Machine Learning the Quantum-Chemical Properties of Metal–Organic Frameworks for Accelerated Materials Discovery with a New Electronic Structure Database

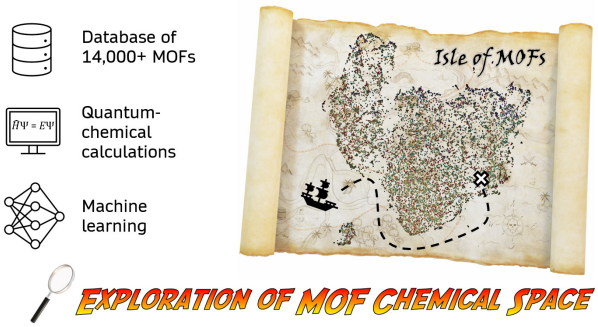
Andrew S. Rosen,[1,*] Shaelyn M. Iyer,[1] Debmalya Ray,[2] Zhenpeng Yao,[3] Alán Aspuru-Guzik,[3,4,5] Laura Gagliardi,[6] Justin M. Notestein,[1] Randall Q. Snurr[1]

[1]Department of Chemical and Biological Engineering, Northwestern University, 2145 Sheridan Rd., Evanston, IL 60208 (USA)
[2]Department of Chemistry, Chemical Theory Center, and Minnesota Supercomputing Institute, University of Minnesota, 207 Pleasant St. SE, Minneapolis, MN 55455 (USA)
[3]Chemical Physics Theory Group, Department of Chemistry and Department of Computer Science, University of Toronto, Toronto, Ontario M5S 3H6, Canada
[4]Vector Institute for Artificial Intelligence, Toronto, Ontario M5S 1M1, Canada
[5]Lebovic Fellow, Canadian Institute for Advanced Research (CIFAR) Senior Fellow, Toronto, Ontario M5S 1M1, Canada
[6]Department of Chemistry, Pritzker School of Molecular Engineering, James Franck Institute, Chicago Center for Theoretical Chemistry, The University of Chicago, 5735 S. Ellis Ave., Chicago, IL 60637 (USA)

*E-mail: rosen@u.northwestern.edu

Abstract. Metal–organic frameworks (MOFs) are a widely investigated class of crystalline solids with tunable structures that make it possible to impart specific chemical functionality tailored for a given application. However, the enormous number of possible MOFs that can be synthesized makes it difficult to determine which materials would be the most promising candidates, especially for applications governed by electronic structure properties that are often computationally demanding to simulate and time-consuming to probe experimentally. Here, we have developed the first publicly available quantum-chemical database for MOFs (the "QMOF database"), which consists of properties derived from density functional theory (DFT) for over 14,000 experimentally synthesized MOFs. Throughout this study, we demonstrate how this new database can be used to identify MOFs with targeted electronic structure properties. As a proof-of-concept, we use the QMOF database to evaluate the performance of several machine learning models for the prediction of DFT-computed band gaps and find that crystal graph convolutional neural networks are capable of achieving superior predictive performance, making it possible to circumvent computationally expensive quantum-chemical calculations. We also show how unsupervised learning methods can aid the discovery of otherwise subtle structure–property relationships using the computational findings in this work. We conclude by highlighting several MOFs with low band gaps, a challenging task given the electronically insulating nature of most MOF structures. The data and predictive models generated in this work, as well as the database of MOF structures, should be highly useful to other researchers interested in the predictive design and discovery of MOFs for the many applications dictated by quantum-chemical phenomena.

Keywords: metal–organic framework, machine learning, density functional theory, database, band gap, electronic structure



Introduction

Over the last several years, significant attention has been focused on the design of novel metal–organic frameworks (MOFs), a class of materials composed of discrete inorganic nodes connected to one another via organic linkers. One of the main advantages of MOFs is that they often have predictable and atomically defined structures with properties that are directly related to the choice of underlying metal and organic building blocks.[1] In this way, it becomes possible to impart physical and chemical functionality specifically tailored for a given application of interest.[2] To date, tens of thousands of MOFs have been synthesized,[3,4] and a nearly unlimited number can be proposed[5–7] by considering different combinations of constituent building blocks. Due to the enormous set of possible framework compositions, structures, and resulting properties, it remains difficult to discover truly top-performing MOFs for a particular application based solely on chemical intuition, conventional trial-and-error experimental testing, or serendipity alone.

High-throughput computational screening approaches based on classical simulations have proven extremely useful for more efficiently exploring the vast combinatorial space of MOF structures.[8,9] Recently, the large quantities of data generated during these computational screening studies have led to the development of machine learning (ML) models[10] that can accelerate the MOF design and discovery process even further. ML-assisted screening studies have been successfully applied to the discovery of MOFs suitable for H_2 storage,[11–13] CO_2 separation/capture,[14–16] and numerous other applications predominantly in the area of gas storage and separations.[9,17,18] Nonetheless, similar efforts remain almost entirely unexplored for the countless applications governed by quantum chemistry,[19] such as those that are based on the electronic, optical, magnetic, and/or catalytic properties of MOFs. Beyond the sheer number of possible MOFs that can be realized, the large number of atoms in MOF crystal structures often makes it computationally demanding to carry out even moderate-scale quantum-chemical screening studies, further magnifying the need for ML approaches in this area.

To date, the most relevant studies focused on training ML models to predict the quantum-chemical properties of MOFs are those of Raza et al.[20] and Korolev et al.[21] who independently developed ML models that can predict the partial atomic charges of MOFs in the Computation-Ready, Experimental (CoRE) MOF database.[22,23] Beyond these fundamental studies on partial charge prediction, however, there remains a significant gap in the literature, particularly for the discovery of MOFs with desired electronic structure properties. To the best of our knowledge, the only prior work in this area is that of He et al.[24] who trained binary classification models to predict whether inorganic solids in the Open Quantum Materials Database (OQMD)[25,26] are metallic or nonmetallic. Without retraining on MOF data, a multi-model voting procedure was then used to predict the metallic or nonmetallic behavior of 2932 MOFs in the CoRE MOF database,[22] which do not have computed band gaps. Of the six identified materials with near-zero band gaps at the PBE level of theory,[27] all are best-described as metal–cyanide/thiocyanate cluster complexes and none have H atoms in the structure. This is likely due in large part to the extreme differences between the OQMD, which consists almost entirely of inorganic compounds, and the CoRE MOF database. Furthermore, the fidelity of the metallic materials was not considered, leading to highlighted structures such as $[CdC_4]_n$ that should actually be $[Cd(CN)_2]_n$.[28]

In stark contrast with the existing literature on MOFs, significant progress has been made in the development of ML models to aid in the quantum-chemical screening of a wide range of inorganic and molecular compounds.[29–36] One of the fundamental features underlying much of this work has been the use of high-throughput density functional theory[37] (DFT) workflows to construct large-scale electronic structure property databases, such as those developed for inorganic solids[26,38–44] and molecular systems.[45–49] The synergistic combination of high-throughput DFT databases and ML has led to the discovery of a diverse range of materials with sought-after properties, including efficient organic light-emitting diodes,[50] superhard inorganic materials,[51] and thermally conductive polymers,[52] among many others.[35] With this in mind, there is a significant need for an analogous database of DFT-computed material properties for MOFs so that new ML models can be developed to rapidly predict their electronic structure properties. High-

throughput screening, database generation, and subsequent ML model development are crucial components for realizing the full potential of reticular chemistry[53] and accelerating materials discovery in general.[54–57]

In the present study, we leverage a recently developed high-throughput periodic DFT workflow tailored for MOF structures[58] to construct a large-scale database of MOF quantum-mechanical properties. This publicly available dataset[59] – the Quantum MOF (QMOF) database – contains computed properties for 14,204 experimentally characterized MOF structures after structure relaxation via DFT, including but not limited to optimized geometries, energies, band gaps, charge densities, partial charges, spin densities, and bond orders. We anticipate the QMOF database will serve two primary purposes: 1) materials discovery using the as-deposited data; 2) the evaluation and development of novel ML algorithms to reduce, or circumvent altogether, the need for otherwise expensive DFT calculations.

To demonstrate the utility of the data generated via the high-throughput DFT workflow, we use the QMOF database to develop several ML models for the prediction of MOF band gaps from nothing more than an encoding of the experimental (i.e. unrelaxed) crystal structures. Beyond serving as a proof-of-concept, an ML model that can predict MOF band gaps is particularly desirable, as most MOFs are known to be electronically insulating,[60] which limits their potential use in electrocatalysis, sensing, energy storage, and other applications where some degree of electrical conductivity is necessary.[60–64] We identify a top-performing band gap regression model based on a crystal graph convolutional neural network[65] and show how dimensionality reduction techniques can be used to discover overarching structure–property relationships to identify MOFs with targeted electronic structure properties. We conclude by highlighting several iron MOFs with low band gaps identified for the first time in this work.

Results

Generation and Overview of the QMOF Database

Prior to carrying out any periodic DFT calculations, a dataset of starting structures must be assembled. There are several databases of MOF structures that have been published to date.[3–6,22,66] However, it is imperative to note that existing databases of synthesized MOFs cannot be used as-is for quantum-chemical screening purposes. If even a single atom is missing or duplicated in a MOF crystal structure, the resulting DFT calculations are unlikely to be physically meaningful. Put another way, the simulation unit cell is expected to be charge-neutral unless otherwise specified, and any additional or missing electron in the system ruins the integrity of the resulting charge density, and therefore all the quantum-chemical properties derived from it. These situations can arise as a result of deficiencies in the deposited experimental crystal structure and/or in the dataset curation process when generating a database of MOF crystal structures. Therefore, in this work we aim to start with as "clean" a dataset as reasonably possible, one we will refer to as a suitably "DFT-ready" dataset of MOFs.

We considered the list of materials identified as MOFs in the Cambridge Structural Database (CSD)[3,67,68] and the 2019 Computation-Ready, Experimental (CoRE) MOF database.[4] All starting structures were taken directly from the CSD,[69] and free (i.e. unbound) solvents were automatically removed from the frameworks. From this set of experimental crystal structures, we constructed a DFT-ready dataset of 42,362 non-disordered MOF structures ("CSD-42362") after an extensive suite of automated fidelity checks, as summarized in Figure S1. In contrast with existing databases of experimental MOF structures,[3,4,22,66] this process better accounts for issues such as omitted H atoms, unresolved disorder, deleted framework atoms, lone (i.e. unbonded) atoms, an improper number of charge-balancing ions, and other structural issues that have been discussed in several recent studies.[68,70–75] Of these 42,362 structures, a subset of 24,013 materials with 150 atoms or fewer per primitive cell were considered such that large-scale screening could be carried out. After completing the high-throughput DFT screening procedure, static single-point (SP) calculations on the un-relaxed structures were successfully completed for 19,691 materials ("CSD-19691-SP"), and full structure relaxations (including cell volume and atomic positions) were successfully carried out for 14,204 materials ("CSD-14204-opt"). All periodic DFT calculations during this dataset construction process were

carried out at the PBE-D3(BJ)[27,76,77] level of theory using the Vienna *ab initio* Simulation Package (VASP)[78,79]. Additional methodological details regarding the dataset construction, DFT calculations, and ML methods can be found in the Supporting Information.

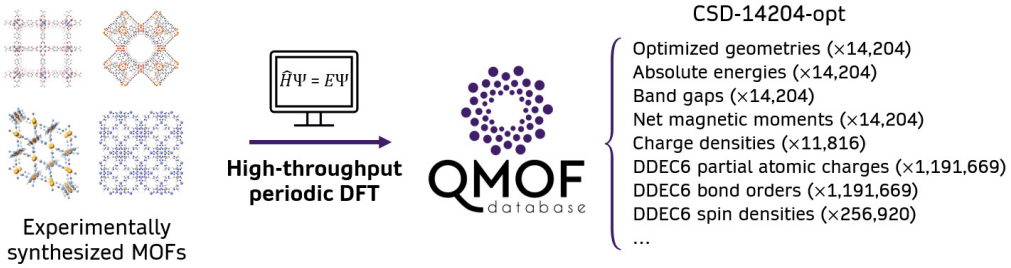


Figure 1. Selected DFT-computed properties for the structurally relaxed MOFs made available in the QMOF database. The DDEC6 properties are enumerated on a per-atom basis.

