Data-driven matching of experimental crystal structures and gas adsorption isotherms of metal-organic frameworks[†]

Daniele Ongari,[‡] Leopold Talirz,[‡,¶] Kevin Maik Jablonka,[‡] Daniel W. Siderius,[*,§] and Berend Smit[*,‡]

‡*Laboratory of Molecular Simulation (LSMO), Institut des Sciences et Ingénierie Chimiques, Ecole Polytechnique Fédérale de Lausanne (EPFL), Rue de l'Industrie 17, CH-1951 Sion, Valais, Switzerland*

¶*Theory and Simulation of Materials (THEOS), Faculté des Sciences et Techniques de l'Ingénieur, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland*

§*Chemical Sciences Division, National Institute of Standards and Technology, 100 Bureau Dr MS 8320, Gaithersburg, Maryland, USA 20899-8320*

E-mail: daniel.siderius@nist.gov; berend.smit@epfl.ch

Abstract

Porous metal-organic frameworks are a class of materials with great promise in gas separation and gas storage applications. Due to the high dimensional space of materials science and engineering, computational screening techniques have long been an important part of the scientific toolbox. However, a broad validation of molecular simulations

[†]Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States. Certain commercially available items may be identified in this paper. This identification does not imply recommendation by NIST, nor does it imply that it is the best available for the purposes described.

in these materials is impeded by the lack of a connection between databases of gas adsorption experiments and databases of the atomic crystal structure of corresponding materials. This work aims to connect the gas adsorption isotherms of metal-organic frameworks collected in the NIST/ARPA-E Database of Novel and Emerging Adsorbent Materials to the corresponding crystal structures in the Cambridge Structural Database. With tens of thousands of isotherms and crystal structures reported to date, an automatic approach is needed to establish this link, which we describe in this paper. As a first application and consistency check, we compare the pore volume measured from low-temperature argon or nitrogen isotherms to the geometrical pore volume computed from the crystal structure. Overall, 545 argon or nitrogen isotherms could be matched to a corresponding crystal structure. We find that the pore volume computed via the two complementary methods shows acceptable agreement only in about 35 % of these cases. We provide the subset of isotherms measured on these materials as a seed for a future, more complete reference data set for computational studies.

Introduction

Porous materials are employed in many applications, such as gas storage,[1] separations,[2] catalysis,[3,4] and sensing.[5,6] For many years, the porous materials research and development community was dominated by activated carbons and zeolites, but over the last two decades the field has expanded grown enormously thanks to the discovery of porous metal-organic frameworks (MOFs),[7] and covalent organic frameworks (COFs).[8–10] In this work, we are interested in the gas adsorption properties of MOFs: crystalline frameworks constructed from metal nodes and organic linkers. At present, crystal structures for over 10 000 different porous MOFs have been reported in the Cambridge Structural Database (CSD),[11] and many computational studies have performed molecular simulations to predict the gas adsorption properties of these materials starting from the reported crystal structure.[12]

The NIST/ARPA-E Database of Novel and Emerging Adsorbent Materials (NIST-ISODB),[13]

on the other hand, is the world's largest public collection of experimental gas adsorption isotherms. Its over 30 000 isotherms cover a wide range of adsorbent materials including not only MOFs, COFs, and zeolites, but also activated carbons and amorphous porous polymers, thus making it a treasure trove for data-driven analysis.[14,15] The database also holds great potential for the integration of data-driven approaches with physics-based models – if a link between an adsorption isotherm in the NIST-ISODB and the crystal structure of the corresponding MOF in the CSD can be established.[16]

For example, in 2017, Sholl et al. investigated differences between independent measurements of CO_2 isotherms on the same MOFs, providing guidelines on how to identify confidence thresholds, assigning ratings for consistency and reproducibility, and comparing the experimental data to simulated CO_2 isotherms.[15] The study was limited to materials for which adsorption measurements had been independently repeated by several groups: 27 MOFs with two or more independently measured CO_2 isotherms (211 isotherms in total). For a few tens of MOFs, the link between the adsorbent in the NIST-ISODB and the crystal structure in the CSD could therefore be established manually, for example by relying on the conventional names of the MOFs or on the CSD reference codes reported in the publications. The same group extended the analysis in 2020 with focus on alcohols, comparing simulations of methanol and ethanol in four MOFs with the interval of confidence obtained from experimental isotherms via their protocol.[17]

In this work, we aim to extend this link to as many gas adsorption isotherms for MOFs as possible, irrespective of the target application: how many MOF identities can we link both to experimentally resolved crystal structures (CSD) and to experimental gas adsorption isotherms (NIST-ISODB) by relying on the metadata present in the two databases? This raises a further question: how can we gauge whether the linked crystal structure is a reasonable model for the experimental sample on which the adsorption study is performed? In order to address this question, we recall that the first step of an adsorption study typically involves the characterization of the adsorbent's pore volume by recording a nitrogen

or argon adsorption isotherm at low temperature.[18] At the same time, the pore volume can also be directly computed from the crystal structure and, being foremost a geometric property,[19] carries less uncertainty related to force field parameters than, e.g., the simulation of a CO_2 adsorption isotherm.[1] These calculations assume a defect-free, infinite crystal, which is impossible to obtain experimentally. While some deviations from the experimental pore volume can therefore be expected, large deviations indicate that the crystal structure is not representative of the actual sample. We therefore propose using the pore volume as a basic consistency check.

In the following, we explore different routes to establishing the structure-isotherm match. We report on statistics, and perform the consistency check described before, comparing the computed and measured pore volumes. We discuss the different reasons that lead to mismatches, and provide a reference set of isotherms with linked crystal structure deemed suitable for comparison with molecular simulations. The reader can inspect all the steps performed within this pipeline, by browsing the Jupyter Notebooks provided in the GitHub repository `https://github.com/danieleongari/matching_isodb_csd`.

Methods

Matching isotherms with structures

Overview and inspection of the NIST-ISODB and CSD

This analysis is based on the NIST-ISODB version as available from the official GitHub repository on September 2021,[20] containing 35 482 digitised isotherms for 7386 adsorbent materials and 280 different adsorbate molecules.

The NIST-ISODB was initially conceived as a list of publications on novel adsorbents with minimal metadata. Of the 4128 indexed publications, $\approx 80\,\%$ are associated with digitized

[1] As long as the crystal remains rigid upon adsorption (which we implicitly assume here), the geometric pore volume is determined by the atomic radii of the framework atoms, which carry less uncertainty than force field potentials.

isotherms, mostly obtained from measurements, but also from fitting to experimental data or from molecular simulations. The collection of metadata on the method used to obtain isotherms started only at a later stage, first at the level of the publication and then at the level of individual isotherms. We include only isotherms that are marked themselves as "experimental" or that are contained in a reference that is marked to contain exclusively experimental isotherms. After excluding isotherms coming from simulations or models and isotherms of unknown origin (unspecified, or the reference paper is denoted to contain both), we end up with 21 375 experimental isotherms (60 % of the total 35 482).

The number of different adsorbate molecules is limited, and mapping their conventional names to their chemical formula is straightforward. The NIST-ISODB includes a mapping of different synonyms for the same gas, such as water and H_2O, to the same InChIKey that uniquely identifies the gas molecule.

As for the adsorbent materials, the NIST-ISODB includes MOFs, zeolites, activated carbons, covalent organic frameworks, and other classes of materials. Adsorbents are identified by a name, which is typically taken from the figure caption of the digitized adsorption isotherm and therefore poorly standardized. While 3.2 % and 7.0 % of the names contain the keywords "zeolite" and "carbon", respectively, classifying the remaining adsorbents based on their name is difficult to classify in an automated way from their name. We can expect that a large portion of them are MOFs due to the large quantity of different MOFs reported up to now, but the reader can refer to the work of Cai et al. for more statistics about the manual classification of a sample of 333 adsorbents from gas adsorption data indexing including the NIST-ISODB.[21]

The NIST-ISODB also includes a mapping of synonyms for adsorbents. For example, the CuBTC MOF has also been reported as Basolite C300, C300, Cu-BTC, $Cu_3(BTC)_2$, CuBTC, and MOF-199, and in the NIST-ISODB all the corresponding isotherms are assigned to the same adsorbent. We want to emphasize from the beginning that constructing such a mapping for adsorbents, however, is more complex since the conventional names for MOFs can be

ambiguous: this is a challenge that will recur over and over in this work. For example, MOFs such as MOF-74 or MIL-53 can be synthesized with different metal nodes and the NIST-ISODB record does not always specify the identity of the coordination metal (e.g., "MIL-53" instead of "MIL-53(Cr)"). Furthermore, some authors use generic names like "MOF-1" intended only to enumerate materials within the particular study. We therefore excluded ambiguous adsorbent names from the NIST-ISODB mapping of synonyms in order to avoid matching materials by synonym that do not have the same crystal structure. Interested readers can find a full list of excluded names and a detailed discussion in Section 1 of the Supporting Information.

Among the 21 375 experimental isotherms collected in the NIST-ISODB, the main adsorbates are nitrogen (5153), carbon dioxide (4366), hydrogen (2540), methane (2212) and water (607). Of particular interest for the present study are the experimental isotherms for nitrogen at 77 K (a total of 4003) and for argon at 87 K (a total of 140).

We note that the NIST-ISODB can contain multiple, different isotherms for the same publication, adsorbent, adsorbate, and temperature. Publications with two or three isotherms recorded under identical conditions often stem from the digitization of multiple figures that display the same isotherm data in different pressure ranges or compared to different adsorbate or adsorbents. Publications with three or more "duplicates" often report adsorption in adsorbents synthesized under different conditions, where the isotherm serves as a benchmark of the quality of the material (e.g., Figure 7 in Ref. 22), or to illustrate the effect of some post-synthetic treatment (e.g., Figure 4 in Ref. 23). In these cases, therefore, the name of the MOF is reported as the same for the digitization, but the sample does contain differences. Other reasons include reporting multiple adsorption-desorption cycles on the same sample. The experimental reproducibility of isotherms in the NIST-ISODB for the same gas-adsorbent pair has been discussed in the literature, e.g., by Sholl and coauthors.[15] Clearly, however, isotherms that have been measured to study the effect of different synthesis conditions[24] or post-synthetic treatments can be expected to differ from each other, unlike

measurements of samples that have been synthesized to reproduce the same material, or studies on the same sample by different groups. A notable case is the NIST inter-laboratory study of methane adsorption in zeolite Y, which reports 109 isotherms at the same conditions: different research groups were asked to independently measure the uptake at 298 K as a way of investigating the reproducibility of the measurement.[25] We postpone the filtering of isotherms recorded under apparently identical conditions to a later stage, when more information on these can inform a rational protocol for their selection.

