ARC-MOF: A Diverse Database of Metal-Organic Frameworks with DFT-Derived Partial Atomic Charges and Descriptors for Machine Learning

Jake Burner, Jun Luo, Andrew White, Adam Mirmiran, Ohmin Kwon, Peter G. Boyd, Stephen Maley, Marco Gibaldi, Scott Simrod, Victoria Ogden, Tom K. Woo*

Department of Chemistry and Biomolecular Sciences, University of Ottawa, 10 Marie Curie Private,
Ottawa K1N 6N5, Canada

*Corresponding author email: Tom.Woo@uottawa.ca

Abstract: Metal-organic frameworks (MOFs) are a class of crystalline materials composed of metal nodes or clusters connected via semi-rigid organic linkers. Owing to their high surface area, porosity, and tunability, MOFs have received significant attention for numerous applications such as gas separation and storage. Atomistic simulations and data-driven methods (e.g., machine learning) have been successfully employed to screen large databases and successfully develop new experimentally synthesized and validated MOFs for CO_2 capture. To enable data-driven materials discovery for any application, the first (and arguably most crucial) step is database curation. This work introduces the ***ab** initio* **R**EPEAT **c**harge MOF (ARC-MOF) database. This is a database of ~280,000 MOFs which have been either experimentally characterized or computationally generated, spanning all publicly available MOF databases. A key feature of ARC-MOF is that it contains DFT-derived electrostatic potential fitted partial atomic charges for each MOF. Additionally, ARC-MOF contains pre-computed descriptors for out-of-the-box machine learning applications. An in-depth analysis of the diversity of ARC-MOF with respect to the currently mapped design space of MOFs was performed – a critical, yet commonly overlooked aspect of previously reported MOF databases. Using this analysis, balanced subsets from ARC-MOF for various machine learning purposes have been identified. Other chemical and geometric diversity analyses are presented, with an analysis on the effect of charge assignment method on atomistic simulation of gas uptake in MOFs.

INTRODUCTION

Metal-organic frameworks (MOFs) are a class of crystalline materials composed of organic and inorganic structural building units (SBUs) and are often synthesized via self-assembly. A hallmark of MOFs is their nanoporosity resulting in record specific surface areas, with a theoretical upper limit of at least 14600 m^2/g.[1] Another key feature of MOFs is their functional tunability, arising from the fact that there is a seemingly infinite number of possible combinations of organic and inorganic SBUs, resulting in a virtually boundless design space. In addition to the chemical diversity of MOFs arising from the identity of the SBUs, the SBUs can be arranged in different net topologies allowing for significant geometric diversity. Thousands of distinct network topologies have been identified in experimentally characterized materials.[2] Due to these exceptional properties (large surface areas, high porosity, and tunability) MOFs have attracted significant interest for a wide variety of applications ranging from catalysis to drug delivery.[3–6] With the urgent and serious concerns of climate change, one high-profile application of MOFs has been in gas separation and storage,[7–11] where MOFs are now being commercialized for industrial scale CO_2 capture from combustion flue gases.[12,13]

Rational design of materials for a specific application is more desirable than trial-and-error development, which can be a long and expensive process. To this end, various atomistic simulation methods such as such as density functional theory (DFT) calculations have become invaluable tools in accelerating the materials and molecular design process.[14–16] For gas separation and storage, atomistic grand canonical Monte Carlo (GCMC) simulations have been shown to provide relatively accurate estimates of the gas adsorption properties of MOFs if the structure is known.[16] The rapid increase in materials data and vast improvements in computer hardware have enabled the application of high throughput screening and data driven methods to further accelerate materials discovery.[17–19] Recently, machine learning (ML) and other data-driven methods have been used to successfully develop new experimentally synthesized and validated MOFs for CO_2 capture.[20] To enable data-driven materials discovery for any application, the first (and perhaps most crucial) step is the curation of a database. To date, tens of thousands of MOFs have been synthesized,[21] as recorded by the Cambridge Crystallographic Data Centre (CCDC). The CCDC collates and curates the Cambridge Structural Database (CSD), a repository of small-molecule and metal-organic crystal structures. In 2017, the CCDC collected MOF and MOF-like structures into a separate database called the CSD MOF Subset.[21] However, the procedure used to generate this database led to the inclusion of a large fraction of MOFs which are not porous, even though permanent porosity is a key requirement for many MOF applications such as gas separation. A smaller, yet more popular, database named the Computation-Ready Experimental (CoRE) MOF database, was created in 2014 composed of ~5,000 MOFs from the CSD.[22] In 2016, Nazarian *et al.* created the “CoRE MOF 2014-DDEC” database, in which DFT-derived partial atomic charges were determined for about half of the CoRE MOF 2014 database.[23] The database was initially validated by using GCMC simulations of methane adsorption. An advantage this database presented over others is that it included DFT-derived charges, which enable improved accuracy of GCMC simulations when modeling adsorption of molecules where electrostatic interactions are important, such as CO_2. A subsequent study performed a DFT-level structural relaxation of the MOFs in the CoRE 2014 database to investigate the effect of solvent removal on the adsorption properties of the materials.[24] The CoRE 2014 database was recently expanded to ~14,000 MOFs from the CSD (CoRE 2019 Database), and is currently the largest source of permanently porous experimentally characterized MOFs.[25] Chen and Manz screened the CoRE 2019 database to determine structures with isolated atoms, overlapping atoms, and hypo/hypervalent atoms.[26] Notably, the authors identified thousands of chemically invalid structures, demonstrating the requirement for a more thorough structure-checking procedure.

Databases of experimentally realized MOFs such as the CoRE and CSD databases are preferable for identifying high-performing materials since they contain synthetically feasible structures. However, the number of permanently porous MOFs suitable for gas separation is still rather limited. Thus, databases of computationally generated or hypothetical MOF (hMOF) structures have been created to explore more combinations of SBUs than currently exist for experimentally synthesized MOFs.[20,27–30] The approaches to generate hMOF structures can be broadly classified as either "bottom-up" or "top-down". Bottom-up approaches involve sequentially connecting SBUs until a periodic crystal structure is obtained. Top-down approaches start with a given topology and map SBUs onto the topological net to generate a periodic crystal structure. In 2012, Wilmer *et al.* developed a bottom-up approach which entailed recombining 102 building blocks derived from already-synthesized MOFs.[27] Using predictable assembly of building blocks into MOFs, the authors generated every possible structure within specified constraints (including geometric rules from existing MOFs), resulting in a database ("Wilmer database") of ~138,000 hMOFs arranged in six topologies. This pioneering work resulted in the first large database of MOFs to be constructed and was initially screened for methane uptake. Gurnani *et al*. recently used the database to generate ML models to predict methane adsorption isotherms of MOFs.[31] The authors expanded the diversity of the database by substituting new metal clusters into the existing structures. Surprisingly, the model could be successfully extended to metal chemistries absent from the training set. In 2016, Boyd and Woo developed the topology-based crystal constructor (ToBasCCo) algorithm – a top-down approach which uses graph theory to generate hMOFs using only an underlying topological net, a combination of SBUs, and the bond-forming sites of the SBUs.[32] Importantly, this method allows for the generation of a more geometrically diverse set of structures. Using this algorithm, a MOF database ("Boyd-Woo database") containing over 300,000 structures and over 1000 net topologies has been constructed.[20,28] Boyd *et al*. later screened and mined the database to identify strong, hydrophobic

CO_2 binding sites which led to the experimental synthesis of water-stable MOFs for post-combustion CO_2 capture that still functioned well at high relative humidities.[20] Other top-down and bottom-up databases of hypothetical MOFs have also been developed, including databases where structures were created by functionalizing the pores of a parent MOF.[33] While hMOF databases allow for access to a wider combination of organic and inorganic SBUs, they frequently possess dramatically poorer diversity with respect to the inorganic SBUs when compared to experimental databases.[34,35] Depending on the property of interest, using such databases could introduce unwanted biases which reduce efficiency of screening and/or transferability of machine learning models. Thus far, few studies have focused on correcting this lack of diversity in hMOF databases, with the exception of the work by Majumdar *et al.*,[34] in which new hMOFs were generated with underrepresented inorganic SBUs to expand the metal chemistry diversity of the existing hMOF space.

The computational screening of MOFs for gas separation and storage applications requires the determination of adsorption properties of the materials, which is usually performed via atomistic GCMC simulations. These simulations use a set of parameters and equations to model interatomic potentials (i.e., a "force field"). Most commonly, fixed partial atomic charges are used to model the electrostatic interactions. Although many force field parameters are generally transferrable from one MOF to the next, the partial atomic charges of a specific MOF framework are not. As such, before a GCMC simulation can be performed, one must determine partial atomic charges for the MOF framework. Ideally, these charges would be derived from first principles DFT calculations, especially when electrostatic interactions play a key role in adsorption of a particular guest molecule. The most appropriate partial atomic charges for evaluating interaction energies are so-called electrostatic potential fitted charges where the atomic charges of the system are selected such that they optimally reproduce the electrostatic potential determined from the DFT calculation. The REPEAT method[36] was the first method for deriving such charges from periodic DFT calculations. However, MOFs typically have hundreds of atoms per unit cell. As such, a periodic DFT calculation to determine the wave function and electrostatic potential can take hours on modern computing resources. Therefore, when performing a large-scale screening of a database using GCMC simulations, using DFT-derived charges can become prohibitively expensive.[37] This problem prompted the use of empirical charge-assignment methods, such as the charge equilibration (QEq) method[38] and the split-charge equilibration (SQE) method,[39] which allow charges in MOFs to be generated within seconds. The drawback of these methods is that they must be parameterized to provide a good approximation for the electrostatic potential of the material. Several parameterizations of these methods have been developed for MOFs[40,41] including the MEPO (MOF electrostatic potential optimized) sets[42,43] which were fit to reproduce DFT-computed electrostatic potentials (ESPs) from a large training set of MOFs. The MEPO-QEq parameters have been used to generate partial atomic charges for a number high-throughput screening studies[44–48]. While significant effort has gone into optimizing parameters for these empirical charge assignment methods, the use of these methods on materials which deviate from structures in the original training sets will yield questionable charges.[37] More recently, machine learned (ML) partial atomic charge models have been developed for MOFs by Raza *et al*.[49] and Kancharlapalli *et al*.[50] that both provide much improved results over the QEq and SQE methods. Both models were trained to reproduce DFT-derived DDEC charges on MOFs from the CoRE MOF 2014-DDEC database. In the study by Raza *et al.*, a set of 2,266 charge-labeled MOFs were partitioned into training, development, and test sets in a 70:10:20 split, while Kancharlapalli *et al.* used 2,974 MOFs in a 20:80 split for the test and training sets, respectively. Notably, Kancharlapalli *et al.* removed atoms with identical chemical environments to eliminate duplicate atoms in the training and test sets. To ensure broad applicability of these ML models, it would be ideal to train ML models on large and diverse datasets.

Even with computed charge parameters available, GCMC simulations typically require millions of guest-host interaction energy calculations to determine a single point on an adsorption isotherm. As such, this type of screening falls under the category of “brute force” methods. Brute force screening of large databases of materials can range from extremely compute-intensive to intractable. Consequently, ML has been used to create models which directly relate the structure of a material to a target property (e.g., CO_2 uptake). Such models would bypass the need for a GCMC simulation altogether and allow one to compute a property for a given material at rates which are orders of magnitude faster than traditional methods. ML models can learn how to relate chemical or structural descriptors of a MOF to its performance. Descriptors should, at a minimum, uniquely characterize the material and be easier to calculate than the property itself. A descriptor should also be invariant to rotations, translations, and atom ordering. Many studies have focused on using ML to predict adsorption properties of MOFs.[45,51,52] However, one major step in these studies entails the calculation of the descriptors. An ideal database of MOFs would be "ML-ready", with pre-computed descriptors and target data to facilitate the development of new ML models.

