules to search for breaking points that ensures the most efficient deconstruction. For instance, if an atom connected to the metal is a nitrogen atom as illustrated in Fig. S5b, we break the nitrogen-metal bond but also ensure that this nitrogen atom and the metal are not part of a porphyrin internal ring.



**(a)** cutting at

**(b)** largest cavity diameter

**Fig. S5** Illustration of bonds that are broken during the MOF deconstruction process. The red bond indicates the bonds, which are broken. X represents all atoms that are connected to the metal. M represents metal atoms that are not part of a porphyrin or ferrocene group. $\alpha$ represents atoms that are immediately connected to the atom that is bonded to the metal while $\beta$ represents the atom that bonded to the identified functional groups. R presents the entirely molecular systems

The deconstruction is done in such a rigorous manner that we record the atom indices of each unique SBU, meaning that we know all the instances in which a particular building unit is presence in the MOF. Moreover, we record all the breaking points also known as point of extensions, which provide information about the coordination number of the SBU. Finally, we also implemented rules to identify topological classes of the metal SBU, such as rods, paddlewheels, IRMOFs etc.

### S-2.2 Metal clusters and organic ligands

The metal clusters and organic ligands are useful because they provide a more complete chemically sensible picture of MOFs, since the organic ligands are IUPAC recognised chemicals. Deconstructing MOFs into metal clusters and organic ligands is easier. Here, we begin by finding organic ligands and then break at the atom connected to the metal. We do this such that oxo oxygens, which often form the metal clusters are not affected. However, when clusters are not found a list of unique metals are returned.

To facilitate further computational analysis, we implemented a wrapper function that unwraps systems around lattice coordinates, making them easily convertible into XYZ format. Additionally, cheminformatic descriptors such as SMILES strings, InChI, and InChIKey are computed for each building unit. We also connected it to the PubChem API to extract the IUPAC name of each unique ligand within a MOF. The deconstruction code is freely available and can be accessed from **https://github.com/bafgreat/mofstructure.git**. The documentation of the code is also available **https://bafgreat.github.io/mofstructure**.

## S-2.3 Classification of metal secondary building units

**mofstructure** was observed to be very robust and is able to successfully deconstruct MOFs with rodlike topologies like MOF-74, which most existing packages failed to deconstruct. Moreover, our code is capable of identifying common metal secondary building units (SBUs) such as rod-like SBUs, paddlewheels, the IRMOF series, and ferrocenes as illustrated in Fig. S6.

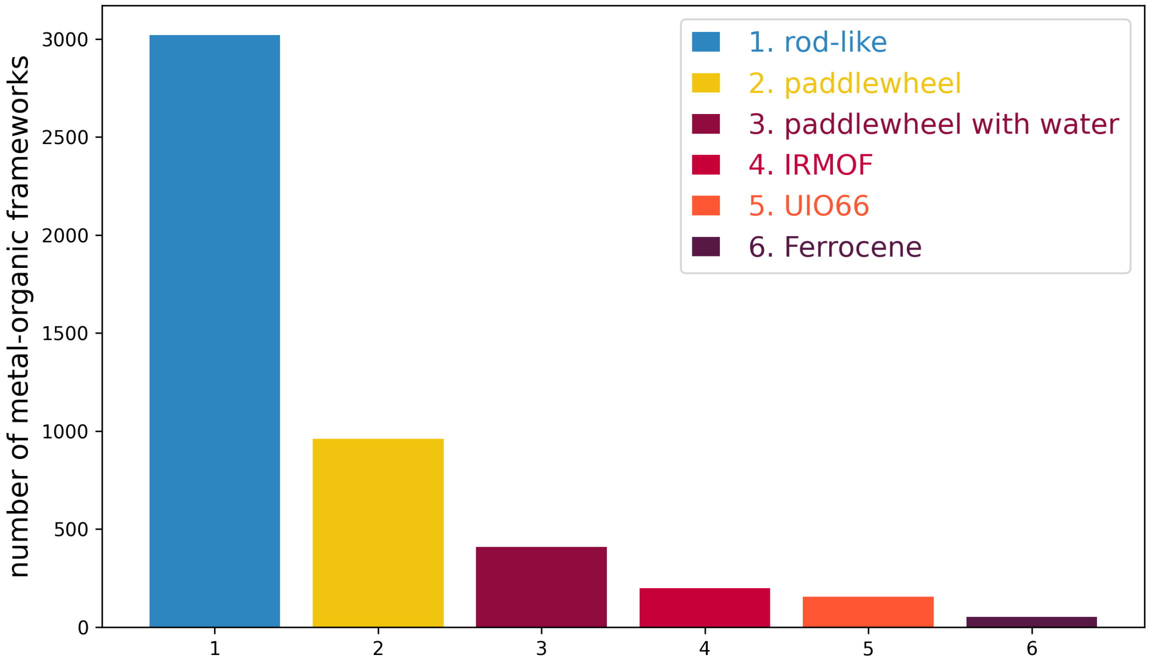


**Fig. S6** Illustration of the distribution of the six most common metal SBUs that are currently identified by our MOF deconstruction method. The color code and the number illustrates the number of MOFs containing these SBUs in our database.

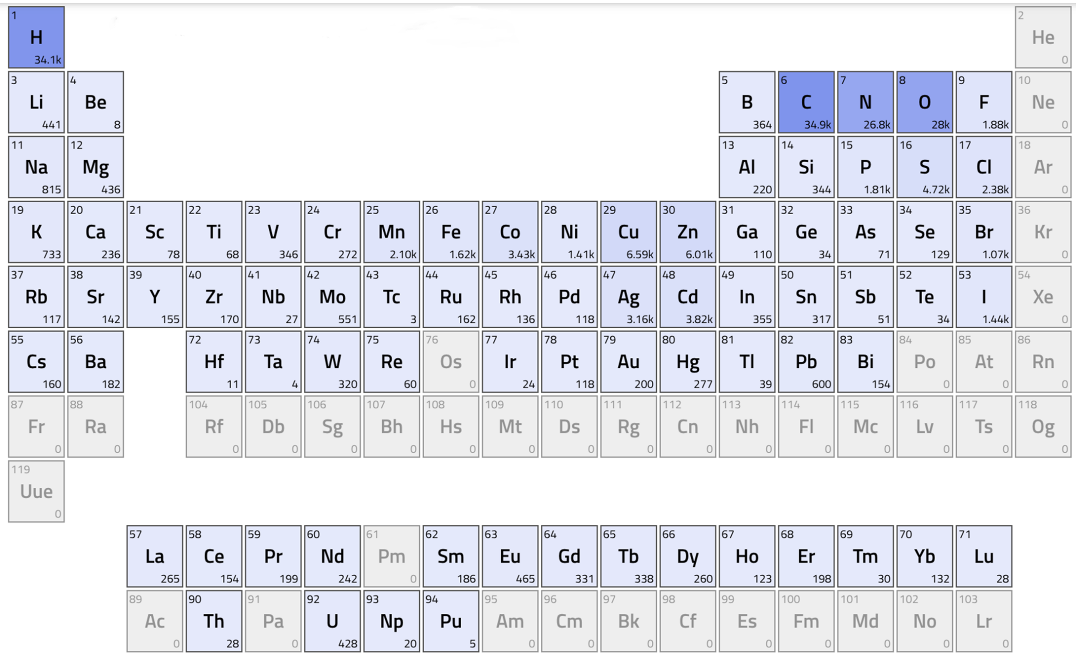
As observed in Fig. S6, our code identified more than 3,000 MOFs that contain rod-like SBUs, which are SBUs that can expand infinitely in one dimension. Anhydrous and hydrated paddlewheel SBUs (which contain either one or two water molecules at the pillar side) were the second and third most common SBUs. MOFs with paddlewheel SBUs were found in more than 1,500 MOFs. Additionally, we identified 198 IRMOF series structures whose metal SBUs are similar to that of MOF-5, as illustrated in Fig. S4. These are commonly known to form the isoreticular series (IRMOF series). Finally, we were also able to classify 156 MOFs containing the Zr-cluster, similar to UIO-66 and 50 ferrocene-containing MOFs.

We are constantly improving **mofstructure**, therefore, if a reader is interested in compiling a dataset for any particular building unit that we have not yet identified, please feel free to contact us. We will implement an identification scheme and provide you with all MOFs containing the specified SBU.

## S-2.4 Integration of Data in NOMAD

As part of our mission to establish a FAIR database for MOFs, we integrated the tools implemented in this study into NOMAD. We called it the Porosity Normaliser, which automatically identifies porous systems uploaded in NOMAD, removes unbound guest molecules, computes their porosity, classify the systems as **MOFs**, **COFs**, **zeolites** or **porous systems**. Systems identified as MOFs are then deconstructed into their building units while mapping the instance of each building units to their parent MOF.

We then uploaded the input and output of all geometry optimised systems into NOMAD, which enabled us to create the second USE CASE in NOMAD, which can directly be searched from this link **https://nomad-lab.eu/prod/v1/gui/search/mofs**. We also plotted a periodic table that illustrates the distribution of the different MOFs based on their elemental composition as shown in Fig. S7.