In carrying out the high-throughput periodic DFT calculations, several properties were computed at each stage of the workflow, a selection of which are shown Figure 1. Of these, band gaps are likely to be of interest for electronic and optical properties, especially in the search for (semi)conducting MOFs[60,62,80,81] or screening for photocatalytic materials.[82] Electronic energies, particularly if converted to formation energies, may provide insight into the relative stability of MOFs.[83] Machine learning the charge density[84] is of interest as a way to bypass the Kohn-Sham equations of DFT[85–87] and can be incorporated as a feature to predict a variety of additional quantum-chemical properties.[88,89] DDEC6 partial atomic charges,[90–92] bond orders,[93] and spin densities[90,91] have a wide range of potential use-cases, from describing electrostatic interactions in classical simulations of MOFs[23] to serving as descriptors to better understand trends in catalytic reactions.[94,95] Furthermore, the DFT-optimized structures can be used as starting points for further quantum-chemical calculations and for analyzing geometric properties of MOFs. In addition to the curated data mentioned in Figure 1, all raw output data from the DFT calculations are made publicly available so other unforeseen properties of interest can be readily tabulated and investigated.

Prior to highlighting how this data can be used in practice, we first investigated several properties of the QMOF database generated in this work. As shown in Figure 2a, the CSD-14204-opt dataset contains MOFs with chemical elements that span nearly the entire periodic table, which is beneficial for the development of transferable ML models. When looking at the geometries before and after structure relaxation, we find that 97.1% of the DFT-optimized MOFs had a change in cell volume less than 10% (Figure 2b), suggesting that the removal of free solvent does not drastically alter the structural properties for the majority of the MOFs in this work. The distribution of DFT-computed band gaps for the fully optimized structures is shown in Figure 2c and indicates that there is a wide spread of values from nearly 0 eV to 6.45 eV. The band gaps are not normally distributed and instead are bimodal, with peaks centered around 0.9 eV and 2.9 eV. This can be attributed to different distributions associated with closed- and open-shell materials in the CSD-14204-opt dataset (Figure 2c), the latter of which have significantly lower band gaps at the PBE-D3(BJ) level of theory on average. With regards to partial atomic charges, a wide spread of values is also obtained (Figure S4a). In comparing the partial atomic charges before and after structure relaxation, we find that 92.4% of the ~1.19 million data points have an absolute difference less than 0.05 q_e, and 98.8% of the points have an absolute difference less than 0.1 q_e (Figure S4b). As has been observed on a smaller scale in prior work,[23,66] it can be safely assumed that the partial charges remain essentially unchanged upon structure relaxation in most cases.

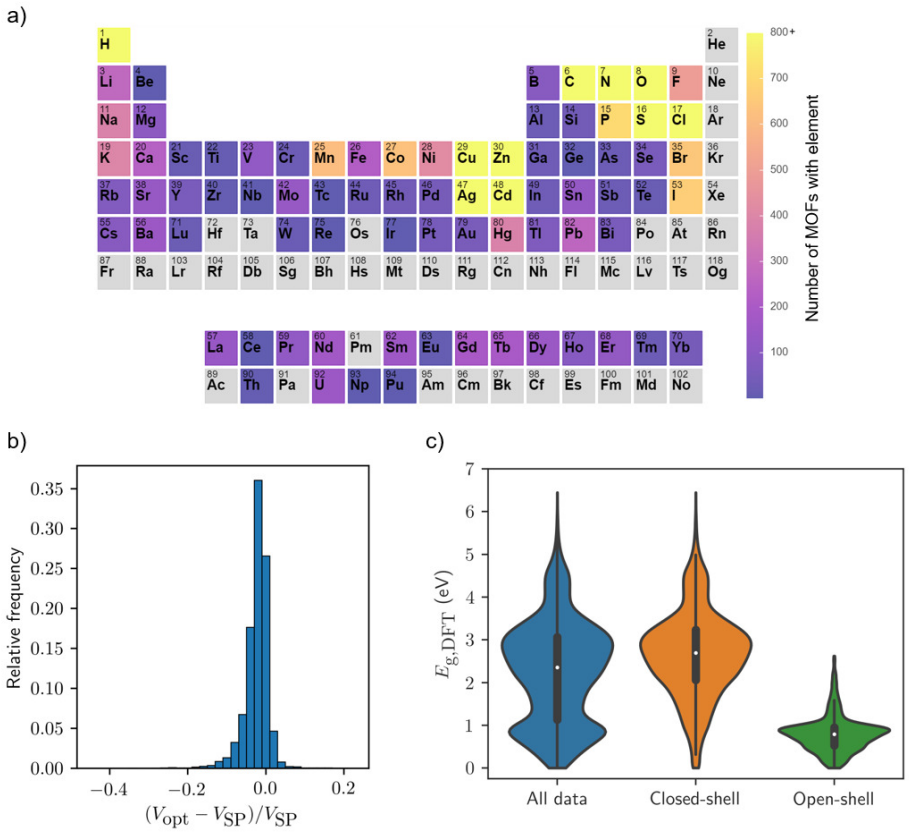


Figure 2. a) Number of MOFs in the CSD-14204 dataset containing a given element. All elements that occur in greater than 800 structures are capped at 800 for ease of visualization. These include: C (14,204), H (14,204), N (11,526), O (11,458), Cu (2,617), S (2,493), Cd (2,272), Zn (2,250), Cl (1,575), and Ag (1,120). Elements in gray are not present in any structure. b) Histogram of the fractional change in cell volume before (V_{SP}) and after (V_{opt}) structure relaxation for the MOFs in the CSD-14204-opt dataset. c) Violin plot of DFT-computed band gaps, $E_{g,DFT}$, in the CSD-14204-opt dataset. Separate distributions are shown for the entire dataset (14,204 entries), the closed-shell MOFs (10,995 entries), and the open-shell MOFs (3,209 entries). Open-shell character is defined here as having a DDEC6 atomic spin density with a magnitude greater than 0.1. A box plot, showing the extrema and interquartile range, is included in each violin, with the median marked by a white dot.

As a brief demonstration for how the data generated via the high-throughput DFT workflow could be used directly, we identified any materials with high-spin Fe species, which we defined as having an absolute DDEC6 spin density greater than 3.5 from the DFT calculations. High-spin Fe complexes are known to be promising for oxidation catalysis, in particular for the activation of strong C–H bonds, and recent work has focused on stabilizing such motifs in MOFs for this purpose.[96–98] We further constrained the list of MOFs to only include those with a pore-limiting diameter greater than 3.6 Å (i.e. the kinetic diameter of N_2). This search resulted in six frameworks, as shown in Figure S5: Fe(bpz) (H_2bpz = 4,4′-bipyrazole) (refcode: ACODAA),[99] Fe_2(dobdc) (H_4dobdc = 2,5-dihydroxybenzene-1,4-dicarboxylic acid) with (refcode: SARHAW) and without (refcode: COKNOH) bound propylene,[100,101] Fe_2Cl_2(bbta) (H_2bbta = 1*H*,5*H*-benzo(1,2-d:4,5-d′)bistriazole) (refcode: HAYYUE),[102] and geometrically distinct, expanded conformations of Fe(bdp) (H_2bdp = 1,4-benzenedipyrazole) (refcodes: QUPZIM01, QUPZIM02).[103] Providing validation of this screening approach, Fe_2(dobdc) has already been shown to oxidize ethane to ethanol,[96,104] and Fe_2Cl_2(bbta) has been computationally investigated for oxidation reactions,[105,106] with experimental evidence suggesting that both MOFs have high-spin Fe sites.[96,102]

Machine Learning Models for Band Gap Prediction

Beyond analyzing the DFT-computed properties directly, the QMOF database now makes it possible to train a wide range of ML models specifically tailored for MOFs, which are likely to have their own distinct feature space compared to isolated molecules and inorganic solids. This serves two primary purposes. The first is more theoretical: featurization methods (i.e. how each MOF structure is encoded) and ML algorithms that are well-suited for other materials may not be equally suitable for MOFs, so this database of quantum-chemical properties can serve as a testing ground to benchmark new ML methods. The Materials Project[38] and OQMD[25,26] in particular have accelerated this research direction for inorganic solids, and the QM9 dataset[45,107] (as one example) has done the same for small molecule chemistry. The second purpose of this new database is to apply these rapid yet accurate ML models to accelerate the materials discovery process, now with the ability to train these models directly on properties computed for MOFs.

In this work, we have chosen to develop an ML regression model that can rapidly predict the DFT-computed band gaps of MOFs. Specifically, we aim to predict the computed band gaps of the DFT-optimized structures from the un-optimized, experimentally resolved MOF crystal structures such that no quantum-chemical calculations need to be carried out. To achieve this, all ML models are trained on the band gaps of the DFT-optimized structures but take representations of the corresponding unrelaxed experimental structures as the input. Since the development of an ML model that can predict the band gaps of MOF crystal structures has not been achieved before, we trained several ML models using a variety of common featurization methods to benchmark each approach. These featurization methods are graphically summarized in Figure 3 for a representative material IRMOF-1 (IRMOF = isoreticular MOF),[108] also known as MOF-5 (Figure 3a). For the purposes of training ML models throughout this work, we specifically focus on a de-duplicated subset of the 13,058 materials in the QMOF database that have gone through the full periodic DFT volume relaxation process ("CSD-13058-opt"), as described in Figure S1.

The simplest featurization methods considered in this work are the feature sets of He et al.[24] (with 45 statistical attributes of elemental properties, denoted "Stoichiometric-45") and Meredig and Agrawal et al.[109] (with 103 attributes describing the elemental fractions from H–Lr and 17 statistical attributes of elemental properties, denoted "Stoichiometric-120"), which rely solely on the chemical composition of each material (Figure 3b). In addition, we consider several structure-sensitive featurization approaches, including the sine Coulomb matrix[110] that encodes pairwise electrostatic interactions between nuclei in a material (Figure 3c, Equation S4) and the orbital field matrix[111] that encodes the distribution of valence electrons in each coordination environment of a material (Figure 3d). The smooth overlap of atomic positions (SOAP)[112,113] is another structure-sensitive descriptor considered in this work and can be used to compute the similarity between a pair of local atomic environments by representing the atoms as Gaussians (i.e. "smoothed positions"), which are then summed to produce a density field (Equation S5). The overlap of these density fields, integrated over all three-dimensional rotations (Equations S6–S7), are compared between structures to generate a kernel matrix describing the similarity between every pair of MOF structures (Figure S2 and Equations S8–S9). In all of the aforementioned examples, these features are used to develop a kernel ridge regression[114] (KRR) model (Equations S1–S3). Motivated by prior work on inorganic solids, we also investigated the use of a crystal graph convolutional neural network (CGCNN),[65] wherein an approximate crystal graph is generated for each MOF, with each node in the graph representing an atom and each edge representing the bonds that connect the atoms (Figure 3f). More detailed descriptions and full methodological details for each featurization method and ML model architecture can be found in the Supporting Information.