In this paper, we introduce a diverse database of 280,000 experimentally characterized and hypothetical MOFs from various sources with DFT derived partial atomic charges for GCMC screening, as well as descriptors and simulated adsorption data that can be used for developing machine learning models. We call this database the ***ab** initio* **R**EPEAT **c**harge (ARC-MOF) database, and it includes DFT derived REPEAT charges for each MOF and 515 descriptors that encompass three different classes, for each material in the database. These descriptor classes can be generalized as geometric, atomic-property-weighted radial distribution functions (AP-RDFs)[53], and revised autocorrelations (RACs)[35] . The RACs and geometric descriptors are then used in an in-depth diversity analysis to evaluate the chemical diversity of ARC-MOF against the current known design space of MOFs – a critical, yet commonly overlooked detail in past publications of MOF databases. Other chemical and geometric diversity analyses are presented, with an analysis on the effect of charge assignment method on GCMC simulated gas uptakes in MOFs.

COMPUTATIONAL METHODS

Partial Atomic Charge Calculations

DFT-Derived Partial Atomic Charges: Partial atomic charges were calculated from periodic density functional theory (DFT) calculations performed on each MOF using the REPEAT method.[36] REPEAT fits partial atomic charges of each atom such that they collectively minimize the difference between the electrostatic potential (ESP) resulting from a DFT calculation of the system and the ESP from the fitted charges. Periodic DFT calculations were performed on all MOFs using VASP version 5.4.4[54,55] with the Perdew–Burke–Ernzerhof (PBE) functional[56] and the projector augmented-wave (PAW) method.[57] A 3x3x3 Monkhorst-Pack sampling of k-points in the Brillouin zone was used for all MOFs whose smallest cell vector length is less than 14 Å. For MOFs whose cell vectors were all greater than 14 Å only the Γ-point was sampled. A random sampling of 3000 MOFs whose minimum cell-vector length was between 13.5-14.0 Å, showed that the REPEAT charges calculated with a Γ-point sampling had a mean absolute deviation (MAD) from those determined with a 3x3x3 sampling of k-points of only 0.0024 e. The same comparison on a random sampling of 3000 MOFs whose minimum cell-vector length was less than 7 Å was found to have a MAD of 0.0156 e, which is an order of magnitude higher. Since insufficient Brillouin zone sampling is more acute for smaller cells, this shows that our choice of sampling only the Γ-point for MOFs with cells vectors larger than 14 Å is justified. The VASP calculations used a planewave cut-off of 300 eV and a lower than default FFT grid density (VASP keyword 'PREC=Low') to save computing time. To justify this choice of lower precision settings for the DFT calculations we have calculated the REPEAT charges for a set of MOFs using a 400 eV plane wave cut-off, a 'PREC=Normal' setting, and a 6x6x6 Monkhorst-Pack sampling of k-points. The REPEAT charges derived from the higher precision calculations were then compared to the charges calculated using the low precision settings (PREC=Low and a 3x3x3 sampling of k-points for MOFs possessing a cell vector less than 14 Å and Γ-point sampling otherwise). For a diverse set of 111 MOFs taken from various databases, with 16 different metals (Cu, Ni, Fe, Mo, Cd, Zn, V, Co, Mg, Mn, In, B, Zr, Ag, Al and La) and at least 12 different topologies, the difference in REPEAT charges using the two precision settings was found to have a MAD of only 0.0032 e over 10572 atomic centers. Further, the average maximum absolute deviation over all MOFs was only found to be 0.015 e. For the REPEAT calculations, no restraints were used but the default van der Waals radii were scaled by a factor of 0.9.

Only MOFs that were expected to have neutral frameworks were selected from the various sources for ARC-MOF, such that MOFs that were explicitly labelled as charged were excluded. Additionally, most MOFs were assumed to be closed-shell and calculated using a restricted-DFT formalism. Any MOFs containing the elements Cr, Mn, Fe, Co, Ni, Gd, Dy, Ho, Er that commonly have unpaired electrons were calculated using an unrestricted-DFT formalism.

Empirical Partial Atomic Charges: To determine the effect of charge-assignment method on adsorption properties of MOFs, MOF electrostatic potential optimized (MEPO) empirical charge assignment methods were compared to REPEAT charges. For these charges, the charge equilibration (MEPO-QEq[58]) and split charge equilibration (SQE-MEPO[43]) methods were used, and are described in detail elsewhere.

Structure Validation

To ensure ARC-MOF contains only chemically reasonable structures, the oxidation state of each metal atom in each structure was determined. Structures containing metals with impossible, unknown, and/or non-integer oxidation states were excluded from this study. Structures which possessed atom-pair distances of less than 70% of the sum of their covalent bond radii were discarded as bad to avoid including structures which possess overlapping atoms. Lattice parameters were also checked to avoid structures with unrealistically small cell vectors. The smallest cell vector in the CSD MOF database which only contains experimentally synthesized MOFs, was found to be 3.22 Å[59]. Therefore, only MOFs with unit cell vectors greater than this length were included in ARC-MOF. Finally, structures with hypercoordinated main-group elements were discarded. For this check, structure graphs were generated for each MOF with the metal atoms removed, since metals frequently involve hypercoordinated oxygen and hydrogen. Then, MOFs containing carbon atoms with more than 4 bonds, oxygen atoms with more than two bonds, and halogen/hydrogen atoms with more than one bond were discarded. For experimental MOFs containing counterions, perchlorate and hydronium ions were excluded from this check.

Chemical Substructure Identification

Trends in the incidence of various chemical substructures (i.e., chemical moieties and organic functional groups) in ARC-MOF were determined using bond connectivity data in the crystallographic information files (CIFs) of the structures. For structures which had CIF files missing bond connectivity tables, the requisite bond connectivity tables were generated using the nearest-neighbour algorithm proposed by Isayev *et al.*[60], as implemented in the Python cheminformatics library pymatgen (version 2022.0.5)[61]. In this nearest-neighbour algorithm, adjacent atoms are considered bonded if they shared a Voronoi facet, and the distance between them was less than their summed Cordero covalent radii[62] plus an additional bond tolerance value. A bond tolerance value of 0.25 Å was selected to represent only strong interatomic interactions[62] and to limit hypercoordination at carbon atom and metal atom sites. The substructures were extracted by identifying bonded atom pairs within its connectivity table, then evaluating each atom's first coordination sphere. Only substructures not bound to metal atoms were considered in this analysis. These fragments were compared against a representative library of chemical substructures to identify which ones were present within each ARC-MOF structure.

Descriptor Calculations

Geometric Descriptors: Geometric descriptors of all materials were calculated with Zeo++[63] version 0.3.0 with the high accuracy flag and a probe radius of 1.86 Å. The accessible surface area, accessible volume, and probe-occupiable volume were calculated with 2000, 50000, and 2000 Monte Carlo steps, respectively. A list of the 20 geometric properties evaluated for each MOF is given in the supporting information.

Revised autocorrelation function descriptors: RACs are provided as descriptors for ML studies and are additionally used in this work to evaluate the diversity of our database in comparison to other MOF databases. The RACs were computed according to the methodology described by Moosavi *et al*.[35] RACs are products or differences of properties computed on a non-weighted chemical graph. An example of a difference-based RAC is:

$$ {}_{scope}^{start}P_d^{diff} = \sum_{i}^{start} \sum_{j}^{scope} \left(P_i - P_j\right) \delta(d_{i,j}; d) $$

Where d is the "depth" of the RAC, equal to the bond-wise path distance from the starting atom and $d_{i,j}$ is the bond-wise path distance between atoms i and j. δ is the Kronecker delta, equal to unity when $d_{i,j}$ is equal to d, and equal to zero otherwise. P_i and P_j represent atomic properties of atom i and j, respectively. The start and scope lists for each type of RAC and the RACs included in this work are given in the supporting information. The properties considered for the RACs in this work are electronegativity, nuclear charge, atom identity, connectivity, and covalent radii. For the linker connecting atom and functional group RACs, an additional property (polarizability) is included. To compute the RACs, the molSimplify code written by Kulik and coworkers was used.[64] The start and scope atom lists used in this work are given in Table S2.

Atomic-property-weighted radial distribution function (AP-RDF) descriptors: An in-depth definition of AP-RDF descriptors[53] and their variants is given elsewhere[45]. Briefly, the AP-RDF descriptor gives an atomic property weighted probability of finding atom pairs separated by a given distance inside a unit cell of a MOF. In this work, we define the AP-RDF descriptor by the following equation:

$$ RDF^P(R) = f \sum_{i,j}^{atom\ pairs} P_i P_j e^{-B(r_{ij}-R)^2} $$

Where r_{ij} is the minimum image distance of all atom pairs in the MOF, P_i and P_j are atomic properties of atom i and j, respectively, B is a Gaussian smoothing parameter, and f is a normalization factor. In previous work,[53] the optimal value of B was found to be 10. Each AP-RDF was normalized by the number of framework atoms in the unit cell (N_{atoms}) and scaled by a factor of 0.001 (i.e., $f = 0.001/\mathrm{N_{atoms}}$). The AP-RDF descriptors were calculated using a distance range of 2.0–30.0 Å in 113 steps that linearly increase from 0.004425–0.5 Å, yielding a total of 339 AP-RDF descriptor values for each MOF.

Diversity analysis

The analysis presented herein was accomplished by expanding upon the methodology proposed by Moosavi *et al*.[35] The chemistry-specific RAC descriptors for a) metal-centre chemistry; b) ligand chemistry; and c) functional group chemistry, as well as the six geometric descriptors shown in Table 2 were employed in this analysis. RACs are graph-based descriptors which describe the chemistry of the various "building blocks" of MOFs. In previous work, a depth of 2 was used to describe the metal chemistry. In our opinion, this is too local to the metal, as there can be some chemically distinct local metal environments if one considers a depth greater than two. An example of this is a metal bound to a substituted versus unsubstituted pyridine ligand. For these reasons, a depth of three was chosen for all RAC descriptors used in this analysis. The Uniform Manifold Approximation and Projection (UMAP) technique[65] as implemented in the RAPIDS cuML Python library[66] was used to reduce the high dimensional data for each type of descriptor to only two dimensions for visualization purposes (e.g., the 20-dimensional metal-centre RAC descriptors were reduced to two dimensions using UMAP). Hyperparameters used for the generation of these plots are given in the supporting material. Due to practical limitations of using UMAP on large datasets, a random set of 50K MOFs was selected to represent the entire set of 480K MOFs (~10% of the entire dataset) for the UMAP plots. While these plots allow for qualitative analysis of the data, quantifiable metrics are desirable. For this purpose, three metrics used by Moosavi *et al*.[35] were employed in this analysis. The first metric, disparity (D), quantifies the spread of the data compared to another dataset. Thus, this metric uses the UMAP plots to obtain a ratio between the 2D area occupied by the coloured points (ARC-MOF) and the 2D area occupied by all the points (entire design space). This metric is equal to unity for a subset which spans the exact same space as the overall set. The other two metrics, variety (V) and balance (B) depend on the determination of clusters. HDBSCAN[67] as implemented in the hdbscan Python library[68] was used to cluster the MOFs based on RAC descriptors (see SI for the clusters of MOFs). Consequently, these cluster-based metrics are independent of the UMAP plots (unlike the disparity). Variety is a measure of how many clusters are occupied by the subset versus the number of total clusters and is equal to unity for a subset which has at least one structure in each cluster. Finally, the balance is computed using Pielou's evenness, a well-behaved transformation of Shannon entropy, which is equal to unity for a completely balanced dataset. Therefore, this is the only absolute metric (i.e., does not involve a comparison of the subset to the total set). Detailed information on the calculation of these metrics, as well as the code to compute these metrics, can be found in the SI. Farthest point sampling[69] (also known as max/min sampling) was used to select a more balanced set of structures for each unbalanced type of chemistry in ARC-MOF (i.e., functional group chemistry and metal chemistry). Balanced subsets of ARC-MOF are provided in the SI, as well as code to perform the farthest point sampling.

