Alarming structural error rates in MOF databases used in data driven workflows identified via a novel metal oxidation state-based method

Andrew White, Marco Gibaldi, Jake Burner, R. Alex Mayo, Tom K. Woo*

Department of Chemistry and Biomolecular Sciences
University of Ottawa, Ottawa, Canada

*to whom correspondence should be addressed: twoo@uottawa.ca

https://doi.org/10.26434/chemrxiv-2024-ftsv3 **ORCID:** https://orcid.org/0000-0003-0073-3901 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Metal-organic frameworks (MOFs) are a diverse class of porous materials composed of inorganic nodes joined by organic linkers, currently under investigation for a wide range of applications including gas storage and separation where they have been commercialized.[1–3] Given the labor-intensive nature of synthesizing and testing individual MOFs, high-throughput computational screening and machine learning (ML) methods are increasingly viewed as essential for facilitating MOF development.[4–6] However, the structural fidelity of the "computation-ready" MOF databases used in such studies remains largely unquantified. We introduce MOSAEC, an algorithm that detects chemically invalid structures on the basis of metal oxidation states. MOSAEC was manually validated against ~16k MOF structures from the popular CoRE database, and was found to flag erroneous structures with 95% accuracy. Systematic examination of 14 leading experimental and hypothetical MOF databases containing >1.9 million MOFs reveals concerning structural error rates, exceeding 40% in most cases.

Owing to their highly tailorable and modular nature, the MOF chemical space is vast and growing. Estimates indicate over 100k MOFs have been synthesized to date,[4] and millions more hypothetical MOFs (hMOFs) have been constructed *in silico*.[7–9] To facilitate high-throughput computational workflows, MOF structures must be collected, curated, and preprocessed to render them "computation-ready". Such curated datasets provide the raw material for hundreds of data driven studies, evidenced by rapidly increasing citations of "computation-ready" MOF databases, nearly 300 annual citations in 2023 (Figure 1). This represents a large and growing body of work relying on "computation-ready" databases with the implicit assumption that the structures therein are chemically reasonable.

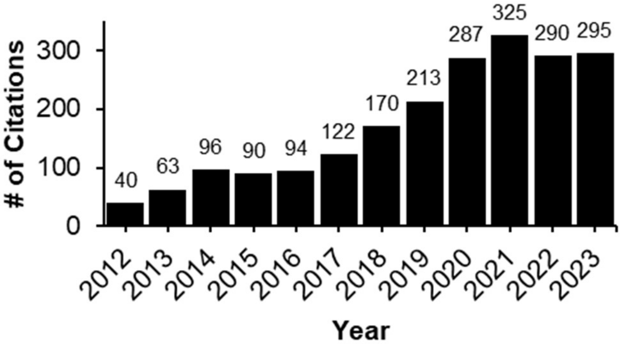

Figure 1. Citations of 'computation-ready' MOF database papers[5,9–39] since 2012.

However, manual validation of a representative sample of structures is not a requirement to publish a “computation-ready” database, and the fidelity of these MOF databases has been increasingly questioned, particularly the most widely cited CoRE databases which contain experimentally characterized structures taken from the CSD and automatically processed to prepare them for computation.[12,40–44] A commentary from Zarabadi-Poor and Marek[42] highlighted that the automated structure processing can overlook or even introduce structural errors. Chen and Manz[43] screened the CoRE-2019 MOF database for specific errors including overlapping atoms and select bonding irregularities, flagging 25% as problematic. Gibaldi et al. manually examined the structural building units (SBUs) used in hMOF construction and discovered error rates as high as 50% in the inorganic SBUs.[12] Despite these concerns, it is common practice to use ‘computation-ready’ databases ‘as-is’, without accounting for structural fidelity.

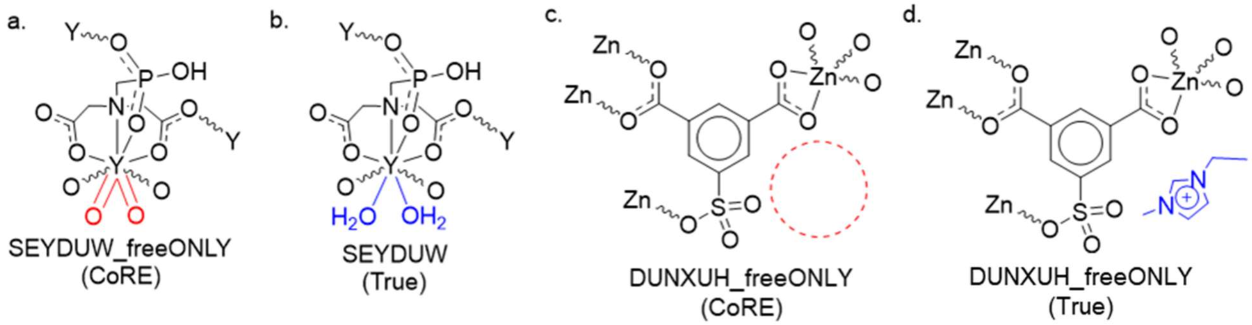

Figure 2. Juxtaposed chemical representations of the “computation-ready” and true reported structures for two CoRE-2019 MOFs: (a-b) SEYDUW, and (c/d) DUNXUH.

