FAIR-MOFs : A Comprehensive Database for Accelerating the Discovery and Synthesis of Metal-Organic Frameworks

A.D. Dinga Wonanke,[*a,b,e] Antonio Longa,[f] Asha Pankajakshan,[h] 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] *Department of Physics and CSMB, Humboldt-Universität zu Berlin, 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*

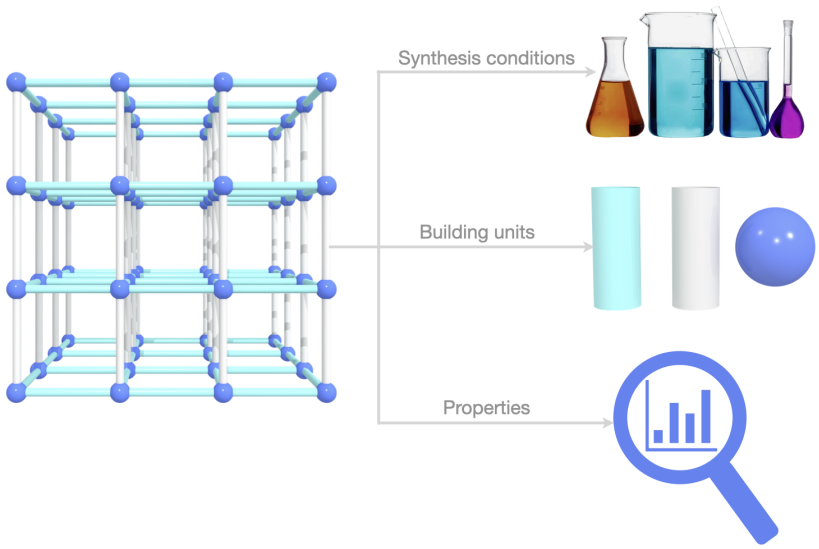
[g] *Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom*

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

Abstract

Metal-organic frameworks (MOFs) are a versatile class of materials with applications in gas storage, separations, and catalysis. Despite extensive research, one of the key factors hindering their broader deployment is the

absence of reproducible and scalable synthetic protocols. To address this, we introduce FAIR-MOFs, a database designed to be Findable, Accessible, Interoperable, and Reusable (FAIR), comprising 45,700 curated experimental structures, 33,361 geometry-optimised structures, and 4,161 entries linked to reported synthesis conditions. Analysis of the dataset showed that the propensity of open-metal sites in MOFs is statistically associated with reaction temperature, metal salts, ligands, topology, and metal secondary building unit. Furthermore, we developed a retrosynthetic recommender that captures literature co-usage patterns among solvents, metal salts, and ligands. For any given component, the system suggests compatible reagents and retrieves MOFs prepared under similar conditions. Finally, we trained a graph-based neural network integrated with our MOF deconstruction module to predict most probable metal salts, ligands and solvents directly from 3D structures of experimental or hypothetical MOFs. Using this model, we successfully synthesised MOFs randomly selected from hypothetical MOF databases illustrating the potential of FAIR-MOFs to accelerate the discovery and enable data-driven synthesis of MOFs.



1 Introduction

The conventional synthetic approach in reticular chemistry and other branches of chemistry often relies on repetitive trial-and-error syntheses. This is particularly unreliable because syntheses are rarely reproducible on the first attempt and are difficult to scale and optimise for industrial application. Against this backdrop, metal-organic frameworks (MOFs), which are porous crystalline materials assembled from metal nodes and organic ligands have over the last two decades been dubbed as the most promising materials for gas and energy storage, catalysis, separations, and drug delivery. Their modularity and tunable functionality has attracted scientific interest, with more than 126,000 publications indexed in Web of Science. However, this interest has not yet translated into real-world industrial-scale application primarily because of synthetic unreliability and scale-up challenges. Moreover, although computational screening has predicted millions of high-performing MOFs for diverse applications, only approximately 120,000 MOFs have been synthesized and refined to the standards required for deposition in the Cambridge Structural Database (CSD). [1–10] This stark contrast highlights the excessively high activation barrier between current computational predictions and experimental syntheses.

In an attempt to bridge this gap, several databases have been developed. The CORE MOF series, Quantum MOF (QMOF), and MOSAEC-DB curated structures from the CSD and made them computationally ready [11–15]. Recent efforts, such as MOFChecker and MOFDescribe, have further improved data quality and standardisation. [16, 17] MOFChecker introduced automated routines for structural validation and curation of MOFs to ensure their computational readiness. Meanwhile, MOFDescribe have established a reproducible ecosystem for featurising and benchmarking MOFs for machine learning applications. Although, these

resources have been instrumental in speeding up simulations and property predictions, they do not address the direct linkage between structure and experimental synthesis conditions, which remains the key gap in material discovery to expedite bench synthesis.

The use of natural language processing (NLP) and large language models (LLMs) to extract experimental synthetic conditions from published literature has greatly advanced the automation of chemical knowledge discovery. [18–25] For example, DigiMOF by Moghadam *et al.* and the approach of Kim *et al.* systematically extract reagents (metal salts, linkers, solvents) and reaction parameters from journal articles. [19, 25] Yaghi *et al.* reported a ChatGPT-based framework trained on 700 MOF papers to automatically identify synthesis conditions from textual descriptions. [20] Similarly, Tsotsalas *et al.* developed a random forest model (SynMOF) trained on 900 MOFs that predicts reaction parameters outperforming expert intuition, [18] while Pils *et al.* applied the SyCoFinder genetic algorithm to optimize synthesis conditions for HKUST-1 SURMOFs. [26]

Despite this progress, existing text-mining and recommender systems share a critical limitation: they primarily operate on paragraph-level extraction of synthesis information and link it to the textual occurrence of a MOF name, without explicitly mapping those synthesis conditions to the corresponding crystal structure. This is problematic because the synthesis route that yields a specific crystal structure with its correct topology, pore geometry, and coordination environment is crucial for reproducibility and targeted material design. In practice, material discovery typically begins with the prediction of hypothetical 3D structures, making it fundamentally important to establish a direct mapping between 3D structures and their experimental synthesis conditions. Without such structure-synthesis linkage,

current methods cannot predict how a hypothetical MOF could be synthesised, nor can they ensure that recommended conditions lead to the desired phase or topology.

To address this limitation, we developed FAIR-MOFs, a comprehensive Findable, Accessible, Interoperable, and Reusable database that establishes a direct correspondence between the crystal structures of MOFs and their reported synthesis conditions. The database integrates computation-ready and experimentally verified MOFs with their geometric properties, topologies, and structural building units. As a proof of concept, we trained a graph-based neural network that learns reagent and condition patterns directly from the 3D atomic graphs of MOFs to enable the prediction of synthesis routes for both known and hypothetical structures. We further developed an interactive co-usage analytics platform to visualise empirical associations among metal salts, linkers, and solvents that enable users to explore compatible reagent combinations and retrieve experimentally realised frameworks synthesised under similar conditions. We also implemented a robust search engine that provides a natural-language search interface to enable researchers to query the database using domain-specific terms without coding expertise. Finally, we performed an in-depth statistical and topological analyses of the curated dataset to decipher key synthetic descriptors that govern the formation of MOFs with open metal sites.

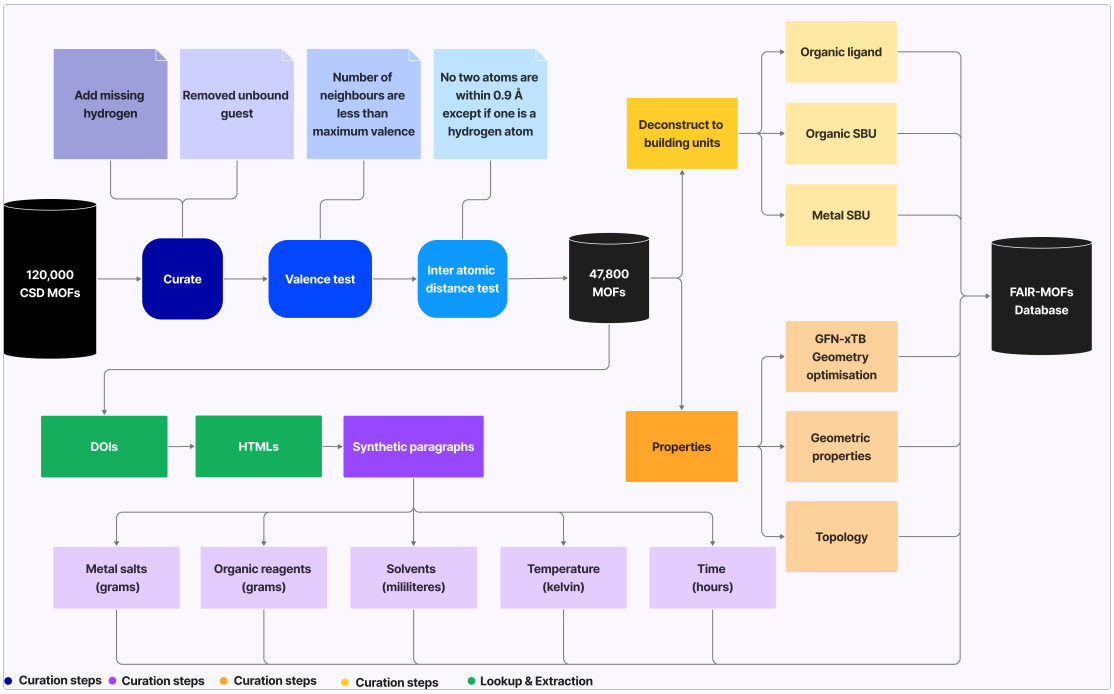


Fig. 1 Workflow for implementing the FAIR-MOF database. The process begins with curating the MOF subset from the Cambridge Structural Database, followed by deconstructing each MOF into its unique building units, computing properties, and text mining the synthetic conditions. Synthetic conditions are extracted by reading the HTML file of a journal article, parsing it into a list of paragraphs, and applying a machine learning model that identifies paragraphs describing the synthesis procedure. Regular expressions and sentence transformers are then applied to extract the reagents, their exact quantities, the time, the temperature, and the synthetic method.