Grand Canonical Monte Carlo Simulations

Grand canonical Monte Carlo (GCMC) simulations were performed using an in-house code based on the DL_POLY molecular dynamics package to calculate equilibrium gas adsorption properties. Guest-framework interactions were modeled using Lennard-Jones (LJ) potentials to account for steric and dispersion interactions, and the fixed partial atomic charge model was used to model non-bonded electrostatic interactions. The LJ parameters for the frameworks were assigned from the UFF force field. The REPEAT method, QEq and SQE methods were used to assign partial charges to the framework atoms to allow for comparison between the charge-assignment methods. Lorenz-Berthelot mixing rules were used for the determination of LJ parameters between atoms of different types. The intermolecular potential parameters for CO_2 were taken from Garcia-Sánchez *et al.*[70] The parameters for H_2S were taken from Kamath *et al.*[71] The parameters for N_2 were developed in-house to reproduce experimental N_2 uptake isotherms in MOFs.[72] The parameters for CH_4 were taken from Martin *et al.*[73] Finally, the parameters for H_2 were taken from Belof *et al*.[74] For the GCMC simulations used in comparison of charge-assignment methods, 200,000 Monte Carlo (MC) steps were performed, split evenly between equilibration and production. 10,000 MOFs were selected to perform the charge-assignment comparison. Additionally, GCMC screening of ARC-MOF was performed to obtain gas uptake relevant to five different gas separations. For these simulations, 10,000 MC cycles were performed, split evenly between equilibration and production, and the REPEAT charges reported in this work were used for the partial charges of the framework atoms. The gas uptake data with standard deviations, heat of adsorption data, as well as working capacity and selectivity for these gas separation conditions is given in the supporting information and is intended for use as target properties for future ML studies. The following processes were used to determine the relevant gas separation conditions given in the SI: a) methane purification; b) post-combustion VSA; c) precombustion PSA; d) methane storage PSA; and e) landfill gas VPSA.

RESULTS AND DISCUSSION

ARC-MOF Composition: ARC-MOF contains both experimentally characterized and hypothetical MOF structures taken from several different sources. For simplicity, we will refer to each source as a database, and have assigned each a unique label – DBx where 'x' is a identifier number starting from zero. Table 1 summarizes the composition of each database used to curate ARC-MOF, including the number of organic SBUs, inorganic SBUs, functional groups, and topologies used to construct the MOFs in each database, if available. The total number of MOFs from each database is given, with the number of MOFs remaining after structure checking, and the number of MOFs from each database present in ARC-MOF. CIF files of all structures in ARC-MOF are available (see SI), with the filename comprising the 'DBx-' prefix from Table 1 followed by the name of the structure given in the original source. DB0 represents the hypothetical database previously developed by our group, often referred to as the Boyd-Woo database[20]. This is a database constructed from both the top-down and bottom-up MOF construction methodologies. However, most structures in the database were made using the top-down ToBasCCo code developed by Boyd and Woo.[32] All structures in the Boyd-Woo database were optimized with the UFF force field[75] where the geometry of the metal and atoms directly bonded to the metal were frozen. DB1 is an hMOF database constructed by Lan *et al.*[30] using their top-down crystal construction algorithm[76] to generate hypothetical ionic liquid/MOF composites. DB2 was created by Colón *et al.*[29] using their top-down MOF builder code Topologically Based Crystal Constructor (ToBaCCo)[29] and optimized the structures using molecular mechanics. DB3 contains hMOFs generated using ToBaCCo by Anderson *et al.*[77],optimized using UFF. The organic SBUs used in this database consist only of functionalized benzene dicarboxylate (BDC) ligands, with the following functional groups: amino, bromo, fluoro, hydroxyl, nitro, methyl, thiol and trifluoromethyl groups. DB4, reported by Gómez-Gauldrón *et al*,[78] consists of hypothetical MOFs featuring $Zr_6O_4(OH_4)(CO_2)_{12}$ and $Zr_6O_4(OH_4)(CO_2)_8(OH)_4$ SBUs. 12 ditopic organic SBUs were combined with the inorganic SBUs to create MOFs with the fcu topology and 36 tetratopic organic SBUs were combined with the inorganic SBUs to create MOFs with ftw, csq, and scu topologies. The MOFs were built and optimized in Materials Studio[79] using the Crystal Builder and Forcite modules, respectively. DB5 is a subset of the Wilmer database[27] containing only MOFs with a unique combination of interpenetration capacity, actual interpenetration level, inorganic node, primary linker, secondary linker, and functional groups, as curated by Chung *et al*[80]. DB6 contains hMOFs based on the Cu-paddlewheel SBU possessing *pcu* topology.[81] An additional 560 unfunctionalized parent versions of the MOFs were also generated. DB7 is an hMOF database generated by Majumdar *et al.*, with the focus of using inorganic SBUs not commonly observed in hMOF databases.[34] DB8 is a set of hMOFs reported by Anderson and Gómez-Gauldrón, in which the Cu-paddelwheel SBU was combined with tetratopic organic SBUs using a modified version of ToBaCCo. The resulting structures were optimized in LAMMPs using the UFF4MOF force field. DB10 contains hMOFs created using ToBaCCo and optimized using UFF.[82] DB12 is the CoRE 2019 database.[25] Bao *et al.* created DB13, which contains MOFs comprised of organic SBUs evolved using a genetic algorithm to maximize deliverable methane capacity from an initial population of 57,815 commercially available molecules.[83] DB14 is the CSD-MOF database that contains structures from the non-disordered subset of the Cambridge Structural Database.[21] Only structures with the H-atoms added were sampled from and structures with the same CSD code that we included from the CoRE databases were not added to ARC-MOF to avoid duplicate structures. DB15 refers to new structures created for this work using a top-down approach as implemented in Pormake[84]. To sample MOFs from each database, the structures were sorted based on the number of atoms in unit cell. DFT calculations were first calculated on the smallest members of each database moving to larger structures incrementally until the DFT calculations were no longer feasible with the computing resources available. For DB1, which contains over 300K MOFs, the above procedure was followed on a random subset of ~75000 structures. Structures where the wave function would not properly converge were discarded. All MOFs in ARC-MOF were assumed to have a charge neutral framework, and so MOFs from the various sources that were specified as being charged were not sampled.

Table 1. The labeling used to identify sources of structures to construct the ARC-MOF database, compositions, the number of MOFs remaining after the structure checking procedure, and the number of MOFs present from each database in ARC-MOF. The percentages in parentheses represent the percentage contribution of each database to ARC-MOF. Information on database composition is given only for hMOF databases.

		Database Composition					**No. of MOFs after structure check**	
label	**Reference**	**No. organic SBUs**	**No. inorganic SBUs**	**No. functional groups**	**No. topologies**	**Total No. of MOFs**		**No. of MOFs in ARC-MOF (%)**
DB0	Boyd *et al.*[20,32]	175	23	50	1,166	358,398	263,218	203,025 (72.2)
DB1	Lan *et al.*[30]	32	17	9	18	303,992	181,885	23,267 (8.3)
DB2	Colón *et al.*[29]	60	14	0	41	13,514	3,920	199 (0.1)
DB3	R. Anderson *et al.*[77]	12	4	8	15	426	358	123 (0.0)
DB4	Gómez-Gauldrón *et al.*[78]	48	2	--	4	204	48	25 (0.0)
DB5	Chung *et al.*[80]	82	5	12	6	51,163	27,022	22,366 (8.0)
DB6	Li *et al.*[81]	41	1	3	1	11,555	10,944	9,092 (3.2)
DB7	Majumdar *et al.*[34]	95	14	--	52	23,891	12,316	6,955 (2.5)
DB8	R. Anderson and Gómez-Gauldrón[85]	7	1	--	19	126	122	8 (0.0)
DB10	G. Anderson *et al.*[82]	5	4	--	17	105	78	22 (0.0)
DB12	CoRE 2019[25]	--				15,613	8,379	7,145 (2.5)
DB13	Bao *et al.*[83]	100	5	--	9	8,629	6,180	5,165 (1.8)
DB14	CSD MOF[21]	--				10,636	4,070	72 (0.0)
DB15	This work	195	19	--	37	7,708	2,841	2,146 (0.8)
Total		--				806,520	521,381	279,610

Chemical and Geometric Diversity of ARC-MOF

We now examine both the chemical and geometric diversity of ARC-MOF, with a comparison to other databases of MOFs. Figure 1 shows the distribution of common geometric descriptors within ARC-MOF in comparison to the CoRE 2019,[25] CSD-MOF,[86] QMOF[87], and Majumdar[34] databases. The distributions plotted are for the entire databases. Since it is difficult to determine the maxima and minima of the geometric parameters from the plots shown in Figure 1, these are given in Table 2 for each database.

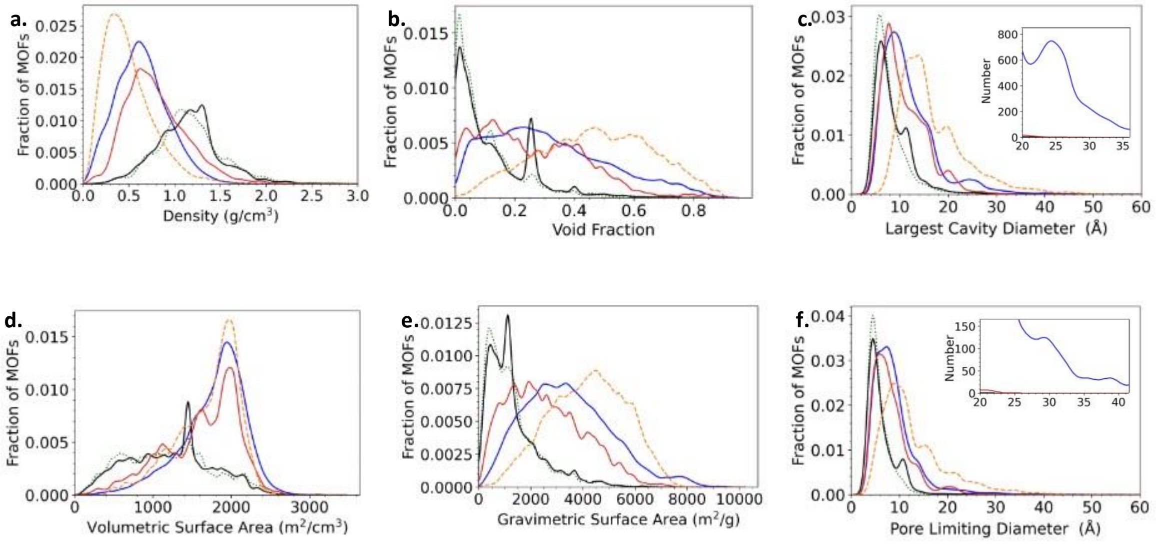

Figure 1. Distribution of common geometric parameters within the ARC-MOF database (solid blue), the full CoRE 2019 database[25] (dotted green), the CSD MOF database[88] (solid black), the QMOF database[87] (solid red), and the Majumdar database[34] (dashed orange). The insets in c) and f) show the absolute number of MOFs. The fraction of structures with zero surface areas, void fraction or diameters is not plotted, although they are included in the calculation of the distributions.