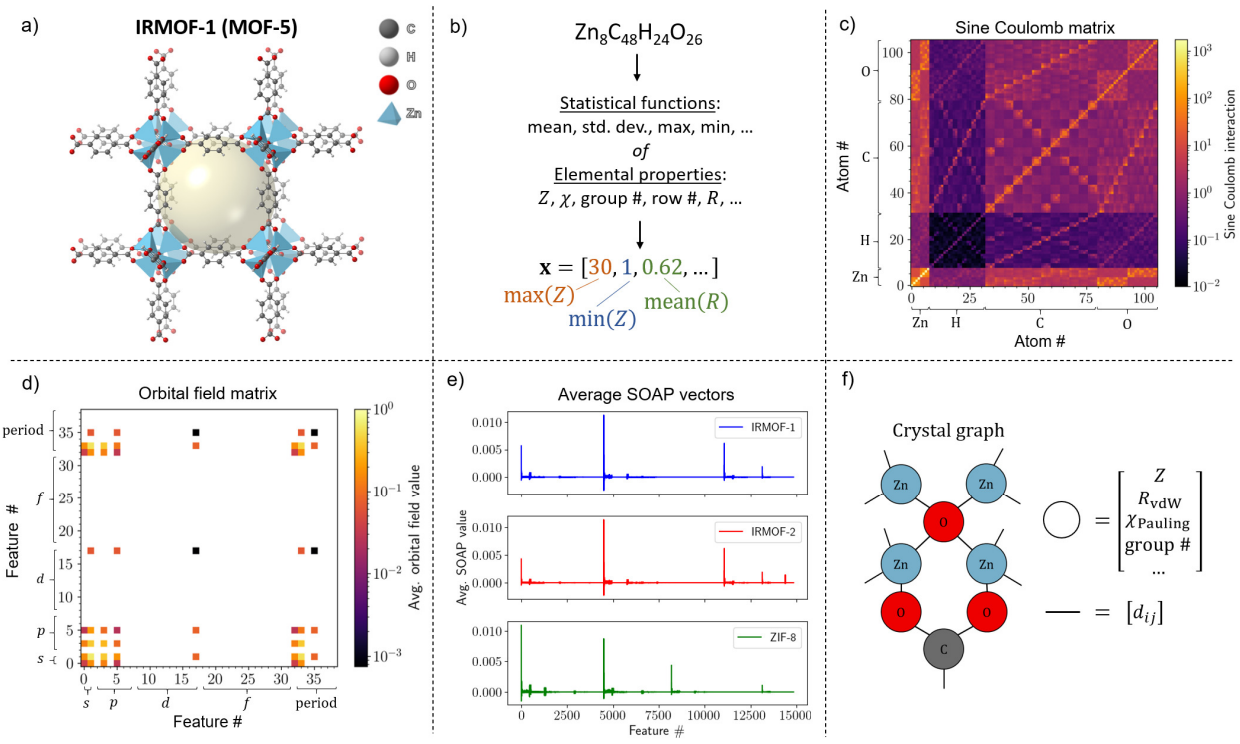


Figure 3. Visualization of various featurization methods applied to the unit cell of IRMOF-1. a) IRMOF-1 structure. b) Examples of composition-based features. c) Sine Coulomb matrix showing the interaction values between each pair of atoms. d) Orbital field matrix showing the average interaction value between each pair of orbital- or period-based features. Only non-zero values are shown. e) Averaged SOAP fingerprint of IRMOF-1 compared to IRMOF-2 and ZIF-8. Taking the dot product of any two vectors yields an unnormalized similarity score. f) Schematic of a crystal graph with example node (circle) and edge (line) embeddings (only a representative portion is shown for clarity).

As shown in Table 1, the KRR models trained on composition-based features (i.e. Stoichiometric-45 and Stoichiometric-120) are able to capture some of the band gap trends with mean absolute errors (MAEs) of 0.44 eV (with respect to the DFT-computed values) on the out-of-sample testing set. Nonetheless, these methods are still quite limited for regression purposes given that they do not encode any information about the structural properties of the MOF. In terms of structure-sensitive methods, taking an eigenvalue spectrum of the sine Coulomb matrix fares worse than the stoichiometry-based features, yielding a testing set MAE of 0.55 eV (Table 1). This can likely be traced back to the required use of zero-padding in the sine Coulomb matrix to ensure constant-length feature vectors between MOFs with different numbers of atoms per unit cell. The KRR model using a flattened orbital field matrix as the feature set is more accurate than the model based on the sine Coulomb matrix but shows only a minor improvement over the stoichiometry-based features. Overall, SOAP performs the best of all tested KRR descriptor sets, with an MAE of 0.37 eV and $R^2 = 0.81$ on the testing set. The marked improvement in performance with SOAP is especially clear when comparing the parity plots of the different KRR models (Figure S6).

Table 1. Summary of the mean absolute error (MAE), R^2, and Spearman rank-order correlation coefficient (ρ) for several machine learning methods to predict the computed band gaps of MOFs from their deposited crystal structures with free solvent removed. Kernel ridge regression was used for all featurization methods except for the crystal graphs of CGCNN, for which a convolutional neural network was constructed. The testing set statistics are shown, averaged over five runs (using different random seeds for data splitting) with ±1 standard deviation shown. For all models, 80% of the CSD-13058 dataset was used for training. The MAE for a dummy model that predicts the mean band gap (2.241 eV) for all the MOFs is shown for reference.

ML Method	MAE (eV)	R^2	ρ
Constant mean model	0.984	—	—
Sine Coulomb matrix	0.551 ± 0.013	0.625 ± 0.019	0.780 ± 0.014
Stoichiometric-45	0.443 ± 0.011	0.746 ± 0.012	0.846 ± 0.009
Stoichiometric-120	0.438 ± 0.009	0.748 ± 0.010	0.848 ± 0.006
Orbital field matrix	0.423 ± 0.007	0.761 ± 0.007	0.865 ± 0.003
SOAP	0.367 ± 0.007	0.807 ± 0.010	0.903 ± 0.005
CGCNN	0.272 ± 0.004	0.885 ± 0.007	0.934 ± 0.004

Notably, the CGCNN significantly outperforms all the aforementioned KRR models, achieving an MAE of 0.27 eV and $R^2 = 0.89$ (Table 1). As a point of reference, a trivial model that simply predicts the mean band gap for every MOF would have an MAE of 0.98 eV, indicating that CGCNN captures much of the underlying chemistry. The performance of the CGCNN model for MOF band gaps is comparable, if not slightly better, than state-of-the-art ML band gap models trained on the inorganic solids of the OQMD and Materials Project database as well as the organic crystals of the Organic Materials Database (OMDB).[41,65,115,116] It is also worth noting that the experimentally measured band gaps of MOFs can vary by several tenths of an eV depending on the synthesis and post-treatment conditions,[117] so an MAE less than 0.3 eV is likely to be sufficiently accurate for identifying structure–property trends and for ranking material candidates, the latter of which is further justified by the CGCNN's high Spearman rank-order correlation coefficient of $\rho = 0.93$. For context, it takes ~7 minutes (6 minutes for a one-time encoding of the crystal graphs and 1 minute to evaluate the neural network) on a modern laptop computer to predict the band gaps of all 13,058 MOFs in the CSD-13058 set using the CGCNN model. In stark contrast, it took approximately 1.5 million hours (~170 years) of computing time on the Stampede2 supercomputer[118,119] to carry out the structure relaxations and compute the band gaps via DFT.

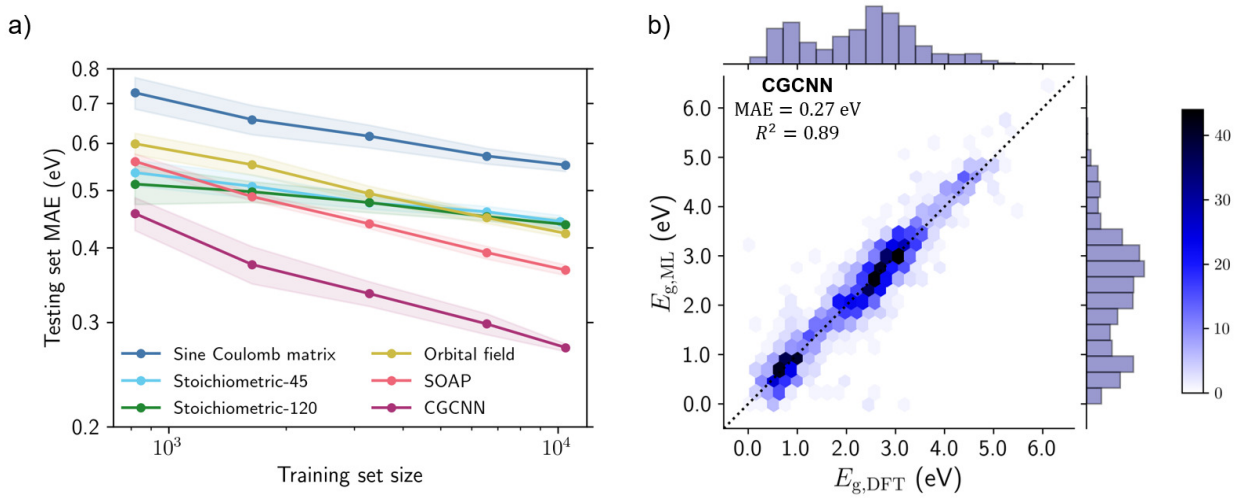


Figure 4. a) Mean absolute error (MAE) for testing set band gap predictions as a function of training set size for various machine learning methods. Each point represents the average value and the shaded region represents ±1 standard deviation of five runs (with different random seeds for data splitting). The data are shown on log–log axes. In all cases, 80% of the total dataset size was used for training. b) Testing set parity plot for the CGCNN model with hexagonal binning, comparing the machine learning band gaps, $E_{g,ML}$, to the DFT-computed band gaps, $E_{g,DFT}$. The color bar indicates the number of MOFs in each bin, and the line of parity is shown as a dashed line. Histograms summarizing the distribution of $E_{g,DFT}$ and $E_{g,ML}$ data are displayed parallel to the x- and y-axes, respectively.

The learning curves for each of the six models are shown in Figure 4a, highlighting the testing set MAE as a function of the training set size. Of all the individual models, CGCNN has both the largest learning rate and the lowest MAE regardless of training set size. While SOAP has a comparable testing set MAE to the

simpler stoichiometric models when trained on ~1600 MOFs, SOAP has a significantly higher learning rate such that it performs much better for larger training set sizes (although still underperforms compared to CGCNN). Reassuringly, the MAEs of the top-performing CGCNN and SOAP methods have not plateaued with respect to the training set size over the range of values considered in this work (i.e. up to ~10^4 training points). This indicates that both CGCNN and SOAP are capable of encoding the MOF crystal structures with sufficient uniqueness between structures and that the performance of the ML algorithms could be further improved if a greater number of training examples were provided. The testing set parity plot for the CGCNN trained on 80% of the CSD-13058 MOF dataset is shown in Figure 4b. As one would expect based on the relatively low MAE and high R^2, the agreement with the DFT predictions is generally quite strong, and this is true across the full range of band gap values.

Dimensionality Reduction for Structure–Property Analysis

While the kernel-based methods have a higher MAE than CGCNN when predicting MOF band gaps, one of their main advantages is that the underlying descriptors can be readily used for dimensionality reduction – an unsupervised learning task that can cluster structurally similar MOFs in feature space for the purposes of identifying structure–property relationships. Using the uniform manifold approximation and projection (UMAP) algorithm to carry out the dimensionality reduction,[120,121] the distance between each MOF in the reduced space can be related to the distance in feature space, such that clusters of points tend to have similar structures (Equation S10). By overlaying the DFT-computed band gaps over the UMAP, regions of low and high band gap emerge, making it possible to identify otherwise subtle structure–property trends.

