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

Andrew S. Rosen,¹,* Shaelyn M. Iyer,¹ Debmalya Ray,² Zhenpeng Yao,³ Alán Aspuru-Guzik,³,⁴,⁵ Laura Gagliardi,⁶ Justin M. Notestein,¹ Randall Q. Snurr¹

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

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

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

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

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

High-throughput computational screening approaches based on classical simulations have proven extremely useful for more efficiently exploring the vast combinatorial space of MOF structures. Recently, the large quantities of data generated during these computational screening studies have led to the development of machine learning (ML) models that can accelerate the MOF design and discovery process even further. ML-assisted screening studies have been successfully applied to the discovery of MOFs suitable for H₂ storage, CO₂ separation/capture, and numerous other applications predominantly in the area of gas storage and separations. Nonetheless, similar efforts remain almost entirely unexplored for the countless applications governed by quantum chemistry, such as those that are based on the electronic, optical, magnetic, and/or catalytic properties of MOFs. Beyond the sheer number of possible MOFs that can be realized, the large number of atoms in MOF crystal structures often makes it computationally demanding to carry out even moderate-scale quantum-chemical screening studies, further magnifying the need for ML approaches in this area.

To date, the most relevant studies focused on training ML models to predict the quantum-chemical properties of MOFs are those of Raza et al. and Korolev et al. who independently developed ML models that can predict the partial atomic charges of MOFs in the Computation-Ready, Experimental (CoRE) MOF database. 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. who trained binary classification models to predict whether inorganic solids in the Open Quantum Materials Database (OQMD) are metallic or nonmetallic. Without retraining on MOF data, a multi-model voting procedure was then used to predict the metallic or nonmetallic behavior of 2932 MOFs in the CoRE MOF database, which do not have computed band gaps. Of the six identified materials with near-zero band gaps at the PBE level of theory, 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 [Cd(CN)₂]₄ that should actually be [Cd(CN)₃]₂.

In stark contrast with the existing literature on MOFs, significant progress has been made in the development of ML models to aid in the quantum-chemical screening of a wide range of inorganic and molecular compounds. One of the fundamental features underlying much of this work has been the use of high-throughput density functional theory (DFT) workflows to construct large-scale electronic structure property databases, such as those developed for inorganic solids and molecular systems. 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, superhard inorganic materials, and thermally conductive polymers, among many others. With this in mind, there is a significant need for an analogous database of DFT-computed material properties for MOFs so that new ML models can be developed to rapidly predict their electronic structure properties. High-
throughput screening, database generation, and subsequent ML model development are crucial components for realizing the full potential of reticular chemistry\textsuperscript{53} and accelerating materials discovery in general.\textsuperscript{54–57}

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

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

Results

Generation and Overview of the QMOF Database

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

We considered the list of materials identified as MOFs in the Cambridge Structural Database (CSD)\textsuperscript{3,67,68} and the 2019 Computation-Ready, Experimental (CoRE) MOF database.\textsuperscript{4} All starting structures were taken directly from the CSD,\textsuperscript{69} and free (i.e. unbound) solvents were automatically removed from the frameworks. From this set of experimental crystal structures, we constructed a DFT-ready dataset of 42,362 non-disordered MOF structures (“CSD-42362”) after an extensive suite of automated fidelity checks, as summarized in Figure S1. In contrast with existing databases of experimental MOF structures,\textsuperscript{3,4,22,66} this process better accounts for issues such as omitted H atoms, unresolved disorder, deleted framework atoms, lone (i.e. unbonded) atoms, an improper number of charge-balancing ions, and other structural issues that have been discussed in several recent studies.\textsuperscript{68,70–75} Of these 42,362 structures, a subset of 24,013 materials with 150 atoms or fewer per primitive cell were considered such that large-scale screening could be carried out. After completing the high-throughput DFT screening procedure, static single-point (SP) calculations on the un-relaxed structures were successfully completed for 19,691 materials (“CSD-19691-SP”), and full structure relaxations (including cell volume and atomic positions) were successfully carried out for 14,204 materials (“CSD-14204-opt”). All periodic DFT calculations during this dataset construction process were
carried out at the PBE-D3(BJ)\textsuperscript{27,76,77} level of theory using the Vienna \textit{ab initio} Simulation Package (VASP)\textsuperscript{78,79}. Additional methodological details regarding the dataset construction, DFT calculations, and ML methods can be found in the Supporting Information.

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

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

Prior to highlighting how this data can be used in practice, we first investigated several properties of the QMOF database generated in this work. As shown in Figure 2a, the CSD-14204-opt dataset contains MOFs with chemical elements that span nearly the entire periodic table, which is beneficial for the development of transferable ML models. When looking at the geometries before and after structure relaxation, we find that 97.1\% of the DFT-optimized MOFs had a change in cell volume less than 10\% (Figure 2b), suggesting that the removal of free solvent does not drastically alter the structural properties for the majority of the MOFs in this work. The distribution of DFT-computed band gaps for the fully optimized structures is shown in Figure 2c and indicates that there is a wide spread of values from nearly 0 eV to 6.45 eV. The band gaps are not normally distributed and instead are bimodal, with peaks centered around 0.9 eV and 2.9 eV. This can be attributed to different distributions associated with closed- and open-shell materials in the CSD-14204-opt dataset (Figure 2c), the latter of which have significantly lower band gaps at the PBE-D3(BJ) level of theory on average. With regards to partial atomic charges, a wide spread of values is also obtained (Figure S4a). In comparing the partial atomic charges before and after structure relaxation, we find that 92.4\% of the \~1.19 million data points have an absolute difference less than 0.05 \(q_e\), and 98.8\% of the points have an absolute difference less than 0.1 \(q_e\) (Figure S4b). As has been observed on a smaller scale in prior work,\textsuperscript{23,66} it can be safely assumed that the partial charges remain essentially unchanged upon structure relaxation in most cases.
Figure 2. a) Number of MOFs in the CSD-14204 dataset containing a given element. All elements that occur in greater than 800 structures are capped at 800 for ease of visualization. These include: C (14,204), H (14,204), N (11,526), O (11,458), Cu (2,617), S (2,493), Cd (2,272), Zn (2,250), Cl (1,575), and Ag (1,120). Elements in gray are not present in any structure. b) Histogram of the fractional change in cell volume before ($V_{SP}$) and after ($V_{opt}$) structure relaxation for the MOFs in the CSD-14204-opt dataset. Separate distributions are shown for the entire dataset (14,204 entries), the closed-shell MOFs (10,995 entries), and the open-shell MOFs (3,209 entries). Open-shell character is defined here as having a DDEC6 atomic spin density with a magnitude greater than 0.1. A box plot, showing the extrema and interquartile range, is included in each violin, with the median marked by a white dot.