Through our own screening work, we encountered numerous structures given in these “computation-ready” databases with impossible or unlikely metal oxidation states (MOSs). For example, the CoRE structure SEYDUW_freeONLY (Figure 2a) has the yttrium center in a 7+ oxidation state, which is impossible given Y only has 3 valence electrons. An examination of the SEYDUW source publication[45] reveals that the oxo-ligands (red in Figure 2a) in the CoRE structure should be aqua ligands. This error likely occurred because structures deposited in the Cambridge Structure Database (CSD), from which the CoRE structures are derived, are often missing protons due to known limitations in the X-ray diffraction technique.[46] The correct structure shown in Figure 2b yields Y in its typical MOS (3+). Another “computation-ready” structure identified as problematic on the basis of MOS is the CoRE structure DUNXUH_freeONLY

https://doi.org/10.26434/chemrxiv-2024-ftsv3 ORCID: https://orcid.org/0000-0003-0073-3901 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0

(Figure 2c). The CoRE structure that is labelled as neutral framework gives Zn in the 3+ oxidation state, which is essentially unknown for stable Zn compounds. Examination of the source publication for DUNXUH[47] reveals the MOF is an anionic framework balanced by imidazolium cations (Figure 2d). If one includes the correct framework charge in computing the MOS of DUNXUH, then Zn carries the familiar 2+ oxidation state. Critically, neither SEYDUW_freeONLY nor DUNXUH_freeONLY were flagged as problematic in previous screenings of structural fidelity.

While differences in the MOS may seem minor, they have significant consequences for the material's electronic structure and can result in drastically different computed properties. For example, we found the erroneous SEYDUW_freeONLY to be a strong candidate for direct air capture of CO_2, whereas the true structure is mediocre in this respect. Additionally, since SEYDUW_freeONLY presents an impossible MOS, it is essentially non-chemical. Errors of this kind in a MOF database are akin to a drug candidate database containing molecules with 6-coordinate carbon atoms–they are not chemically valid and not synthetically accessible. Indeed, it is common practice among inorganic chemists to use the MOS heuristic to evaluate the feasibility of proposed structures or catalytic cycles.

From the cases above and others encountered in our work, it became clear that impossible or unlikely MOSs could be used as a basis to algorithmically detect structural problems in MOFs that go undetected with traditional methods based on interatomic distances or carbon valence bond sums.[48] With the wide diversity of chemical environments around the metal centers in MOFs, the challenge is to develop a robust and automated way to calculate the MOSs in any MOF which is both accurate and sensitive to errors in the structure "as-given". There are several published methods, including accurate ML approaches, for predicting what the MOSs are *expected to be*, irrespective of structural errors. Existing methods are either based on empirical valence-bond sums derived from metal-ligand bond lengths or are ML models trained exclusively on known/valid MOSs.[49–54] These methods cannot be used to identify structural errors because they would, for example, assign Y to be 3+ in both the erroneous and true SEYDUW structures (Figures 2a and 2b) because Y is almost always found in the 3+ oxidation state.

We have developed a novel algorithm called MOSAEC (Metal Oxidation State Automated Error Checker), which is the first generalized method that computes the MOSs for a structure "as-given" such that it can be used to detect problematic structures. MOSAEC considers the bonding environment at metal rather than the valence state of carbon and is therefore sensitive to otherwise undetectable errors such as those in Figure 2. Further, we have used MOSAEC to interrogate the structural fidelity of 14 computation-ready experimental and hMOF databases (Table 1) reflective of the current state of the field including older but widely cited datasets such as CoRE, as well as more recently published datasets reflective of contemporary developments in MOF structure curation. Over 1.9 million MOF structures have been evaluated with MOSAEC to answer the important question: what fraction of structures in published 'computation-ready' MOF databases have structural errors and are therefore not chemically valid?

Table 1. List and brief descriptions of the "computation-ready" datasets screened in this study, including distinguishing labels used for discussion.

label	**ref.**	**year**[a]	**description**
CoRE-2019	[23]	2019	2019 CoRE database including structures unmodified from the CSD
CoRE-2014	[30]	2014	2014 CoRE database
CSD MOF	[5,55]	2021	Publicly available "computation ready" MOF set provided by the CSD
QMOF	[56]	2021*	DFT optimized experimental structures, some hMOFs recently added
1D-MOF	[13]	2022	Set of experimental 1D coordination polymers
Majumdar	[15]	2021	hMOF dataset targeting structural diversity
Anderson-Gomez	[57]	2021	hMOF dataset targeting thermodynamic feasibility
Wilmer	[33]	2012	hMOF dataset built with a "bottom-up" tinkertoy approach
TobaCCo	[9]	2017	hMOF dataset built with a topology-based constructor
Boyd-Woo	[58]	2014	hMOF dataset built with both bottom-up and topology-based methods
ARC MOF	[37]	2023*	Hybrid dataset compiled from existing databases
MOFTransformer	[8]	2023	Set used to train the MOFTransformer model built with topology method
StableMOF	[35]	2023	hMOF dataset targeting thermodynamic stability
ODAC23	[20]	2023	Structures taken from CoRE + algorithmically defect-engineered MOFs

[a] Year published. * indicates the dataset is continually updated.