Moving to analyse the CSD database of crystal structures, in the present study we employed version 542 (2020.3), released in February 2021. Its MOF subset includes 105 922 structures, 8034 of which are assigned one or more conventional names (such as CuBTC, MOF-5, or UiO-66). Analogous to isotherms in the NIST-ISODB, one publication may be associated with multiple variants of a MOF crystal structure with the same name in the CSD. For example, the in-situ study by Breogán Pato-Doldán et al.[26] reported 1853 crystallographic information files (CIFs) of different MOF-74 analogues at various CO_2 loading conditions. In the absence of an automated method for selecting a representative structure, this selection eventually requires manual inspection. In order to limit the effort involved, we exclude CSD entries if there are more than three other entries of the same conventional name from the same paper. The reasoning behind keeping three structures instead of just one is to be able to evaluate the uncertainty on the computed pore volume from different measurements of the crystal structures.

The following subsections describe the matching procedure, which is summarized in Figure 1.

Matching by conventional names

The most straightforward way of matching an adsorbent from the NIST-ISODB to a MOF crystal structure in the CSD is by conventional name of the MOF, i.e., when any of the synonyms related to a NIST-ISODB's adsorbent matches with any synonym on the CSD

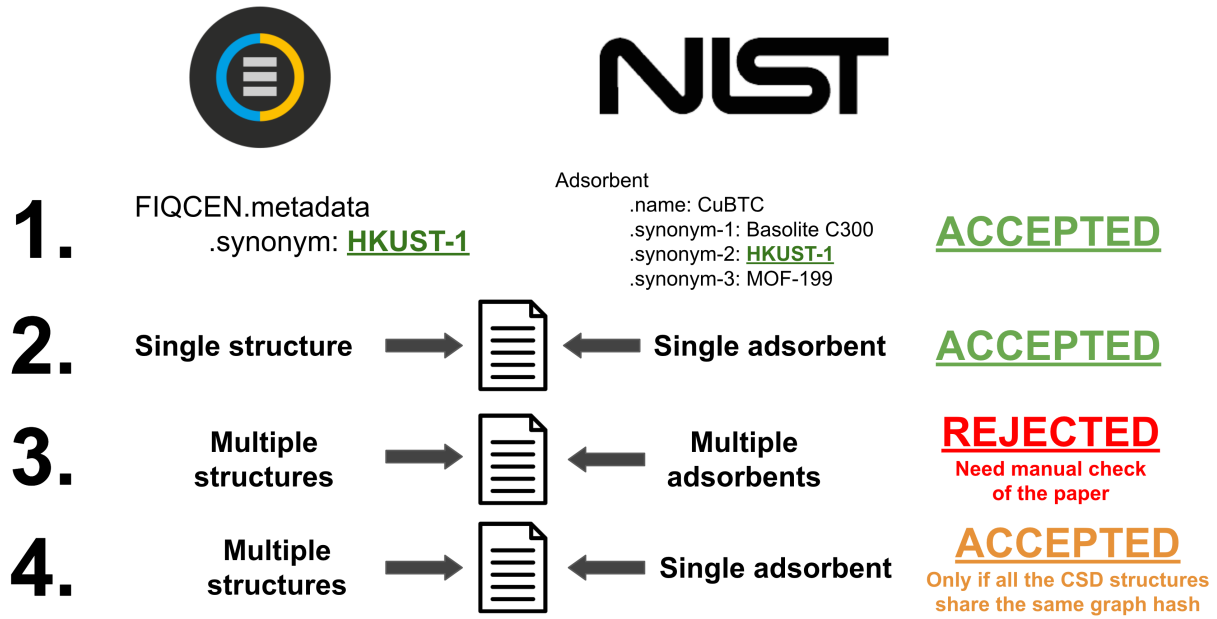

Figure 1: Matching MOF crystal structures from the CSD to adsorbents in the NIST-ISODB: First match by conventional name. The remaining three approaches attempt a match by reference to a common published article that reported both the crystal structure(s) as well as the adsorption isotherm(s).

side.

One downside of this approach lies in the low number of conventional names reported on the CSD side. In many cases, authors do not specify the conventional name used in the publication as metadata during the deposition to the CSD, meaning that a conventional name may have been used on the NIST-ISODB side that is not present in the CSD. In other cases, the publication reporting the crystal structure may not mention a conventional name either, in which case the name reported in the NIST-ISODB is typically chosen as the chemical formula of the unit cell, for example $C_{66}H_{50}B_2CoCu_6N_{18}O_{24}$ or $C_{11}H_{12}GdO_{11}$. In these cases, it is clearly harder to identify the same material in other publications, as it will likely be labelled with a conventional name or with a formula in a different format. Moreover, two MOFs can have a the same formula but different topology, giving rise to different gas adsorption properties.

Before comparing names, we perform a set of normalization steps on the name strings:

1. Convert to lower case (e.g., to match bio-MOF with Bio-MOF, and UIO with UiO)

2. Remove dashes and spaces (e.g., ZIF-8 vs ZIF 8 vs ZIF8)

3. Remove name of common adsorbate from adsorbent name (e.g., "Zn-MOF-74 CO2" becomes "Zn-MOF-74")

4. Remove "a" and ' suffixes used to label activated materials (e.g., UTSA-36a, or SNU-71')

The first normalization step led to 12 additional matches, one of which turned out to be a false positive upon manual inspection: in CD-MOF-1, "CD" stands for cyclodextrin,[27] while in Cd-MOF-1, "Cd" stands for Cadmium. While this was the only false positive, this finding suggests that case mismatches should continue to be inspected manually in future updates.

The third step becomes necessary since the conventional name of the materials in the CSD sometimes contains the name of the adsorbed gas as a suffix. We compiled a list of common adsorbates (see Supporting Information, Section 2, for the full list) and remove them from the name.

This approach provided matches for only 334 out of the 7386 NIST-ISODB adsorbents. The most frequently reported MOFs in this set are ZIF-8, UiO-67, IRMOF-1 and CuBTC (HKUST-1), reported respectively in 12, 6, 5 and 5 distinct publications. For example, CuBTC is associated (via this conventional name or any of its synonyms) with the CSD entries FIQCEN (deposited together with the original 1999 paper),[28] BOPAN, DIHVIB, DOTSOV42, and LUDLED. One should note that these are the most interesting cases for our study, as they will allow us to compare structures and isotherms measured from different studies for the same MOF. However, only a minority of the MOFs we could match by synonym are present in different articles and thus have distinct measurements (22 in total), which motivates our search for other ways to match structures with NIST-ISODB adsorbents.

Matching by DOI

A complementary approach to match an adsorbent from the NIST-ISODB to a MOF crystal structure in the CSD is to use the DOI to identify entries coming from the same publication. After converting all DOIs to lower case, we find an additional 476 matches between NIST-ISODB adsorbents and any CSD entries in the MOF-subset that were not already matched by synonym. These matches fall into three categories, depending on how many CSD entries and NIST-ISODB adsorbents are linked to the publication used for the matching.

In the first category, the publication is associated with exactly one CSD entry and one NIST-ISODB adsorbent, making it reasonable to assume a direct match. We count 319 NIST-ISODB adsorbents added to the successful matches because of this one-to-one reference.

In the second category, the reference paper is associated with one NIST-ISODB adsorbent but multiple CSD entries: 157 NIST adsorbents are mapped to 583 CSD entries.

Reasons for this include:

1. Crystal structures reported at different gas loading conditions or temperatures, with possibly a significant structural change if the MOF is particularly flexible.

2. Crystal structures reported in both the solvated and activated state, or with different solvents.

3. Crystal structures of different MOFs are reported but isotherms were measured (or digitized) only for one material, for example, if only one of the materials was porous.

The first two reasons are related to the presence of an adsorbate or solvent in the structure. Since we are interested in the properties of the underlying framework, we automatically removed adsorbates from the crystal structure as described in the next section, and then compared their atomic structure graph (see section "Comparing structure graphs"). For only 30 of 164 paper references, the structure graph of all structures was identical and we identified that structure as a match. For example Ref. 29 is associated to 3 different CSD entries of the same MOF but measured at different solvent compositions, which in this case

are reported to affect the pore opening: we will detect flexible structures at a later stage, but at this point we are interested to group structures with the same chemical identity, i.e., the same composition and topology. As for the remaining articles, the structure graph hashes (SGHs) identified different desolvated structures which were discarded for the time being, since only a manual check of the report would allow to select a representative structure to match with the NIST-ISODB adsorbent.

In the third category, the reference paper is associated with multiple NIST-ISODB adsorbents (and, usually, multiple CSD entries): we count 330 articles falling in this category, which are associated to 937 CSD entries. For example, the DOI of Ref. 30 is associated with CSD entries DANWOF and DANWUL, and with isotherms in the NIST-ISODB for two MOFs labelled MOF-235 and MOF-236. Since these names are not reported in the CSD metadata, the association between isotherm and CSD entry is lost and could only be recovered by a careful reading of the publication: it would be a substantial effort considered the many CSD structures in this category. Therefore, we decided to discard these ambiguous matches.

To recap, by matching both the conventional name and the DOI, we were able to identify the structure for 683 NIST-ISODB adsorbents, linked to 842 CSD entries: of these, as previously reported, about one half are matched by same name and the other half by one-to-one DOI match.

CSD structure analysis

Comparing structure graphs

The CSD contains a sizeable number of identical crystal structures reported under different reference codes.[31,32] In the following, we describe how to flag these based on comparing their atomic structure graphs.

Before performing a gas adsorption study, it is common practice to activate the adsorbent, resulting in a more porous structure as solvent molecules are removed. In the CSD,

structures are often reported with solvent molecules still present inside the pores. Different research groups may synthesize the same material with different solvents or report crystal structures at different activation stages. For this reason, we computationally removed all the free and coordinated solvents using the algorithm provided with the original release of the CSD MOF subset.[33] In particular, we only removed solvent molecules listed on a list of common molecules provided by the CSD API. This avoids the removal of (necessary) charge-counterbalancing ions, or even the removal of some parts of the crystal structure.[34]

After removing the solvent, we compute the primitive cell of the crystal structure using pymatgen[35] and spglib[36] (with a symmetry tolerance of 0.1). We then use the VESTA[37] cutoffs for bond distances to construct a structure graph (see Figure 2), in which every atom is a node and the edges are the bonds inferred using the VESTA cutoff heuristic.