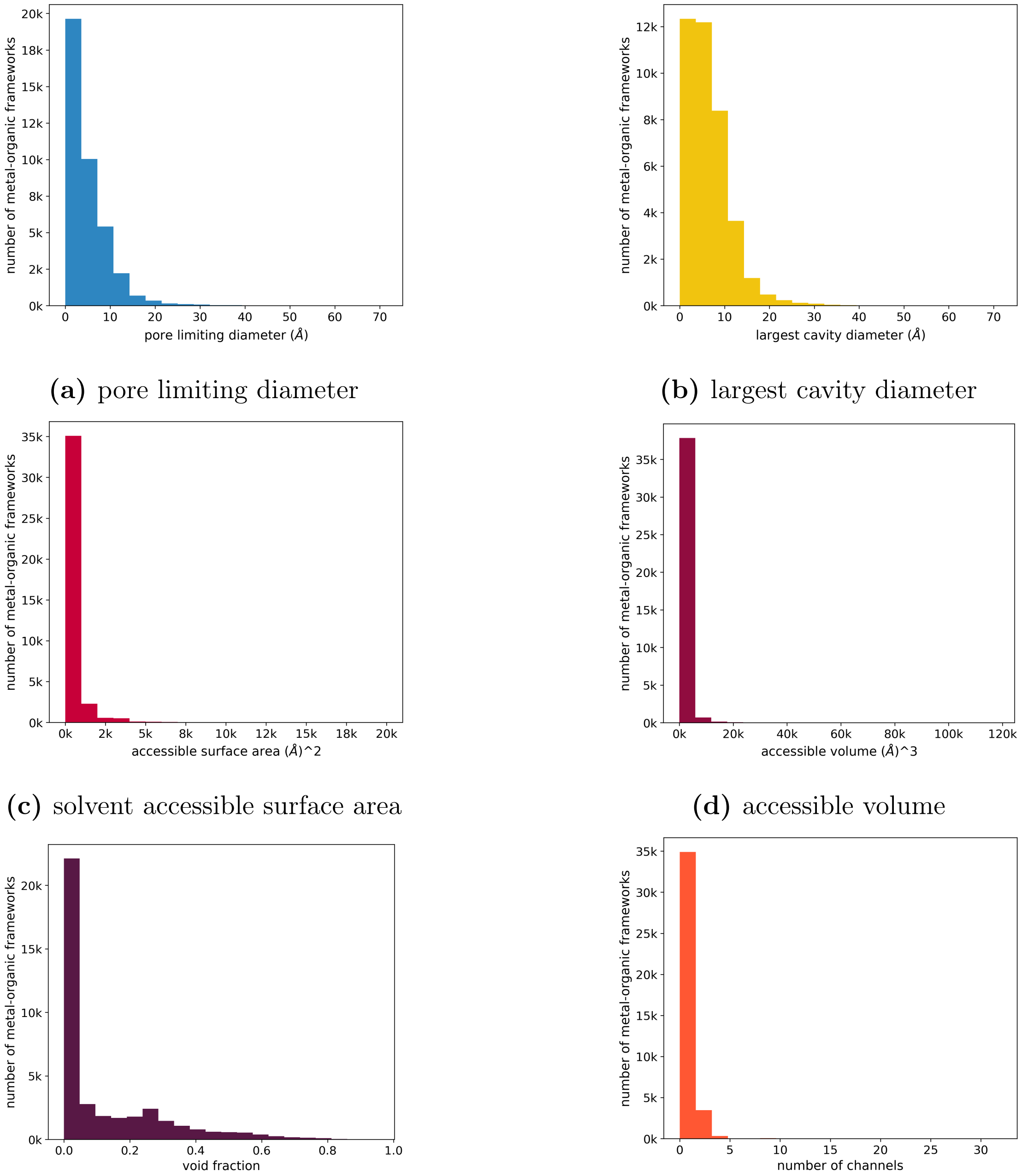
| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1<br>H<br>34.1k | | | | | | | | | | | | | | | | | 2<br>He<br>0 |
| 3<br>Li<br>441 | 4<br>Be<br>8 | | | | | | | | | | | 5<br>B<br>364 | 6<br>C<br>34.9k | 7<br>N<br>26.8k | 8<br>O<br>28k | 9<br>F<br>1.88k | 10<br>Ne<br>0 |
| 11<br>Na<br>815 | 12<br>Mg<br>436 | | | | | | | | | | | 13<br>Al<br>220 | 14<br>Si<br>344 | 15<br>P<br>1.81k | 16<br>S<br>4.72k | 17<br>Cl<br>2.38k | 18<br>Ar<br>0 |
| 19<br>K<br>733 | 20<br>Ca<br>236 | 21<br>Sc<br>78 | 22<br>Ti<br>68 | 23<br>V<br>346 | 24<br>Cr<br>272 | 25<br>Mn<br>2.10k | 26<br>Fe<br>1.62k | 27<br>Co<br>3.43k | 28<br>Ni<br>1.41k | 29<br>Cu<br>6.59k | 30<br>Zn<br>6.01k | 31<br>Ga<br>110 | 32<br>Ge<br>34 | 33<br>As<br>71 | 34<br>Se<br>129 | 35<br>Br<br>1.07k | 36<br>Kr<br>0 |
| 37<br>Rb<br>117 | 38<br>Sr<br>142 | 39<br>Y<br>155 | 40<br>Zr<br>170 | 41<br>Nb<br>27 | 42<br>Mo<br>551 | 43<br>Tc<br>3 | 44<br>Ru<br>162 | 45<br>Rh<br>136 | 46<br>Pd<br>118 | 47<br>Ag<br>3.16k | 48<br>Cd<br>3.82k | 49<br>In<br>355 | 50<br>Sn<br>317 | 51<br>Sb<br>51 | 52<br>Te<br>34 | 53<br>I<br>1.44k | 54<br>Xe<br>0 |
| 55<br>Cs<br>160 | 56<br>Ba<br>182 | | 72<br>Hf<br>11 | 73<br>Ta<br>4 | 74<br>W<br>320 | 75<br>Re<br>60 | 76<br>Os<br>0 | 77<br>Ir<br>24 | 78<br>Pt<br>118 | 79<br>Au<br>200 | 80<br>Hg<br>277 | 81<br>Tl<br>39 | 82<br>Pb<br>600 | 83<br>Bi<br>154 | 84<br>Po<br>0 | 85<br>At<br>0 | 86<br>Rn<br>0 |
| 87<br>Fr<br>0 | 88<br>Ra<br>0 | | 104<br>Rf<br>0 | 105<br>Db<br>0 | 106<br>Sg<br>0 | 107<br>Bh<br>0 | 108<br>Hs<br>0 | 109<br>Mt<br>0 | 110<br>Ds<br>0 | 111<br>Rg<br>0 | 112<br>Cn<br>0 | 113<br>Nh<br>0 | 114<br>Fl<br>0 | 115<br>Mc<br>0 | 116<br>Lv<br>0 | 117<br>Ts<br>0 | 118<br>Og<br>0 |
| 119<br>Uue<br>0 | | | | | | | | | | | | | | | | | |
| | | 57<br>La<br>265 | 58<br>Ce<br>154 | 59<br>Pr<br>199 | 60<br>Nd<br>242 | 61<br>Pm<br>0 | 62<br>Sm<br>186 | 63<br>Eu<br>465 | 64<br>Gd<br>331 | 65<br>Tb<br>338 | 66<br>Dy<br>260 | 67<br>Ho<br>123 | 68<br>Er<br>198 | 69<br>Tm<br>30 | 70<br>Yb<br>132 | 71<br>Lu<br>28 | |
| | | 89<br>Ac<br>0 | 90<br>Th<br>28 | 91<br>Pa<br>0 | 92<br>U<br>428 | 93<br>Np<br>20 | 94<br>Pu<br>5 | 95<br>Am<br>0 | 96<br>Cm<br>0 | 97<br>Bk<br>0 | 98<br>Cf<br>0 | 99<br>Es<br>0 | 100<br>Fm<br>0 | 101<br>Md<br>0 | 102<br>No<br>0 | 103<br>Lr<br>0 | |

**Fig. S7** Illustration of the different atomic representation of the geometry optimised MOFs found in the MOF database. The elements highlighted in blue corresponds to atoms found in the MOFs, while those highlighted in grey correspond to atoms not present in the MOF database. Each element is represented by its atomic number at the top left and the number of this atom present in the database is found at the bottom right. This database will be continuously updated, consequently always checkout updates in NOMAD.

When navigating through the MOF USE CASE in NOMAD, only 17,000 MOFs will be seen. This is because for the sake of efficiency the Porosity Normaliser filters out systems with more than 1000 atoms in the unit cell and systems with a PLD below 1.86 Å.

### S-2.5 Distribution of geometric properties in FAIR-MOFs

The distribution of geometric properties computed from our python implementation of zeo++ are illustrated in Fig. S9.



**(a)** pore limiting diameter

**(b)** largest cavity diameter

**(c)** solvent accessible surface area

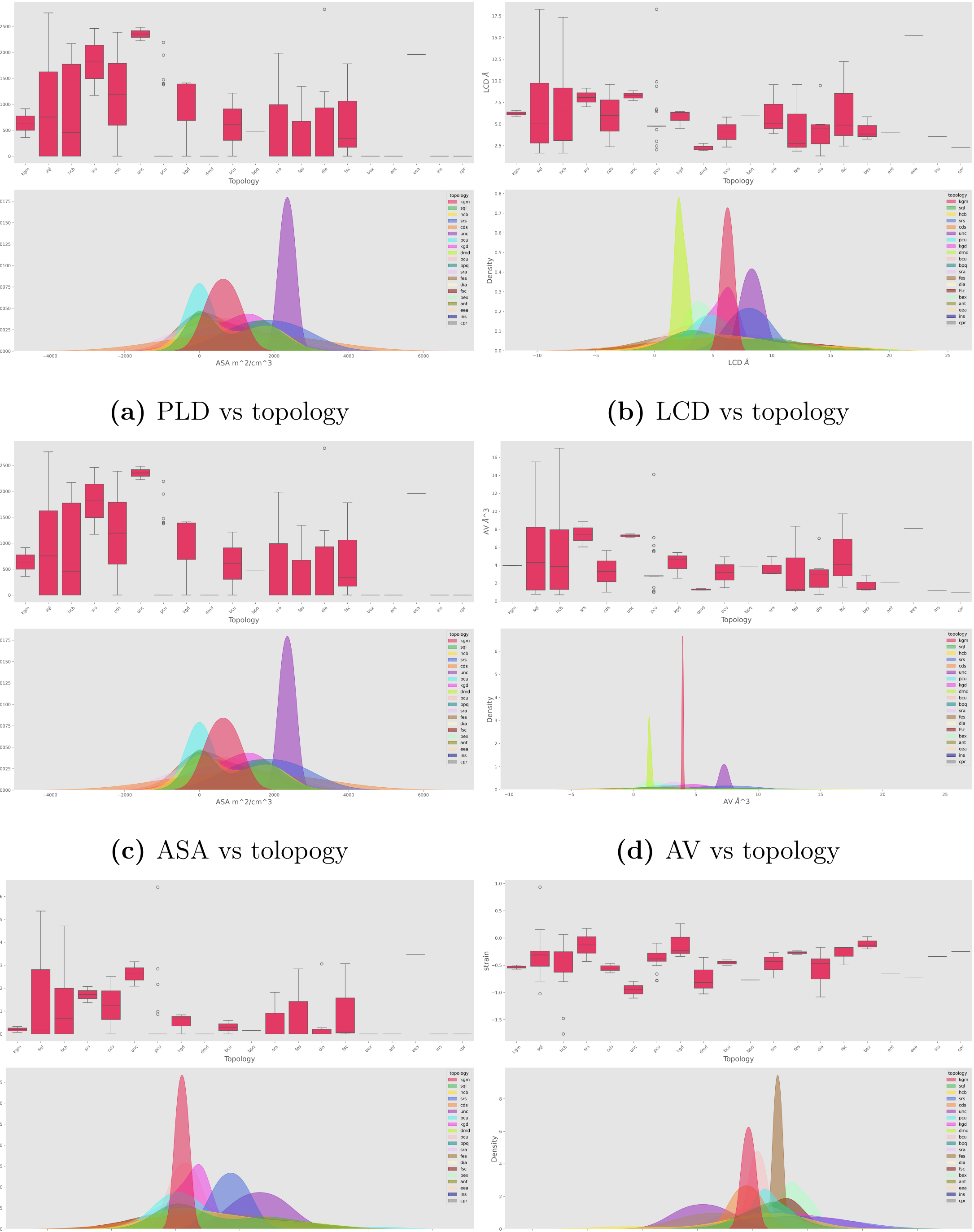
**(d)** accessible volume

**(e)** void fraction

**(f)** number of channels

**Fig. S8** Distribution of geometric properties of MOF. The histograms illustrates the number of MOFs possessing the computed properties from a-f. The **k** preceding each number corresponds to a thousand multiplied by the number.

