Supplementary Information

MOSAEC-DB: A comprehensive database of experimental metal-organic frameworks with verified chemical accuracy suitable for molecular simulations.

Marco Gibaldi[1], Anna Kapeliukha[1,2], Andrew White[1], Jun Luo[1`], Robert Alex Mayo[1], Jake Burner[1`], Tom K. Woo[1,*]

[1] Department of Chemistry and Biomolecular Sciences,

University of Ottawa, 10 Marie Curie Private, Ottawa K1N 6N5, Canada

[2] Educational and Scientific Institute of High Technologies,

Taras Shevchenko National University of Kyiv, 4-g Hlushkova Avenue, Kyiv 03022, Ukraine

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

Candidate Structure Retrieval.

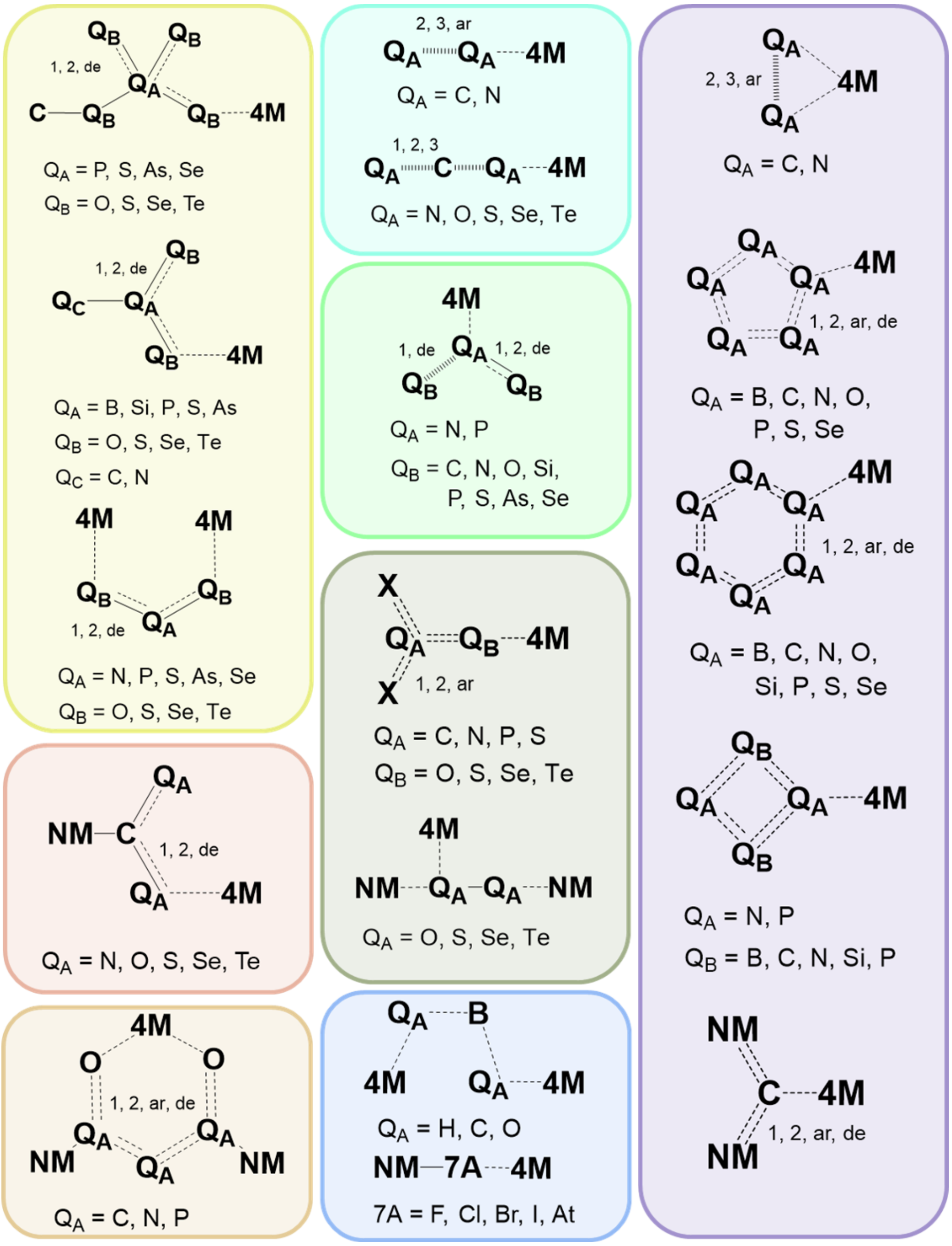

Figure S1. Summary of the chemical substructure criteria employed when searching for additional crystal structure candidates within the Cambridge Structural Database (CSD) using Conquest[1]. Annotations regarding general atom and bond type definitions are provided when necessary. Dotted lines represent bonds of any type (i.e. single, double, triple, aromatic, delocalized, *etc.*) The atom labels 4M and NM represent any metal atom and any non-metal atom, respectively. Implementation of these criteria yielded ca. 19.9k additional candidates beyond those present in the CSD MOF subset[2] as of CSD version 5.4.5 (including updates up to March 2024).[3]

Solvent Removal Algorithm (SAMOSA)—Examples.