As a brief demonstration for how the data generated via the high-throughput DFT workflow could be used directly, we identified any materials with high-spin Fe species, which we defined as having an absolute DDEC6 spin density greater than 3.5 from the DFT calculations. High-spin Fe complexes are known to be promising for oxidation catalysis, in particular for the activation of strong C–H bonds, and recent work has focused on stabilizing such motifs in MOFs for this purpose.96–98 We further constrained the list of MOFs to only include those with a pore-limiting diameter greater than 3.6 Å (i.e. the kinetic diameter of N$_2$). This search resulted in six frameworks, as shown in Figure S5: Fe(bpz) (H$_2$bpz = 4,4′-bipyrazole) (refcode: ACODAA),99 Fe$_2$(dobdc) (H$_4$dobdc = 2,5-dihydroxybenzene-1,4-dicarboxylic acid) with (refcode: SARHAW) and without (refcode: COKNOH) bound propylene,100,101 Fe$_2$Cl$_2$(bbta) (H$_2$bbta = 1H,5H-benzo(1,2-d:4,5-d′)bistriazole) (refcode: HAYYUE),102 and geometrically distinct, expanded conformations of Fe(bdp) (H$_2$bdp = 1,4-benzenedipryrazole) (refcodes: QUPZIM01, QUPZIM02).103 Providing validation of this screening approach, Fe$_2$(dobdc) has already been shown to oxidize ethane to ethanol,96,104 and Fe$_2$Cl$_2$(bbta) has been computationally investigated for oxidation reactions,105,106 with experimental evidence suggesting that both MOFs have high-spin Fe sites.96,102
Machine Learning Models for Band Gap Prediction

Beyond analyzing the DFT-computed properties directly, the QMOF database now makes it possible to train a wide range of ML models specifically tailored for MOFs, which are likely to have their own distinct feature space compared to isolated molecules and inorganic solids. This serves two primary purposes. The first is more theoretical: featurization methods (i.e. how each MOF structure is encoded) and ML algorithms that are well-suited for other materials may not be equally suitable for MOFs, so this database of quantum-chemical properties can serve as a testing ground to benchmark new ML methods. The Materials Project\textsuperscript{18} and OQMD\textsuperscript{25,26} in particular have accelerated this research direction for inorganic solids, and the QM9 dataset\textsuperscript{45,107} (as one example) has done the same for small molecule chemistry. The second purpose of this new database is to apply these rapid yet accurate ML models to accelerate the materials discovery process, now with the ability to train these models directly on properties computed for MOFs.

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

The simplest featurization methods considered in this work are the feature sets of He et al.\textsuperscript{24} (with 45 statistical attributes of elemental properties, denoted “Stoichiometric-45”) and Meredig and Agrawal et al.\textsuperscript{109} (with 103 attributes describing the elemental fractions from H–Lr and 17 statistical attributes of elemental properties, denoted “Stoichiometric-120”), which rely solely on the chemical composition of each material (Figure 3b). In addition, we consider several structure-sensitive featurization approaches, including the sine Coulomb matrix\textsuperscript{110} that encodes pairwise electrostatic interactions between nuclei in a material (Figure 3c, Equation S4) and the orbital field matrix\textsuperscript{111} that encodes the distribution of valence electrons in each coordination environment of a material (Figure 3d). The smooth overlap of atomic positions (SOAP)\textsuperscript{112,113} is another structure-sensitive descriptor considered in this work and can be used to compute the similarity between a pair of local atomic environments by representing the atoms as Gaussians (i.e. “smoothed positions”), which are then summed to produce a density field (Equation S5). The overlap of these density fields, integrated over all three-dimensional rotations (Equations S6–S7), are compared between structures to generate a kernel matrix describing the similarity between every pair of MOF structures (Figure S2 and Equations S8–S9). In all of the aforementioned examples, these features are used to develop a kernel ridge regression\textsuperscript{114} (KRR) model (Equations S1–S3). Motivated by prior work on inorganic solids, we also investigated the use of a crystal graph convolutional neural network (CGCNN),\textsuperscript{65} wherein an approximate crystal graph is generated for each MOF, with each node in the graph representing an atom and each edge representing the bonds that connect the atoms (Figure 3f). More detailed descriptions and full methodological details for each featurization method and ML model architecture can be found in the Supporting Information.
Figure 3. Visualization of various featurization methods applied to the unit cell of IRMOF-1. a) IRMOF-1 structure. b) Examples of composition-based features. c) Sine Coulomb matrix showing the interaction values between each pair of atoms. d) Orbital field matrix showing the average interaction value between each pair of orbital- or period-based features. Only non-zero values are shown. e) Averaged SOAP fingerprint of IRMOF-1 compared to IRMOF-2 and ZIF-8. Taking the dot product of any two vectors yields an unnormalized similarity score. f) Schematic of a crystal graph with example node (circle) and edge (line) embeddings (only a representative portion is shown for clarity).