## S-2.6 Distribution of geometric properties with respect to topologies



**(a)** PLD vs topology

**(b)** LCD vs topology

**(c)** ASA vs tolopogy

**(d)** AV vs topology

**(e)** Void fraction vs topology

**(f)** Strain vs topology

**Fig. S9** Distribution of geometric properties with the 20 most common topologies.

### S-3 Computational Details

The atomic positions in our curated MOF databased were relaxed by performing a geometry optimisation of all the systems at their fixed experimental lattice coordinates. We used the GFN-xTB, which is an efficient approximation of the Density Functional Theory (DFT) and designed to produce reasonable Geometries, Frequencies, and Non-covalent interactions for diverse chemical systems consisting of elements from the periodic table, Z $\leq$ 86. [12] This method was recently shown to be a powerful method for computing different properties of MOFs and was recently recommended as a suitable method for screening of hypothetical MOFs. [13, 14] However, there were 600 MOFs in our curated database that had one or more atoms with Z $\gg$ 86. For these systems, the geometry optimisation was performed using Grimme3 dispersion corrected [15] generalised gradient corrected Perdew, Kieron and Burke exchange-correlation functional PBE-D3, [16] alongside the doubly polarised triple-zeta, TZ2P, basis set. [17] The PBE-D3/TZ2P level of theory was used because it provides a good comprise between computational accuracy and time, [18–20] thus it can consequently be used to compute reasonably correct geometries of large periodic systems like MOFs. All computations were perform using the GFN-xTB and PBE-D3/TZ2P implementation in the Amsterdam Modelling Suits (AMS) package version ADF2019.305. [21]

# S-4 Experimental synthetic conditions

## S-4.1 Synthesis paragraphs

A machine learning model was implemented to predict the paragraphs in text files that describe experimental synthetic procedure. The module was implemented such that when any HTML file or plain text is parsed, it returns a list of synthesis paragraphs identified in the text.

To enable the implementation of this machine learning model, we randomly selected 500 HTML files corresponding to 500 journal articles. We then split the content of these files into list of paragraphs and then literally read through each of these paragraphs. We encoded paragraphs describing synthesis conditions with a value of 1 and 0 for all other paragraphs. This data was then used to perform a sentiment analysis for predicting synthetic paragraphs.

The sentiment analysis was performed using **bag of words** classification. In this approach a huge sparse matrix is created, which stores the counts of each word found in all the 500 journal articles. We used both the **Count Vectorizer** and the **term frequencies inverse document frequency** (TF-IDF) in creating the sparse matrix. We then used a 5 fold Stratified KFold cross-validation for training and validation due to the unbalanced nature of the data given that only 3.5 % of the 44,232 paragraphs corresponded to synthetic paragraphs. Finally we implemented six ML models, Logistic regression, Multinomial naive bayes, Support vector machine, Decision tree, Random forest and Neural network, to perform the paragraph classification. Each of these models were combined with either the Count Vectorizer or TF-IDF from which the best model was use for predicting synthetic paragraphs.

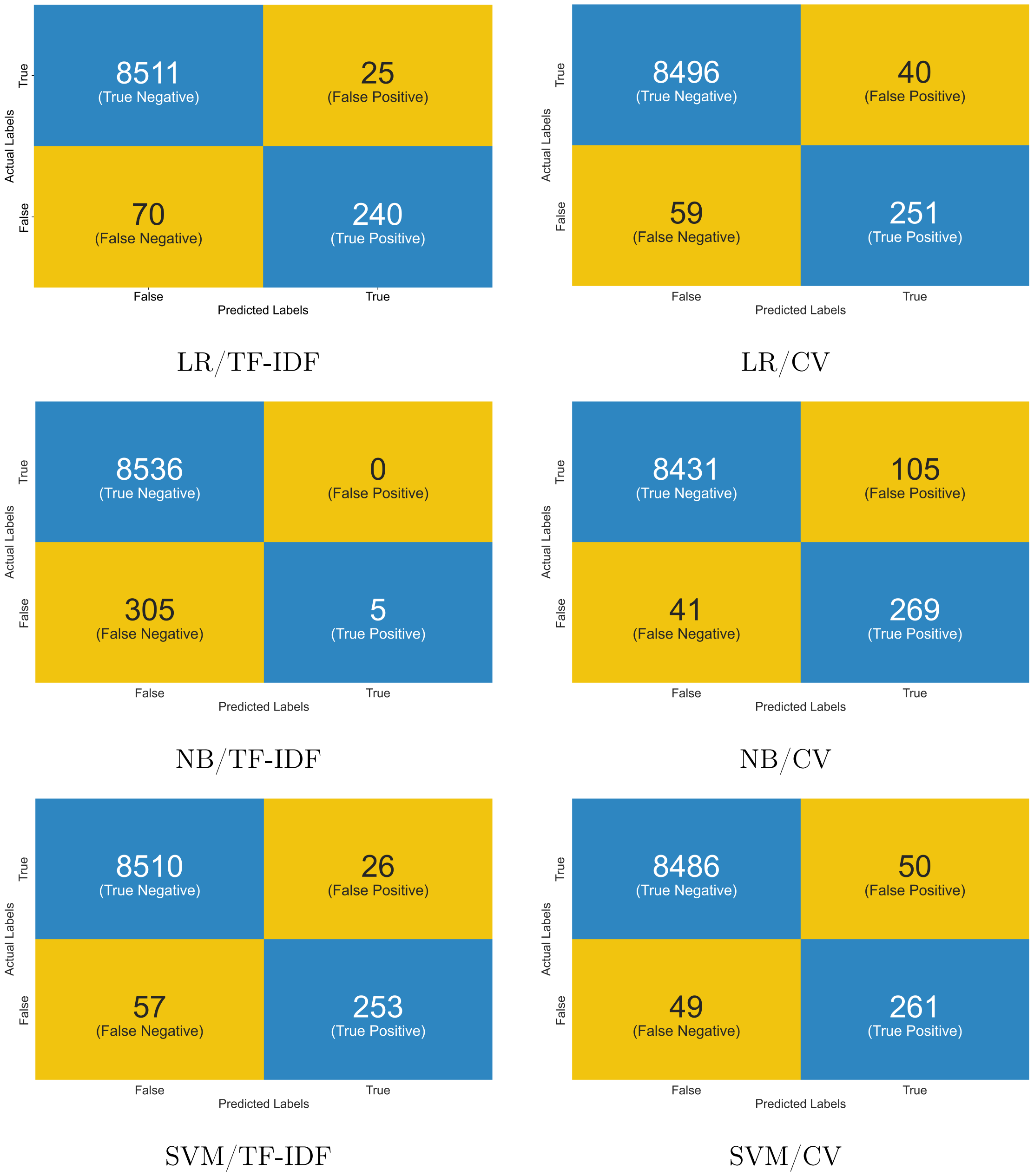
### S-4.2 Evaluation of model for predicting synthesis paragraphs

**Table S2 Performance of supervised machine-learning models for synthesis paragraph classification.** Comparison of six classifiers trained using TF-IDF or count-vector (CV) encodings. Accuracy, ROC-AUC, F-score, recall, and precision are reported for each model. The neural network (NN) trained with TF-IDF achieved the highest overall performance.

| Model | Vectorizer | Accuracy | ROC-AUC | F-score | Recall | Precision |
|---|---|---|---|---|---|---|
| LR | TF-IDF | 0.990 | 0.892 | 0.842 | **1.00** | 0.788 |
| LR | CV | 0.989 | 0.902 | 0.839 | **1.00** | 0.808 |
| NB | TF-IDF | 0.966 | 0.511 | 0.043 | **1.00** | 0.022 |
| NB | CV | 0.983 | 0.912 | 0.779 | **1.00** | 0.836 |
| SVM | TF-IDF | 0.990 | 0.906 | 0.855 | **1.00** | 0.814 |
| SVM | CV | 0.988 | 0.906 | 0.832 | **1.00** | 0.818 |
| DT | TF-IDF | 0.985 | 0.883 | 0.787 | **1.00** | 0.773 |
| DT | CV | 0.986 | 0.886 | 0.794 | **1.00** | 0.779 |
| RF | TF-IDF | 0.990 | 0.882 | 0.845 | **1.00** | 0.766 |
| RF | CV | 0.990 | 0.886 | 0.851 | **1.00** | 0.773 |
| **NN** | **TF-IDF** | **0.991** | **0.912** | **0.863** | **1.00** | **0.830** |
| NN | CV | 0.989 | 0.907 | 0.844 | **1.00** | 0.819 |

## S-4.3 Synthetic condition

We successfully text mined 30,711 set of unique synthetic conditions where each set of conditions was composed of the metal salts, organic reagent, solvent, the concentration of each chemical used, reaction time, crystallisation time, stability temperature, crystallisation temperature, synthesis method, stability time, MOF