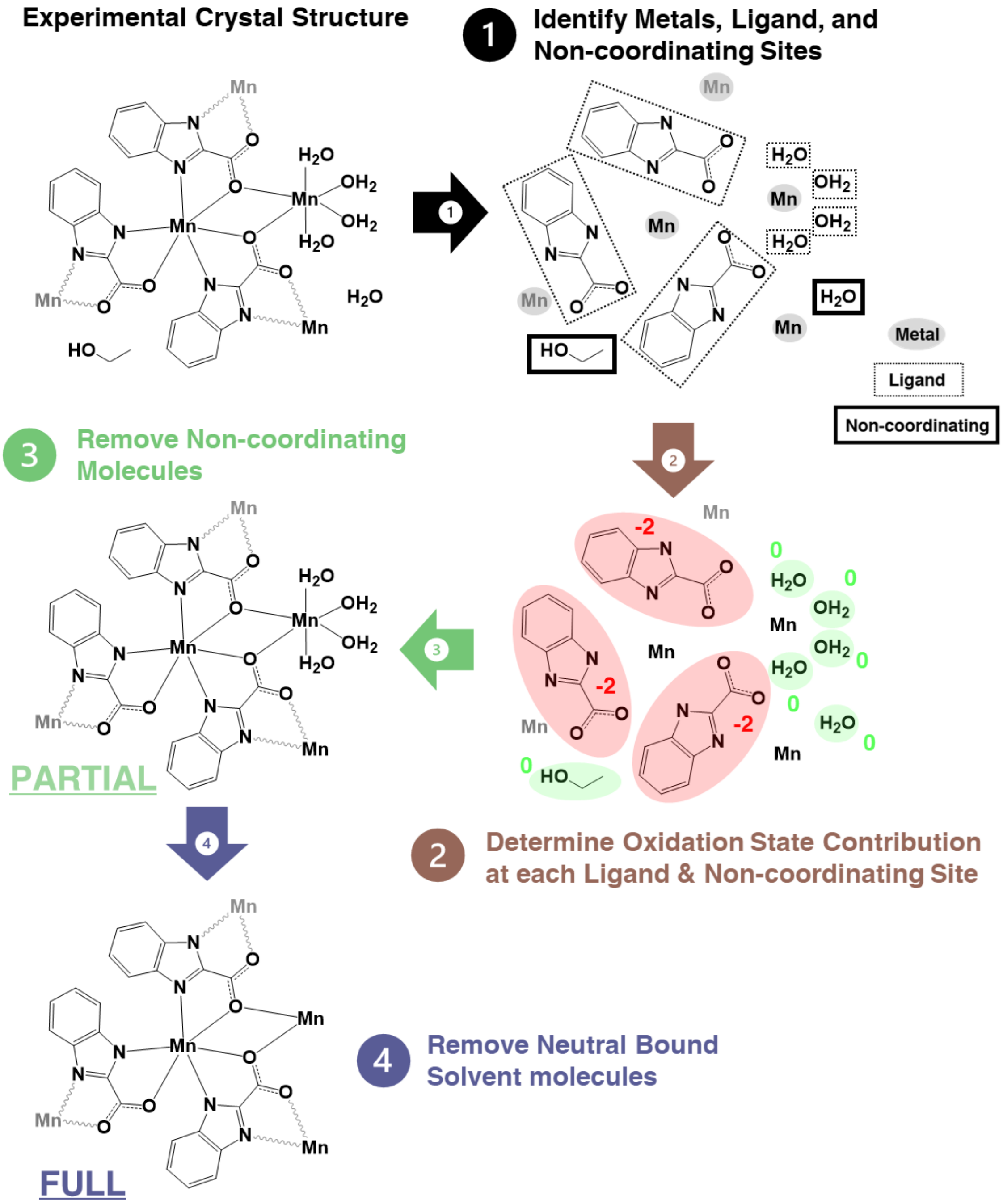
Experimental Crystal Structure
1 Identify Metals, Ligand, and Non-coordinating Sites
Metal
Ligand
Non-coordinating
2 Determine Oxidation State Contribution at each Ligand & Non-coordinating Site
-2
0
3 Remove Non-coordinating Molecules
PARTIAL
4 Remove Neutral Bound Solvent molecules
FULL
Mn
N
O
H2O
OH2
HO

Figure S2. Demonstration of the oxidation state-informed solvent removal protocol (SAMOSA) applied to generate fully and partially activated, neutral crystal structures. Chemical structure diagrams of a representative crystal structure (CSD Refcode: UYODOC) highlight the contribution of each individual step to the general workflow.

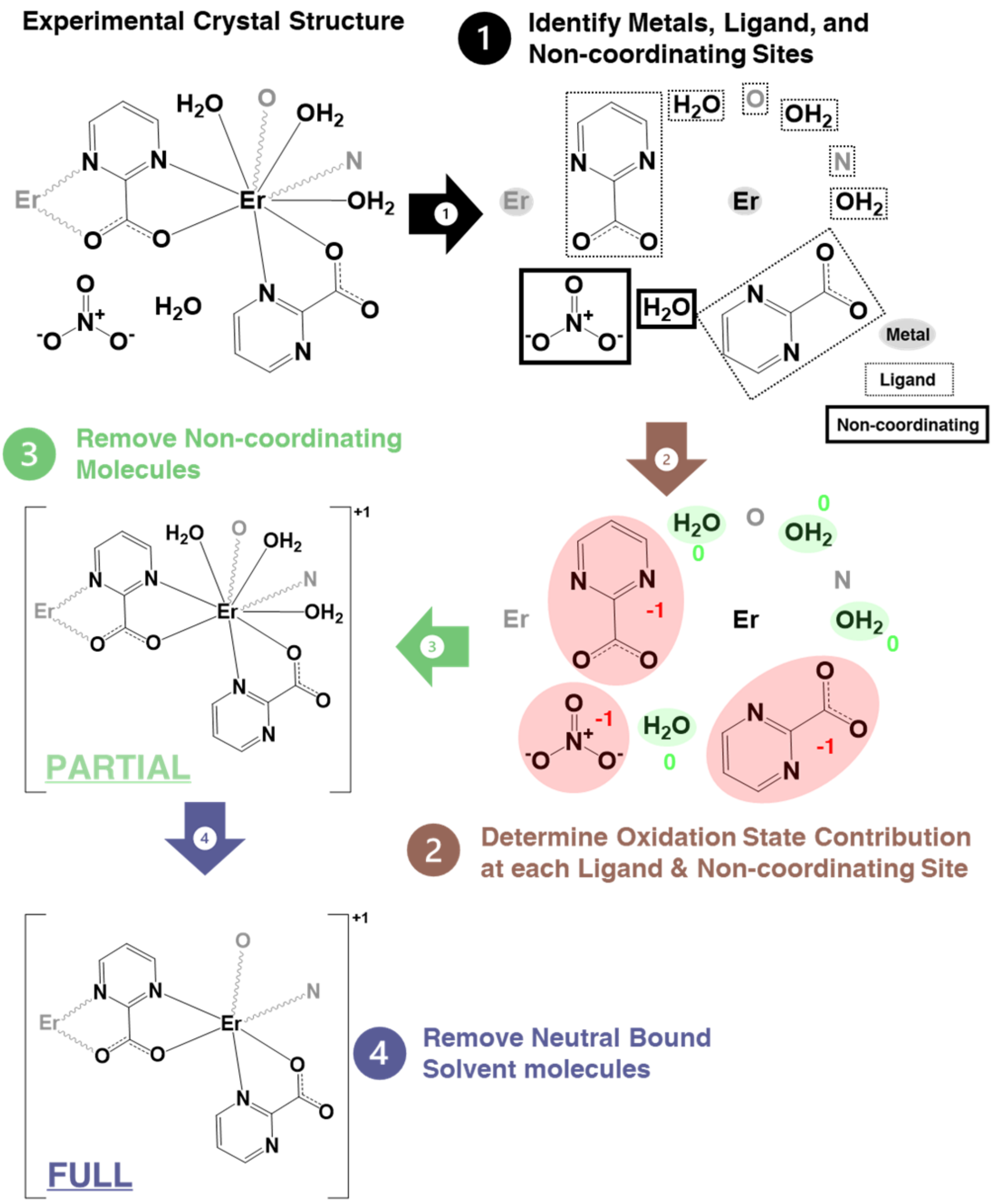
Experimental Crystal Structure
1 Identify Metals, Ligand, and Non-coordinating Sites
Metal
Ligand
Non-coordinating
2 Determine Oxidation State Contribution at each Ligand & Non-coordinating Site
3 Remove Non-coordinating Molecules
PARTIAL
+1
4 Remove Neutral Bound Solvent molecules
FULL
+1
Er
N
O
H_2O
OH_2
N^+
O^-
-1
0

