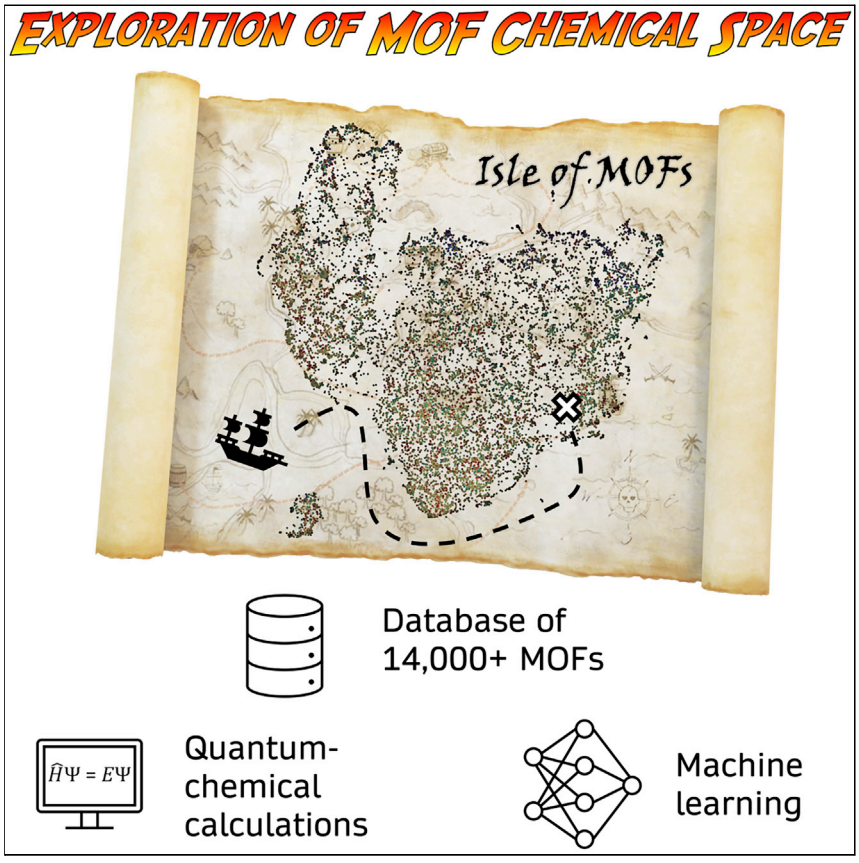
Matter

CellPress
OPEN ACCESS

Article

Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery



Using a new database of electronic structure properties derived from high-throughput density functional theory calculations for thousands of metal–organic frameworks (MOFs), we benchmark a variety of machine learning models to accurately and rapidly predict MOF band gaps. Unsupervised dimensionality reduction techniques are also used to map the MOF feature space and identify otherwise subtle structure–property relationships. We anticipate that machine learning models derived from this new database will accelerate the discovery of promising MOFs with targeted quantum-chemical properties.

Understanding
Dependency and conditional studies on material behavior

Andrew S. Rosen, Shaelyn M. Iyer, Debmalya Ray, ..., Laura Gagliardi, Justin M. Notestein, Randall Q. Snurr

rosen@u.northwestern.edu

HIGHLIGHTS

New publicly available database of electronic structure properties for 14,000+ MOFs

Machine learning models can rapidly and accurately predict computed MOF band gaps

A crystal graph convolutional neural network achieves high predictive performance

Several MOFs with low band gaps are computationally identified

Rosen et al., Matter 4, 1578–1597
May 5, 2021
https://doi.org/10.1016/j.matt.2021.02.015

CellPress
OPEN ACCESS

Matter

Article

Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery

Andrew S. Rosen,[1,7,8,*] Shaelyn M. Iyer,[1] Debmalya Ray,[2] Zhenpeng Yao,[3] Alán Aspuru-Guzik,[3,4,5] Laura Gagliardi,[6] Justin M. Notestein,[1] and Randall Q. Snurr[1]

SUMMARY

The modular nature of metal–organic frameworks (MOFs) enables synthetic control over their physical and chemical properties, but it can be difficult to know which MOFs would be optimal for a given application. High-throughput computational screening and machine learning are promising routes to efficiently navigate the vast chemical space of MOFs but have rarely been used for the prediction of properties that need to be calculated by quantum mechanical methods. Here, we introduce the Quantum MOF (QMOF) database, a publicly available database of computed quantum-chemical properties for more than 14,000 experimentally synthesized MOFs. Throughout this study, we demonstrate how machine learning models trained on the QMOF database can be used to rapidly discover MOFs with targeted electronic structure properties, using the prediction of theoretically computed band gaps as a representative example. We conclude by highlighting several MOFs predicted to have low band gaps, a challenging task given the electronically insulating nature of most MOFs.

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.[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,[8] 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 proved extremely useful for more efficiently exploring the vast combinatorial space of MOF structures.[9,10] Recently, the large quantities of data generated during these computational screening studies have led to the development of machine learning (ML) models[11] that can accelerate the MOF design and discovery process even further. ML-assisted screening studies have been successfully applied to

Progress and potential

Metal–organic frameworks (MOFs) are a class of crystalline solids with tunable structures that make it possible to impart chemical functionality tailored for a given application. A virtually unlimited number of MOFs can be realized, making it difficult to know which would be top-performing candidates. This is especially true for the many applications governed by quantum mechanical properties, which are computationally demanding to simulate and time-consuming to probe experimentally. Here, we present a database of quantum mechanical properties for thousands of MOFs, the result of over 170 years of computing time. With this new database, we developed machine learning models that can predict the electronic properties of MOFs in mere seconds without the need for expensive quantum mechanical simulations and powerful supercomputers. We anticipate that machine learning models constructed using this database will accelerate the discovery of MOFs for a variety of long-standing challenges.

the discovery of MOFs suitable for H_2 storage,[12–14] CO_2 separation/capture,[15–17] and numerous other applications predominantly (though not exclusively)[18,19] in the area of gas storage and separations.[10,20,21] Nonetheless, similar efforts remain almost entirely unexplored for the many applications in which the properties of interest are best described by quantum mechanical models,[22] such as those 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.[23] and Korolev et al.,[24] who independently developed ML models that can predict the partial atomic charges of MOFs in the Computation-Ready, Experimental (CoRE) MOF database.[25,26] 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.,[27] who trained binary classification models to predict whether inorganic solids in the Open Quantum Materials Database (OQMD)[28,29] are metallic or nonmetallic. Without retraining on MOF data, a multimodel voting procedure was then used to predict the metallic or nonmetallic behavior of 2,932 MOFs in the CoRE MOF database,[25] which do not have computed band gaps. Of the six identified materials with near-zero band gap at the PBE level of theory,[30] 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 $[Cd(CN)_2]_n$.[31]

In stark contrast to the existing literature on MOFs, significant progress has been made in the development of ML models that can accelerate the quantum-chemical screening process for a wide range of inorganic and molecular compounds.[32–39] One of the fundamental features underlying much of this work has been the use of high-throughput density functional theory[40] (DFT) workflows to construct large-scale electronic structure–property databases, such as those developed for inorganic solids[29,41–47] and molecular systems.[48–52] 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,[53] superhard inorganic materials,[54] and thermally conductive polymers,[55] among many others.[38] 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 for the rapid prediction of MOF electronic structure properties. High-throughput screening, database generation, and subsequent ML model development are crucial components for realizing the full potential of reticular chemistry[56] and accelerating materials discovery in general.[57–60]

In the present study, we leverage a recently developed high-throughput periodic DFT workflow tailored for MOF structures[61] to construct a large-scale database of MOF quantum mechanical properties. This publicly available dataset—the Quantum MOF (QMOF) database[62]—contains computed properties for 15,713 experimentally characterized MOFs after structure relaxation via DFT, including but not limited

[1]Department of Chemical and Biological Engineering, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, USA

[2]Department of Chemistry, Chemical Theory Center, and Minnesota Supercomputing Institute, University of Minnesota, 207 Pleasant Street SE, Minneapolis, MN 55455, USA

[3]Chemical Physics Theory Group, Department of Chemistry and Department of Computer Science, University of Toronto, Toronto, ON M5S 3H6, Canada

[4]Vector Institute for Artificial Intelligence, Toronto, ON M5S 1M1, Canada

[5]Lebovic Fellow, Canadian Institute for Advanced Research (CIFAR) Senior Fellow, Toronto, ON 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 South Ellis Avenue, Chicago, IL 60637, USA

[7]Lead contact

[8]Twitter: @Andrew_S_Rosen

*Correspondence: rosen@u.northwestern.edu

https://doi.org/10.1016/j.matt.2021.02.015

to optimized geometries, energies, band gaps, charge densities, density of states, partial charges, spin densities, and bond orders. We anticipate that the QMOF database will serve two primary purposes: (1) materials discovery using the as-deposited data; and (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, drastically decreasing the number of computationally demanding quantum mechanical simulations that would need to be carried out in future screening studies. 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,[63] which limits their potential use in electrocatalysis, sensing, energy storage, and other applications for which some degree of electrical conductivity is necessary.[63–67] We identify a top-performing band gap regression model based on a crystal graph convolutional neural network[68] and show how dimensionality reduction techniques can be used to discover overarching structure–property relationships for the identification of MOFs with targeted electronic structure properties. We conclude by highlighting several Fe MOFs with low band gaps identified for the first time in this work.

RESULTS AND DISCUSSION

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.[3–6,25,69] 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; 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 a comparatively "clean" dataset of crystal structures for high-throughput computational investigation, one we will refer to as a suitably "DFT-ready" dataset of MOFs.

We considered the list of materials identified as MOFs from both the Cambridge Structural Database (CSD) MOF subset[3] and the 2019 Computation-Ready, Experimental (CoRE) MOF database,[4] the latter of which contains a relatively small number of MOFs not present in the former. All starting structures were taken directly from the CSD by querying the corresponding CSD reference code ("refcode"), and free (i.e., unbound) solvents were automatically removed from the frameworks. We chose to take the initial structures directly from the CSD as a matter of consistency and so that we could make use of valuable CSD metadata[70] (e.g., unresolved atoms, charged structures) associated with each deposited crystal structure. From this set of experimental crystal structures, we constructed a smaller DFT-ready subset of 42,349 nondisordered MOF structures ("QMOF-42349") after an extensive suite of automated fidelity checks, as summarized in Figure S1. This process serves to filter out many problematic MOFs with omitted H atoms, fractional occupancies, deleted

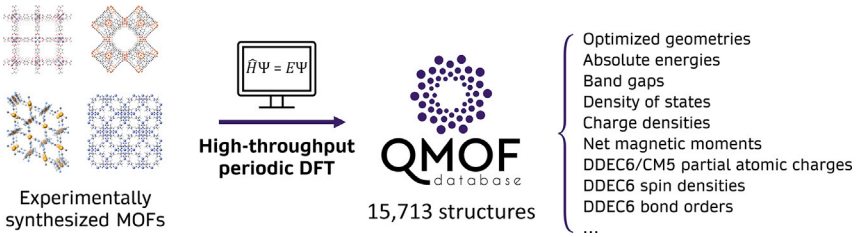


Figure 1. Overview of the QMOF database

Selected DFT-computed properties for the structurally relaxed MOFs made available in the QMOF database.

framework atoms, lone (i.e., unbonded) atoms, overlapping atoms, an improper number of charge-balancing ions, and other structural issues that have been discussed in several recent studies.[70–76] Of these 42,349 experimental crystal structures, a subset of materials with 300 atoms or fewer per primitive cell was considered such that high-throughput DFT calculations could be carried out in an efficient manner. Full structure relaxations (including cell volume and atomic positions) were carried out via a multistage workflow[61] (Table S2) at the PBE-D3(BJ)[30,77,78] level of theory with the Vienna *ab initio* Simulation Package.[79,80] Additional methodological details regarding the dataset construction, DFT calculations, and ML methods can be found in the supplemental information.

The high-throughput periodic DFT workflow was successfully completed for 15,713 MOFs, and several DFT-computed properties were tabulated following the structure relaxation process, a selection of which are listed in 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[63,65,81,82] or screening for photocatalytic materials.[83] Electronic energies, particularly if converted to formation energies, may provide insight into the relative stability of MOFs.[84] Machine learning the charge density[85] is a potential way to bypass a large portion of the calculations performed with Kohn-Sham DFT.[86–88] Both the charge density and density of states can provide insight into the electronic structure in addition to serving as promising features to predict a variety of other quantum-chemical properties.[89,90] Partial atomic charges,[91–94] bond orders,[95] and spin densities[91,92] have a wide range of potential use cases, from describing electrostatic interactions in classical simulations of MOFs[26,96] to serving as descriptors to better understand trends in catalytic reactions[97–99] and small-molecule binding.[100] 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 output data from the DFT calculations are made publicly available so that other properties of interest can be readily investigated.

Prior to highlighting how these data can be used in practice, we first investigated several properties of the QMOF database. As shown in Figure 2A, the QMOF database contains MOFs with chemical elements that span nearly the entire periodic table, which is beneficial for the development of transferable ML models. As anticipated, there is also a large number of MOFs in the QMOF database containing Cu, Zn, and Cd, which constitute the three most common types of inorganic nodes in the MOF literature.[75] Nonetheless, we note that some types of MOFs are currently under-represented in the QMOF database due in part to the dataset curation process, which filters out any MOFs with missing atomic coordinates or partial occupancies. These situations are likely to arise in MOFs with complex proton topologies

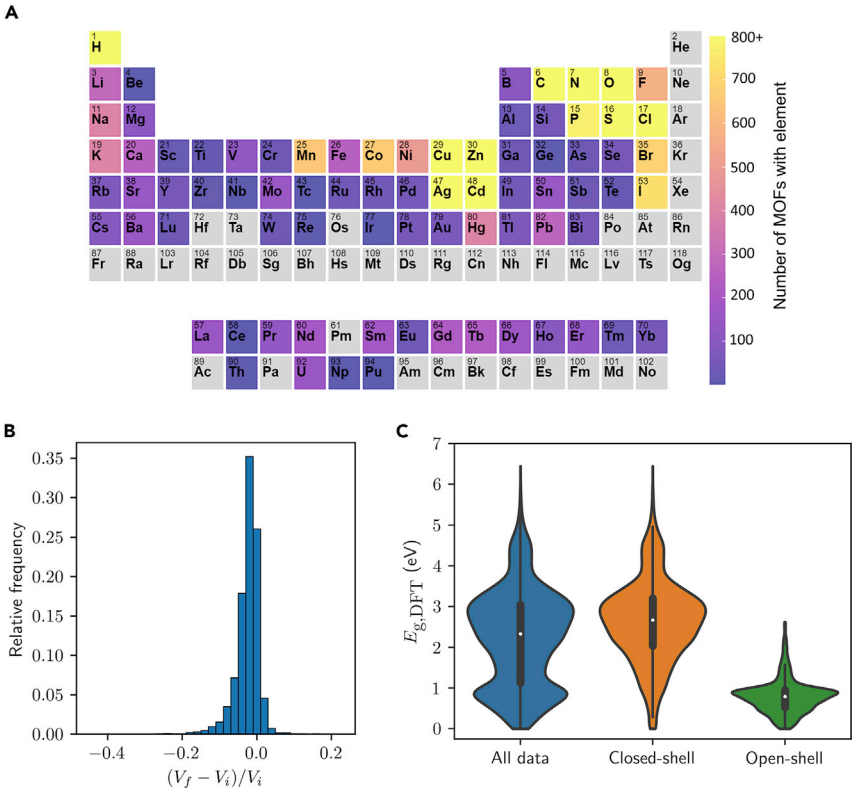


Figure 2. Distribution of properties within the QMOF database

(A) Number of MOFs in the QMOF database containing a given element. All elements that occur in greater than 800 structures are capped at 800 for ease of visualization. These include: C (15,713), H (15,713), N (12,892), O (12,821), Cu (2,882), S (2,684), Zn (2,665), Cd (2,538), Cl (1,687), and Ag (1,213). Elements in gray are not present in any structure.

(B) Histogram of the fractional change in cell volume before (V_i) and after (V_f) structure relaxation at the PBE-D3(BJ) level of theory for the MOFs in the QMOF database.

(C) Violin plots of the DFT-computed band gaps, $E_{g,DFT}$, at the PBE-D3(BJ) level of theory for the MOFs in the QMOF database. Separate distributions are shown for the entire dataset (15,713 entries), the closed-shell MOFs (12,169 entries), and the open-shell MOFs (3,544 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.

that cannot be easily resolved from X-ray diffraction alone (e.g., Zr- and Hf-based MOFs),[101,102] MOFs with defects in the crystal structure,[103] and MOFs that have undergone postsynthetic functionalization.[104,105]

When looking at the geometries before and after structure relaxation, we find that 96.6% of the DFT-optimized MOFs had a change in cell volume of less than 10% (Figure 2B), suggesting that the removal of free solvent does not drastically alter the structural properties for most of the MOFs in this work. In the case of flexible MOFs, multiple conformations are often included in the QMOF database, which is important since they may exhibit different electronic structure properties.[106] As depicted in Figure S3, three distinct conformations of the flexible MOF Fe(bdp) (H_2bdp = 1,4-benzenedipyrazole)[107] are included in the QMOF database (refcodes: QUPZIM, QUPZIM01, QUPZIM02), one of which is on the extreme end of the distribution shown in Figure 2A (see also Table S5). This is not surprising given that high pressures of CH_4 are needed to stabilize the given open-pore configuration of Fe(bdp).[107]

The distribution of DFT-computed band gaps for the fully optimized structures at the PBE-D3(BJ) level of theory 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 QMOF database (Figure 2C), the latter of which have significantly lower band gaps at the PBE-D3(BJ) level of theory on average. With regard to partial atomic charges, a wide spread of values is also obtained (Figure S5A). On comparing the partial atomic charges before and after structure relaxation, we find that 92.4% of the ~1.2 million data points have an absolute difference of less than 0.05 q_e, and 98.9% of the points have an absolute difference of less than 0.1 q_e (Figure S5B). As has been observed on a smaller scale in prior work,[26,69] it can be safely assumed that the partial charges remain essentially unchanged upon structure relaxation in most cases.

As a brief demonstration for how the data generated via the high-throughput DFT workflow could be used directly, we identified any porous framework materials with high-spin Fe species following the high-throughput DFT workflow. 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.[108–110] This query of the QMOF database resulted in six unique MOFs, as shown in Figure S6. Providing validation of this screening approach, two of the six MOFs—Fe_2(dobdc) (H_4dobdc = 2,5-dihydroxybenzene-1,4-dicarboxylic acid) (refcode: COKNOH)[111] and Fe_2(dobpdc) (H_4dobpdc = 4,4′-dihydroxy-(1,1′-biphenyl)-3,3′-dicarboxylic acid) (refcode: MALSIE)[112]—have already been shown to oxidize strong C–H bonds.[108,113,114] Another two of the six MOFs—Fe_2Cl_2(bbta) (H_2bbta = 1*H*,5*H*-benzo(1,2-d:4,5-d′)bistriazole) (refcode: HAYYUE)[115] and Fe_2Cl_2(btdd) (H_2btdd = bis(1*H*-1,2,3-triazolo[4,5-*b*],[4′,5′-*i*])dibenzo[1,4]dioxin) (refcode: HAYZAL)[115]—have been computationally investigated for their use in oxidation reactions.[97,116,117] Prior experimental studies suggest that the aforementioned MOFs exhibit high-spin Fe sites.[108,114,115]

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 with isolated molecules and inorganic solids. This serves two primary purposes. The first is more theoretical: featurization methods (i.e., how each MOF structure is numerically 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[41] and OQMD[28,29] in particular have accelerated this research direction for inorganic solids, and the QM9 dataset[48,118] (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 unoptimized, 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

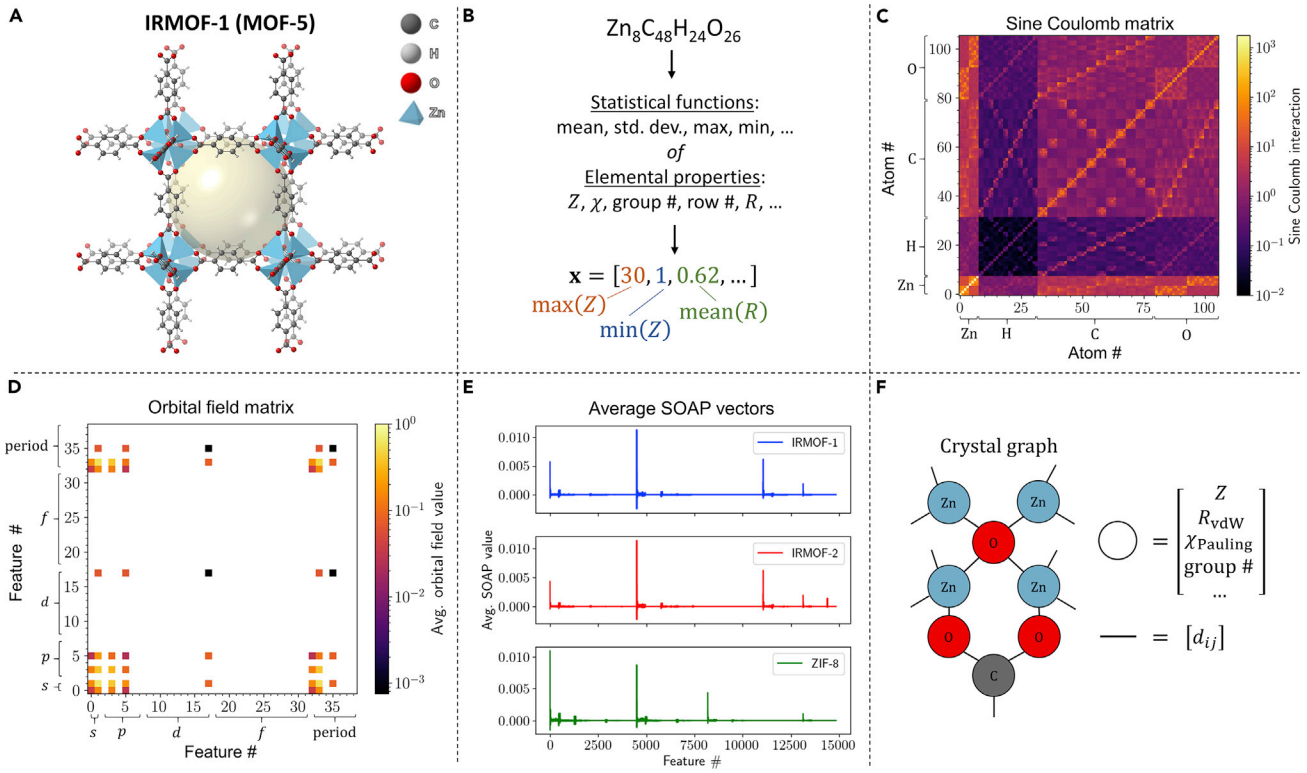


Figure 3. Visualization of various featurization methods applied to IRMOF-1

(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 nonzero values are shown.

(E) Averaged SOAP fingerprint of IRMOF-1 compared with 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).

corresponding unrelaxed experimental structures as the input. Since the development of an ML regression model that can predict the band gaps of MOF crystal structures has not been achieved previously, 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),[119] 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 14,482 materials in the QMOF database ("QMOF-14482") that have gone through the full periodic DFT volume-relaxation process.