LR/TF-IDF

LR/CV

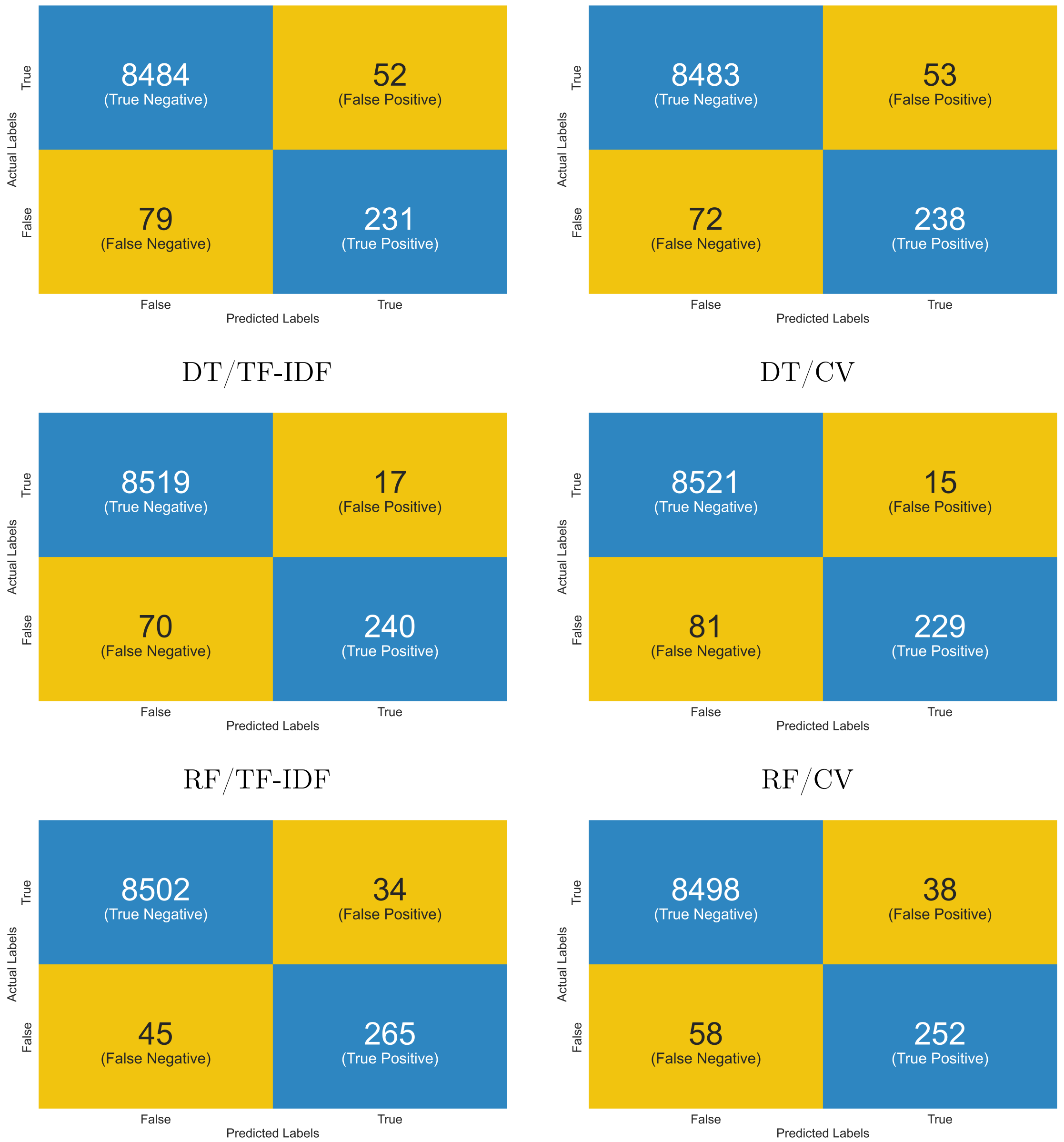
NB/TF-IDF

NB/CV

SVM/TF-IDF

SVM/CV

**Fig. S10** Confusion matrices for classification of synthesis paragraphs using six machine-learning algorithms with two vectorization schemes (TF-IDF and CV). Panels show the classification performance of logistic regression (LR), naive Bayes (NB), support vector machine (SVM), decision tree (DT), random forest (RF), and a neural network (NN), using TF-IDF or CV encodings. All models demonstrate relatively good recall for identifying synthesis paragraphs.



DT/TF-IDF

DT/CV

RF/TF-IDF

RF/CV

NN/TF-IDF

NN/CV

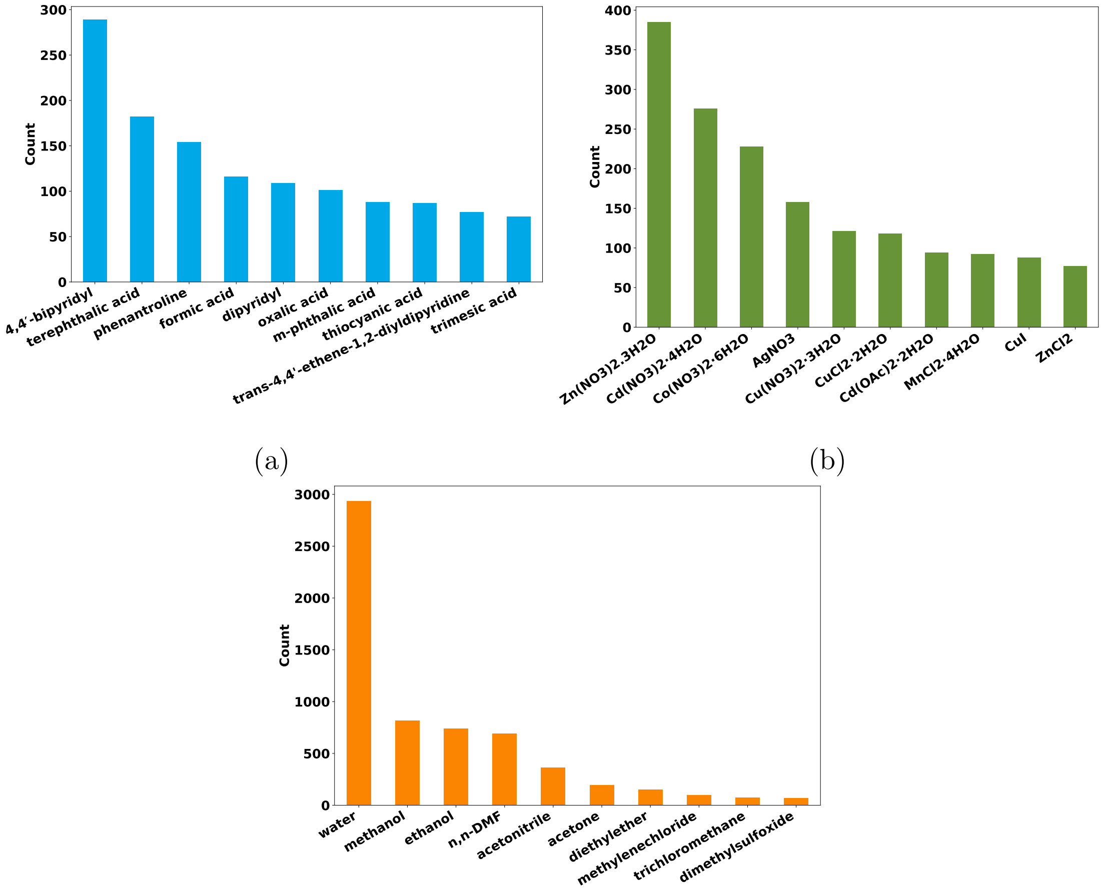
**Fig. S10 Confusion matrices (continued).**

name and warnings if reported. We then successfully mapped the unique set of experimental synthetic conditions of 4,161 crystal structures.

To achieve this task, we began by identifying articles that reported a single MOF, whose crystal structure was uploaded in the CSD. We then checked that the organic ligands matched the linker in the MOF structure and that the metal salt corresponded to the central metal ions. The most tricky process was to match structures in journals that reported several MOFs and especially those where a single metal salt was used with varying organic ligands and vice versa. To achieve this, the previous process was performed with also inspecting their IUPAC to check for solvents used. We also checked if a name was reported for the MOF in the paragraph and whether this name matched the Alias of the MOF reported in the CSD. It is also important to note that not every MOF in the CSD have an Alias, which is the common naming convention generally known for the MOF like MOF-5. On the other hand, it was impossible to blindly match structures where the same ligands and metal salts were used with variation of concentrations or temperature.

Finally, it is worth noting that our main assumption was based each experimental paragraph described a separate synthesis. Therefore each set of synthetic conditions represent the conditions extracted from a given paragraph.

Distributions of top-5 organic linkers, metal salts and solvents are illustrated in Fig. S11.



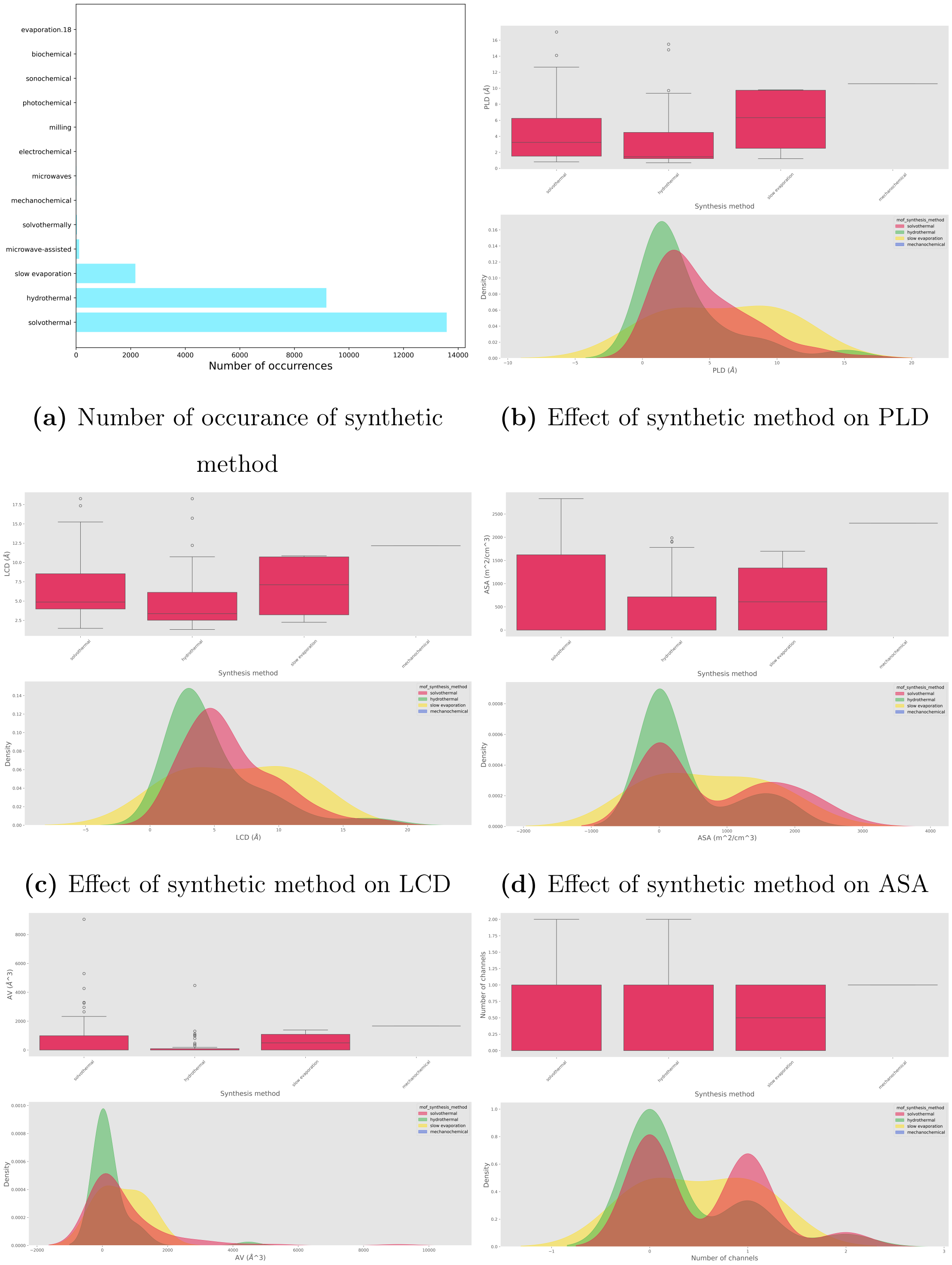
(a)

(b)

(c)

**Fig. S11** Illustration of the distribution of the top 10 most frequently occurring organic ligands (a), metal salts (b) and solvents (c). To ensure consistent representation, all metal salts were converted to their chemical formulae, and ligands and solvents to InChIKeys. This approach was necessary because chemical names are not canonical, meaning a single compound may be referred to by multiple names. In contrast, each compound has a unique InChIKey.