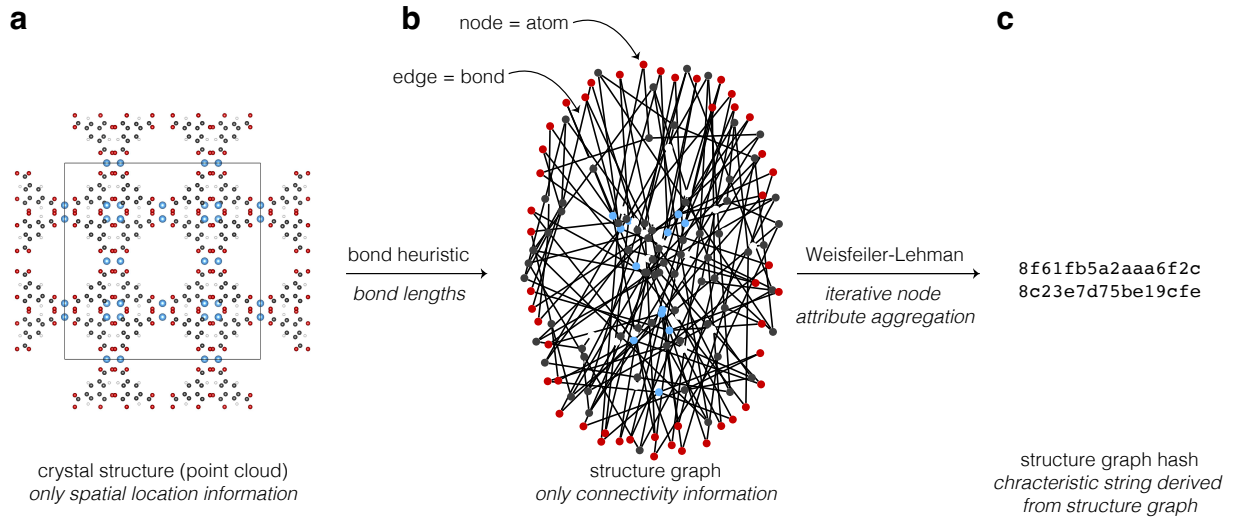

Figure 2: From crystal structure (a) to structure graph hash (c). The crystal structure specifies the location of atoms in three dimensional space. First, a periodic structure graph (b) is created based on heuristic bond length thresholds. This graph no longer contains information about the positions but only about the connectivity (shown as black edges) between the atoms (shown as blue nodes), thus making it robust against small changes in the atom positions or cell dimensions. Second, we use the Weisfeiler-Lehman algorithm to aggregate information about the neighborhood of each node and compute a characteristic fixed-length hash of the graph (more information in Figure S6 of the Supporting Information). Comparing two graphs for identity then reduces to comparing their structure graph hash.

Given the structure graph, we then compute the Weisfeiler-Lehman[38] graph hash of the

undirected structure graph using networkx,[39] using the atom types as node labels. The structure graph hash (SGH) of two crystal structures should be identical if and only if the bond network of the two structures is identical, allowing us to identify duplicates simply by comparing their SGH.

We also compute the hash of the undecorated graphs (ignoring atom types), which allows us to identify structures that only differ by the metal (e.g., Co-MOF-74 vs Ni-MOF-74).

Consistency checks on names and crystal structures

Matching CSD structures via their SGH provides a complementary method to matching them via other metadata, such as the DOI of the publication or their conventional name. This provides us with an opportunity to perform a consistency check.

First, we investigate those CSD entries that include a conventional name but were not on the list of synonyms of the NIST-ISODB and therefore could only be matched by DOI. It is instructive to list the classes of reasons for why no match was established:

- One name explicitly reports the metal, the other does not (e.g., DUT-49(Cu) vs DUT-49)
- The conventional name used for a material is not yet known as a synonym in the NIST-ISODB (e.g., UHM-30 vs $Cu_3(NH_2btc)_2$)
- The CSD reports a the conventional name, while the chemical formula is used in the NIST-ISODB (e.g., $C_{26}H_8Cu_2N_2O_{12}$ and SNU-50)

We note that these issues could be addressed on the CSD side by adopting stricter rules for the reporting of conventional names, as well as by expanding the list of known synonyms in the NIST-ISODB.

As a second consistency check, we took all the CSD entries linked to a given NIST-ISODB adsorbent and checked whether the SGH of all structures was identical. Manual inspection revealed that in a minority of cases the same MOF name was indeed used to

identify different structures in independent reports: this is usually the case for generic names chosen for enumeration purposes within the publication (e.g., MOF-1, Cd-MOF-1, PCP-1/2/3). More elaborate names can clash as well, however: for example the name "ZJU-21" was used to identify both a Cu-based MOF in 2014[40] and a Zn-based MOF in 2016.[41] Both reports are from authors affiliated with Zhejiang University (acronym "ZJU") but do not share any co-author. However, in most of the cases, different SGHs point to slight differences in the reporting the crystal structure of the "same material": crystal structures with/without disorder in the framework, disorder in the solvent which therefore was not recognised as a known solvent molecule and not removed by the "computational activation", or the presence/absence of hydrogen atoms in the reported CIF.

In response, we removed CSD entries that can not be uniquely matched with a NIST-ISODB adsorbent as well as entries with overlapping atoms. This operation was done manually, and the exclusions are tracked in the GitHub repository associated to this project.[42]

Finally, we compared the SGH of all structures once more, but with a different purpose: to identify cases where the SGH of two CSD structure was the same but the name was different. For example, only experienced scientists in the MOF field are likely to recognize MAF-4 as a synonym for ZIF-8.:[43] this analysis allowed us to add this new synonym to the NIST-ISODB adsorbent definition.

Informed by the previous visual inspection, we run a further check on the crystal structures: remember that the most of them do not have a second structure to compare for the verification. We further removed CSD structures with atomic overlaps (38) and lone molecules (80, possibly disordered or unrecognized solvent). We also checked for the presence of hydrogen atoms (in certain cases not explicitly included in the crystal structures deposited at the CSD) but given the low impact of hydrogen in the calculation of the internal volume and the weight of the crystal, we did not exclude these 47 defective structures files.

After the data cleaning of this section, we are left with 569 NIST-ISODB adsorbents,

matched with 666 CSD entries.

Pore volume comparison

A key motivation of this study is that the connection between an isotherm and a crystal structure enables the comparison to predictions from molecular simulations. Molecular simulations typically represent the adsorbent as an infinite perfect crystal, while experimental samples may include defects, amorphous regions or regions where the sample is only partly activated. Comparing the experimental to the theoretical pore volume thus provides a first consistency check for the periodic crystal representation.

Experimentally, the pore volume is routinely determined from the adsorption isotherm for nitrogen at 77 K or argon at 87 K using the Gurvich Rule.[18] Therefore, a large number of these low-temperature characterization isotherms are available in the NIST-ISODB: for 291 of the 569 matched adsorbents at least one characterization isotherm is reported. This ratio is slightly higher than in the NIST-ISODB overall[2] but it still forces us to exclude a substantial number of MOFs, and for the future of the NIST-ISODB we suggest placing additional focus on providing these key characterization data in digital form.

Rather than relying on the experimental pore volume reported by the authors, we can now use the characterization isotherms in order to consistently apply the same methodology for computing the pore volume across all structures. Figure 3 shows the application of our methodology.

We extract the experimental pore volume from the average uptake (of nitrogen at 77 K or argon at 87 K) in the 0.6 bar to 0.8 bar range. Only a few characterization isotherms (29 over 860) were excluded because no pressure points were falling within this range. The Gurvich rule states that the density of the saturated nitrogen (or argon) in the pores is equal to its liquid density ($\rho_{N_2}^{\text{liq}}$; equal to 28.83 mol L^{-1} and 34.98 mol L^{-1} for nitrogen and argon,

[2] 36.2% of adsorbents in the NIST-ISODB are associated with nitrogen or argon characterization isotherms.

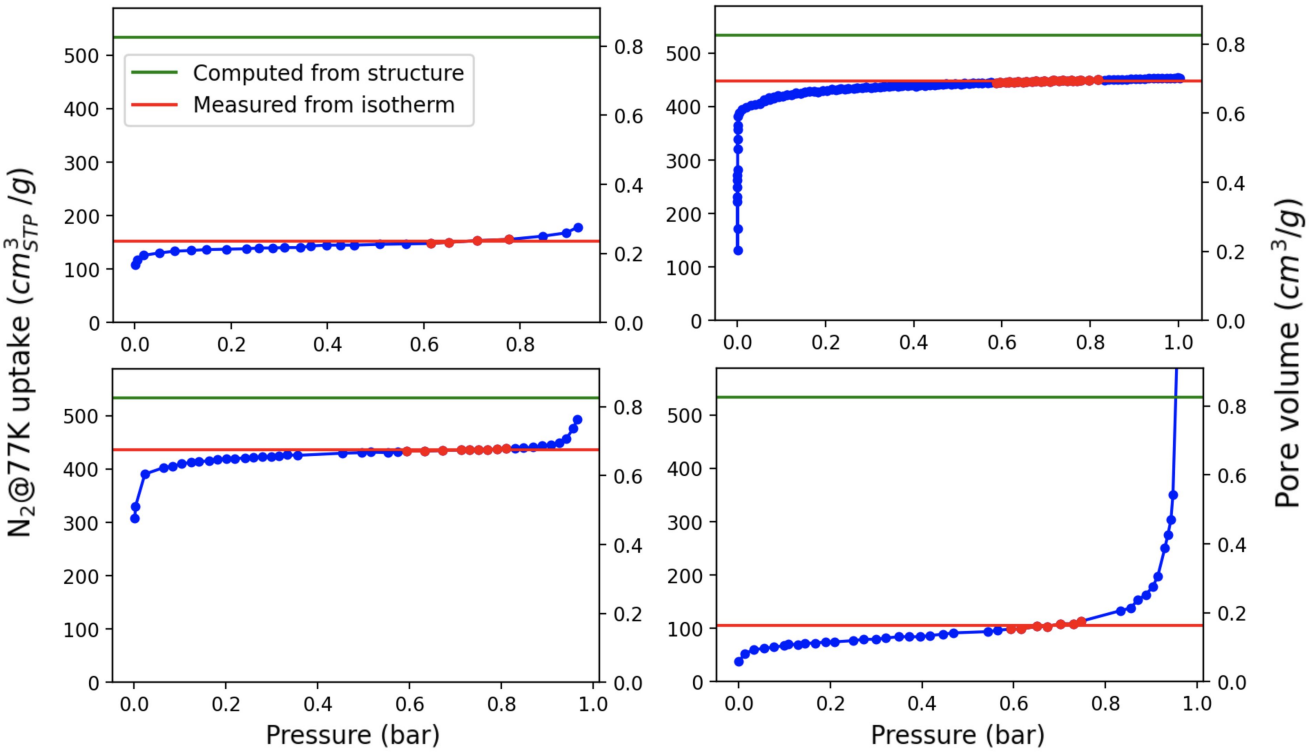

Figure 3: Comparison of measured and computed pore volume for the NIST-ISODB adsorbent CuBTC from four different isotherms. The average nitrogen uptake in the 0.6 bar to 0.8 bar range (red markers) is used to measure the experimental gravimetric pore volume (red line), and it is compared to the geometric pore volume computed from the crystal structures (green line): all the CuBTC structures we previously found lead to a similar pore volume of 0.81–0.83 $\mathrm{cm^3\,g^{-1}}$. Note how relating the two calculations of the pore volume, immediately gives an idea on the quality of the sample for which the isotherm was measured.

respectively), regardless of the shape of the internal void network or the chemistry of the crystal structure.[44] [3]

Under this assumption, the pore volume (v_{pore}) of the adsorbent is computed as:

$$v_{\mathrm{pore}} = \frac{n_{\mathrm{N_2}}^{\mathrm{ads,sat}}}{\rho_{\mathrm{N_2}}^{\mathrm{liq}}}, \tag{1}$$

where the adsorbate uptake $n_{\mathrm{N_2}}^{\mathrm{ads,sat}}$ is converted to units coherent with $\rho_{\mathrm{N_2}}^{\mathrm{liq}}$.