2 Results and Discussion

2.1 Structural Data

The workflow for implementing the FAIR-MOF database is summarised in Fig. 1. The database is composed of 45,700 curated experimental crystal structures of MOFs. A full description of the curation methodology and the tools used can be found in **S-1** of the Electronic Supporting Information (ESI). From the 45,700 structures, we successfully performed a geometry optimisation on 33,361 crystal structures. Systems containing more than 5,000 atoms within the unit cell were computationally intractable so their structures were not optimised.

The distributions of the unoptimised and geometry-optimised datasets are illustrated as Uniform Manifold Approximation and Projection (UMAP) plots in Fig. 2 and Fig. 3, respectively where each properties are represented by colors. Overall, both datasets exhibit similar property distributions. From Fig. 2a–b and Fig. 3a–b, the pore-limiting diameter (PLD) and largest cavity diameter (LCD) span 0–70 Å with 90 % of MOFs falling within the range 0–19 Å, and approximately 9 % between 20–38 Å. Fig. 2c and Fig. 3c show that the accessible surface area (ASA) spans 0–3500 m^2 cm^{-3} with about 70 % of MOFs lying between 0–900 m^2 cm^{-3}, 20 % between 900–1900 m^2 cm^{-3}, 9 % between 1900–2900 m^2 cm^{-3}, and only 1 % exceeding 3000 m^2 cm^{-3}. From Fig. 2d and Fig. 3d, the accessible pore volume (AV) ranges from 0–120,000 $Å^3$ and 0–80,000 $Å^3$ for the unoptimised and optimised datasets respectively. However more than 98 % of MOFs in both datasets fall below 22,000 $Å^3$. Furthermore, it can be observed from Fig. 2e and Fig. 3e, that over 70 % of MOFs have void fractions below 0.22, 20% between 0.22–0.42, 8 % between 0.42–0.68, and only 2 % exceed 0.70. Finally, it can be observed from Fig. 2f and Fig. 3f that the maximum coordination number of metals is 12. More

than 80 % of MOFs have coordination numbers of 4 or 6, while around 20 % have a coordination number of 8.

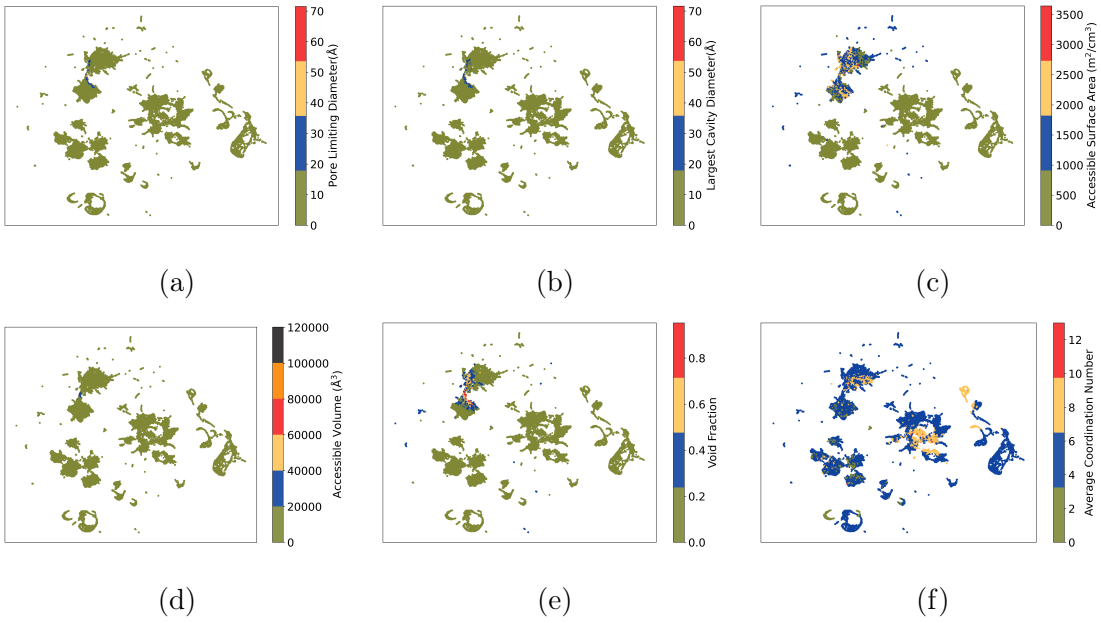


Fig. 2 Illustrations of the data distribution for (a) pore-limiting diameter (PLD), (b) largest cavity diameter (LCD), (c) accessible surface area (ASA), (d) accessible pore volume (AV), (e) void fraction, and (f) average metal coordination number for the unoptimised dataset containing 45,700 curated MOFs are shown. Each panel displays a UMAP projection onto two principal components using the cosine metric (n_neighbors = 20, min_dist = 0.1), with points colored by the respective property.

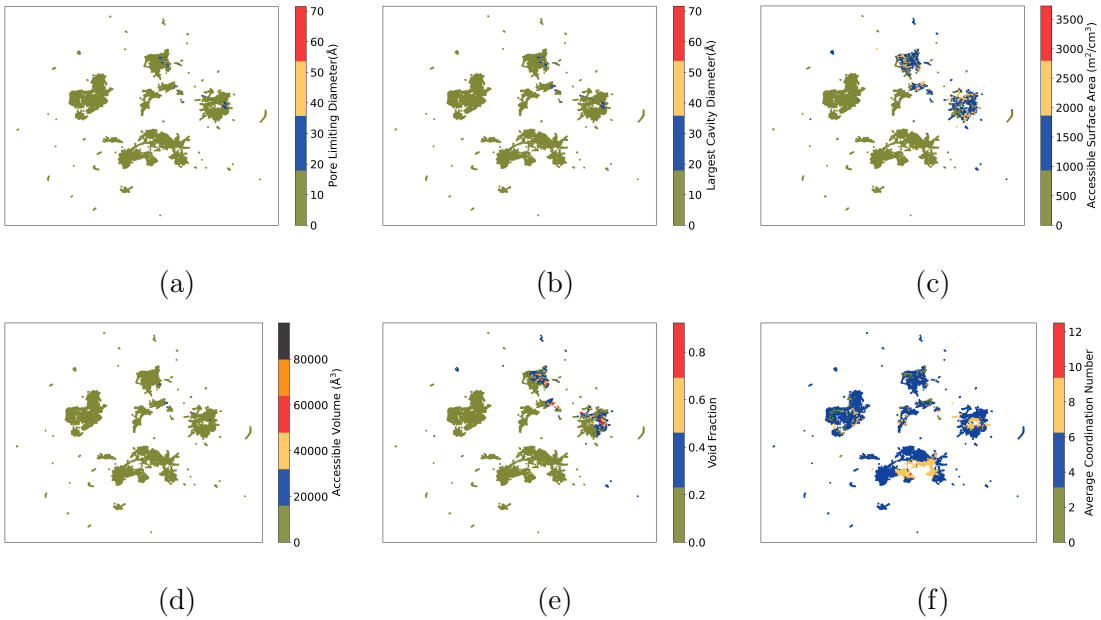


Fig. 3 Illustrations of the data distribution for (a) pore-limiting diameter (PLD), (b) largest cavity diameter (LCD), (c) accessible surface area (ASA), (d) accessible pore volume (AV), (e) void fraction, and (f) average metal coordination number for the geometry-optimised dataset containing 33,361 curated MOFs are shown. Each panel displays a UMAP projection onto two components using the cosine metric (n_neighbors = 15, min_dist = 0.1), with points colored by the respective property.

2.2 mofstructure

We implemented a robust MOF and scalable MOF deconstruction Python module called **mofstructure**. Details about the algorithm and methods are described in **S-2** of the ESI or the extensive **mofstructure documentation**. The **mofstructure** package provides four principal methods:

1. **Efficiently removes** unbound guest molecules found in porous materials, including Covalent Organic Frameworks (COFs), zeolites, and Metal-Organic

Frameworks (MOFs).

2. **Deconstructs MOFs** into their unique metal secondary building units (SBUs), organic SBUs, and organic ligands. It also computes cheminformatic identifiers for each building unit, connects to the PubChem API to retrieve the names and chemical properties of organic ligands found in MOFs, determines the topology of the SBUs and their number of points of extension, and encodes atom indices for all building units to enable traceability back to the parent MOF.