Examination of the plots in Figure 1 reveals that the databases containing hypothetical MOFs (QMOF, Majumdar and ARC-MOF) have distributions of geometric parameters that are more similar to one another compared to purely experimental MOF databases (CoRE and CSD-MOF). The hypothetical databases have distributions that are more skewed towards structures with larger pore diameters, surface areas and void fractions, indicating that they contain materials that are more porous than the two purely experimental databases. This is consistent with the fact that the experimental MOF databases are also higher in density, on average. It would be prudent to specifically compare the ARC-MOF and QMOF databases since DFT calculations have been performed on all materials in each. In general, the two databases have similar distributions for the six geometric parameters shown in Figure 1. Perhaps the most apparent difference is that ARC-MOF has a more normal distribution of void fractions and gravimetric surface areas as compared to QMOF which has more even distributions for these parameters in the low value regimes. Close inspection of the distribution of largest cavity diameter and pore limiting diameter might suggest that the QMOF database has more structures with larger pores. However, it is important to realize that ARC-MOF is a much larger database than QMOF. The inset plots in Figures 1c and 1f show the absolute numbers of MOFs for those parameters, rather than the fraction. (Figure S1 provides plots of the absolute values for all six geometric parameters). This reveals ARC-MOF has a greater number of MOFs over the ranges of large diameters than QMOF. Additionally, Table 2 shows that ARC-MOF has maximum diameters that are more than twice that of QMOF.

Table 2. Maxima and minima of selected geometric properties for ARC-MOF and other MOF databases.

Database		density (g/cm^3)	volumetric surface area (m^2/cm^3)	gravimetric surface area (m^2/g)	volume fraction	largest cavity diameter (Å)	pore limiting diameter (Å)
ARC-MOF	max	6.20	3474.6	10218.6	0.95	83.1	81.1
	min	0.02	0	0	0	1.6	0.1
CoRE	max	4.16	3150.3	8308.7	0.89	71.6	71.5
	min	0.06	0	0	0	2.7	1.1
CSD	max	4.06	3152.9	6621.1	0.803	71.6	71.5
	min	0.13	0	0	0	2.7	0.5
QMOF	max	2.88	2878.2	7437.8	0.876	33.7	30.1
	min	0.08	0	0	0	1.9	0.9
Majumdar	max	2.09	2723.6	8013.1	0.92	56.6	54.2
	min	0.05	0	0	0	4.1	3.0

A second analysis that was performed is related to the chemical substructures present in ARC-MOF. This data is summarized in Figure 2, with the raw data given in the supporting material. The most common moieties were found to be *aromatics* (93.0 % of all MOFs with 47.4 % containing an N-heterocyclic aromatic), *alkynes* (33.1 %), *alkenes* (25.3 %), *halogens* (25.5 %), and *amines* (24.2 %). Other significant functional groups — including *acids*, *alcohols*, *amides*, *sulfides*, etc. — were also present in the database structures in relatively smaller quantities. The results of our database functional group analysis appear consistent with the moieties most frequently observed in organic linker SBUs employed in MOF syntheses and those included in hypothetical SBU libraries.

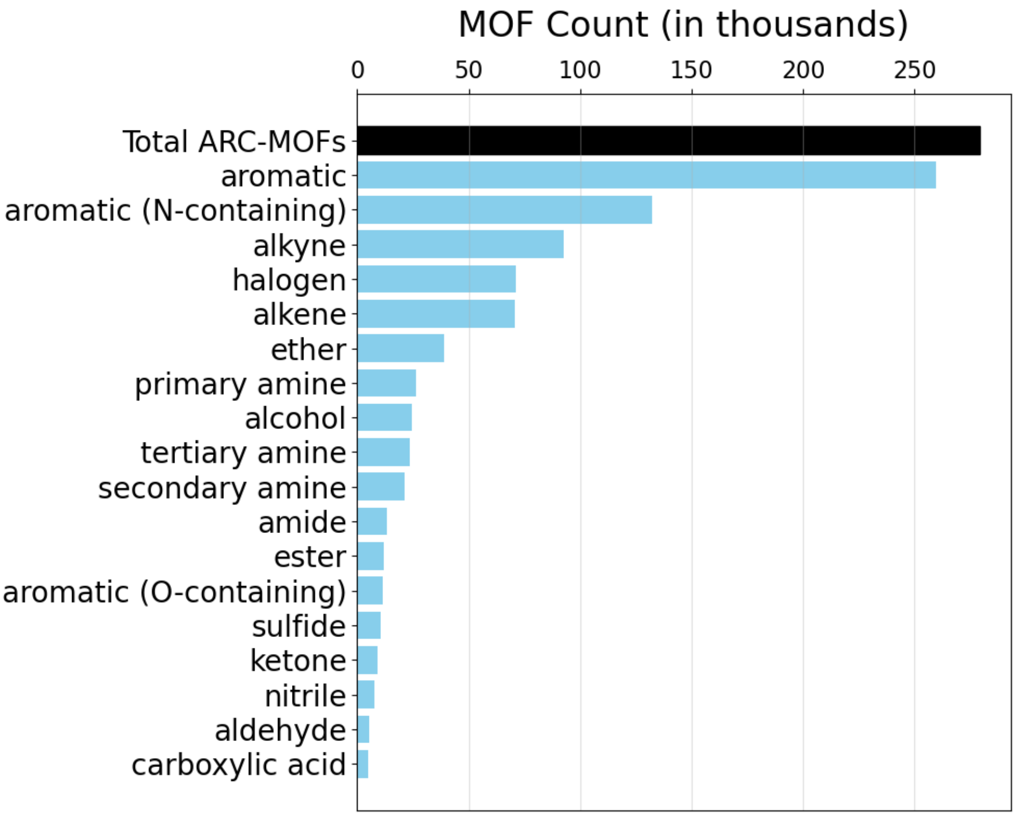

Figure 2. Comparison between the incidence of representative chemical substructures in ARC-MOF. Only substructures not bound to metal atoms were considered in this analysis.

Topological data (i.e., information about the node-linker connectivity and symmetry) is available in the filenames of over 210 000 structures in the ARC-MOF database. Of the MOFs whose topology is known, the frequency of the twenty most common net topologies in ARC-MOF are summarized in Table 3. The majority of MOFs in ARC-MOF are identified as possessing *pcu* (30.1 %), *fsc* (21.5 %), *nbo* (20.4 %), or *pts* (17.6 %) net topologies. However, the database consists of more than 426 unique net topologies, with 31 topologies being represented in a minimum of 50 structures.

Table 3. Top 20 occurring net topologies in ARC-MOF. Only MOFs containing topology information in their filename are considered.

Topology	*Count*	*%*	*Topology*	*Count*	*%*
pcu	63 485	30.1	pto	342	0.2
fsc	45 297	21.5	qtz	246	0.1
nbo	42 924	20.4	tfo	228	0.1
pts	37 068	17.6	tfz	211	0.1
sra	10 772	5.1	dia	200	0.1
cds	1 409	0.6	xbe	144	0.1
lvt	1 190	0.6	qzd	144	0.1
bcu	741	0.4	sxc	139	0.1
fof	690	0.3	kag	125	0.1
acs	682	0.3	other	4 143	2.0
hxg	511	0.2			

Biases are frequently introduced in the curation of any MOF database which affects the transferability of ML models and efficiency of large-scale screening. For these reasons, the importance of the chemical diversity of MOF dataset in any high-throughput workflow cannot be overstated. For example, Moosavi *et al.* have previously demonstrated that there is a substantially lower diversity of the metal chemistry of hMOFs compared to the design space covered by experimental databases.[35] Therefore, as a final step to investigating the diversity of the structures in ARC-MOF, an in-depth analysis using revised autocorrelation descriptors (RACs) and geometric descriptors was performed, as described in the methodology section. In this work, we define the 'entire' design space as the collection of all MOFs taken from the sources listed in Table 1 that passed our structure checking protocols (~532K MOFs) as outlined in the methodology section. ARC-MOF is a subset of this design space. The chemistry-specific RAC descriptors for a) metal-centre chemistry; b) ligand chemistry; and c) functional group chemistry, as well as the six geometric descriptors shown in Table 2 were used in this analysis. Structures for which a periodic graph could not be constructed were not included in the analysis since RACs could not be computed for these structures (~9% of the overall space – 49,402 MOFs). Thus, 480K MOFs were used in the following analysis. While we regard this as an inherent limitation of using RACs for this analysis, these are presently the best descriptors available in the literature for this purpose.

Ideally, one would aim to construct a scatterplot of these descriptors and visualize a "map" of the chemistry of the design space as described by the RACs and thereby analyze the chemical diversity. However, since the RACs are a high-dimensional representation of the design space, one must reduce this space to a 2- or 3-dimensional space to allow for visualization. For this task, Uniform Manifold Approximation and Projection (UMAP) was used to display the high-dimensional data in a two-dimensional space for each type of descriptor to generate so-called UMAP plots (Figure 3). Due to practical limitations of using UMAP on large datasets, a random set of 50K MOFs was selected to represent the entire set of 480K MOFs (~10% of the entire dataset). The UMAP plots show how well the subset (ARC-MOF) covers the entire known design space, and how uniformly this space is mapped (i.e., whether clusters exist in the data). On the two axes, kernel density estimate (KDE) plots of the UMAP data for each dimension show the density of the points in the space, giving a qualitative measure of the balance of the dimension-reduced data. The radar plots in Figure 3 show the value of each diversity metric (disparity, variety, and balance), which are described in the methodology section. In Figure 3, coloured points on the UMAP plots and coloured lines on the radar plots represent ARC-MOF, while grey points and lines represent the entire design space.

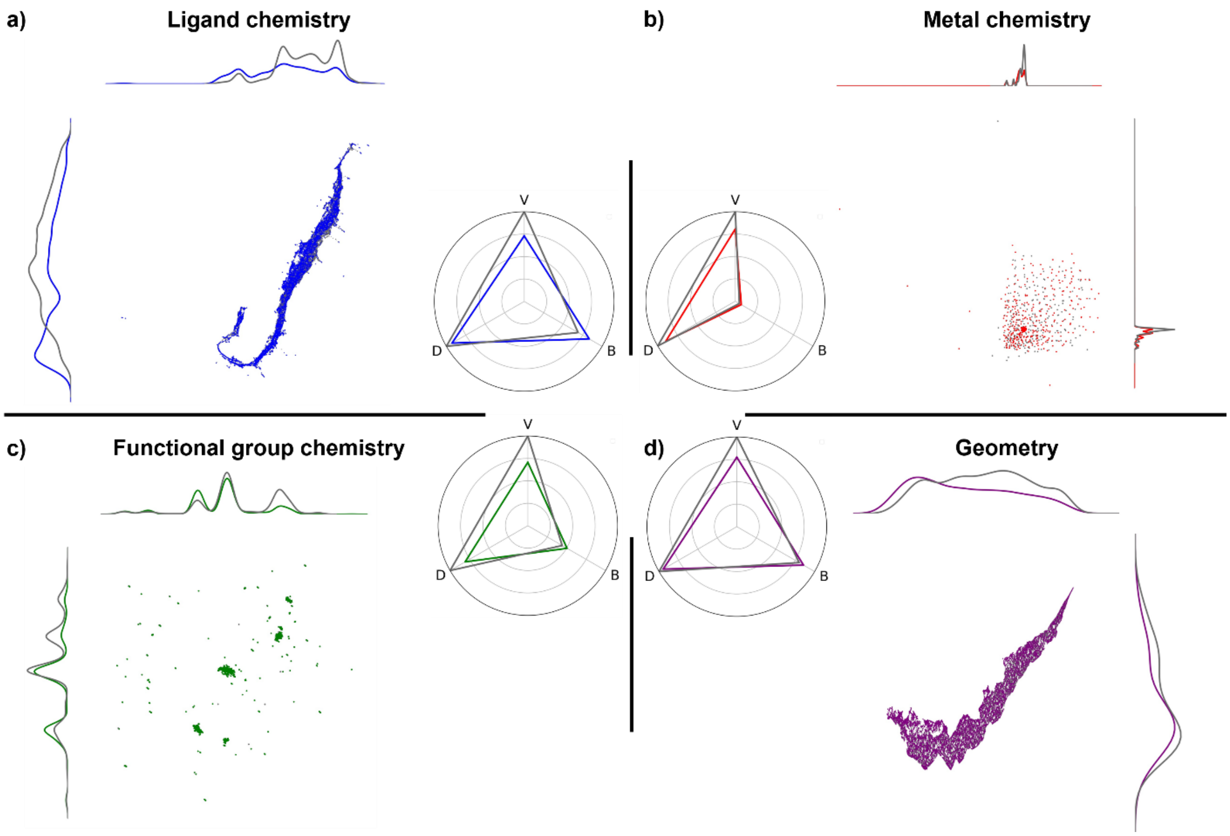

Figure 3. Two-dimensional UMAP projection of descriptors of 50,000 random MOFs and corresponding radar plots showing diversity metrics for the a) ligand chemistry; b) metal chemistry; c) functional group chemistry; and d) geometry. RAC descriptors were used for the chemical diversity (a, b, c) and geometric descriptors were used for d. Structures present in ARC-MOF are represented by coloured points, which is overlaid on the entire design space, represented by grey points. The diversity metrics shown on the radar plots are variety (V), disparity (D), and balance (B), where grey again represents the entire set of MOFs and colour represents the ARC-MOF subset. Only functionalized MOFs were considered for c), and only MOFs with non-zero accessible surface area are shown in d).