The simplest featurization methods considered in this work are the feature sets of He et al.[27] (with 45 statistical attributes of elemental properties, denoted "Stoichiometric-45") and Meredig and Agrawal et al.[120] (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[121] that encodes pairwise electrostatic interactions between nuclei in a material (Figure 3C and Equation S4) and the orbital field matrix[122] that encodes the distribution of valence electrons in each coordination environment of a material (Figure 3D). The smooth overlap of

Table 1. Benchmarking the performance of different machine learning models

ML method	MAE (eV)	R^2	ρ
Constant mean model	0.973	–	–
Sine Coulomb matrix	0.529 ± 0.008	0.643 ± 0.012	0.787 ± 0.008
Stoichiometric-45	0.437 ± 0.004	0.743 ± 0.006	0.842 ± 0.004
Stoichiometric-120	0.433 ± 0.010	0.750 ± 0.009	0.847 ± 0.005
Orbital field matrix	0.417 ± 0.008	0.763 ± 0.010	0.863 ± 0.003
SOAP	0.357 ± 0.008	0.822 ± 0.010	0.910 ± 0.003
CGCNN	0.274 ± 0.008	0.876 ± 0.011	0.932 ± 0.005

Summary of the testing-set mean absolute error (MAE), coefficient of determination (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 QMOF-14482 dataset was used for training. The MAE for a dummy model that predicts the mean band gap (2.220 eV) for all the MOFs is shown for reference.

atomic positions (SOAP)[123,124] is another structure-sensitive descriptor considered in this work, which can be used to compute the similarity between a pair of local atomic environments—and, by extension, a pair of structures—by representing the atoms as Gaussians (i.e., "smoothed positions") and comparing the spatial overlap in the resulting atomic density fields (Figure S2 and Equations S5–S9). In all of the aforementioned examples, these features are used to develop a kernel ridge regression[125] (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),[68] 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 supplemental information.

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 0.43–0.44 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 0.53 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 0.36 eV and $R^2 = 0.82$ 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 S7).

Notably, CGCNN significantly outperforms all the aforementioned KRR models, achieving an MAE of 0.27 eV and $R^2 = 0.88$ (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 0.97 eV, suggesting that CGCNN captures much of the underlying chemistry. The

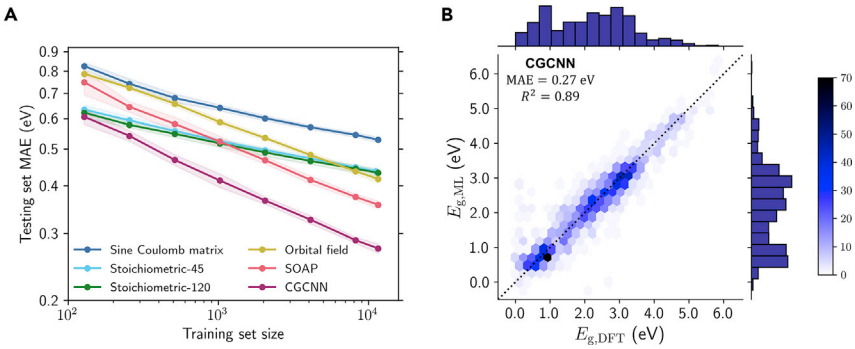


Figure 4. Learning curves and CGCNN parity plot

(A) Mean absolute error (MAE) for band gap predictions on the testing set as a function of training-set size for various machine learning methods. Each point represents the average MAE over five runs with different random seeds for data splitting, and the shaded region represents ±1 standard deviation. The data are shown on log–log axes.

(B) Testing-set parity plot for the CGCNN model with hexagonal binning, comparing the machine learning band gaps, $E_{g,ML}$, with the PBE-D3(BJ) band gaps of the DFT-optimized structures, $E_{g,DFT}$. 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.

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 inorganic solids from the OQMD and Materials Project as well as the organic crystals from the Organic Materials Database (OMDB).[44,68,126,127] It is also worth noting that the experimentally measured band gaps of MOFs can vary by several tenths of an electronvolt depending on the synthesis and post-treatment conditions.[128] As such, an MAE less than 0.3 eV is promising for the identification of structure–property trends and for sorting material candidates by band gap, the latter of which is further justified by the CGCNN's high Spearman rank-order correlation coefficient of ρ = 0.93. For context, it took ~8 min (7 min for a one-time encoding of the crystal graphs and 1 min to evaluate the neural network) on a modern laptop computer to predict the band gaps of all 14,482 MOFs in the QMOF-14482 set using the CGCNN model. In stark contrast, it took over 1.5 million hours (~170 years) of computing time on the Stampede2 supercomputer[129,130] to carry out the structure relaxations and compute the band gaps via DFT.

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 the lowest MAE regardless of training-set size. While SOAP has a worse testing-set MAE than simpler stoichiometric models when trained on fewer than 1,000 MOFs, SOAP has a significantly higher learning rate such that it performs much better for larger training-set sizes (although it still underperforms compared with 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 QMOF-14482 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 quite strong, and this is generally true across the full range of band gap values.

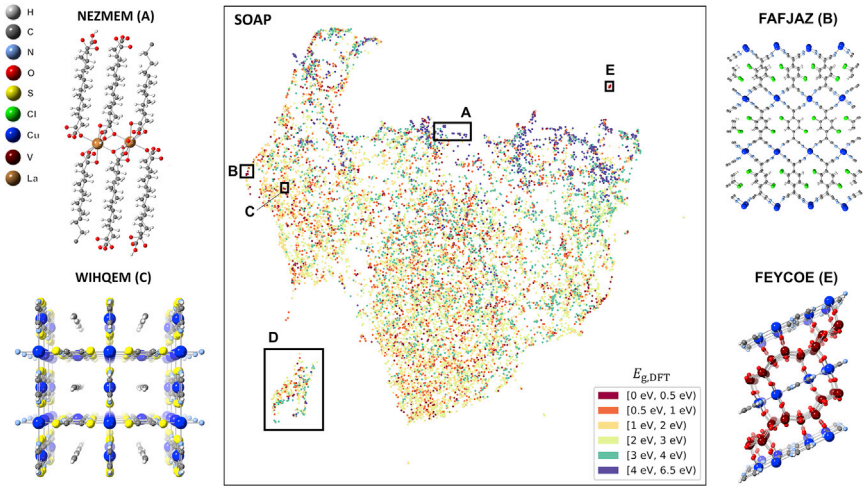


Figure 5. UMAP based on the SOAP average similarity kernel

Unsupervised structural dimensionality reduction performed using UMAP, with a distance matrix obtained from the SOAP average similarity kernel of the unrelaxed structures in the QMOF-14482 dataset. The PBE-D3(BJ) band gaps of the DFT-optimized structures, $E_{g,DFT}$, are overlaid on the UMAP. Selected MOFs in the projection are highlighted.

Dimensionality reduction for structure–property analysis

While the kernel-based methods have a higher MAE than CGCNN when predicting MOF band gaps, the underlying descriptors can still be used for dimensionality reduction, an unsupervised learning task that can cluster structurally similar MOFs in feature space for the purposes of identifying interpretable structure–property relationships. Using the uniform manifold approximation and projection (UMAP) algorithm to carry out the dimensionality reduction,[131,132] 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 gaps can emerge, making it possible to identify otherwise subtle structure–property trends.

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., refcode: NEZMEM),[133] 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., refcodes: BISVUW,[134] FAFJAZ[135]). 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)[136] is in the QMOF-14482 dataset, and it is known to be one of the rare examples of a three-dimensional, porous framework that exhibits room-temperature electrical conductivity.[136] 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)[137] (region C), which has also been studied for its conductive properties.[138,139] In general, we find that 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 S6). Modifications to the SOAP encoding that better account for the discrete building-block nature of MOFs, such as variations

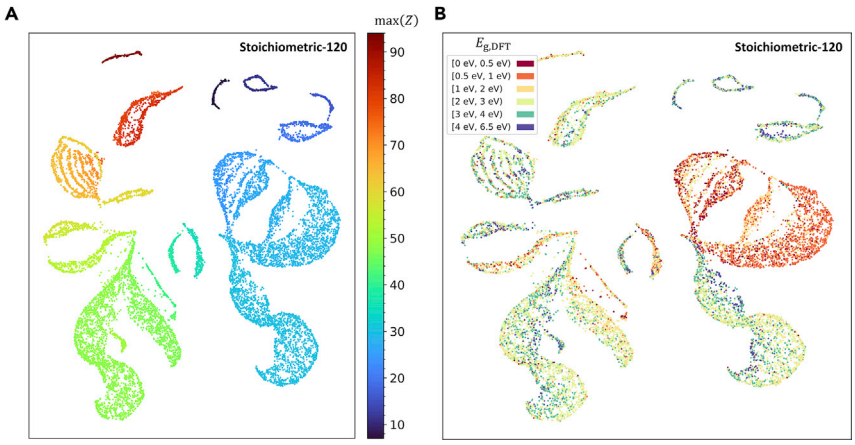


Figure 6. UMAP based on the Stochiometric-120 features

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 QMOF-14482 dataset. The (A) maximum atomic number in each structure, max(Z), and (B) PBE-D3(BJ) band gaps for the corresponding DFT-optimized structures, $E_{g,DFT}$, are overlaid on the UMAPs.

on the recently developed coarse-grained SOAP (cg-SOAP) method,[140] may yield improvements in the future.

Similar to what has been done in prior work with revised autocorrelation functions,[141] we can use the SOAP similarity kernel to understand the diversity of structures in the QMOF-14482 dataset 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., refcodes: MUQCEH,[142] HADMOR[143]) or metal-fluoride species (e.g., refcode: EMEJAJ[144]), which leads to a large difference in the average SOAP fingerprint compared with 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., refcodes: FEYCOE,[145] XEHYEP[146]).

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 largely partitions the MOF chemical space by 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 is one useful way to obtain a global view of the QMOF database. For instance, we find that the QMOF-14482 dataset closely overlaps with both the larger QMOF-42349 dataset it was drawn from and the separate CoRE MOF 2019 database[4] based on the reduced space of Stoichiometric-120 features (Figures S8 and S9). The data in Figure 6B also emphasize how the Zn-containing MOFs (east cluster, max(Z) = 30) tend to have lower band gaps than MOFs with first-row transition metals (south-east cluster, max(Z) = 23–29) at the PBE-D3(BJ) level of theory. To

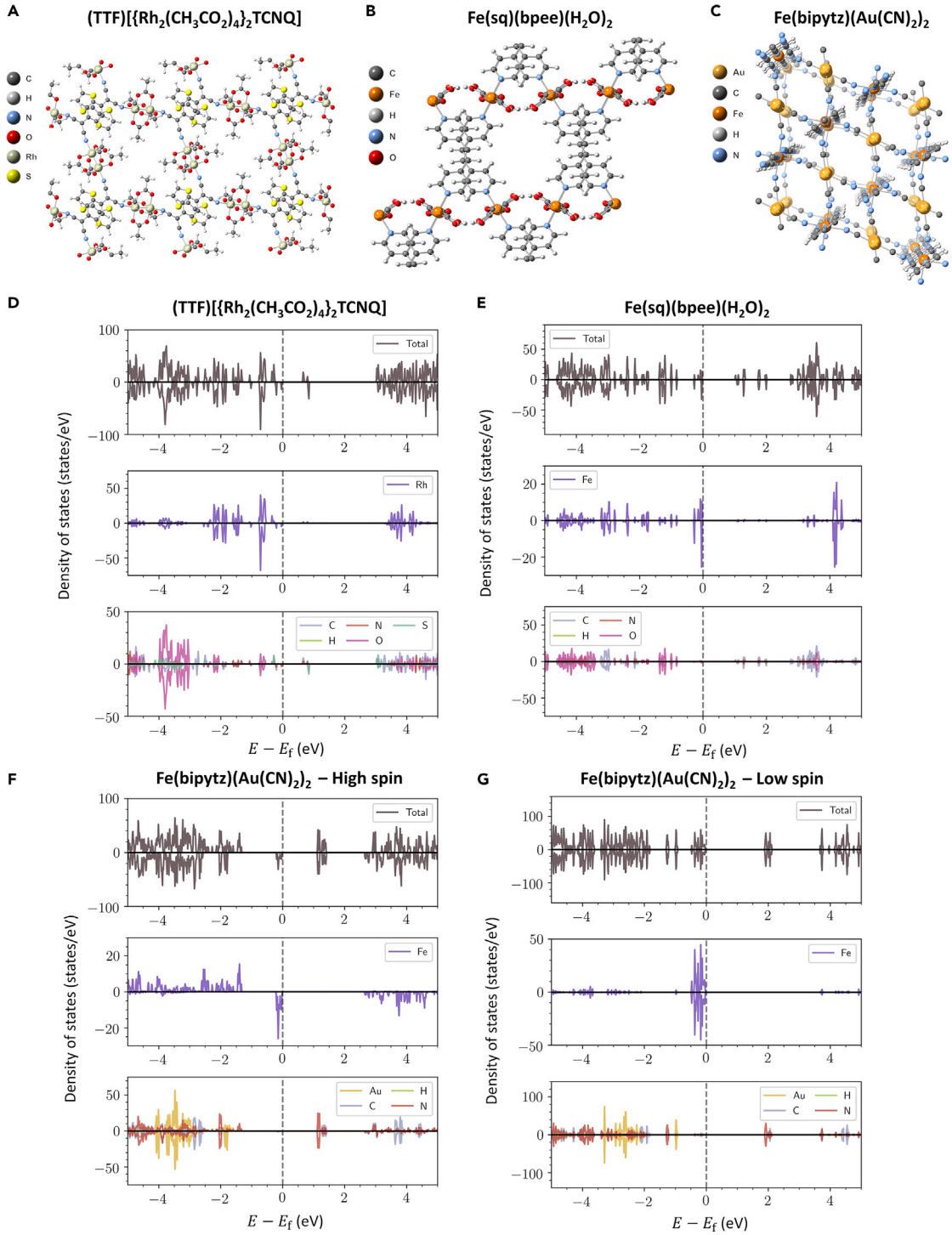


Figure 7. HSE06-D3(BJ) density of states for selected MOFs

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.

enable additional data exploration, interactive versions of the UMAPs are available in the supporting dataset.[62]

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.[147,148] As such, we carried out full structure relaxations and corresponding band gap calculations using the hybrid-level HSE06-D3(BJ) functional[149–151] 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,[128] UiO-66 [UiO = Universitetet i Oslo],[152] ZIF-8 [ZIF = zeolitic imidazolate framework[153]]).[63,147] Generally, lower band gaps are necessary to support electrical conductivity (although it is not the sole factor required for achieving high electrical conductivities).[63]

When the CGCNN model is used to predict the band gaps of all 42,349 structures that compose the QMOF-42349 dataset, one of the lowest band gap materials is predicted to be $Ag(DCl)_2$ (DCl = 2,5-Cl,Cl-*N*,*N*′-dicyanoquinone diamine) (refcode: OTARUX),[154] which is known from experiments to exhibit metallic character via organic radicals that connect the Ag(I) cations.[154] The introduction of radical or redox-active linking units is a well-established strategy to increase the electrical conductivity of framework materials.[63] Although $Ag(DCl)_2$ is arguably best described as a coordination polymer, one notable MOF in the QMOF-42349 dataset with a low predicted band gap and a radical-containing linker is (TTF)[{$Rh_2(CH_3CO_2)_4$}$_2$TCNQ] (TTF = tetrathiafulvalene) (refcode: WAQMEJ)[155]—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, one of the most electronically insulating structures in the QMOF-42349 dataset based on CGCNN-predicted band gap is the nonporous coordination polymer Sr[$C_2H_4(SO_3)_2$] (refcode: GUTYAW),[156] which has an HSE06-D3(BJ) band gap of 8.36 eV (Figure S12).

Consistent with prior experimental work,[157] we also find several Fe-containing materials in the QMOF-42349 dataset with low band gaps, many of which have not yet 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),[158] 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).

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),[159] 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),[159] 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 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 with 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)$(Au(CN)_2)_2$—than those of several Fe-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).[157,160] 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)[62] via a high-throughput periodic DFT workflow.[61] DFT-computed geometries, energies, band gaps, densities of states, partial charges, spin densities, bond orders, and related electronic structure properties are made publicly available.[62] 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 unoptimized MOF crystal structures. A CGCNN[68] 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 SOAP[123,124] and composition-based features[120] 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 in which 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,[126,161] multitask learning,[162] and Δ-ML[163] 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 areas 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[164] or molecular mechanics force fields[165]) with the hopes of achieving high accuracies for MOF structures.

We conclude by noting that the QMOF database should be considered a living resource; several updates to the QMOF database are planned for 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 the QMOF database will accelerate the materials design and discovery process while being specifically tailored for the chemical space of experimentally realized MOF structures.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources and materials should be directed to and will be fulfilled by the lead contact, Andrew S. Rosen (rosen@u.northwestern.edu).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

The landing page for the QMOF database can be found at the following GitHub repository: https://github.com/arosen93/QMOF. Data associated with the QMOF database is hosted via Figshare and has the following permanent DOI: 10.6084/m9.figshare.13147324. All data associated with this work is made publicly available, including results from the DFT calculations, Python scripts to reproduce the ML analyses, code to reproduce the automated DFT screening process, and other related resources.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.matt.2021.02.015.

ACKNOWLEDGMENTS

A.S.R. was supported by a fellowship award through the National Defense Science and Engineering Graduate Fellowship Program, sponsored by the Air Force Research Laboratory (AFRL), the Office of Naval Research, and the Army Research Office. 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 US 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 AFRL (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.

AUTHOR CONTRIBUTIONS

A.S.R.: conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing – original draft, writing – review & editing, visualization, supervision, project administration, funding acquisition. S.M.I.: methodology, software, formal analysis, investigation, data curation, writing – review & editing. D.R.: formal analysis, writing – review & editing. Z.Y.: writing – review & editing. A.A.-G.: writing – review & editing, supervision, funding acquisition. L.G.: writing – review & editing, supervision, funding acquisition. J.M.N.: writing – review & editing, supervision. R.Q.S.: writing – review & editing, supervision, funding acquisition.

DECLARATION OF INTERESTS

R.Q.S. has a financial interest in the start-up company NuMat Technologies, which is seeking to commercialize MOFs. A.A.-G. is a member of the advisory board for *Matter*.

Received: November 11, 2020
Revised: January 16, 2021
Accepted: February 17, 2021
Published: April 5, 2021

REFERENCES

1. Yaghi, O.M., Kalmutzki, M.J., and Diercks, C.S. (2020). Introduction to Reticular Chemistry: Metal-Organic Frameworks and Covalent Organic Frameworks (John Wiley & Sons).

2. Kalmutzki, M.J., Hanikel, N., and Yaghi, O.M. (2018). Secondary building units as the turning point in the development of the reticular chemistry of MOFs. Sci. Adv. *4*, eaat9180.

3. Moghadam, P.Z., Li, A., Wiggin, S.B., Tao, A., Maloney, A.G.P., Wood, P.A., et al. (2017). Development of a Cambridge Structural Database subset: a collection of metal-organic frameworks for past, present, and future. Chem. Mater. *29*, 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., et al. (2019). Advances, updates, and analytics for the computation-ready, experimental metal-organic framework database: CoRE MOF 2019. J. Chem. Eng. Data *64*, 5985–5998.

5. Wilmer, C.E., Leaf, M., Lee, C.Y., Farha, O.K., Hauser, B.G., Hupp, J.T., and Snurr, R.Q. (2012). Large-scale screening of hypothetical metal-organic frameworks. Nat. Chem. *4*, 83–89.

6. Colón, Y.J., Gómez-Gualdrón, D.A., and Snurr, R.Q. (2017). Topologically guided, automated construction of metal-organic frameworks and their evaluation for energy-related applications. Cryst. Growth Des. *17*, 5801–5810.

7. Boyd, P.G., and Woo, T.K. (2016). A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. CrystEngComm *18*, 3777–3792.

8. Ejsmont, A., Andreo, J., Lanza, A., Galarda, A., Macreadie, L., Wuttke, S., Canossa, S., Ploetz, E., and Goscianska, J. (2020). Applications of reticular diversity in metal-organic frameworks: an ever-evolving state of the art. Coord. Chem. Rev. *430*, 213655.

9. Colón, Y.J., and Snurr, R.Q. (2014). High-throughput computational screening of metal-organic frameworks. Chem. Soc. Rev. *43*, 5735–5749.

10. Jablonka, K.M., Ongari, D., Moosavi, S.M., and Smit, B. (2020). Big-data science in porous materials: materials genomics and machine learning. Chem. Rev. *120*, 8066–8129.

11. Chibani, S., and Coudert, F.-X. (2020). Machine learning approaches for the prediction of materials properties. APL Mater. *8*, 80701.

12. Anderson, G., Schweitzer, B., Anderson, R., and Gomez-Gualdron, D.A. (2018). Attainable volumetric targets for adsorption-based hydrogen storage in porous crystals: molecular simulation and machine learning. J. Phys. Chem. C *123*, 120–130.

13. Bucior, B.J., Bobbitt, N.S., Islamoglu, T., Goswami, S., Gopalan, A., Yildirim, T., Farha, O.K., Bagheri, N., and Snurr, R.Q. (2019). Energy-based descriptors to rapidly predict hydrogen storage in metal-organic frameworks. Mol. Syst. Des. Eng. *4*, 162–174.

14. 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., and Smit, B. (2017). Materials Genome in action: identifying the performance limits of physical hydrogen storage. Chem. Mater. *29*, 2844–2854.

15. Anderson, R., Rodgers, J., Argueta, E., Biong, A., and Gomez-Gualdron, D.A. (2018). Role of pore chemistry and topology in the CO_2 capture capabilities of MOFs: from molecular simulation to machine learning. Chem. Mater. *30*, 6325–6337.

16. Dureckova, H., Krykunov, M., Aghaji, M.Z., and Woo, T.K. (2019). 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 *123*, 4133–4139.

17. Yao, Z., Sanchez-Lengeling, B., Bobbitt, N.S., Bucior, B.J., Kumar, S.G.H., Collins, S.P., et al. (2021). Inverse design of nanoporous

crystalline reticular materials with deep generative models. Nat. Mach. Intell. 76–86.

18. Moghadam, P.Z., Rogge, S.M.J., Li, A., Chow, C.-M., Wieme, J., Moharrami, N., Aragones-Anglada, M., Conduit, G., Gomez-Gualdron, D.A., Van Speybroeck, V., and Fairen-Jimenez, D. (2019). Structure-mechanical stability relations of metal-organic frameworks via machine learning. Matter *1*, 219–234.

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

20. Shi, Z., Yang, W., Deng, X., Cai, C., Yan, Y., Liang, H., Liu, Z., and Qiao, Z. (2020). Machine-learning-assisted high-throughput computational screening of high performance metal-organic frameworks. Mol. Syst. Des. Eng. *5*, 725–742.

21. Chong, S., Lee, S., Kim, B., and Kim, J. (2020). Applications of machine learning in metal-organic frameworks. Coord. Chem. Rev. *423*, 213487.

22. Mancuso, J.L., Mroz, A.M., Le, K.N., and Hendon, C.H. (2020). Electronic structure modeling of metal-organic frameworks. Chem. Rev. *120*, 8641–8715.

23. Raza, A., Sturluson, A., Simon, C., and Fern, X. (2020). Message passing neural networks for partial charge assignment to metal-organic frameworks. J. Phys. Chem. C *124*, 19070–19082.

24. Korolev, V.V., Mitrofanov, A., Marchenko, E.I., Eremin, N.N., Tkachenko, V., and Kalmykov, S.N. (2020). Transferable and extensible machine learning derived atomic charges for modeling hybrid nanoporous materials. Chem. Mater. *32*, 7822–7831.