### S-4.4 Effect of synthetic method on geometric properties



**(a)** Number of occurance of synthetic method

**(b)** Effect of synthetic method on PLD

**(c)** Effect of synthetic method on LCD

**(d)** Effect of synthetic method on ASA

**(e)** Effect of synthetic method on AV

**(f)** Effect of synthetic method on number of channels

**Fig. S12** Effect of different synthetic methods on various goemetric properties.

### S-4.5 Metal salts

The most challenging aspect in analysing metal salts results from the varied and non-standardised ways in which different authors choose to refer to the same metal salts. For instance, certain authors refer to **CuCl** as either Copper (I) chloride, Cuprous chloride, Copper (i) chloride, or Copper [i] chloride. While all these names refer to the same compound, they introduce variability that complicates feature extraction for data analysis and for implementing machine learning models for prediction of synthetic conditions. For this reason, we implement a module that maps various names of metal salts to their chemical formulas. This enables a more standardised naming convention. Moreover, we would like to encourage authors to use the chemical formulas of metal salts when writing the synthetic sections of their papers.

Another challenging aspect in analysing metal salts result from the way in which some authors choose to report hydration. Many authors use vague statements like we used hydrated copper sulphate. It is important to note that accounting for the number of hydrated water molecules is important to ensure reproducibility.

Nonetheless, we were able to successfully text mine 35,875 metal salts corresponding to 793 standardised unique salts. A plot of the 5 most common salts used for the 6 most common metals is shown in Fig. S13.

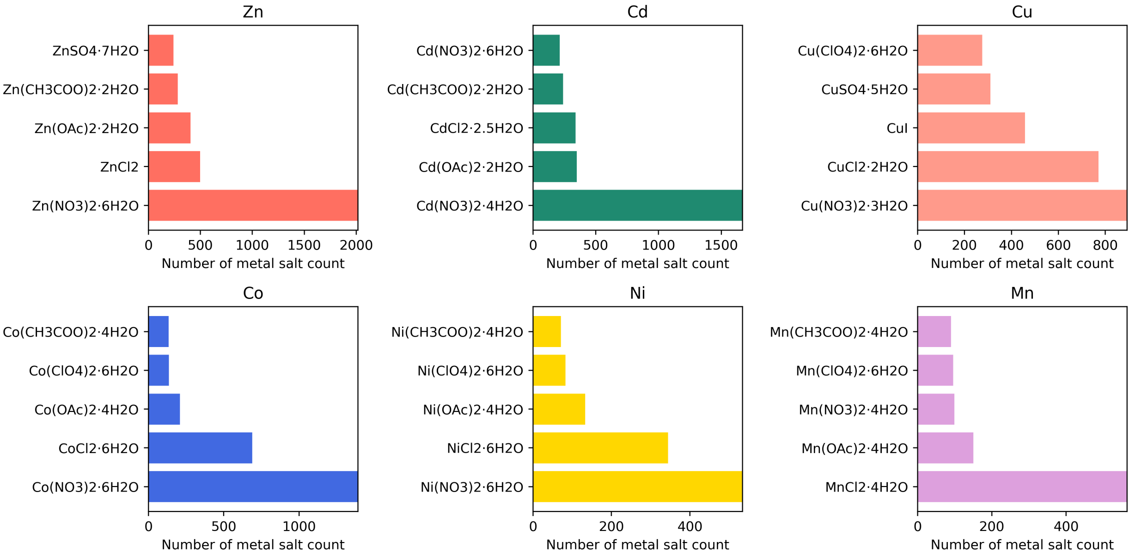
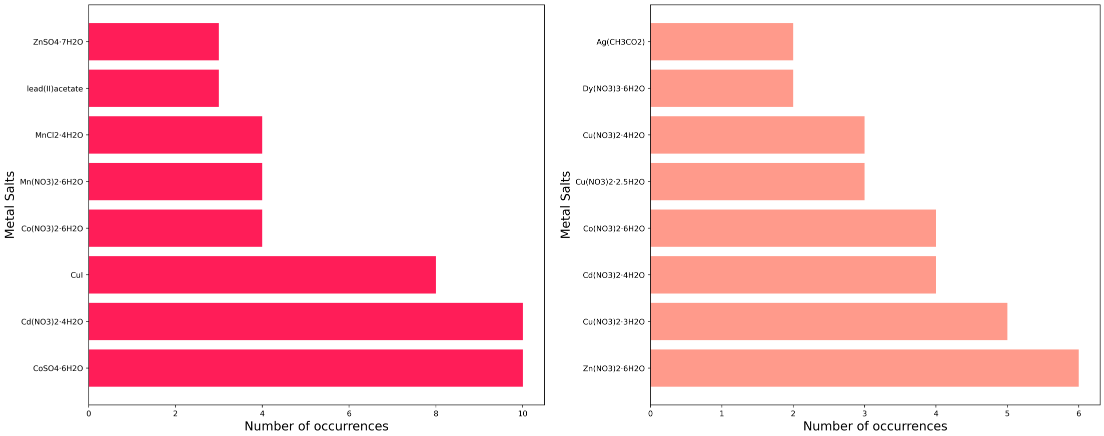


**Fig. S13** Illustration of the 5 most used metal salts for the synthesis of MOFs. The plot represents the salts used in the 6 most common metals, which are all represented in different colours.

It can be observed from Fig. S13 that most MOFs are commonly made from zinc, cadmium, copper, cobalt, nickel and manganese salts. Moreover, it can be observed that hydrated nitrates are the most common salts, which shows a direct relationship between reticular chemistry and coordination chemistry wherein hydrated nitrates provides a source of metal ions that can coordinate with organic linkers to form stable frameworks with the water molecules sometimes participating to enhance stability and functionality of the resulting of the resulting MOFs. [22]

To further explore the importance of the type of metal salts used in the synthesis of MOFs, we investigated various metal salts used in the synthesis of MOFs with rod-like SBUs and paddlewheels. From the 4,161 synthetic conditions which we correctly mapped to their crystal structures, we compiled 131 synthetic conditions for MOFs with rod-like SBUs and 52 conditions for MOFs with paddlewheel

SBUs. We then plotted the eight most common types of metal salts used in the synthesis of these MOFs as illustrated in Fig. S14. It can be observed from Fig. S14 that most MOFs with rod-like SBUs are made from hydrated sulphates and nitrates of Co, Cd, and Mn. Meanwhile, majority of MOFs with paddlewheel SBU were mostly made from different hydrated metal nitrates, which further highlights the importance of hydrated nitrates in MOF chemistry. Consequently, we believe that a future experimental study should be carried out to investigate how different hydrates of a given metal salt affects the crystal structure, morphology and properties of the resulting MOF. Such a study would provide an invaluable insight in understanding one of the many multidimensional experimental factors that affects synthetic outcomes.



**(a)** Synthesis of rod MOFs

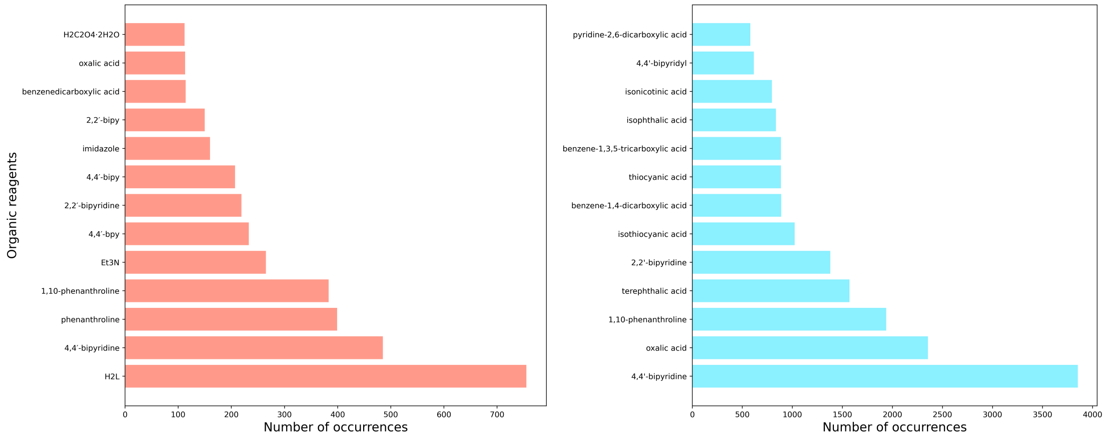
**(b)** Synthesis of MOF with paddlewheel metal SBUs

**Fig. S14** Representation of the 8 most commonly used metal salts used in the synthesis of MOFs with rodlike SBUs (a) and MOFs with a paddlewheel SBUs (b)

### S-4.6 Organic reagents

We text-mined 22,805 organic ligands, corresponding to 4,398 unique organic ligands, from their respective journal articles. However, extracting organic reagents proved to be more challenging compared to other reagents, primarily due to the ambiguous manner in which most scientists have become accustomed to writing.

For instance, in many articles, the chemical names of organic reagents are not explicitly written. Instead, the reagents are drawn and given a label L, which is later used in the chemical formula of the MOF and referred to as $H_2L$ throughout the text. In other cases, there are several inconsistencies in abbreviations. For example, 4,4'-bipyridine is often inconsistently abbreviated as 4,4'-bpy or 4',4'-bipy. On the other hand, some authors prefer to use empirical formulas while others prefer chemical names. A typical example of this inconsistency is oxalic acid, which is often written as $H_2C_2O_42H_2O$. An illustration of this ambiguity is represented in Fig. S15a, where we plotted the 13 most occuring organic reagents extracted from journal articles.