Figure 3 suggests ARC-MOF is sufficiently diverse with respect to the presently known design space. This is perhaps an unsurprising conclusion given ARC-MOF is a near random sample of 54% of this space, and as such should have approximately equal diversity compared to the overall space. Clearly, the absence of substantial grey regions in Figure 3 suggests ARC-MOF covers the same design space as the overall set of MOFs (this is reflected in the disparity metric). The variety and balance are also approximately equal between the two sets. However, while the balance of the geometry and ligand chemistry of ARC-MOF are almost equal to unity (suggesting a nearly uniform distribution), this is not the case for the metal chemistry and functional group chemistry. Since the balance and variety calculations are independent of the UMAP plots in Figure

3, clustering was done on the entire dataset, rather than the random 50K MOF sample, to achieve a more accurate determination of balance of ARC-MOF. The results of this analysis are summarized in Table 4. Table 4 shows a very low balance of 0.01 for both ARC-MOF and the entire design space, in agreement with the results of Figure 3b (Table 4). The imbalance of metal chemistry in ARC-MOF (and the overall design space) is expected, as most MOF structures are hypothetical (i.e., computationally generated). In hypothetical database generation, there is typically a bias towards small number (usually in the tens or even less) of inorganic SBUs that are used to produce the structures, but hundreds of inorganic SBUs have been experimentally incorporated into MOFs and are found in ARC-MOF. Although the balance of the functional group chemistry within ARC-MOF (and the entire design space) is significantly higher than that of the metal chemistry, it is still not particularly high with a value of only 0.24 in ARC-MOF and 0.20 in the entire design space (Table 4).

Table 4. Balance and variety of the design space (~480K MOFs) and ARC-MOF (~288K MOFs).

Type of MOF diversity		Balance	Variety
ligand chemistry	All	0.55	1.00
	ARC-MOF	0.78	0.74
metal chemistry	All	0.01	1.00
	ARC-MOF	0.01	0.76
functional group chemistry	All	0.20	1.00
	ARC-MOF	0.24	0.76
geometry	All	0.66	1.00
	ARC-MOF	0.83	0.74

Two common ways to address data imbalance are to either add duplicate/artificial data to the underrepresented clusters (upsampling) or remove data from the overrepresented clusters (downsampling). In this work, farthest point sampling[69] (also known as max/min sampling) was used to select a more balanced set of structures for each unbalanced type of chemistry in ARC-MOF (i.e., functional group chemistry and metal chemistry). Notably, the RACs used for the farthest point sampling depended on the type of chemistry (e.g., only the metal-centre RACs were used for sampling the MOFs with most diverse metal chemistry). Farthest point sampling is a greedy algorithm which iteratively samples points from a set, such that the sampled points optimally cover the dataset (i.e., they are “spread out”). Using this technique, in combination with a clustering technique (for unclassified data), one can determine a subset which has maximum size and balance for a particular type of chemistry. The diversity metrics and UMAP plots for these balanced subsets are shown in Figure 4. For the ligand chemistry and geometry, a farthest-point-sampled set of 100,000 MOFs was used to demonstrate the diversity was unaffected, since these types of RACs are already well-balanced over the entire set. It was determined from this analysis that a subset of 150,000 MOFs was reasonable to obtain a set of MOFs with balanced functional group chemistry. Likewise, a set of 20,000 MOFs was found to have balanced metal chemistry. These "diverse subsets" are provided in the supporting material and labelled the ARC-MOF diverse-ligand, diverse-functional-group, diverse-metal and diverse-geometry subsets. We anticipate the ARC-MOF diverse subsets will be useful for training machine learning models which rely particularly on these types of chemistry, (e.g., applications in catalysis or low-pressure gas adsorption).

Figure 4 shows how in particular for metal chemistry, a small fraction (~5%) can sufficiently represent the diversity of the entire space and demonstrates the high imbalance of metal chemistry in the available (hypothetical) databases. The same conclusion can be drawn from Figure 5, which shows the 10 most common a) organic; and b) inorganic substructures present in ARC-MOF, and their corresponding frequencies. The top 10 most common organic substructures comprise ~1% of the entire ARC-MOF database, while the top 10 most common inorganic substructures are found in ~67% of the entire ARC-MOF database. Furthermore, the top 10 inorganic substructures are chemically similar, being composed primarily of variations of copper and zinc paddlewheels. A similar result is observed for the entire design space, in which 60% of the MOFs contain the top 10 most common inorganic substructures (Figure S3). If the same analysis is done on the diverse-metal subset (20K MOFs), only ~2% of the MOFs contain the top 10 most common inorganic substructures. Consequently, we conclude it is critical that future ML and screening studies of MOFs curate their own datasets from MOF databases in a way that maximizes not only the number of structures in the dataset, but also the balance of the descriptor space of interest (e.g., the metal chemistry). Otherwise, it is unlikely such models will extend to these MOFs with underrepresented descriptor vectors.

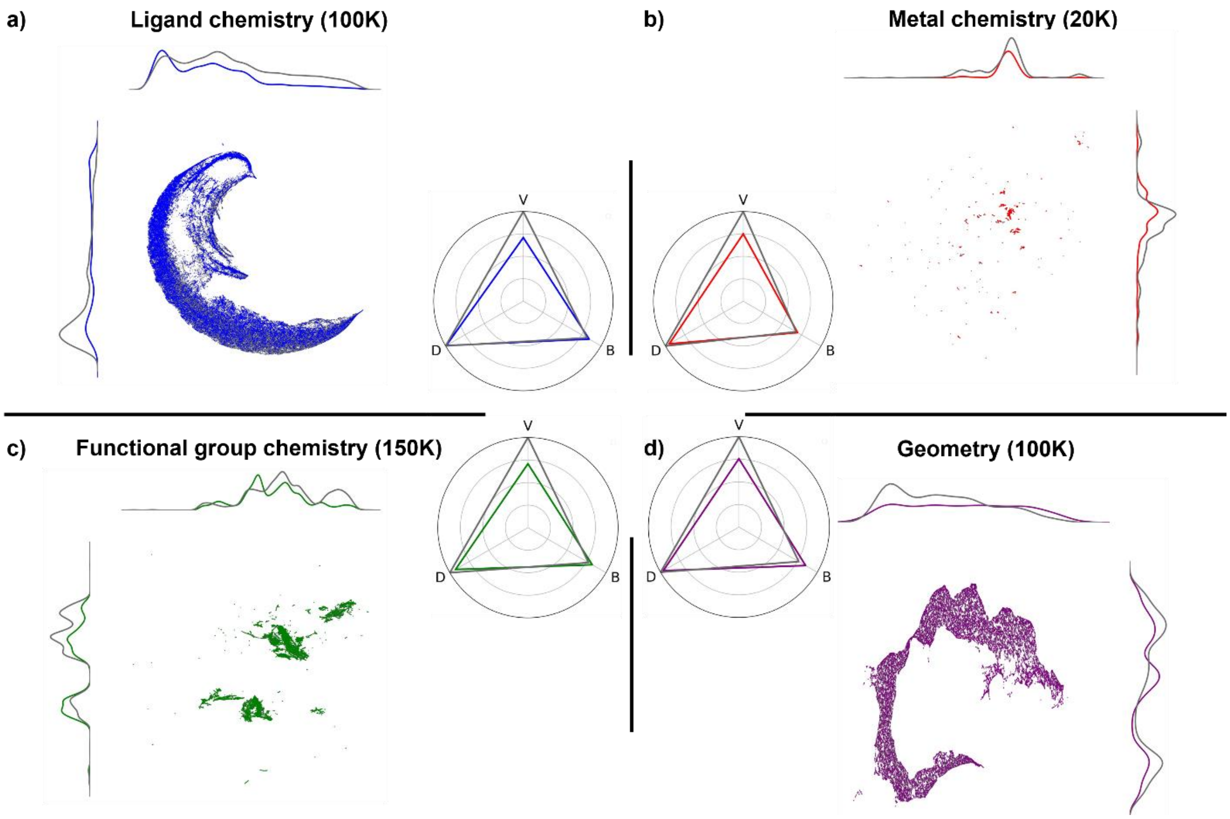

Figure 4. Two-dimensional UMAP projection of descriptors of a varying number of MOFs sampled using farthest point sampling and corresponding radar plots showing diversity metrics for the a) ligand chemistry; b) metal chemistry; c) functional group chemistry; and d) geometry. RAC descriptors were used for the chemical diversity (a, b, c) and geometric descriptors were used for d. Structures present in ARC-MOF are represented by coloured points, which is overlaid on the entire design space, represented by grey points. The diversity metrics shown on the radar plots are variety (V), disparity (D), and balance (B), where grey again represents the entire set of MOFs and colour represents the ARC-MOF subset. Only functionalized MOFs were considered for c), and only MOFs with non-zero accessible surface area are shown in d).