25. Chung, Y.G., Camp, J., Haranczyk, M., Sikora, B.J., Bury, W., Krungleviciute, V., Yildirim, T., Farha, O.K., Sholl, D.S., and Snurr, R.Q. (2014). Computation-ready, experimental metal-organic frameworks: a tool to enable high-throughput screening of nanoporous crystals. Chem. Mater. *26*, 6185–6192.

26. Nazarian, D., Camp, J.S., and Sholl, D.S. (2016). A comprehensive set of high-quality point charges for simulations of metal-organic frameworks. Chem. Mater. *28*, 785–793.

27. He, Y., Cubuk, E.D., Allendorf, M.D., and Reed, E.J. (2018). Metallic metal-organic frameworks predicted by the combination of machine learning methods and ab initio calculations. J. Phys. Chem. Lett. *9*, 4562–4569.

28. Saal, J.E., Kirklin, S., Aykol, M., Meredig, B., and Wolverton, C. (2013). Materials design and discovery with high-throughput density functional theory: the Open Quantum Materials Database (OQMD). JOM *65*, 1501–1509.

29. Kirklin, S., Saal, J.E., Meredig, B., Thompson, A., Doak, J.W., Aykol, M., et al. (2015). The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. npj Comput. Mater. *1*, 15010.

30. Perdew, J.P., Burke, K., and Ernzerhof, M. (1996). Generalized gradient approximation made simple. Phys. Rev. Lett. *77*, 3865–3868.

31. Abrahams, B.F., Hardie, M.J., Hoskins, B.F., Robson, R., and Williams, G.A. (1992). 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. *114*, 10641–10643.

32. von Lilienfeld, O.A., Müller, K.-R., and Tkatchenko, A. (2020). Exploring chemical compound space with quantum-based machine learning. Nat. Rev. Chem. *4*, 347–358.

33. Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., and Walsh, A. (2018). Machine learning for molecular and materials science. Nature *559*, 547–555.

34. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., and Kim, C. (2017). Machine learning in materials informatics: recent applications and prospects. npj Comput. Mater. *3*, 54.

35. Ward, L., and Wolverton, C. (2017). Atomistic calculations and materials informatics: a review. Curr. Opin. Solid State Mater. Sci. *21*, 167–176.

36. Chen, C., Zuo, Y., Ye, W., Li, X., Deng, Z., and Ong, S.P. (2020). A critical review of machine learning of energy materials. Adv. Energy Mater. *10*, 1903242.

37. Morgan, D., and Jacobs, R. (2020). Opportunities and challenges for machine learning in materials science. Annu. Rev. Mater. Res. *50*, 71–103.

38. Saal, J.E., Oliynyk, A.O., and Meredig, B. (2020). Machine learning in materials discovery: confirmed predictions and their underlying approaches. Annu. Rev. Mater. Res. *50*, 49–69.

39. Suh, C., Fare, C., Warren, J.A., and Pyzer-Knapp, E.O. (2020). Evolving the materials Genome: how machine learning is fueling the next generation of materials discovery. Annu. Rev. Mater. Res. *50*, 1–25.

40. Jain, A., Hautier, G., Moore, C.J., Ping Ong, S., Fischer, C.C., Mueller, T., Persson, K.A., and Ceder, G. (2011). A high-throughput infrastructure for density functional theory calculations. Comput. Mater. Sci. *50*, 2295–2310.

41. Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., and Persson, K.A. (2013). The Materials Project: a materials Genome approach to accelerating materials innovation. APL Mater. *1*, 11002.

42. 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., et al. (2012). AFLOWLIB.ORG: a distributed materials properties repository from high-throughput ab initio calculations. Comput. Mater. Sci. *58*, 227–235.

43. Winther, K.T., Hoffmann, M.J., Boes, J.R., Mamun, O., Bajdich, M., and Bligaard, T. (2019). Catalysis-hub.org, an open electronic structure database for surface reactions. Sci. Data *6*, 75.

44. Borysov, S.S., Geilhufe, R.M., and Balatsky, A.V. (2017). Organic Materials Database: an open-access online database for data mining. PLoS One *12*, e0171501.

45. Landis, D.D., Hummelshoj, J.S., Nestorov, S., Greeley, J., Dulak, M., Bligaard, T., Nørskov, J.K., and Jacobsen, K.W. (2012). The computational materials repository. Comput. Sci. Eng. *14*, 51–57.

46. Draxl, C., and Scheffler, M. (2018). NOMAD: the FAIR concept for big data-driven materials science. MRS Bull. *43*, 676–682.

47. Chanussot, L., Das, A., Goyal, S., Lavril, T., Shuaibi, M., Riviere, M., et al. (2020). The Open Catalyst 2020 (OC20) dataset and community challenges. arXiv, 2010.09990.

48. Ramakrishnan, R., Dral, P.O., Rupp, M., and Von Lilienfeld, O.A. (2014). Quantum chemistry structures and properties of 134 kilo molecules. Sci. Data *1*, 140022.

49. Montavon, G., Rupp, M., Gobre, V., Vazquez-Mayagoitia, A., Hansen, K., Tkatchenko, A., Müller, K.-R., and Von Lilienfeld, O.A. (2013). Machine learning of molecular electronic properties in chemical compound space. N. J. Phys. *15*, 95003.

50. Smith, J.S., Isayev, O., and Roitberg, A.E. (2017). ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. Sci. Data *4*, 170193.

51. Blau, S., Spotte-Smith, E., Wood, B., Dwaraknath, S., and Persson, K. (2020). Accurate, automated density functional theory for complex molecules using on-the-fly error correction. ChemRxiv. https://doi.org/10.26434/chemrxiv.13076030.

52. Balcells, D., and Skjelstad, B.B. (2020). The tmQM dataset—quantum geometries and properties of 86k transition metal complexes. J. Chem. Inf. Model. *60*, 6135–6146.

53. 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., et al. (2016). Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. Nat. Mater. *15*, 1120–1127.

54. Mansouri Tehrani, A., Oliynyk, A.O., Parry, M., Rizvi, Z., Couper, S., Lin, F., Miyagi, L., Sparks, T.D., and Brgoch, J. (2018). Machine learning directed search for ultraincompressible, superhard materials. J. Am. Chem. Soc. *140*, 9844–9853.

55. Wu, S., Kondo, Y., Kakimoto, M., Yang, B., Yamada, H., Kuwajima, I., et al. (2019). Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. npj Comput. Mater. *5*, 66.

56. Lyu, H., Ji, Z., Wuttke, S., and Yaghi, O.M. (2020). Digital reticular chemistry. Chem *6*, 2219–2241.

57. Flores-Leonar, M.M., Mejía-Mendoza, L.M., Aguilar-Granda, A., Sanchez-Lengeling, B., Tribukait, H., Amador-Bedolla, C., and Aspuru-Guzik, A. (2020). Materials acceleration platforms: on the way to

autonomous experimentation. Curr. Opin. Green. Sustain. Chem. *25*, 100370.

58. Tabor, D.P., Roch, L.C., Saikin, S.K., Kreisbeck, C., Sheberla, D., Montoya, J.H., Dwaraknath, S., Aykol, M., Ortiz, C., Tribukait, H., et al. (2018). Accelerating the discovery of materials for clean energy in the era of smart automation. Nat. Rev. Mater. *3*, 5–20.

59. Coley, C.W., Eyke, N.S., and Jensen, K.F. (2020). Autonomous discovery in the chemical sciences part I: progress. Angew. Chem. Int. Ed. *59*, 22858–22893.

60. Coley, C.W., Eyke, N.S., and Jensen, K.F. (2020). Autonomous discovery in the chemical sciences part II: outlook. Angew. Chem. Int. Ed. *59*, 23414–23436.

61. Rosen, A.S., Notestein, J.M., and Snurr, R.Q. (2019). Identifying promising metal-organic frameworks for heterogeneous catalysis via high-throughput periodic density functional theory. J. Comput. Chem. *40*, 1305–1318.

62. QMOF Database. https://dx.doi.org/10.6084/m9.figshare.13147324.

63. Xie, L.S., Skorupskii, G., and Dinca, M. (2020). Electrically conductive metal-organic frameworks. Chem. Rev. *120*, 8536–8580.

64. Baumann, A.E., Burns, D.A., Liu, B., and Thoi, V.S. (2019). Metal-organic framework functionalization and design strategies for advanced electrochemical energy storage devices. Commun. Chem. *86*, https://doi.org/10.1038/s42004-019-0184-6.

65. D'Alessandro, D.M. (2016). Exploiting redox activity in metal-organic frameworks: concepts, trends and perspectives. Chem. Commun. *52*, 8957–8971.

66. Downes, C.A., and Marinescu, S.C. (2017). Electrocatalytic metal-organic frameworks for energy applications. ChemSusChem *10*, 4374–4392.

67. Allendorf, M.D., Dong, R., Feng, X., Kaskel, S., Matoga, D., and Stavila, V. (2020). Electronic devices using open framework materials. Chem. Rev. *120*, 8581–8640.

68. Xie, T., and Grossman, J.C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys. Rev. Lett. *120*, 145301.

69. Nazarian, D., Camp, J.S., Chung, Y.G., Snurr, R.Q., and Sholl, D.S. (2017). Large-scale refinement of metal-organic framework structures using density functional theory. Chem. Mater. *29*, 2521–2528.

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

71. Zarabadi-Poor, P., and Marek, R. (2019). Comment on "Database for CO_2 separation performances of MOFs based on computational materials screening.". ACS Appl. Mater. Interfaces *11*, 16261–16265.

72. Barthel, S., Alexandrov, E.V., Proserpio, D.M., and Smit, B. (2018). Distinguishing metal-organic frameworks. Cryst. Growth Des. *18*, 1738–1747.

73. Altintas, C., Avci, G., Daglar, H., Azar, A.N.V., Erucar, I., Velioglu, S., and Keskin, S. (2019). 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 *7*, 9593–9608.

74. Velioglu, S., and Keskin, S. (2020). Revealing the effect of structure curations on the simulated CO_2 separation performances of MOFs. Mater. Adv. *1*, 341–353.

75. 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., and Snurr, R.Q. (2019). Identification schemes for metal-organic frameworks to enable rapid search and cheminformatics analysis. Cryst. Growth Des. *19*, 6682–6697.

76. Chen, T., and Manz, T.A. (2020). Identifying misbonded atoms in the 2019 CoRE metal-organic framework database. RSC Adv. *10*, 26944–26951.

77. Grimme, S., Antony, J., Ehrlich, S., and Krieg, H. (2010). A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H–Pu. J. Chem. Phys. *132*, 154104.

78. Grimme, S., Ehrlich, S., and Goerigk, L. (2011). Effect of the damping function in dispersion corrected density functional theory. J. Comput. Chem. *32*, 1456–1465.

79. Kresse, G., and Furthmüller, J. (1996). Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. Phys. Rev. B *54*, 11169–11186.

80. Kresse, G., and Joubert, D. (1999). From ultrasoft pseudopotentials to the projector augmented-wave method. Phys. Rev. B *59*, 1758–1775.

81. Hendon, C.H., Tiana, D., and Walsh, A. (2012). Conductive metal-organic frameworks and networks: fact or fantasy? Phys. Chem. Chem. Phys. *14*, 13120–13132.

82. Sun, L., Campbell, M.G., and Dincă, M. (2016). Electrically conductive porous metal-organic frameworks. Angew. Chem. Int. Ed. *55*, 3566–3579.

83. Singh, A.K., Montoya, J.H., Gregoire, J.M., and Persson, K.A. (2019). Robust and synthesizable photocatalysts for CO_2 reduction: a data-driven materials discovery. Nat. Commun. *10*, 443.

84. Yang, L.-M., Ravindran, P., Vajeeston, P., Svelle, S., and Tilset, M. (2013). A quantum mechanically guided view of Cd-MOF-5 from formation energy, chemical bonding, electronic structure, and optical properties. Microporous Mesoporous Mater. *175*, 50–58.

85. Gong, S., Xie, T., Zhu, T., Wang, S., Fadel, E.R., Li, Y., and Grossman, J.C. (2019). Predicting charge density distribution of materials using a local-environment-based graph convolutional network. Phys. Rev. B *100*, 184103.

86. Grisafi, A., Fabrizio, A., Meyer, B., Wilkins, D.M., Corminboeuf, C., and Ceriotti, M. (2018). Transferable machine-learning model of the electron density. ACS Cent. Sci. *5*, 57–64.

87. Chandrasekaran, A., Kamal, D., Batra, R., Kim, C., Chen, L., and Ramprasad, R. (2019). Solving the electronic structure problem with machine learning. npj Comput. Mater.22 https://doi.org/10.1038/s41524-019-0162-7.

88. Kamal, D., Chandrasekaran, A., Batra, R., and Ramprasad, R. (2020). A charge density prediction model for hydrocarbons using deep neural networks. Mach. Learn. Sci. Technol. *1*, 25003.

89. Kolb, B., Lentz, L.C., and Kolpak, A.M. (2017). Discovering charge density functionals and structure-property relationships with PROPhet: a general framework for coupling machine learning and first-principles methods. Sci. Rep. *7*, 1192.

90. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S., and Ramprasad, R. (2013). Accelerating materials property predictions using machine learning. Sci. Rep. *3*, 2810.

91. Manz, T.A., and Limas, N.G. (2016). Introducing DDEC6 atomic population analysis: part 1. Charge partitioning theory and methodology. RSC Adv. *6*, 47771–47801.

92. Limas, N.G., and Manz, T.A. (2016). Introducing DDEC6 atomic population analysis: part 2. Computed results for a wide range of periodic and nonperiodic materials. RSC Adv. *6*, 45727–45747.

93. Limas, N.G., and Manz, T.A. (2018). Introducing DDEC6 atomic population analysis: part 4. Efficient parallel computation of net atomic charges, atomic spin moments, bond orders, and more. RSC Adv. *8*, 2678–2707.

94. Marenich, A.V., Jerome, S.V., Cramer, C.J., and Truhlar, D.G. (2012). Charge Model 5: an extension of hirshfeld population analysis for the accurate description of molecular interactions in gaseous and condensed phases. J. Chem. Theor. Comput. *8*, 527–541.

95. Manz, T.A. (2017). Introducing DDEC6 atomic population analysis: part 3. Comprehensive method to compute bond orders. RSC Adv. *7*, 45552–45581.

96. Haldoupis, E., Nair, S., and Sholl, D.S. (2012). Finding MOFs for highly selective CO_2/N_2 adsorption using materials screening based on efficient assignment of atomic point charges. J. Am. Chem. Soc. *134*, 4313–4323.

97. Rosen, A.S., Notestein, J.M., and Snurr, R.Q. (2019). Structure-activity relationships that identify Metal-organic framework catalysts for methane activation. ACS Catal. *9*, 3576–3587.

98. Yang, B., Wu, X.-P., Gagliardi, L., and Truhlar, D.G. (2019). Methane functionalization by an Ir(III) catalyst supported on a metal-organic framework: an alternative explanation of steric confinement effects. Theor. Chem. Acc. *138*, 107.

99. Sours, T., Patel, A., Nørskov, J., Siahrostami, S., and Kulkarni, A. (2020). Circumventing scaling relations in oxygen electrochemistry using metal-organic frameworks. J. Phys. Chem. Lett. *11*, 10029–10036.

100. Rosen, A.S., Mian, M.R., Islamoglu, T., Chen, H., Farha, O.K., Notestein, J.M., and Snurr, R.Q. (2020). 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. *142*, 4317–4328.

101. Planas, N., Mondloch, J.E., Tussupbayev, S., Borycz, J., Gagliardi, L., Hupp, J.T., Farha, O.K., and Cramer, C.J. (2014). Defining the proton topology of the Zr_6-based metal-organic framework NU-1000. J. Phys. Chem. Lett. *5*, 3716–3723.

102. Klet, R.C., Liu, Y., Wang, T.C., Hupp, J.T., and Farha, O.K. (2016). Evaluation of Brønsted acidity and proton topology in Zr- and Hf-based metal-organic frameworks using potentiometric acid-base titration. J. Mater. Chem. A *4*, 1479–1485.

103. Ren, J., Ledwaba, M., Musyoka, N.M., Langmi, H.W., Mathe, M., Liao, S., and Pang, W. (2017). Structural defects in metal-organic frameworks (MOFs): formation, detection and control towards practices of interests. Coord. Chem. Rev. *349*, 169–197.

104. Deria, P., Mondloch, J.E., Karagiaridi, O., Bury, W., Hupp, J.T., and Farha, O.K. (2014). Beyond post-synthesis modification: evolution of metal-organic frameworks via building block replacement. Chem. Soc. Rev. *43*, 5896–5912.

105. Syed, Z.H., Sha, F., Zhang, X., Kaphan, D.M., Delferro, M., and Farha, O.K. (2020). Metal-organic framework nodes as a supporting platform for tailoring the activity of metal catalysts. ACS Catal. *10*, 11556–11566.

106. Ling, S., and Slater, B. (2015). Unusually large band gap changes in breathing metal-organic framework materials. J. Phys. Chem. C *119*, 16667–16677.

107. 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., et al. (2015). Methane storage in flexible metal-organic frameworks with intrinsic thermal management. Nature *527*, 357–361.

108. 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., et al. (2014). Oxidation of ethane to ethanol by N_2O in a metal-organic framework with coordinatively unsaturated iron(II) sites. Nat. Chem. *6*, 590–595.

109. Vogiatzis, K.D., Haldoupis, E., Xiao, D.J., Long, J.R., Siepmann, J.I., and Gagliardi, L. (2016). Accelerated computational analysis of metal-organic frameworks for oxidation catalysis. J. Phys. Chem. C *120*, 18707–18712.

110. 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., et al. (2018). Isolated Fe sites in metal-organic framework catalyze the direct conversion of methane to methanol. ACS Catal. *8*, 5542–5548.

111. 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., et al. (2014). Comprehensive study of carbon dioxide adsorption in the metal-organic frameworks M_2(dobdc) (M = Mg, Mn, Fe, Co, Ni, Cu, Zn). Chem. Sci. *5*, 4569–4581.

112. Gygi, D., Bloch, E.D., Mason, J.A., Hudson, M.R., Gonzalez, M.I., Siegelman, R.L., et al. (2016). Hydrogen storage in the expanded pore metal-organic frameworks M_2(dobpdc) (M = Mg, Mn, Fe, Co, Ni, Zn). Chem. Mater. *28*, 1128–1138.

113. Verma, P., Vogiatzis, K.D., Planas, N., Borycz, J., Xiao, D.J., Long, J.R., et al. (2015). Mechanism of oxidation of ethane to ethanol at iron(IV)-Oxo sites in magnesium-diluted Fe_2(dobdc). J. Am. Chem. Soc. *137*, 5770–5781.

114. Xiao, D.J., Oktawiec, J., Milner, P.J., and Long, J.R. (2016). Pore environment effects on catalytic cyclohexane oxidation in expanded Fe_2(dobdc) analogues. J. Am. Chem. Soc. *138*, 14371–14379.

115. Reed, D.A., Keitz, B.K., Oktawiec, J., Mason, J.A., Runčevski, T., Xiao, D.J., Darago, L.E., Crocellà, V., Bordiga, S., and Long, J.R. (2017). A spin transition mechanism for cooperative adsorption in metal-organic frameworks. Nature *550*, 96–100.

116. Rosen, A.S., Notestein, J.M., and Snurr, R.Q. (2020). 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. *132*, 19662–19670.

117. Rosen, A.S., Notestein, J.M., and Snurr, R.Q. (2020). Comparing GGA, GGA+*U*, and meta-GGA functionals for redox-dependent binding at open metal sites in Metal-organic frameworks. J. Chem. Phys. *152*, 224101.

118. Ruddigkeit, L., Van Deursen, R., Blum, L.C., and Reymond, J.-L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. J. Chem. Inf. Model. *52*, 2864–2875.

119. Li, H., Eddaoudi, M., O'Keeffe, M., and Yaghi, O.M. (1999). Design and synthesis of an exceptionally stable and highly porous metal-organic framework. Nature *402*, 276–279.

120. Meredig, B., Agrawal, A., Kirklin, S., Saal, J.E., Doak, J.W., Thompson, A., Zhang, K., Choudhary, A., and Wolverton, C. (2014). Combinatorial screening for new materials in unconstrained composition space with machine learning. Phys. Rev. B *89*, 94104.

121. Faber, F., Lindmaa, A., von Lilienfeld, O.A., and Armiento, R. (2015). Crystal structure representations for machine learning models of formation energies. Int. J. Quant. Chem. *115*, 1094–1101.

122. Lam Pham, T., Kino, H., Terakura, K., Miyake, T., Tsuda, K., Takigawa, I., and Chi Dam, H. (2017). Machine learning reveals orbital interaction in materials. Sci. Technol. Adv. Mater. *18*, 756–765.

123. Bartók, A.P., Kondor, R., and Csányi, G. (2013). On representing chemical environments. Phys. Rev. B *87*, 184115.

124. De, S., Bartók, A.P., Csányi, G., and Ceriotti, M. (2016). Comparing molecules and solids across structural and alchemical space. Phys. Chem. Chem. Phys. *18*, 13754–13769.

125. Pronobis, W., and Müller, K.-R.. Kernel methods for quantum chemistry. In Machine Learning Meets Quantum Physics; Schütt K.T., Chmiela S., von Lilienfeld O.A., Tkatchenko A., Tsuda K., Müller K.-R., Springer; pp 25–36.

126. Chen, C., Ye, W., Zuo, Y., Zheng, C., and Ong, S.P. (2019). Graph networks as a universal machine learning framework for molecules and crystals. Chem. Mater. *31*, 3564–3572.

127. Olsthoorn, B., Geilhufe, R.M., Borysov, S.S., and Balatsky, A.V. (2019). Band gap prediction for large organic crystal structures with machine learning. Adv. Quan. Technol. *2*, 1900023.

128. Gascon, J., Hernández-Alonso, M.D., Almeida, A.R., van Klink, G.P.M., Kapteijn, F., and Mul, G. (2008). Isoreticular MOFs as efficient photocatalysts with tunable band gap: an operando FTIR study of the photoinduced oxidation of propylene. ChemSusChem *1*, 981–983.

129. Stanzione, D., Barth, B., Gaffney, N., Gaither, K., Hempel, C., Minyard, T., Mehringer, S., Wernert, E., Tufo, H., Panda, D., and Teller, P. (2017). 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, D. Hart, ed. (Association for Computing Machinery). https://doi.org/10.1145/3093338.3093385.

130. Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., Hazlewood, V., Lathrop, S., Lifka, D., Peterson, G.D., et al. (2014). XSEDE: accelerating scientific discovery. Comput. Sci. Eng. *16*, 62–74.

131. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. arXiv, 1802.03426.

132. Leland, M., John, H., Nathaniel, S., and Lukas, G. (2018). UMAP: uniform manifold approximation and projection. J. Open Source Softw. *3*, 861.

133. Zeng, X.-Z., Zhang, A.-Y., Bu, D., and Li, Y.-W. (2013). Hydrothermal synthesis, structure and thermal properties of a novel three-dimensional La(III)-Sebacate framework. Chin. J. Struct. Chem. *32*, 120–124.

134. Zhang, Z., Zhao, H., Matsushita, M.M., Awaga, K., and Dunbar, K.R. (2014). A new metal-organic hybrid material with intrinsic resistance-based bistability: monitoring in situ room temperature switching behavior. J. Mater. Chem. C *2*, 399–404.