Figure S3. Demonstration of the oxidation state-informed solvent removal protocol (SAMOSA) applied to generate fully and partially activated, charged crystal structures. Chemical structure diagrams of a representative crystal structure (CSD Refcode: KERWUC) highlight the contribution of each individual step to the general workflow.

Structural Error Analysis (MOSAEC)—Details.

Computed ligand formal changes and their implied metal oxidation states were employed as a metric for flagging problematic periodic crystal structures according to the protocol first discussed in reports of the MOSAEC[4] technique. Instances of likely structural errors (e.g., omission of protons or other atoms, crystallographic disorder, omitted charge-balancing ions, hypovalent and hypervalent atoms, overlapping atoms, *etc.*) in the input crystal structure are deduced based on the fact that if the unit cell is assumed to be neutral, the resulting metal oxidation states should follow accepted rules and formalisms (i.e. oxidation states should not exceed the atom's valence, should be integer values, should be frequently observed in experimental chemistry, *etc.*). Supplying the crystallographic information of any crystal structure containing metals as input, this process flags any structures containing metal sites that violates these generally accepted chemistry principles as highly likely to contain some form of structural error. A key advantage over previous approaches which relied on machine learning of known oxidation or empirical bond valence methods to compute values from the metal-ligand bonding environment (e.g., bond distances) lies in the fact that MOSAEC plainly computes the metal oxidation states for the structure *as given*, including any structural problems. Comprehensive explanations regarding the design and operation of the core MOSAEC functions are provided in previous accounts[4,5]; however, we provide a concise step-by-step rundown of the oxidation state calculation and error flagging processes below:

(i). Atomic coordinates are read in from the crystallographic information file (.cif) and bonding is assigned according to the CSD Python API[3] bond assignment algorithm.

(ii). Metal and ligand atoms are separately identified through analysis the bond connectivity of each atom in the crystal structure.

(iii). Formal charges of each nonmetal ligand atom are calculated according to an idealized bond valence sum method, wherein only the assigned bond orders (i.e. single bonds contributing 1, double bonds contributing 2, and so on) are applied. This differs from traditional bond valence sum techniques which employ empirical bond orders according to atom identity and interatomic distances and can produce non-integer values. Note that the metal-ligand bonds, which may vary considerably, are not involved at all when computing a given metal's oxidation state.

(iv). Each ligand's overall contribution to the oxidation state is then calculated as the sum of formal charges across all its atoms.

(v). Binding networks of metal atom sites defined as those connected through charged ligands possessing the ability to delocalize (share) their charge contribution through conjugation are identified using functions that recursively walk through the bonding paths connecting metals through ligand atoms.

(vi). Ligand charge contributions are then distributed across the available metals in the crystal structure according to several charge accounting principles i.e. ranging from local distribution where only attached ligand atoms contribute to a metal's oxidation state to fully global distribution where no restrictions are placed on how the charges can be distributed between all metals' oxidation states.

(vii). Calculated oxidation states of each metal atom are then evaluated according to rules derived from chemical insights regarding what constitutes a valid oxidation state. Namely, that oxidation states should be integer values that do need exceed the metal's available valence electron counts, and typically they should correspond to oxidation states which are observed commonly in experimental characterization. CSD metadata was parsed to assess the relative population of various oxidation states for each metal element. Oxidation states appearing in more than 1% of the metals' reported structures were deemed sufficiently probable, while those below this threshold were classified as improbable. Note that each unique metal atom site possesses distinct error flags to describe the varying metal environments present in the structure.

(viii). Finally, the overall quality of structure can be assessed by analyzing any error flags associated with the constituent metal atoms. Structure containing acceptable oxidation states at all metal atoms are classified as likely free from structural errors, while those possessing one of more flagged metal atom sites indicative of irregular oxidation states are classified as likely to contain structural errors.

The MOSAEC workflow accuracy in flagging erroneous crystal structures and properly computing the metal oxidation states was evaluated through inspection of a manually labelled validation datasets containing thousands of MOF crystal structures.[4] This investigation found that MOSAEC was exceptionally well suited towards the classification of erroneous structures with accuracies of 95% when designating a given crystal structure as chemically invalid, and it possessed reasonable accuracy of approximately 85% when designating a structure as chemical valid. Thus, the MOSAEC designation of the crystal structure validity was used as the primary filter to eliminate faulty structures.

Duplicate Crystal Structures—Examples.