A large majority of characterization isotherms report the adsorbate uptake in $\mathrm{cm^3(STP)/g}$, thus yielding the gravimetric pore volume. In cases where the isotherm is reported in vol-

[3]While widely applicable to MOFs,[1,19] exceptions to the Gurvich rule can occur in narrow micropores or channels, for example in the case of commensurate adsorption.[45]

umetric units for the adsorbent instead, the adsorbent density (which is not recorded in the NIST-ISODB) would be needed to obtain the gravimetric pore volume. While the density could be computed from the linked crystal structure, we decided to exclude these cases (less than 1 % of the characterization isotherms finally selected) in order to avoid additional methodological uncertainty as the authors may have used a different value for the density of the material than the one we would compute.

On the other hand, the pore volume can be computed from the crystal structure.[19] Here, we choose the geometrical pore volume, which is an upper limit to the probe accessible pore volume. It is intuitively defined as all volume inside the unit cell that is not occupied by the atoms of the framework (described as hard spheres with Bondi's van der Waals radii).[46] We note that certain pore pockets in adsorbents can be inaccessible to the adsorbing molecule due to narrow connection channels, which is not reflected in the geometric pore volume. However, which pores are inaccessible computationally can be highly sensitive to parameters such as the kinetic radius of the molecule, its diffusion kinetics, the atomic radii used to model the framework, and the assumption of a rigid crystal (i.e., no "saloon-door" effect[47]). In the Supporting Information we compare different definitions of the pore volume, and conclude that their choice has negligible impact on the final statistics obtained in this work, leading us to prefer the geometric pore volume definition.

We compute the geometric pore volume using Zeo++[48] from the experimental structures retrieved from the CSD database after computational desolvation via the CSD Python API.[49] If more than one crystal structure matches a given NIST-ISODB adsorbent, we select the crystal structure with the largest geometric pore volume as a reference. Supplementary Figure SI-1 reports the geometric pore volumes for the structures of those adsorbents associated with more than one CSD entry. The difference in pore volume is often small, i.e., below 10 % in more than 80 % of the cases.

When multiple characterization isotherms for a given adsorbent are available *from the same paper*, it is tempting to attribute these differences to experimental uncertainty or

inaccuracies of the digitization. However, inspection of some of these articles reveals that in most cases such isotherms are used for the characterization of different synthesis or activation attempts, and the isotherm with the *maximum* pore volume typically corresponds to the optimal procedure. We acknowledge that the maximum pore volume does not unequivocally imply optimal crystallinity – for example, the presence of defects can also lead to higher pore volumes[50] – however, recognizing such cases goes beyond the scope of our automated comparison. For the sake of consistency, in cases of multiple isotherms, we therefore selected the one giving the maximum pore volume and discarded all others from the same article.

Results and discussion

Having established the link between adsorbents, crystal structures and experimental isotherms for 569 MOFs, we can analyse how the measurement and calculations compare for the same material. Figure 4 compares the measured pore volumes as calculated from nitrogen and argon isotherms to the geometric pore volume computed from the crystal structures.

Figure 5 shows the same data in the form of a histogram of the ratio between measured and geometric pore volume. As the geometric pore volume will overestimate the measured value,[19] and considering some uncertainty, one would expect those materials that are fully activated and nearly fully crystalline to fall in the 0.75–1.1 range for this measured/computed ratio. It is encouraging to see that we observe a peak in this range, accounting for ≈35 % of the measured pore volumes. However, the majority of materials falls outside this range. It is interesting to investigate these materials in more detail, dividing them into three rough categories.

The first category are ratios close to zero (i.e. < 0.1 ratio), which account for ≈10 % of the samples. These measurements report negligible or no uptake of nitrogen or argon. Figure 5 shows that most of the materials close to the x-axis (near-zero measured pore volume) also have a below-average geometric pore volume in the range of 0.25–0.5 $\mathrm{cm^3\,g^{-1}}$. With such a

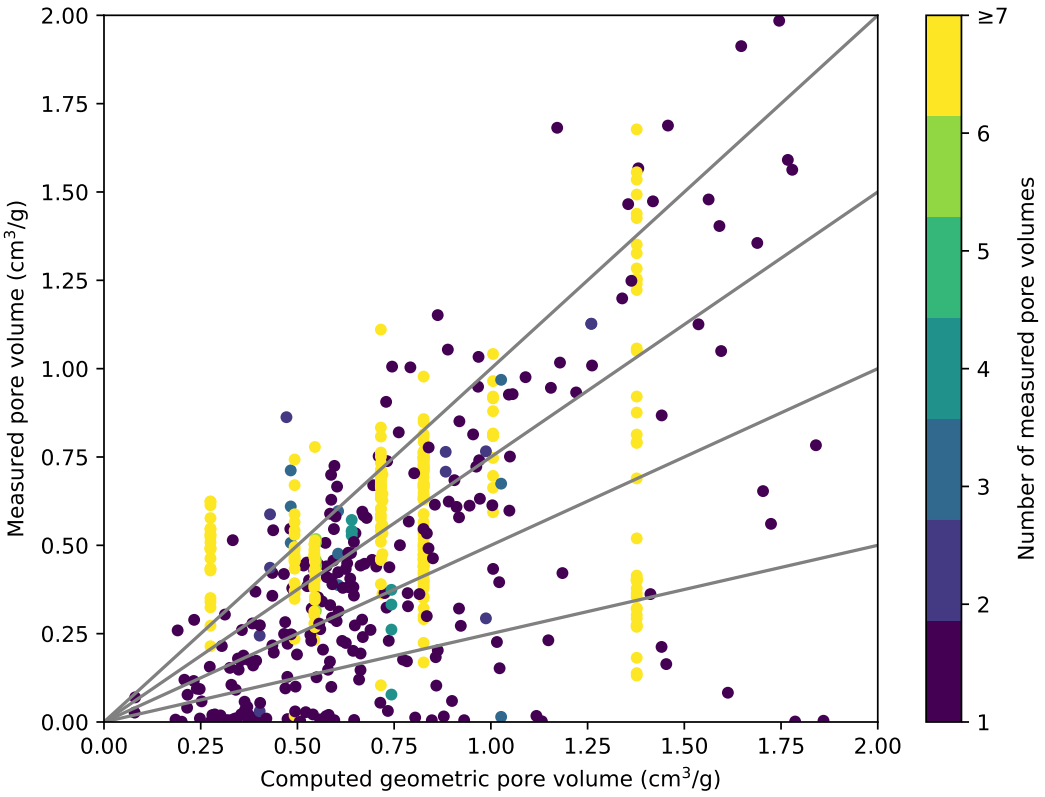

Figure 4: Comparison of geometric pore volume computed from the crystal structure to measured pore volume obtained from nitrogen isotherms at 77 K or argon isotherms at 87 K. If more than one crystal structure is available for the same material, the one with the largest geometric pore volume is used as a reference. The color scale indicates the number of papers containing characterization isotherms for a given material, from one (dark blue) to 8 and more (yellow). The most reported MOFs, with vertically aligned yellow markers, are MIL-53, UIO-66, MOF-74, ZIF-8, CuBTC, MIL-100 and IRMOF-1 in order of increasing computed volume (see Table 1 for more details). Grey lines indicate a ratio of measured to computed pore volume of 100 %, 75 %, 50 %, and 25 %.

small pore volume computed from the crystal structure, one could suspect small interstices where the nitrogen (or argon) may not fit or permeate.

To investigate this further, in Figure 6 we zoom into the structures with a measured pore volume below 0.1 $cm^3\,g^{-1}$ and compare them to the pore-limiting diameter computed for the structure.

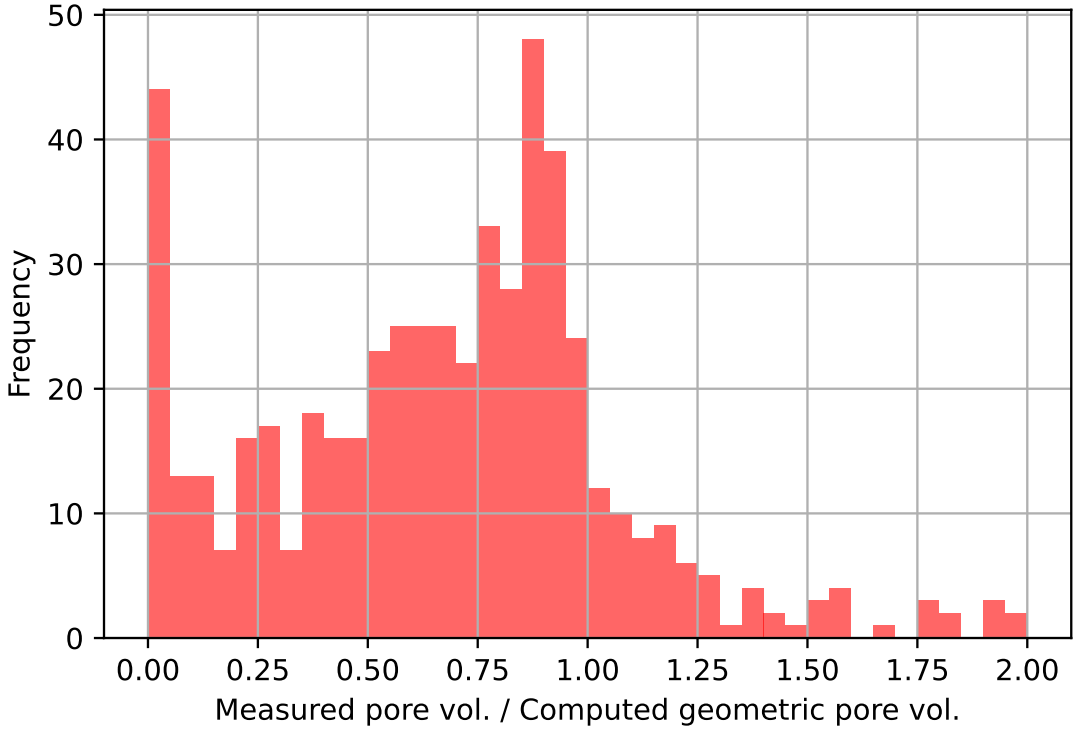

Figure 5: Histogram of the ratio between the measured pore volume and the geometric pore volume. A total of ca. 1000 values have been sampled in this graph.