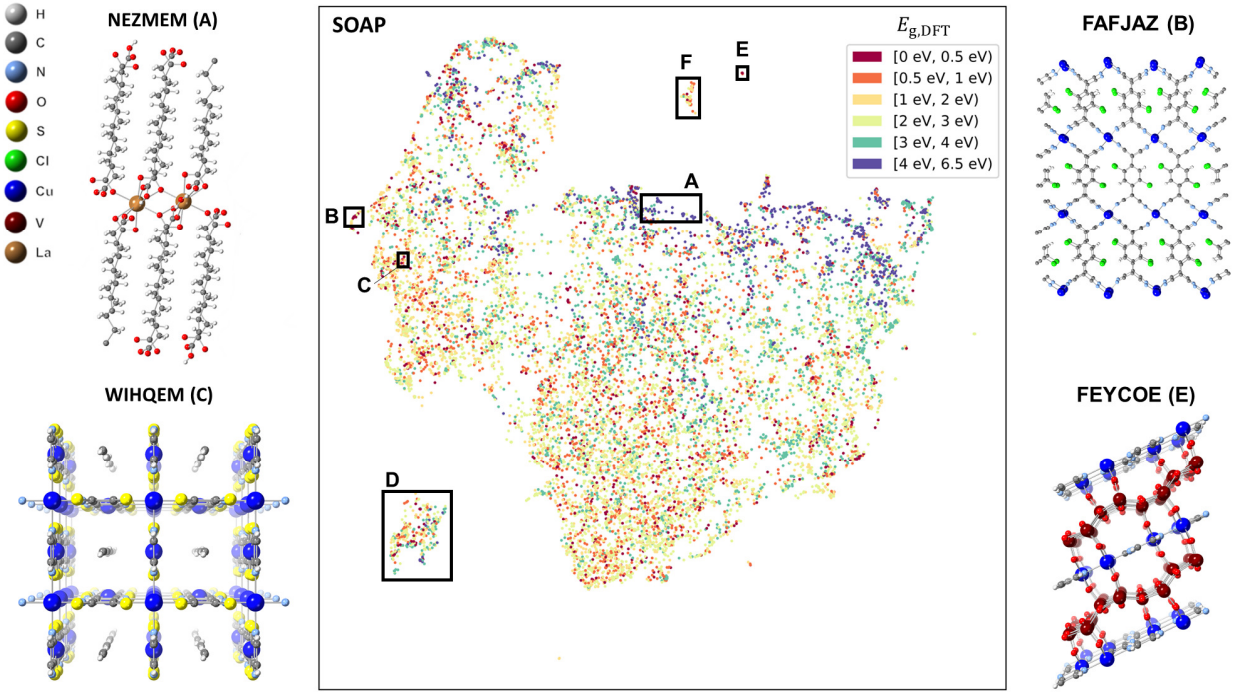


Figure 5. Unsupervised structural dimensionality reduction performed using UMAP, with a distance matrix obtained from the average SOAP similarity kernel of the structures in the CSD-13058-SP dataset. The computed band gaps of the DFT-optimized structures, $E_{g,DFT}$, are overlaid on the UMAP. Selected MOFs in the projection are highlighted.

As an example, selecting several MOFs in region A of the SOAP-based UMAP (Figure 5) yields materials with long, linear alkane-based linkers (e.g. refcodes NEZMEM[122], ROKZOI[123]), which consistently have high band gaps regardless of the coordinating metal. The low band gap MOFs are more scattered throughout the reduced feature space, but as one example, region B of Figure 5 contains framework materials with linkers consisting of various TCNQ (TCNQ = 7,7,8,8-tetracyano-quinodimethane) derivatives, with several

of these materials previously shown to have high electrical conductivities (e.g. BISVUW[124], FAFJAZ[125]). The projection in Figure 5 can be used to find MOFs that are structurally similar to a given material of interest as well. For instance, $Cu[Ni(pdt)_2]\cdot C_2H_2$ (pdt^{2-} = 2,3-pyrazinedithiolate) (refcode: HIVPOU[126]) is in the CSD-13058 dataset, and it is known to be one of the rare examples of a three-dimensional, porous framework that exhibits room temperature electrical conductivity.[126] Perhaps unsurprisingly, one of the closest points to $Cu[Ni(pdt)_2]\cdot C_2H_2$ is the isostructural framework $Cu[Cu(pdt)_2]\cdot C_2H_2$ (refcode: WIHQEM[127]) (region C), which has also been studied for its conductive properties.[128,129] In general, we find the SOAP-based UMAP places greater emphasis on the similarity of the organic linkers rather than the metal identity, likely due to the averaging scheme used in the generation of the similarity kernel (Table S5). Modifications to the SOAP encoding that better account for the discrete building block nature of MOFs, such as variations on the recently developed coarse-grained SOAP (cg-SOAP) method,[130] may yield improvements in the future.

Similar to what has been done in prior work with revised autocorrelation functions,[131] we can use the SOAP similarity kernel to understand the diversity of structures in the CSD-13058 database and identify structural outliers. The most apparent example is the isolated cluster of points in region D of Figure 5. Investigation of these crystal structures indicates that they are predominantly frameworks with high fluorine content, such as MOFs with fluorinated linkers (e.g. MUQCEH[132], HADMOR[133]) or metal–fluoride species (e.g. EMEJAJ[134]), which leads to a large difference in the average SOAP fingerprint compared to most other MOFs in the dataset. The isolated region E of Figure 5 where there is a low band gap cluster contains polyoxovanadate-based MOFs, some of which have already been investigated for their conductive and electrocatalytic properties (e.g. FEYCOE[135], XEHYEP[136]). Similarly, the nearby isolated region F contains molybdenum oxide-based frameworks (e.g. LUYQUT[137], SASCIB[138]).

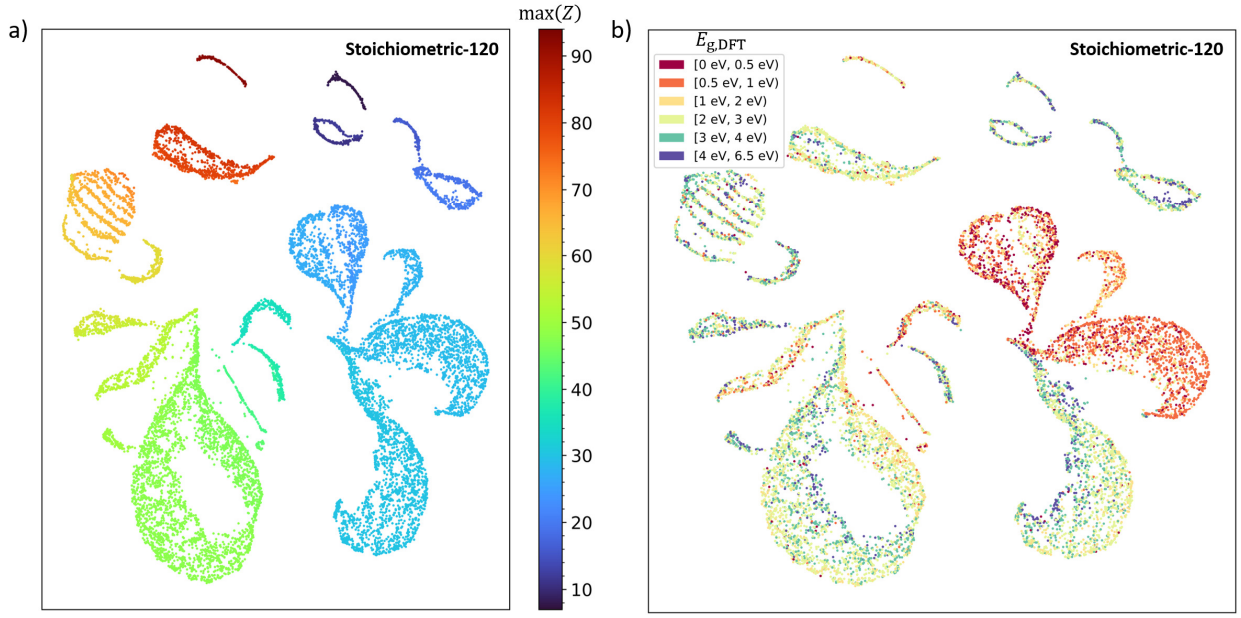


Figure 6. Unsupervised dimensionality reduction performed using UMAP, with a distance matrix obtained using a Euclidean distance metric of the Stoichiometric-120 encodings for the structures in the CSD-13058-SP dataset. The (a) maximum atomic number in each structure, $\max(Z)$, and (b) computed band gaps for the corresponding DFT-optimized structures, $E_{g,DFT}$, are overlaid on the UMAPs.

While the SOAP-based UMAP is useful for identifying local trends in feature space, significantly greater clustering is observed when using the Stoichiometric-120 encoding. As is evident in Figure 6a, the UMAP based on the Stoichiometric-120 encoding partitions the MOF chemical space based on the maximum atomic number in each chemical formula. The variations within a given cluster are due to more subtle

differences in the elemental fractions and compositional features that compose the Stoichiometric-120 descriptor. Notably, the band gaps are well-separated between and within each cluster in the reduced space (Figure 6b). For these reasons, the Stoichiometric-120 UMAP can be a valuable tool for obtaining a global view of the QMOF database. For instance, we find that the CSD-13058 dataset closely overlaps with both the larger CSD-42362 dataset it was drawn from and the separate CoRE MOF 2019 database[4] based on the reduced space of Stoichiometric-120 features (Figures S7 and S8). To enable additional data exploration, we have made interactive versions of the UMAPs available in the supporting dataset.[59]

Highlighting Notable Low Band Gap MOFs

We conclude by highlighting several framework materials identified in this work that have low band gaps, motivated in part by the search for a greater number of (semi)conducting MOFs. It should be noted that while the PBE-D3(BJ) level of theory makes it possible to generate a sufficiently large database for the purposes of ML model development and to identify structure–property relationships, it is known to underestimate band gaps like essentially all generalized gradient approximation functionals.[139,140] As such, we carried out full structure relaxations and corresponding band gap calculations using the hybrid-level HSE06-D3(BJ) functional[141,142] on select materials to generate more accurate band gap predictions. As a point of reference, materials with band gaps in excess of ~4 eV are often classified as electronic insulators, including many of the most commonly studied MOFs (e.g. MOF-5,[117] UiO-66 (UiO = Universitetet i Oslo),[143] ZIF-8 (ZIF = zeolitic imidazolate framework)[144]).[60,139] Generally, lower band gaps are necessary to support electrical conductivity (although it is not the sole factor required for achieving high electrical conductivities[60]).

When the CGCNN model is used to predict the band gaps of all 42,362 structures that compose the CSD-42362 dataset, the lowest band gap material is predicted to be $Ag(DCl)_2$ (DCl = 2,5-Cl,Cl-*N*,*N*′-dicyanoquinone diamine) (refcode: OTARUX[145]), which is known from experiments to exhibit metallic character via organic radicals that connect the Ag(I) cations.[145] The introduction of radical or redox-active linking units is a well-established strategy to increase the electrical conductivity of framework materials.[60] Although $Ag(DCl)_2$ is arguably best described as a coordination polymer, one notable MOF in the CSD-42362 dataset with a low predicted band gap and a radical-containing linker is $(TTF)[\{Rh_2(CH_3CO_2)_4\}_2TCNQ]$ (TTF = tetrathiafulvalene) (refcode: WAQMEJ[146]) – a pillared layer framework material built from Rh(II) paddlewheels and a TTF–TCNQ charge-transfer salt (Figure 7a). The HSE06-D3(BJ) band gap for this material is found to be particularly small with a value of 0.71 eV, which can be directly attributed to a reduced conduction band minimum (CBM) from the TTF and TCNQ components (Figure 7d). Furthermore, the valence band maximum (VBM) also exhibits hybridization between the 4*d* orbitals of Rh and 2*p* orbitals of C and N atoms belonging to the radical TCNQ linker, which is important for applications involving electron transport. In contrast, the most insulating structure in the CSD-42362 set based on CGCNN-predicted band gap is the non-porous coordination polymer $Sr[C_2H_4(SO_3)_2]$ (refcode: GUTYAW[147]), which has an HSE06-D3(BJ) band gap of 8.36 eV (Figure S11).

Consistent with prior experimental work,[148] we also find several Fe-containing materials in the CSD-42362 dataset with low band gaps, many of which have not been studied for their electronic properties. One representative example is $Fe(sq)(bpee)(H_2O)_2$ (bpee = 1,2-bis(4-pyridyl)ethylene; sq = squarate) (refcode: RAXNEK[149]), shown in Figure 7b, which has a band gap of 1.06 eV at the HSE06-D3(BJ) level of theory. The high-spin Fe(II) species in an octahedral crystal field with $t_{2g}^4e_g^2$ electron configuration dominate the VBM in this material, whereas the bpee linker (as opposed to the bridging sq species or inorganic node) make up the conduction band edge (Figure 7e).