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

Table 1. Summary of the mean absolute error (MAE), \( R^2 \), and Spearman rank-order correlation coefficient (\( \rho \)) for several machine learning methods to predict the computed band gaps of MOFs from their deposited crystal structures with free solvent removed. Kernel ridge regression was used for all featurization methods except for the crystal graphs of CGCNN, for which a convolutional neural network was constructed. The testing set statistics are shown, averaged over five runs (using different random seeds for data splitting) with ±1 standard deviation shown. For all models, 80% of the CSD-13058 dataset was used for training. The MAE for a dummy model that predicts the mean band gap (2.241 eV) for all the MOFs is shown for reference.
| ML Method            | MAE (eV)       | $R^2$  | $\rho$       |
|----------------------|----------------|--------|--------------|
| Constant mean model  | 0.984          | —      | —            |
| Sine Coulomb matrix  | 0.551 ± 0.013  | 0.625 ± 0.019 | 0.780 ± 0.014 |
| Stoichiometric-45    | 0.443 ± 0.011  | 0.746 ± 0.012 | 0.846 ± 0.009 |
| Stoichiometric-120   | 0.438 ± 0.009  | 0.748 ± 0.010 | 0.848 ± 0.006 |
| Orbital field matrix | 0.423 ± 0.007  | 0.761 ± 0.007 | 0.865 ± 0.003 |
| SOAP                 | 0.367 ± 0.007  | 0.807 ± 0.010 | 0.903 ± 0.005 |
| CGCNN                | 0.272 ± 0.004  | 0.885 ± 0.007 | 0.934 ± 0.004 |

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

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

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

**Dimensionality Reduction for Structure–Property Analysis**

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

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

As an example, selecting several MOFs in region A of the SOAP-based UMAP (Figure 5) yields materials with long, linear alkane-based linkers (e.g. refcodes NEZMEM\textsuperscript{122}, ROKZOI\textsuperscript{123}), which consistently have high band gaps regardless of the coordinating metal. The low band gap MOFs are more scattered throughout the reduced feature space, but as one example, region B of Figure 5 contains framework materials with linkers consisting of various TCNQ (TCNQ = 7,7,8,8-tetracyano-quinodimethane) derivatives, with several
of these materials previously shown to have high electrical conductivities (e.g. BISVUW\textsuperscript{124}, FAFJAZ\textsuperscript{125}). The projection in Figure 5 can be used to find MOFs that are structurally similar to a given material of interest as well. For instance, Cu\{Ni(pdt)\}_2\cdot C_2H_2 (pdt\textsuperscript{2} = 2,3-pyrazinedithiolate) (refcode: HIVPOU\textsuperscript{126}) is in the CSD-13058 dataset, and it is known to be one of the rare examples of a three-dimensional, porous framework that exhibits room temperature electrical conductivity.\textsuperscript{126} Perhaps unsurprisingly, one of the closest points to Cu\{Ni(pdt)\}_2\cdot C_2H_2 is the isostructural framework Cu\{Cu(pdt)\}_2\cdot C_2H_2 (refcode: WIHQEM\textsuperscript{127}) (region C), which has also been studied for its conductive properties.\textsuperscript{128,129} In general, we find the SOAP-based UMAP places greater emphasis on the similarity of the organic linkers rather than the metal identity, likely due to the averaging scheme used in the generation of the similarity kernel (Table S5). Modifications to the SOAP encoding that better account for the discrete building block nature of MOFs, such as variations on the recently developed coarse-grained SOAP (cg-SOAP) method,\textsuperscript{130} may yield improvements in the future.

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

![Figure 6. Unsupervised dimensionality reduction performed using UMAP, with a distance matrix obtained using a Euclidean distance metric of the Stoichiometric-120 encodings for the structures in the CSD-13058-SP dataset. The (a) maximum atomic number in each structure, $\text{max}(\text{Z})$, and (b) computed band gaps for the corresponding DFT-optimized structures, $E_{\text{g, DFT}}$, are overlaid on the UMAPs.](image)
differences in the elemental fractions and compositional features that compose the Stoichiometric-120 descriptor. Notably, the band gaps are well-separated between and within each cluster in the reduced space (Figure 6b). For these reasons, the Stoichiometric-120 UMAP can be a valuable tool for obtaining a global view of the QMOF database. For instance, we find that the CSD-13058 dataset closely overlaps with both the larger CSD-42362 dataset it was drawn from and the separate CoRE MOF 2019 database\(^4\) based on the reduced space of Stoichiometric-120 features (Figures S7 and S8). To enable additional data exploration, we have made interactive versions of the UMAPs available in the supporting dataset.\(^{59}\)

**Highlighting Notable Low Band Gap MOFs**

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

When the CGCNN model is used to predict the band gaps of all 42,362 structures that compose the CSD-42362 dataset, the lowest band gap material is predicted to be Ag(DCl)\(_2\) (DCl = 2,5-Cl,Cl-N',N'-dicyanoquinone diamine) (refcode: OTARUX\(^{145}\)), which is known from experiments to exhibit metallic character via organic radicals that connect the Ag(I) cations.\(^{145}\) The introduction of radical or redox-active linking units is a well-established strategy to increase the electrical conductivity of framework materials.\(^{60}\)