The pore-limiting diameter is the diameter of the largest molecule that can diffuse through the structure, and Figure 6(b) plots the experimental uptake as a function of the pore-limiting diameter. For structures with a pore-limiting diameter below the kinetic diameter of nitrogen or argon, zero uptake is expected, and these MOFs can be reasonably considered as non-porous. When the pore-limiting diameter is close to the size of the adsorbate, kinetics of diffusion can still be very slow, leading to negligible uptake at low temperature. However, we notice a significant fraction of structures with pore-limiting diameter much larger than the kinetic diameter of nitrogen. In these cases, the measurement may have been conducted on a material that collapsed after solvent removal, solvent removal may have been unsuccessful, or the presence of floating counter-ions may not have been reported in the crystal structure. Another possible explanation is that the synthesis of the porous material was unsuccessful, and the authors reported the nitrogen isotherm for the nonporous sample to document a

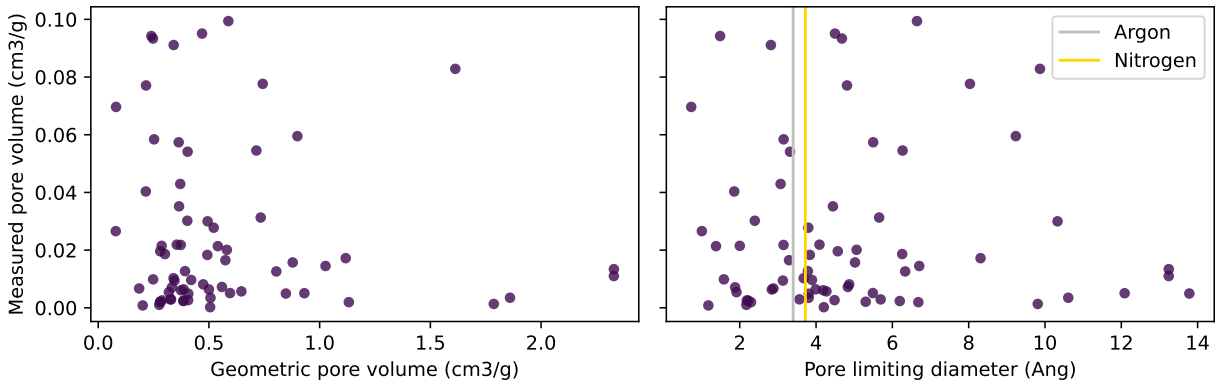

Figure 6: Comparison of near-zero measured pore volumes ($< 0.1\,\mathrm{cm^3\,g^{-1}}$) versus geometric pore volume and pore limiting diameter of the corresponding structures. The kinetic diameters of argon and nitrogen are shown as vertical lines in the graph on the right.

failed attempt.

The second category includes the ratios between 0.1 and 0.75, for which the measured pore volume is substantially lower than the geometric pore volume. Some of the hypotheses mentioned above still apply: partial desolvation, presence of counterions, presence of unreacted ligands trapped in the pores, or partially collapsing (flexible) structures upon activation. We note that that the presence of a guest molecule or a partial collapse of the structure not only reduces the available space for the nitrogen/argon probe molecule but also increases the apparent density of the material: it affects both the numerator and the denominator in the measurement of the pore volume.

The third category includes the pore volume measurements that exceed the geometric value, ≈11.7 % of the total. Possible hypotheses include an error in the structure-isotherm match or the crystal structure itself, a significant presence of defects in the crystal (e.g., missing ligands), or an unreliable measurement due to significant uptake on the surface of the crystal (e.g., small and packed crystals or jagged surfaces that create mesoporous interstices for probe molecules to adsorb outside the bulk of the crystal). MOFs with very strong bonds between nodes and ligands are known to display higher percentages of missing ligands: typical examples being UiO-66 and MOFs constructed from the Zr_6O_8 secondary building unit.[51]

We emphasize that the possible explanations of the observed deviation listed above are hypotheses based on our experience and the inspection of individual cases in this work. In particular, this study has allowed us to identify materials for which characterization isotherms are reported by several independent studies and to compare them. The eight materials with the highest number of characterization isotherms (yellow markers in Figure 3) are listed in Table 1. Histograms for the pore volume for each individual material are shown in Figure 7, and the full set of nitrogen isotherms are plotted in Figure 8.

Table 1: MOFs with the highest number of characterization isotherms available. If more than one crystal structure was available, the one with the maximum pore volume was selected. The measured pore volume was averaged over all available nitrogen and argon isotherms.

Adsorbent	N. of Isotherms	Computed Pore Vol. ($cm^3\,g^{-1}$)	Measured Pore Vol. ($cm^3\,g^{-1}$)
MIL-53(Al)	22	0.28	0.47 ± 0.11
UiO-66	20	0.49	0.45 ± 0.15
Ni-MOF-74	11	0.54	0.39 ± 0.09
Zn-MOF-74	9	0.55	0.44 ± 0.14
Mg-MOF-74	7	0.72	0.56 ± 0.11
ZIF-8	35	0.72	0.64 ± 0.15
CuBTC	78	0.83	0.58 ± 0.17
MIL-100(Fe)	13	1.01	0.82 ± 0.12
IRMOF-1	40	1.38	0.78 ± 0.48

The examples of IRMOF-1 and CuBTC show a high spread of saturation uptakes (and therefore measured pore volumes), with an apparent multi-modal distribution. In this context, the geometrical pore volume from the crystal structure is an absolute reference that helps pinpoint which reports involve highly crystalline and fully activated materials – an insight that would be difficult to gain from relative statistical analysis alone (e.g., using the method of Sholl and co-workers).[15]

For CuBTC, the material with the most characterization isotherms (78 in total), we proceeded inspecting isotherm and paper, manually. No significant error related to the digitization process or the extraction of the pore volume from the isotherm was found (and the pore volume reported by the authors, when present, was similar to the one computed

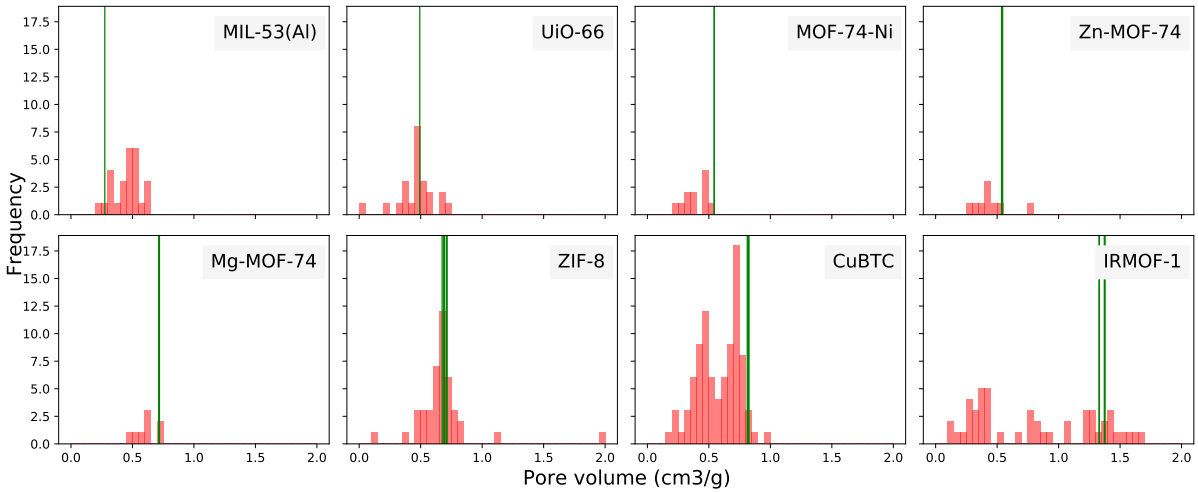

Figure 7: Histograms of the measured pore volumes for the MOFs with the highest number of characterization isotherms available. The computed geometric pore volumes from the matched CSD structures are shown by the vertical lines in green.

by us). We also note that while many of the reports contained CuBTC modifications or composite materials, the isotherms flagged as related to CuBTC in the NIST-ISODB indeed referred to the pristine version of the MOF, since it is often reported as a benchmark before further modification. However, this evidence suggest that algorithms that try to parse these values from the manuscripts via natural language processing may be particularly prone to errors, requiring elaborate tuning or the supervision by an expert reader.[52–54]

When the characterization isotherms indicated a weakly porous CuBTC ($< 0.4\,\mathrm{cm^3\,g^{-1}}$), this fact was usually mentioned by the authors (e.g., in the case of pellets,[55] or alternative synthesis routes[56]). Among the isotherms from which we computed a low pore volume in the range of $0.4\,\mathrm{cm^3\,g^{-1}}$ to $0.5\,\mathrm{cm^3\,g^{-1}}$, some authors attribute the low porosity to partial activation.[57] In many other cases, however, authors did not recognize the pore volume of their sample as low, despite it being less than half of the theoretical pore volume of the perfect, solvent-free crystal (as well as some of the highest reported experimental values). We can only speculate that they may have been influenced by the numerous other reports of pore volumes in this range and did aim to consult an independent benchmark. Going forward, we suggest to consider adsorption analyses and conclusions drawn from works on

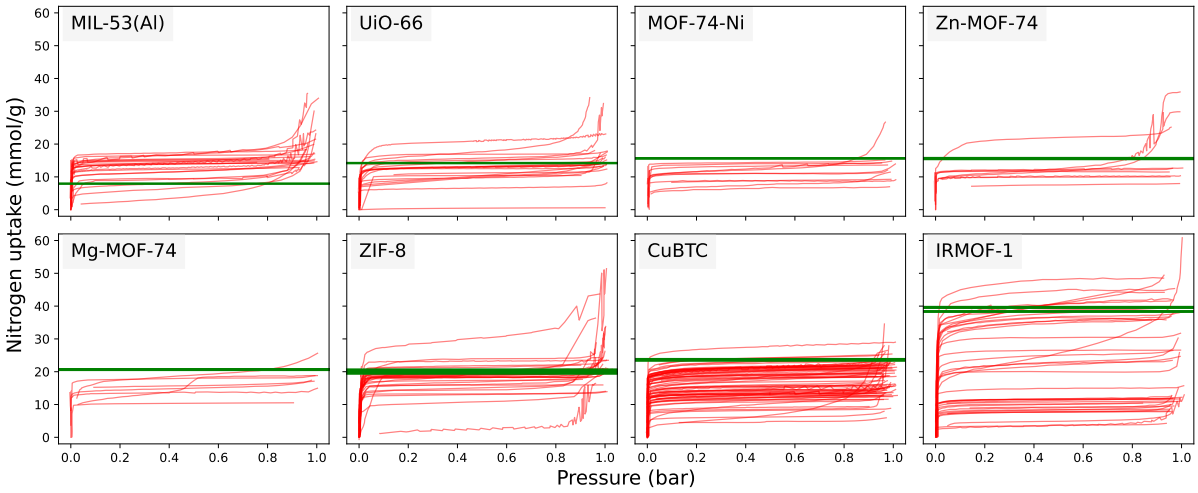

Figure 8: Nitrogen isotherms for the selection of most documented MOFs. Horizontal green lines indicate the pore volume computed from the crystal structures, in units of mmol of nitrogen per gram of material.

low-porosity CuBTC samples with caution.