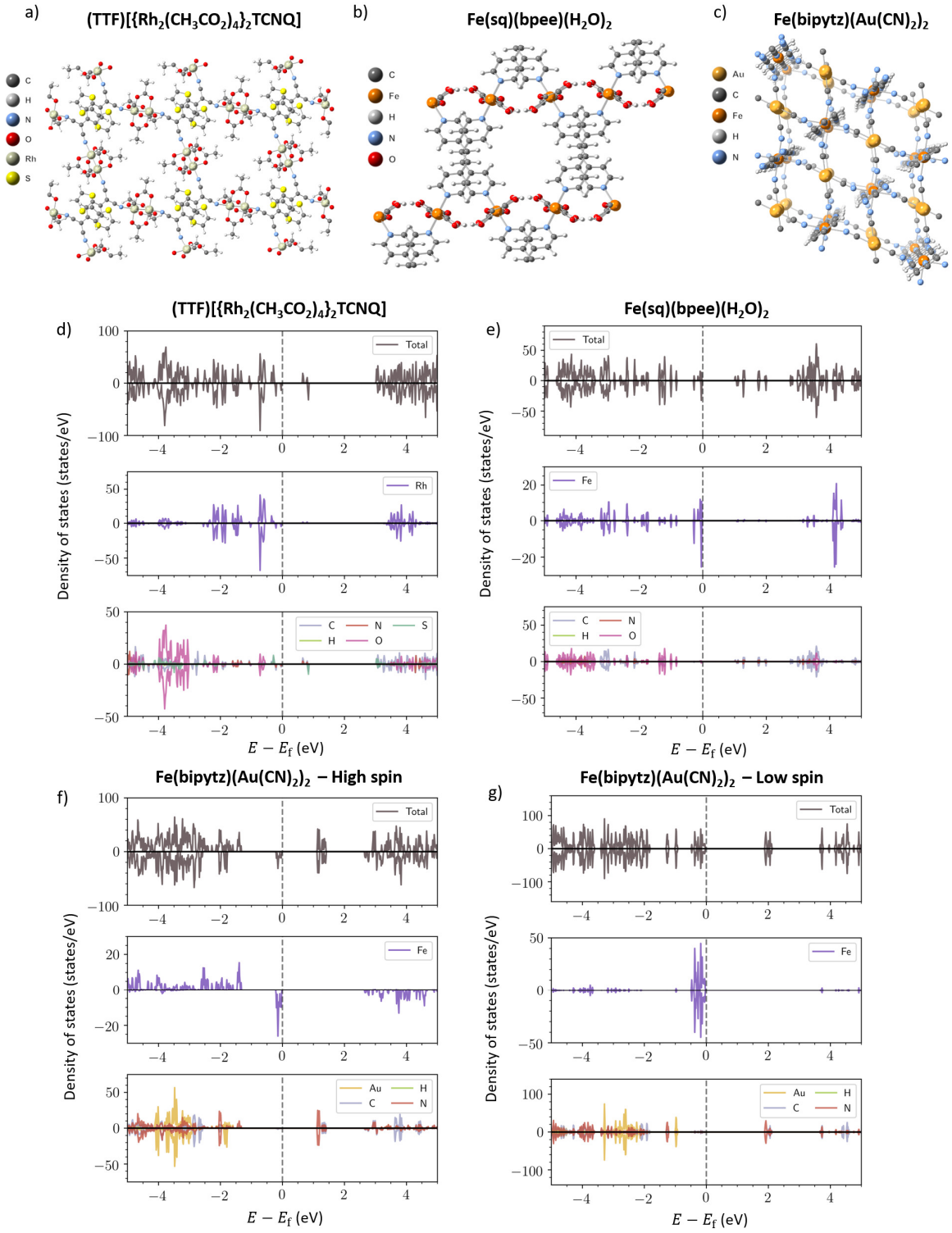


Figure 7. Structures of a) (TTF)[{$Rh_2(CH_3CO_2)_4$}$_2$TCNQ], b) Fe(sq)(bpee)(H_2O)$_2$, and c) Fe(bipytz)($Au(CN)_2$)$_2$. Total and projected density of states (DOS) at the HSE06-D3(BJ) level of theory for d) (TTF)[{$Rh_2(CH_3CO_2)_4$}$_2$TCNQ], e) Fe(sq)(bpee)(H_2O)$_2$, f) Fe(bipytz)($Au(CN)_2$)$_2$ (high spin), and g) Fe(bipytz)($Au(CN)_2$)$_2$ (low spin). The energy, E, in eV is shown with respect to the Fermi level, E_f. DOS values above and below zero refer to the spin-up and spin-down channels, respectively.

Another noteworthy example is the three-dimensional porous framework material Fe(bipytz)$(Au(CN)_2)_2$ (bipytz = 3,6-bis(4-pyridyl)-1,2,4,5-tetrazine) (refcode: LOJLAZ[150]), shown in Figure 7c. At the HSE06-D3(BJ) level of theory, we find that the high spin state exhibits a band gap of 1.17 eV (Figure 7f) – similar to that of Fe(sq)(bpee)$(H_2O)_2$. The projected density of states indicates that the Au(I) species are unrelated to the relatively low band gap; instead, the low band gap can be attributed to the combination of Fe(II) and bipytz linker. Fe(bipytz)$(Au(CN)_2)_2$ is known to be a spin-crossover framework (with a sharp spin transition around 290 K),[150] and we find the low spin HSE06-D3(BJ) band gap to be 1.95 eV (Figure 7g), suggesting that the material may have tunable electronic properties as a function of temperature. For the low spin case, the VBM is composed of Fe 3*d* orbitals and the CBM is composed of N 2*p* orbitals. The reduction in band gap from low spin to high spin state can be rationalized on the basis of crystal field theory. In the high spin state, the Fe(II) centers have a $t_{2g}^4e_g^2$ electronic configuration, whereas in the low spin state they have a $t_{2g}^6e_g^0$ electron configuration. This occupation of the e_g orbitals in the high spin state is directly related to the predicted ~0.8 eV reduction in the band gap compared to the low spin state. For both highlighted Fe-containing frameworks, the band gaps are lower – or comparable in the low spin state for Fe(bipytz)$(Au(CN)_2)$ – to those of several iron-containing MOFs that have been studied for their conductive properties, such as Fe_2(dobdc), Fe_2(dsbdc) (H_4dsbdc = 2,5-disulfhydrylbenzene-1,4-dicarboxylic acid), and Fe(bpz).[99,148] Collectively, these findings demonstrate the practical utility of the QMOF database for identifying MOFs with targeted quantum-chemical properties.

Conclusion

In this work, we have developed a database of quantum-chemical properties for over 14,000 MOF structures (the "QMOF database")[59] via a high-throughput periodic DFT workflow. DFT-computed geometries, energies, band gaps, partial charges, spin densities, bond orders, and related electronic structure properties are made publicly available. We highlight how this database can be used to identify MOFs with targeted electronic structure properties and then develop several ML models to predict the DFT-computed band gaps using descriptors derived from the un-optimized MOF crystal structures. A crystal graph convolutional neural network (CGCNN)[65] is found to achieve high predictive performance for this task, making it possible to circumvent large numbers of computationally expensive DFT calculations in future studies. While not as accurate as CGCNN for regression purposes, we show that both the smooth overlap of atomic positions (SOAP)[112,113] and composition-based features[109] can be used to discover otherwise subtle structure–property relationships in the QMOF database via unsupervised dimensionality reduction techniques. Finally, we show how top-performing ML models generated from the database of DFT-computed properties can be used to aid in the discovery of MOFs with desired quantum-chemical properties – in this case, discovering MOFs with low band gaps that could be suitable candidates to consider further for applications where electrical conductivity is necessary.

Importantly, the QMOF database now makes it possible to pursue several important research directions that are reliant on a large database of quantum-chemical properties for MOFs beyond those directly discussed in this work. For instance, with the success of transfer learning,[115,151] multi-task learning,[152] and Δ-ML[153] methods in materials research, the QMOF database can serve as a valuable resource to increase the accuracy – and reduce the required training set size – for ML models tasked with the prediction of new MOF properties not present in the QMOF database. Since the output of any ML models will depend on the chosen density functional approximation, related transfer learning approaches may also prove useful in generalizing ML model predictions to other levels of theory using the PBE-D3(BJ) data as a starting point. Instead of relying on representation approaches that were originally designed for inorganic solids or small molecules, the QMOF database can also be used to develop better methods for the encoding of MOF structures in ML models. Even outside the area of high-throughput DFT screening, data mining, and ML, there are countless possible use-cases for the QMOF database. As just one example, the DFT-generated properties in the QMOF database could be used to develop and/or benchmark (semi-)empirical methods (e.g. tight binding approaches[154] or molecular mechanics force fields[155]) with the hopes of achieving high

accuracies for MOF structures. Finally, we note that the QMOF database should be considered a living resource; several additions to the QMOF database are planned in the future, and we welcome the development of subsets, modifications, and supplements to the database that suit the diverse needs of the MOF community. With all this in mind, we anticipate that this database will accelerate the material design and discovery process while being specifically tailored for the chemical space of experimentally realized MOF structures.

Acknowledgments

A.S.R. was supported by a fellowship award through the National Defense Science and Engineering Graduate (NDSEG) Fellowship Program, sponsored by the Air Force Research Laboratory (AFRL), the Office of Naval Research (ONR) and the Army Research Office (ARO). A.S.R. also acknowledges support by a Ryan Fellowship from the International Institute for Nanotechnology and a Presidential Fellowship from The Graduate School at Northwestern University. This work was further supported by the U.S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences and Biosciences through the Nanoporous Materials Genome Center under Award Number DE-FG02-17ER16362. The authors acknowledge computing support from the Extreme Science and Engineering Discovery Environment (XSEDE) Stampede2 through allocation CTS180057 supported by National Science Foundation grant number ACI-1548562 (A.S.R, R.Q.S.), the Quest high-performance computing (HPC) facility at Northwestern University (A.S.R., S.M.I., R.Q.S.), the Mustang HPC environment via the Department of Defense High Performance Computing Modernization Program at the Air Force Research Laboratory (A.S.R.), and the Minnesota Supercomputing Institute at the University of Minnesota (D.R., L.G.). A.A.-G. thanks Dr. Anders G. Frøseth for his generous support.

Notes

R.Q.S. has a financial interest in the start-up company NuMat Technologies, which is seeking to commercialize metal–organic frameworks.

References

(1) Yaghi, O. M.; Kalmutzki, M. J.; Diercks, C. S. *Introduction to Reticular Chemistry: Metal-Organic Frameworks and Covalent Organic Frameworks*; John Wiley & Sons: Weinheim, Germany, 2019.

(2) Kalmutzki, M. J.; Hanikel, N.; Yaghi, O. M. Secondary Building Units as the Turning Point in the Development of the Reticular Chemistry of MOFs. *Sci. Adv.* **2018**, *4* (10), eaat9180.

(3) 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.

(4) 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.

(5) 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.

(6) 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.

(7) Boyd, P. G.; Woo, T. K. A Generalized Method for Constructing Hypothetical Nanoporous Materials of Any Net Topology from Graph Theory. *CrystEngComm* **2016**, *18* (21), 3777–3792.

(8) Colón, Y. J.; Snurr, R. Q. High-Throughput Computational Screening of Metal–Organic Frameworks. *Chem. Soc. Rev.* **2014**, *43* (16), 5735–5749.