To validate the accuracy of MOSAEC's ability to detect problem structures, we manually evaluated the publicly available portion of CoRE-2019 (ca. 16k MOFs) for structural errors. 49% of these structures were manually categorized as "good", such that the structure matched that described in the source literature and 51% were determined to be "bad", such that the structure differed from the literature description. Structures flagged as "bad" by MOSAEC were manually confirmed to be "bad" in 95% of cases. Structures flagged as "good" by MOSAEC were confirmed

https://doi.org/10.26434/chemrxiv-2024-ftsv3 **ORCID:** https://orcid.org/0000-0003-0073-3901 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

to match the literature-reported structure in 86% of cases—indicating a high sensitivity but lower specificity. The lower specificity is attributed to cases where a structural error renders the ‘as given’ structure different from the literature, but the structure still has acceptable MOSs. This can occur for metals with multiple common oxidation states such as Fe or Mo. Nevertheless, high error sensitivity suggests that if MOSAEC flags a significant proportion of structures as “bad”, a dataset can be regarded as having low structural fidelity. Detailed discussion of the validation of MOSAEC is provided in the Supplementary Information.

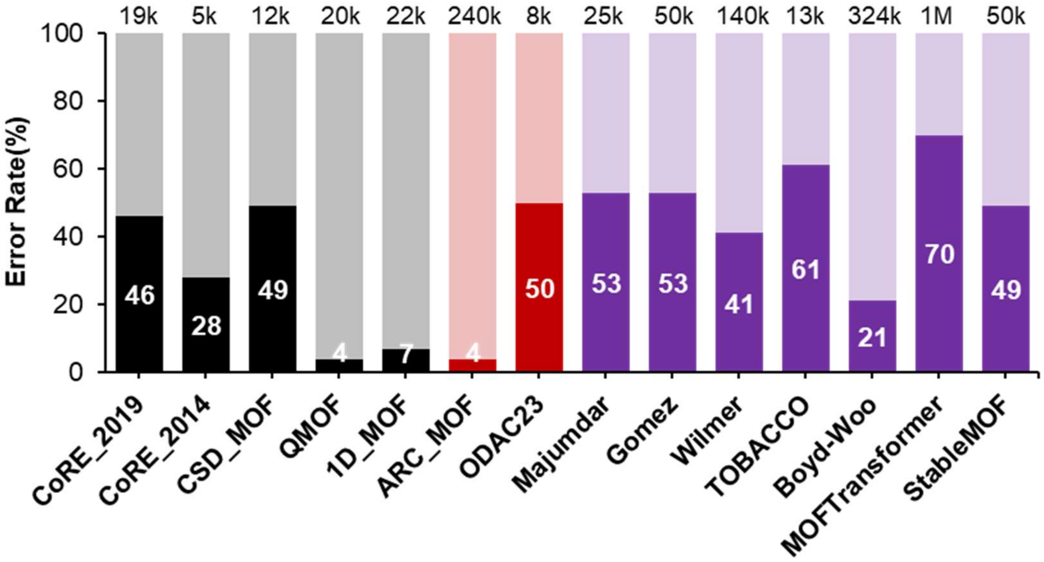

Figure 3. Error rates across 14 MOF databases. Databases composed solely or primarily of experimental structures are coloured black, databases of hMOFs purple, and hybrid (both experimental and hypothetical) databases colored red. Total number of MOFs per dataset indicated above corresponding bar.

Figure 3 shows the structural error rate determined by MOSAEC for 14 computation-ready MOF databases. Overall, the error rates are alarming - over 40% for most databases. Concerningly, CoRE-2019, which remains the most widely cited and screened MOF database has an error rate of 46%. This is nearly double the error rate of ca. 24% identified by Chen and Manz[59] in their evaluation of CoRE-2019 (We note that 23% of the structures flagged as problematic by Chen were actually ‘good’ structures). Unfortunately, with such large error rates, any previous screening studies using these databases should be revisited considering that some of the ‘hits’ may contain structural errors and could be false positives. Additionally, large portions of chemical space, as

https://doi.org/10.26434/chemrxiv-2024-ftsv3 **ORCID:** https://orcid.org/0000-0003-0073-3901 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

represented by the correct structures, remain unscreened. The impact of using databases with large error rates when training ML models is difficult to generalize as it will depend on how sensitive descriptors and/or target properties are to the structure errors among other factors. Unfortunately, one cannot predict the impact without retraining the model with a corrected dataset and comparing to results obtained with the original, flawed database. Thus, ML models trained on databases with large error rates should be revisited and possibly retrained on databases with erroneous structures excluded.

Algorithmically curated experimental datasets with the lowest error rates, QMOF (4%) and 1D-MOF (7%), arise from more cautious database construction workflows. Where the CoRE and CSD-MOF databases try to identify and repair errors (*e.g.* resolving disorder, restoring missing protons, etc.), QMOF and 1D-MOF outright exclude structures with potential errors and make minimal, if any, modifications to the original crystallographic information files (*cifs*). Indeed, modifications to the original *cif* often produce negative outcomes *e.g.*, structures in CSD-MOF which underwent automatic proton restoration (denoted with the suffix "_H" in that dataset) have double the error rate of their unaltered counterparts. ARC-MOF, a hybrid dataset containing structures sampled from a variety of experimental and hMOF databases, also has a low error rate of 4%. Its low error rate is to be expected as a development version of MOSAEC was utilized to discard erroneous structures during database construction, details in SI.