**(a)** Extracted from journal articles

**(b)** Extracted from IUPAC names

**Fig. S15** Illustration of top 13 most commonly used organic ligands used in the synthesis of MOFs. (a) represents the organic reagents that were extracted directly from journal articles and (b) represents journals that were derived from IUPAC names of the MOFs.

These ambiguities make it impossible to utilise this data for any future data analysis or machine learning implementation. Therefore to circumvent this problem, we decided to seek an alternative approach for obtaining a more standardised representation of organic reagents. While studying the IUPAC names of all the MOFs uploaded in the CSD, we observed that deriving organic ligands directly from the IUPAC names would be more effective since the names of the organic ligands are often preserved in the IUPAC naming convention of coordination compounds except with only minor changes in their suffixes.

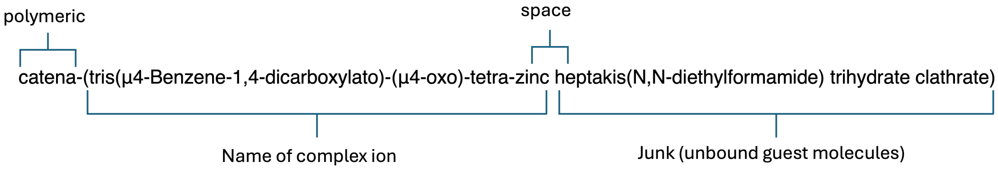

polymeric
space
catena-(tris(μ4-Benzene-1,4-dicarboxylato)-(μ4-oxo)-tetra-zinc heptakis(N,N-diethylformamide) trihydrate clathrate)
Name of complex ion
Junk (unbound guest molecules)

**Fig. S16** Structure of the IUPAC name of a metal organic framework extracted from the CSD

From the IUPAC naming convention for MOFs, as illustrated in the example in Fig. S16, the names of MOFs often begin with the prefix catena, implying a polymeric system. This is often followed by the name of the complex ion or coordination entity, which typically ends with the name of the central metal ions. After the space, everything that follows is not directly bound to the coordination entity. Hence, the name of the ligand can be directly extracted from the coordination entity because in coordination chemistry, the name of the ligand often precedes the central metal ion.

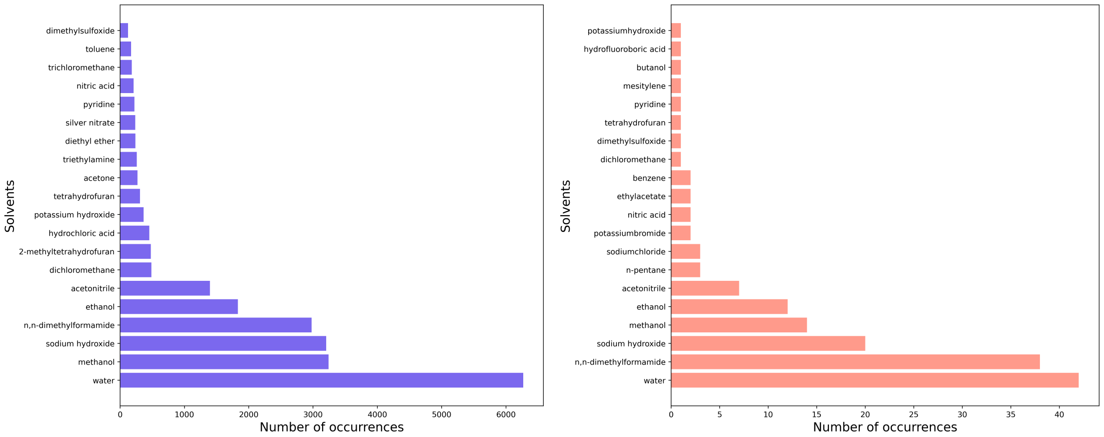
In the example in Fig. S16, the organic ligand is benzene-1,4-dicarboxylato, which corresponds to benzene-1,4-dicarboxylate derived from benzene-1,4-dicarboxylic acid. Consequently, we implemented a Python code that reads every IUPAC name and returns the names of the exact organic ligands. We were able to derive the exact organic ligands for all the hundreds of thousands of MOFs found in the CSD. We plotted the 13 most occurring organic ligands derived from IUPAC names as represented in Fig. S15b. It can be observed from Fig. S15b that the names of the reagents are more standardised. Consequently, we also computed the InChIKeys for all these organic reagents using OPSIN and PubChem Python APIs to avoid ambiguities that often results from using common names and IUPAC names. [23–26]

However during this process, we also identified several errors in the IUPAC names of MOFs uploaded in the CSD. A compiled list of some of these errors are found section **S-1.4**.

Unfortunately, this approach also proved unrealistic because numerous minor typographical inconsistencies in the reported IUPAC names prevented reliable conversion to SMILES strings for name standardisation. Consequently, we adopted an alternative strategy using our MOF deconstruction algorithm, mofstructure, to extract chemical names directly from the crystal structure. In this workflow, we isolate the molecular fragment, compute its InChIKey and SMILES strings, and then query the PubChem API to retrieve the corresponding IUPAC name. This method proved both more accurate and substantially faster than text-mining,Äìbased approaches.

### S-4.7 Solvents

Text mining solvents was one of the most straightforward approach. We did this by compiling a list of all possible solvents that are commonly used in synthetic chemistry as well as their chemical formulas. We then implemented regex patterns to match these solvents and their quantities. Finally, we mapped all abbreviated solvents to their chemical names. The list of the 20 most occurring solvents are illustrated in Fig. S17.



**(a)** All extracted solvents

**(b)** Rod MOF synthesis

**Fig. S17** Illustration of top 20 most commonly used solvents used in the synthesis of MOFs. (a) Represents the 20 most occurring solvents in our database and (b) represents the 20 most occurring solvents used in the synthesis of MOFs with rodlike SBUs.

It can be observed from Fig. S17a that the 7 most occurring solvents used in the synthesis of MOFs with occurrence exceeding 1000 are : water, methanol, sodium hydroxide, n,n-dimethylformamide, ethanol, acetonitrile and dichloromethane respectively. Interestingly, these solvents are all highly polar characterised by high dielectric constants and dipole moments. Their high polarity is necessary to significantly contribute to their effectiveness in dissolving majority of the organic ligands and metal salts that are often used in the synthesis of MOFs.

We were also intrigued to know the solvents that are predominantly used in the synthesis of MOFs with rodlike topologies, since these MOFs are well known for their stability. As illustrated in figure Fig. S17b, it can be observed that the 7 most common solvents in the synthesis of MOFs with rodlike SBUs are: water, n,n-dimethylformamide, sodium hydroxide, methanol, ethanol, acetonitrile and n-pentane. These solvents are also highly polar apart from n-pentane. Notably, all these solvents, except for n-pentane, are highly polar. It is also worth noting that most polar solvents act as ligands themselves by forming a coordination bond with the central metal ion, which could be advantageous or disadvantageous depending on the application.

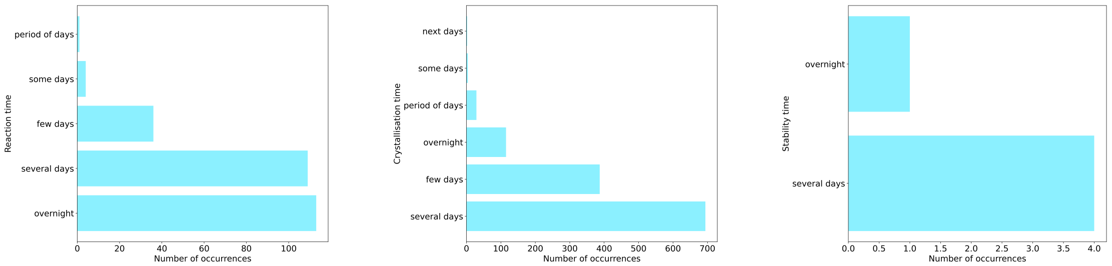
### S-4.8 Synthetic methods

S-4.8 Extracting the different synthetic methods used in the synthesis of MOFs was another straightforward approach. We started by compiling a list of all synthetic methods commonly used in reticular chemistry. Then we implement regex patterns to match this patterns in each synthetic paragraph. However, in most cases the names of the synthetic methods are not explicitly written. In such cases, we checked whether there was any heating mentioned in the paragraph and whether water was present as solvent or not. In paragraphs where heating was mentioned in the presence of water, our module returned the synthetic method as hydrothermal and those where heating was mention in the absence of water but in the presence of other solvents, the method was reported as solvothermal. We also looked out for how the synthesis was performed. For instance synthesis carried out in the presence of a microwave was reported as microwave assisted method.

### S-4.9 Time and ambiguity in reporting time

We implemented a method to find crystallization time, stability time and reaction time from each synthetic paragraph where time was reported. We then converted this time to hours since they are sometimes reported in days, weeks and hours. The reaction time generally include all the time spent performing the reaction including the time of mixing. On the other hand the stability time refers to how long the structure was stable either in solvent or during thermal analysis. And finally, the crystallisation time corresponded to the time after which crystal started forming. These quantities were not always collected from all paragraphs because several authors do not always report each of these time.

However, the most challenging aspect of text mining and processing time resulted from the ambiguous and imprecise way in which most authors report time in the literature. These imprecisions are illustrated in Fig. S18. As shown in Fig. S18a, some authors use terms like overnight or several days to report the duration of their synthesis, while others use several days or a few days to report the time of crystallisation. We strongly encourage authors to be more precise when reporting time, as time is crucial for controlling the synthesis and properties of MOFs.



**(a)** Ambiguity in reporting reaction time

**(b)** Ambiguity in reporting crystallisation time

**(c)** Ambiguity in reporting stability time

**Fig. S18** Illustration of the imprecise manner in which several authors report time in journal articles. Each plot shows the number of occurrence of each of these terms. (a) Represents some of the imprecise ways in which reaction time is reported (b) Represents some of the imprecise ways in which crystallisation time is reported and (c) Represents some of the imprecise ways in which stability of the MOF is reported

To gain a clearer understanding of how different people interpret these imprecise terms, we conducted a survey where we asked participants to assign specific durations to these terms.

### S-4.10 Temperature, reaction quantities and warning

We also text mine temperatures from each synthetic paragraphs. The temperature extracted were the crystallisation temperature, drying temperature, melting temperature, stability temperature and reaction temperature. These temperatures were extracted by checking their units and also looking what the words used in the sentences in which these temperatures were referred. For instance if a temperature is found in a sentence and the word crystallisation precedes or proceeds the value, the temperature will be considered as crystallisation temperature. We also converted the values of all temperatures to Kelvin.

However, authors frequently report temperature simply as room temperature. Although this may appear logically acceptable, it is scientifically imprecise and physically inaccurate, as room temperature varies substantially with weather conditions, season, and geographic location. For instance, in humid coastal cities such as Douala or Limbe in Cameroon, room temperature can range from 30–35 °C during the dry season and 25–30 °C during the rainy season, which will be markedly different from those in cities in Germany or the United Kingdom. This variability highlights the need for explicit temperature reporting to ensure reproducibility in synthesis.