Since incomplete activation was the most-cited reason for low pore volumes, we modelled the expected pore volume for CuBTC in the presence of several solvents. As shown in Table 2, partial activation can certainly explain the large reductions observed in CuBTC (but other hypotheses are also plausible).

Table 2: Geometric pore volume of partially solvated CuBTC. The desolvated FIQCEN structure[28] is used as a reference. Water and dimethylformamide (DMF) loading is expressed as a ratio of Cu metal sites. Note that the presence of solvent molecules both reduces the pore space and increase the weight of the sample, thus having a compound effect on the gravimetric pore volume.

System	Geometric Pore Volume ($cm^3\,g^{-1}$)
Bare framework	0.822
Half Cu-sites occupied by water	0.761
All Cu-sites occupied by water	0.706
Half Cu-sites occupied by DMF	0.605
All Cu-sites occupied by DMF	0.446

Among the isotherms showing too high pore volume, we identified one case of a typo in the units of the isotherm graph (mmol/g instead of the more plausible cm^3(STP)/g), resulting

in an experimental pore volume that dramatically exceeded the geometric one (11.34 vs. $0.82\,\mathrm{cm^3\,g^{-1}}$).[58] In another case, the experimental value of $0.978\,\mathrm{cm^3\,g^{-1}}$ slightly exceeds the geometric pore volume of CuBTC, and the manuscript also reports a very high BET surface area of $2327\,\mathrm{m^2\,g^{-1}}$, the largest ever reported to our knowledge.[59] While we are not able to determine in retrospect what led to this large value – for example, the calibration of the instrument, a material with large defects, or an imprecise BET calculation[60] – our analysis points at potential benefits from checking the geometric pore volume of the sample before moving on to measure the adsorption of other gases.

Agrawal et al. recently analysed the repeat synthesis of popular MOFs.[61] Since the indicators of popularity included the number of reports in the NIST-ISODB, their list of the six most-studied MOFs, unsurprisingly, contains four of the MOFs listed in Table 1 (UIO-66, CuBTC, ZIF-8 and IRMOF-1). The two missing MOFs did not make it to the end of our pipeline: MIL-101(Cr) was excluded because of the ambiguity with MIL-101 and MIL-101(Al) (see Table S1), and for MOF-177 we identified only three characterization isotherms vs. 8 references mentioned in their analysis. For the four MOFs that both lists have in common, Figure 7 and the histogram of reported BET surface areas in the Fig. 2 of the work by Agrawal et al. show clear qualitative similarities: a wide distribution between 0 and the theoretical reference for CuBTC and IRMOF-1, and the presence of samples with higher-than-theoretical reference BET (or pore volume) for UIO-66 and ZIF-8.

Finally, we briefly comment on the pore volume distributions of the other seven MOFs for which independent isotherms were reported:

- MIL-53(Al) is known for swelling upon adsorption, thus opening its pores. The crystal structure we matched in this study is a closed-pore model with $0.28\,\mathrm{cm^3\,g^{-1}}$ of geometrical pore volume (refcode: SABWAU01).[62] Using an open-pore model instead, e.g., DOYBEA,[63] we obtain a geometric pore volume of $0.55\,\mathrm{cm^3\,g^{-1}}$, about twice its closed-pore configuration. Most of the measured pore volumes fall inside the range between the open and closed-pore model.

- UiO-66 shows a distribution of measured pore volumes around the geometric pore volume, while one would expect the geometric pore volume to be the upper bound. Indeed, the article that reports the highest measured pore volume[64] ($0.74\,\mathrm{cm^3\,g^{-1}}$ as computed by our protocol and $0.8\,\mathrm{cm^3\,g^{-1}}$ as reported in the article) mentions that 2.3/12 of the BDC ligands were found to be missing, much higher than the normally observed ratio (1/12). As we mentioned before, defects are the most likely reason for a measured pore volume exceeding the geometric pore volume of the perfect crystal.

- The Ni, Zn and Mg analogs of the MOF-74 family display similar trends, with measured pore volumes in the range of 50 % to 100 % of the geometric one. Only for Zn-MOF-74 one isotherm exceeds the theoretical maximum uptake. The reported BET surface area for the structure is also very high: $1474\,\mathrm{m^2\,g^{-1}}$[65] while other reports do not go beyond $948\,\mathrm{m^2\,g^{-1}}$.[66,67] Our analysis flags this measurement not only as an outlier but also as a theoretically unlikely.

- ZIF-8 also has a distribution of measured pore volumes in the 50 % to 100 % range with respect to the geometric pore volume. One can spot in Figure 8 some outliers that should likely be double-checked from the the experimental side.

- IRMOF-1 is the second-most reproduced sample after CuBTC. It is surprising to observe the wide spread of measured values; only the comparison with the crystal structure allows to evaluate the quality of the sample and its desolvation. In a 2015 report, Sarkisov showed how computationally generated defects in the IRMOF-1 crystal structure impact gas adsorption, explaining deviations between experimental isotherms and those computed from the perfect crystal.[68]

Conclusions and Recommendations

Figure 9 summarizes the results of this work: starting from 105 922 structures identified as MOFs in the CSD and 4143 low-temperature N_2 or Ar adsorption isotherms provided by the NIST-ISODB, the automated pipeline was able to match a total of 545 characterization isotherms, corresponding to 291 MOFs.

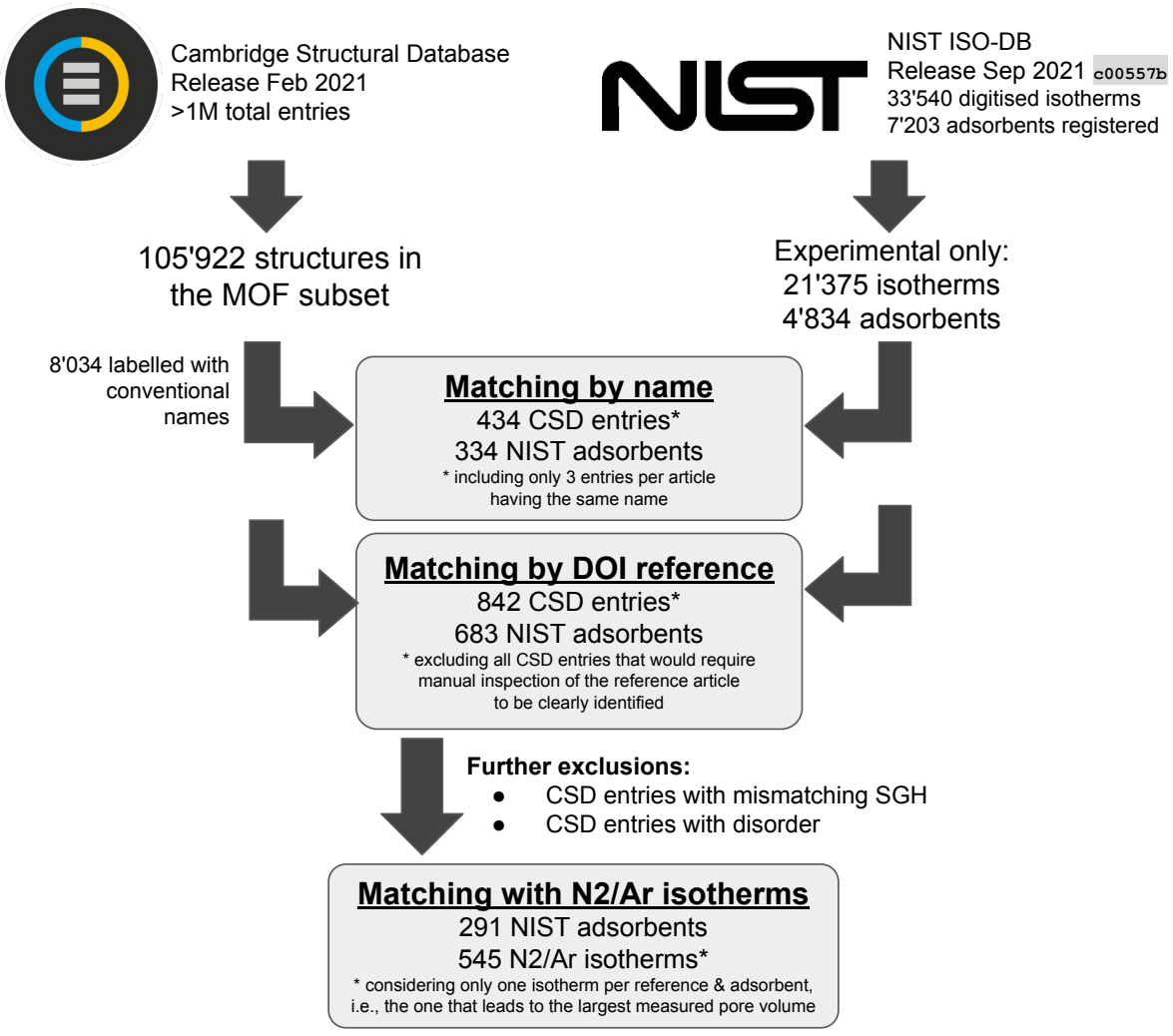

Figure 9: Final summary of the pipeline.

When comparing the measured to the geometric pore volume for all matched characterization isotherms, their ratio falls in the 75 % to 110 % range only in 35 % of the cases. In the remaining reports, the isotherms were likely measured on MOFs that were non-porous (because of narrow channels, unremoved solvent or collapsing upon desolvation) or that displayed other significant deviations from their expected theoretical crystal structure. In these MOFs, we can not expect molecular simulations to accurately predict the experimental uptake for any adsorbate, since the pore volume of the sample does not match the one of the provided crystal structure. The full pipeline is provided in the GitHub repository (`https://github.com/danieleongari/matching_isodb_csd`) in the form of Jupyter notebooks. The repository also contains the refcodes of CSD structures and the filenames of the NIST-ISODB isotherms matched in this work.

The NIST-ISODB has been collected by researchers and summer students, painstakingly digitizing thousands of adsorption isotherms from figures of academic papers. For those willing to contribute to this digitization effort, a tool has been developed as part of this work that streamlines the digitization process and makes it easy to submit new isotherms to the NIST-ISODB.[69] Going forward, however, we hope that the need for this digitization procedure will gradually recede as standard practices for reporting adsorption isotherm data are established.

For instance, the Allotrope format,[70] AniML,[71] Unified Data Model,[72] or the JCAMP-DX standard[73–76] provide not only a standardized serialization format but also standardized vocabularies for many techniques (however, at the moment, not for gas adsorption isotherms). Some of these formats (e.g., the Allotrope format) even support contextualizing the data by referencing ontologies, which can enable powerful semantic search.

The output files generated by adsorption information vary from manufacturer to manufacturer, contain different amounts and types of metadata, and are generally not published even in their native forms.[77] This characteristic of adsorption data is a large obstacle to more efficient and more accurate entry of adsorption isotherms into repositories such as

NIST-ISODB.