ODAC23 is also a hybrid dataset in that it contains ~5K experimental structures taken primarily from CoRE-2019, and an additional 3740 defect-engineered versions of the CoRE structures. Using these structures, ~40 million single point DFT calculations were performed with the intent of providing the training data for ML interatomic potentials. We find that 50% of the DFT input structures used in ODAC23 are found to contain structural errors. A quick analysis reveals that the defect engineering algorithm was a significant source of structural errors. For example, 62% of the structures that were 'good' prior to defect engineering, contained errors after the defects were introduced (see SI for further details). While modelling defects is increasingly viewed as important to correctly predicting the behavior of real MOF samples, it is important to maintain the chemical integrity of structures when introducing defects - charge balancing, for example, should be maintained.

https://doi.org/10.26434/chemrxiv-2024-ftsv3 **ORCID:** https://orcid.org/0000-0003-0073-3901 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

The majority of hMOF datasets evaluated (purple in Figure 3) were built using SBUs extracted from the CoRE 2014/2019 databases. As such, it is not surprising that the hMOF database error rates generally exceed 40%, including the newest hMOF datasets evaluated - the MOFtransformer (70%) and Stable-MOF (48%) datasets. The Boyd-Woo and Wilmer-hMOF databases are not constructed from CoRE SBUs and have the lowest (though still high) error rates among pure hMOF datasets at 21% and 40%, respectively. Future work involving construction of hMOFs should consider using the HEALED database of SBUs[12] or similarly validated SBU set for hMOF construction with care taken to account for the framework charge.

Although the hMOF databases have alarmingly high error rates, this does not necessarily invalidate the results of the studies from which they originate. Take for example the Stable-MOF database. The datamining used to identify SBU combinations that yielded MOFs with high thermodynamic stability is likely still valid, as is the set of MOFs proposed (provided that the structures are corrected). However, any computational screening performed on the uncorrected database is likely to give false positives due to the structural errors.

In summary, this study challenges the customary assumption that “computation-ready” MOF databases contain chemically realistic structures. Application of a novel MOS-based error screening algorithm on various MOF databases with >1.9 M structures (combined) demonstrated alarming error rates exceeding 40% in most databases including the popular CoRE databases. Utilization of these high-error databases without modification in subsequent computational studies is heavily discouraged to limit chemically unrealistic simulations. Further, results of previous studies employing error-stricken databases may require reconsideration in light of the pervasive structural errors. We find that the QMOF, 1D-MOF, ARC MOF, databases are relatively error-free and should be used instead of the popular CoRE and CSD-MOF databases. Alternatively, MOSEAC provides a lightweight means to automatically detect and remove flawed structures from existing databases. Future construction of hMOF databases should use a validated SBU set, such as HEALED and any automated defect-engineering algorithms should be evaluated to determine if they introduce structural errors. These revelations regarding inadequacies pervading computational MOF databases arm researchers with the knowledge necessary to leverage the

https://doi.org/10.26434/chemrxiv-2024-ftsv3 **ORCID:** https://orcid.org/0000-0003-0073-3901 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

power of high throughput screening and data driven approaches in a manner faithful to the true, experimental chemistry.

Online Methods

The detailed inner working of the MOSAEC algorithm will be the subject of another manuscript, but representative examples illustrating the core concepts and details of implementation are provided in the Supplementary Information. MOSAEC is freely available for non-commercial use upon request. Structures are flagged as problematic or ‘bad’ by MOSAEC if they feature one or more MOSs that are: i) impossible, ii) unknown, iii) non-integer, iv) zero, or v) improbable (<1% reported in that oxidation state for a given metal).

Supplementary Information: The Supplementary Information is available free of charge at [*doi link*].

- Details of the MOSAEC algorithm and validation are provided.
- The manual validation sets used to check oxidation state accuracy, error sensitivity, and error flag accuracy.
- A complete list of structures flagged by MOSAEC as being problematic in each databased screened in this work.

Competing Interests: A European patent application has been made for a method for assigning metal oxidation states.

Acknowledgements: Financial support from the Natural Sciences and Engineering Research Council of Canada (DISCOVERY Grant), the University of Ottawa, MITACS (Accelerate), and TotalEnergies is appreciated, and computing resources from the Canada Foundation for Innovation, TotalEnergies and the Digital Research Alliance of Canada.

References:

(1) Chen, Z.; Wasson, M. C.; Drout, R. J.; Robison, L.; Idrees, K. B.; Knapp, J. G.; Son, F. A.; Zhang, X.; Hierse, W.; Kühn, C.; Marx, S.; Hernandez, B.; Farha, O. K. The State of the Field: From Inception to Commercialization of Metal–Organic Frameworks. *Faraday Discuss.* **2021**, *225* (0), 9–69. https://doi.org/10.1039/D0FD00103A.