The pointwise distance distribution (PDD) metric utilized to identify duplicated crystal structures in MOSAEC-DB was implemented using the average-minimum-distance[6,7] Python package. Evaluations of the PDD score were only performed on pairs of MOF crystal structures sharing an empirical formula. Careful inspection of hundreds of MOF crystal structure pairs was undertaken to determine an ideal value of the PDD metric to distinguish between strongly matching and unique crystal structures so as to minimize false assignment of duplicate structure pairs. This investigation ultimately determined that PDD values below 0.15 are sufficient to confidently conclude that the structures in question are highly correlated duplicates. An example of one such pair of nearly identical crystal structures reported as distinct entities in the CSD is provided in Figure S4a, while two dissimilar crystal structures exhibiting a high PDD value are shown in Figure S4b.

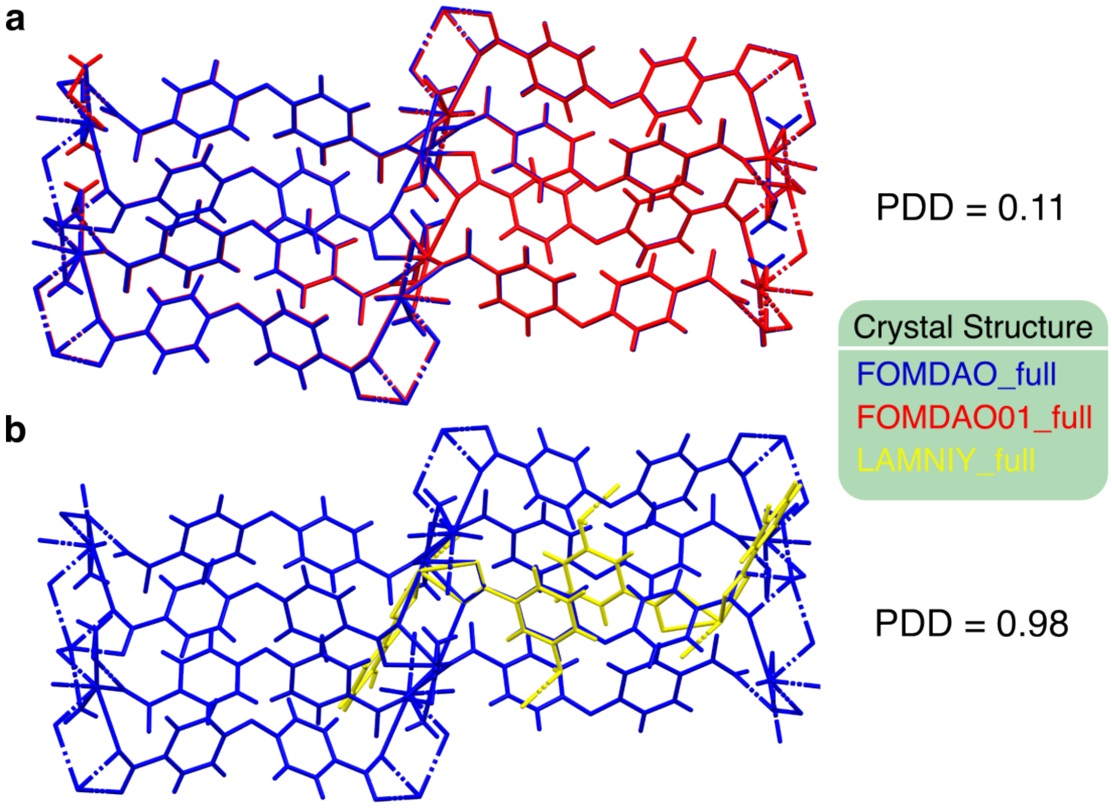

Figure S4. Overlaid crystal structures of pairs of MOSAEC-DB MOFs possessing common empirical chemical formulas and their corresponding pointwise distance distribution (PDD) similarity scores. Instance (a) depicts a case of two near identical crystal structures while instance (b) illustrates disparate crystal structures, which correspond to low and high values of PDD score, respectively.

Additional Chemical Substructure and Metal Site Analysis.