Although Ag(DCl)\(_2\) is arguably best described as a coordination polymer, one notable MOF in the CSD-42362 dataset with a low predicted band gap and a radical-containing linker is (TTF)[{Rh\(_2\)(CH\(_3\)CO\(_2\))\(_4\)}\(_2\)TCNQ] (TTF = tetrathiafulvalene) (refcode: WAQMEJ\(^{146}\)) – a pillared layer framework material built from Rh(II) paddlewheels and a TTF–TCNQ charge-transfer salt (Figure 7a). The HSE06-D3(BJ) band gap for this material is found to be particularly small with a value of 0.71 eV, which can be directly attributed to a reduced conduction band minimum (CBM) from the TTF and TCNQ components (Figure 7d). Furthermore, the valence band maximum (VBM) also exhibits hybridization between the 4d orbitals of Rh and 2p orbitals of C and N atoms belonging to the radical TCNQ linker, which is important for applications involving electron transport. In contrast, the most insulating structure in the CSD-42362 set based on CGCNN-predicted band gap is the non-porous coordination polymer Sr[C\(_2\)H\(_4\)(SO\(_3\))\(_2\)] (refcode: GUTYAW\(^{147}\)), which has an HSE06-D3(BJ) band gap of 8.36 eV (Figure S11).

Consistent with prior experimental work,\(^{148}\) we also find several Fe-containing materials in the CSD-42362 dataset with low band gaps, many of which have not been studied for their electronic properties. One representative example is Fe(sq)(bpee)(H\(_2\)O)\(_2\) (bpee = 1,2-bis(4-pyridyl)ethylene; sq = squarate) (refcode: RAXNEK\(^{149}\)), shown in Figure 7b, which has a band gap of 1.06 eV at the HSE06-D3(BJ) level of theory. The high-spin Fe(II) species in an octahedral crystal field with t\(_{2g}\)e\(_{g}\)\(^2\) electron configuration dominate the VBM in this material, whereas the bpee linker (as opposed to the bridging sq species or inorganic node) make up the conduction band edge (Figure 7e).
Figure 7. Structures of a) (TTF)[(Rh₂(CH₃CO₂)₄)_₂TCNQ], b) Fe(sq)(bpee)(H₂O)₂, and c) Fe(bipytz)(Au(CN)₂). Total and projected density of states (DOS) at the HSE06-D3(BJ) level of theory for d) (TTF)[(Rh₂(CH₃CO₂)₄)_₂TCNQ], e) Fe(sq)(bpee)(H₂O)₂, f) Fe(bipytz)(Au(CN)₂) (high spin), and g) Fe(bipytz)(Au(CN)₂) (low spin). The energy, $E$, in eV is shown with respect to the Fermi level, $E_F$. DOS values above and below zero refer to the spin-up and spin-down channels, respectively.
Another noteworthy example is the three-dimensional porous framework material Fe(bipytz)(Au(CN)$_2$)$_2$ (bipytz = 3,6-bis(4-pyridyl)-1,2,4,5-tetrazine) (recode: LOJLAA$^{150}$), shown in Figure 7c. At the HSE06-D3(BJ) level of theory, we find that the high spin state exhibits a band gap of 1.17 eV (Figure 7f) – similar to that of Fe(sq)(bpee)(H$_2$O)$_2$. The projected density of states indicates that the Au(I) species are unrelated to the relatively low band gap; instead, the low band gap can be attributed to the combination of Fe(II) and bipytz linker. Fe(bipytz)(Au(CN)$_2$)$_2$ is known to be a spin-crossover framework (with a sharp spin transition around 290 K),$^{150}$ and we find the low spin HSE06-D3(BJ) band gap to be 1.95 eV (Figure 7g), suggesting that the material may have tunable electronic properties as a function of temperature. For the low spin case, the VBM is composed of Fe 3$d$ orbitals and the CBM is composed of N 2$p$ orbitals. The reduction in band gap from low spin to high spin state can be rationalized on the basis of crystal field theory. In the high spin state, the Fe(II) centers have a $t_{2g}^4e_g^2$ electronic configuration, whereas in the low spin state they have a $t_{2g}^6e_g^0$ electron configuration. This occupation of the $e_g$ orbitals in the high spin state is directly related to the predicted ~0.8 eV reduction in the band gap compared to the low spin state. For both highlighted Fe-containing frameworks, the band gaps are lower – or comparable in the low spin state for Fe(bipytz)(Au(CN)$_2$)$_2$ – to those of several iron-containing MOFs that have been studied for their conductive properties, such as Fe$_2$(dobdc), Fe$_2$(dsbdc) (H$_4$dsbdc = 2,5-disulfhydrylbenzene-1,4-dicarboxylic acid), and Fe(bpz).$^{99,148}$ Collectively, these findings demonstrate the practical utility of the QMOF database for identifying MOFs with targeted quantum-chemical properties.

**Conclusion**

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

Importantly, the QMOF database now makes it possible to pursue several important research directions that are reliant on a large database of quantum-chemical properties for MOFs beyond those directly discussed in this work. For instance, with the success of transfer learning,$^{115,151}$ multi-task learning,$^{152}$ and Δ-ML$^{153}$ methods in materials research, the QMOF database can serve as a valuable resource to increase the accuracy – and reduce the required training set size – for ML models tasked with the prediction of new MOF properties not present in the QMOF database. Since the output of any ML models will depend on the chosen density functional approximation, related transfer learning approaches may also prove useful in generalizing ML model predictions to other levels of theory using the PBE-D3(BJ) data as a starting point. Instead of relying on representation approaches that were originally designed for inorganic solids or small molecules, the QMOF database can also be used to develop better methods for the encoding of MOF structures in ML models. Even outside the area of high-throughput DFT screening, data mining, and ML, there are countless possible use-cases for the QMOF database. As just one example, the DFT-generated properties in the QMOF database could be used to develop and/or benchmark (semi-)empirical methods (e.g. tight binding approaches$^{154}$ or molecular mechanics force fields$^{155}$) with the hopes of achieving high
accuracies for MOF structures. Finally, we note that the QMOF database should be considered a living resource; several additions to the QMOF database are planned in the future, and we welcome the development of subsets, modifications, and supplements to the database that suit the diverse needs of the MOF community. With all this in mind, we anticipate that this database will accelerate the material design and discovery process while being specifically tailored for the chemical space of experimentally realized MOF structures.