(2) Lin, J.-B.; Nguyen, T. T. T.; Vaidhyanathan, R.; Burner, J.; Taylor, J. M.; Durekova, H.; Akhtar, F.; Mah, R. K.; Ghaffari-Nik, O.; Marx, S.; Fylstra, N.; Iremonger, S. S.; Dawson,

https://doi.org/10.26434/chemrxiv-2024-ftsv3 **ORCID:** https://orcid.org/0000-0003-0073-3901 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

K. W.; Sarkar, P.; Hovington, P.; Rajendran, A.; Woo, T. K.; Shimizu, G. K. H. A Scalable Metal-Organic Framework as a Durable Physisorbent for Carbon Dioxide Capture. *Science (80-.).* **2021**, *374* (6574), 1464–1469. https://doi.org/10.1126/science.abi7281.

(3) Ozin, G.; Ye, J.; Bauchman, E. CALF-20: A Carbon Capture Success Story. *Advanced Science News*. January 2022.

(4) 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), 1890–1900. https://doi.org/10.1021/acscentsci.0c00988.

(5) 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* (7), 2618–2625. https://doi.org/10.1021/acs.chemmater.7b00441.

(6) Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B. Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chem. Rev.* **2020**, *120* (16), 8066–8129. https://doi.org/10.1021/acs.chemrev.0c00004.

(7) Lee, S.; Kim, B.; Cho, H.; Lee, H.; Lee, S. Y.; Cho, E. S.; Kim, J. Computational Screening of Trillions of Metal–Organic Frameworks for High-Performance Methane Storage. *ACS Appl. Mater. Interfaces* **2021**, *13* (20), 23647–23654. https://doi.org/10.1021/acsami.1c02471.

(8) Kang, Y.; Park, H.; Smit, B.; Kim, J. A Multi-Modal Pre-Training Transformer for Universal Transfer Learning in Metal–Organic Frameworks. *Nat. Mach. Intell.* **2023**, *5* (3), 309–318. https://doi.org/10.1038/s42256-023-00628-2.

(9) Colón, Y. J.; Gómez-Gualdrón, D. A.; Snurr, R. Q. Topologically Guided, Automated Construction of Metal–Organic Frameworks and Their Evaluation for Energy-Related Applications. *Cryst. Growth Des.* **2017**, *17* (11), 5801–5810. https://doi.org/10.1021/acs.cgd.7b00848.

https://doi.org/10.26434/chemrxiv-2024-ftsv3 **ORCID:** https://orcid.org/0000-0003-0073-3901 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(10) Villajos, J. A.; Bienert, M.; Gugin, N.; Emmerling, F.; Maiwald, M. A Database to Select Affordable MOFs for Volumetric Hydrogen Cryoadsorption Considering the Cost of Their Linkers. *Mater. Adv.* **2023**, *4* (18), 4226–4237. https://doi.org/10.1039/d3ma00315a.

(11) Lin, W.; Yu, Z.; Jiang, K.; Liang, H.; Xiong, X.; Chen, G. Data-Mining Based Assembly of Promising Metal-Organic Frameworks on Xe/Kr Separation. *Sep. Purif. Technol.* **2023**, *304*, 122357. https://doi.org/10.1016/j.seppur.2022.122357.

(12) Gibaldi, M.; Kwon, O.; White, A.; Burner, J.; Woo, T. K. The HEALED SBU Library of Chemically Realistic Building Blocks for Construction of Hypothetical Metal–Organic Frameworks. *ACS Appl. Mater. Interfaces* **2022**, *14* (38), 43372–43386. https://doi.org/10.1021/acsami.2c13100.

(13) Gharagheizi, F.; Yu, Z.; Sholl, D. S. Curated Collection of More than 20,000 Experimentally Reported One-Dimensional Metal-Organic Frameworks. *ACS Appl. Mater. Interfaces* **2022**, *14* (37), 42258–42266. https://doi.org/10.1021/ACSAMI.2C12485.

(14) Krokidas, P.; Karozis, S.; Moncho, S.; Giannakopoulos, G.; Brothers, E. N.; Kainourgiakis, M. E.; Economou, I. G.; Steriotis, T. A. Data Mining for Predicting Gas Diffusivity in Zeolitic-Imidazolate Frameworks (ZIFs). *J. Mater. Chem. A Mater. Energy Sustain.* **2022**, *10* (26), 13697–13703. https://doi.org/10.1039/d2ta02624d.

(15) Majumdar, S.; Moosavi, S. M.; Jablonka, K. M.; Ongari, D.; Smit, B. Diversifying Databases of Metal Organic Frameworks for High-Throughput Computational Screening. *ACS Appl. Mater. Interfaces* **2021**, *13* (51), 61004–61014. https://doi.org/10.1021/acsami.1c16220.

(16) Beauregard, N.; Pardakhti, M.; Srivastava, R. In Silico Evolution of High-Performing Metal Organic Frameworks for Methane Adsorption. *J. Chem. Inf. Model.* **2021**, *61* (7), 3232–3239. https://doi.org/10.1021/acs.jcim.0c01479.