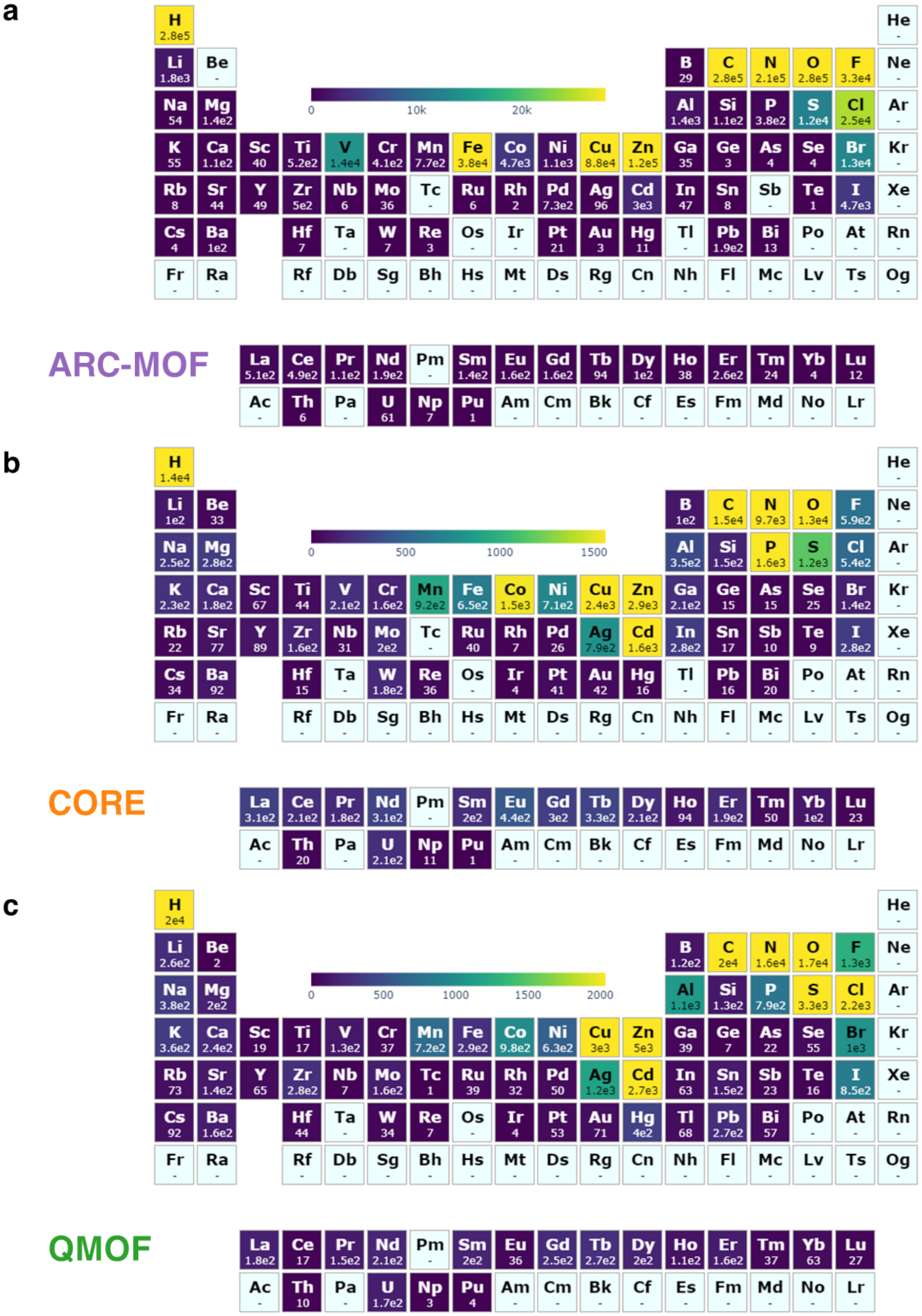

Figure S5. Quantity of crystal structures containing given elements across the periodic table in alternative MOF databases: (a) ARC-MOF, (b) CoRE 2019, and (c) QMOF. A colour gradient maximum equal to 10% of each respective databases' total structure count is enforced to facilitate visualization (generated *via* pymatviz[8]).

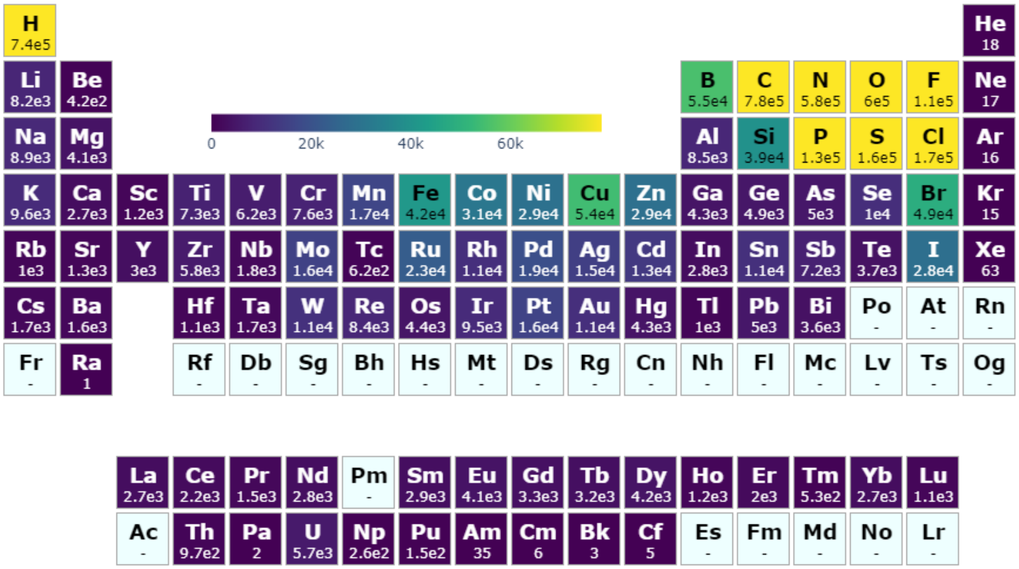

Figure S6. Quantity of crystal structures containing given elements across the periodic table in the entirety of the CSD crystal structure repository (version 5.4.5). A colour gradient maximum equal to 10% of each respective databases' total structure count is enforced to facilitate visualization (generated *via* pymatviz[8]).

Table S1. Comprehensive summary of the chemical substructure frequency analysis within MOSAEC-DB. Analogous chemical substructures—such as primary, secondary and tertiary amines or various heterocyclic aromatics—were considered separately. Any considered substructure with zero occurrence following confirmation between the bond connectivity analyses and Conquest[1] substructure search results was omitted.

Chemical Substructure	*Representation*		*Chemical Substructure*	*Representation*	
	Structure Count	*Frequency* (%)		*Structure Count*	*Frequency* (%)
aromatic	91 843	73.79	sulfate	1 585	1.27
aromatic (N-containing)	56 430	45.34	ketone	1 108	0.89
carboxylate	50 905	40.90	tertiary phosphine	985	0.79
alkoxide	37 822	30.39	alkyne	895	0.72
tertiary amine	15 206	12.22	phosphate	509	0.41
secondary amine	11 196	9.00	thiolate	419	0.34
halide	10 471	8.41	disulfide	396	0.32
alkene	9 306	7.48	imine	327	0.26
halogen	8 001	6.43	aldehyde	71	0.06
ether	7 479	6.01	ester	36	0.03
primary amine	6 380	5.13	nitrite	16	0.01
carboxylic acid	6 180	4.96	sulfonic acid	16	0.01
amide	5 171	4.15	sulfinate	11	0.01
sulfide	3 865	3.11	thiol	6	0.00(4)
nitrate	2 593	2.08	secondary phosphine	2	0.00(1)
alcohol	1 848	1.48			

Table S2. Analysis of the atom identity most commonly observing open metal sites (OMS) within the MOSAEC, ARC-MOF, QMOF, and CoRE 2019 databases.