(9) 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.
(10) Chibani, S.; Coudert, F.-X. Machine Learning Approaches for the Prediction of Materials Properties. *APL Mater.* **2020**, *8* (8), 80701.
(11) 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* **2018**, *123* (1), 120–130.
(12) Bucior, B. J.; Bobbitt, N. S.; Islamoglu, T.; Goswami, S.; Gopalan, A.; Yildirim, T.; Farha, O. K.; Bagheri, N.; Snurr, R. Q. Energy-Based Descriptors to Rapidly Predict Hydrogen Storage in Metal–Organic Frameworks. *Mol. Syst. Des. Eng.* **2019**, *4* (1), 162–174.
(13) Thornton, A. W.; Simon, C. M.; Kim, J.; Kwon, O.; Deeg, K. S.; Konstas, K.; Pas, S. J.; Hill, M. R.; Winkler, D. A.; Haranczyk, M.; Smit, B. Materials Genome in Action: Identifying the Performance Limits of Physical Hydrogen Storage. *Chem. Mater.* **2017**, *29* (7), 2844–2854.
(14) Anderson, R.; Rodgers, J.; Argueta, E.; Biong, A.; Gómez-Gualdrón, D. A. Role of Pore Chemistry and Topology in the CO_2 Capture Capabilities of MOFs: From Molecular Simulation to Machine Learning. *Chem. Mater.* **2018**, *30* (18), 6325–6337.
(15) 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), 4133–4139.
(16) Yao, Z.; Sanchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O.; Snurr, R. Q.; Aspuru-Guzik, A. Inverse Design of Nanoporous Crystalline Reticular Materials with Deep Generative Models. *ChemRxiv* **2020**. https://doi.org/10.26434/chemrxiv.12186681.v2.
(17) Shi, Z.; Yang, W.; Deng, X.; Cai, C.; Yan, Y.; Liang, H.; Liu, Z.; Qiao, Z. Machine-Learning-Assisted High-Throughput Computational Screening of High Performance Metal–Organic Frameworks. *Mol. Syst. Des. Eng.* **2020**, *5* (4), 725–742.
(18) Chong, S.; Lee, S.; Kim, B.; Kim, J. Applications of Machine Learning in Metal-Organic Frameworks. *Coord. Chem. Rev.* **2020**, *423*, 213487.
(19) Mancuso, J. L.; Mroz, A. M.; Le, K. N.; Hendon, C. H. Electronic Structure Modeling of Metal–Organic Frameworks. *Chem. Rev.* **2020**, *120* (16), 8641–8715.
(20) Raza, A.; Sturluson, A.; Simon, C.; Fern, X. Message Passing Neural Networks for Partial Charge Assignment to Metal-Organic Frameworks. *J. Phys. Chem. C* **2020**, *124* (35), 19070–19082.
(21) Korolev, V. V; Mitrofanov, A.; Marchenko, E. I.; Eremin, N. N.; Tkachenko, V.; Kalmykov, S. N. Transferable and Extensible Machine Learning Derived Atomic Charges for Modeling Hybrid Nanoporous Materials. *Chem. Mater.* **2020**, *32* (18), 7822–7831.
(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.
(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), 785–793.
(24) He, Y.; Cubuk, E. D.; Allendorf, M. D.; Reed, E. J. Metallic Metal–Organic Frameworks Predicted by the Combination of Machine Learning Methods and Ab Initio Calculations. *J. Phys. Chem. Lett.* **2018**, *9* (16), 4562–4569.
(25) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *Jom* **2013**, *65* (11), 1501–1509.
(26) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): Assessing the Accuracy of DFT Formation Energies. *npj Comput. Mater.* **2015**, *1* (1), 1–15.

(27) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865–3868.
(28) Abrahams, B. F.; Hardie, M. J.; Hoskins, B. F.; Robson, R.; Williams, G. A. Topological Rearrangement within a Single Crystal from a Honeycomb Cadmium Cyanide $[Cd(CN)_2]_n$ 3D Net to a Diamond Net. *J. Am. Chem. Soc.* **1992**, *114* (26), 10641–10643.
(29) von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring Chemical Compound Space with Quantum-Based Machine Learning. *Nat. Rev. Chem.* **2020**, *4*, 347–358.
(30) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559* (7715), 547–555.
(31) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *npj Comput. Mater.* **2017**, *3* (1), 1–13.
(32) Ward, L.; Wolverton, C. Atomistic Calculations and Materials Informatics: A Review. *Curr. Opin. Solid State Mater. Sci.* **2017**, *21* (3), 167–176.
(33) Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Deng, Z.; Ong, S. P. A Critical Review of Machine Learning of Energy Materials. *Adv. Energy Mater.* **2020**, *10* (8), 1903242.
(34) Morgan, D.; Jacobs, R. Opportunities and Challenges for Machine Learning in Materials Science. *Annu. Rev. Mater. Res.* **2020**, *50*, 71–103.
(35) Saal, J. E.; Oliynyk, A. O.; Meredig, B. Machine Learning in Materials Discovery: Confirmed Predictions and Their Underlying Approaches. *Annu. Rev. Mater. Res.* **2020**, *50* (1), 49–69.
(36) Suh, C.; Fare, C.; Warren, J. A.; Pyzer-Knapp, E. O. Evolving the Materials Genome: How Machine Learning Is Fueling the Next Generation of Materials Discovery. *Annu. Rev. Mater. Res.* **2020**, *50*, 1–25.
(37) Jain, A.; Hautier, G.; Moore, C. J.; Ping Ong, S.; Fischer, C. C.; Mueller, T.; Persson, K. A.; Ceder, G. A High-Throughput Infrastructure for Density Functional Theory Calculations. *Comput. Mater. Sci.* **2011**, *50* (8), 2295–2310.
(38) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1* (1), 11002.
(39) Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R. H.; Nelson, L. J.; Hart, G. L. W.; Sanvito, S.; Buongiorno-Nardelli, M.; Mingo, N.; Levy, O. AFLOWLIB.ORG: A Distributed Materials Properties Repository from High-Throughput Ab Initio Calculations. *Comput. Mater. Sci.* **2012**, *58*, 227–235.
(40) Winther, K. T.; Hoffmann, M. J.; Boes, J. R.; Mamun, O.; Bajdich, M.; Bligaard, T. Catalysis-Hub. Org, an Open Electronic Structure Database for Surface Reactions. *Sci. Data* **2019**, *6* (1), 1–10.
(41) Borysov, S. S.; Geilhufe, R. M.; Balatsky, A. V. Organic Materials Database: An Open-Access Online Database for Data Mining. *PLoS One* **2017**, *12* (2).
(42) Landis, D. D.; Hummelshoj, J. S.; Nestorov, S.; Greeley, J.; Dulak, M.; Bligaard, T.; Nørskov, J. K.; Jacobsen, K. W. The Computational Materials Repository. *Comput. Sci. Eng.* **2012**, *14* (6), 51–57.
(43) Draxl, C.; Scheffler, M. NOMAD: The FAIR Concept for Big Data-Driven Materials Science. *MRS Bull.* **2018**, *43* (9), 676–682.
(44) Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; Palizhati, A.; Sriram, A.; Wood, B.; Yoon, J.; Parikh, D.; Zitnick, C. L.; Ulissi, Z. The Open Catalyst 2020 (OC20) Dataset and Community Challenges. *arXiv. arXiv:2010.09990* **2020**.
(45) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 140022.
(46) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Machine Learning of Molecular Electronic Properties in Chemical Compound Space. *New J. Phys.* **2013**, *15* (9), 95003.

(47) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1, A Data Set of 20 Million Calculated off-Equilibrium Conformations for Organic Molecules. *Sci. Data* **2017**, *4*, 170193.
(48) Blau, S.; Spotte-Smith, E.; Wood, B.; Dwaraknath, S.; Persson, K. Accurate, Automated Density Functional Theory for Complex Molecules Using On-the-Fly Error Correction. *ChemRxiv* **2020**. https://doi.org/10.26434/chemrxiv.13076030.v1.
(49) Balcells, D.; Skjelstad, B. B. The TmQM Dataset - Quantum Geometries and Properties of 86k Transition Metal Complexes. *ChemRxiv* **2020**. https://doi.org/10.26434/chemrxiv.12894818.v1.
(50) Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazakim, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I.; Baldo, M.; Adam, R. P.; Aspuru-Guzik, A. Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016**, *15* (10), 1120–1127.
(51) Mansouri Tehrani, A.; Oliynyk, A. O.; Parry, M.; Rizvi, Z.; Couper, S.; Lin, F.; Miyagi, L.; Sparks, T. D.; Brgoch, J. Machine Learning Directed Search for Ultraincompressible, Superhard Materials. *J. Am. Chem. Soc.* **2018**, *140* (31), 9844–9853.
(52) Wu, S.; Kondo, Y.; Kakimoto, M.; Yang, B.; Yamada, H.; Kuwajima, I.; Lambard, G.; Hongo, K.; Xu, Y.; Shiomi, J.; Schick, C.; Morikawa, J.; Yoshida, R. Machine-Learning-Assisted Discovery of Polymers with High Thermal Conductivity Using a Molecular Design Algorithm. *npj Comput. Mater.* **2019**, *5* (1), 1–11.
(53) Lyu, H.; Ji, Z.; Wuttke, S.; Yaghi, O. M. Digital Reticular Chemistry. *Chem* **2020**, *6* (9), 2219–2241.
(54) Flores-Leonar, M. M.; Mejía-Mendoza, L. M.; Aguilar-Granda, A.; Sanchez-Lengeling, B.; Tribukait, H.; Amador-Bedolla, C.; Aspuru-Guzik, A. Materials Acceleration Platforms: On the Way to Autonomous Experimentation. *Curr. Opin. Green Sustain. Chem.* **2020**, *25*, 100370.
(55) Tabor, D. P.; Roch, L. C.; Saikin, S. K.; Kreisbeck, C.; Sheberla, D.; Montoya, J. H.; Dwaraknath, S.; Aykol, M.; Ortiz, C.; Tribukait, H.; Amador-Bedolla, C.; Brabec, C. J.; Maruyama, B.; Persson, K. A.; Aspuru-Guzik, A. Accelerating the Discovery of Materials for Clean Energy in the Era of Smart Automation. *Nat. Rev. Mater.* **2018**, *3* (5), 5–20.
(56) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angew. Chem. Int. Ed.* **2019**. https://doi.org/10.1002/anie.201909987.
(57) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part II: Outlook. *Angew. Chem. Int. Ed.* **2019**. https://doi.org/10.1002/anie.201909989.
(58) Rosen, A. S.; Notestein, J. M.; Snurr, R. Q. Identifying Promising Metal−Organic Frameworks for Heterogeneous Catalysis via High-Throughput Periodic Density Functional Theory. *J. Comput. Chem.* **2019**, *40* (12), 1305–1318.
(59) QMOF Database. https://github.com/arosen93/QMOF, https://dx.doi.org/10.6084/m9.figshare.13147324.
(60) Xie, L. S.; Skorupskii, G.; Dincă, M. Electrically Conductive Metal–Organic Frameworks. *Chem. Rev.* **2020**, *120* (16), 8536–8580.
(61) Baumann, A. E.; Burns, D. A.; Liu, B.; Thoi, V. S. Metal–Organic Framework Functionalization and Design Strategies for Advanced Electrochemical Energy Storage Devices. *Commun. Chem.* **2019**, *2* (1), 1–14.
(62) D'Alessandro, D. M. Exploiting Redox Activity in Metal–Organic Frameworks: Concepts, Trends and Perspectives. *Chem. Commun.* **2016**, *52* (58), 8957–8971.
(63) Downes, C. A.; Marinescu, S. C. Electrocatalytic Metal–Organic Frameworks for Energy Applications. *ChemSusChem* **2017**, *10* (22), 4374–4392.
(64) Allendorf, M. D.; Dong, R.; Feng, X.; Kaskel, S.; Matoga, D.; Stavila, V. Electronic Devices Using Open Framework Materials. *Chem. Rev.* **2020**, *120* (16), 8581–8640.
(65) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120* (14), 145301.
(66) 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.
(67) Cole, J. C.; Wiggin, S.; Stanzione, F. New Insights and Innovation from a Million Crystal Structures in the Cambridge Structural Database. *Struct. Dyn.* **2019**, *6* (5), 54301.
(68) Li, A.; Bueno-Perez, R.; Wiggin, S.; Fairen-Jimenez, D. Enabling Efficient Exploration of Metal–Organic Frameworks in the Cambridge Structural Database. *CrystEngComm* **2020**. https://doi.org/10.1039/D0CE00299B.
(69) 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* **2002**, *58* (3–1), 389–397.
(70) Zarabadi-Poor, P.; Marek, R. Comment on "Database for CO_2 Separation Performances of MOFs Based on Computational Materials Screening." *ACS Appl. Mater. Interfaces* **2019**, *11* (18), 16261–16265.
(71) Barthel, S.; Alexandrov, E. V; Proserpio, D. M.; Smit, B. Distinguishing Metal−Organic Frameworks. *Cryst. Growth Des.* **2018**, *18* (3), 1738–1747.
(72) Altintas, C.; Avci, G.; Daglar, H.; Azar, A. N. V.; Erucar, I.; Velioglu, S.; Keskin, S. An Extensive Comparative Analysis of Two MOF Databases: High-Throughput Screening of Computation-Ready MOFs for CH_4 and H_2 Adsorption. *J. Mater. Chem. A* **2019**, *7* (16), 9593–9608.
(73) Velioglu, S.; Keskin, S. Revealing the Effect of Structure Curations on the Simulated CO_2 Separation Performances of MOFs. *Mater. Adv.* **2020**, *1* (3), 341–353.
(74) Bucior, B. J.; Rosen, A. S.; Haranczyk, M.; Yao, Z.; Ziebel, M. E.; Farha, O. K.; Hupp, J. T.; Siepmann, J. I.; Aspuru-Guzik, A.; Snurr, R. Q. Identification Schemes for Metal–Organic Frameworks to Enable Rapid Search and Cheminformatics Analysis. *Cryst. Growth Des.* **2019**, *19* (11), 6682–6697.
(75) Chen, T.; Manz, T. A. Identifying Misbonded Atoms in the 2019 CoRE Metal–Organic Framework Database. *RSC Adv.* **2020**, *10* (45), 26944–26951.
(76) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H−Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104.
(77) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32* (7), 1456–1465.
(78) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54* (16), 11169–11186.
(79) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B* **1999**, *59* (3), 1758–1775.
(80) Hendon, C. H.; Tiana, D.; Walsh, A. Conductive Metal–Organic Frameworks and Networks: Fact or Fantasy? *Phys. Chem. Chem. Phys.* **2012**, *14* (38), 13120–13132.
(81) Sun, L.; Campbell, M. G.; Dincă, M. Electrically Conductive Porous Metal–Organic Frameworks. *Angew. Chem. Int. Ed.* **2016**, *55* (11), 3566–3579.
(82) Singh, A. K.; Montoya, J. H.; Gregoire, J. M.; Persson, K. A. Robust and Synthesizable Photocatalysts for CO_2 Reduction: A Data-Driven Materials Discovery. *Nat. Commun.* **2019**, *10*, 443.
(83) Yang, L.-M.; Ravindran, P.; Vajeeston, P.; Svelle, S.; Tilset, M. A Quantum Mechanically Guided View of Cd-MOF-5 from Formation Energy, Chemical Bonding, Electronic Structure, and Optical Properties. *Microporous Mesoporous Mater.* **2013**, *175*, 50–58.
(84) Gong, S.; Xie, T.; Zhu, T.; Wang, S.; Fadel, E. R.; Li, Y.; Grossman, J. C. Predicting Charge Density Distribution of Materials Using a Local-Environment-Based Graph Convolutional Network. *Phys. Rev. B* **2019**, *100* (18), 184103.
(85) Grisafi, A.; Fabrizio, A.; Meyer, B.; Wilkins, D. M.; Corminboeuf, C.; Ceriotti, M. Transferable Machine-Learning Model of the Electron Density. *ACS Cent. Sci.* **2018**, *5* (1), 57–64.
(86) Chandrasekaran, A.; Kamal, D.; Batra, R.; Kim, C.; Chen, L.; Ramprasad, R. Solving the