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

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

References
(1) Yaghi, O. M.; Kalmutzki, M. J.; Diercks, C. S. Introduction to Reticular Chemistry: Metal-Organic Frameworks and Covalent Organic Frameworks; John Wiley & Sons: Weinheim, Germany, 2019.
(2) Kalmutzki, M. J.; Hanikel, N.; Yaghi, O. M. Secondary Building Units as the Turning Point in the Development of the Reticular Chemistry of MOFs. Sci. Adv. 2018, 4 (10), eaat9180.
(3) Moghadam, P. Z.; Li, A.; Wiggin, S. B.; Tao, A.; Maloney, A. G. P.; Wood, P. A.; Ward, S. C.; Fairen-Jimenez, D. Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future. Chem. Mater. 2017, 29 (7), 2618–2625.
(4) Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S.; Slater, B.; Siepmann, J. I.; Sholl, D. S.; Snurr, R. Q. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. J. Chem. Eng. Data 2019, 64 (12), 5985–5998.
(5) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-Scale Screening of Hypothetical Metal–Organic Frameworks. Nat. Chem. 2012, 4 (2), 83–89.
(6) Colón, Y. J.; Gómez-Gualdrón, D. A.; Snurr, R. Q. Topologically Guided, Automated Construction of Metal–Organic Frameworks and Their Evaluation for Energy-Related Applications. Cryst. Growth Des. 2017, 17 (11), 5801–5810.
(7) Boyd, P. G.; Woo, T. K. A Generalized Method for Constructing Hypothetical Nanoporous Materials of Any Net Topology from Graph Theory. CrystEngComm 2016, 18 (21), 3777–3792.
(8) Colón, Y. J.; Snurr, R. Q. High-Throughput Computational Screening of Metal–Organic Frameworks. Chem. Soc. Rev. 2014, 43 (16), 5735–5749.
(9) Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B. Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. Chem. Rev. 2020, 120 (16), 8066–8129.

(10) Chibani, S.; Coudert, F.-X. Machine Learning Approaches for the Prediction of Materials Properties. APL Mater. 2020, 8 (8), 80701.

(11) Anderson, G.; Schweitzer, B.; Anderson, R.; Gómez-Gualdrón, D. A. Attainable Volumetric Targets for Adsorption-Based Hydrogen Storage in Porous Crystals: Molecular Simulation and Machine Learning. J. Phys. Chem. C 2018, 123 (1), 120–130.

(12) Bucior, B. J.; Bobbitt, N. S.; Islamoglu, T.; Goswami, A.; Yildirim, T.; Farha, O. K.; Bagheri, N.; Snurr, R. Q. Energy-Based Descriptors to Rapidly Predict Hydrogen Storage in Metal–Organic Frameworks. Mol. Syst. Des. Eng. 2019, 4 (1), 162–174.

(13) Thornton, A. W.; Simon, C. M.; Kim, J.; Kwon, O.; Deeg, K. S.; Konstas, K.; Pas, S. J.; Hill, M. R.; Winkler, D. A.; Haranczyk, M.; Smit, B. Materials Genome in Action: Identifying the Performance Limits of Physical Hydrogen Storage. Chem. Mater. 2017, 29 (7), 2844–2854.

(14) Anderson, R.; Rodgers, J.; Argueta, E.; Biong, A.; Gómez-Gualdrón, D. A. Role of Pore Chemistry and Topology in the CO$_2$ Capture Capabilities of MOFs: From Molecular Simulation to Machine Learning. Chem. Mater. 2018, 30 (18), 6325–6337.

(15) Dureckova, H.; Krykunov, M.; Aghaji, M. Z.; Woo, T. K. Robust Machine Learning Models for Predicting High CO$_2$ Working Capacity and CO$_2$/H$_2$ Selectivity of Gas Adsorption in Metal Organic Frameworks for Precombustion Carbon Capture. J. Phys. Chem. C 2019, 123 (7), 4133–4139.

(16) Yao, Z.; Sanchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O.; Snurr, R. Q.; Aspuru-Guzik, A. Inverse Design of Nanoporous Crystalline Reticular Materials with Deep Generative Models. ChemRxiv 2020. https://doi.org/10.26434/chemrxiv.12186681.v2.

(17) Shi, Z.; Yang, W.; Deng, X.; Cai, C.; Yan, Y.; Liang, H.; Liu, Z.; Qiao, Z. Machine-Learning-Assisted High-Throughput Computational Screening of High Performance Metal–Organic Frameworks. Mol. Syst. Des. Eng. 2020, 5 (4), 725–742.

(18) Chong, S.; Lee, S.; Kim, B.; Kim, J. Applications of Machine Learning in Metal-Organic Frameworks. Coord. Chem. Rev. 2020, 423, 213487.

(19) Mancuso, J. L.; Mroz, A. M.; Le, K. N.; Hendon, C. H. Electronic Structure Modeling of Metal–Organic Frameworks. Chem. Rev. 2020, 120 (16), 8641–8715.

(20) Raza, A.; Sturluson, A.; Simon, C.; Fern, X. Message Passing Neural Networks for Partial Charge Assignment to Metal-Organic Frameworks. J. Phys. Chem. C 2020, 124 (35), 19070–19082.