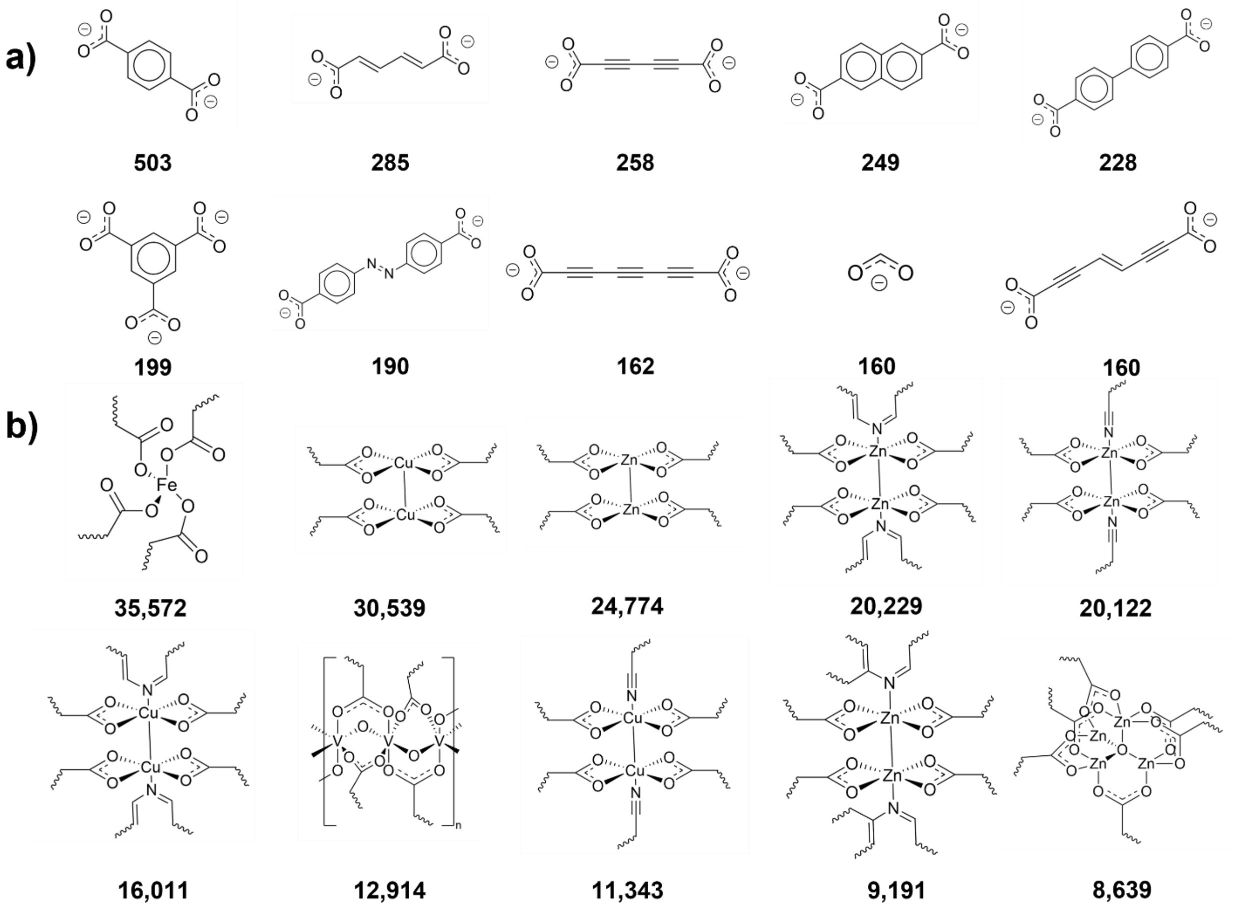

Figure 5. The top 10 most common a) organic; and b) inorganic substructures in ARC-MOF, as determined through a clustering analysis of the a) ligand; and b) metal-centre RAC descriptors of the entire database. The numbers under each substructure represent the frequency of the respective substructure in ARC-MOF.

REPEAT Charge Analysis

Partial atomic charges were derived from a DFT calculation of each MOF using the REPEAT method of Campañá et al.,[36] which is an electrostatic potential (ESP) fitted charge method. In these methods, the ESP from a first principles (DFT) calculation is evaluated on grid points outside atomic radii of the atoms in the system. The partial charges on the atoms are then fit to minimize the difference between the ESP due to the DFT calculation and that due to the point charges themselves at each of the grid points. Histograms showing the distribution of the REPEAT charges in ARC-MOF for the most common elements are shown in Figure 6. REPEAT charge statistics for all elements present in the database are available in Table S5.

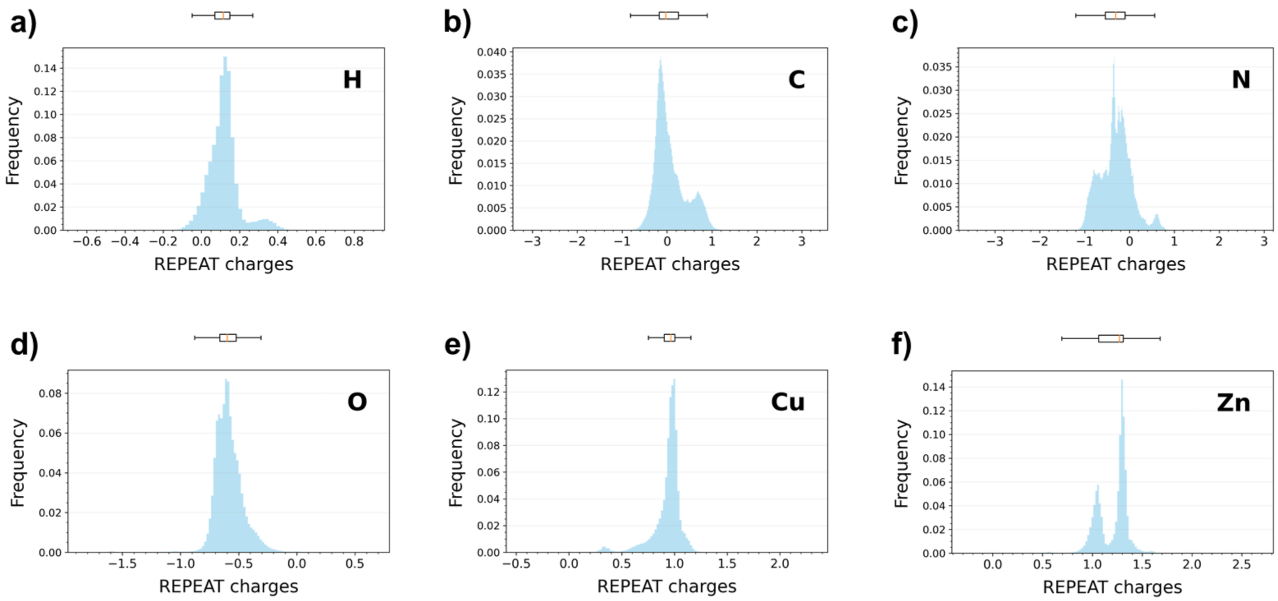

Figure 6. Histograms presenting the distribution of calculated REPEAT charges by element for (a) H, (b) C, (c) N, (d) O, (e) Cu, and (f) Zn. Each histogram represents the cumulative total of atoms extracted from all MOFS in the ARC-MOF database. The charges are given in units of e. Overhead boxplots illustrate the position of the REPEAT charge median (yellow line) and the interquartile range (box width) for each element.

To evaluate the importance of DFT-quality charges for atomistic simulations, we obtained gas uptake data for CO_2 and H_2S under various conditions using GCMC simulations with two different charge assignment methods. CO_2 pressures of 0.15 bar and 1 bar were chosen due to their relevance to post-combustion carbon capture conditions[89] and 0.4 mbar due to its relevance to direct-air carbon capture conditions[90]. To demonstrate the importance of charge assignment method on gas adsorption simulations of highly polar guests, we performed a similar analysis on H_2S. Pressures ranging from 1 bar to 0.001 bar (1000 ppm) were chosen, as a 1000 ppm concentration of H_2S is both the immediate lethal concentration in humans[91] and is frequently observed in biogas applications[92].

In this analysis, we treat REPEAT as being the most reliable charge method for obtaining gas uptakes using GCMC simulations. Therefore, linear regressions were performed between gas uptakes obtained using REPEAT charges and gas uptakes obtained using QEq. These plots and additional statistical information are available in the supporting material. Slope, R^2 correlation, and Spearman rank coefficients from these plots are given in Table 4. In this case, slope is considered a measure of accuracy while R^2 correlation is considered a measure of precision. Both metrics are equal 1 when the gas uptakes obtained using the empirical charge assignment method (i.e., SQE or QEq) are equal to the gas uptakes obtained using the REPEAT partial charges. When the rank coefficient is equal to 1, the order of adsorption in REPEAT is conserved in QEq data.

Table 4 demonstrates that charge dependence on gas uptake increases greatly at low pressures and for highly polar guests. Particularly, QEq is only able to reasonably reproduce gas uptakes obtained using REPEAT charges when the pressure is greater than or equal to 1 bar (m=0.90 for CO_2 and m=0.84 for H_2S). When considering hydrogen sulfide at low pressures (0.001 bar), QEq performs poorly in reproducing the gas uptake as simulated using REPEAT charges (m=0.07, R^2=0.03). This result underpins the value of a database with DFT-derived partial atomic charges for GCMC simulations where electrostatic interactions dominate such as simulations performed at low pressure, or for simulations involving polar guests such as water.

Table 5. Linear regression parameters for correlation between GCMC uptakes simulated using MEPO-QEq versus REPEAT charges for 10,000 MOFs. The R^2 correlation parameter is determined from the Pearson R correlation coefficient. The Spearman R parameter is determined from the Spearman rank correlation coefficient.

Guest	Pressure (bar)	Linear Regression Parameters		
		Slope	R^2	Spearman R
CO_2	1.00	0.90	0.92	0.97
	0.15	0.86	0.94	0.97
	0.0004	0.22	0.22	0.96
H_2S	1.00	0.84	0.85	0.93
	0.1	0.63	0.75	0.93
	0.01	0.28	0.46	0.91
	0.001	0.07	0.03	0.90

Descriptors and Target Data

A final goal of this work was to compute descriptors and target data relevant to gas adsorption of MOFs to enable “out of the box” use of ARC-MOF in future screening or machine learning studies.

To this end, we have computed AP-RDF descriptors for the entire database, and RAC descriptors for most MOFs in ARC-MOF. RAC descriptors could not be computed for all MOFs in ARC-MOF due to limitations of the structural graph generation method implemented in MolSimplify. AP-RDFs have previously been shown to be optimal descriptors for gas adsorption studies of MOFs, particularly for CO_2 separations.[45,46] In addition to these descriptors, target data was computed corresponding to gas adsorption properties (e.g., gas uptake, working capacity, selectivity) of all MOFs in ARC-MOF for five gas separation processes, namely a) methane purification; b) postcombustion VSA; c) precombustion PSA; d) methane storage PSA; and e) landfill gas VPSA. The specific conditions of these separations can be found in the supporting material.

CONCLUSIONS

The ARC-MOF database contains ~280,000 thoroughly structure-checked MOFs which have been either experimentally characterized or computationally generated, spanning all publicly available MOF databases, with DFT-derived partial atomic charges for each MOF. Additionally, ARC-MOF contains pre-computed descriptors for out-of-the-box machine learning applications. An in-depth analysis of the diversity of ARC-MOF with respect to the currently mapped design space of MOFs was performed – a critical, yet commonly overlooked aspect in past publications of MOF databases. Chemical insights about ARC-MOF were drawn from this analysis – primarily that ARC-MOF sufficiently spans the overall chemical space, and that it is sufficiently balanced with respect to geometric properties, as well as ligand chemistry. However, ARC-MOF suffers from being highly unbalanced with respect to metal chemistry, a well-known flaw of current hypothetical MOF databases. Thus, a subset of the overall design space which is balanced with respect to metal chemistry has been curated, for machine learning or screening applications where metal chemistry is important (e.g., chemisorptive processes). As a result of the high imbalance of metal chemistry in the hypothetical databases, we conclude it is critical that future ML and screening studies of MOFs curate their own datasets from MOF databases in a way that maximizes not only the number of structures in the dataset, but also the balance of the descriptor space of interest (e.g., the metal chemistry). It was shown that farthest point sampling is an effective method to do this, and an efficient code to perform this sampling is provided in the SI. A key goal of this work is not only for researchers to make use of ARC-MOF, but also adapt the proposed methodology for curating balanced datasets for machine learning and screening applications. Finally, we have shown the utility of the ARC-MOF for applications where high quality partial atomic charges are desired (e.g., simulations involving polar guests, or simulations at low pressure). In these types of simulations, there is very poor correlation between GCMC-simulated uptakes using empirical charge assignment methods versus DFT-quality partial atomic charges.

SUPPORTING INFORMATION

The ARC-MOF database (.CIF files with charges) and associated descriptors/adsorption target properties, all structures used in the diversity analysis, clusters used in the diversity analysis, and balanced subsets are available on Zenodo (https://doi.org/10.5281/zenodo.6908727). The codes used for the diversity analysis are available on the uOttawa Woo Lab GitHub page: https://github.com/uowoolab.

ACKNOWLEDGEMENTS

Financial support from the Natural Sciences and Engineering Research Council of Canada, the University of Ottawa, MITACS, and TotalEnergies is greatly appreciated, as well as the computing resources provided by the Canada Foundation for Innovation, TotalEnergies, Compute Canada, and the US DOE HPC4EI program.

REFERENCES

(1) Farha, O. K.; Eryazici, I.; Jeong, N. C.; Hauser, B. G.; Wilmer, C. E.; Sarjeant, A. a; Snurr, R. Q.; Nguyen, S. T.; Yazaydın, a Ö.; Hupp, J. T. Metal-Organic Framework Materials with Ultrahigh Surface Areas: Is the Sky the Limit? *J. Am. Chem. Soc.* **2012**, *134* (36), 15016–15021. https://doi.org/10.1021/ja3055639.