3. **Includes a Python wrapper for Zeo++**, which enables the computation of geometric properties such as pore size, accessible surface area, and accessible volume.

4. **Computes open-metal sites and metal coordination environments**, and includes a function to detect overlapping atoms in crystal structures.

Moreover, the **mofstructure** module is capable of classifying the metal secondary building units (SBUs) into eight classes: **rodlike**, **paddlewheel with water**, **paddlewheel**, **MOF32**, **IRMOF**, **UIO66**, and **ferrocenelike**. These correspond, respectively, to MOFs with rodlike topologies (which are known to be periodic in one dimension); MOFs whose SBUs adopt a paddlewheel geometry, with or without coordinated water molecules; the MOF-32 series; the IRMOF series; zirconium cluster-based MOFs; and MOFs containing ferrocene.

The output from **mofstructure** provides a rich schema of searchable quantities that users can use to identify specific MOFs based on properties such as pore limiting diameter (PLD), largest cavity diameter (LCD), accessible surface area (ASA), accessible volume (AV), void fraction, number of channels, organic ligand (IUPAC name, InChIKey, or SMILES string), metal type, metal coordination

number, number of point of extension in building units, presence of open-metal sites, topology, and secondary building unit type.

The robustness and scalabity of **mofstructure** was evaluated by applying on both our geometry-optimised and unoptimised dataset. The module can also be applied to remove unbound guest molecules and compute porosity of other **porous systems** such as **COFs** and **zeolites**. In comparison to other existing deconstruction methods, such as *mofid*, [27] **mofstructure** is more robust and can reliably deconstruct rodlike systems such as MOF-74, where existing methods fail. However, for systems that do not have well-defined building units (e.g., Refcode: ABAFUH illustrated in Fig. 4), only their organic ligands will be returned. The SBU and consequently the topology, cannot be extracted for these systems.

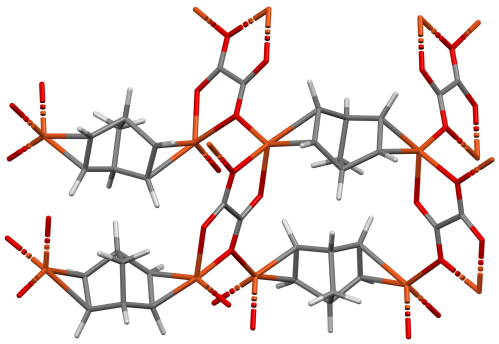
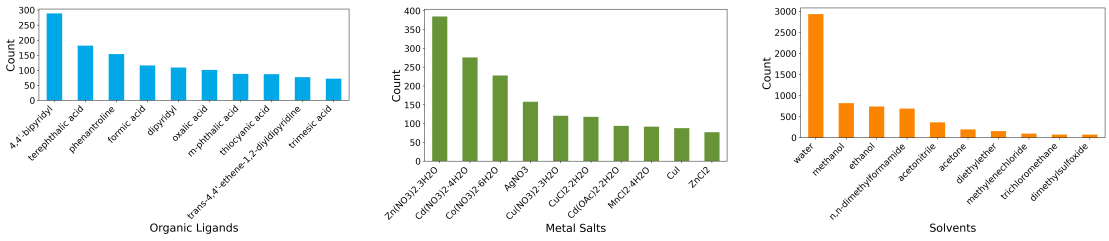


Fig. 4 Illustration of a MOF (refcode: ABAFUH) that cannot currently be fully deconstructed into secondary building units using the `mofstructure` Python module. At present, only the organic ligands and metal clusters can be extracted. More robust rules are being developed to enable the accurate deconstruction of SBUs across a wider range of MOF topologies.

2.3 Text Data

The extraction of synthesis conditions from publications and their mapping to exact crystal structures was the major bottleneck in this study. A full description of the challenges and workarounds, as well as details of all extracted quantities, is provided in **S-5** of the ESI. We began by mining 47,521 unique DOIs from the CSD, corresponding to journal articles describing the synthesis of the MOFs deposited in the CSD. This enabled us to text-mine 1,743 unique organic ligands, 793 unique metal salts, and 78 unique solvents, which were mapped to 4,161 MOFs, whose distributions are illustrated in Fig. 5.



(a) (b) (c)

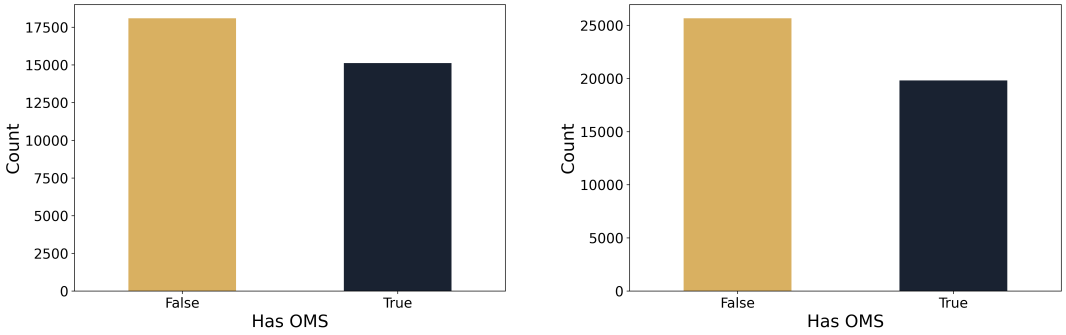
Fig. 5 Illustration of the distribution of the top 10 most frequently occurring organic ligands (a), metal salts (b), and solvents (c). To ensure consistent representation, all metal salts were converted to their chemical formulae, and ligands and solvents to InChIKeys. This approach was necessary because chemical names are not canonical, meaning a single compound may be referred to by multiple names. In contrast, each compound has a unique InChIKey.

2.4 Use Case

2.4.1 Synthetic Factors Affecting Open-Metal Sites

The presence of open-metal sites (OMS) is ubiquitous in MOFs. As shown in Fig. 6, more than 40% of MOFs in both the geometry-optimised (33,361 MOFs)

and unoptimised (45,700 MOFs) datasets possess OMS. Beyond their prevalence, the presence of OMS plays a significant effect on adsorption enthalpies, catalytic turnover, and chemical stability, as they function as strong, localized Lewis acid sites. This behavior is well established in prototypical systems such as HKUST-1, the M-MOF-74 series, and MIL-101-type frameworks, where removal of terminal ligands exposes metal centers that strongly bind small molecules such as CO_2, H_2, NH_3.



(a) (b)

Fig. 6 Illustration of the data counts of open (dark blue) and closed (yellow) systems in the geometry-optimized dataset (a) and the pristine, curated experimental dataset without relaxed structures (b), containing 33,361 and 45,700 MOFs, respectively.

Despite its importance, there is so far no well-established synthetic heuristic for predicting which reaction conditions are most likely to yield open metal sites. Most studies rely on empirical observations, like post-synthetic drying in strongly coordinating solvents such as DMF or ethanol, or the presence of uncoordinated linkers that introduce defects and missing-coordination sites. These strategies, while effective in certain cases, remain system-specific and do not provide a systematic approach across the chemical space of MOFs.

We move beyond this piecemeal approach by performing an in-depth statistical analysis of the FAIR-MOF dataset to identify key synthetic factors that correlate with OMS formation. In particular, we explore the effects of reaction temperature (Fig. 7a), time (Fig. 7b), metal salts (Fig. 7c), organic ligands (Fig. 7d), solvents (Fig. 7e), topology (Fig. 7f), and SBU type (Fig. 7g).