Electronic Structure Problem with Machine Learning. *npj Comput. Mater.* **2019**, *5* (1), 1–7.
(87) Kamal, D.; Chandrasekaran, A.; Batra, R.; Ramprasad, R. A Charge Density Prediction Model for Hydrocarbons Using Deep Neural Networks. *Mach. Learn. Sci. Technol.* **2020**, *1* (2), 25003.
(88) Kolb, B.; Lentz, L. C.; Kolpak, A. M. Discovering Charge Density Functionals and Structure-Property Relationships with PROPhet: A General Framework for Coupling Machine Learning and First-Principles Methods. *Sci. Rep.* **2017**, *7* (1), 1–9.
(89) Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating Materials Property Predictions Using Machine Learning. *Sci. Rep.* **2013**, *3* (1), 1–6.
(90) Manz, T. A.; Limas, N. G. Introducing DDEC6 Atomic Population Analysis: Part 1. Charge Partitioning Theory and Methodology. *RSC Adv.* **2016**, *6* (53), 47771–47801.
(91) Limas, N. G.; Manz, T. A. Introducing DDEC6 Atomic Population Analysis: Part 2. Computed Results for a Wide Range of Periodic and Nonperiodic Materials. *RSC Adv.* **2016**, *6* (51), 45727–45747.
(92) Limas, N. G.; Manz, T. A. Introducing DDEC6 Atomic Population Analysis: Part 4. Efficient Parallel Computation of Net Atomic Charges, Atomic Spin Moments, Bond Orders, and More. *RSC Adv.* **2018**, *8* (5), 2678–2707.
(93) Manz, T. A. Introducing DDEC6 Atomic Population Analysis: Part 3. Comprehensive Method to Compute Bond Orders. *RSC Adv.* **2017**, *7* (72), 45552–45581.
(94) Rosen, A. S.; Notestein, J. M.; Snurr, R. Q. Structure−Activity Relationships That Identify Metal−Organic Framework Catalysts for Methane Activation. *ACS Catal.* **2019**, *9*, 3576–3587.
(95) Yang, B.; Wu, X.-P.; Gagliardi, L.; Truhlar, D. G. Methane Functionalization by an Ir(III) Catalyst Supported on a Metal–Organic Framework: An Alternative Explanation of Steric Confinement Effects. *Theor. Chem. Acc.* **2019**, *138* (9), 107.
(96) Xiao, D. J.; Bloch, E. D.; Mason, J. A.; Queen, W. L.; Hudson, M. R.; Planas, N.; Borycz, J.; Dzubak, A. L.; Verma, P.; Lee, K.; Bonino, F.; Crocellà, V.; Yano, J.; Bordiga, S.; Truhlar, D. G.; Gagliardi, L.; Brown, C. M.; Long, J. R. Oxidation of Ethane to Ethanol by N_2O in a Metal−Organic Framework with Coordinatively Unsaturated Iron(II) Sites. *Nat. Chem.* **2014**, *6* (7), 590–595.
(97) Vogiatzis, K. D.; Haldoupis, E.; Xiao, D. J.; Long, J. R.; Siepmann, J. I.; Gagliardi, L. Accelerated Computational Analysis of Metal−Organic Frameworks for Oxidation Catalysis. *J. Phys. Chem. C* **2016**, *120* (33), 18707–18712.
(98) Osadchii, D.; Olivos Suarez, A. I.; Szécsényi, Á.; Li, G.; Nasalevich, M. A.; Dugulan, A. I.; Serra-Crespo, P.; Hensen, E. J. M.; Veber, S. L.; Fedin, M. V; Sankar, G.; Pidko, E. A.; Gascon, J. Isolated Fe Sites in Metal−Organic Framework Catalyze the Direct Conversion of Methane to Methanol. *ACS Catal.* **2018**, *8* (6), 5542–5548.
(99) Spirkl, S.; Grzywa, M.; Reschke, S.; Fischer, J. K. H.; Sippel, P.; Demeshko, S.; von Nidda, H.-A.; Volkmer, D. Single-Crystal to Single-Crystal Transformation of a Nonporous Fe(II) Metal−Organic Framework into a Porous Metal-Organic Framework via a Solid-State Reaction. *Inorg. Chem.* **2017**, *56* (20), 12337–12347.
(100) Queen, W. L.; Hudson, M. R.; Bloch, E. D.; Mason, J. A.; Gonzalez, M. I.; Lee, J. S.; Gygi, D.; Howe, J. D.; Lee, K.; Darwish, T. A.; James, M.; Peterson, V. K.; Teat, S. J.; Smit, B.; Neaton, J. B.; Long, J. R.; Brown, C. M. Comprehensive Study of Carbon Dioxide Adsorption in the Metal−Organic Frameworks M_2(dobdc)(M = Mg, Mn, Fe, Co, Ni, Cu, Zn). *Chem. Sci.* **2014**, *5* (12), 4569–4581.
(101) Bloch, E. D.; Queen, W. L.; Krishna, R.; Zadrozny, J. M.; Brown, C. M.; Long, J. R. Hydrocarbon Separations in a Metal−Organic Framework with Open Iron(II) Coordination Sites. *Science* **2012**, *335* (6076), 1606–1610.
(102) Reed, D. A.; Keitz, B. K.; Oktawiec, J.; Mason, J. A.; Runčevski, T.; Xiao, D. J.; Darago, L. E.; Crocellà, V.; Bordiga, S.; Long, J. R. A Spin Transition Mechanism for Cooperative Adsorption in Metal−Organic Frameworks. *Nature* **2017**, *550* (7674), 96–100.
(103) Mason, J. A.; Oktawiec, J.; Taylor, M. K.; Hudson, M. R.; Rodriguez, J.; Bachman, J. E.;

Gonzalez, M. I.; Cervellino, A.; Guagliardi, A.; Brown, C. M.; Llewellyn, P. L.; Masciocchi, N.; Long, J. R. Methane Storage in Flexible Metal−Organic Frameworks with Intrinsic Thermal Management. *Nature* **2015**, *527* (7578), 357–361.
(104) Verma, P.; Vogiatzis, K. D.; Planas, N.; Borycz, J.; Xiao, D. J.; Long, J. R.; Gagliardi, L.; Truhlar, D. G. Mechanism of Oxidation of Ethane to Ethanol at Iron(IV)-Oxo Sites in Magnesium-Diluted Fe_2(DOBDC). *J. Am. Chem. Soc.* **2015**, *137* (17), 5770–5781.
(105) Rosen, A. S.; Notestein, J. M.; Snurr, R. Q. High-Valent Metal-Oxo Species at the Nodes of Metal−Triazolate Frameworks: The Effects of Ligand-Exchange and Two-State Reactivity for C−H Bond Activation. *Angew. Chem. Int. Ed.* **2020**, *132* (44), 19662–19670.
(106) Rosen, A. S.; Mian, M. R.; Islamoglu, T.; Chen, H.; Farha, O. K.; Notestein, J. M.; Snurr, R. Q. Tuning the Redox Activity of Metal−Organic Frameworks for Enhanced, Selective O_2 Binding: Design Rules and Ambient Temperature O_2 Chemisorption in a Cobalt−Triazolate Framework. *J. Am. Chem. Soc.* **2020**, *142* (9), 4317–4328.
(107) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864–2875.
(108) Li, H.; Eddaoudi, M.; O'Keeffe, M.; Yaghi, O. M. Design and Synthesis of an Exceptionally Stable and Highly Porous Metal−Organic Framework. *Nature* **1999**, *402* (6759), 276–279.
(109) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial Screening for New Materials in Unconstrained Composition Space with Machine Learning. *Phys. Rev. B* **2014**, *89* (9), 94104.
(110) Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Crystal Structure Representations for Machine Learning Models of Formation Energies. *Int. J. Quantum Chem.* **2015**, *115* (16), 1094–1101.
(111) Lam Pham, T.; Kino, H.; Terakura, K.; Miyake, T.; Tsuda, K.; Takigawa, I.; Chi Dam, H. Machine Learning Reveals Orbital Interaction in Materials. *Sci. Technol. Adv. Mater.* **2017**, *18* (1), 756–765.
(112) Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B* **2013**, *87* (18), 184115.
(113) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing Molecules and Solids across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* **2016**, *18* (20), 13754–13769.
(114) Pronobis, W.; Müller, K.-R. Kernel Methods for Quantum Chemistry. In *Machine Learning Meets Quantum Physics*; Springer, 2020; pp 25–36.
(115) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31* (9), 3564–3572.
(116) Olsthoorn, B.; Geilhufe, R. M.; Borysov, S. S.; Balatsky, A. V. Band Gap Prediction for Large Organic Crystal Structures with Machine Learning. *Adv. Quantum Technol.* **2019**, *2* (7–8), 1900023.
(117) Gascon, J.; Hernández-Alonso, M. D.; Almeida, A. R.; van Klink, G. P. M.; Kapteijn, F.; Mul, G. Isoreticular MOFs as Efficient Photocatalysts with Tunable Band Gap: An Operando FTIR Study of the Photoinduced Oxidation of Propylene. *ChemSusChem* **2008**, *1* (12), 981–983.
(118) Stanzione, D.; Barth, B.; Gaffney, N.; Gaither, K.; Hempel, C.; Minyard, T.; Mehringer, S.; Wernert, E.; Tufo, H.; Panda, D.; Teller, P. Stampede 2: The Evolution of an XSEDE Supercomputer. In *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*; 2017; pp 1–8.
(119) Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazlewood, V.; Lathrop, S.; Lifka, D.; Peterson, G. D.; Roskies, R.; Scott, J. R.; Wilkins-Diehr, N. XSEDE: Accelerating Scientific Discovery. *Comput. Sci. Eng.* **2014**, *16* (5), 62–74.
(120) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv. arXiv:1802.03426* **2018**.
(121) Leland, M.; John, H.; Nathaniel, S.; Lukas, G. UMAP: Uniform Manifold Approximation and