OMS Structure Frequency			
MOSAEC-DB	ARC-MOF	QMOF	CoRE
Cu (13.56 %)	Zn (38.13 %)	Cu (11.62 %)	Cu (12.23 %)
Zn (7.52 %)	Cu (12.78 %)	Zn (7.97 %)	Zn (10.61 %)
Ag (5.77 %)	Co (1.06 %)	Ag (5.49 %)	Cd (4.60%)
Cd (4.52 %)	Cd (0.78 %)	Cd (2.21 %)	Co (4.26 %)
Co (3.96 %)	Li (0.65 %)	Hg (1.90 %)	Ag (4.07 %)
Mn (2.47 %)	Ni (0.30%)	K (1.40 %)	Mn (2.83 %)
Ni (2.42 %)	Pd (0.26 %)	Pb (1.29 %)	Ni (2.10 %)
Pb (1.85 %)	Fe (0.19 %)	Ni (0.89 %)	Eu (1.49 %)

Additional Geometric Property Analysis.

Table S3. Summary of the geometry property statistics (i.e. mean, standard deviations, and range) computed within the MOSAEC, ARC-MOF, CoRE 2019, QMOF, and CSD MOF databases.

database		volume fraction	LCD (Á)	PLD (Á)	gravimetric surface area (m^2 g^{-1})	volumetric surface area (m^2 cm^{-3})	density (g cm^{-3})
MOSAEC DB	avg.	0.04 ± 0.10	4.1 ± 2.3	2.6 ± 2.0	526 ± 1062	517 ± 866	1.61 ± 0.53
	range	[0.0, 0.84]	[0.0, 42.8]	[0.0, 38.4]	[0, 24961]	[0, 5050]	[0.13, 8.97]
ARC-MOF	avg.	0.37 ± 0.19	11.1 ± 4.8	8.4 ± 3.6	3529 ± 1329	2057 ± 470	0.69 ± 0.30
	range	[0.0, 0.93]	[1.9, 68.9]	[0.5, 67.4]	[0, 8344]	[0, 3727]	[0.06, 3.91]
CoRE	avg.	0.12 ± 0.13	6.7 ± 3.7	4.7 ± 2.8	1245 ± 1130	1298 ± 827	1.38 ± 0.50
	range	[0.0, 0.92]	[2.7, 71.6]	[0.7, 71.5]	[0, 7999]	[0, 3622]	[0.06, 5.18]
QMOF	avg.	0.07 ± 0.15	4.5 ± 4.1	3.0 ± 3.5	638 ± 1271	475 ± 806	1.70 ± 0.65
	range	[0.0, 0.90]	[0.8, 44.9]	[0.0, 44.4]	[0, 6832]	[0, 3147]	[0.09, 5.44]
CSD Collection	avg.	0.11 ± 0.13	6.9 ± 3.6	4.7 ± 3.2	1081 ± 1089	1092 ± 835	1.34 ± 0.43
	range	[0.0, 0.84]	[2.7, 71.6]	[0.5, 71.5]	[0, 6415]	[0, 3630]	[0.13, 4.19]

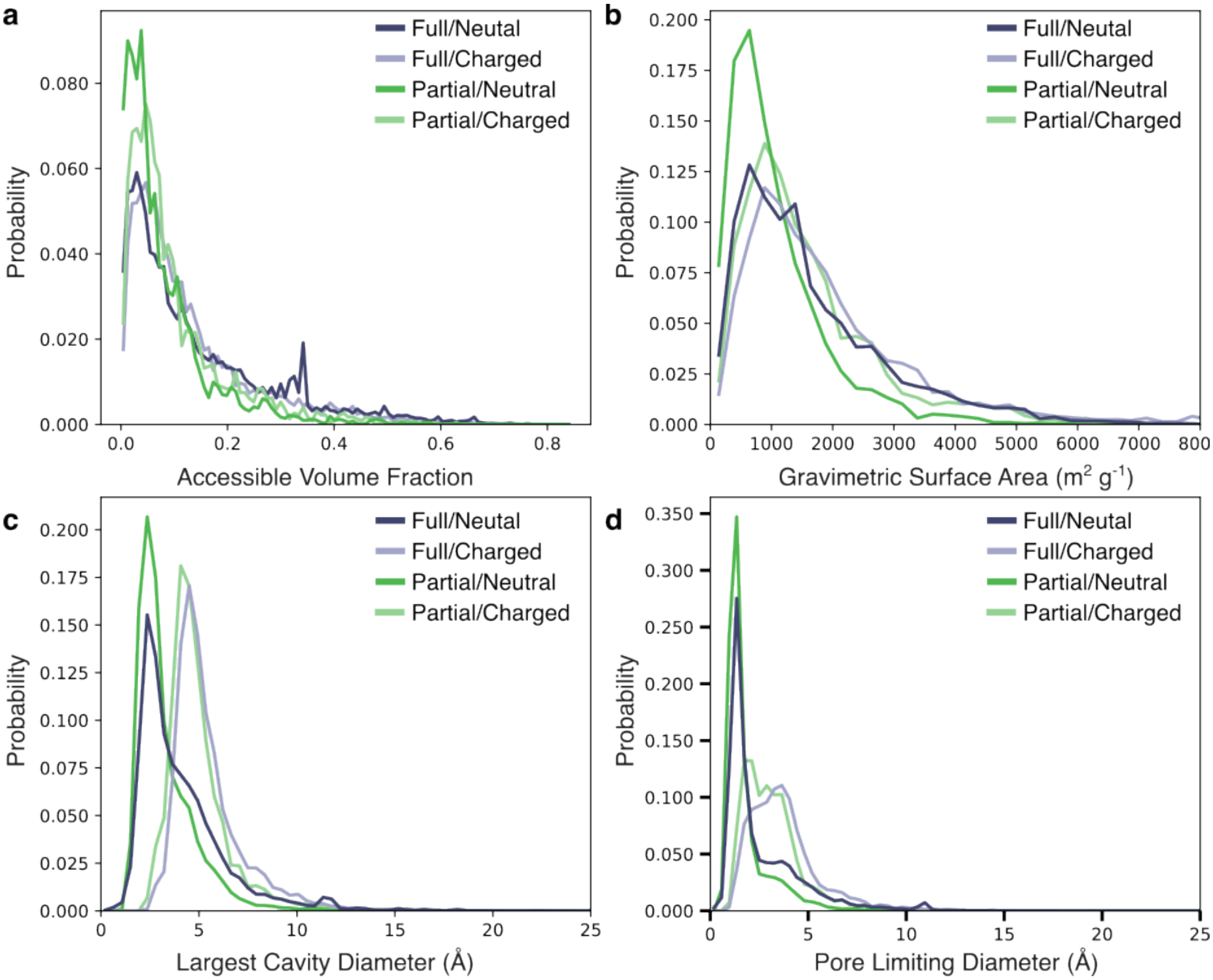

Figure S7. Distributions of geometric properties within MOSAEC-DB plotted according to their activation states (*i.e.* full vs partial) and framework charge status (*i.e.* neutral vs. charged). Distributions of (a) accessible volume fraction, (b) gravimetric surface area, (c) largest cavity diameter (LCD), and (d) pore limiting diameter (PLD) were calculated by Zeo++ using a probe radius corresponding to the kinetic diameter of an H_2 molecule (1.45 Å). Zero values are excluded from the analysis for clarity.