(21) Korolev, V. V; Mitrofanov, A.; Marchenko, E. I.; Eremin, N. N.; Tkachenko, V.; Kalmykov, S. N. Transferable and Extensible Machine Learning Derived Atomic Charges for Modeling Hybrid Nanoporous Materials. Chem. Mater. 2020, 32 (18), 7822–7831.

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

(23) Nazarian, D.; Camp, J. S.; Sholl, D. S. A Comprehensive Set of High-Quality Point Charges for Simulations of Metal–Organic Frameworks. Chem. Mater. 2016, 28 (3), 785–793.

(24) He, Y.; Cubuk, E. D.; Allendorf, M. D.; Reed, E. J. Metallic Metal–Organic Frameworks Predicted by the Combination of Machine Learning Methods and Ab Initio Calculations. J. Phys. Chem. Lett. 2018, 9 (16), 4562–4569.

(25) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredith, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). Jom 2013, 65 (11), 1501–1509.

(26) Kirklin, S.; Saal, J. E.; Meredith, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): Assessing the Accuracy of DFT Formation Energies. npj Comput. Mater. 2015, 1 (1), 1–15.
(27) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. Phys. Rev. Lett. 1996, 77 (18), 3865–3868.
(28) Abrahams, B. F.; Hardie, M. J.; Hoskins, B. F.; Robson, R.; Williams, G. A. Topological Rearrangement within a Single Crystal from a Honeycomb Cadmium Cyanide [Cd(CN)$_2$]$_n$ 3D Net to a Diamond Net. J. Am. Chem. Soc. 1992, 114 (26), 10641–10643.
(29) von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring Chemical Compound Space with Quantum-Based Machine Learning. Nat. Rev. Chem. 2020, 4, 347–358.
(30) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. Nature 2018, 559 (7715), 547–555.
(31) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. npj Comput. Mater. 2017, 3 (1), 1–13.
(32) Ward, L.; Wolverton, C. Atomistic Calculations and Materials Informatics: A Review. Curr. Opin. Solid State Mater. Sci. 2017, 21 (3), 167–176.
(33) Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Deng, Z.; Ong, S. P. A Critical Review of Machine Learning of Energy Materials. Adv. Energy Mater. 2020, 10 (8), 1903242.
(34) Morgan, D.; Jacobs, R. Opportunities and Challenges for Machine Learning in Materials Science. Annu. Rev. Mater. Res. 2020, 50, 71–103.
(35) Saal, J. E.; Oliynyk, A. O.; Meredig, B. Machine Learning in Materials Discovery: Confirmed Predictions and Their Underlying Approaches. Annu. Rev. Mater. Res. 2020, 50 (1), 49–69.
(36) Suh, C.; Fare, C.; Warren, J. A.; Pyzer-Knapp, E. O. Evolving the Materials Genome: How Machine Learning Is Fueling the Next Generation of Materials Discovery. Annu. Rev. Mater. Res. 2020, 50, 1–25.
(37) Jain, A.; Hautier, G.; Moore, C. J.; Ping Ong, S.; Fischer, C. C.; Mueller, T.; Persson, K. A.; Ceder, G. A High-Throughput Infrastructure for Density Functional Theory Calculations. Comput. Mater. Sci. 2011, 50 (8), 2295–2310.
(38) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. APL Mater. 2013, 1 (1), 11002.
(39) Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R. H.; Nelson, L. J.; Hart, G. L. W.; Sanvito, S.; Buongiorno-Nardelli, M.; Mingo, N.; Levy, O. AFLWLIB.ORG: A Distributed Materials Properties Repository from High-Throughput Ab Initio Calculations. Comput. Mater. Sci. 2012, 58, 227–235.
(40) Winther, K. T.; Hoffmann, M. J.; Boes, J. R.; Mamun, O.; Bajdich, M.; Bliagaard, T. Catalysis-Hub. Org, an Open Electronic Structure Database for Surface Reactions. Sci. Data 2019, 6 (1), 1–10.
(41) Borysov, S. S.; Geilhufe, R. M.; Balatsky, A. V. Organic Materials Database: An Open-Access Online Database for Data Mining. PLoS One 2017, 12 (2).
(42) Landis, D. D.; Hummelshoj, J. S.; Nestorov, S.; Greeley, J.; Dulak, M.; Bliagaard, T.; Nørskov, J. K.; Jacobsen, K. W. The Computational Materials Repository. Comput. Sci. Eng. 2012, 14 (6), 51–57.
(43) Draxl, C.; Scheffler, M. NOMAD: The FAIR Concept for Big Data-Driven Materials Science. MRS Bull. 2018, 43 (9), 676–682.
(44) Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; Palizhati, A.; Sriram, A.; Wood, B.; Yoon, J.; Parikh, D.; Zitnicks, C. L.; Ulissi, Z. The Open Catalyst 2020 (OC20) Dataset and Community Challenges. arXiv. arXiv:2010.09990 2020.
(45) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. Sci. Data 2014, 1, 140022.
(46) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Machine Learning of Molecular Electronic Properties in Chemical Compound Space. New J. Phys. 2013, 15 (9), 95003.
(47) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1, A Data Set of 20 Million Calculated off-Equilibrium Conformations for Organic Molecules. Sci. Data 2017, 4, 170193.

(48) Blau, S.; Spotte-Smith, E.; Wood, B.; Dwaraknath, S.; Persson, K. Accurate, Automated Density Functional Theory for Complex Molecules Using On-the-Fly Error Correction. ChemRxiv 2020. https://doi.org/10.26434/chemrxiv.13076030.v1.

(49) Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazakim, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I.; Baldo, M.; Adam, R. P.; Aspuru-Guzik, A. Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. Nat. Mater. 2016, 15 (10), 1120–1127.