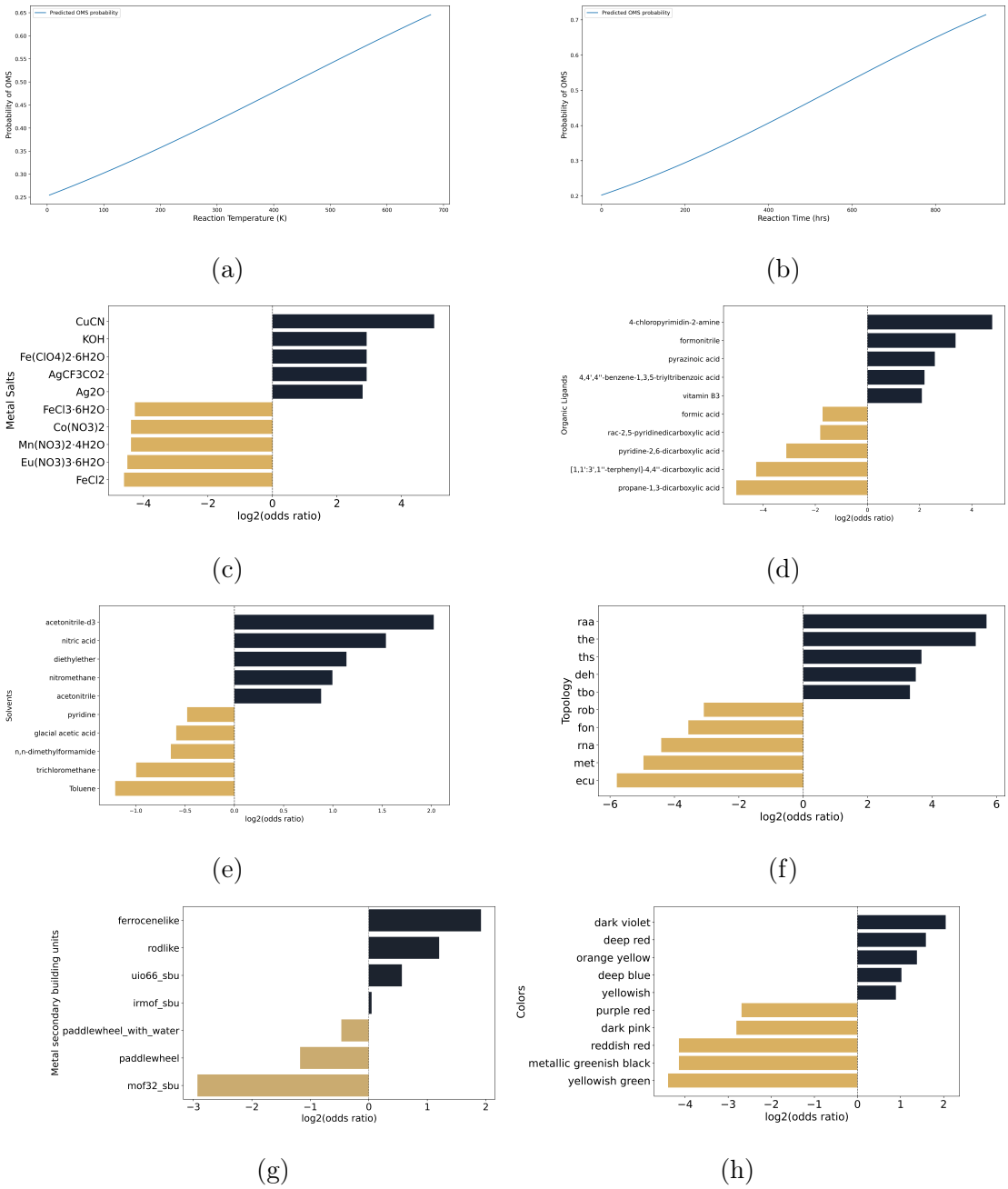


Fig. 7 Illustration of the probability of obtaining MOFs with OMS as a function of (a) reaction temperature at median reaction time and (b) reaction time at median temperature. Panels (c)-(h) show the association, expressed as log2(odds ratio), between the propensity for OMS formation and various factors: metal salts (c), organic ligands (d), solvents (e), topology (f), SBU type (g), and color of the MOF. Black bars indicate a positive association with OMS formation, while dark yellow bars indicate a negative association. Each bar chart displays the five most and five least influential quantities.

The effect of temperature and time on OMS formation was assessed using a logistic regression model in which the probability of OMS formation was modeled as a function of reaction temperature and reaction time with the presence or absence of OMS treated as a binary response variable. It can be observed from Fig. 7a that there is a positive linear relationship between reaction temperature and the likelihood of OMS formation. This suggests that syntheses conducted at higher temperatures such as solvothermal and hydrothermal methods are more likely to result in MOFs with OMS than low-temperature methods like electrochemical and mechanochemical synthesis as observed in **S-6.1** of the ESI. In contrast, it can be observed from Fig. 7b that there is a negative linear relationship between reaction time and the probability of OMS formation. At short reaction times, the probability of OMS formation is approximately 50%, indicating that there is no clear trend. However, as the reaction time increases, the probability of OMS formation decreases linearly.

On the other hand, the impact of the other parameters on OMS formation was evaluated by performing a statistical enrichment analysis. For each parameter a 2×2 contingency table was constructed to compare the frequency of each label in OMS-containing versus non-OMS structures. Odds ratios (OR) were computed with the Haldane-Anscombe correction to avoid division by zero and expressed as $\log_2(\mathrm{OR})$. Positive $\log_2(\mathrm{OR})$ values indicate high association and negative values indicate low association to OMS formation. The statistical significance was assessed using Fisher's exact test with Benjamini-Hochberg false discovery rate correction.

Fig. 7c correspond to the top five most and least influential metal salts with the highest (black bars) and lowest (dark yellow bars) odds of forming MOFs with

OMS. It can be observed from Fig. 7c that MOFs formed with CuCN have the highest odds of forming OMS with a $\log_2(\mathrm{OR})$ of approximately 5. This means that MOFs synthesised with CuCN have about $2^5 \approx 32$-fold higher odds of forming OMS compared to those synthesised without CuCN. Similarly, MOFs formed with KOH, Fe(ClO4)2.6H2O, AgCF3COO and Ag2O all have $\log_2(\mathrm{OR}) > 3$, corresponding to at least a 8-fold enrichment in OMS formation. On the contrary, MOFs synthesised with FeCl3.6H2O, Co(NO3)2, Mn(NO3)2.4H2), Eu(NO3)3.6H2O and FeCl2 all have a $\log_2(\mathrm{OR}) < -4$, indicating that they have an $\approx$ 16-fold odds of forming MOFs with no OMS.

It can also be observed from Fig. 7d that the choice of ligands also has a significant impact on the odds of forming OMS. For instance, MOFs synthesised using 4-chloropyrimidine-2-amine ligands have approximately a 32-fold higher odds of forming OMS. Meanwhile, MOFs synthesised using propane-1,3-dicarboxylic acid have comparatively higher odds of not forming OMS. Furthermore, it can be observed from Fig. 7e-f that while the choice of solvent has a negligible effect on the odds of forming OMS, the choice of topology plays the most significant role. It can be observed in Fig. 7f that MOFs with **raa**, **the**, and **ths** topologies have greater than 32-fold odds of having OMS, while MOFs with **ecu**, **rna**, and **met** topologies have similarly high odds of forming MOFs with no OMS. Concomitantly, we also observed from Fig. 7g-h that the type of metal SBU and the color of the MOF also have an impact on the odds of OMS formation.

2.4.2 Metal salt prediction

From the 3D structure of a MOF, one can often unambiguously determine the organic ligands involved. Using our **mofstructure** Python module, we accurately identify ligand fragments, compute their InChIKeys, and query the PubChem API

to retrieve IUPAC names.

However, currently there is no method to directly deduce the exact identity or hydration state of the metal salt used in synthesis solely from the 3D structure. Yet, the identity of the metal salt is critical in the synthesis of MOFs. For instance, in cyclodextrin-based MOFs, using KOH versus potassium benzoate as the potassium source leads to distinctly different crystal systems, body-centred cubic versus trigonal respectively, which result in markedly different textural properties such as BET surface area and pore volume.

Consequently, we decided to address this gap by developing a graph neural network (GNN) model that directly predicts the identity of the metal salt based solely on features extracted from the 3D structure. To achieve this, we constructed graph-based representations of MOFs tailored for machine learning. Each MOF is encoded as an undirected graph, where the nodes represent the atoms, and the edges capture chemical bonds derived from the crystal structure. For periodic systems like MOFs, global graph-level features, such as lattice parameters, are incorporated.

To assess the utility of these graph-based representations for predicting the metal salts used in MOF synthesis, we formulated the task as a multi-class classification problem across the salts, with each MOF encoded as a structure graph comprising 3,054 instances. The baseline representation combined the structural graph with node, edge, and lattice features, to which global descriptors were incrementally added. Specifically, we first evaluated each descriptor individually (concentration of elemental species, metal coordination number, crystal system, open metal sites (OMS), and space group). Further details on the model and its implementation

are provided in 4.6.

Table 1 Top-5 and Top-3 classification accuracy (%) for metal salt prediction using different combinations of global features. All models include graph, lattice, node, and edge features. Highest accuracy are in bold, and the global highest is in bold-blue. Top-3 and Top-5 respectively correspond to the frequency with which experimentally-reported salt appears within top 3 and top 5 recommended synthesis conditions.

Exp	Conc. Species	CN	Crystal sys.	OMS	S. Group	Top-3 Acc.	Top-5 Acc.
1	✗	✗	✗	✗	✗	55.88 ±1.64	65.03 ±1.39
2	✓	✗	✗	✗	✗	**66.43** ±1.67	**76.91** ±1.30
3	✗	✓	✗	✗	✗	65.10 ±0.52	76.86 ±0.67
4	✗	✗	✓	✗	✗	55.69 ±2.16	65.88 ±2.68
5	✗	✗	✗	✓	✗	55.36 ±0.87	65.88 ±1.07
6	✗	✗	✗	✗	✓	52.88 ±2.23	62.68 ±1.54
7	✓	✓	✗	✗	✗	67.06 ±1.83	76.54 ±0.89
8	✓	✗	✓	✗	✗	67.52 ±0.87	77.45 ±1.81
9	✓	✗	✗	✓	✗	**68.37** ±1.04	**77.92** ±1.33
10	✓	✗	✗	✗	✓	67.71 ±0.72	77.45 ±0.72
11	✓	✓	✗	✓	✗	**68.24** ±1.77	76.73 ±1.10
12	✓	✗	✓	✓	✗	67.19 ±0.84	**77.71** ±0.81
13	✓	✗	✗	✓	✓	67.32 ±1.40	77.52 ±0.70
14	✓	✓	✓	✓	✗	66.27 ±1.14	76.01 ±0.53
15	✓	✓	✗	✓	✓	**68.50** ±1.77	**77.97** ±0.94
16	✓	✓	✓	✓	✓	67.78 ±1.55	76.93 ±0.76

Exp: Unique training experiment. Conc. Species: Concentration of each element in the structure (similar to empirical formula). CN: Metal coordination number. Crystal sys.: Crystal system. OMS: Open metal site. S. Group: Space Group symmetry.