Table S4. Summary of the geometry property statistics (i.e. mean, standard deviations, and range) computed within the various subcategories (i.e. activation and framework charge states) of the MOSAEC database.

MOSAEC DB subset		volume fraction	LCD (Å)	PLD (Å)	gravimetric surface area (m^2 g^{-1})	volumetric surface area (m^2 cm^{-3})	density (g cm^{-3})
Full/ Neutral	avg.	0.05 ± 0.10	4.1 ± 2.476	2.7 ± 2.2	522 ± 1056	510 ± 863	1.63 ± 0.56
	range	[0.0, 0.84]	[0.0, 42.8]	[0.0, 38.4]	[0., 24961]	[0, 5050]	[0.13, 8.97]
Full/ Charged	avg.	0.09 ± 0.12	5.5 ± 1.8	3.7 ± 1.8	1255 ± 1479	1193 ± 1058	1.29 ± 0.44
	range	[0.0, 0.76]	[1.9, 30.1]	[0.6, 28.5]	[0, 11396]	[0, 4852]	[0.23, 4.27]
Partial/ Neutral	avg.	0.01 ± 0.04	3.3 ± 1.6	1.8 ± 1.3	154 ± 482	176 ± 483	1.75 ± 0.44
	range	[0.0, 0.63]	[0.9, 32.1]	[0.1, 31.6]	[0, 5839]	[0, 3911]	[0.33, 5.28]
Partial/ Charged	avg.	0.05 ± 0.08	4.9 ± 1.6	3.2 ± 1.6	791 ± 1141	782 ± 918	1.35 ± 0.44
	range	[0.0, 0.63]	[2.2, 29.1]	[0.6, 28.2]	[0, 7690]	[0, 3728]	[0.30, 4.24]

Additional Structural Diversity Analysis.

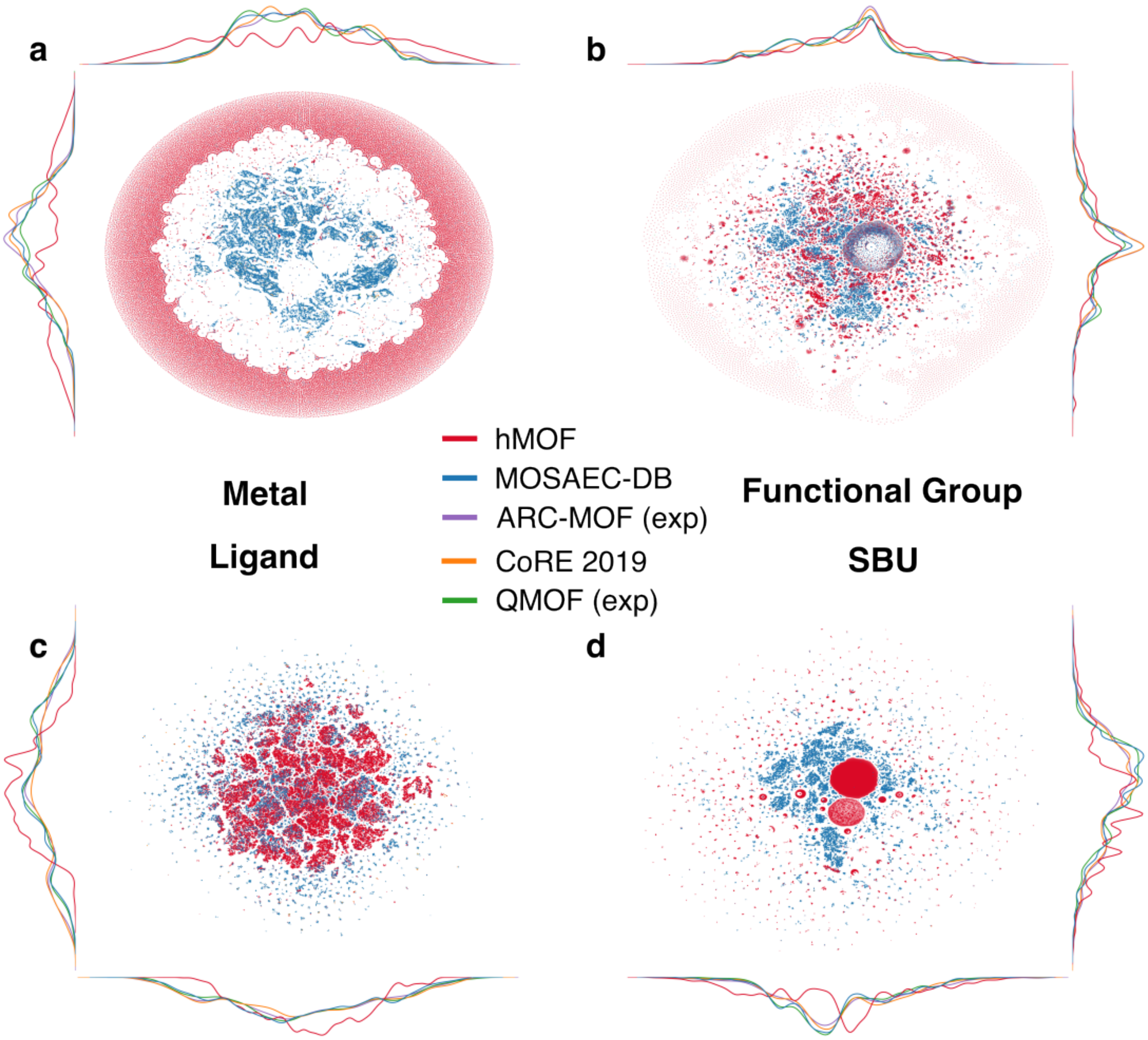

Figure S8. Analysis of the revised autocorrelation (RACs) descriptor space represented in the MOSAEC, ARC-MOF, CoRE 2019, and QMOF databases with the hypothetical MOFs from all databases plotted separately. The (a) metal, (b) functional group, (c) ligand, and (d) SBU RAC descriptors subcategories are distinctly characterized. Dimensionality reduction is performed using the t-SNE algorithm on the combined descriptor space of all relevant structures.