Evans et al. have, however, demonstrated how to convert the output files of three manufacturers' instruments plus the NIST-ISODB JSON format into a common format, the "adsorption information file" (AIF) that allows for machine-facilitated comparison of isotherms without extensive human intervention.[77,78] The AIF format does not intend to include all possible metadata regarding an isotherm measurement, but *enough* to allow comparison of (ostensibly) equilibrium isotherms. The AIF format has been approved as an IUPAC Project,[79] which will facilitate its development both for other manufacturers' instruments and generic isotherm data (which could include output of molecular simulations) as well as leverage and revise other IUPAC resources such as the IUPAC Gold Book.[80] The development version of the AIF is available for use even prior to completion of the IUPAC project and we highly encourage that authors release their isotherm data in the AIF development format in the supplementary information of papers without delay.

The adoption of a standard like the AIF will likely lead to both increased availability and quality of adsorption data, but on its own will not necessarily address the general challenge encountered in this work which concerns establishing links between related but independently maintained data sources, such as the CSD and the NIST-ISODB, as one can envision that similar arguments hold for matching of the structure with an IR, NMR, or XPS spectrum, and the oxidation state,[81] or the color of the crystal.[82]

Importantly, the need for matching of entries in different databases could be avoided by providing metadata using unique resource identifiers (URIs), such as the ORCID for the author of an entry, or the link to a structure in the CSD, and when the MOF is not reported in the CSD there exist technologies to uniquely identify structures.[32] Additionally, open research data repositories such as Zenodo,[83] the Open Science Framework[84] or the NIST-ISODB allow to publish data under a persistent identifier such as a DOI. If these recommendations were adopted for all chemical measurements, a part of this article would not have been written, as the matching between the identity of MOFs, their crystal structure

and the measured isotherms would have been certainly trivial and more extensively available. It is not the aim of this work to advocate for one standard or another, but we do advocate for the need for data interoperability conventions that are accepted by the community, and are a condition for publication. Only then we can avoid that similar articles, focusing on the mining of a certain property of interest, need to appear, as extensive querying would become an easy operation for computational and experimental researchers.

Supporting Information Available

List of ambiguous MOF names excluded. List of adsorbate strings stripped. Comparison of geometric pore volumes between analog structures. List of isotherms for the most reported adsorbents. Measured pore volume by year of publication. Comparison of pore volumes computed via different protocols. Illustration of the Weisfeiler-Lehman algorithm.

Acknowledgments

The research in this article was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 666983, MaGic), the MARVEL National Centre for Competence in Research funded by the Swiss National Science Foundation (grant agreement ID 51NF40-182892), and is part of the PrISMa Project (No 299659), which is funded through the ACT programme (Accelerating CCS Technologies, Horizon2020 Project No 294766).

Literature Cited

(1) Mason, J. A.; Veenstra, M.; Long, J. R. Evaluating metal–organic frameworks for natural gas storage. Chem. Sci. **2014**, 5, 32–51.

(2) Wang, C.; Wang, Y.; Ge, R.; Song, X.; Xing, X.; Jiang, Q.; Lu, H.; Hao, C.; Guo, X.; Gao, Y.; Jiang, D. A 3D covalent organic framework with exceptionally high iodine capture capability. Chem. - A Eur. J. **2018**, 24, 585–589.

(3) Zhu, L.; Liu, X.-Q.; Jiang, H.-L.; Sun, L.-B. Metal–Organic Frameworks for Heterogeneous Basic Catalysis. Chem. Rev. **2017**, acs.chemrev.7b00091.

(4) Bavykina, A.; Kolobov, N.; Khan, I. S.; Bau, J. A.; Ramirez, A.; Gascon, J.

Metal–Organic Frameworks in Heterogeneous Catalysis: Recent Progress, New Trends, and Future Perspectives. Chem. Rev. **2020**, acs.chemrev.9b00685.

(5) Liu, X.; Huang, D.; Lai, C.; Zeng, G.; Qin, L.; Wang, H.; Yi, H.; Li, B.; Liu, S.; Zhang, M.; Deng, R.; Fu, Y.; Li, L.; Xue, W.; Chen, S. Recent advances in covalent organic frameworks (COFs) as a smart sensing material. Chem. Soc. Rev. **2019**, 48, 5266–5302.

(6) Kreno, L. E.; Leong, K.; Farha, O. K.; Allendorf, M.; Van Duyne, R. P.; Hupp, J. T. Metal–Organic Framework Materials as Chemical Sensors. Chem. Rev. **2012**, 112, 1105–1125.

(7) Furukawa, H.; Cordova, K. E.; O'Keeffe, M.; Yaghi, O. M. The chemistry and applications of metal-organic frameworks. Science **2013**, 341.

(8) Diercks, C. S.; Yaghi, O. M. The atom, the molecule, and the covalent organic framework. Science **2017**, 355.

(9) Ongari, D.; Yakutovich, A. V.; Talirz, L.; Smit, B. Building a Consistent and Reproducible Database for Adsorption Evaluation in Covalent–Organic Frameworks. ACS Cent. Sci. **2019**, 5, 1663–1675.

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

(11) Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S.; Slater, B.; Siepmann, J. I.; Sholl, D. S.; Snurr, R. Q. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. J. Chem. Eng. Data **2019**, 64, 5985–5998.

(12) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge structural database. Acta Crystallogr. Sect. B Struct. Sci. **2016**, 72, 171–179.

(13) Siderius, D., Shen, V., Johnson III, R., van Zee, R., Eds. NIST/ARPA-E Database of Novel and Emerging Adsorbent Materials; National Institute of Standards and Technology: Gaithersburg, MD, 20899, 2014.

(14) Iacomi, P.; Llewellyn, P. L. Data Mining for Binary Separation Materials in Published Adsorption Isotherms. Chem. Mater. **2020**, 32, 982–991.

(15) Park, J.; Howe, J. D.; Sholl, D. S. How Reproducible Are Isotherm Measurements in Metal–Organic Frameworks? Chem. Mater. **2017**, 29, 10487–10495.

(16) Jablonka, K. M.; Zasso, M.; Patiny, L.; Marzari, N.; Pizzi, G.; Smit, B.; Yakutovich, A. V. Connecting lab experiments with computer experiments: Making "routine" simulations routine. ChemRxiv **2021**,

(17) Bingel, L. W.; Chen, A.; Agrawal, M.; Sholl, D. S. Experimentally verified alcohol adsorption isotherms in nanoporous materials from literature meta-analysis. J. Chem. Eng. Data **2020**, 65, 4970–4979.

(18) Thommes, M.; Kaneko, K.; Neimark, A. V.; Olivier, J. P.; Rodriguez-Reinoso, F.; Rouquerol, J.; Sing, K. S. Physisorption of gases, with special reference to the evaluation of surface area and pore size distribution (IUPAC Technical Report). Pure Appl. Chem. **2015**, 87, 1051–1069.

(19) Ongari, D.; Boyd, P. G.; Barthel, S.; Witman, M.; Haranczyk, M.; Smit, B. Accurate characterization of the pore volume in microporous crystalline materials. Langmuir **2017**, 33, 14529–14538.

(20) NIST-ISODB, `https://github.com/NIST-ISODB/isodb-library/commit/c00557bf5173a83a7f0bde73bb96a162c2ce9f12`, accessed on October 22nd 2021.

(21) Cai, X.; Gharagheizi, F.; Bingel, L. W.; Shade, D.; Walton, K. S.; Sholl, D. S. A collection of more than 900 gas mixture adsorption experiments in porous materials from literature meta-analysis. Industrial & Engineering Chemistry Research **2020**, 60, 639–651.

(22) Sarawade, P.; Tan, H.; Polshettiwar, V. Shape-and morphology-controlled Sustainable synthesis of Cu, Co, and in metal organic frameworks with high CO2 capture capacity. ACS Sustainable Chem. Eng. **2013**, 1, 66–74.

(23) Gadipelli, S.; Travis, W.; Zhou, W.; Guo, Z. A thermally derived and optimized structure from ZIF-8 with giant enhancement in CO 2 uptake. Energy Environ. Sci. **2014**, 7, 2232–2238.

(24) Moosavi, S. M.; Chidambaram, A.; Talirz, L.; Haranczyk, M.; Stylianou, K. C.; Smit, B. Capturing chemical intuition in synthesis of metal-organic frameworks. Nat. Commun. **2019**, 10, 539.

(25) Nguyen, H. G. T.; Sims, C. M.; Toman, B.; Horn, J.; van Zee, R. D.; Thommes, M.; Ahmad, R.; Denayer, J. F.; Baron, G. V.; Napolitano, E., et al. A reference high-pressure CH 4 adsorption isotherm for zeolite Y: results of an interlaboratory study. Adsorption **2020**, 26, 1253–1266.

(26) Pato-Doldán, B.; Rosnes, M. H.; Dietzel, P. D. An In-Depth Structural Study of the Carbon Dioxide Adsorption Process in the Porous Metal–Organic Frameworks CPO-27-M. ChemSusChem **2017**, 10, 1710–1719.

(27) Smaldone, R. A.; Forgan, R. S.; Furukawa, H.; Gassensmith, J. J.; Slawin, A. M.; Yaghi, O. M.; Stoddart, J. F. Metal–organic frameworks from edible natural products. Angew. Chem. **2010**, 49, 8630–8634.

(28) Chui, S. S.-Y.; Lo, S. M.-F.; Charmant, J. P.; Orpen, A. G.; Williams, I. D. A chemically

functionalizable nanoporous material [Cu3 (TMA) 2 (H2O) 3] n. Science **1999**, 283, 1148–1150.

(29) Xiao, J.; Wu, Y.; Li, M.; Liu, B.-Y.; Huang, X.-C.; Li, D. Crystalline structural intermediates of a breathing metal–organic framework that functions as a luminescent sensor and gas reservoir. Chem. - Eur. J. **2013**, 19, 1891–1895.

(30) Ohashi, M.; Yagyu, A.; Xu, Q.; Mashima, K. Metathesis Approach to Linkage of Two Tetraplatinum Cluster Units: Synthesis, Characterization, and Dimerization of [Pt4 (μ-OCOCH3) 7 (μ-OCO (CH2) n CH= CH2)](n= 0–3). Chem. Lett. **2006**, 35, 954–955.

(31) Barthel, S.; Alexandrov, E. V.; Proserpio, D. M.; Smit, B. Distinguishing Metal–Organic Frameworks. Cryst. Growth Des. **2018**, 18, 1738–1747.

(32) Bucior, B. J.; Rosen, A. S.; Haranczyk, M.; Yao, Z.; Ziebel, M. E.; Farha, O. K.; Hupp, J. T.; Siepmann, J. I.; Aspuru-Guzik, A.; Snurr, R. Q. Identification Schemes for Metal–Organic Frameworks To Enable Rapid Search and Cheminformatics Analysis. Cryst. Growth Des. **2019**, 19, 6682–6697.