The results, summarised in Table 1, reveal several consistent trends. Incorporating concentration of species dramatically improved performance over the baseline, raising Top-3 and Top-5 accuracies from 55.88 % and 65.03 % to 66.43 % and 76.91

%, respectively (Exp. 2). The Top-3 and Top-5 accuracies here correspond top the frequency with which experimentally-reported salt appears within top 3 and top 5 recommended synthesis conditions respectively. The concentration of species consistently emerged as the most important feature for obtaining better predictions across all settings. Adding further descriptors provided incremental improvements, with the best-performing model (Exp. 15) achieving 68.50 % Top-3 and 77.97 % Top-5 accuracy when combining concentration of species, coordination number, OMS, and space group. Curiously, the full-feature configuration (Exp. 16) did not yield additional gains, suggesting that some descriptors introduce redundancy rather than complementary information.

In comparison with prior synthesis recommender systems, this study provides a more generalisable representation. SynMOF, which uses tabulated reaction parameters, is restricted to a limited subset of linker-metal combinations. SyCoFinder, based on genetic optimisation, explores parameter space iteratively for a single MOF family (HKUST-1 SURMOFs). Our graph-based framework, by contrast, learns transferable patterns across more than 4,000 distinct MOFs and achieves Top-5 accuracies approaching 78 %, which surpasses domain-restricted models without requiring predefined reaction templates. Furthermore, unlike text-driven ChatGPT synthesis extraction tools, our model is structure-aware and quantitatively predictive rather than descriptive.

2.4.3 An interactive tool for exploring synthetic components

To provide a complete workflow for discovery and to expedite synthesis, we implemented **cheminteraction**, which is an interactive web application that transforms curated co-usage data into a bench-facing search and visualization interface. For any given ligand, solvent or metal salt, the tool returns literature-derived

co-occurrence patterns, alongside an interactive hierarchical network and ranked tables of the most frequent partners. For example, querying m-nitrobenzoic acid, the application immediately displays the solvents and metal salts most commonly reported with this ligand as observed in Fig. 8.

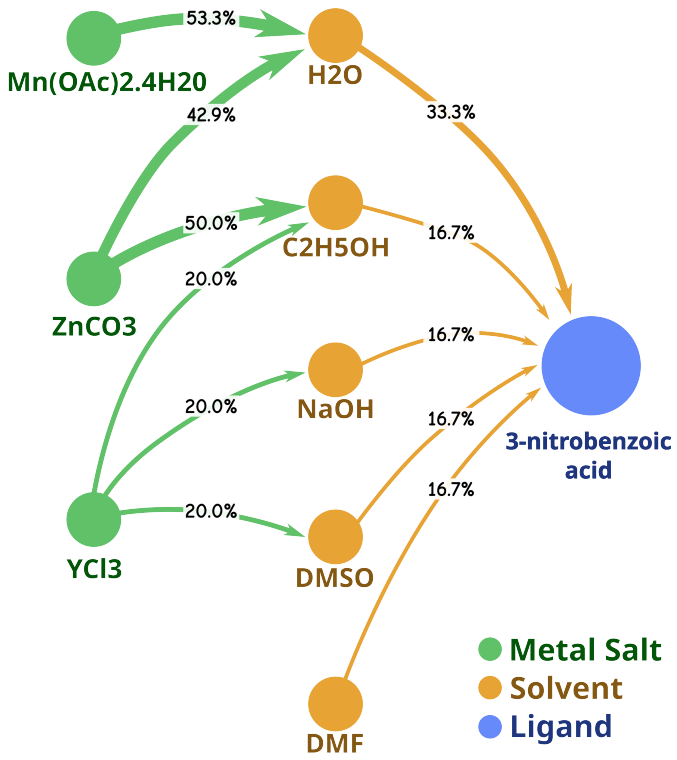


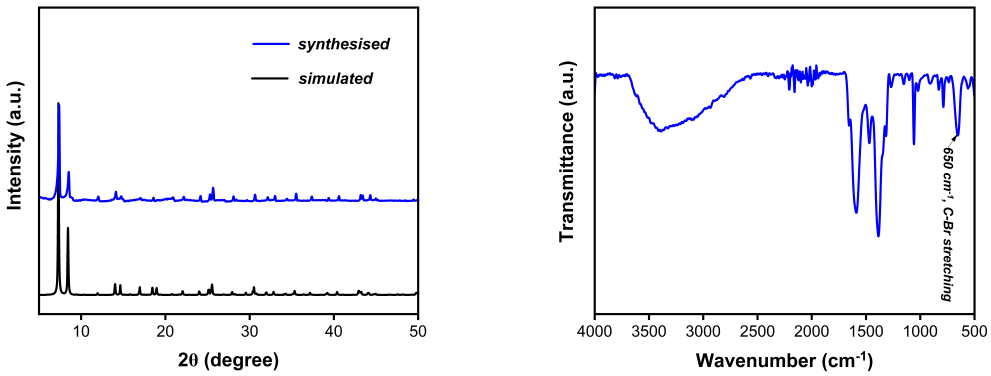
Fig. 8 Example web-app. Search results for the ligand m-nitrobenzoic acid, together with the associated solvents and metal salts.

The application also expedites the purchase of each reagent by linking directly to catalogue endpoints and site-scoped queries. In addition, the crystal structures of MOFs from the co-usage dataset are returned alongside their geometric properties and direct journal links. Beyond reagent co-usage, **cheminteraction** also bundles the geometry-optimized structures of FAIR-MOFs and provides a streamlined interface that enables rapid structure search and retrieval across commonly used fields, including refcode, DOI, ligand name or InChIKey, ligand abbreviation,

metal salt, topology, pore metrics (PLD/LCD), open-metal-site flag, and free-text keywords. The platform further supports CSV and CIF export for downstream analysis.

2.4.4 Experimental Synthesis

To validate the potential of our model to accelerate the discovery and synthesis of MOFs, we decided to randomly select a hypothetical MOF from the qMOF data, which was extracted from the **GMOF** with id **qmof-6031bc0**. [28] The reported Powder X-ray Diffraction (PXRD) and Fourier Transform Infrared (FTIR) spectroscopy are illustrated in Fig. 9 **(**a) and **b** respectively.



(a) (b)

Fig. 9 Powder X-ray Diffraction (a) and Fourier Transform Infrared Spectroscopy (b) of a synthesised MOF, whose synthetic condition was predicted using fairsyncondition. The blue plot corresponds to the experimentally synthesised data, while the black plot corresponds to the simulated data.

The PXRD confirmed the phase purity of the synthesised MOF. The experimental pattern closely matches the simulated profile derived from single-crystal data, which is consistent with the phase predicted by the Fairmofsyncondition report.

The FTIR spectroscopy further validated the molecular structure and functional groups present in the MOF. The spectrum exhibits a distinct C–Br stretching band at ~ 650 cm^{-1}, confirming successful incorporation of the brominated linker. Prominent bands at 1592 cm^{-1} and 1384 cm^{-1} correspond to the asymmetric and symmetric stretching modes of carboxylate (COO^-) groups, respectively, which is indicative of coordination to the Zr_6 clusters. An additional band at ~ 1057 cm^{-1} is attributed to C-O stretching vibrations.

Analysis of the crystal structure shows that the MOF corresponds to **UiO-66-Br$_2$**, in which each $[Zr_6O_4(OH)_4]^{12+}$ node is 12-connected to 2,5-dibromoterephthalate (Br_2-BDC) linkers. Whereas prior syntheses have almost exclusively used $ZrCl_4$ as the metal precursor, our model instead predicted $ZrOCl_2 \cdot 8\,H_2O$ as the optimal precursor, which yielded pure phase of crystalline **UiO-66-Br$_2$**.

2.4.5 Towards complete prediction of synthetic conditions

Despite the predictive power of the FAIR-MOFs database in recommending viable synthetic pathways, our overarching goal is to advance the discovery and expedite the synthesis of MOFs by predicting the full set of synthesis conditions including reagent concentrations, temperature, pressure, time, humidity and synthetic method. So far, with only 4,160 structures in the database mapped to experimental synthetic conditions, the feasibility of such comprehensive predictions remains limited. Nevertheless, our results indicate that this goal is achievable with expanded datasets.

To quantify the impact of data availability, we trained a graph neural network (GNN) on progressively larger fractions of the dataset (20 – 100 %) and evaluated its performance on three key prediction tasks: ligands, metal salts, and solvents.

The performance was benchmarked against a random baseline that computes the top-k accuracy (with $k = 10$) of a model selecting k random classes from a uniform distribution thereby providing a realistic lower bound that accounts for class imbalance and task difficulty.

The GNN consistently outperformed this baseline across all targets, with the performance gap widening as more training data became available (Fig. 10). Notably, ligand prediction exhibited the largest gain, with improvements exceeding 1900 % at full data capacity, while metal salt prediction surpassed 1400 %. By contrast, solvent prediction showed more modest gains. This is understandable, as there is only limited variation among solvents, indicating that the dataset already provides near-complete coverage of all possible solvents.