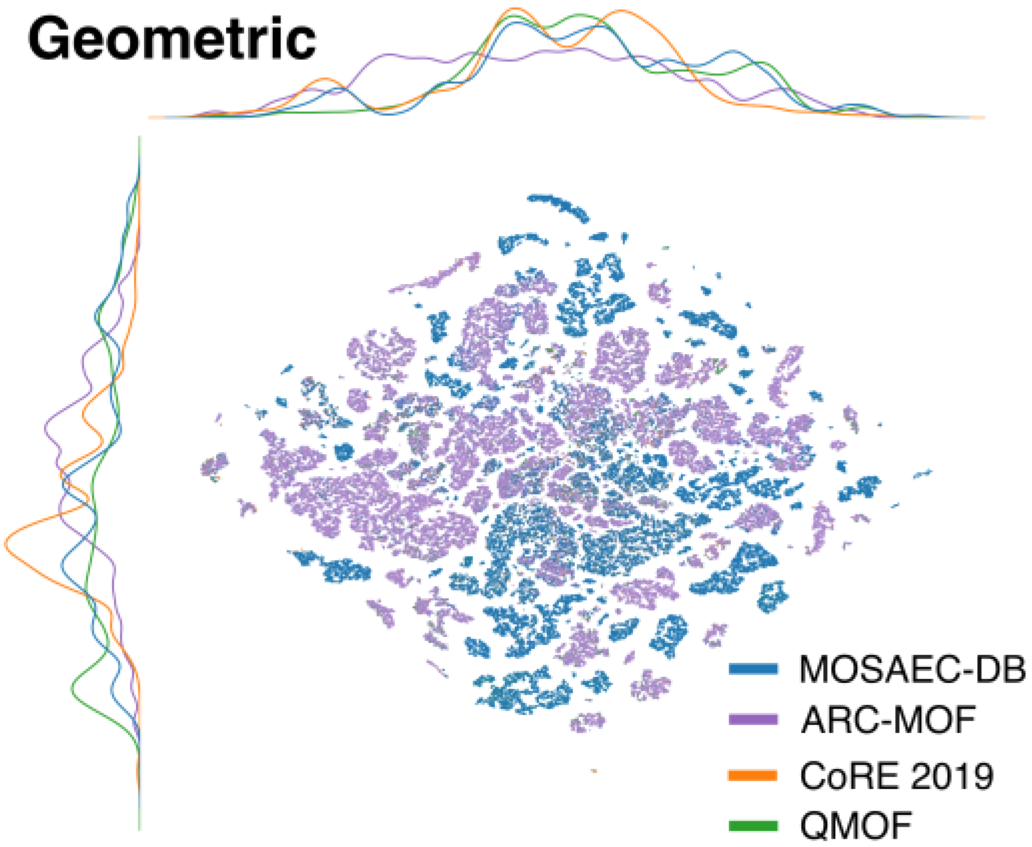

Figure S9. Comparison of the two-dimensional projections of geometric descriptor space computed within the MOSAEC, ARC-MOF, CoRE 2019, and QMOF databases. The geometric properties considered in this analysis include pore-limiting diameter (PLD), largest cavity diameter (LCD), as well as all accessible, non-accessible, and probe-occupiable surface areas, volumes, and volume fractions available within Zeo++[9] pore geometry analysis. Dimensionality reduction is performed using the t-SNE algorithm on the combined descriptor space of all relevant structures.

References

(1) Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R. New Software for Searching the Cambridge Structural Database and Visualizing Crystal Structures. *Acta Crystallogr. Sect. B Struct. Sci.* **2002**, *58* (3), 389–397. https://doi.org/10.1107/S0108768102003324.

(2) 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.

(3) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2016**, *72* (2), 171–179. https://doi.org/10.1107/S2052520616003954.

(4) White, A.; Gibaldi, M.; Burner, J.; Woo, T. K. Alarming Structural Error Rates in MOF Databases Used in Computational Screening Identified via a Novel Metal Oxidation State-Based Method. *ChemRxiv* **2024**. https://doi.org/10.26434/chemrxiv-2024-ftsv3.

(5) Gibaldi, M.; Kapeliukha, A.; White, A.; Woo, T. Incorporation of Ligand Charge and Metal Oxidation State Considerations into the Computational Solvent Removal and Activation of Experimental Crystal Structures Preceding Molecular Simulation. *ChemRxiv* **2024**. https://doi.org/10.26434/chemrxiv-2024-7vq41.

(6) Widdowson, D.; Mosca, M. M.; Pulido, A.; Cooper, A. I.; Kurlin, V. Average Minimum Distances of Periodic Point Sets – Foundational Invariants for Mapping Periodic Crystals. *MATCH Commun. Math. Comput. Chem.* **2022**, *87* (3), 529–559. https://doi.org/10.46793/match.87-3.529W.

(7) Widdowson, D. E.; Kurlin, V. A. Resolving the Data Ambiguity for Periodic Crystals. *Adv. Neural Inf. Process. Syst.* **2022**, *35* (NeurIPS), 1–14.

(8) Riebesell, J.; Yang, H.; Goodall, R.; Baird, S. G. Pymatviz: Visualization Toolkit for Materials Informatics. 2022. https://doi.org/10.5281/zenodo.7486816.

(9) 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* (1), 134–141. https://doi.org/10.1016/j.micromeso.2011.08.020.