(50) Mansouri Tehrani, A.; Oliynyk, A. O.; Parry, M.; Rizvi, Z.; Couper, S.; Lin, F.; Miyagi, L.; Sparks, T. D.; Brugoch, J. Machine Learning Directed Search for Ultraincompressible, Superhard Materials. J. Am. Chem. Soc. 2018, 140 (31), 9844–9853.

(51) Wu, S.; Kondo, Y.; Kakimoto, M.; Yang, B.; Yamada, H.; Kuwajima, I.; Lambard, G.; Hongo, K.; Xu, Y.; Shiomi, J.; Schick, C.; Morikawa, J.; Yoshida, R. Machine-Learning-Assisted Discovery of Polymers with High Thermal Conductivity Using a Molecular Design Algorithm. npj Comput. Mater. 2019, 5 (1), 1–11.

(52) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part I: Progress. Angew. Chem. Int. Ed. 2019. https://doi.org/10.1002/anie.201909987.

(53) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part II: Outlook. Angew. Chem. Int. Ed. 2019. https://doi.org/10.1002/anie.201909989.

(54) Rosen, A. S.; Notestein, J. M.; Snurr, R. Q. Identifying Promising Metal–Organic Frameworks for Heterogeneous Catalysis via High-Throughput Periodic Density Functional Theory. J. Comput. Chem. 2019, 40 (12), 1305–1318.

(55) QMOF Database. https://github.com/arosen93/QM0F, https://dx.doi.org/10.6084/m9.figshare.13147324.

(56) Xie, L. S.; Skorupskii, G.; Dincă, M. Electrically Conductive Metal–Organic Frameworks. Chem. Rev. 2020, 120 (16), 8536–8580.

(57) Baumann, A. E.; Burns, D. A.; Liu, B.; Thoi, V. S. Metal–Organic Framework Functionalization and Design Strategies for Advanced Electrochemical Energy Storage Devices. Commun. Chem. 2019, 2 (1), 1–14.

(58) D’Alessandro, D. M. Exploiting Redox Activity in Metal–Organic Frameworks: Concepts, Trends and Perspectives. Chem. Commun. 2016, 52 (58), 8957–8971.

(59) Downes, C. A.; Marinescu, S. C. Electrocatalytic Metal–Organic Frameworks for Energy Applications. ChemSusChem 2017, 10 (22), 4374–4392.

(60) Allendorf, M. D.; Dong, R.; Feng, X.; Kaskel, S.; Matoga, D.; Stavila, V. Electronic Devices Using Open Framework Materials. Chem. Rev. 2020, 120 (16), 8581–8640.

(61) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. Phys. Rev. Lett. 2018, 120 (14), 145301.

(62) Nazarian, D.; Camp, J. S.; Chung, Y. G.; Snurr, R. Q.; Sholl, D. S. Large-Scale Refinement of
Metal–Organic Framework Structures Using Density Functional Theory. Chem. Mater. 2017, 29 (6), 2521–2528.

(67) Cole, J. C.; Wiggin, S.; Stanzione, F. New Insights and Innovation from a Million Crystal Structures in the Cambridge Structural Database. Struct. Dyn. 2019, 6 (5), 54301.

(68) Li, A.; Bueno-Perez, R.; Wiggin, S.; Fairen-Jimenez, D. Enabling Efficient Exploration of Metal–Organic Frameworks in the Cambridge Structural Database. CrystEngComm 2020. https://doi.org/10.1039/D0CE00299B.

(69) Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kesseler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor. R. New Software for Searching the Cambridge Structural Database and Visualizing Crystal Structures. Acta Crystallogr. Sect. B 2002, 58 (3–1), 389–397.

(70) Zarabadi-Poor, P.; Marek, R. Comment on “Database for CO$_2$ Separation Performances of MOFs Based on Computational Materials Screening.” ACS Appl. Mater. Interfaces 2019, 11 (18), 16261–16265.

(71) Barthel, S.; Alexandrov, E. V; Proserpio, D. M.; Smit, B. Distinguishing Metal–Organic Frameworks. Cryst. Growth Des. 2018, 18 (3), 1738–1747.

(72) Altintas, C.; Avci, G.; Daglar, H.; Azar, A. N. V.; Ercan, I.; Velioglu, S.; Keskin, S. An Extensive Comparative Analysis of Two MOF Databases: High-Throughput Screening of Computation-Ready MOFs for CH$_4$ and HE Adsorption. J. Mater. Chem. A 2019, 7 (16), 9593–9608.

(73) Velioglu, S.; Keskin, S. Revealing the Effect of Structure Curations on the Simulated CO$_2$ Separation Performances of MOFs. Mater. Adv. 2020, 1 (3), 341–353.

(74) Bucior, B. J.; Rosen, A. S.; Haranczyk, M.; Yao, Z.; Ziebel, M. E.; Farha, O. K.; Hupp, J. T.; Siepmann, J. J.; Aspuru-Guzik, A.; Snurr, R. Q. Identification Schemes for Metal–Organic Frameworks to Enable Rapid Search and Cheminformatics Analysis. Cryst. Growth Des. 2019, 19 (11), 6682–6697.

(75) Chen, T.; Manz, T. A. Identifying Misbonded Atoms in the 2019 CoRE Metal–Organic Framework Database. RSC Adv. 2020, 10 (45), 26944–26951.

(76) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H–Pu. J. Chem. Phys. 2010, 132 (15), 154104.

(77) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. J. Comput. Chem. 2011, 32 (7), 1456–1465.

(78) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. Phys. Rev. B 1996, 54 (16), 11169–11186.

(79) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. Phys. Rev. B 1999, 59 (3), 1758–1775.

(80) Hendon, C. H.; Tiana, D.; Walsh, A. Conductive Metal–Organic Frameworks and Networks: Fact or Fantasy? Phys. Chem. Chem. Phys. 2012, 14 (38), 13120–13132.