(33) Moghadam, P. Z.; Li, A.; Wiggin, S. B.; Tao, A.; Maloney, A. G. P.; Wood, P. A.; Ward, S. C.; Fairen-Jimenez, D. Development of a Cambridge structural database subset: a collection of metal-organic frameworks for past, present, and future. Chem. Mater. **2017**, 29, 2618–2625.

(34) Zarabadi-Poor, P.; Marek, R. Comment on "Database for CO2 Separation Performances of MOFs Based on Computational Materials Screening". ACS Appl. Mater. Interfaces **2019**, 11, 16261–16265.

(35) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A

robust, open-source python library for materials analysis. Comput. Mater. Sci. **2013**, 68, 314–319.

(36) Togo, A.; Tanaka, I. `Spglib`: a software library for crystal symmetry search. 2018.

(37) Momma, K.; Izumi, F. VESTA 3for three-dimensional visualization of crystal, volumetric and morphology data. J. Appl. Crystallogr. **2011**, 44, 1272–1276.

(38) Shervashidze, N.; Schweitzer, P.; van Leeuwen, E. J.; Mehlhorn, K.; Borgwardt, K. M. Weisfeiler-Lehman Graph Kernels. J. Mach. Learn. Res. **2011**, 12, 2539–2561.

(39) Hagberg, A. A.; Schult, D. A.; Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. Proceedings of the 7th Python in Science Conference. Pasadena, CA USA, 2008; pp 11 – 15.

(40) Yang, X.; Zou, C.; He, Y.; Zhao, M.; Chen, B.; Xiang, S.; O'keeffe, M.; Wu, C. A Stable Microporous Mixed-Metal Metal–Organic Framework with Highly Active Cu2+ Sites for Efficient Cross-Dehydrogenative Coupling Reactions. Chem. Eur. J **2014**, 20, 1447–1452.

(41) Zhao, D.; Yue, D.; Jiang, K.; Cui, Y.; Zhang, Q.; Yang, Y.; Qian, G. Ratiometric dual-emitting MOF⊃dye thermometers with a tunable operating range and sensitivity. J. Mat. Chem. C **2017**, 5, 1607–1613.

(42) GitHub repository of the project, `https://github.com/danieleongari/matching_isodb_csd`, accessed on December 6th 2021.

(43) Wu, H.; Gong, Q.; Olson, D. H.; Li, J. Commensurate adsorption of hydrocarbons and alcohols in microporous metal organic frameworks. Chem. Rev. **2012**, 112, 836–868.

(44) Gurvich, L. Physicochemical attractive force. J. Phys. Chem. Soc. Russ **1915**, 47.

(45) Banerjee, D.; Simon, C. M.; Plonka, A. M.; Motkuri, R. K.; Liu, J.; Chen, X.; Smit, B.; Parise, J. B.; Haranczyk, M.; Thallapally, P. K. Metal–organic framework with optimally selective xenon adsorption and separation. Nat. Commun. **2016**, 7, 1–7.

(46) Bondi, A. v. van der Waals volumes and radii. J. Phys. Chem. **1964**, 68, 441–451.

(47) Kolokolov, D. I.; Stepanov, A. G.; Jobic, H. Mobility of the 2-Methylimidazolate Linkers in ZIF-8 Probed by 2H NMR: Saloon Doors for the Guests. J. Phys. Chem. C **2015**, 119, 27512–27520.

(48) Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. Microporous Mesoporous Mater. **2012**, 149, 134–141.

(49) Li, A.; Bueno-Perez, R.; Wiggin, S.; Fairen-Jimenez, D. Enabling efficient exploration of metal–organic frameworks in the Cambridge Structural Database. CrystEngComm **2020**, 22, 7152–7161.

(50) Jiao, Y.; Liu, Y.; Zhu, G.; Hungerford, J. T.; Bhattacharyya, S.; Lively, R. P.; Sholl, D. S.; Walton, K. S. Heat-treatment of defective UiO-66 from modulated synthesis: Adsorption and stability studies. J. Phys. Chem. C **2017**, 121, 23471–23479.

(51) Sholl, D. S.; Lively, R. P. Defects in metal–organic frameworks: challenge or opportunity? J. Phys. Chem. Lett. **2015**, 6, 3437–3444.

(52) Park, S.; Kim, B.; Choi, S.; Boyd, P. G.; Smit, B.; Kim, J. Text mining metal-organic framework papers. J. Chem. Inf. Model. **2018**, 58, 244–251.

(53) Nandy, A.; Duan, C.; Kulik, H. J. Using Machine Learning and Data Mining to Leverage Community Knowledge for the Engineering of Stable Metal–Organic Frameworks. J. Am. Chem. Soc. **2021**, 143, 17535–17547.

(54) Luo, Y.; Bag, S.; Zaremba, O.; Andreo, J.; Wuttke, S.; Tsotsalas, M.; Friederich, P. MOF Synthesis Prediction Enabled by Automatic Data Mining and Machine Learning. ChemRxiv **2021**,

(55) Li, J.; Yang, J.; Li, L.; Li, J. Separation of CO2/CH4 and CH4/N2 mixtures using MOF-5 and Cu3 (BTC) 2. J. Energy Chem. **2014**, 23, 453–460.

(56) Yang, H.; Orefuwa, S.; Goudy, A. Study of mechanochemical synthesis in the formation of the metal–organic framework Cu3 (BTC) 2 for hydrogen storage. Microporous Mesoporous Mater. **2011**, 143, 37–45.

(57) Zhao, Y.; Seredych, M.; Zhong, Q.; Bandosz, T. J. Aminated graphite oxides and their composites with copper-based metal–organic framework: in search for efficient media for CO 2 sequestration. RSC Adv. **2013**, 3, 9932–9941.

(58) Macias, E. E.; Ratnasamy, P.; Carreon, M. A. Catalytic activity of metal organic framework Cu3 (BTC) 2 in the cycloaddition of CO2 to epichlorohydrin reaction. Catal. Today **2012**, 198, 215–218.

(59) Tian, F.; Zhang, X.; Chen, Y. Highly selective adsorption and separation of dichloromethane/trichloromethane on a copper-based metal–organic framework. RSC Adv. **2016**, 6, 31214–31224.

(60) Osterrieth, J. How Reproducible Are Surface Areas Calculated from the BET Equation? Preprint at https://doi.org/10.26434/chemrxiv.14291644.v2. ChemRxiv **2021**,

(61) Agrawal, M.; Han, R.; Herath, D.; Sholl, D. S. Does repeat synthesis in materials chemistry obey a power law? Proc. Natl. Acad. Sci. **2020**, 117, 877–882.

(62) Ortiz, G.; Chaplais, G.; Paillaud, J.-L.; Nouali, H.; Patarin, J.; Raya, J.; Marichal, C. New insights into the hydrogen bond network in Al-MIL-53 and Ga-MIL-53. J. Phys. Chem. C **2014**, 118, 22021–22029.

(63) Alvarez, E.; Guillou, N.; Martineau, C.; Bueken, B.; Van de Voorde, B.; Le Guillouzer, C.; Fabry, P.; Nouar, F.; Taulelle, F.; De Vos, D., et al. The structure of the aluminum fumarate metal–organic framework A520. Angew. Chem. **2015**, 127, 3735–3739.

(64) Hu, Z.; Faucher, S.; Zhuo, Y.; Sun, Y.; Wang, S.; Zhao, D. Combination of optimization and metalated-ligand exchange: an effective approach to functionalize UiO-66 (Zr) MOFs for CO2 separation. Chem. - Eur. J. **2015**, 21, 17246–17255.

(65) Adhikari, A. K.; Lin, K.-S. Synthesis, fine structural characterization, and CO2 adsorption capacity of metal organic frameworks-74. J. Nanosci. Nanotechnol. **2014**, 14, 2709–2717.

(66) Ruano, D.; Díaz-García, M.; Alfayate, A.; Sánchez-Sánchez, M. Nanocrystalline M–MOF-74 as Heterogeneous Catalysts in the Oxidation of Cyclohexene: Correlation of the Activity and Redox Potential. ChemCatChem **2015**, 7, 674–681.

(67) Diaz-Garcia, M.; Mayoral, A.; Diaz, I.; Sanchez-Sanchez, M. Nanoscaled M-MOF-74 materials prepared at room temperature. Cryst. Growth Des. **2014**, 14, 2479–2487.

(68) Sarkisov, L. Molecular simulation of low temperature argon adsorption in several models of IRMOF-1 with defects and structural disorder. Dalton Trans. **2016**, 45, 4203–4212.

(69) NIST-ISODB Digitalizer interface, `https://github.com/NIST-ISODB/isotherm-digitizer-panel`, accessed on October 22nd 2021.

(70) Allotrope Foundation , Allotrope Data Format. 2021; `https://www.allotrope.org/`.

(71) Schäfer, B. Data Exchange in the Laboratory of the Future. Wiley Analytical Science **2018**,

(72) Pistoia Alliance , UDM. 2021; `https://github.com/PistoiaAlliance/UDM`.

(73) Grasselli, J. G. Jcamp-Dx, a Standard Format for Exchange of Infrared Spectra in Computer Readable Form. 2016; https://doi.org/10.1515/iupac.63.0111.

(74) Lampen, P.; Hillig, H.; Davies, A. N.; Linscheid, M. JCAMP-DX for Mass Spectrometry. Appl. Spectrosc. **1994**, 48, 1545–1552.

(75) Baumbach, J. I.; Davies, A. N.; Lampen, P.; Schmidt, H. JCAMP-DX. A standard format for the exchange of ion mobility spectrometry data (IUPAC Recommendations 2001). Pure Appl. Chem. **2001**, 73, 1765–1782.

(76) Davies, A. N.; Lampen, P. JCAMP-DX for NMR. Appl. Spectrosc. **1993**, 47, 1093–1099.

(77) Evans, J. D.; Bon, V.; Senkovska, I.; Kaskel, S. A universal standard archive file for adsorption data. Langmuir **2021**, 37, 4222–4226.

(78) GitHub repository AIF, https://github.com/AIF-development-team/adsorptioninformationformat, accessed on December 6th 2021.

(79) IUPAC Project: Standardized Reporting of Gas Adsorption Isotherms, https://iupac.org/project/2021-016-1-024, accessed on December 6th 2021.

(80) The IUPAC Compendium of Chemical Terminology (IUPAC Gold Book), https://doi.org/10.1351/goldbook, accessed on December 6th 2021.

(81) Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B. Using collective knowledge to assign oxidation states of metal cations in metal–organic frameworks. Nat. Chem. **2021**, 13, 771–777.

(82) Jablonka, K. M.; Moosavi, S. M.; Asgari, M.; Ireland, C.; Patiny, L.; Smit, B. A data-driven perspective on the colours of metal–organic frameworks. Chem. Sci. **2021**, 12, 3587–3598.

(83) Zenodo, `https://zenodo.org`, accessed on October 22nd 2021.

(84) Open Science Framework, `https://osf.io`, accessed on October 22nd 2021.

Table of Contents graphic