135. Lopez, N., Zhao, H., Ota, A., Prosvirin, A.V., Reinheimer, E.W., and Dunbar, K.R. (2010). Unprecedented binary semiconductors based on TCNQ: single-crystal X-ray studies and physical properties of $Cu(TCNQX_2)$ X = Cl, Br. Adv. Mater. *22*, 986–989.

136. Aubrey, M.L., Kapelewski, M.T., Melville, J.F., Oktawiec, J., Presti, D., Gagliardi, L., and Long, J.R. (2019). Chemiresistive detection of gaseous hydrocarbons and interrogation of charge transport in Cu[Ni(2,3-pyrazinedithiolate)$_2$] by gas adsorption. J. Am. Chem. Soc. *141*, 5005–5013.

137. Peng, Y.-L., Pham, T., Li, P., Wang, T., Chen, Y., Chen, K.-J., Forrest, K.A., Space, B., Cheng, P., Zaworotko, M.J., and Zhang, Z. (2018). Robust ultramicroporous metal-organic frameworks with benchmark affinity for acetylene. Angew. Chem. Int. Ed. *57*, 10971–10975.

138. Takaishi, S., Hosoda, M., Kajiwara, T., Miyasaka, H., Yamashita, M., Nakanishi, Y., et al. (2009). Electroconductive porous coordination polymer Cu[Cu(pdt)$_2$] composed of donor and acceptor building units. Inorg. Chem. *48*, 9048–9050.

139. Kobayashi, Y., Jacobs, B., Allendorf, M.D., and Long, J.R. (2010). Conductivity, doping, and redox chemistry of a microporous dithiolene-based metal-organic framework. Chem. Mater. *22*, 4120–4122.

140. Nicholas, T.C., Goodwin, A., and Deringer, V.L. (2020). Understanding the geometric diversity of inorganic and hybrid frameworks through structural coarse-graining. Chem. Sci. *11*, 12580–12587.

141. Moosavi, S.M., Nandy, A., Jablonka, K.M., Ongari, D., Janet, J.P., Boyd, P.G., Lee, Y., Smit, B., and Kulik, H.J. (2020). Understanding the diversity of the metal-organic framework ecosystem. Nat. Commun. *11*, 4068.

142. Hulvey, Z., Furman, J.D., Turner, S.A., Tang, M., and Cheetham, A.K. (2010). Dimensionality trends in metal-organic frameworks containing perfluorinated or nonfluorinated benzenedicarboxylates. Cryst. Growth Des. *10*, 2041–2043.

143. Taylor, M.K., Runcevski, T., Oktawiec, J., Gonzalez, M.I., Siegelman, R.L., Mason, J.A., Ye, J., Brown, C.M., and Long, J.R. (2016). Tuning the adsorption-induced phase change in the flexible metal-organic framework Co(bdp). J. Am. Chem. Soc. *138*, 15019–15026.

144. Cui, X., Chen, K., Xing, H., Yang, Q., Krishna, R., Bao, Z., Wu, H., Zhou, W., Dong, X., Han, Y., et al. (2016). Pore chemistry and size control in hybrid porous materials for acetylene capture from ethylene. Science *353*, 141–144.

145. Li, S., Zhang, L., Lu, B., Yan, E., Wang, T., Li, L., Wang, J., Yu, Y., and Mu, Q. (2018). A new polyoxovanadate-based metal-organic framework: synthesis, structure and photo-/electro-catalytic properties. New J. Chem. *42*, 7247–7253.

146. Yan, B., Luo, J., Dube, P., Sefat, A.S., Greedan, J.E., and Maggard, P.A. (2006). Spin-gap formation and thermal structural studies in reduced hybrid layered vanadates. Inorg. Chem. *45*, 5109–5118.

147. Choudhuri, I., and Truhlar, D.G. (2019). HLE17: an efficient way to predict band gaps of complex materials. J. Phys. Chem. C *123*, 17416–17424.

148. Borlido, P., Aull, T., Huran, A.W., Tran, F., Marques, M.A.L., and Botti, S. (2019). Large-scale benchmark of exchange–correlation functionals for the determination of electronic band gaps of solids. J. Chem. Theor. Comput. *15*, 5069–5079.

149. Heyd, J., Scuseria, G.E., and Ernzerhof, M. (2003). Hybrid functionals based on a screened Coulomb potential. J. Chem. Phys. *118*, 8207–8215.

150. Krukau, A.V., Vydrov, O.A., Izmaylov, A.F., and Scuseria, G.E. (2006). Influence of the exchange screening parameter on the performance of screened hybrid functionals. J. Chem. Phys. *125*, 224106.

151. Moellmann, J., and Grimme, S. (2014). DFT-D3 study of some molecular crystals. J. Phys. Chem. C *118*, 7615–7621.

152. Valenzano, L., Civalleri, B., Chavan, S., Bordiga, S., Nilsen, M.H., Jakobsen, S., Lillerud, K.P., and Lamberti, C. (2011). Disclosing the complex structure of UiO-66 metal organic framework: a synergic combination of experiment and theory. Chem. Mater. *23*, 1700–1718.

153. Saliba, D., Ammar, M., Rammal, M., Al-Ghoul, M., and Hmadeh, M. (2018). Crystal growth of ZIF-8, ZIF-67, and their mixed-metal derivatives. J. Am. Chem. Soc. *140*, 1812–1823.

154. Naito, T., Kakizaki, A., Inabe, T., Sakai, R., Nishibori, E., and Sawa, H. (2011). Growth of nanocrystals in a single crystal of different materials: a way of giving function to molecular crystals. Cryst. Growth Des. *11*, 501–506.

155. Sekine, Y., Tonouchi, M., Yokoyama, T., Kosaka, W., and Miyasaka, H. (2017). Built-in TTF-TCNQ charge-transfer salts in π-stacked pillared layer frameworks. CrystEngComm *19*, 2300–2304.

156. Salami, T.O., Patterson, S.N., Jones, V.D., Masello, A., and Abboud, K.A. (2009). 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. *12*, 1150–1153.

157. Sun, L., Hendon, C.H., Park, S.S., Tulchinsky, Y., Wan, R., Wang, F., Walsh, A., and Dincă, M. (2017). Is iron unique in promoting electrical conductivity in MOFs? Chem. Sci. *8*, 4450–4457.

158. Manna, S.C., Zangrando, E., Ribas, J., and Chaudhuri, N.R. (2005). Squarato-bridged polymeric networks of iron(II) with N-donor coligands: syntheses, crystal structures and magnetic properties. Inorgan. Chim. Acta *358*, 4497–4504.

159. Clements, J.E., Price, J.R., Neville, S.M., and Kepert, C.J. (2014). Perturbation of spin crossover behavior by covalent post-synthetic modification of a porous metal-organic framework. Angew. Chem. Int. Ed. *126*, 10328–10332.

160. Spirkl, S., Grzywa, M., Reschke, S., Fischer, J.K.H., Sippel, P., Demeshko, S., von Nidda, H.-A., and Volkmer, D. (2017). 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. *56*, 12337–12347.

161. Lee, J., and Asahi, R. (2020). Transfer learning for materials informatics using crystal graph convolutional neural network. arXiv, 2007.09932.

162. Sanyal, S., Balachandran, J., Yadati, N., Kumar, A., Rajagopalan, P., Sanyal, S., and Talukdar, P. (2018). MT-CGCNN: integrating crystal graph convolutional neural network with multitask learning for material property prediction. arXiv, 1811.05660.

163. Ramakrishnan, R., Dral, P.O., Rupp, M., and von Lilienfeld, O.A. (2015). Big data meets quantum chemistry approximations: the Δ-machine learning approach. J. Chem. Theor. Comput. *11*, 2087–2096.

164. Bannwarth, C., Ehlert, S., and Grimme, S. (2019). GFN2-xTB—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. J. Chem. Theor. Comput. *15*, 1652–1671.

165. Spicher, S., and Grimme, S. (2020). Robust atomistic modeling of materials, organometallic, and biochemical systems. Angew. Chem. Int. Ed. *59*, 15665–15673.

Matter, Volume *4*

Supplemental information

Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery

Andrew S. Rosen, Shaelyn M. Iyer, Debmalya Ray, Zhenpeng Yao, Alán Aspuru-Guzik, Laura Gagliardi, Justin M. Notestein, and Randall Q. Snurr

Contents

Section A: Supplemental Experimental Procedures

Publicly Available Data

Please refer to the following GitHub page for an overview of how to access the QMOF database as well as for additional scripts/tools needed to reproduce the machine learning results presented in this study: https://github.com/arosen93/QMOF. Data associated with the QMOF database is hosted via Figshare and has the following permanent DOI: 10.6084/m9.figshare.13147324.

Dataset Construction

Dataset Summary

A summary of the dataset construction process is shown in Figure S1, with the important datasets in this work summarized in Table S1.

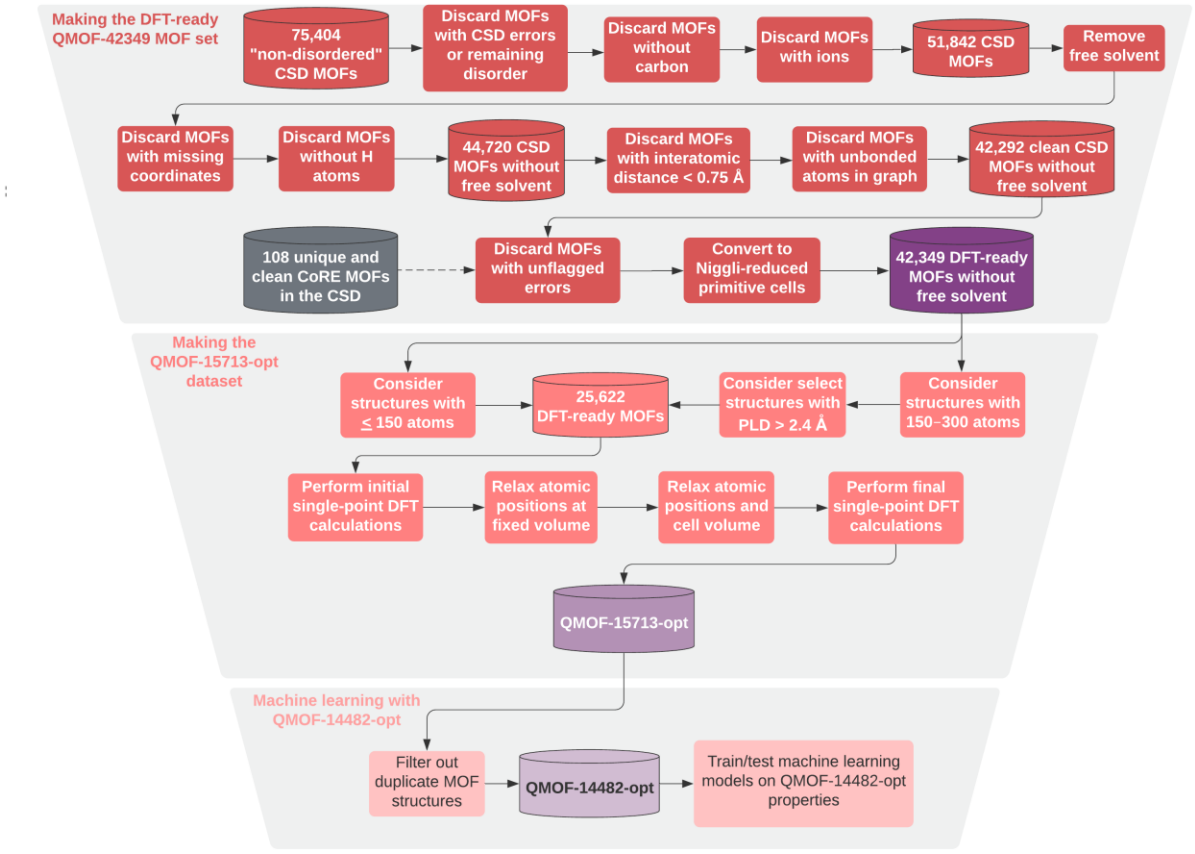


Figure S1. Workflow for generating the dataset of DFT-ready MOF structures and DFT-computed properties. Important datasets discussed throughout this study are highlighted in purple.

Table S1. Summary of the important datasets discussed throughout this work. All de-duplicated subsets are made using Pymatgen's `StructureMatcher` utility[1] to flag identical materials in the parent set.

Description	Name
Un-optimized, DFT-ready MOF structures.	QMOF-42349
DFT results for structures that passed all stages of the workflow. DFT-derived properties are those associated with the fully optimized ("opt") structures.	QMOF-15713-opt
De-duplicated subset of QMOF-15713-opt.	QMOF-14482-opt

The DFT-Ready, Free Solvent Removed QMOF-42349 Dataset

Obtaining the initial structures

In this work, we chose to take all crystal structures from the Cambridge Structural Database (CSD). As discussed below, existing databases of "cleaned" metal–organic framework (MOF) structures that have been widely used for grand canonical Monte Carlo simulations often contain several structural fidelity issues that can significantly impact the quality of density functional theory (DFT) calculations. Furthermore, starting from the unmodified CSD structures makes it possible to use ConQuest[2] for more complicated filtering stages that take into account CSD meta-data not necessarily present in the individual crystallographic information files (CIFs). The CSD also contains many more MOF structures than existing "pre-cleaned" experimental MOF databases, making it easier to generate a large database of computed properties for subsequent machine learning studies.

Disorder and error handling

To construct the dataset of MOFs to study with DFT, we began with the Aug. 2019 release of the CSD and considered the 75,404 structures that are part of the "non-disordered" MOF subset.[3] These structures lack disorder in the framework atoms but can potentially have disorder in the remaining species (e.g. free or coordinating solvents). We used ConQuest to remove any structures that were flagged as having any remaining disorder to increase the likelihood that the resulting CIFs would be physically reasonable for DFT calculations. While the CoRE MOF database attempts to automatically resolve disorder, this automated procedure is prone to occasional errors[4–6] and so we instead neglect any disordered materials in the present study. Additionally, we use ConQuest to remove any structures with CSD-flagged errors in the crystal structure.

Carbon requirement

We ensured that all structures contain at least 1 carbon atom, as this is an inherent requirement to yield a MOF. Several structures in the CoRE MOF 2019 database lack carbon atoms, many of which are best-described as inorganic metal–phosphate frameworks (e.g. refcodes ABETAE[7], BEFLIJ[8]).

Ion handling

We did not consider any structures that were flagged as having ions, as identified via ConQuest. This step is crucial, as it is often difficult to experimentally resolve all the charge-balancing ions, and many of these structures are therefore not charge-neutral. A structure with the incorrect number of electrons makes the resulting DFT calculations unphysical. This is a potential cause of inaccurate calculation results when screening MOF databases.[4–6]

Solvent removal

We chose to remove free (i.e. unbound) solvent molecules from each structure but retained solvent bound to the metal centers. We chose not to remove bound solvent, as automated scripts to remove bound solvent have been shown to incorrectly remove framework atoms on occasion.[4] The removal of bound solvent can also lead to undesirable charge-balancing issues. For instance, the structure with refcode ASAHEJ[9] in the "all solvent removed" subset of the CoRE MOF 2019 database is missing its terminal oxo ligands because they were incorrectly assumed to be bound water molecules. Another motivating factor for only removing free solvent is that it may not be feasible to remove bound solvent during the thermal activation procedure for some MOFs. Here, we removed all free solvents that have identical SMILES strings as the molecules included in the CSD list of solvents.[3]

Missing 3D coordinates

Following removal of free solvent, we used ConQuest to filter out any structures that have an atom flagged as having missing 3D coordinates. When structures are downloaded directly from the CSD, they may be missing atoms that were not able to be assigned based on X-ray diffraction (XRD). For instance, the MOF with refcode ADATAC[10] has terminal water groups bound to metal centers, but the unmodified CIF is missing the H atoms on the water ligands. Similarly, many Zr-containing frameworks are known to have complicated proton topologies, such that it can be difficult to distinguish between terminal oxo, hydroxo,

and water ligands from XRD alone.[11] If these H atoms are not included (or an incorrect number are included), this will lead to charge-balancing issues with the overall structure. In addition, based on the CIF alone, it can be difficult to tell if a terminal O atom should be an oxo, hydroxo, or water ligand, and this can complicate the solvent removal procedure if the user wishes to remove bound solvent. We note that, in cases where the CSD entry is appropriately annotated, it may be possible to retain more structures by adding the corresponding H atoms via the CSD Python API in future studies.[12]

No H atoms

Structures without any H atoms (following removal of free solvent) were discarded. While, in principle, a MOF could have a linker without H atoms, the more common scenario is that the H atoms were simply omitted from the structure, leaving behind highly unphysical organic groups. This is a well-established limitation with existing databases of MOF crystal structures.[4–6]

Short interatomic distances

Any structures with an interatomic distance less than 0.75 Å were discarded after the above filtering procedures. This can often happen if the structure has disorder that was not appropriately flagged in the CSD entry (e.g. partial occupancies were not supplied). Nearly overlapping atoms will also create challenges for the structure relaxation algorithms.

Lone atoms

After the above procedures, we used Pymatgen to generate crystal graphs of every MOF using the CrystalNN algorithm[13,14] and removed any structures that had lone (i.e. unbonded) atoms in the graph. As an example, this is necessary to remove structures like CAXVOO,[15] which has lone H atoms in the pores of the crystal structure (which should actually be H_2).

After all of the above procedures, this resulted in 42,292 MOFs (Figure S1).

Additional Structures Identified as MOFs in the CoRE MOF Database

To supplement this list of structures, we also considered the MOFs identified during the construction of v.1.1.2 of the 2019 CoRE MOF database.[16] The CoRE MOF structures were not used directly in this work. Rather, the corresponding CSD refcodes were identified and run through the aforementioned filtering procedure for consistency. The 2019 CoRE MOF database contains a maximum of 14,142 structures identified as MOFs, of which 13,544 can be found in the Aug. 2019 release of the CSD. Of these 13,544 structures, a total of 3,788 have no disorder, no errors, do not contain ions, and contain carbon. The majority of the structures removed in this process had disorder in the CIF. Of these 3,788 structures, 2,844 of them had H atoms and no missing coordinates. 2,699 MOFs were left after removing structures with lone atoms in the crystal graphs and ensuring that there were no interatomic distances less than 0.75 Å. Of these, only 108 were unique refcodes when compared to the 42,292 taken directly from the list of MOFs in the CSD MOF subset. This resulted in a combined dataset of 42,400 refcodes.

QMOF-42349

After the above procedure, we removed 51 additional structures that had disorder or missing H atoms not flagged via the automated ConQuest search, the majority of which have been mentioned in prior work.[12] Finally, this left us with a suitably DFT-ready dataset containing a grand total of 42,349 structures, which we refer to as the **QMOF-42349** dataset. The list of refcodes for the QMOF-42349 dataset, the script to remove free solvent, and the intermediate lists of refcodes are available with the supporting dataset.[17] All the CIFs in the QMOF-42349 dataset were converted to their Niggli-reduced primitive unit cells using Pymatgen prior to carrying out the DFT calculations.[1]

Completed Job Statistics to Yield the QMOF-15713-opt Dataset

From the Niggli-reduced QMOF-42349 dataset, we started by selecting MOFs with $\leq$ 150 atoms to ensure that a large number of DFT calculations could be carried out. A total of 24,002 structures fit this criterion. Of the 24,002 MOFs with $\leq$ 150 atoms per Niggli-reduced unit cell considered for the high-throughput periodic DFT screening, 19,308 successfully completed the initial single-point (i.e. static) calculation, and a

total of 14,170 successfully completed every step of the workflow. While some calculations did not complete due to wall-time limits and related resource limitations, the majority of the incomplete calculations can be attributed to not meeting the strict 10^{-6} eV self-consistent field (SCF) convergence tolerance in 150 iterations during the initial single-point calculation. Many of these cases would likely have the SCF converged after a few steps of the geometry optimization, as it is common for the first few steps to require the largest number of SCF cycles to reach convergence. However, to be on the cautious side and to maximize overall resource usage, we did not consider them further or run them for a greater number of SCF iterations. We refer to the computed properties of the 14,170 DFT-optimized structures as the QMOF-14170-opt ("opt" = optimized) dataset. The corresponding single-point data on the starting structures is referred to as the QMOF-14170-SP ("SP" = single-point) dataset.

The above procedure was how v1 of the QMOF database was generated. Since its initial release, we decided to expand the database further by including structures not already in the QMOF-14170-opt dataset. Specifically, we identified unique structures not in the QMOF-14170-opt dataset with a pore-limiting diameter greater than 2.4 Å (prior to structure relaxation) and increased the limit on the maximum atoms per cell from 150 to 300. 2185 new structures fitting these criteria were added to the DFT workflow, of which 1543 made it through the entire structure relaxation procedure.

Collectively, a grand total of 15,713 structures completed the structure relaxation workflow and are included in the current version (v4) of the QMOF database. We refer to the completed structural relaxations as the **QMOF-15713-opt** dataset.

De-Duplication to Yield the QMOF-14482-opt Dataset

Prior to this point, duplicate structures were not removed, as slight variations in the input geometry could potentially lead to different optimized structures, and the definition of unique will ultimately depend on the application of interest. Nonetheless, for training machine learning (ML) models, it is important to have a diverse dataset, and identical structures may lead to unrepresentative testing statistics. Therefore, we used Pymatgen's `StructureMatcher` tool (using the default algorithm) on the 15,713 initial, un-relaxed structures and their relaxed counterparts to identify a unique subset of 14,482 structures. For Niggli-reduced primitive cells of two structures, the `StructureMatcher` scales the two lattice volumes, aligns the crystal lattices, and compares the atomic distances. While other methods, such as MOFid/MOFkey[18] or a comparison of the underlying crystal graphs, could be used to identify unique MOFs based on their building blocks, here we used a geometrically sensitive structure matching approach so that identical nodes/linkers but different geometries would still be considered as separate entities in the dataset. For instance, QUPZIM, QUPZIM01, and QUPZIM02 are the same MOF with the same composition and connectivity, but the first is the closed-pore analogue of the latter two, and the latter two are conformationally distinct.[19] All three are included in the de-duplicated subset, as they can potentially have different electronic structure properties (as has been shown for other flexible frameworks in the literature[20]). We acknowledge that no matter what approach is taken, there will nonetheless be a few MOFs in the dataset with very similar structures. Other de-duplication schemes are always possible, and we encourage users to consider different approaches depending on their intended use-case.

Of the 15,713 structures, 14,482 were classified as unique and used for machine learning. The quantum-chemical properties of these 14,482 structures are collectively referred to as the **QMOF-14482-opt** dataset.

High-Throughput Periodic DFT Screening

VASP Details

Plane-wave, periodic density functional theory calculations were carried out using the Vienna *ab initio* Simulation Package (VASP) v.5.4.4.[21,22] The widely used and computationally tractable PBE exchange-correlation functional[23] with Grimme's D3 dispersion correction[24] and Becke–Johnson (BJ) damping[25] was used to generate a sufficiently large dataset for the purposes of training machine learning models. PBE with dispersion corrections has been shown to accurately capture the geometries of MOFs.[26,27] Based on prior benchmarking work,[28] the following parameters were generally used for the results presented in this study (see Table S2 for more details). A 520 eV plane-wave kinetic energy cutoff was applied with a k-point density of ~1000 per number of atoms, as arranged using Pymatgen 2019.9.16.[1] The VASP-recommended v.54 projector-augmented wave (PAW)[22,29] pseudopotentials were considered for all elements, with the

exception of Li (for which we used the standard 140 eV default cutoff potential for computational simplicity), Eu (for which we use the Eu_3 pseudopotential rather than the Eu_2 pseudopotential since Eu(III) is more common), Yb (for which we use the Yb_3 pseudopotential rather than the Yb_2 pseudopotential since Yb(III) is more common), and W (for which we use the _sv pseudopotential since the _pv pseudopotential is not included in the v.54 PAW set). All elements have a default cutoff of $\leq$ 520 eV when multiplied by 1.3 (to prevent Pulay stresses upon volume relaxation[30]). Structure relaxations were considered converged when the net force on each atom is below 0.03 eV/Å.