(2) O'Keeffe, M.; Peskov, M. A.; Ramsden, S. J.; Yaghi, O. M. The Reticular Chemistry Structure Resource (RCSR) Database of, and Symbols for, Crystal Nets. *Acc. Chem. Res.* **2008**, *41* (12), 1782–1789. https://doi.org/10.1021/AR800124U/ASSET/IMAGES/LARGE/AR-2008-00124U_0004.JPEG.

(3) Yang, D.; Gates, B. C. Catalysis by Metal Organic Frameworks: Perspective and Suggestions for Future Research. *ACS Catal.* **2019**, *9* (3), 1779–1798.

https://doi.org/10.1021/acscatal.8b04515.

(4) Park, J. G.; Collins, B. A.; Darago, L. E.; Runčevski, T.; Ziebel, M. E.; Aubrey, M. L.; Jiang, H. Z. H.; Velasquez, E.; Green, M. A.; Goodpaster, J. D.; Long, J. R. Magnetic Ordering through Itinerant Ferromagnetism in a Metal–Organic Framework. *Nat. Chem.* **2021**, *13* (6), 594–598. https://doi.org/10.1038/s41557-021-00666-6.

(5) Jian, M.; Qiu, R.; Xia, Y.; Lu, J.; Chen, Y.; Gu, Q.; Liu, R.; Hu, C.; Qu, J.; Wang, H.; Zhang, X. Ultrathin Water-Stable Metal-Organic Framework Membranes for Ion Separation. *Sci. Adv.* **2020**, *6* (23), eaay3998. https://doi.org/10.1126/sciadv.aay3998.

(6) Mokri, N.; Sepehri, Z.; Faninam, F.; Khaleghi, S.; Kazemi, N. M.; Hashemi, M. Chitosan-Coated Zn-Metal-Organic Framework Nanocomposites for Effective Targeted Delivery of LNA-Antisense MiR-224 to Colon Tumor: In Vitro Studies. *Gene Ther.* **2021**, 1–11. https://doi.org/10.1038/s41434-021-00265-7.

(7) Sadiq, M. M.; Batten, M. P.; Mulet, X.; Freeman, C.; Konstas, K.; Mardel, J. I.; Tanner, J.; Ng, D.; Wang, X.; Howard, S.; Hill, M. R.; Thornton, A. W. A Pilot-Scale Demonstration of Mobile Direct Air Capture Using Metal-Organic Frameworks. *Adv. Sustain. Syst.* **2020**, 2000101. https://doi.org/10.1002/adsu.202000101.

(8) Wen, H. M.; Liao, C.; Li, L.; Alsalme, A.; Alothman, Z.; Krishna, R.; Wu, H.; Zhou, W.; Hu, J.; Chen, B. A Metal-Organic Framework with Suitable Pore Size and Dual Functionalities for Highly Efficient Post-Combustion CO 2 Capture. *J. Mater. Chem. A* **2019**, *7* (7), 3128–3134. https://doi.org/10.1039/c8ta11596f.

(9) Burns, T. D.; Pai, K. N.; Subraveti, S. G.; Collins, S. P.; Krykunov, M.; Rajendran, A.; Woo, T. K. Prediction of MOF Performance in Vacuum Swing Adsorption Systems for Postcombustion CO2 Capture Based on Integrated Molecular Simulations, Process Optimizations, and Machine Learning Models. *Environ. Sci. Technol.* **2020**, *54* (7), 4536–4544. https://doi.org/10.1021/acs.est.9b07407.

(10) Schoedel, A.; Ji, Z.; Yaghi, O. M. The Role of Metal–Organic Frameworks in a Carbon-Neutral Energy Cycle. *Nat. Energy* **2016**, *1*, 16034.

(11) Bae, Y.-S. S.; Snurr, R. Q. Development and Evaluation of Porous Materials for Carbon

Dioxide Separation and Capture. *Angew. Chemie - Int. Ed.* **2011**, *50* (49), 11586–11596. https://doi.org/10.1002/anie.201101891.

(12) McCaffrey, R.; Edwards, P.; Perilli, D.; Winskell, J. Global Cement Magazine: Reaching New Heights. *Global Cement Magazine*. Epsom, Surrey, UK December 2020, pp 20–22.

(13) 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, 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.

(14) Meng, Y. S.; Arroyo-De Dompablo, M. E. First Principles Computational Materials Design for Energy Storage Materials in Lithium Ion Batteries. *Energy Environ. Sci.* **2009**, *2* (6), 589–609. https://doi.org/10.1039/b901825e.

(15) Jain, A.; Shin, Y.; Persson, K. A. Computational Predictions of Energy Materials Using Density Functional Theory. *Nat. Rev. Mater.* **2016**, *1* (1), 1–13. https://doi.org/10.1038/natrevmats.2015.4.

(16) Duren, T.; Bae, Y. S.; Snurr, R. Q. Using Molecular Simulation to Characterise Metal-Organic Frameworks for Adsorption Applications. *Chem. Soc. Rev.* **2009**, *38* (5), 1237–1247. https://doi.org/Doi 10.1039/B803498m.

(17) Jennings, P. C.; Lysgaard, S.; Hummelshøj, J. S.; Vegge, T.; Bligaard, T. Genetic Algorithms for Computational Materials Discovery Accelerated by Machine Learning. *npj Comput. Mater.* **2019**, *5* (1), 1–6. https://doi.org/10.1038/s41524-019-0181-4.

(18) Li, J.; Liu, J.; Baronett, S. A.; Liu, M.; Wang, L.; Li, R.; Chen, Y.; Li, D.; Zhu, Q.; Chen, X. Q. Computation and Data Driven Discovery of Topological Phononic Materials. *Nat. Commun.* **2021**, *12* (1), 1–12. https://doi.org/10.1038/s41467-021-21293-2.

(19) 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.

(20) Boyd, P. G.; Chidambaram, A.; García-Díez, E.; Ireland, C. P.; Daff, T. D.; Bounds, R.; Gładysiak, A.; Schouwink, P.; Moosavi, S. M.; Maroto-Valer, M. M.; others. Data-Driven Design of Metal–Organic Frameworks for Wet Flue Gas CO2 Capture. *Nature* **2019**, *576* (7786), 253–256.

(21) 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.

(22) 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.

(23) 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.

(24) 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.

(25) 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.

(26) 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.

(27) 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* (2), 83–89. https://doi.org/10.1038/nchem.1192.

(28) Aghaji, M. Z.; Fernandez, M.; Boyd, P. G.; Daff, T. D.; Woo, T. K. Quantitative Structure–Property Relationship Models for Recognizing Metal Organic Frameworks (MOFs) with High CO2 Working Capacity and CO2/CH4 Selectivity for Methane Purification. *Eur. J. Inorg. Chem.* **2016**, *2016* (27), 4505–4511. https://doi.org/10.1002/ejic.201600365.

(29) 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.

(30) 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* **2019**, *7* (20), 12556–12564. https://doi.org/10.1039/c9ta01752f.

(31) Gurnani, R.; Yu, Z.; Kim, C.; Sholl, D. S.; Ramprasad, R. Interpretable Machine Learning-Based Predictions of Methane Uptake Isotherms in Metal-Organic Frameworks. *Chem. Mater.* **2021**, *33*, 3552. https://doi.org/10.1021/acs.chemmater.0c04729.

(32) Boyd, P. G. P. G. P. G.; Woo, T. K. 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.

(33) Collins, S. P.; Daff, T. D.; Piotrkowski, S. S.; Woo, T. K. Materials Design by Evolutionary Optimization of Functional Groups in Metal-Organic Frameworks. *Sci. Adv.* **2016**, *2* (11). https://doi.org/10.1126/sciadv.1600954.

(34) 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.

(35) Moosavi, S. M.; Nandy, A.; Jablonka, K. M.; Ongari, D.; Janet, J. P.; Boyd, P. G.; Lee, Y.; Smit, B.; Kulik, H. J. Understanding the Diversity of the Metal-Organic Framework Ecosystem. *Nat. Commun.* **2020**, *11* (1), 1–10. https://doi.org/10.1038/s41467-020-17755-8.

(36) Campañá, C.; Mussard, B.; Woo, T. K. Electrostatic Potential Derived Atomic Charges for Periodic Systems Using a Modified Error Functional. *J. Chem. Theory Comput.* **2009**, *5* (10), 2866–2878. https://doi.org/10.1021/ct9003405.

(37) Ongari, D.; Boyd, P. G.; Kadioglu, O.; Mace, A. K.; Keskin, S.; Smit, B. Evaluating Charge Equilibration Methods To Generate Electrostatic Fields in Nanoporous Materials. *J. Chem. Theory Comput* **2019**, *15*, 30. https://doi.org/10.1021/acs.jctc.8b00669.

(38) Rappe, A. K.; Goddard, W. A. Charge Equilibration for Molecular-Dynamics Simulations. *J. Phys. Chem.* **1991**, *95* (8), 3358–3363.

(39) Nistor, R. A.; Polihronov, J. G.; Müser, M. H.; Mosey, N. J. A Generalization of the Charge Equilibration Method for Nonmetallic Materials. *J. Chem. Phys.* **2006**, *125* (9), 094108. https://doi.org/10.1063/1.2346671.

(40) Wilmer, C. E.; Kim, K. C.; Snurr, R. Q. An Extended Charge Equilibration Method. *J. Phys. Chem. Lett.* **2012**, *3* (17), 2506–2511. https://doi.org/10.1021/JZ3008485/SUPPL_FILE/JZ3008485_SI_002.ZIP.

(41) Manz, T. A.; Sholl, D. S. Chemically Meaningful Atomic Charges That Reproduce the Electrostatic Potential in Periodic and Nonperiodic Materials. *J. Chem. Theory Comput.* **2010**, *6* (8), 2455–2468. https://doi.org/Doi 10.1021/Ct100125x.

(42) Kadantsev, E. S. E. S.; Boyd, P. G. P. G.; Daff, T. D. T. D.; Woo, T. K. T. K. Fast and Accurate Electrostatics in Metal Organic Frameworks with a Robust Charge Equilibration Parameterization for High-Throughput Virtual Screening of Gas Adsorption. *J. Phys. Chem. Lett.* **2013**, *4* (18), 3056–3061. https://doi.org/10.1021/jz401479k.

(43) Collins, S. P.; Woo, T. K. Split-Charge Equilibration Parameters for Generating Rapid Partial Atomic Charges in Metal–Organic Frameworks and Porous Polymer Networks for High-Throughput Screening. *J. Phys. Chem. C* **2017**, *121* (1), 903–910.

https://doi.org/10.1021/acs.jpcc.6b10804.

(44) Yao, Z.; Sánchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O. K.; Snurr, R. Q.; Aspuru-Guzik, A. Inverse Design of Nanoporous Crystalline Reticular Materials with Deep Generative Models. *Nat. Mach. Intell.* **2021**, *3* (1), 76–86. https://doi.org/10.1038/s42256-020-00271-1.

(45) Burner, J.; Schwiedrzik, L.; Krykunov, M.; Luo, J.; Boyd, P. G.; Woo, T. K. High-Performing Deep Learning Regression Models for Predicting Low-Pressure CO 2 Adsorption Properties of Metal–Organic Frameworks. *J. Phys. Chem. C* **2020**, acs.jpcc.0c06334. https://doi.org/10.1021/acs.jpcc.0c06334.

(46) Dureckova, H.; Krykunov, M.; Aghaji, M. Z.; Woo, T. K. Robust Machine Learning Models for Predicting High Co 2 Working Capacity and Co 2 /h 2 Selectivity of Gas Adsorption in Metal Organic Frameworks for Precombustion Carbon Capture. *J. Phys. Chem. C* **2019**, *123* (7). https://doi.org/10.1021/acs.jpcc.8b10644.