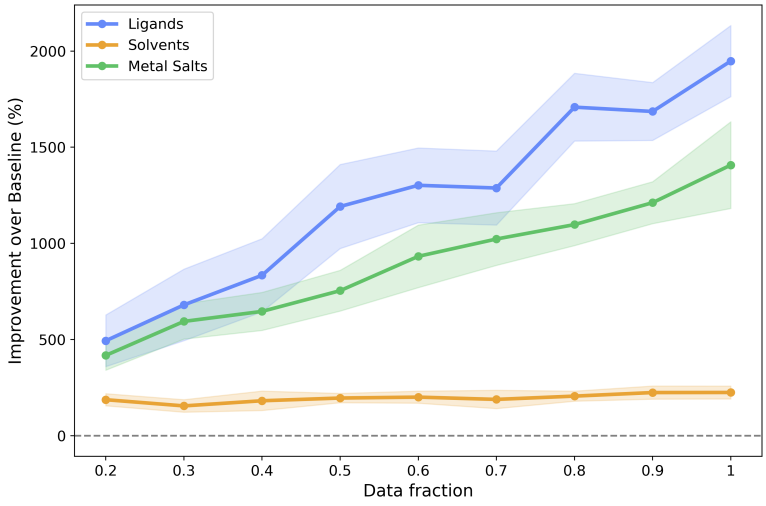


Fig. 10 Percentage improvement over random baseline for ligand, metal salt, and solvent prediction tasks across increasing training data fractions. Shaded regions represent standard deviation across multiple random seeds.

3 Conclusion

In this work, we present FAIR-MOFs, the most extensive resource to date that bridges the gap between the structure, properties, and synthesis of metal–organic frameworks. The data set unifies curated, computation-ready structures with experimentally reported synthesis conditions, providing a foundation for accelerated simulation, discovery, and reproducible synthesis.

We developed the mofstructure module for automated MOF deconstruction, enabling direct identification of organic ligands, metal secondary building units, geometric descriptors, and cheminformatic identifiers from crystallographic data. In parallel, the fairmofsyncondition workflow performs large-scale text mining of published synthesis to extract reagents, conditions, and quantities, thereby resolving the long-standing challenge of linking literature data to specific CSD entries. Together, these tools provide an interoperable structure–synthesis map that connects atomic-scale representations with underlying bench synthesis.

The integration of synthesis-aware structural descriptors further revealed statistically significant correlations between reaction conditions and functional properties, notably the formation of open metal sites (OMS). Because OMS critically influence adsorption, catalysis, and framework stability, identifying the synthetic parameters that promote their formation demonstrates the potential of FAIR-MOFs to guide experimental design.

Building on this foundation, we implemented a graph-base neural network to predict key synthesis components, including ligands, metal salts, and solvents, directly from 3D structural graphs. The performance of the model scales systematically with dataset size, highlighting the potential of data expansion for comprehensive

synthesis prediction. Guided by these predictions, we successfully synthesised a MOF selected from a hypothetical database, thereby validating the practical utility of our model in translating structural representations into realistic synthetic routes. The accompanying interactive web platform completes the retrosynthetic loop by providing a visual representation of reagent co-usage patterns.

In summary, the integration of structural deconstruction, geometric characterisation, text-mined synthesis knowledge, and graph-based predictive modelling establishes a new paradigm for structure-to-synthesis translation in reticular chemistry. Unlike previous frameworks that focuses data curation or property prediction, FAIR-MOFs closes the synthesis-structure-property loop thus paving the way for self-driving synthesis platforms and generative models capable of proposing not only new MOF architectures but also viable synthetic routes. This unified framework positions FAIR-MOFs as a cornerstone resource for the emerging era of autonomous reticular chemistry.

4 Methods

4.1 Data extraction

The crystal structures of all MOFs were extracted from the CSD (November 2022 release), which is known to host a MOF-subset of approximately 118,313 MOFs. A detailed description of the procedure is presented in section **S-1.1** of the Electronic Supporting Information (ESI). We used the CSD API to normalise the bonds in each structure and fill in missing hydrogen atoms as desrcibed in **S-1.3-1**. Since more than 52 % of MOFs in the CSD have unbound guest molecules in their crystal structures, [29] we decided to implement a robust model that searches for unbound guest molecules and then removes them. Our method reads in the crystal

structure from any file format and for any porous material then creates a graph and uses Depth-first search graph algorithm to compute connected components. The connected components correspond to a list of all unconnected entities in the system. The method then searches for the component containing the periodic systems and considers every other unconnected component as unbound guest. This approach is robust in comparison to existing guest removal methods and can be used for any porous periodic system such as Zeolites, Covalent-organic frameworks (COFs) and MOFs. We then implemented a workflow for filtering out systems containing overlapping atoms as fully described in section **S-1.3-3** of the ESI. 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 45,700 MOFs.

4.2 Structure relaxation

The geometry of all the curated structures containing elements with $Z \leq 86$ was performed at GFN-xTB level of theory and systems in which any element had a $Z \gg 86$ was computed at PBE-D3/TZP level of theory. Moreover, the thermodynamic stability of each MOF was computed at these level of theory both at the ligand/cluster and sbu building units. Further details on the structure relaxation is full described in section **S-3** of the ESI.

4.3 MOF deconstruction

To obtain unique building units, topology and cheminformatic identifiers, we developed a robust MOF deconstruction module call **mofstructure**. This module has two primary procedures to systematically deconstruct MOFs into their constituent building units, as depicted in Fig. S4 of section of the ESI. Full description of the algorithm is provided in in section **S-2** of the ESI. 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**

4.4 Geometric properties

The geometric properties, which mainly involves the porosity of the MOFs were computed using zeo++. [30] To enable an easy use of zeo++ across every computer architecture and facile implementation into NOvel MAterials Discovery (NOMAD), we decided to implement a python binary from the original c++ implementation of zeo++ called pyzeo. This module can be downloaded from the following link **https://github.com/nomad-coe/pyzeo.git**.

All geometric properties were computed using the high-accuracy flag with a probe volume of 1.86 Å and 10,000 simulation cycles. Note that this code has been implemented in NOMAD and can compute the porosity of any porous material. This is advantageous to both experimental and computational chemist for performing high-throughput screening of porous systems at minimal computational cost.

4.4.1 Dataset distribution through Uniform Manifold Approximation and Projection

We performed a Uniform Manifold Approximation and Projection (UMAP) on the MOF dataset to provide a visual illustration of the representation and distribution. To achieve this, we started by constructing a high-dimensional feature space that integrates chemical, structural, and porosity descriptors. Specifically, chemical descriptors included metal and organic secondary building units encoded as InChIKeys, as well as metal SBU type (e.g paddlewheel, rodlike e.t.c), the number of organic and metal SBUs, and their points of extension. Structural descriptors comprised density, unit cell volume, number of metals, presence of open metal

sites, average metal coordination number, and the fraction of open metal sites. Porosity descriptors included pore-limiting diameter (PLD), largest cavity diameter (LCD), accessible surface area (ASA), accessible pore volume, void fraction and the number of channels.

Categorical features (e.g. InChIKeys, SBU type) were encoded as one-hot vectors, while continuous features were standardised. These two blocks were then combined into a sparse high-dimensional feature matrix. Finally, UMAP with the cosine metric ($n_{\text{neighbors}} = 20$, min_dist $= 0.1$ for unoptimsed dataset) and ($n_{\text{neighbors}} = 20$, min_dist $= 0.1$ for geometry-optimsed dataset) were applied to project this feature space into two dimensions.

4.5 Text-Mining Experimental Synthetic Conditions

To map crystal structures to their synthetic conditions, we created a module to automate the extraction of synthesis conditions from journal articles. During extraction of crystal structures from the CSD, we also extracted the Digital Object Identifier(DOI) of the journal articles that describes the synthesis and characterisation of the structure. We then used Puppeteer to read HyperText Markup Language, HTML files of open access journals and paid journals from our various institutions. [18]

An overview of the synthesis condition extraction is illustrated in the in Fig. 1. As illustrated in Fig. 1, we implemented a spaCy module that reads an HTML file and return a list of plain texts, wherein each item in the list corresponds to different paragraphs. We then implemented a machine learning model to predict paragraphs describing experimental synthesis. A series of regular expressions were then implemented to extract reaction time, temperature, synthetic method and the

exact quantities used for organic reagents, metal salts and solvents. The module can be downloaded from **https://github.com/bafgreat/mofsyncondition.git** and a detailed explanation of the procedure is provided in section **S-3** of the ESI.

4.6 Metal Salt Prediction Graph Neural Model

GNNs [31–40] have emerged as a powerful paradigm for learning directly from graph-structured data, where entities are represented as nodes and their relationships as edges. In the context of crystalline materials and MOFs, GNNs are particularly well-suited because they can exploit the natural graph representation of periodic atomic structures: atoms act as nodes, chemical bonds define edges, and lattice descriptors encode long-range periodicity [41–43]. Unlike traditional feature engineering, which relies on handcrafted descriptors, GNNs can automatically integrate both local coordination environments and global structural information [44, 45], making them attractive for predicting synthetic precursors such as metal salts.