The accurate-precision keyword was enabled in VASP. Gaussian smearing of the band occupancies with a smearing width of 0.01 eV was applied, with extrapolation back to the 0 K limit. Symmetry operations were disabled. The SCF was converged using the "Fast" algorithm, which is a mixture of the Davidson and residual minimization method–direct inversion in the iterative subspace (RMM-DIIS) algorithms.[31] If the SCF did not converge to 10^{-6} eV within 150 iterations, the calculation was aborted and the results not considered in this work. In some cases, challenging SCF convergence can be attributed to an incorrect structure, oftentimes a result of a structure that is not charge-neutral. Spin-polarization was considered in a similar manner as several previous DFT-computed property databases.[32,33] Here, any d-block metals (excluding Zn, Cd, and Hg) were initialized with a magnetic moment of 5 μ_B. All f-block elements (excluding Lu and Lr) were initialized with 7 μ_B. All other elements were not initialized with any spin. We note that in VASP, the magnetic moments can freely change throughout the SCF convergence procedure, reaching a local minimum configuration once converged.

Breakdown of Sequential Steps in Periodic DFT Workflow

Each calculation was broken down into five sequential stages, similar to what has been described and benchmarked previously.[28] These stages generally include: 1) An initial, high-accuracy single-point calculation (520 eV cutoff, ~1000 k-points per number of atoms); 2) A (coarse accuracy) relaxation of the atomic positions (default plane-wave kinetic energy cutoff, ~100 k-points per number of atoms, 0.05 eV/Å force tolerance); 3) A medium-accuracy relaxation of the cell volume and atomic positions (520 eV cutoff and ~100 k-points per number of atoms, 0.03 eV/Å force tolerance); 4) A high-accuracy relaxation of the cell volume and atomic positions (520 eV cutoff, ~1000 k-points per number of atoms, 0.03 eV/Å force tolerance); 5) A final high-accuracy single-point calculation of the fully optimized structure using the aforementioned settings. If the SCF did not converge for any step within 150 iterations, that calculation was not considered to be complete, and the remaining steps of the workflow were not carried out. If any individual stage of the workflow took greater than 2 hours per MOF, the job was canceled, and the remaining stages were also not carried out. On-the-fly error-handling was used to correct for warnings and errors should they appear,[28] but if for any reason the job crashed and could not be successfully continued, that MOF was also not considered further. The VASP input parameters are summarized in Table S2. All VASP calculations were carried out using the Atomic Simulation Environment (ASE) 3.19.0b1.[34] Band gaps were obtained using `pymatgen.io.vasp.outputs.Eigenval()` with an occupancy tolerance of 10^{-8}. Partial atomic charges, spin densities, and effective bond orders were computed using the density-derived electrostatic and chemical (DDEC6) method[35–38] as implemented in Chargemol 09-26-2017.[39] Charge Model 5 (CM5) charges[40] were also computed using Chargemol 09-26-2017. PyMOFScreen commit #e9768a5 was used to manage and carry out the automated DFT calculations.[41]

Table S2. ASE input arguments for the VASP calculators used in the screening workflow, excluding file I/O-related keywords.* Note that the appropriate pseudopotentials can be automatically selected with `setups={'base':'recommended','Li':'', 'Eu':'_3','Yb':'_3', 'W':'_sv'}`.

Flag	Stage 1	Stage 2**	Stage 3	Stage 4	Stage 5
`xc`	'PBE'	'PBE'	'PBE'	'PBE'	'PBE'
`ivdw`	12	12	12	12	12
`encut`	520[a]	400	520	520	520
`kppa`[b]	1000[a]	100	100	1000	1000
`isif`	—	2	3	3	—
`ibrion`	—	2	2	2	—
`prec`	'Accurate'	'Accurate'	'Accurate'	'Accurate'	'Accurate'
`ismear`	0	0	0	0	0

sigma	0.01	0.01	0.01	0.01	0.01
ediff	1E-6	1E-4	1E-6	1E-6	1E-6
algo	'Fast'	'Fast'	'Fast'	'Fast'	'Fast'
nelm	150	150	150	150	150
nelmin	3	3	3	3	—
lreal	False	False[c]	False[c]	False	False
nsw	0	250[d]	30[e]	30[e]	0
ediffg	—	-0.05	-0.03	-0.03	—
lorbit	11	11	11	11	11
isym	0	0	0	0	0
symprec	1E-8	1E-8	1E-8	1E-8	1E-8

[a]For structures in v1 of the QMOF database (QMOF-14170), "Stage 1" (i.e. the static calculation on the starting structure) used the same settings as "Stage 5" (i.e. the static calculation on the optimized structure) to enable direct comparisons between properties (e.g. Figure S5). For subsequent additions to the database, "Stage 1" used the same `encut` and `kppa` as in "Stage 2" for increased computational efficiency.
[b]`kppa` = k-point density (i.e. number of k-points/number of atoms), computed with the `automatic_density()` tool in Pymatgen. The choice of whether the grid should be Γ-centered or not (i.e. `gamma=True` or `gamma=False`) and how the k-points are distributed among the three lattice dimensions are also determined based on this Pymatgen utility.
[c]Switches to `lreal='Auto'` if the VASP output file suggests doing so due to a large unit cell (only for Stages 2 – 3).
[d]The max 250-cycle relaxations of atomic positions were sequentially repeated until the force tolerance given by `|ediffg|` (0.05 eV/Å) was achieved.
[e]The max 30-cycle volume relaxations were sequentially repeated until the force tolerance given by `|ediffg|` (0.03 eV/Å) was achieved. For Stage 4, after this process was completed, a final max 100-cycle volume relaxation was carried out for good measure.
[*]Slight changes to the input parameters that do not affect the accuracy of the results may occur during the workflow to correct for errors on-the-fly. For instance, the conjugate-gradient (CG) algorithm (`ibrion=2`) often leads to a bracketing error when the potential energy surface is flat, and in such a scenario the geometry optimization algorithm automatically switches to the Fast Inertial Relaxation Engine (FIRE)[42] (`ibrion=3, iopt=7, potim=0`).
[**]The coarse-accuracy update of the atomic positions is preceded by an initial relaxation using the `BFGSLineSearch` algorithm in ASE until the maximum net force is less than 10 eV/Å. Empirically, we have found that this algorithm is better at resolving high forces without the structure "exploding" when compared to the CG algorithm.

Further Investigation of Selected MOFs

For select calculations, we use a hybrid-level functional to improve the quality of the band gap predictions. Using the PBE-D3(BJ) wavefunction and structure as a starting point, the HSE06-D3(BJ) level of theory[43–45] was used to re-relax the unit cell shape, volume, and atomic positions. Due to the high computational cost when running periodic DFT calculations with hybrid functionals, a looser force tolerance of 0.05 eV/Å was adopted. For all HSE06-D3(BJ) calculations, the VASP-recommended preconditioned conjugate gradient "all bands simultaneous update of orbitals" algorithm[46–48] (`algo='All'`) was used to converge the SCF, the SCF convergence was set to a slightly looser value of 10^{-5} eV, and the density of states (DOS) was evaluated with 3000 grid-points. For increased computational efficiency, the HSE06-D3(BJ) structure relaxations were occasionally carried out using a smaller k-point grid than the single-point calculation used to evaluate the band gap and DOS for select materials (Table S3). All other settings remain unchanged from "Stage 5" of Table S2.

Table S3. k-point grids for selected MOFs at the HSE06-D3(BJ) level of theory. k-points (low) and (high) refer to the structure relaxation and subsequent electronic structure analysis, respectively.

CSD Refcode	k-points (low)	k-points (high)
LOJLAZ	$2 \times 2 \times 1$	$2 \times 2 \times 1$
RAXNEK	$2 \times 1 \times 1$	$3 \times 2 \times 1$
WAQMEJ	$2 \times 1 \times 1$	$3 \times 1 \times 1$

GUTYAW	$2 \times 2 \times 1$	$4 \times 4 \times 1$

Table S4. HSE06-D3(BJ) primitive cell lattice parameters compared with experiment. Note that any free solvent present in the crystal structure was removed from the framework in the DFT calculations. LS = low spin; HS = high spin.

CSD Refcode		a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)
LOJLAZ-LS	Theory	10.03	10.03	15.06	90.0	90.0	74.8
	Exp.	10.07	10.07	15.10	90.0	90.0	73.6
LOJLAZ-HS	Theory	10.40	10.40	15.46	89.9	90.0	71.0
	Exp.	10.38	10.38	15.50	90.0	90.0	70.1
RAXNEK	Theory	7.99	11.56	12.47	117.6	99.4	90.0
	Exp.	7.98	11.68	12.79	117.2	100.2	90.0
WAQMEJ	Theory	7.98	13.65	14.12	91.6	99.1	90.1
	Exp.	8.28	13.81	14.08	91.1	97.1	90.5
GUTYAW	Theory	4.88	4.88	14.90	86.9	86.9	66.9
	Exp.	4.97	4.97	14.98	87.0	87.0	66.4

The HSE06-D3(BJ) lattice parameters for these materials are shown in Table S4, and the relevant spin states are discussed below.

1. LOJLAZ, Fe(bipytz)(Au(CN)$_2$)$_2$ (bipytz = 3,6-bis(4-pyridyl)-1,2,4,5-tetrazine): This material has Fe(II) and Au(I) species. Experimentally, it has been shown that LOJLAZ is a spin-crossover framework that has a low-to-high spin transition with increasing temperature.[49] We consider both spin states in this work as a matter of consistency with the spin-crossover behavior observed experimentally. For reference, the high spin state is predicted to be 42 kJ/mol (per cell) more stable than the low spin state at the HSE06-D3(BJ) level of theory.
2. RAXNEK, Fe(sq)(bpee)(H$_2$O)$_2$ (bpee = 1,2-bis(4-pyridyl)ethylene; sq = squarate): This material has Fe(II) species, which are known to exist in the high spin state with antiferromagnetic coupling.[50] At the HSE06-D3(BJ) level of theory, a high-spin ground state is found. Both ferromagnetic and antiferromagnetic states were found to have comparable structures and energies, so we model the latter as a matter of consistency with the reported experiments.
3. WAQMEJ, (TTF)[{Rh$_2$(CH$_3$CO$_2$)$_4$}$_2$TCNQ]: This material is reported to have diamagnetic (formally) Rh(II) dimers with antiferromagnetically coupled TTF–TCNQ species such that the net magnetic moment is zero.[51] A spin-unrestricted state with a net magnetic moment of zero was found to be the ground state at the HSE06-D3(BJ) level of theory.
4. GUTYAW, Sr[C$_2$H$_4$(SO$_3$)$_2$]: This material has Sr(II) cations, and the framework is modeled as spin-restricted based on its structure.

Additional Software and Hardware Details for Machine Learning

Regression-based machine learning model development was carried out using scikit-learn v.0.23.2 and the standard SciPy stack with NumPy v.1.19.2, pandas 1.1.5, Matplotlib 3.3.2, and Seaborn v.0.11.1.[52–57] Pymatgen[1] v.2020.12.3 was used to analyze structures and generate descriptors. The smooth overlap of atomic positions (SOAP) features were computed using DScribe v.0.4.0.[58] The sine Coulomb matrix and Meredig and Agrawal et al.[59] features were generated using Matminer v.0.6.4.[60] Crystal graph convolutional neural networks (CGCNNs) were based on the work of Xie and Grossman[61] and used PyTorch v.1.6.0[62] for constructing and evaluating the neural networks. Specifically, our CGCNN code is built upon commit #d612a69 of the CGCNN code[63] with slight variations, as reflected in our fork of the CGCNN code.[64] This fork saves the crystal graphs to .pkl files so they can be read in as-needed instead of needing to be re-computed when the memory cache is filled. This is a common problem with MOF crystal graphs given the large size of the unit cells. A branch of this revision[65] also makes it possible to use Pymatgen-computed crystal graphs rather than those based on a fixed number of neighbors, although we did not observe any improvement when using a crystal graph based on the CrystalNN algorithm.[13,14] This is potentially because

there are many crystal structures in the QMOF database that are connected in 1D or 2D, such that there are disconnected regions of the Pymatgen-generated crystal graph. PTable Trends v.2.0[66] was used to generate a heat map over the periodic table. Zeo++ v.0.3 was used for the pore diameter calculations using the "high accuracy" flag.[67] PyProcar v.5.6.1 was used to parse the DOS data.[68] Timing data for the machine learning models are reported using Python 3.8.5 on a laptop with an Intel Core i7-9750H CPU. For the CGCNNs, CUDA v.10.1 was used to enable GPU support with an NVIDIA GeForce RTX-2070 (Max-Q Design) graphics card.

Dataset Handling for Training and Evaluating Machine Learning Models

Unless otherwise stated 80% of the 14,482 data points was reserved for training while 20% was held-out for testing of the kernel ridge regression (KRR) models. To optimize the hyperparameters and determine the optimal ML models, 5-fold cross-validation of the training set was applied for KRR. Due to the higher computational cost when training neural network models, for CGCNN, 80% of the data was reserved for training, 10% was held-out for validation, and 10% was reserved for testing. In all cases, performance of the models on the testing data was not inspected until the end of the project when ideal models were determined on the basis of the validation process. Data splitting was done via purely random sampling. To account for minor variations in model performance due to sampling bias, all the performance statistics in Table 1 and Figure 4A are reported as averages over five separate runs with different random seeds for the data splitting (arbitrarily chosen in advance to be 42, 125, 267, 541, and 582). Elsewhere, a constant seed is used for consistency (chosen in advance to be 42).

Learning Curves

For the learning curves in Figure 4A, training set sizes of 2^7, 2^8, 2^9, 2^{10}, 2^{11}, 2^{12}, 2^{13}, and 80% of the full dataset of 14,482 data points were investigated. Powers of 2 were chosen to allow for equidistant spacing on a logarithmic grid. For internal consistency, the same testing set was used (for a given data-splitting seed) regardless of training set size. The same validation set was also used (for a given data-splitting seed) for the CGCNN models. For the KRR models, 20% of the full QMOF-14482-opt dataset was held-out for testing. For the CGCNN models, 10% of the full QMOF-14482-opt dataset was used for validation, and 10% of the full QMOF-14482-opt dataset was held-out for testing.

Kernel Ridge Regression

KRR combines the kernel trick with ridge regression.[69] Like all regression methods, the goal of KRR is to predict a response variable $\mathbf{y}$ from a set of individual input vectors $\mathbf{x}$ (which, when combined, form a feature matrix $\mathbf{X}$ containing an encoding of each individual material). KRR, being a kernel method, achieves this by transforming $\mathbf{X}$ into a kernel matrix $\mathbf{K}$ that describes the similarity between every pair of materials in $\mathbf{X}$. In this way, KRR has a closed-form solution given by

$$\mathbf{w} = (\mathbf{K}_{\text{train}} + \lambda \mathbf{I})^{-1} \mathbf{y}_{\text{train}} \tag{S1}$$

where $\mathbf{w}$ is the vector of model weights, $\mathbf{K}_{\text{train}}$ is the training set kernel matrix, λ is the regularization hyperparameter, $\mathbf{I}$ is the identity matrix, and $\mathbf{y}_{\text{train}}$ is the training set values to predict. For scikit-learn's implementation of KRR, a parameter α is supplied, which is defined as $\alpha \equiv \lambda/2$.

With the model weights obtained, new values can be predicted via

$$\mathbf{y}_{\text{ML}} = \mathbf{K}_{\text{test}} \mathbf{w} \tag{S2}$$

where $\mathbf{y}_{\text{ML}}$ are the ML-predicted $\mathbf{y}$ values for a new kernel matrix of the testing set $\mathbf{K}_{\text{test}}$. For N training samples and M testing samples, $\mathbf{K}_{\text{train}}$ will have dimensions of $(N \times N)$ and $\mathbf{K}_{\text{test}}$ will have dimensions of $(M \times N)$. Here, $\mathbf{K}_{\text{train}}$ represents the similarity between every pair of structures in the training set, whereas $\mathbf{K}_{\text{test}}$ represents the similarity between each structure in the training set and each structure in the testing set. The transformation of $\mathbf{X} \rightarrow \mathbf{K}$ can be achieved by one of several kernel functions. For all kernel methods (except for SOAP), we use a Laplacian kernel function, k, given by

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \left\| \mathbf{x}_i - \mathbf{x}_j \right\|_1\right) \tag{S3}$$

where γ serves as an adjustable KRR model hyperparameter. In the case of SOAP, a similarity kernel $\mathbf{K}$ is directly generated and so there is no need for further transformation.

In all cases throughout his work, $\mathbf{y}_{\text{train}}$ refers to the DFT-computed band gaps of the DFT-optimized structures, whereas $\mathbf{X}$ refers to the encodings of the corresponding unrelaxed crystal structures.

Featurization Methods for KRR

As mentioned in the main text, several featurization methods were pursued for generating the feature matrices $\mathbf{X}$ for use with KRR, which we summarize in this section. For all non-SOAP featurization methods, a min-max scaler was applied during the KRR process, such that each feature was scaled to the range 0 – 1.

Description: Stoichiometric-120 Features

The Meredig and Agrawal et al.[59] feature set ("Stoichiometric-120") is a composition-based descriptor that was originally developed for formation energy predictions of inorganic solids in the Open Quantum Materials Database[32,70] (OQMD). In this work, the descriptor set has 120 attributes. 103 of these encode the elemental composition via the fraction of each unique element from H–Lr in the MOF. The remaining attributes are the mean atomic weight, mean group number, mean period number, maximum difference in atomic number, mean atomic number, range in atomic radii, mean atomic radius, range in electronegativities, mean electronegativity, the average number of s, p, d, and f valence electrons, and the composition-weighted fraction of s, p, d, and f valence electrons.

Description: Stoichiometric-45 Features

The He et al.[71] feature set ("Stoichiometric-45") is a composition-based descriptor that has been used to classify if inorganic solids in the OQMD are metallic or non-metallic. The descriptor set has 45 attributes. These consist of 9 elemental properties (atomic number, group number, period number, electronegativity, electron affinity, melting temperature, boiling temperature, density, and ionization energy) and five statistical quantities of each (arithmetic mean, geometric mean, standard deviation, maximum, and minimum) computed for each structure. The tabulated data was taken from the Wolfram Knowledgebase, which we accessed with Mathematica 11.3.0. It is important to note that several of these attributes (e.g. melting and boiling temperatures, density) are ill-defined for single atoms. In the Wolfram Knowledgebase, these values are generally defined as being for stable bulk forms at ambient conditions. Furthermore, the electron affinities and ionization energies were chosen to be for the addition or removal of a single electron, respectively. We also chose to place the lanthanides and actinides in a fictitious group 19. We note that 477 out of the 14482 MOF structures in the QMOF-14482 dataset did not have fingerprints generated due to missing tabulated data for one or more of the elements in the structure and were therefore not considered with this featurization method.

Description: Sine Coulomb Matrix Eigenspectrum

The sine Coulomb matrix[72] is a structure-based featurization method where a pair-wise interaction matrix M_{ij} is generated by the following formula:

$$M_{ij} = \begin{cases} 0.5 Z_i^{2.4}, & i = j \\ \dfrac{Z_i Z_j}{\left| \mathbf{B} \cdot \sum_{k=\{x,y,z\}} \hat{\mathbf{e}}_k \sin^2\left(\pi \mathbf{B}^{-1} \cdot \left(\mathbf{R}_i - \mathbf{R}_j\right)\right)\right|}, & i \neq j \end{cases} \tag{S4}$$

where i and j are two atoms in the structure, Z_i is the atomic number of i, $\mathbf{B}$ is a matrix formed by the lattice vectors, $\hat{\mathbf{e}}_k$ are the Cartesian unit vectors, and $\mathbf{R}_i - \mathbf{R}_j$ is the distance vector between atoms i and j. The sine Coulomb matrix is dependent on the number of atoms a given structure has, so to ensure a square matrix is generated, it is padded with zeros to match the maximum number of atoms in the dataset (i.e. 300 atoms in the case of the QMOF-14482 dataset). Since a feature vector for each material is needed for KRR, only the (sorted) eigenvalues of the sine Coulomb matrix are returned such that the descriptor becomes one-dimensional for each structure with a length of $n_{\max \text{atoms}}$ (i.e. 300 in this case). This approach was chosen instead of flattening the sine Coulomb matrix because the resulting feature length would otherwise

be extremely large, as the sine Coulomb matrix for each material has dimensions $n_{\max \text{ atoms}} \times n_{\max \text{ atoms}}$ (i.e. 90,000 total entries upon flattening).

Description: Orbital Field Matrix

The orbital field matrix[73] encodes each atom in a structure by a constant-length vector representing the valence subshells of the atomic environments in each structure. To do so, each atom in a structure is represented via its (neutral) electron configuration. This electron configuration is turned into a numerical vector via a one-hot encoding scheme using a dictionary composed of the possible valence subshell orbitals and their occupancies (i.e. s^1, s^2, p^1, p^2, …, p^6, d^1, d^2, …, d^{10}, f^1, f^2, …, f^{14}). This is a 32-entry one-hot encoding. As implemented in matminer,[60] we supplemented this 32-entry encoding with 7 extra entries that represent the one-hot encodings of the period number for the element (with lanthanides in period 6 and actinides in period 7). These atomic one-hot encoding vectors are then used to construct one-hot encoding vectors for each atomic local environment (i.e. an atom center and its coordinating atoms). This is achieved by defining a 39×39 matrix obtained by multiplying the one-hot encoding vector of the central atom and a given coordinating atom. The orbital field matrix for an atomic environment is then the sum of these matrices between the center atom and each of its coordinating atoms, scaled by a distance function. This distance function is the inverse of the bond distance multiplied by a weighting factor (using the solid angle determined by the Voronoi polyhedra between the center atom and each neighbor). Each atomic environment orbital field matrix is converted into a structural orbital field matrix by averaging the atomic environment matrices across every atomic site such that each structure is described by an averaged 39×39 matrix, which is flattened to a 1521-length encoding. Additional details can be found in the original work by Pham and coworkers.[73]

Description: Average SOAP Kernel

SOAP is a featurization method that encodes information about local atomic environments in a structure, which can then be used with an appropriate kernel function to measure the structural similarity between every pair of structures in a given dataset. For full details regarding SOAP, we refer the reader to the original paper on the use of SOAP for structure comparison[74] and the brief summary and implementation of SOAP in the original DScribe paper,[58] which we summarize below. We note that we have adopted much of the nomenclature from Musil and coworkers[75] to clarify the description of the SOAP kernel.

We start by representing a given structure using local atomic densities ρ, separately defined for each atomic element Z. The local density of atoms within a chemical environment χ_i (i.e. a spherical region centered around atom i) is described as a sum of Gaussians placed at the central atom and the neighboring atoms within a cutoff region r_{cut}. Mathematically, this is expressed as

$$\rho_{\chi_i}^{Z}(\mathbf{r}) = \sum_{k} \exp\left(-\frac{|\mathbf{r} - \mathbf{R}_k|^2}{2\sigma^2}\right) \tag{S5}$$

where σ is the standard deviation of the Gaussians, and $|\mathbf{r} - \mathbf{R}_k|$ describes the distance between atom k, $\mathbf{R}_k$, and position vector $\mathbf{r}$. The origin, $\boldsymbol{r} = \mathbf{0}$, is centered on the local point of interest (i.e. atom i). The summation is carried out for all atoms k with atomic number Z in the structure that are within radius r_{cut} from atom i.