(47) Qiao, Z.; Peng, C.; Zhou, J.; Jiang, J. High-Throughput Computational Screening of 137953 Metal–Organic Frameworks for Membrane Separation of a CO 2 /N 2 /CH 4 Mixture. *J. Mater. Chem. A* **2016**, *4* (41), 15904–15912. https://doi.org/10.1039/C6TA06262H.

(48) Wang, R.; Zhong, Y.; Bi, L.; Yang, M.; Xu, D. Accelerating Discovery of Metal-Organic Frameworks for Methane Adsorption with Hierarchical Screening and Deep Learning. *ACS Appl. Mater. Interfaces* **2020**, *12* (47), 52797–52807. https://doi.org/10.1021/ACSAMI.0C16516/SUPPL_FILE/AM0C16516_SI_001.PDF.

(49) Raza, A.; Sturluson, A.; Simon, C. M.; Fern, X. Message Passing Neural Networks for Partial Charge Assignment to Metal–Organic Frameworks. *J. Phys. Chem. C* **2020**, *124* (35), 19070–19082. https://doi.org/10.1021/acs.jpcc.0c04903.

(50) Kancharlapalli, S.; Gopalan, A.; Haranczyk, M.; Snurr, R. Q. Fast and Accurate Machine Learning Strategy for Calculating Partial Atomic Charges in Metal-Organic Frameworks. *J. Chem. Theory Comput.* **2021**, *17* (5), 3052–3064. https://doi.org/10.1021/acs.jctc.0c01229.

(51) Pardakhti, M.; Nanda, P.; Srivastava, R. Impact of Chemical Features on Methane Adsorption by Porous Materials at Varying Pressures. *J. Phys. Chem. C* **2020**, *124* (8), 4534–4544. https://doi.org/10.1021/acs.jpcc.9b09319.

(52) Anderson, R.; Biong, A.; Gómez-Gualdrón, D. A. Adsorption Isotherm Predictions for Multiple Molecules in MOFs Using the Same Deep Learning Model. *J. Chem. Theory Comput.* **2020**, *16* (2), 1271–1283. https://doi.org/10.1021/acs.jctc.9b00940.

(53) Fernandez, M.; Trefiak, N. R. N. R.; Woo, T. K. T. K. Atomic Property Weighted Radial Distribution Functions Descriptors of Metal-Organic Frameworks for the Prediction of Gas Uptake Capacity. *J. Phys. Chem. C* **2013**, *117* (27), 14095–14105. https://doi.org/10.1021/jp404287t.

(54) Kresse, G.; Hafner, J. Ab Initio Molecular Dynamics for Liquid Metals. *Phys. Rev. B* **1993**, *47*, 558.

(55) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B* **1999**, *59*, 1758.

(56) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865–3868.

(57) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50*, 17953.

(58) Kadantsev, E. S.; Boyd, P. G.; Daff, T. D.; Woo, T. K. Fast and Accurate Electrostatics in Metal Organic Frameworks with a Robust Charge Equilibration Parameterization for High-Throughput Virtual Screening of Gas Adsorption. *J. Phys. Chem. Lett.* **2013**, *4*, 3056–3061. https://doi.org/10.1021/jz401479k.

(59) Huang, Y. G.; Mu, B.; Schoenecker, P. M.; Carson, C. G.; Karra, J. R.; Cai, Y.; Walton, K. S. A Porous Flexible Homochiral SrSi2 Array of Single-Stranded Helical Nanotubes Exhibiting Single-Crystal-to-Single-Crystal Oxidation Transformation. *Angew. Chemie Int. Ed.* **2011**, *50* (2), 436–440. https://doi.org/10.1002/ANIE.201004921.

(60) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals. *Nat. Commun.* **2017**, *8*, 1–12. https://doi.org/10.1038/ncomms15679.

(61) 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. https://doi.org/10.1016/j.commatsci.2012.10.028.

(62) Cordero, B.; Gómez, V.; Platero-Prats, A. E.; Revés, M.; Echeverría, J.; Cremades, E.; Barragán, F.; Alvarez, S. Covalent Radii Revisited. *J. Chem. Soc. Dalt. Trans.* **2008**, No. 21, 2832–2838. https://doi.org/10.1039/b801115j.

(63) 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.

(64) Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. MolSimplify: A Toolkit for Automating Discovery in Inorganic Chemistry. *J. Comput. Chem.* **2016**, *37* (22), 2106–2117. https://doi.org/10.1002/JCC.24437.

(65) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints* **2018**. https://doi.org/10.48550/arxiv.1802.03426.

(66) Nolet, C. J.; Lafargue, V.; Raff, E.; Nanditale, T.; Oates, T.; Zedlewski, J.; Patterson, J. Bringing UMAP Closer to the Speed of Light with GPU Acceleration. *ArXiv e-prints* **2020**. https://doi.org/10.48550/arxiv.2008.00325.

(67) Campello, R. J. G. B.; Moulavi, D.; Zimek, A.; Sander, J. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Trans. Knowl. Discov. from Data* **2015**, *10* (1). https://doi.org/10.1145/2733381.

(68) Wang, Y.; Yu, S.; Gu, Y.; Shun, J. Fast Parallel Algorithms for Euclidean Minimum Spanning Tree and Hierarchical Spatial Clustering. In *Proceedings of the 2021 International Conference on Management of Data*; ACM: New York, NY, USA, 2021; pp 1982–1995. https://doi.org/10.1145/3448016.

(69) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics*

1969, *11* (1), 137–148. https://doi.org/10.1080/00401706.1969.10490666.

(70) Garcia-Sanchez, A.; Ania, C. O.; Parra, J. B.; Dubbeldam, D.; Vlugt, T. J. H. H.; Krishna, R.; Calero, S.; García-Sánchez, A.; Ania, C. O.; Parra, J. B.; Dubbeldam, D.; Vlugt, T. J. H. H.; Krishna, R.; Calero, S. Transferable Force Field for Carbon Dioxide Adsorption in Zeolites. *J. Phys. Chem. C* **2009**, *113* (20), 8814–8820. https://doi.org/10.1021/jp810871f.

(71) Kamath, G.; Lubna, N.; Potoff, J. J. Effect of Partial Charge Parametrization on the Fluid Phase Behavior of Hydrogen Sulfide. *J. Chem. Phys.* **2005**, *123* (12), 124505. https://doi.org/10.1063/1.2049278.

(72) Provost, B. An Improved N2 Model for Predicting Gas Adsorption in MOFs and Using Molecular Simulation to Aid in the Interpretation of SSNMR Spectra of MOFs, University of Ottawa, 2014.

(73) Martin, M. G.; Siepmann, J. I. Transferable Potentials for Phase Equilibria. 1. United-Atom Description of n-Alkanes. *J. Phys. Chem. B* **1998**, *102*, 2569–2577. https://doi.org/10.1021/jp049459w.

(74) Belof, J. L.; Stern, A. C.; Space, B. An Accurate and Transferable Intermolecular Diatomic Hydrogen Potential for Condensed Phase Simulation. *J. Chem. Theory Comput.* **2008**, *4* (8), 1332–1337. https://doi.org/10.1021/CT800155Q.

(75) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. Uff, a Full Periodic-Table Force-Field for Molecular Mechanics and Molecular-Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114* (25), 10024–10035. https://doi.org/DOI: 10.1021/ja00051a040.

(76) Lan, Y.; Han, X.; Tong, M.; Huang, H.; Yang, Q.; Liu, D.; Zhao, X.; Zhong, C. Materials Genomics Methods for High-Throughput Construction of COFs and Targeted Synthesis. *Nat. Commun. 2018 91* **2018**, *9* (1), 1–10. https://doi.org/10.1038/s41467-018-07720-x.

(77) Anderson, R.; Rodgers, J.; Argueta, E.; Biong, A.; Gómez-Gualdrón, D. A. Role of Pore Chemistry and Topology in the CO2 Capture Capabilities of MOFs: From Molecular Simulation to Machine Learning. *Chem. Mater.* **2018**, *30* (18), 6325–6337. https://doi.org/10.1021/acs.chemmater.8b02257.

(78) Gomez-gualdron, D. A.; Gutov, O. V; Krungleviciute, V.; Borah, B.; Mondloch, J. E.; Hupp, J. T.; Yildirim, T.; Farha, O. K.; Snurr, R. Q. Computational Design of Metal − Organic Frameworks Based on Stable Zirconium Building Units for Storage and Delivery of Methane. **2014**.

(79) Materials Studio, BIOVIA, San Diego.

(80) Chung, Y. G.; Gómez-Gualdrón, D. A.; Li, P.; Leperi, K. T.; Deria, P.; Zhang, H.; Vermeulen, N. A.; Stoddart, J. F.; You, F.; Hupp, J. T.; Farha, O. K.; Snurr, R. Q. In Silico Discovery of Metal-Organic Frameworks for Precombustion CO 2 Capture Using a Genetic Algorithm. *Sci. Adv.* **2016**, *2* (10), e1600909. https://doi.org/10.1126/sciadv.1600909.

(81) Li, S.; Chung, Y. G.; Simon, C. M.; Snurr, R. Q. High-Throughput Computational Screening of Multivariate Metal-Organic Frameworks (MTV-MOFs) for CO2 Capture. *J. Phys. Chem. Lett.* **2017**, *8* (24), 6135–6141. https://doi.org/10.1021/acs.jpclett.7b02700.

(82) Anderson, G.; Schweitzer, B.; Anderson, R.; Gómez-Gualdrón, D. A. Attainable Volumetric Targets for Adsorption-Based Hydrogen Storage in Porous Crystals: Molecular Simulation and Machine Learning. *J. Phys. Chem. C* **2019**, *123* (1), 120–130. https://doi.org/10.1021/acs.jpcc.8b09420.

(83) Bao, Y.; Martin, R. L.; Haranczyk, M.; Deem, M. W. In Silico Prediction of MOFs with High Deliverable Capacity or Internal Surface Area. *Phys. Chem. Chem. Phys.* **2015**, *17* (18), 11962–11973. https://doi.org/10.1039/c5cp00002e.

(84) 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.

(85) Anderson, R.; Gómez-Gualdrón, D. A. Increasing Topological Diversity during Computational "Synthesis" of Porous Crystals: How and Why. *CrystEngComm* **2019**, *21* (10), 1653–1665. https://doi.org/10.1039/c8ce01637b.

(86) 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.

(87) 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.

(88) 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.

(89) Wang, M.; Lawal, A.; Stephenson, P.; Sidders, J.; Ramshaw, C. Post-Combustion CO2 Capture with Chemical Absorption: A State-of-the-Art Review. *Chem. Eng. Res. Des.* **2011**, *89* (9), 1609–1624. https://doi.org/10.1016/j.cherd.2010.11.005.

(90) Petersen, H. A.; Luca, O. R. Application-Specific Thermodynamic Favorability Zones for Direct Air Capture of Carbon Dioxide. *Phys. Chem. Chem. Phys.* **2021**, *23* (22), 12533–12536. https://doi.org/10.1039/d1cp01670a.

(91) Ng, P. C.; Hendry-Hofer, T. B.; Witeof, A. E.; Brenner, M.; Mahon, S. B.; Boss, G. R.; Haouzi, P.; Bebarta, V. S. Hydrogen Sulfide Toxicity: Mechanism of Action, Clinical Presentation, and Countermeasure Development. *J. Med. Toxicol.* **2019**, *15* (4), 287–294. https://doi.org/10.1007/s13181-019-00710-5.

(92) Rasi, S.; Läntelä, J.; Rintala, J. Trace Compounds Affecting Biogas Energy Utilisation – A Review. *Energy Convers. Manag.* **2011**, *52* (12), 3369–3375. https://doi.org/10.1016/j.enconman.2011.07.005.