(17) Li, A.; Perez, R. B.; Wiggin, S.; Ward, S. C.; Wood, P. A.; Fairen-Jimenez, D. The Launch of a Freely Accessible MOF CIF Collection from the CSD. *Matter* **2021**, *4* (4),

https://doi.org/10.26434/chemrxiv-2024-ftsv3 **ORCID:** https://orcid.org/0000-0003-0073-3901 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

1105–1106. https://doi.org/10.1016/j.matt.2021.03.006.

(18) Gu, C.; Yu, Z.; Liu, J.; Sholl, D. S. Construction of an Anion-Pillared MOF Database and the Screening of MOFs Suitable for Xe/Kr Separation. *ACS Appl. Mater. Interfaces* **2021**, *13* (9), 11039–11049. https://doi.org/10.1021/acsami.1c00152.

(19) Anderson, R.; Gomez-Gualdron, D. A. Large-Scale Free Energy Calculations on a Computational Metal-Organic Frameworks Database: Toward Synthetic Likelihood Predictions. *Chem. Mater.* **2020**, *32* (19), 8106–8119. https://doi.org/10.1021/acs.chemmater.0c00744.

(20) Sriram, A.; Choi, S.; Yu, X.; Brabson, L. M.; Das, A.; Ulissi, Z.; Uyttendaele, M.; Medford, A. J.; Sholl, D. S. The Open DAC 2023 Dataset and Challenges for Sorbent Discovery in Direct Air Capture. **2023**.

(21) Moghadam, P. Z.; Li, A.; Liu, X. W.; Bueno-Perez, R.; Wang, S. D.; Wiggin, S. B.; Wood, P. A.; Fairen-Jimenez, D. Targeted Classification of Metal-Organic Frameworks in the Cambridge Structural Database (CSD). *Chem. Sci.* **2020**, *11* (32), 8373–8387. https://doi.org/10.1039/d0sc01297a.

(22) Zhou, F.; Zheng, B.; Liu, D.; Wang, Z.; Yang, Q. Large-Scale Structural Refinement and Screening of Zirconium Metal-Organic Frameworks for H2S/CH4 Separation. *ACS Appl. Mater. Interfaces* **2019**, *11* (50). https://doi.org/10.1021/acsami.9b17885.

(23) 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* (12), 5985–5998. https://doi.org/10.1021/acs.jced.9b00835.

(24) Lan, Y.; Yan, T.; Tong, M.; Zhong, C. Large-Scale Computational Assembly of Ionic Liquid/MOF Composites: Synergistic Effect in the Wire-Tube Conformation for Efficient CO2/CH4 Separation. *J. Mater. Chem. A Mater. Energy Sustain.* **2019**, *7* (20), 12556–12564. https://doi.org/10.1039/c9ta01752f.

https://doi.org/10.26434/chemrxiv-2024-ftsv3 **ORCID:** https://orcid.org/0000-0003-0073-3901 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(25) Altintas, C.; Avci, G.; Daglar, H.; Nemati Vesali Azar, A.; Velioglu, S.; Erucar, I.; Keskin, S. Database for CO 2 Separation Performances of MOFs Based on Computational Materials Screening. *ACS Appl. Mater. Interfaces* **2018**, *10* (20), 17257–17268. https://doi.org/10.1021/acsami.8b04600.

(26) Zhang, C.; Lan, Y.; Guo, X.; Yang, Q.; Zhong, C. Materials Genomics-Guided Ab Initio Screening of MOFs with Open Copper Sites for Acetylene Storage. *AIChE J.* **2018**, *64* (4). https://doi.org/10.1002/aic.16025.

(27) Nazarian, D.; Camp, J. S.; Chung, Y. G.; Snurr, R. Q.; Sholl, D. S. Large-Scale Refinement of Metal-Organic Framework Structures Using Density Functional Theory. *Chem. Mater.* **2017**, *29* (6), 2521–2528. https://doi.org/10.1021/acs.chemmater.6b04226.

(28) Nazarian, D.; Camp, J. S.; Sholl, D. S. A Comprehensive Set of High-Quality Point Charges for Simulations of Metal-Organic Frameworks. *Chem. Mater.* **2016**, *28* (3). https://doi.org/10.1021/acs.chemmater.5b03836.

(29) Boyd, P. G.; Woo, T. K. A Generalized Method for Constructing Hypothetical Nanoporous Materials of Any Net Topology from Graph Theory. *CrystEngComm* **2016**, *18* (21), 3777–3792. https://doi.org/10.1039/C6CE00407E.

(30) Chung, Y. G.; Camp, J.; Haranczyk, M.; Sikora, B. J.; Bury, W.; Krungleviciute, V.; Yildirim, T.; Farha, O. K.; Sholl, D. S.; Snurr, R. Q. Computation-Ready, Experimental Metal–Organic Frameworks: A Tool To Enable High-Throughput Screening of Nanoporous Crystals. *Chem. Mater.* **2014**, *26* (21), 6185–6192. https://doi.org/10.1021/cm502594j.

(31) Lim, Y.; Kim, B.; Kim, J. Data-Driven Design of Flexible Metal-Organic Frameworks for Gas Storage. *Chem. Mater.* **2024**, *36* (11), 5465–5473. https://doi.org/10.1021/acs.chemmater.4c00398.