4.6.1 Model architecture

Our predictive framework encodes each MOF as an undirected attributed graph. Node features include atomic identities and local environments, while edge features represent interatomic interactions. To capture periodicity, lattice parameters are explicitly incorporated through a separate encoder [46]. In addition, we augment the representation with global synthesis-aware descriptors extracted using `mofstructure`: atomic composition one-hot vectors, metal coordination environments, crystal system encodings, space group symmetries, and open-metal-site (OMS) indicators. The core model is a multi-branch GNN. A stack of GINE convolutional layers [47] processes the structural graph, producing graph-level embeddings via global mean pooling. A formal mathematical definition of the

message-passing framework and the GINE update rule is provided in **S-8**. In parallel, a multilayer perceptron (MLP) encodes the lattice matrix, while additional lightweight MLP modules transform the global descriptors. All embeddings are concatenated into a joint latent representation and passed through a final MLP classifier that outputs class probabilities over all possible salts. Dropout layers [48] and batch normalization [49] are applied throughout to improve generalization [50].

4.6.2 Feature ablation and integration

To quantify the role of different structural descriptors, we implemented an ablation protocol in which global features were added incrementally. Starting from the baseline (graph + lattice only), we evaluated the predictive power of each descriptor separately (atomic composition, coordination number, crystal system, OMS, and space group), followed by progressively richer combinations. This systematic evaluation (see Table 1 demonstrates that elemental composition consistently yields the largest accuracy gains, while other descriptors provide incremental improvements. The best-performing configuration integrated composition, coordination number, OMS, and space group, achieving 68.5% Top-3 accuracy and 78.0% Top-5 accuracy, substantially outperforming the baseline.

4.6.3 Training protocol

All models were implemented in PyTorch Geometric [51]. The dataset was split into training/validation/test sets (70%/15%/15%), and experiments were repeated across five random seeds. Models were trained with Adam optimizer (learning rate 10^{-3}, weight decay 10^{-4}), cross-entropy loss, and early stopping based on validation accuracy with a patience of 50 epochs. Hyperparameters such as hidden dimension (64-128), dropout rate (0.2-0.35), and MLP depth were tuned via grid search. This ensured that performance differences reflected the utility of struc-

tural descriptors rather than optimisation artifacts. Training was conducted on a workstation equipped with an Intel® Core i9-14900HX CPU (32 threads, 5.8 GHz max clock) and an NVIDIA GeForce RTX 4070 GPU.

4.7 Experimental synthetic condition

$ZrOCl_2 \cdot 8H_2O$ (97 mg, 0.03 mmol) and 2,5-dibromoterephthalic acid (Br_2BDC) (97 mg, 0.03 mmol) were mixed with 4 mL of glacial acetic acid and 10 mL of *N,N*-dimethylformamide (DMF) in a vial and sonicated for 10 minutes. The resulting solution was transferred to a Teflon-lined autoclave and heated at 120 °C for 24 hours. A white, cloudy precipitate was obtained, washed with DMF, and air-dried.

The organic linker predicted in `qmof-6031bc0` is 2,5-chloroterephthalate. However, this ligand was not available in our laboratory but the brominated analogue, 2,5-dibromoterephthalate was availabe hence we instead used the brominated analogue for the synthesis of the MOF. Predictions for both linker variants are provided in**S-9** of the ESI.

Powder X-ray Diffraction (PXRD): PXRD measurements were performed in transmission mode using a Stoe StadiP diffractometer equipped with a Cu Kα radiation source ($\lambda = 1.54186$ Å) and a Mythen photodetector. The powder sample was mounted on Scotch tape for analysis. Data acquisition and processing were carried out using the `WinXPOW` software package.

Infrared (IR) Spectroscopy: FTIR spectra were recorded using a Thermo Scientific Nicolet iS50 spectrometer. MOF powder was placed directly on a diamond ATR (attenuated total reflectance) crystal for measurement. Data acquisition and

processing were performed using the OMNIC software suite.

5 Data availability

The data and codes from this study have been carefully compiled and can be searched, downloaded and installed from the following repositories.

5.1 GitHub

All the code implemented in this study can be installed from the following GitHub links. They have been well documented with a very easy installation and usability guide.

1. The MOF deconstruction, porosity, open-metal site and structural analysis **mofstructure** can be downloaded from: **https://github.com/bafgreat/mofstructure.git**
2. The synthesis prediction **fairmofsyncondition** can be downloaded from: **https://github.com/bafgreat/fairmofsyncondition.git**
3. The text mining tool **mofsyncondition** for extracting synthetic conditions can be downloaded from: **https://github.com/bafgreat/mofsyncondition.git**

mofstructure and **fairmofsyncondition** can be pip installed from **mofstructure** and **fairmofsyncondition**. They all have commandline arguments which can facilitate quick usage.

5.2 Search app

We implemented a search app where users can query MOFs using all possible keys words. **cheminteraction.com**

5.3 Zenodo

To enable rapid and seamless use of this data for further computation and data analysis, we have compiled various useful data and uploaded on Zenodo.

1. The cif files for experimental and geometry-optimised structures can downloaded from:
2. The experimental synthetic conditions can all be downloaded from:
3. All building units and geometric properties can be downloaded from:
4. The machine learning model for identifying experimental paragraphs in journal articles can be downloaded from
5. All processed data used for analysis can be downloaded from:

5.4 NOMAD

Both inputs and out of all geometry optimisations were compiled and uploaded into NOMAD from which we created the MOF use case and the MOF dataset. We also implemented a schema to host all experimental synthetic conditions. The data can be search in the following links

1. MOF use case: **https://nomad-lab.eu/prod/v1/gui/search/mofs**.
2. MOF dataset: **https://dx.doi.org/10.17172/NOMAD/2023.11.17-2**

5.5 Command line Tools

CML tools for mofstructure

```
mofstructure MOF.cif
```

Deconstruct MOFs into building units, compute cheminformatic identifiers, porosity metrics, and open metal sites (OMS)

```
mofstructure_building_units MOF.cif
```

Deconstruct MOFs into building units and compute cheminformatic identifiers

```
mofstructure_database cif_folders
```

Batch process: deconstruct MOFs, compute cheminformatic identifiers, porosity metrics, and OMS for a folder of CIF files

```
mofstructure_oms File/folder
```

Compute OMS for a CIF file or folder of porous materials containing metals, e.g., MOFs or zeolites

```
mofstructure_curate File/folder
```

Remove unbound guest molecules and provide a report on system integrity; works for MOFs, COFs, and zeolites; accepts single files or folders

```
mofstructure_porosity File/folder
```

Compute geometric properties of porous materials (MOFs, COFs, zeolites) and work on single files or folders).

CML tools for fairmofsyncondition

```
fairmofsyncondition_syncon my_mof.cif
```

Predict synthesis conditions for a given MOF structure

```
iupac2cheminfor 'water'
```

Convert IUPAC name to InChIKey

```
cheminfo2iupac -n 'O' -o filename
```

Convert SMILES or InChIKey to IUPAC name

```
struct2iupac H2O.xyz
```

extract InChIKey, SMILES, and IUPAC name from a 3D molecular structure; works only for molecules, not periodic structures

5.6 Survey

The survey on what people mean when they use terms like "several days", "few days" is found in the following link **https://forms.gle/wJETchYFMu1vzzQV8**. We also encourage readers to participate in this survey.

6 Acknowledgment

ADADW acknowledges funding from the European Union Horizon research and innovation programme under the the Marie Skłodowska-Curie grant agreement (No. 101107360). The authors also acknowledge the Deutsche Forschungsgemeinschaft (DFG, Germany) for funding through the NFDI consortium FAIRmat, project 460197019. ADADW gratefully acknowledges the computing time made available for this project on the high-performance computer Noctua2 at the NHR Centre Paderborn Center for Parallel Computing (PC2). This centre is jointly supported by the Federal Ministry of Research, Technology, and Space and the state governments participating in the National High-Performance Computing (NHR) joint funding programme. ADADW also acknowledges the computing time made available on the high-performance computer at the NHR Centre of TU Dresden. This

centre is jointly supported by the Federal Ministry of Education and Research and the state governments participating in the NHR (www.nhr-verein.de/unsere-partner). Via our membership of the UK's HEC Materials Chemistry Consortium, which is funded by EPSRC (EP/L000202), this work used the UK Materials and Molecular Modelling Hub for computational resources, MMM Hub, which is partially funded by EPSRC (EP/T022213/1, EP/W032260/1 and EP/P020194/1)

7 Author Contributions

A.D.D. Wonanke conceived and designed the project, developed the core methodology, implemented the data-curation and text-mining pipelines, designed and contributed to training the graph-neural network models, and wrote the manuscript.

A. Longa and **P. Lio** contributed to the implementated and optimisation of the graph-neural network models for metal-salt prediction and fully wrote the ML section of the manuscript.

T. Heine, **M.A. Addicoat**, **D. Crittenden**, and C. Wöll contributed to project design, FAIR-data architecture, and overall supervision.

L. Himanen, **A.N. Ladines**, **J.A. Márquez**, and **M. Scheidgen** contributed to the FAIR design framework and integration of the dataset within the NOMAD infrastructure.