Projection. *J. Open Source Softw.* **2018**, *3* (29), 861.
(122) Zeng, X.-Z.; Zhang, A.-Y.; Bu, D.; Li, Y.-W. Hydrothermal Synthesis, Structure and Thermal Properties of a Novel Three-Dimensional La(III)-Sebacate Framework. *Chinese J. Struct. Chem.* **2013**, *32* (1), 120–124.
(123) Martínez-Casado, F. J.; Ramos-Riesco, M.; Rodríguez-Cheda, J. A.; Cucinotta, F.; Fernández-Martínez, A.; Garrido, L.; Matesanz, E.; Marchese, L. Short Lead(II) Soaps: From Weakly Fluorescent Crystals to Strongly Phosphorescent and Structurally Varied Vitreous Phases. A Thermal, Structural and Spectroscopic Study. *J. Mater. Chem. C* **2014**, *2* (44), 9489–9496.
(124) Zhang, Z.; Zhao, H.; Matsushita, M. M.; Awaga, K.; Dunbar, K. R. A New Metal–Organic Hybrid Material with Intrinsic Resistance-Based Bistability: Monitoring in Situ Room Temperature Switching Behavior. *J. Mater. Chem. C* **2014**, *2* (2), 399–404.
(125) Lopez, N.; Zhao, H.; Ota, A.; Prosvirin, A. V; Reinheimer, E. W.; Dunbar, K. R. Unprecedented Binary Semiconductors Based on TCNQ: Single-Crystal X-Ray Studies and Physical Properties of $Cu(TCNQX_2)$ X = Cl, Br. *Adv. Mater.* **2010**, *22* (9), 986–989.
(126) Aubrey, M. L.; Kapelewski, M. T.; Melville, J. F.; Oktawiec, J.; Presti, D.; Gagliardi, L.; Long, J. R. Chemiresistive Detection of Gaseous Hydrocarbons and Interrogation of Charge Transport in $Cu[Ni(2,3\text{-Pyrazinedithiolate})_2]$ by Gas Adsorption. *J. Am. Chem. Soc.* **2019**, *141* (12), 5005–5013.
(127) Peng, Y.-L.; Pham, T.; Li, P.; Wang, T.; Chen, Y.; Chen, K.-J.; Forrest, K. A.; Space, B.; Cheng, P.; Zaworotko, M. J.; Zhang, Z. Robust Ultramicroporous Metal–Organic Frameworks with Benchmark Affinity for Acetylene. *Angew. Chem. Int. Ed.* **2018**, *57* (34), 10971–10975.
(128) Takaishi, S.; Hosoda, M.; Kajiwara, T.; Miyasaka, H.; Yamashita, M.; Nakanishi, Y.; Kitagawa, Y.; Yamaguchi, K.; Kobayashi, A.; Kitagawa, H. Electroconductive Porous Coordination Polymer $Cu[Cu(pdt)_2]$ Composed of Donor and Acceptor Building Units. *Inorg. Chem.* **2009**, *48* (19), 9048–9050.
(129) Kobayashi, Y.; Jacobs, B.; Allendorf, M. D.; Long, J. R. Conductivity, Doping, and Redox Chemistry of a Microporous Dithiolene-Based Metal–Organic Framework. *Chem. Mater.* **2010**, *22* (14), 4120–4122.
(130) Nicholas, T. C.; Goodwin, A.; Deringer, V. L. Understanding the Geometric Diversity of Inorganic and Hybrid Frameworks through Structural Coarse-Graining. *Chem. Sci.* https://doi.org/10.1039/D0SC03287E.
(131) 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*, 4068.
(132) Hulvey, Z.; Furman, J. D.; Turner, S. A.; Tang, M.; Cheetham, A. K. Dimensionality Trends in Metal–Organic Frameworks Containing Perfluorinated or Nonfluorinated Benzenedicarboxylates. *Cryst. Growth Des.* **2010**, *10* (5), 2041–2043.
(133) Taylor, M. K.; Runčevski, T.; Oktawiec, J.; Gonzalez, M. I.; Siegelman, R. L.; Mason, J. A.; Ye, J.; Brown, C. M.; Long, J. R. Tuning the Adsorption-Induced Phase Change in the Flexible Metal–Organic Framework Co(bdp). *J. Am. Chem. Soc.* **2016**, *138* (45), 15019–15026.
(134) Cui, X.; Chen, K.; Xing, H.; Yang, Q.; Krishna, R.; Bao, Z.; Wu, H.; Zhou, W.; Dong, X.; Han, Y.; Li, B.; Ren, Q.; Zaworotko, M. J.; Chen, B. Pore Chemistry and Size Control in Hybrid Porous Materials for Acetylene Capture from Ethylene. *Science* **2016**, *353* (6295), 141–144.
(135) Li, S.; Zhang, L.; Lu, B.; Yan, E.; Wang, T.; Li, L.; Wang, J.; Yu, Y.; Mu, Q. A New Polyoxovanadate-Based Metal–Organic Framework: Synthesis, Structure and Photo-/Electro-Catalytic Properties. *New J. Chem.* **2018**, *42* (9), 7247–7253.
(136) Yan, B.; Luo, J.; Dube, P.; Sefat, A. S.; Greedan, J. E.; Maggard, P. A. Spin-Gap Formation and Thermal Structural Studies in Reduced Hybrid Layered Vanadates. *Inorg. Chem.* **2006**, *45* (13), 5109–5118.
(137) Lysenko, A. B.; Senchyk, G. A.; Domasevitch, K. V; Hauser, J.; Fuhrmann, D.; Kobalz, M.; Krautscheid, H.; Neves, P.; Valente, A. A.; Gonçalves, I. S. Synthesis and Structural Elucidation

of Triazolylmolybdenum(VI) Oxide Hybrids and Their Behavior as Oxidation Catalysts. *Inorg. Chem.* **2015**, *54* (17), 8327–8338.
(138) Wang, X.-L.; Zhang, S.; Wang, X.; Zhang, H.-X.; Xu, N.; Li, T.-J. Three β-Octamolybdate-Based Supramolecular Hybrids Constructed from a Bis-Imidazolyl-Bis-Amide Ligand: Fast and Selective Adsorption Activities of Organic Dyes. *New J. Chem.* **2017**, *41* (5), 2178–2185.
(139) Choudhuri, I.; Truhlar, D. G. HLE17: An Efficient Way To Predict Band Gaps of Complex Materials. *J. Phys. Chem. C* **2019**, *123* (28), 17416–17424.
(140) Borlido, P.; Aull, T.; Huran, A. W.; Tran, F.; Marques, M. A. L.; Botti, S. Large-Scale Benchmark of Exchange–Correlation Functionals for the Determination of Electronic Band Gaps of Solids. *J. Chem. Theory Comput.* **2019**, *15* (9), 5069–5079.
(141) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. Hybrid Functionals Based on a Screened Coulomb Potential. *J. Chem. Phys.* **2003**, *118* (18), 8207–8215.
(142) Moellmann, J.; Grimme, S. DFT-D3 Study of Some Molecular Crystals. *J. Phys. Chem. C* **2014**, *118* (14), 7615–7621.
(143) Valenzano, L.; Civalleri, B.; Chavan, S.; Bordiga, S.; Nilsen, M. H.; Jakobsen, S.; Lillerud, K. P.; Lamberti, C. Disclosing the Complex Structure of UiO-66 Metal Organic Framework: A Synergic Combination of Experiment and Theory. *Chem. Mater.* **2011**, *23* (7), 1700–1718.
(144) Saliba, D.; Ammar, M.; Rammal, M.; Al-Ghoul, M.; Hmadeh, M. Crystal Growth of ZIF-8, ZIF-67, and Their Mixed-Metal Derivatives. *J. Am. Chem. Soc.* **2018**, *140* (5), 1812–1823.
(145) Naito, T.; Kakizaki, A.; Inabe, T.; Sakai, R.; Nishibori, E.; Sawa, H. Growth of Nanocrystals in a Single Crystal of Different Materials: A Way of Giving Function to Molecular Crystals. *Cryst. Growth Des.* **2011**, *11* (2), 501–506.
(146) Sekine, Y.; Tonouchi, M.; Yokoyama, T.; Kosaka, W.; Miyasaka, H. Built-in TTF–TCNQ Charge-Transfer Salts in π-Stacked Pillared Layer Frameworks. *CrystEngComm* **2017**, *19* (17), 2300–2304.
(147) Salami, T. O.; Patterson, S. N.; Jones, V. D.; Masello, A.; Abboud, K. A. Synthesis, Characterization, Thermal Study, and Crystal Structure of a New Layered Alkaline Earth Metal Sulfonate: $Sr[C_2H_4(SO_3)_2]$. *Inorg. Chem. Commun.* **2009**, *12* (11), 1150–1153.
(148) Sun, L.; Hendon, C. H.; Park, S. S.; Tulchinsky, Y.; Wan, R.; Wang, F.; Walsh, A.; Dincă, M. Is Iron Unique in Promoting Electrical Conductivity in MOFs? *Chem. Sci.* **2017**, *8* (6), 4450–4457.
(149) Manna, S. C.; Zangrando, E.; Ribas, J.; Chaudhuri, N. R. Squarato-Bridged Polymeric Networks of Iron(II) with N-Donor Coligands: Syntheses, Crystal Structures and Magnetic Properties. *Inorganica Chim. Acta* **2005**, *358* (15), 4497–4504.
(150) Clements, J. E.; Price, J. R.; Neville, S. M.; Kepert, C. J. Perturbation of Spin Crossover Behavior by Covalent Post-Synthetic Modification of a Porous Metal–Organic Framework. *Angew. Chem. Int. Ed.* **2014**, *126* (38), 10328–10332.
(151) Lee, J.; Asahi, R. Transfer Learning for Materials Informatics Using Crystal Graph Convolutional Neural Network. *arXiv. arXiv:2007.09932* **2020**.
(152) Sanyal, S.; Balachandran, J.; Yadati, N.; Kumar, A.; Rajagopalan, P.; Sanyal, S.; Talukdar, P. MT-CGCNN: Integrating Crystal Graph Convolutional Neural Network with Multitask Learning for Material Property Prediction. *arXiv. arXiv:1811.05660* **2018**.
(153) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11* (5), 2087–2096.
(154) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15* (3), 1652–1671.
(155) Spicher, S.; Grimme, S. Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems. *Angew. Chem. Int. Ed.* **2020**, *59* (36), 15665–15673.