(32) Addicoat, M. A.; Coupry, D. E.; Heine, T. AuToGraFS: Automatic Topological Generator for Framework Structures. *J. Phys. Chem. A* **2014**, *118* (40), 9607–9614. https://doi.org/10.1021/jp507643v.

https://doi.org/10.26434/chemrxiv-2024-ftsv3 **ORCID:** https://orcid.org/0000-0003-0073-3901 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(33) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-Scale Screening of Hypothetical Metal-Organic Frameworks. *Nat. Chem.* **2012**, *4*, 83–89. https://doi.org/10.1038/nchem.1192.

(34) Oliveira, F. L.; Cleeton, C.; Neumann Barros Ferreira, R.; Luan, B.; Farmahini, A. H.; Sarkisov, L.; Steiner, M. CRAFTED: An Exploratory Database of Simulated Adsorption Isotherms of Metal-Organic Frameworks. *Sci. Data* **2023**, *10* (1), 230. https://doi.org/10.1038/s41597-023-02116-z.

(35) Nandy, A.; Yue, S.; Oh, C.; Duan, C.; Terrones, G. G.; Chung, Y. G.; Kulik, H. J. A Database of Ultrastable MOFs Reassembled from Stable Fragments with Machine Learning Models. *Matter* **2023**, *6* (5), 1585–1603. https://doi.org/10.1016/j.matt.2023.03.009.

(36) Zhang, Z.; Valente, D. S.; Shi, Y.; Limbu, D. K.; Momeni, M. R.; Shakib, F. A. In Silico High-Throughput Design and Prediction of Structural and Electronic Properties of Low-Dimensional Metal-Organic Frameworks. *ACS Appl. Mater. Interfaces* **2023**, *15* (7), 9494–9507. https://doi.org/10.1021/acsami.2c22665.

(37) Burner, J.; Luo, J.; White, A.; Mirmiran, A.; Kwon, O.; Boyd, P. G.; Maley, S.; Gibaldi, M.; Simrod, S.; Ogden, V.; Woo, T. K. ARC-MOF: A Diverse Database of Metal-Organic Frameworks with DFT-Derived Partial Atomic Charges and Descriptors for Machine Learning. *Chem. Mater.* **2023**, *35* (3), 900–916. https://doi.org/10.1021/acs.chemmater.2c02485.

(38) Bobbitt, N. S.; Shi, K.; Bucior, B. J.; Chen, H.; Tracy-Amoroso, N.; Li, Z.; Sun, Y.; Merlin, J. H.; Siepmann, J. I.; Siderius, D. W.; Snurr, R. Q. MOFX-DB: An Online Database of Computational Adsorption Data for Nanoporous Materials. *J. Chem. Eng. Data* **2023**, *68* (2), 483–498. https://doi.org/10.1021/acs.jced.2c00583.

(39) Sun, J.; Gharagheizi, F.; Fang, H.; Ravikovitch, P. I.; Sholl, D. S. Development of Porous Crystalline Materials for Selective Binding of O2 from Air. *J. Phys. Chem. C* **2023**, *127* (1), 776–787. https://doi.org/10.1021/acs.jpcc.2c07508.

https://doi.org/10.26434/chemrxiv-2024-ftsv3 **ORCID:** https://orcid.org/0000-0003-0073-3901 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(40) Velioglu, S.; Keskin, S. Revealing the Effect of Structure Curations on the Simulated CO 2 Separation Performances of MOFs. *Mater. Adv.* **2020**, *1* (3), 341–353. https://doi.org/10.1039/D0MA00039F.

(41) Daglar, H.; Gulbalkan, H. C.; Avci, G.; Aksu, G. O.; Altundal, O. F.; Altintas, C.; Erucar, I.; Keskin, S. Effect of Metal–Organic Framework (MOF) Database Selection on the Assessment of Gas Storage and Separation Potentials of MOFs. *Angew. Chemie Int. Ed.* **2021**, *60* (14), 7828–7837. https://doi.org/10.1002/anie.202015250.

(42) Zarabadi-Poor, P.; Marek, R. Comment on "Database for CO 2 Separation Performances of MOFs Based on Computational Materials Screening." *ACS Appl. Mater. Interfaces* **2019**, *11* (18), 16261–16265. https://doi.org/10.1021/acsami.8b15684.

(43) Chen, T.; Manz, T. A. Identifying Misbonded Atoms in the 2019 CoRE Metal-Organic Framework Database. *RSC Adv.* **2020**, *10* (45), 26944–26951. https://doi.org/10.1039/d0ra02498h.

(44) Barthel, S.; Alexandrov, E. V; Proserpio, D. M.; Smit, B. Distinguishing Metal–Organic Frameworks. *Cryst. Growth Des.* **2018**, *18* (3), 1738–1747. https://doi.org/10.1021/acs.cgd.7b01663.

(45) Tang, S. F.; Song, J. L.; Li, X. N.; Mao, J. G. Luminescent Lanthanide(III) Carboxylate-Phosphonates with Helical Tunnels. *Cryst. Growth Des.* **2006**, *6* (10), 2322–2326. https://doi.org/10.1021/cg060248l.