A. Pankajakshan and **S. Dehnen** performed the experimental synthesis and characterisation of the MOF.

All authors discussed the results, provided critical feedback and contributed to the final manuscript.

8 Competing Interests

The authors declare no competing interests.

References

[1] P. G. Boyd, Y. Lee, B. Smit, *Nature Reviews Materials* **2017**, *2*, 17037.

[2] P. G. Boyd, A. Chidambaram, E. García-Díez, C. P. Ireland, T. D. Daff, R. Bounds, A. Gładysiak, P. Schouwink, S. M. Moosavi, M. M. Maroto-Valer, J. A. Reimer, J. A. R. Navarro, T. K. Woo, S. Garcia, K. C. Stylianou, B. Smit, *Nature* **2019**, *576*, 253–256.

[3] D. A. Gómez-Gualdrón, Y. J. Colón, X. Zhang, T. C. Wang, Y.-S. Chen, J. T. Hupp, T. Yildirim, O. K. Farha, J. Zhang, R. Q. Snurr, *Energy Environ. Sci.* **2016**, *9*, 3279–3289.

[4] C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp, R. Q. Snurr, *Nature Chemistry* **2012**, *4*, 83–89.

[5] S. Lee, B. Kim, H. Cho, H. Lee, S. Y. Lee, E. S. Cho, J. Kim, *ACS Applied Materials & Interfaces* **2021**, *13*, 23647–23654.

[6] P. G. Boyd, T. K. Woo, *CrystEngComm* **2016**, *18*, 3777–3792.

[7] S. M. Moosavi, A. Nandy, K. M. Jablonka, D. Ongari, J. P. Janet, P. G. Boyd, Y. Lee, B. Smit, H. J. Kulik, *Nature Communications* **2020**, *11*, 4068.

[8] S. Majumdar, S. M. Moosavi, K. M. Jablonka, D. Ongari, B. Smit, *ACS Applied Materials & Interfaces* **2021**, *13*, 61004–61014.

[9] A. Nandy, S. Yue, C. Oh, C. Duan, G. G. Terrones, Y. G. Chung, H. J. Kulik, *Matter* **2023**, *6*, 1585–1603.

[10] P. Z. Moghadam, A. Li, S. B. Wiggin, A. Tao, A. G. P. Maloney, P. A. Wood, S. C. Ward, D. Fairen-Jimenez, *Chemistry of Materials* **2017**, *29*, 2618–2625.

[11] G. Zhao, L. M. Brabson, S. Chheda, J. Huang, H. Kim, K. Liu, K. Mochida, T. D. Pham, Prerna, G. G. Terrones, S. Yoon, L. Zoubritzky, F.-X. Coudert, M. Haranczyk, H. J. Kulik, S. M. Moosavi, D. S. Sholl, J. I. Siepmann, R. Q. Snurr, Y. G. Chung, *Matter* **2025**, *8*.

[12] Y. G. Chung, J. Camp, M. Haranczyk, B. J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O. K. Farha, D. S. Sholl, R. Q. Snurr, *Chemistry of Materials* **2014**, *26*, 6185–6192.

[13] 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, R. Q. Snurr, *Journal of Chemical & Engineering Data* **2019**, *64*, 5985–5998.

[14] A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein, R. Q. Snurr, *Matter* **2021**, *4*, 1578–1597.

[15] M. Gibaldi, A. Kapeliukha, A. White, J. Luo, R. A. Mayo, J. Burner, T. K. Woo, *Chem. Sci.* **2025**, *16*, 4085–4100.

[16] X. Jin, K. M. Jablonka, E. Moubarak, Y. Li, B. Smit, *Digital Discovery* **2025**, *4*, 1560–1569.

[17] K. M. Jablonka, A. S. Rosen, A. S. Krishnapriyan, B. Smit, *ACS Central Science* **2023**, *9*, 563–581.

[18] Y. Luo, S. Bag, O. Zaremba, A. Cierpka, J. Andreo, S. Wuttke, P. Friederich, M. Tsotsalas, *Angewandte Chemie International Edition* **2022**, *61*, e202200242.

[19] L. T. Glasby, K. Gubsch, R. Bence, R. Oktavian, K. Isoko, S. M. Moosavi, J. L. Cordiner, J. C. Cole, P. Z. Moghadam, *Chemistry of Materials* **2023**, *35*, 4510–4524.

[20] Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes, O. M. Yaghi, *Journal of the American Chemical Society* **2023**, *145*, 18048–18062.

[21] S. Park, B. Kim, S. Choi, P. G. Boyd, B. Smit, J. Kim, *Journal of Chemical Information and Modeling* **2018**, *58*, 244–251.

[22] A. Nandy, G. Terrones, N. Arunachalam, C. Duan, D. W. Kastner, H. J. Kulik, *Scientific Data* **2022**, *9*, 74.

[23] T. Gupta, M. Zaki, N. M. A. Krishnan, Mausam, *npj Computational Materials* **2022**, *8*, 102.

[24] Z. Wang, O. Kononova, K. Cruse, T. He, H. Huo, Y. Fei, Y. Zeng, Y. Sun, Z. Cai, W. Sun, G. Ceder, *Scientific Data* **2022**, *9*, 231.

[25] H. Park, Y. Kang, W. Choe, J. Kim, *Journal of Chemical Information and Modeling* **2022**, *62*, 1190–1198.

[26] L. Pilz, M. Koenig, M. Schwotzer, H. Gliemann, C. Wall, M. Tsotsalas, *Advanced Functional Materials* **2024**, *34*, 2404631.

[27] 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, R. Q. Snurr, *Crystal Growth & Design* **2019**, *19*, 6682–6697.

[28] Y. Lan, T. Yan, M. Tong, C. Zhong, *J. Mater. Chem. A* **2019**, *7*, 12556–12564.

[29] P. Z. Moghadam, A. Li, X.-W. Liu, R. Bueno-Perez, S.-D. Wang, S. B. Wiggin, P. A. Wood, D. Fairen-Jimenez, *Chemical Science* **2020**, *11*, 8373–8387.

[30] R. L. Martin, M. Haranczyk, *Crystal Growth & Design* **2014**, *14*, 2431–2440.

[31] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, *IEEE transactions on neural networks* **2008**, *20*, 61–80.

[32] A. Micheli, *IEEE Transactions on Neural Networks* **2009**, *20*, 498–511.

[33] T. N. Kipf, M. Welling in *International Conference on Learning Representations*. https://openreview.net/forum?id=SJU4ayYgl.

[34] W. Hamilton, Z. Ying, J. Leskovec, *Advances in neural information processing systems* **2017**, *30*.

[35] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio', Y. Bengio in *International Conference on Learning Representations*. https://openreview.net/forum?id=rJXMpikCZ.

[36] K. Xu, W. Hu, J. Leskovec, S. Jegelka in *International Conference on Learning Representations*. https://openreview.net/forum?id=ryGs6iA5Km.

[37] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, Y. Shen, *Advances in neural information processing systems* **2020**, *33*, 5812–5823.

[38] Z. Wu, P. Jain, M. Wright, A. Mirhoseini, J. E. Gonzalez, I. Stoica, *Advances in neural information processing systems* **2021**, *34*, 13266–13279.

[39] F. M. Bianchi, V. Lachi, *Advances in neural information processing systems* **2023**, *36*, 71603–71618.

[40] F. Ferrini, A. Longa, A. Passerini, M. Jaeger in *Learning on Graphs Conference*, PMLR, pp. 2–1.

[41] K. Choudhary, T. Yildirim, D. W. Siderius, A. G. Kusne, A. McDannald, D. L. Ortiz-Montalvo, *Computational Materials Science* **2022**, *210*, 111388.

[42] A. S. Rosen, V. Fung, P. Huck, C. T. O'Donnell, M. K. Horton, D. G. Truhlar, K. A. Persson, J. M. Notestein, R. Q. Snurr, *npj Computational Materials* **2022**, *8*, 112.

[43] A. Alfarraj, M. R. Alfuraidan, A. M. P. Peedikakkal, I. O. Sarumi, *Journal of Chemical Information and Modeling* **2025**.

[44] V. P. Dwivedi, A. T. Luu, T. Laurent, Y. Bengio, X. Bresson in *International Conference on Learning Representations*. `https://openreview.net/forum?id=wTTjnvGphYj`.

[45] L. Airale, A. Longa, M. Rigon, A. Passerini, R. Passerone in *Forty-second International Conference on Machine Learning*. `https://openreview.net/forum?id=t3zwUqibMq`.

[46] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl in *International conference on machine learning*, Pmlr, pp. 1263–1272.

[47] W. Hu*, B. Liu*, J. Gomes, M. Zitnik, P. Liang, V. Pande, J. Leskovec in *International Conference on Learning Representations*. `https://openreview.net/forum?id=HJlWWJSFDH`.

[48] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, *The journal of machine learning research* **2014**, *15*, 1929–1958.

[49] S. Ioffe, C. Szegedy in *International conference on machine learning*, pmlr, pp. 448–456.

[50] I. Goodfellow, Y. Bengio, A. Courville, *Deep learning* **2016**, *1*, 161–217.

[51] M. Fey, J. E. Lenssen, *arXiv preprint arXiv:1903.02428* **2019**.