Given local atomic environments i and j in two structures A and B, one can then compute $\rho_{\chi_i^A}^{Z}$ and $\rho_{\chi_j^B}^{Z}$. The structural similarity between two chemical environments in structures A and B, denoted χ_i^A and χ_j^B, is

$$\tilde{k}\left(\chi_i^A, \chi_j^B\right) = \int_{\mathrm{SO}(3)} \left| \sum_{Z} \int_{\mathbb{R}^3} \rho_{\chi_i^A}^{Z}(\mathbf{r}) \rho_{\chi_j^B}^{Z}(\mathbf{r}) \mathrm{d}\mathbf{r} \right|^2 \mathrm{d}\hat{R} \tag{S6}$$

where SO(3) and $\hat{R}$ refer to the group of all three-dimensional rotations. The above expression is necessary to achieve a rotationally invariant descriptor and describes the (squared) overlap of the density fields, integrated over all three dimensional rotations. In practice, the calculation of $\tilde{k}$ is carried out by expanding

$\rho(\mathbf{r})$ using $n_{\max}$ real spherical harmonic and $\ell_{\max}$ radial basis functions.[58,74] We use spherical Gaussian type orbitals (GTOs) for the radial basis function in this work. The expression for $\tilde{k}$ can be normalized via

$$k(\chi_i^A, \chi_j^B) = \frac{\tilde{k}(\chi_i^A, \chi_j^B)}{\sqrt{\tilde{k}(\chi_i^A, \chi_i^A)\tilde{k}(\chi_j^B, \chi_j^B)}} \tag{S7}$$

such that the self-similarity of a given environment, $k(\chi_i^A, \chi_i^A)$ or $k(\chi_j^B, \chi_j^B)$, is equal to 1.

The similarity between all local atomic environments i in structure A and all local environments j in structure B is then given by the general expression

$$C_{ij}(A, B) = k(\chi_i^A, \chi_j^B) \tag{S8}$$

For a pair of structures A and B, we can then compute an average kernel to go from the similarity of local environments to the similarity of global structures. This average kernel function is defined as

$$K(A, B) = \left(\frac{1}{n_A n_B} \sum_{ij} C_{ij}(A, B) \right)^{\xi} \tag{S9}$$

where n_A and n_B are the number of atoms in structure A and B, respectively. Taking the summation over all sites i and j between the pairs of structures and then dividing by the number of atoms in both structures converts these otherwise local similarity scores into a global structural descriptor comparing structures A and B. The variable ξ is an optional model hyperparameter to modify the spread of entries in the kernel matrix, which we include as a tunable parameter during the KRR grid search. This expression for $\mathbf{K}$ can be readily extended for all relevant pairs of structures, which can then be used directly with KRR. An example of the average SOAP similarity kernel for IRMOF-1, IRMOF-2, and ZIF-8 is shown in Figure S2 for reference. Note that $\mathbf{K}$ is normalized such that self-similarity is unity (i.e. $K(A, A) = K(B, B) = 1$).

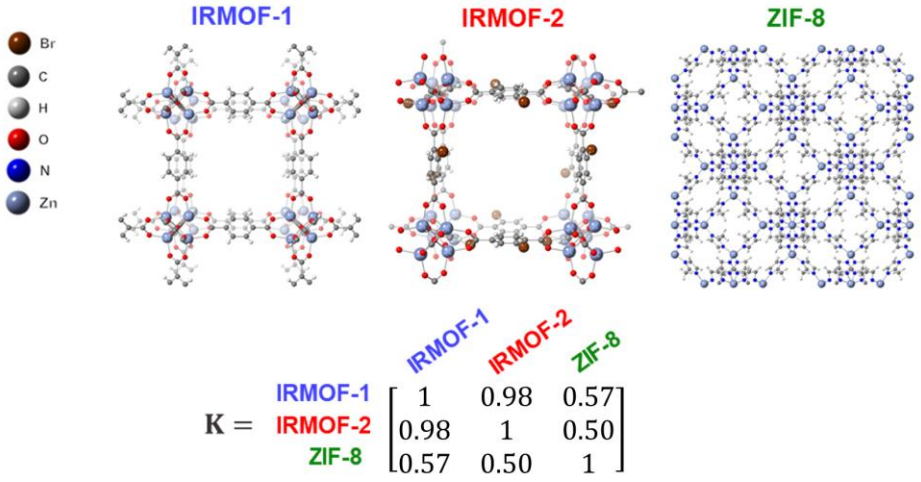


Figure S2. Average (normalized) SOAP similarity kernel for IRMOF-1, IRMOF-2, and ZIF-8. Here, $r_{\text{cut}} = 4$ Å, $\sigma = 0.1$ Å, $\xi = 2$, and $n_{\max} = \ell_{\max} = 9$.

Hyperparameter Tuning for KRR

For featurization methods other than SOAP, a feature matrix $\mathbf{X}$ is generated. As such, a decision must be made for the type of kernel function that should be used. The hyperparameters were identified based on a grid search via 5-fold cross-validation. Initially, we considered linear, Gaussian, and Laplacian kernel functions, of which we eventually decided to use a Laplacian kernel function (Equation S3) where $\gamma = 0.1$ since this consistently yielded the lowest cross-validation mean absolute error (MAE). A value of $\alpha = 0.1$

for the KRR regularization hyperparameter (Equation S1) was chosen for all KRR models except for the sine Coulomb matrix, for which we use α = 0.01.

For the SOAP-based KRR model, there are two KRR parameters to tune: α and ξ. These hyperparameters were also optimized using a grid search via 5-fold cross-validation, for which we decided upon $\alpha = 0.001$ and $\xi = 2$. The SOAP descriptor itself also has several hyperparameters that can be tuned, including but not limited to the distance cutoff for determining local regions within a structure (r_{cut}), the maximum number of radial basis functions (n_{max}), the maximum number of spherical harmonics (ℓ_{max}), and the standard deviation of the Gaussians used to expand the atomic density (σ). Although there are too many parameters to easily carry out an exhaustive grid search, each parameter was independently adjusted, and the parameters that reduced the (average) MAE over the 5-fold cross-validation process were retained. This led to $r_{\mathrm{cut}} = 4$ Å, $\sigma = 0.1$ Å, and $n_{\mathrm{max}} = \ell_{\mathrm{max}} = 9$. All other SOAP hyperparameters and settings were set to the default values in DScribe.

Crystal Graph Convolutional Neural Networks

CGCNN featurizes each crystal structure as an approximate crystal graph, defined such that the nodes are atoms and the edges are the atom connections, accounting for periodic boundary conditions. The crystal graphs are constructed by searching for a maximum set of neighbors within some user-defined cutoff distance. These crystal graphs are then fed as input to a convolutional neural network, wherein convolution and pooling layers convert the crystal graph to a given output, with the weights of the neural network updated to minimize the validation loss. Further details can be found in the original CGCNN paper.[61]

Iterative testing of the various CGCNN hyperparameters led to the following high-performing convolutional neural network configuration with regards to a reduced validation MAE: 5 convolutional layers, 64 hidden atom features in the convolutional layers, 1 fully connected hidden layer after pooling, and 128 hidden features after pooling. A batch size of 16, initial learning rate of 0.01, and stochastic gradient descent optimizer were used. All other settings were the default values, including a neighbor search radius of 8 Å and a maximum of 12 neighbors connected to every node in the graph. The best model obtained within 400 epochs (in terms of validation MAE) was retained. We note that several variations on the original CGCNN algorithm, such as CGCNN with a $\tanh$ activation function[76] and iCGCNN,[77] did not show substantial improvements over the original CGCNN implementation, although a detailed exploration of the hyperparameter space was not carried out.

In the original CGCNN work,[61] the (initial) CGCNN node (i.e. atom) feature vectors were based on one-hot encodings of group number, period number, electronegativity, covalent radius, number of valence electrons, first ionization energy, electron affinity, block, and atomic volume. However, the currently published version of the code contains several inconsistencies in the one-hot encodings compared to that reported in the original text.[78] As such, we regenerated the atom initialization file and made several minor modifications to the initialization process, wherein we: 1) used Pauling electronegativities instead of Sanderson electronegativities; 2) defined the lanthanides and actinides as period 6 and 7 rather than 8 and 9; 3) placed the lanthanides and actinides in a fictitious group 19; 4) used van der Waals radius instead of the covalent radius defined by Cordero and coworkers[79]; 5) removed the atomic volume feature; 6) removed the electron affinity feature. Tabulated values were taken from mendeleev v.0.5.2.[80] Functionally, we found that this process has no apparent change in the performance of the CGCNN models developed in this work, likely because the node vectors are iteratively optimized during the model training process. Nonetheless, the changes were retained. The edge (i.e. bond) feature vectors contain the bond distance between nodes, as in the original CGCNN work.[61]

Dimensionality Reduction

Dimensionality reduction was carried out via the uniform manifold approximation and projection (UMAP) algorithm[81] as implemented in umap v.0.4.6.[82] UMAP constructs a weighted graph of a given dataset in the high-dimensional space and then projects this graph to a lower-dimensional (in this case, two-dimensional) space. Each node of the graph represents a data point, with the edges representing the proximity of each pair of data points in the feature space.[81] The number of neighbors was set to 15 (for SOAP) or 50 (for Stoichiometric-120), and the minimum distance between points was set to 0.1 (for SOAP) and 0.4 (for Stoichiometric-120). All other parameters were set to the default values (for reproducibility, a random seed

of 42 was used). For the connectivity map, edge bundling[83] was enabled to help convey the overall structure by allowing edges to curve and then grouping nearby connections.

The SOAP similarity kernel was converted to a distance matrix $\mathbf{D}$ by invoking the following metric:

$$D_{ij} = \sqrt{K_{ii} + K_{jj} - 2K_{ij}} \therefore \mathbf{D} = \sqrt{2 - 2\mathbf{K}} \tag{S10}$$

since the self-similarity scores K_{ii} and K_{jj} are normalized to 1. A Euclidean distance metric was used with the Stoichiometric-120 descriptor to create the distance matrix. Although maximum atomic number, $\max(Z)$, is not a feature in Stoichiometric-120, it can be directly related to the $\mathrm{range}(Z)$ feature since $\min(Z) = 1$ in every MOF. For this reason, we use the more intuitive $\max(Z)$ feature in Figure 6A.

Methodological Comments for Data Reuse

One of the main motivations for developing the QMOF database is to enable the development/evaluation of new machine learning models. In this case, if the goal is to develop a new machine learning algorithm specifically tailored for MOFs, there are a few comments worth considering.

First, if the purpose is to predict the properties of the optimized MOF structures from the un-optimized structures, then it is important to note that some MOFs have small structural changes before and after optimization whereas others may have somewhat significant changes in the lattice constants (e.g. due to solvent removal). This may (or may not) influence the machine learning process, depending on the property of interest and the way in which the MOFs are encoded. For computational simplicity, we have also assumed a high-spin initial guess for the magnetic moments in a similar manner as the current iterations of the OQMD[32] and Materials Project.[84,33] While this initial guess may converge to a spin state quite different from the high-spin initialization, it is inevitable that several open-shell MOFs in the QMOF database are not at their true magnetic ground states. This is especially the case for MOFs with the possibility of antiferromagnetic ordering.

Another aspect to consider is that, while every effort was made to ensure the initial structures were charge-neutral and accurately constructed, it is inevitable that some structures in the database are not pristine. Oftentimes, this can occur for reasons completely outside the control of the workflow shown in Figure S1, such as if some atoms could not be identified experimentally and were never included in the CIF or CSD entry. The most common scenario is likely omitted/additional H atoms, which are particularly difficult to identify if not already specified in the CSD entry. While many of these instances are filtered out either via the workflow in Figure S1 or due to SCF convergence issues when the DFT calculations were performed, additional filtering steps are always possible and are suggested for new applications of interest. Users are encouraged to flag any identified “structural fidelity” issues on the QMOF database GitHub page.[17]

We also note that the definition of a MOF employed by the CSD MOF subset[3] is significantly less strict than that of the CoRE MOF database.[16] Namely, materials in the CSD MOF subset do not need to be porous, and there is no restriction on the dimensionality of the framework itself. This is arguably ideal from a data inclusivity and model generalizability standpoint but will result in a subset of materials that are perhaps better classified as coordination polymers rather than “conventional” MOFs. For training machine learning models to predict electronic structure properties, it is not expected that the presence or lack of pore space would directly influence the properties of interest. Nonetheless, users only interested in more “conventional” MOF structures may wish to filter the QMOF database by pore size and/or framework dimensionality.

Finally, as mentioned in the text surrounding Figure 2A, some types of MOFs are likely to be underrepresented in the current version of the QMOF database. The most apparent cases are MOFs that have undergone post-synthetic modification, MOFs with defects, and MOFs with metal-oxo clusters containing complex proton topologies. Unsupervised learning methods like those presented in this work can be used to determine if new MOFs of interest overlap in feature space with the MOFs in the QMOF database. For instances where there is not significant overlap, we encourage users to supplement the QMOF database with their own structures of interest. For particularly problematic structures in experimental MOF databases and/or those with elements that appear relatively infrequently in the database (e.g. Zr, Hf, and Al MOFs), we encourage the use of hypothetical MOF construction codes (e.g. ToBaCCo[85,86]) to generate “clean” starting structures suitable for DFT screening.

Section B: Supplemental Figures and Tables

Example Flexible MOF

The experimental and theoretical lattice constants for an example flexible MOF in the QMOF database, Fe(bdp) (H_2bdp = 1,4-benzenedipyrazole)[19] (Figure S3), are shown in Table S5.

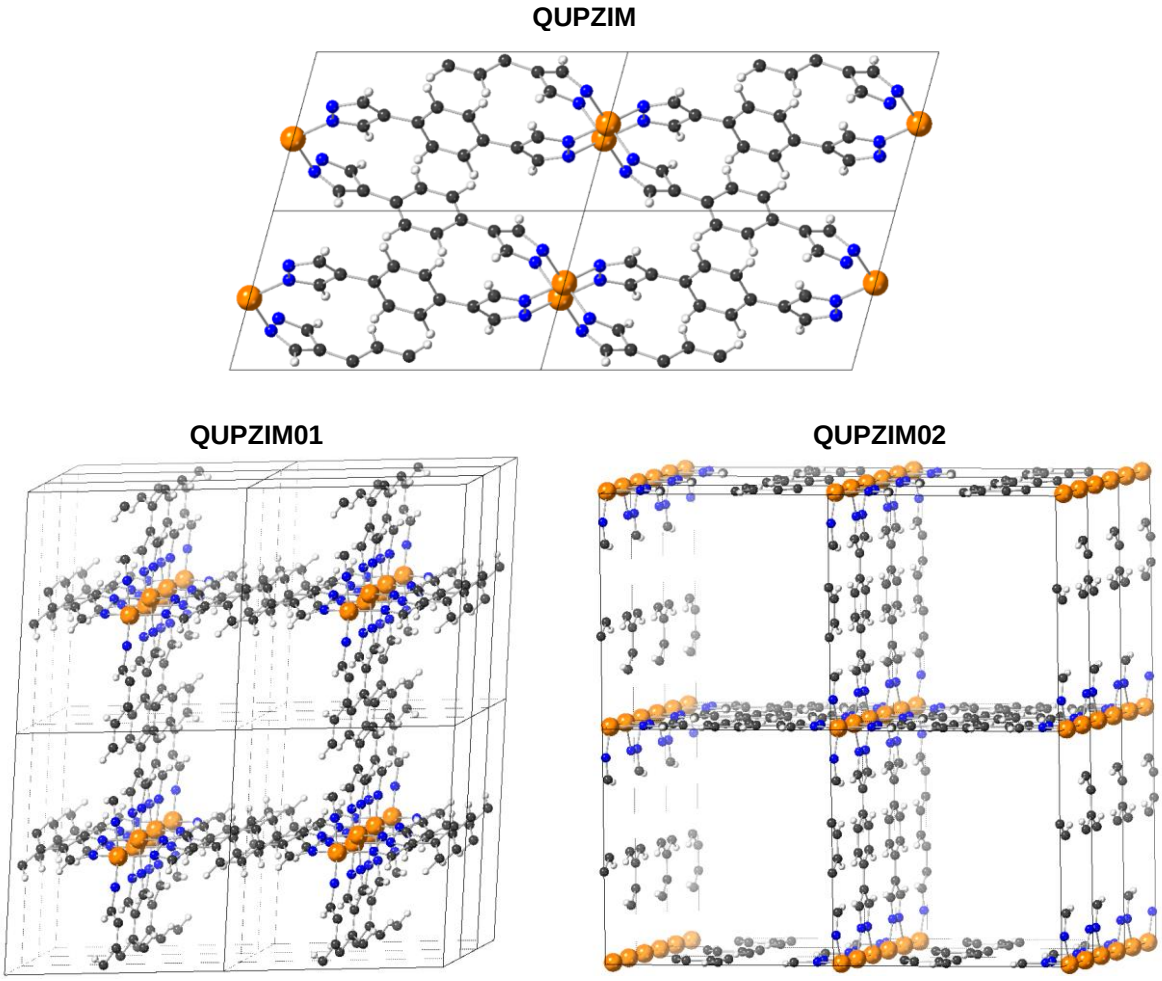


Figure S3. DFT-optimized structures of three different conformations of Fe(bdp) in the QMOF database. Color key: Fe (orange), N (blue), C (gray), H (white).

Table S5. Experimental lattice constants[19] for the flexible MOF Fe(bdp) compared to the PBE-D3(BJ) lattice constants from the QMOF database.

CSD Refcode		a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)
QUPZIM	Exp.	6.89	6.98	13.00	91.6	105.4	90.0
	Theory	6.89	6.55	13.03	91.8	105.3	90.0
QUPZIM01	Exp.	6.95	13.48	13.48	83.2	84.5	84.5
	Theory	5.60	13.68	13.68	87.6	72.3	72.3
QUPZIM02	Exp.	7.20	13.41	13.41	90.0	90.0	90.0
	Theory	7.06	13.47	13.47	90.0	90.0	90.0

Dataset Overview

The average band gaps associated with each element are shown in Figure S4. The partial charges before and after optimization are shown in Figure S5A, indicating that the values do not substantially change upon structure relaxation. We also plot the cumulative frequency of absolute deviations in partial chares before and after structure relaxation in Figure S5B.

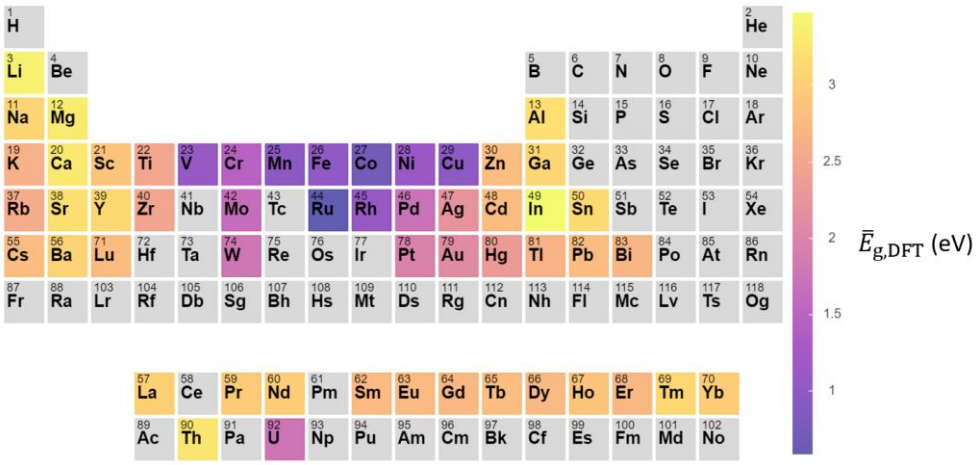


Figure S4. Average DFT-computed band gap at the PBE-D3(BJ) level of theory, $\bar{E}_{g,DFT}$, for MOFs containing a given metal element in the QMOF-14482-opt set. If multiple metal elements are present in a given MOF, the band gap is considered for both elements. Metals with less than 10 entries were excluded.

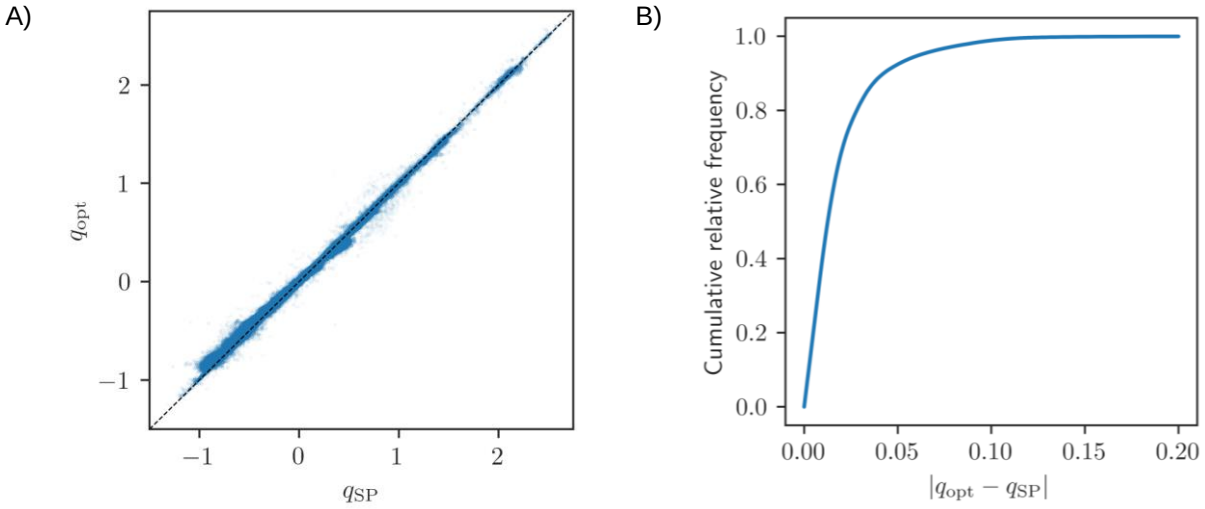


Figure S5. (A) Parity plot comparing the DDEC6 partial atomic charges for the QMOF-14170-opt (q_{opt}) and QMOF-14170-SP (q_{SP}) datasets. (b) The cumulative fraction of DDEC6 partial atomic charges in the QMOF-14170-opt dataset that are within some tolerance, given by $|q_{opt} - q_{SP}|$, of the QMOF-14170-SP dataset.

High-Spin Fe MOFs

MOFs in the QMOF database that contain both high-spin iron species (defined as having a spin density with a magnitude greater than 3.5 based on a DDEC6[35] population analysis) and a pore-limiting diameter greater than 3.6 Å (the kinetic diameter of N_2) following the structure relaxation workflow are shown in Figure S6. These MOFs include: Fe_2(dobdc) (H_4dobdc = 2,5-dihydroxybenzene-1,4-dicarboxylic acid) (refcode: COKNOH)[87] and its expanded pore analogue Fe_2(dobpdc) (H_4dobpdc = 4,4′-dihydroxy-(1,1′-biphenyl)-3,3′-dicarboxylic acid) (refcode: MALSIE),[88] Fe_2Cl_2(bbta) (H_2bbta = 1*H*,5*H*-benzo(1,2-d:4,5-d′)bistriazole) (refcode: HAYYUE)[89] and its expanded pore analogue Fe_2Cl_2(btdd) (H_2btdd = bis(1*H*-1,2,3-triazolo[4,5-*b*],[4′,5′-*i*])dibenzo[1,4]dioxin) (refcode: HAYZAL),[89] Fe(bdp) (refcode: QUPZIM01),[19] and Fe(bpz) (H_2bpz = 4,4′-bipyrazole) (refcode: ACODAA).[90] For brevity, we exclude refcodes that have

identical frameworks but contain guest species in the pores or are different conformations of the same MOF.