In addition to text mining temperatures, we also text mined the quantities of all the reagents used in the synthesis. We implemented a robust model to find chemicals and extract both their quantities and the units of the quantity used in the synthesis. We then converted all these quantities to their SI units to standardised the values for future data analysis and implementation of machine learning models.

Finally, we text mine all warnings about experimental synthetic conditions when reported in the experimental paragraph. For instance we extracted full sentences like:

*CAUTION*! Though while working with the perchlorate compounds described here we have not met with any incident, care should be taken in handling them as perchlorates are potentially explosive. These should not be prepared and stored in large amounts.

This provides a careful warning about dangerous chemicals as well as provide information on hazards measures.

# S-5 Graph Dataset

To assess the structural diversity and topological complexity of the dataset, we computed several statistics across the 4,161 graph instances. As summarised in Table S3, the number of nodes per graph ranges from 13 to over 5,000, with an average of $232.84 \pm 234.83$, reflecting the broad variability in MOF sizes-from small molecular fragments to large extended frameworks. The number of edges follows a similar trend, with a mean of 273.36±277.42 and a range spanning from 14 to 5,664.

Each node encodes a 4-dimensional feature vector capturing atomic-level descriptors, while the average node degree across all graphs is $2.36 \pm 0.24$, consistent with the sparsity typically observed in molecular and crystalline structures. Notably, node degree values vary between 1.88 and 4.86, suggesting heterogeneous connectivity patterns within and across MOFs.

In addition to size and connectivity, we computed the average shortest path length within the largest connected component of each graph. This metric quantifies the typical topological distance between pairs of atoms and provides insight into the compactness of the graph structure. On average, MOF graphs exhibit an internal path length of $8.74 \pm 3.53$, with a minimum of 2.06 and a maximum exceeding 50, indicating the presence of both compact and highly extended topologies.

**Table S3** Summary statistics of the graph-based MOF representations used for machine learning.

| **Metric** | **Mean ± Std** | **Min** | **Max** |
|---|---|---|---|
| Number of nodes per graph | $232.84 \pm 234.83$ | 13 | 5136 |
| Number of edges per graph | $273.36 \pm 277.42$ | 14 | 5664 |
| Average node degree | $2.36 \pm 0.24$ | 1.88 | 4.86 |
| Node feature dimension | 4 | - | - |
| Average shortest path length | $8.74 \pm 3.53$ | 2.06 | 52.03 |

Fig. S19 complements Table S3 by visualising the distribution of key structural properties across the dataset. The left panel shows the distribution of node counts per graph, highlighting that the majority of MOFs contain fewer than 1,000 atoms, although a small number of large frameworks account for the long tail. The right panel displays the distribution of average shortest path lengths, revealing that most graphs have values between 5 and 15, indicative of moderately compact structures, while a few exhibit much higher values due to sparse or extended topologies.

.

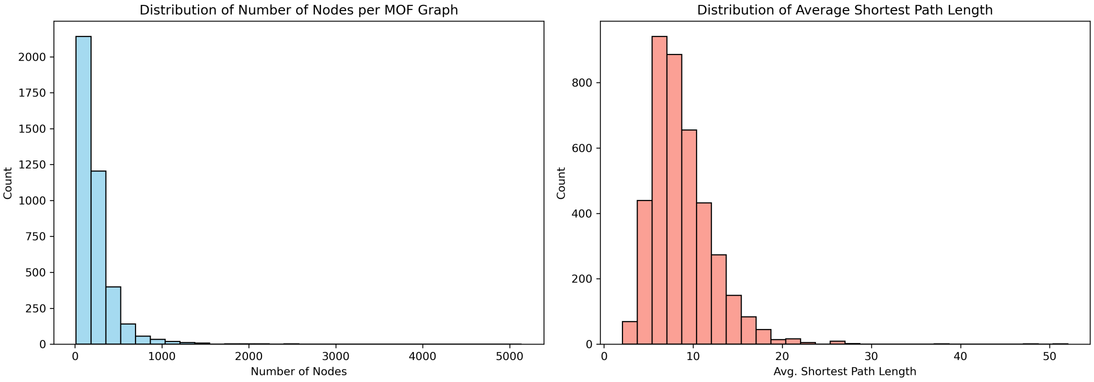


**Fig. S19** Left: Distribution of the number of nodes per MOF graph, illustrating the structural diversity of the dataset. Right: Distribution of average shortest path lengths, reflecting the compactness and topological spread of the MOF structures.

# S-6 Model prediction of synthesis condition

The predicted synthesis parameters obtained from the `fairmofsyncondition` model for both the original `qmof-6031bc0` structure and the brominated analogue synthesised experimentally are presented in and respectively. Each report contains information on thermodynamic stability, crystal symmetry, predicted metal salts, solvents, and organic ligands, along with their associated probabilities. These predictions were used to guide the experimental synthesis described in the manuscript.

## S-6.1 Prediction for `qmof-6031bc0`

```
Predicted Synthetic Data Report
For: qmof-6031bc0
================================================================================
Space group number:        1
Crystal system:            triclinic
```

```
================================================================================
Thermodynamic Stability:  -793.8344422866076 kJ/mol
Organic Ligands
--------------------------------------------------------------------------------
InChIKey                     SMILES                         IUPAC Name
--------------------------------------------------------------------------------
LMOSYFZLPBHEOW-UHFFFAOYSA-N  C1=C(C(=CC(=C1Cl)C(=O)O)Cl)C(=O)O  2,5-dichloroterepht

Top 3 Predicted Metal Salts
--------------------------------------------------------------------------------
Metal Salt                                   Probability
--------------------------------------------------------------------------------
ZrOCl2.8H2O                                       0.4881
Cd(NO3)2.4H2O                                     0.2421
CdCl2.6H2O                                        0.0862
================================================================================
Percentage probability of solvents
--------------------------------------------------------------------------------
Organic ligands
--------------------------------------------------------------------------------
Ligands : 2,5-dichloroterephthalic acid
--------------------------------------------------------------------------------
Solvent                                  Predicted probability
--------------------------------------------------------------------------------
H2O                                                83.33 %
C2H5OH                                             16.67 %
================================================================================
Metal Salts
--------------------------------------------------------------------------------
Metal Salts: ZrOCl2.8H2O
--------------------------------------------------------------------------------
Solvent                                  Predicted probability
--------------------------------------------------------------------------------
AcOH                                                33.33 %
CH3OH                                               33.33 %
DMF                                                 33.33 %
--------------------------------------------------------------------------------
Metal Salts: Cd(NO3)2.4H2O
--------------------------------------------------------------------------------
Solvent                                  Predicted probability
--------------------------------------------------------------------------------
H2O                                                 34.69 %
DMF                                                 14.29 %
CH3OH                                                12.7 %
--------------------------------------------------------------------------------
Metal Salts: CdCl2.6H2O
--------------------------------------------------------------------------------
Solvent                                  Predicted probability
--------------------------------------------------------------------------------
H2O                                                  50.0 %
DMF                                                  25.0 %
C2H5OH                                               12.5 %
================================================================================
Report generated by fairmofsyncondition
```

```
Authors: Dinga Wonanke \& Antonio Longa

================================================================================
```

## S-6.2 Prediction for Brominated Variant

```
Predicted Synthetic Data Report

For: dibromoterephthalate_Zr

================================================================================

Space group number:       1

Crystal system:           triclinic


================================================================================

Thermodynamic Stability:  -895.4167197843163 kJ/mol

Organic Ligands

--------------------------------------------------------------------------------

InChIKey                     SMILES                         IUPAC Name

--------------------------------------------------------------------------------

VUTICWRXMKBOSF-UHFFFAOYSA-N  C1=C(C(=CC(=C1Br)C(=O)O)Br)C(=O)O  2,5-dibromoterephth


Top 3 Predicted Metal Salts

--------------------------------------------------------------------------------

Metal Salt                                   Probability

--------------------------------------------------------------------------------

ZrOCl2.8H2O                                       0.8075

BiCl3                                             0.1151

CaCl2                                             0.0344

================================================================================

Percentage probability of solvents
```

```
--------------------------------------------------------------------------------
Organic ligands
--------------------------------------------------------------------------------
Ligands : 2,5-dibromoterephthalic acid
--------------------------------------------------------------------------------
Solvent                                  Predicted probability
--------------------------------------------------------------------------------
CH3OH                                               50.0 %
H2O                                                 25.0 %
TEA                                                 25.0 %
================================================================================
Metal Salts
--------------------------------------------------------------------------------
Metal Salts: ZrOCl2.8H2O
--------------------------------------------------------------------------------
Solvent                                  Predicted probability
--------------------------------------------------------------------------------
AcOH                                               33.33 %
CH3OH                                              33.33 %
DMF                                                33.33 %
--------------------------------------------------------------------------------
Metal Salts: BiCl3
--------------------------------------------------------------------------------
Solvent                                  Predicted probability
--------------------------------------------------------------------------------
THF                                                42.86 %
H2O                                                28.57 %
```

```
C2H5OH                                             14.29 %
--------------------------------------------------------------------------------
Metal Salts: CaCl2
--------------------------------------------------------------------------------
Solvent                                  Predicted probability
--------------------------------------------------------------------------------
H2O                                                38.46 %
CH3OH                                              23.08 %
C2H5OH                                             15.38 %
================================================================================
Report generated by fairmofsyncondition
Authors: Dinga Wonanke \& Antonio Longa
================================================================================
```