(46) Müller, P.; Herbst-Irmer, R.; Spek, A. L.; Schneider, T. R.; Sawaya, M. R. *Crystal Structure Refinement: A Crystallographer's Guide to SHELXL*; Oxford University Press, 2010; Vol. 9780198570. https://doi.org/10.1093/acprof:oso/9780198570769.001.0001.

(47) Tong, Y. B.; Liu, S. X.; Zou, Y.; Xue, C.; Duan, H. B.; Liu, J. L.; Ren, X. M. Insight into Understanding Dielectric Behavior of a Zn-MOF Using Variable-Temperature Crystal Structures, Electrical Conductance, and Solid-State 13C NMR Spectra. *Inorg. Chem.* **2016**, *55* (22), 11716–11726. https://doi.org/10.1021/ACS.INORGCHEM.6B01759/SUPPL_FILE/IC6B01759_SI_002.

https://doi.org/10.26434/chemrxiv-2024-ftsv3 **ORCID:** https://orcid.org/0000-0003-0073-3901 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

CIF.

(48) Chen, T.; Manz, T. A. Identifying Misbonded Atoms in the 2019 CoRE Metal–Organic Framework Database. *RSC Adv.* **2020**, *10* (45), 26944–26951. https://doi.org/10.1039/D0RA02498H.

(49) 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* (8), 771–777. https://doi.org/10.1038/s41557-021-00717-y.

(50) Reeves, M. G.; Wood, P. A.; Parsons, S. Automated Oxidation-State Assignment for Metal Sites in Coordination Complexes in the Cambridge Structural Database. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2019**, *75* (6), 1096–1105. https://doi.org/10.1107/S2052520619013040.

(51) Shields, G. P.; Raithby, P. R.; Allen, F. H.; Motherwell, W. D. S. S. The Assignment and Validation of Metal Oxidation States in the Cambridge Structural Database. *Acta Crystallogr. Sect. B Struct. Sci.* **2000**, *56* (3), 455–465. https://doi.org/10.1107/S0108768199015086.

(52) Park, H.; Kang, Y.; Choe, W.; Kim, J. Mining Insights on Metal-Organic Framework Synthesis from Scientific Literature Texts. *J. Chem. Inf. Model.* **2022**, *62* (5), 1190–1198. https://doi.org/10.1021/acs.jcim.1c01297.

(53) Shevchenko, A. P.; Smolkov, M. I.; Wang, J.; Blatov, V. A. Mining Knowledge from Crystal Structures: Oxidation States of Oxygen-Coordinated Metal Atoms in Ionic and Coordination Compounds. *J. Chem. Inf. Model.* **2022**, *62* (10), 2332–2340. https://doi.org/10.1021/acs.jcim.2c00080.

(54) Fu, N.; Hu, J.; Feng, Y.; Morrison, G.; Loye, H.-C. zur; Hu, J. Composition Based Oxidation State Prediction of Materials Using Deep Learning Language Models. *Adv. Sci.* **2023**, *10* (28), 2301011. https://doi.org/https://doi.org/10.1002/advs.202301011.

(55) Li, A.; Perez, R. B.; Wiggin, S.; Ward, S. C.; Wood, P. A.; Fairen-Jimenez, D. The Launch of a Freely Accessible MOF CIF Collection from the CSD. *Matter* **2021**, *4* (4),

https://doi.org/10.26434/chemrxiv-2024-ftsv3 **ORCID:** https://orcid.org/0000-0003-0073-3901 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

1105–1106. https://doi.org/10.1016/J.MATT.2021.03.006.

(56) Rosen, A. S.; Iyer, S. M.; Ray, D.; Yao, Z.; Aspuru-Guzik, A.; Gagliardi, L.; Notestein, J. M.; Snurr, R. Q. Machine Learning the Quantum-Chemical Properties of Metal–Organic Frameworks for Accelerated Materials Discovery. *Matter* **2021**, *4* (5), 1578–1597. https://doi.org/10.1016/j.matt.2021.02.015.

(57) Anderson, R.; Gómez-Gualdrón, D. A. Deep Learning Combined with IAST to Screen Thermodynamically Feasible MOFs for Adsorption-Based Separation of Multiple Binary Mixtures. *J. Chem. Phys.* **2021**, *154* (23), 234102. https://doi.org/10.1063/5.0048736/15978101/234102_1_ACCEPTED_MANUSCRIPT.PDF.

(58) Fernandez, M.; Boyd, P. G.; Daff, T. D.; Aghaji, M. Z.; Woo, T. K. Rapid and Accurate Machine Learning Recognition of High Performing Metal Organic Frameworks for CO 2 Capture. *J. Phys. Chem. Lett.* **2014**, *5* (17), 3056–3060. https://doi.org/10.1021/jz501331m.

(59) Chen, T.; Manz, T. A. Identifying Misbonded Atoms in the 2019 CoRE Metal-Organic Framework Database. *RSC Adv.* **2020**, *10* (45), 26944–26951. https://doi.org/10.1039/d0ra02498h.

https://doi.org/10.26434/chemrxiv-2024-ftsv3 **ORCID:** https://orcid.org/0000-0003-0073-3901 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0