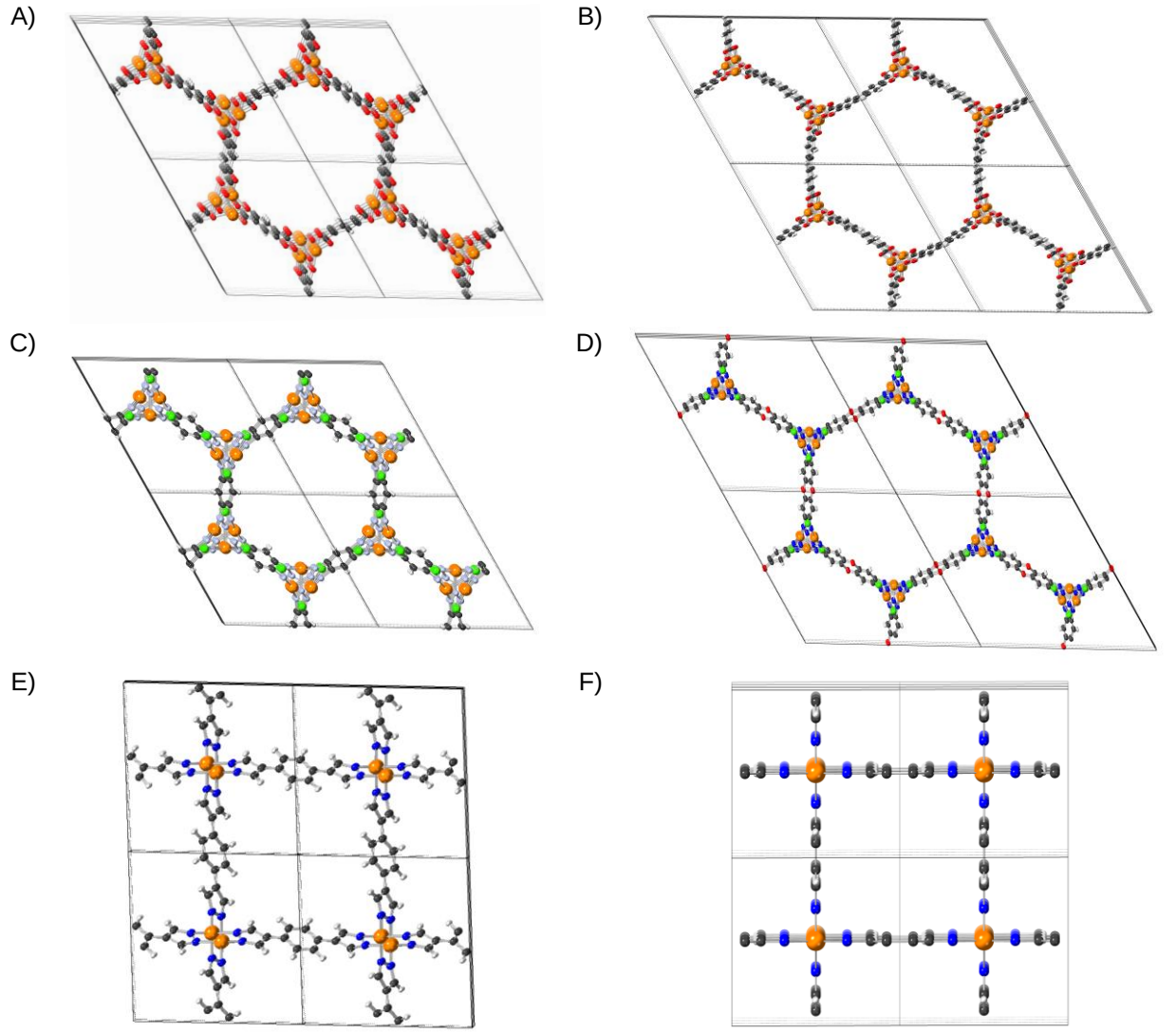


Figure S6. Porous MOFs in the QMOF database with high-spin Fe sites. A) Fe_2(dobdc) (refcode: COKNOH); B) Fe_2(dobpdc) (refcode: MALSIE); C) Fe_2Cl_2(bbta) (refcode: HAYYUE); D) Fe_2Cl_2(btdd) (refcode: HAYZAL) E) Fe(bdp) (refcode: QUPZIM01); F) Fe(bpz) (refcode: ACODAA). Color key: Fe (orange), N (blue), O (red), Cl (green), C (gray), H (white).

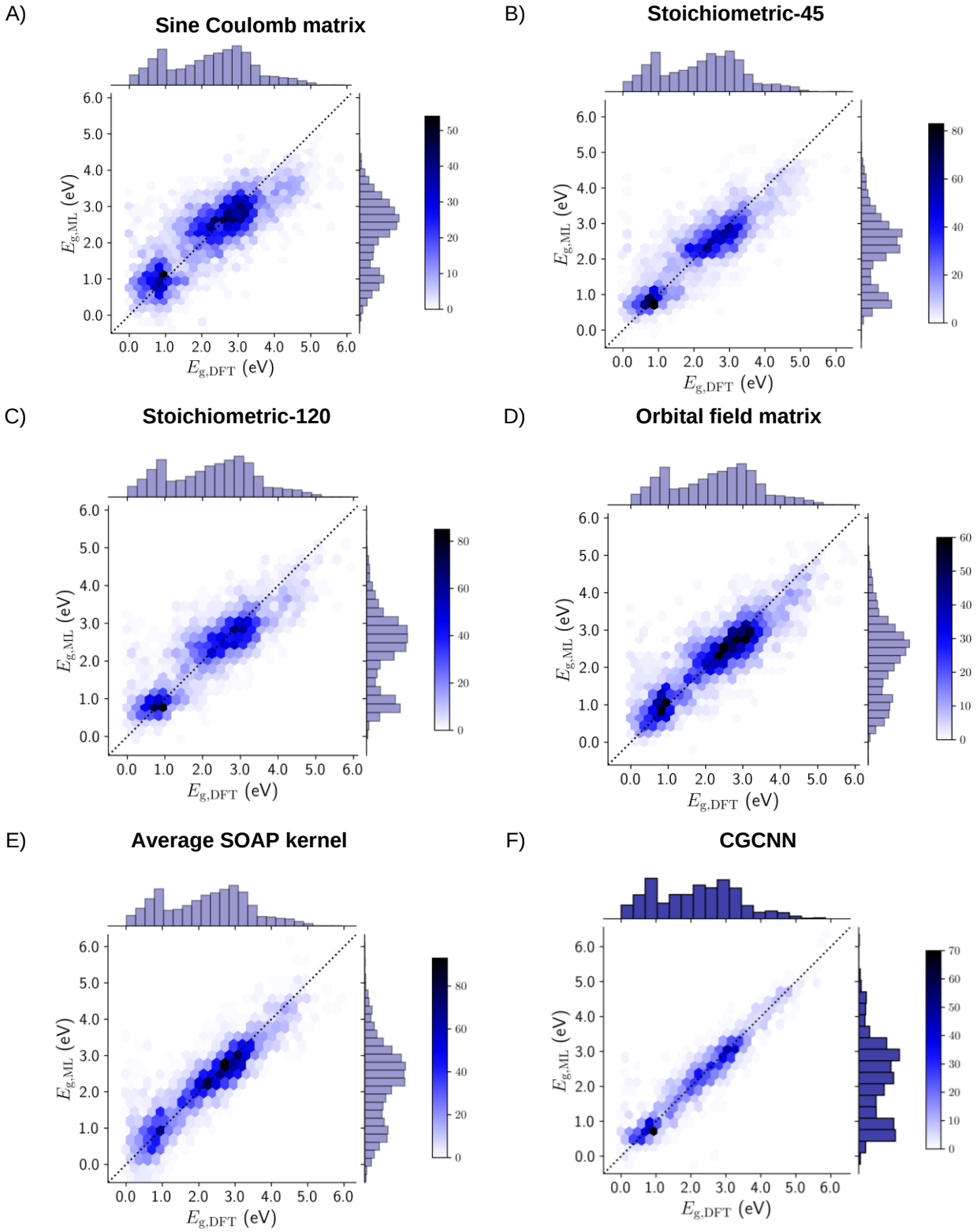


Figure S7. Testing set parity plot for the A) Sine Coulomb matrix, B) Stoichiometric-45, C) Stoichiometric-120, D) Orbital field matrix, E) SOAP, and F) CGCNN machine learning models. The data is presented with hexagonal binning, comparing the machine learning band gaps, $E_{g,ML}$, to the DFT-computed band gaps,

$E_{\mathrm{g,DFT}}$. 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 $E_{\mathrm{g,ML}}$ and $E_{\mathrm{g,DFT}}$ data are displayed parallel to the y- and x-axes, respectively.

Comparing Against ML Band Gap Models for Other Crystalline Materials

It is worth comparing the results of the top-performing ML models in this work against state-of-the-art ML models developed for the band gaps of other crystalline materials in the literature. In the original CGCNN work, the convolutional neural network was able to achieve a testing MAE of 0.39 eV when trained on 16,458 inorganic solids from the OQMD.[61] A different graph network approach – the MatErials Graph Network (MEGNet) – achieved an MAE of 0.38 eV when trained on 36,720 inorganic solids from the Materials Project, which could be reduced to 0.33 eV after transfer learning via a model originally trained on the DFT-computed formation energies of 60,000 inorganic solids.[91] Recently, a global attention graph neural network (GATGNN) achieved MAEs of 0.32 eV and 0.31 eV on the OQMD and Materials Project datasets, respectively.[92] Particularly relevant for the present study, Olsthoorn et al.[93] used a weighted average of SOAP- and SchNet[94]-based ML regression models trained on 10,000 band gaps of organic crystals in the Organic Materials Database (OMDB)[95] to achieve a testing MAE of 0.39 eV. The band gaps in the OMDB work were based on single-point calculations of the as-deposited crystal structures, as geometry optimizations were not carried out. In addition, organic crystals with non-zero net magnetic moments were not considered in the OMDB work.

Additional UMAP Results

It is also worth investigating the degree of overlap in feature space between the QMOF-14482 dataset and the parent QMOF-42349 dataset that the former was drawn from. To carry out this analysis, we used UMAP to project the feature space of the QMOF-42349 dataset to two dimensions. Again, we used the Stoichiometric-120 descriptor to featurize each material. We then highlighted the subset of materials that are also present in the QMOF-14482 dataset. As shown in Figure S8, there is significant overlap between the two datasets, such that we can expect ML models trained on the QMOF-14482 dataset to be applicable to other MOFs deposited in the CSD. We carried out a similar analysis to compare the QMOF-14482 dataset with the CoRE MOF 2019 (v.1.1.3) database.[16] For this purpose, we use the CoRE MOF 2019 database with free solvent removed (i.e. “FSR”), including both the “public” and “internal” subsets but excluding structures flagged as having disorder that could not be refined. Significant overlap in the reduced feature space was observed between both databases (Figure S9).

Figures S10 and S11 show Stoichiometric-120- and SOAP-based UMAPs generated for the QMOF-14482 dataset, respectively, but with edge connections shown to highlight the connectivity between different local regions in the projection. This is particularly notable for the Stoichiometric-120 UMAP, which shows that the individual clusters are connected in sequential order of $\max(Z)$ (Figure 6A, Figure S10). The edge connections for the SOAP-based UMAP also makes the local regions, and their connectivity, clearer.

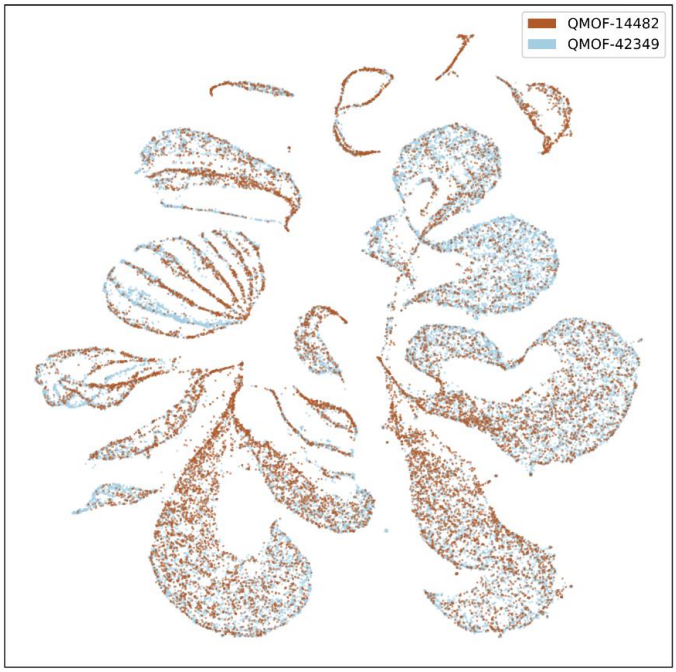


Figure S8. Structural dimensionality reduction performed using UMAP, with a distance matrix obtained from the Euclidean distance of the Stoichiometric-120 encodings for the structures in the QMOF-42349 dataset. The QMOF-14482 subset is overlaid onto the projection.

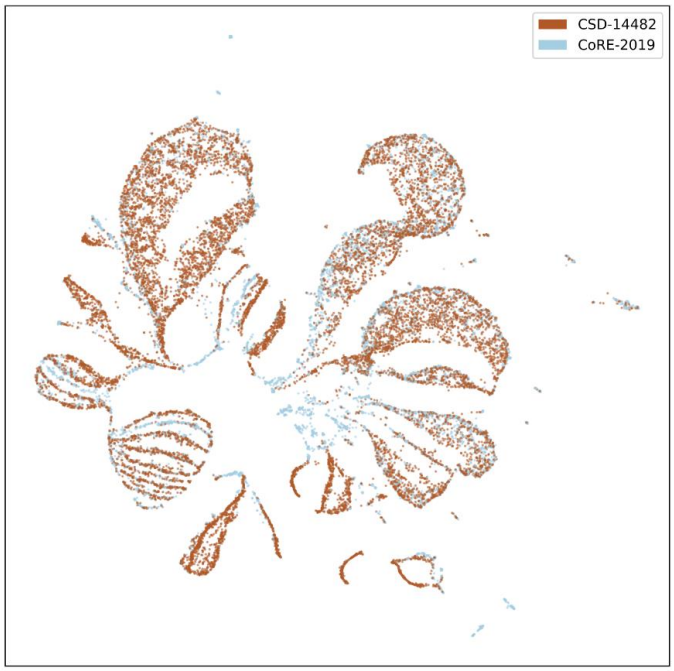


Figure S9. Structural dimensionality reduction performed using UMAP, with a distance matrix obtained from the Euclidean distance of the Stoichiometric-120 encodings for the structures in the QMOF-14482 and CoRE MOF 2019 (free solvent removed) databases.

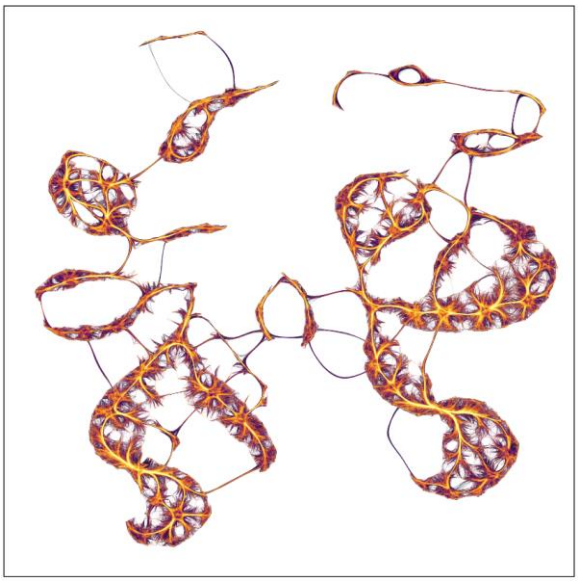


Figure S10. Structural dimensionality reduction performed using UMAP, with a distance matrix obtained from the Euclidean distance of the Stoichiometric-120 encodings of the structures in the QMOF-14482 dataset. The connectivity between points is shown. Brighter colors indicate a greater density of connections.

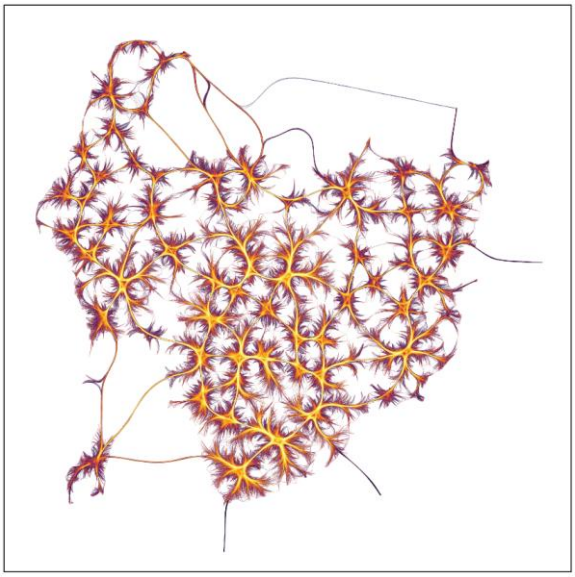


Figure S11. Structural dimensionality reduction performed using UMAP, with a distance matrix obtained from the average SOAP similarity kernel of the (unrelaxed) structures in the QMOF-14482 dataset. The connectivity between points is shown. Brighter colors indicate a greater density of connections.

Electronic Structure of GUTYAW

The projected density of states at the HSE06-D3(BJ) level of theory for $Sr[C_2H_4(SO_3)_2]$ (refcode: GUTYAW[96]) is shown in Figure S12.

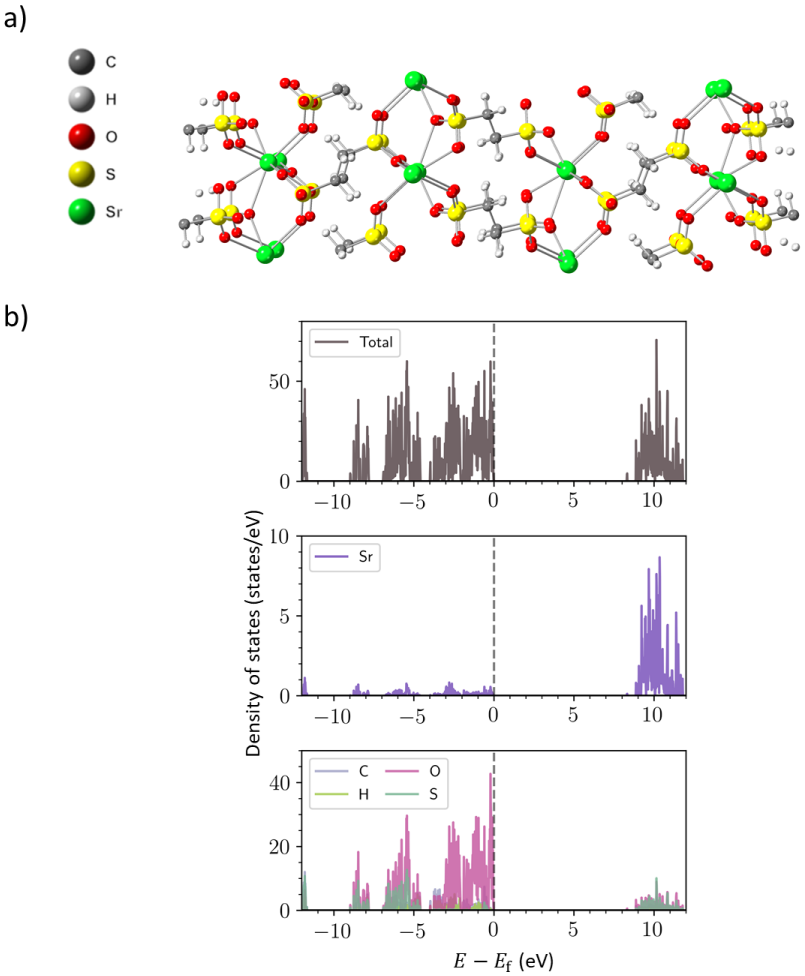


Figure S12. a) Structure of $Sr[C_2H_4(SO_3)_2]$. b) Total and projected density of states. The energy, E, is shown with respect to the Fermi level, E_f.

Limitations of Averaging Schemes

For any featurization method, there are inevitable limitations with how a given set of materials is encoded for machine learning. In the case of an average SOAP kernel, for instance, one limitation is that every atomic environment is weighted equally. To highlight why this may be imperfect, we show the average SOAP similarity kernel for IRMOF-1 with the formula $Zn_4O(bdc)_3$ (bdc = benzene-1,4-dicarboxylate), IRMOF-2 with the formula $Zn_4O(bdc\text{-}Br)_3$, IRMOF-10 with the formula $Zn_4O(bpdc)$ (bpdc = 4,4′-biphenyldicarboxylate), $Zn_2(dobdc)$ (dobdc = 2,5-dihydroxybenzene-1,4-dicarboxylate), and MFU-4*l* (MFU = Metal-Organic Framework Ulm University, *l* = large) with the formula $Zn_5Cl_4(btdd)_3$ (btdd = bis(1,2,3-triazolato-[4,5-b],[4′,5′-i])dibenzo-[1,4]-dioxin)) (Figure S13). While IRMOF-1 and the functionalized analogue IRMOF-2 have nearly identical averaged SOAP features, IRMOF-1 and the elongated analogue IRMOF-10 are quite different (Table S6). In fact, IRMOF-1 and $Zn_2(dobdc)$ are more similar than IRMOF-1 and IRMOF-2 based on the average SOAP kernel (Table S6). This can likely be traced back to the similarity of the linkers (bdc vs. dobdc) in IRMOF-1 and $Zn_2(dobdc)$ despite their very different inorganic nodes (Zn_4O vs. isolated Zn sites) and metal coordination environments (tetrahedral vs. square pyramidal). Weighting the structural similarity of the inorganic nodes and organic linkers by different factors is one approach that

may resolve this issue, aside from trying alternate kernel methods such as the computationally more expensive regularized entropy match (REMatch) kernel.[74] This phenomenon is also expected to limit the performance of other methods that involve simple averaging over a structure, such as the orbital field matrix. With the development of a database of DFT-computed MOF properties, there is a rich opportunity for exploring featurization methods that are constructed specifically for the robust and accurate representation of MOFs.