## S-6.3 Prediction for qmof-835565b

```
Predicted Synthetic Data Report
For: qmof-835565b
================================================================================
Space group number:       1
Crystal system:           triclinic


================================================================================
Thermodynamic Stability:  -386.75146094210396 kJ/mol
Organic Ligands
--------------------------------------------------------------------------------
InChIKey                     SMILES                         IUPAC Name
--------------------------------------------------------------------------------
```

```
NEQFBGHQPUXOFH-UHFFFAOYSA-N  C1=CC(=CC=C1C2=CC=C(C=C2)C(=O)O)C(=O)O  4-(4-carboxyp

Top 3 Predicted Metal Salts

--------------------------------------------------------------------------------

Metal Salt                                   Probability

--------------------------------------------------------------------------------

Zn(NO3)2.3H2O                                     0.9545

ZnCO3                                             0.0182

Zn(NO3)2.6H2O                                     0.0086

================================================================================

Percentage probability of solvents

--------------------------------------------------------------------------------

Organic ligands

--------------------------------------------------------------------------------

Ligands : 4-(4-carboxyphenyl)benzoic acid

--------------------------------------------------------------------------------

Solvent                                  Predicted probability

--------------------------------------------------------------------------------

DMF                                                30.43 %

H2O                                                26.09 %

CH3OH                                              10.87 %

================================================================================

Metal Salts

--------------------------------------------------------------------------------

Metal Salts: Zn(NO3)2.3H2O

--------------------------------------------------------------------------------

Solvent                                  Predicted probability
```

```
--------------------------------------------------------------------------------
H2O                                                34.83 %
DMF                                                19.07 %
C2H5OH                                             10.78 %
--------------------------------------------------------------------------------
Metal Salts: ZnCO3
--------------------------------------------------------------------------------
Solvent                                  Predicted probability
--------------------------------------------------------------------------------
C2H5OH                                              50.0 %
H2O                                                42.86 %
KBr                                                 7.14 %
--------------------------------------------------------------------------------
Metal Salts: Zn(NO3)2.6H2O
--------------------------------------------------------------------------------
Solvent                                  Predicted probability
--------------------------------------------------------------------------------
H2O                                                36.11 %
DMF                                                 25.0 %
C2H5OH                                             11.11 %
================================================================================

Report generated by fairmofsyncondition
Authors: Dinga Wonanke & Antonio Longa
================================================================================
```

## S-6.4 Prediction for qmof-3fb24cf

```
Predicted Synthetic Data Report
For: qmof-3fb24cf
================================================================================
Space group number:       1
Crystal system:           triclinic

================================================================================
Thermodynamic Stability:  -317.66544675314094 kJ/mol
Organic Ligands
--------------------------------------------------------------------------------
InChIKey                     SMILES                         IUPAC Name
--------------------------------------------------------------------------------
RXOHFPCZGPKIRD-UHFFFAOYSA-N  C1=CC2=C(C=CC(=C2)C(=O)O)C=C1C(=O)O  naphthalene-2,6-

Top 3 Predicted Metal Salts
--------------------------------------------------------------------------------
Metal Salt                                   Probability
--------------------------------------------------------------------------------
Zn(NO3)2.3H2O                                     0.9302
ZnCl2                                             0.0220
ZnI2                                              0.0161
================================================================================
Percentage probability of solvents
--------------------------------------------------------------------------------
Organic ligands
--------------------------------------------------------------------------------
```

```
Ligands : naphthalene-2,6-dicarboxylic acid

--------------------------------------------------------------------------------

Solvent                                  Predicted probability

--------------------------------------------------------------------------------

H2O                                                34.62 %

DMF                                                21.15 %

KBr                                                13.46 %

================================================================================

Metal Salts

--------------------------------------------------------------------------------

Metal Salts: Zn(NO3)2.3H2O

--------------------------------------------------------------------------------

Solvent                                  Predicted probability

--------------------------------------------------------------------------------

H2O                                                34.83 %

DMF                                                19.07 %

C2H5OH                                             10.78 %

--------------------------------------------------------------------------------

Metal Salts: ZnCl2

--------------------------------------------------------------------------------

Solvent                                  Predicted probability

--------------------------------------------------------------------------------

H2O                                                29.57 %

KBr                                                16.52 %

C2H5OH                                             13.91 %

--------------------------------------------------------------------------------

Metal Salts: ZnI2
```

```
--------------------------------------------------------------------------------
Solvent                                  Predicted probability
--------------------------------------------------------------------------------
H2O                                                63.64 %
CH3OH                                              18.18 %
DMF                                                 9.09 %
================================================================================

Report generated by fairmofsyncondition
Authors: Dinga Wonanke & Antonio Longa
================================================================================
```

## S-6.5 Explanation of Terms

- **Space group number:** Identifies the symmetry group of the predicted crystal structure according to the International Tables for Crystallography.

- **Crystal system:** Describes the lattice symmetry (e.g., triclinic, monoclinic, orthorhombic).

- **Thermodynamic Stability (kJ/mol):** The predicted formation energy per mole, used as a proxy for the relative stability of the structure. Model trained on GFN-xTB formation energies per formular unit of MOFs.

- **Organic Ligands:** Lists the predicted linker molecules with their InChIKey (unique chemical identifier), SMILES notation, and IUPAC name.

- **Top 3 Predicted Metal Salts:** Lists the metal precursors most likely to yield the MOF, ranked by their predicted probability. Each probability represents the likelihood of successfully synthesising the target MOF using *one* specific metal salt under appropriate conditions. These values should not be interpreted as suggesting that multiple metal salts are to be combined in a single synthesis; rather, they indicate alternative viable precursors, each considered independently.

- **Percentage Probability of Solvents:** Lists the most frequently used solvents in the literature for a given reagent, derived from statistical co-usage analysis. The reported percentage reflects how often each solvent has been associated with that reagent across experimental records. These probabilities should not be interpreted as suggesting that the solvents are to be combined in a single synthesis; each solvent is considered independently as a viable option. When no co-usage data are found in the literature, the model defaults to reporting no data found try water as a general fallback recommendation. It is important to note that this solvent prediction is not generated by a trained AI model but rather represents a statistical measure of historical solvent usage frequencies.

- **Report generated by `fairmofsyncondition`:** All predictions were computed using our trained retrosynthetic recommender model based on literature co-usage patterns extracted from thousands of experimental MOF reports.

# References

[1] Buonsanti, R. Electronic lab notebooks for materials synthesis. *Chemistry of Materials* **35**, 805–806 (2023).

[2] Machina, H. K. & Wild, D. J. Electronic laboratory notebooks progress and challenges in implementation. *SLAS Technology* **18**, 264–268 (2013).

[3] Sud, D. & Kaur, G. A comprehensive review on synthetic approaches for metal-organic frameworks: From traditional solvothermal to greener protocols. *Polyhedron* **193**, 114897 (2021).

[4] Øien-Ødegaard, S., Shearer, G. C., Wragg, D. S. & Lillerud, K. P. Pitfalls in metal–organic framework crystallography: towards more accurate crystal structures. *Chem. Soc. Rev.* **46**, 4867–4876 (2017).

[5] Evans, P. R. An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallographica Section D* **67**, 282–292 (2011).

[6] Bruno, I. J., Shields, G. P. & Taylor, R. Deducing chemical structure from crystallographically determined atomic coordinates. *Acta Crystallographica Section B* **67**, 333–349 (2011).

[7] Moghadam, P. Z. *et al.* Targeted classification of metal–organic frameworks in the cambridge structural database (csd). *Chemical Science* **11**, 8373–8387 (2020).

[8] Larsen, A. H. *et al.* The atomic simulation environment?a python library for working with atoms. *Journal of Physics: Condensed Matter* **29**, 273002 (2017). URL `http://stacks.iop.org/0953-8984/29/i=27/a=273002`.

[9] Bahn, S. R. & Jacobsen, K. W. An object-oriented scripting interface to a legacy electronic structure code. *Comput. Sci. Eng.* **4**, 56–66 (2002).

[10] C., M., F. *et al.* Mercury csd 2.0 - new features for the visualization and investigation of crystal structures. *Journal of Applied Crystallography* **4**, 466–470 (2008). URL `http://scripts.iucr.org/cgi-bin/paper?S0021889807067908`. `https://doi.org/10.1107/S0021889807067908`.

[11] Addicoat, M. A., Coupry, D. E. & Heine, T. Autografs: Automatic topological generator for framework structures. *The Journal of Physical Chemistry A* **118**, 9607–9614 (2014).

[12] Grimme, S., Bannwarth, C. & Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and non-covalent interactions of large molecular systems parametrized for all spd-block elements (z = 1-86). *Journal of Chemical Theory and Computation* **13**, 1989–2009 (2017). URL `https://doi.org/10.1021/acs.jctc.7b00118`. `https://doi.org/10.1021/acs.jctc.7b00118`.

[13] Wonanke, A. D., Bennett, P., Caldwell, L. & Addicoat, M. A. Role of host-guest interaction in understanding polymerisation in metal-organic frameworks. *Frontiers in Chemistry* **9** (2021).

[14] Nurhuda, M., Perry, C. C. & Addicoat, M. A. Performance of gfn1-xtb for periodic optimization of metal organic frameworks. *Phys. Chem. Chem. Phys.* **24**, 10906–10914 (2022).

[15] Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu. *The Journal of Chemical Physics* **132**, 154104 (2010).

URL https://doi.org/10.1063/1.3382344. https://doi.org/10.1063/1.3382344.

[16] Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Physical Review Letters* **77**, 3865–3868 (1996). URL https://link.aps.org/doi/10.1103/PhysRevLett.77.3865.

[17] Lenthe, E. V. & Baerends, E. J. Optimized slater-type basis sets for the elements 1-118. *Journal of computational chemistry* **24 9**, 1142–56 (2003).

[18] Raupach, M. & Tonner, R. A periodic energy decomposition analysis method for the investigation of chemical bonding in extended systems. *The Journal of Chemical Physics* **142**, 194105 (2015). URL https://doi.org/10.1063/1.4919943. https://doi.org/10.1063/1.4919943.

[19] Nazarian, D., Ganesh, P. & Sholl, D. S. Benchmarking density functional theory predictions of framework structures and properties in a chemically diverse test set of metal-organic frameworks. *Journal of Materials Chemistry A* **3**, 22432–22440 (2015). URL http://dx.doi.org/10.1039/C5TA03864B.

[20] Kuc, A. *et al.* Proximity effect in crystalline framework materials: Stacking-induced functionality in mofs and cofs. *Advanced Functional Materials* 1908004–1908017 (2020). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/adfm.201908004. https://onlinelibrary.wiley.com/doi/pdf/10.1002/adfm.201908004.

[21] te Velde, G. *et al.* Chemistry with adf. *Journal of Computational Chemistry* **22**, 931–967 (2001). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.1056. https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.1056.

[22] Vasile, R. L. *et al.* Influence of the synthesis and crystallization processes on the cation distribution in a series of multivariate rare-earth metal-organic frameworks and their magnetic characterization. *Chem Mater* **34**, 7029–7041 (2022).

[23] Rocktäschel, T., Weidlich, M. & Leser, U. Chemspot: a hybrid system for chemical named entity recognition. *Bioinformatics* **28**, 1633–1640 (2012).

[24] Filippov, I. V. & Nicklaus, M. C. Optical structure recognition software to recover chemical information: Osra, an open source solution. *Journal of Chemical Information and Modeling* **49**, 740–743 (2009).

[25] Lowe, D. M., Corbett, P. T., Murray-Rust, P. & Glen, R. C. Chemical name to structure: Opsin, an open source solution. *Journal of Chemical Information and Modeling* **51**, 739–753 (2011).

[26] Kim, S. *et al.* Pubchem 2023 update. *Nucleic Acids Research* **51**, D1373–D1380 (2023).