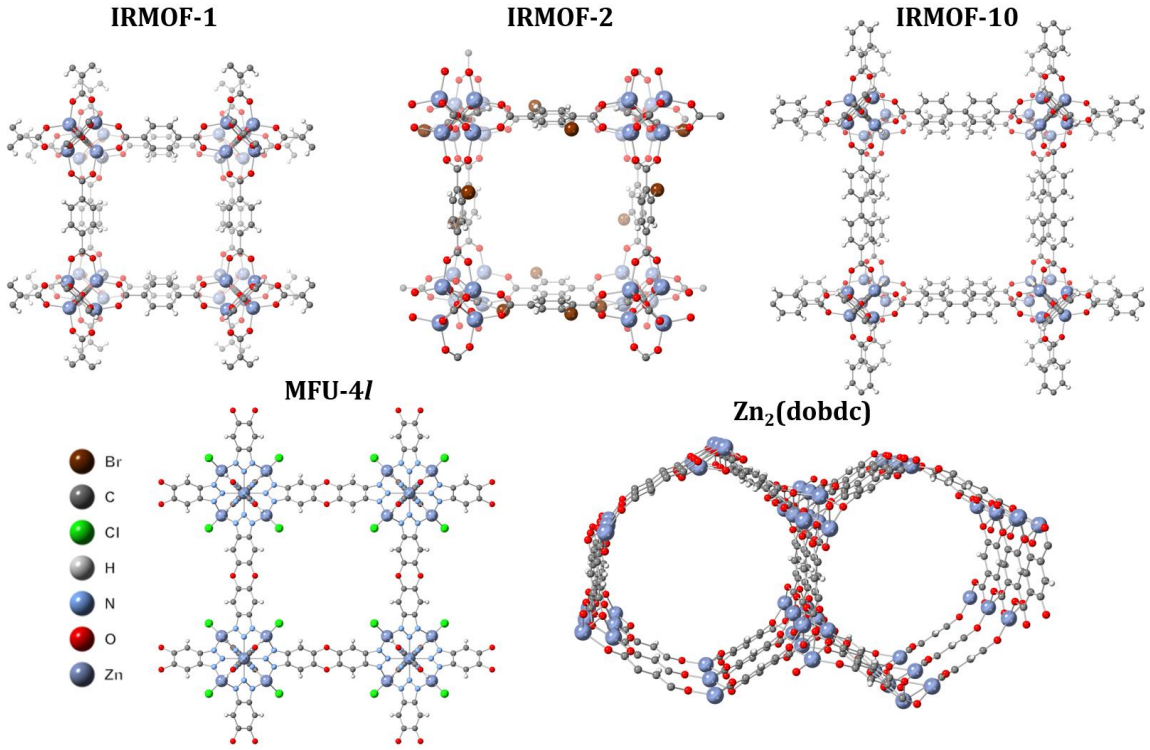


Figure S13. Crystal structures of IRMOF-1, IRMOF-2, IRMOF-10, Zn_2(dobdc) (also known as Zn-MOF-74 and Zn-CPO-27), and MFU-4*l*.

Table S6. Average (normalized) SOAP similarity kernel for IRMOF-1, IRMOF-2, IRMOF-10, Zn_2(dobdc), and MFU-4*l*. Here, $r_{\text{cut}} = 4$ Å, $\sigma = 0.1$ Å, $\xi = 2$, and $n_{\text{max}} = \ell_{\text{max}} = 9$.

	IRMOF-1	IRMOF-2	IRMOF-10	Zn_2(dobdc)	MFU-4*l*
IRMOF-1	1.00	0.98	0.52	0.92	0.73
IRMOF-2	0.98	1.00	0.43	0.94	0.72
IRMOF-10	0.52	0.43	1.00	0.29	0.35
Zn_2(dobdc)	0.92	0.94	0.29	1.00	0.64
MFU-4*l*	0.73	0.72	0.35	0.64	1.00

Supplemental References

(1) Ong, S.P., Richards, W.D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V.L., Persson, K.A., and Ceder, G. (2013). Python Materials Genomics (Pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Comput. Mater. Sci. 68*, 314–319.

(2) Bruno, I.J., Cole, J.C., Edgington, P.R., Kessler, M., Macrae, C.F., McCabe, P., Pearson, J., and Taylor, R. (2002). New Software for Searching the Cambridge Structural Database and Visualizing Crystal Structures. *Acta Crystallogr. Sect. B 58*, 389–397.

(3) Moghadam, P.Z., Li, A., Wiggin, S.B., Tao, A., Maloney, A.G.P., Wood, P.A., Ward, S.C., and Fairen-Jimenez, D. (2017). Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future. *Chem. Mater. 29*, 2618–2625.

(4) Altintas, C., Avci, G., Daglar, H., Azar, A.N.V., Erucar, I., Velioglu, S., and Keskin, S. (2019). 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 7*, 9593–9608.

(5) Velioglu, S., and Keskin, S. (2020). Revealing the Effect of Structure Curations on the Simulated CO_2 Separation Performances of MOFs. *Mater. Adv. 1*, 341–353.

(6) Chen, T., and Manz, T.A. (2020). Identifying Misbonded Atoms in the 2019 CoRE Metal–Organic Framework Database. *RSC Adv. 10*, 26944–26951.

(7) Jouffret, L., Rivenet, M., and Abraham, F. (2011). Linear Alkyl Diamine-Uranium-Phosphate Systems: U(VI) to U(IV) Reduction with Ethylenediamine. *Inorg. Chem. 50*, 4619–4626.

(8) Feng, P., Bu, X., and Stucky, G.D. (1997). Hydrothermal Syntheses and Structural Characterization of Zeolite Analogue Compounds Based on Cobalt Phosphate. *Nature 388*, 735–741.

(9) Liang, L., Zhang, R., Zhao, J., Liu, C., and Weng, N.S. (2016). Two Actinide-Organic Frameworks Constructed by a Tripodal Flexible Ligand: Occurrence of Infinite $\{(UO_2)O_2(OH)_3\}_{4n}$ and Hexanuclear $\{Th_6O_4(OH)_4\}$ Motifs. *J. Solid State Chem. 243*, 50–56.

(10) Eubank, J.F., Nouar, F., Luebke, R., Cairns, A.J., Wojtas, L., Alkordi, M., Bousquet, T., Hight, M.R., Eckert, J., Embs, J.P., Georgiev, P.A., and Eddaoudi, M. (2012). On Demand: The Singular rht Net, an Ideal Blueprint for the Construction of a Metal–Organic Framework (MOF) Platform. *Angew. Chem. 124*, 10246–10250.

(11) Planas, N., Mondloch, J.E., Tussupbayev, S., Borycz, J., Gagliardi, L., Hupp, J.T., Farha, O.K., and Cramer, C.J. (2014). Defining the Proton Topology of the Zr_6-Based Metal–Organic Framework NU-1000. *J. Phys. Chem. Lett. 5*, 3716–3723.

(12) Li, A., Bueno-Perez, R., Wiggin, S., and Fairen-Jimenez, D. (2020). Enabling Efficient Exploration of Metal–Organic Frameworks in the Cambridge Structural Database. *CrystEngComm 22*, 7152–7161.

(13) Zimmermann, N.E.R., and Jain, A. (2020). Local Structure Order Parameters and Site Fingerprints for Quantification of Coordination Environment and Crystal Structure Similarity. *RSC Adv. 10*, 6063–6081.

(14) Pan, H., Ganose, A.M., Horton, M., Aykol, M., Persson, K.A., Zimmermann, N.E.R., and Jain, A. (2020). Benchmarking Coordination Number Prediction Algorithms on Inorganic Crystal Structures. *Inorg. Chem. 60*, 1590–1603.

(15) Queen, W.L., Bloch, E.D., Brown, C.M., Hudson, M.R., Mason, J.A., Murray, L.J., Ramirez-Cuesta, A.J., Peterson, V.K., and Long, J.R. (2012). Hydrogen Adsorption in the Metal–Organic Frameworks Fe_2(dobdc) and $Fe_2(O_2)$(dobdc). *Dalt. Trans. 41*, 4180–4187.

(16) 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., and Snurr, R.Q. (2019). Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal−Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data 64*, 5985–5998.

(17) QMOF Database. https://github.com/arosen93/QMOF, https://dx.doi.org/10.6084/m9.figshare.13147324.

(18) 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., and Snurr, R.Q. (2019). Identification Schemes for Metal–Organic Frameworks to Enable Rapid Search and Cheminformatics Analysis. *Cryst. Growth Des. 19*, 6682–6697.

(19) 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., and Long, J.R.

(2015). Methane Storage in Flexible Metal−Organic Frameworks with Intrinsic Thermal Management. *Nature 527*, 357–361.
(20) Ling, S., and Slater, B. (2015). Unusually Large Band Gap Changes in Breathing Metal–Organic Framework Materials. *J. Phys. Chem. C 119*, 16667–16677.
(21) Kresse, G., and Furthmüller, J. (1996). Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B 54*, 11169–11186.
(22) Kresse, G., and Joubert, D. (1999). From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B 59*, 1758–1775.
(23) Perdew, J.P., Burke, K., and Ernzerhof, M. (1996). Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett. 77*, 3865–3868.
(24) Grimme, S., Antony, J., Ehrlich, S., and Krieg, H. (2010). A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H−Pu. *J. Chem. Phys. 132*, 154104.
(25) Grimme, S., Ehrlich, S., and Goerigk, L. (2011). Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem. 32*, 1456–1465.
(26) Nazarian, D., Ganesh, P., and Sholl, D.S. (2015). Benchmarking Density Functional Theory Predictions of Framework Structures and Properties in a Chemically Diverse Test Set of Metal−Organic Frameworks. *J. Mater. Chem. A 3*, 22432–22440.
(27) Formalik, F., Fischer, M., Rogacka, J., Firlej, L., and Kuchta, B. (2018). Benchmarking of GGA Density Functionals for Modeling Structures of Nanoporous, Rigid and Flexible MOFs. *J. Chem. Phys. 149*, 064110.
(28) Rosen, A.S., Notestein, J.M., and Snurr, R.Q. (2019). Identifying Promising Metal−Organic Frameworks for Heterogeneous Catalysis via High-Throughput Periodic Density Functional Theory. *J. Comput. Chem. 40*, 1305–1318.
(29) Blöchl, P.E. (1994). Projector Augmented-Wave Method. *Phys. Rev. B 50*, 17953–17979.
(30) Pulay, P. (1980). Convergence Acceleration of Iterative Sequences. The Case of SCF Iteration. *Chem. Phys. Lett. 73*, 393–398.
(31) Kresse, G., and Furthmüller, J. (1996). Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Comput. Mater. Sci. 6*, 15–50.
(32) Saal, J.E., Kirklin, S., Aykol, M., Meredig, B., and Wolverton, C. (2013). Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *Jom 65*, 1501–1509.
(33) Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., and Persson, K.A. (2013). The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater. 1*, 11002.
(34) Larsen, A., Mortensen, J., Blomqvist, J., Castelli, I., Christensen, R., Dulak, M., Friis, J., Groves, M., Hammer, B., Hargus, C., Hermes, E., Jennings, P., Jensen, P., Kermode, J., Kitchin, J., Kolsbjerg, E., Kubal, J., Kaasbjerg, K., Lysgaard, S., Maronsson, J., Maxson, T., Olsen, T., Pastewka, L., Peterson, A., Rostgaard, C., Schiøtz, J., Schütt, O., Strange, M., Thygesen, K., Vegge, T., Vilhelmsen, L., Walter, M., Zeng, Z., and Jacobsen, K. (2017). The Atomic Simulation Environment—A Python Library for Working with Atoms. *J. Phys. Condens. Matter 29*, 273002.
(35) Manz, T.A., and Limas, N.G. (2016). Introducing DDEC6 Atomic Population Analysis: Part 1. Charge Partitioning Theory and Methodology. *RSC Adv. 6*, 47771–47801.
(36) Limas, N.G., and Manz, T.A. (2016). Introducing DDEC6 Atomic Population Analysis: Part 2. Computed Results for a Wide Range of Periodic and Nonperiodic Materials. *RSC Adv. 6*, 45727–45747.
(37) Manz, T.A. (2017). Introducing DDEC6 Atomic Population Analysis: Part 3. Comprehensive Method to Compute Bond Orders. *RSC Adv. 7*, 45552–45581.
(38) Limas, N.G., and Manz, T.A. (2018). Introducing DDEC6 Atomic Population Analysis: Part 4. Efficient Parallel Computation of Net Atomic Charges, Atomic Spin Moments, Bond Orders, and More. *RSC Adv. 8*, 2678–2707.
(39) Manz, T.A., and Gabaldon Limas, N. Chargemol program for performing DDEC analysis http://ddec.sourceforge.net/.
(40) Marenich, A. V, Jerome, S. V, Cramer, C.J., and Truhlar, D.G. (2012). Charge Model 5: An Extension of Hirshfeld Population Analysis for the Accurate Description of Molecular Interactions in Gaseous and Condensed Phases. *J. Chem. Theory Comput. 8*, 527–541.

(41) Rosen, A.S., Notestein, J.M., and Snurr, R.Q. PyMOFScreen https://doi.org/10.5281/zenodo.1451873.
(42) Bitzek, E., Koskinen, P., Gähler, F., Moseler, M., and Gumbsch, P. (2006). Structural Relaxation Made Simple. *Phys. Rev. Lett. 97*, 170201.
(43) Heyd, J., Scuseria, G.E., and Ernzerhof, M. (2003). Hybrid Functionals Based on a Screened Coulomb Potential. *J. Chem. Phys. 118*, 8207–8215.
(44) Krukau, A. V, Vydrov, O.A., Izmaylov, A.F., and Scuseria, G.E. (2006). Influence of the Exchange Screening Parameter on the Performance of Screened Hybrid Functionals. *J. Chem. Phys. 125*, 224106.
(45) Moellmann, J., and Grimme, S. (2014). DFT-D3 Study of Some Molecular Crystals. *J. Phys. Chem. C 118*, 7615–7621.
(46) Freysoldt, C., Boeck, S., and Neugebauer, J. (2009). Direct Minimization Technique for Metals in Density Functional Theory. *Phys. Rev. B 79*, 241103.
(47) Teter, M.P., Payne, M.C., and Allan, D.C. (1989). Solution of Schrödinger's Equation for Large Systems. *Phys. Rev. B 40*, 12255–12263.
(48) Bylander, D.M., Kleinman, L., and Lee, S. (1990). Self-Consistent Calculations of the Energy Bands and Bonding Properties of B12C3. *Phys. Rev. B 42*, 1394–1403.
(49) Clements, J.E., Price, J.R., Neville, S.M., and Kepert, C.J. (2014). Perturbation of Spin Crossover Behavior by Covalent Post-Synthetic Modification of a Porous Metal–Organic Framework. *Angew. Chem. Int. Ed. 126*, 10328–10332.
(50) Manna, S.C., Zangrando, E., Ribas, J., and Chaudhuri, N.R. (2005). Squarato-Bridged Polymeric Networks of Iron(II) with N-Donor Coligands: Syntheses, Crystal Structures and Magnetic Properties. *Inorganica Chim. Acta 358*, 4497–4504.
(51) Sekine, Y., Tonouchi, M., Yokoyama, T., Kosaka, W., and Miyasaka, H. (2017). Built-in TTF–TCNQ Charge-Transfer Salts in π-Stacked Pillared Layer Frameworks. *CrystEngComm 19*, 2300–2304.
(52) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res. 12*, 2825–2830.
(53) Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Ilhan, P., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbgret, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods 17*, 261–272.
(54) Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T.E. (2020). Array Programming with NumPy. *Nature 585*, 357–362.
(55) Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng. 9*, 90–95.
(56) McKinney, W. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*; Vol. 445, pp 51–56.
(57) Seaborn http://doi.org/10.5281/zenodo.592845.
(58) Himanen, L., Jäger, M.O.J., Morooka, E. V, Canova, F.F., Ranawat, Y.S., Gao, D.Z., Rinke, P., and Foster, A.S. (2020). DScribe: Library of Descriptors for Machine Learning in Materials Science. *Comput. Phys. Commun. 247*, 106949.
(59) Meredig, B., Agrawal, A., Kirklin, S., Saal, J.E., Doak, J.W., Thompson, A., Zhang, K., Choudhary, A., and Wolverton, C. (2014). Combinatorial Screening for New Materials in Unconstrained Composition Space with Machine Learning. *Phys. Rev. B 89*, 94104.
(60) Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N.E.R., Bajaj, S., Wang, Q., Montoya, J., Chen, J., Bystrom, K., Dylla, M., Chard, K., Asta, M., Persson, K.A., Snyder, J., Foster, I., and Jain, A. (2018). Matminer: An Open Source Toolkit for Materials Data Mining. *Comput. Mater. Sci. 152*, 60–69.
(61) Xie, T., and Grossman, J.C. (2018). Crystal Graph Convolutional Neural Networks for an Accurate

and Interpretable Prediction of Material Properties. *Phys. Rev. Lett. 120*, 145301.
(62) Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*; pp 8024–8035.
(63) Xie, T., and Grossman, J. Crystal Graph Convolutional Neural Networks https://github.com/txie-93/cgcnn.
(64) Rosen, A.S., Notestein, J.M., and Snurr, R.Q. Crystal Graph Convolutional Neural Networks https://github.com/snurr-group/cgcnn.
(65) Rosen, A.S., Notestein, J.M., and Snurr, R.Q. Crystal Graph Convolutional Neural Networks https://github.com/snurr-group/cgcnn/tree/scatter.
(66) Rosen, A.S. PTable Trends.
(67) Willems, T.F., Rycroft, C.H., Kazi, M., Meza, J.C., and Haranczyk, M. (2012). Algorithms and Tools for High-Throughput Geometry-Based Analysis of Crystalline Porous Materials. *Microporous Mesoporous Mater. 149*, 134–141.
(68) Herath, U., Tavadze, P., He, X., Bousquet, E., Singh, S., Munoz, F., and Romero, A.H. (2020). PyProcar: A Python Library for Electronic Structure Pre/Post-Processing. *Comput. Phys. Commun. 251*, 107080.
(69) Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT press.
(70) Kirklin, S., Saal, J.E., Meredig, B., Thompson, A., Doak, J.W., Aykol, M., Rühl, S., and Wolverton, C. (2015). The Open Quantum Materials Database (OQMD): Assessing the Accuracy of DFT Formation Energies. *npj Comput. Mater. 1*, 15010.
(71) He, Y., Cubuk, E.D., Allendorf, M.D., and Reed, E.J. (2018). Metallic Metal–Organic Frameworks Predicted by the Combination of Machine Learning Methods and Ab Initio Calculations. *J. Phys. Chem. Lett. 9*, 4562–4569.
(72) Faber, F., Lindmaa, A., von Lilienfeld, O.A., and Armiento, R. (2015). Crystal Structure Representations for Machine Learning Models of Formation Energies. *Int. J. Quantum Chem. 115*, 1094–1101.
(73) Lam Pham, T., Kino, H., Terakura, K., Miyake, T., Tsuda, K., Takigawa, I., and Chi Dam, H. (2017). Machine Learning Reveals Orbital Interaction in Materials. *Sci. Technol. Adv. Mater. 18*, 756–765.
(74) De, S., Bartók, A.P., Csányi, G., and Ceriotti, M. (2016). Comparing Molecules and Solids across Structural and Alchemical Space. *Phys. Chem. Chem. Phys. 18*, 13754–13769.
(75) Musil, F., De, S., Yang, J., Campbell, J.E., Day, G.M., and Ceriotti, M. (2018). Machine Learning for the Structure–Energy–Property Landscapes of Molecular Crystals. *Chem. Sci. 9*, 1289–1300.
(76) Noh, J., Gu, G.H., Kim, S., and Jung, Y. (2020). Uncertainty-Quantified Hybrid Machine Learning/Density Functional Theory High Throughput Screening Method for Crystals. *J. Chem. Inf. Model. 60*, 1996–2003.
(77) Park, C.W., and Wolverton, C. (2019). Developing an Improved Crystal Graph Convolutional Neural Network Framework for Accelerated Materials Discovery. *arXiv. arXiv:1906.05267*.
(78) Issue #2: atom_init.json https://github.com/txie-93/cgcnn/issues/2.
(79) Cordero, B., Gómez, V., Platero-Prats, A.E., Revés, M., Echeverría, J., Cremades, E., Barragán, F., and Alvarez, S. (2008). Covalent Radii Revisited. *Dalt. Trans.* No. 21, 2832–2838.
(80) Mentel, Ł. Mendeleev - A Python Resource for Properties of Chemical Elements, Ions and Isotopes. 2014.
(81) McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv. arXiv:1802.03426*.
(82) Leland, M., John, H., Nathaniel, S., and Lukas, G. (2018). UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw. 3*, 861.
(83) Hurter, C., Ersoy, O., and Telea, A. Graph Bundling by Kernel Density Estimation. In *Computer Graphics Forum*; Vol. 31, pp 865–874.
(84) Jain, A., Hautier, G., Moore, C.J., Ping Ong, S., Fischer, C.C., Mueller, T., Persson, K.A., and Ceder, G. (2011). A High-Throughput Infrastructure for Density Functional Theory Calculations. *Comput. Mater. Sci. 50*, 2295–2310.
(85) Colón, Y.J., Gómez-Gualdrón, D.A., and Snurr, R.Q. (2017). Topologically Guided, Automated Construction of Metal–Organic Frameworks and Their Evaluation for Energy-Related Applications. *Cryst. Growth Des. 17*, 5801–5810.

(86) Anderson, R., and Gómez-Gualdrón, D.A. (2019). Increasing Topological Diversity during Computational “Synthesis” of Porous Crystals: How and Why. *CrystEngComm 21*, 1653–1665.
(87) 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., and Brown, C.M. (2014). Comprehensive Study of Carbon Dioxide Adsorption in the Metal−Organic Frameworks M_2(dobdc)(M = Mg, Mn, Fe, Co, Ni, Cu, Zn). *Chem. Sci. 5*, 4569–4581.
(88) Gygi, D., Bloch, E.D., Mason, J.A., Hudson, M.R., Gonzalez, M.I., Siegelman, R.L., Darwish, T.A., Queen, W.L., Brown, C.M., and Long, J.R. (2016). Hydrogen Storage in the Expanded Pore Metal−Organic Frameworks M_2(DOBPDC) (M = Mg, Mn, Fe, Co, Ni, Zn). *Chem. Mater. 28*, 1128–1138.
(89) Reed, D.A., Keitz, B.K., Oktawiec, J., Mason, J.A., Runčevski, T., Xiao, D.J., Darago, L.E., Crocellà, V., Bordiga, S., and Long, J.R. (2017). A Spin Transition Mechanism for Cooperative Adsorption in Metal−Organic Frameworks. *Nature 550*, 96–100.
(90) Spirkl, S., Grzywa, M., Reschke, S., Fischer, J.K.H., Sippel, P., Demeshko, S., von Nidda, H.-A., and Volkmer, D. (2017). 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. 56*, 12337–12347.
(91) Chen, C., Ye, W., Zuo, Y., Zheng, C., and Ong, S.P. (2019). Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater. 31*, 3564–3572.
(92) Louis, S.-Y.M., Zhao, Y., Nasiri, A., Wang, X., Song, Y., Liu, F., and Hu, J. (2020). Graph Convolutional Neural Networks with Global Attention for Improved Materials Property Prediction. *Phys. Chem. Chem. Phys. 22*, 18141–18148.
(93) Olsthoorn, B., Geilhufe, R.M., Borysov, S.S., and Balatsky, A. V. (2019). Band Gap Prediction for Large Organic Crystal Structures with Machine Learning. *Adv. Quantum Technol. 2*, 1900023.
(94) Schütt, K.T., Sauceda, H.E., Kindermans, P.-J., Tkatchenko, A., and Müller, K.-R. (2018). SchNet – A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys. 148*, 241722.
(95) Borysov, S.S., Geilhufe, R.M., and Balatsky, A. V. (2017). Organic Materials Database: An Open-Access Online Database for Data Mining. *PLoS One 12*, e0171501.
(96) Salami, T.O., Patterson, S.N., Jones, V.D., Masello, A., and Abboud, K.A. (2009). 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. 12*, 1150–1153.