(81) Sun, L.; Campbell, M. G.; Dincă, M. Electrically Conductive Porous Metal–Organic Frameworks. Angew. Chem. Int. Ed. 2016, 55 (11), 3566–3579.

(82) Singh, A. K.; Montoya, J. H.; Greigore, J. M.; Persson, K. A. Robust and Synthesizable Photocatalysts for CO$_2$ Reduction: A Data-Driven Materials Discovery. Nat. Commun. 2019, 10, 443.

(83) Yang, L.-M.; Ravindran, P.; Vajeeston, P.; Svelle, S.; Tilset, M. A Quantum Mechanically Guided View of Cd-MOF-5 from Formation Energy, Chemical Bonding, Electronic Structure, and Optical Properties. Microporous Mesoporous Mater. 2013, 175, 50–58.

(84) Gong, S.; Xie, T.; Zhu, T.; Wang, S.; Fadel, E. R.; Li, Y.; Grossman, J. C. Predicting Charge Density Distribution of Materials Using a Local-Environment-Based Graph Convolutional Network. Phys. Rev. B 2019, 100 (18), 184103.

(85) Grisafi, A.; Fabrizio, A.; Meyer, B.; Wilkins, D. M.; Corminboeuf, C.; Ceriotti, M. Transferable Machine-Learning Model of the Electron Density. ACS Cent. Sci. 2018, 5 (1), 57–64.

(86) Chandrasekaran, A.; Kamal, D.; Batra, R.; Kim, C.; Chen, L.; Ramprasad, R. Solving the
Electronic Structure Problem with Machine Learning. npj Comput. Mater. 2019, 5 (1), 1–7.
(87) Kamal, D.; Chandrasekaran, A.; Batra, R.; Ramprasad, R. A Charge Density Prediction Model for Hydrocarbons Using Deep Neural Networks. Mach. Learn. Sci. Technol. 2020, 1 (2), 25003.
(88) Kolb, B.; Lentz, L. C.; Kolpak, A. M. Discovering Charge Density Functionals and Structure-Property Relationships with PROPhet: A General Framework for Coupling Machine Learning and First-Principles Methods. Sci. Rep. 2017, 7 (1), 1–9.
(89) Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating Materials Property Predictions Using Machine Learning. Sci. Rep. 2013, 3 (1), 1–6.
(90) Manz, T. A.; Limas, N. G. Introducing DDEC6 Atomic Population Analysis: Part 1. Charge Partitioning Theory and Methodology. RSC Adv. 2016, 6 (53), 47771–47801.
(91) Limas, N. G.; Manz, T. A. Introducing DDEC6 Atomic Population Analysis: Part 2. Computed Results for a Wide Range of Periodic and Nonperiodic Materials. RSC Adv. 2016, 6 (51), 45727–45747.
(92) Manz, T. A. Introducing DDEC6 Atomic Population Analysis: Part 4. Efficient Parallel Computation of Net Atomic Charges, Atomic Spin Moments, Bond Orders, and More. RSC Adv. 2018, 8 (5), 2678–2707.
(93) Manz, T. A. Introducing DDEC6 Atomic Population Analysis: Part 3. Comprehensive Method to Compute Bond Orders. RSC Adv. 2017, 7 (72), 45552–45581.
(94) Rosen, A. S.; Notestein, J. M.; Snurr, R. Q. Structure−Activity Relationships That Identify Metal−Organic Framework Catalysts for Methane Activation. ACS Catal. 2019, 9, 3576–3587.
(95) Yang, B.; Wu, X.-P.; Gagliardi, L.; Truhlar, D. G. Methane Functionalization by an Ir(III) Catalyst Supported on a Metal−Organic Framework: An Alternative Explanation of Steric Confinement Effects. Theor. Chem. Acc. 2019, 138 (9), 107.
(96) Xiao, D. J.; Bloch, E. D.; Mason, J. A.; Queen, W. L.; Hudson, M. R.; Planas, N.; Borycz, J.; Dzubak, A. L.; Verma, P.; Lee, K.; Bonino, F.; Crocellà, V.; Yano, J.; Bordiga, S.; Truhlar, D. G.; Gagliardi, L.; Brown, C. M.; Long, J. R. Oxidation of Ethane to Ethanol by N₂O in a Metal−Organic Framework with Coordinatively Unsaturated Iron(II) Sites. Nat. Chem. 2014, 6 (7), 590–595.
(97) Vogiatzis, K. D.; Haldoupis, E.; Xiao, D. J.; Long, J. R.; Siepmann, J. I.; Gagliardi, L. Accelerated Computational Analysis of Metal−Organic Frameworks for Oxidation Catalysis. J. Phys. Chem. C 2016, 120 (33), 18707–18712.
(98) Osadchii, D.; Olivas Suarez, A. I.; Szécsényi, Á.; Li, G.; Nasalevich, M. A.; Dugulan, A. I.; Serracrespo, P.; Hensen, E. J. M.; Veber, S. L.; Fedin, M. V.; Sankar, G.; Pidko, E. A.; Gascon, J. Isolated Fe Sites in Metal−Organic Framework Catalyze the Direct Conversion of Methane to Methanol. ACS Catal. 2018, 8 (6), 5542–5548.
(99) Spirkl, S.; Grzywa, M.; Reschke, S.; Fischer, J. K. H.; Sippel, P.; Demeshko, S.; von Nidda, H.-A.; Volkmer, D. Single-Crystal to Single-Crystal Transformation of a Nonporous Fe(II) Metal−Organic Framework into a Porous Metal−Organic Framework via a Solid-State Reaction. Inorg. Chem. 2017, 56 (20), 12337–12347.
(100) Queen, W. L.; Hudson, M. R.; Bloch, E. D.; Mason, J. A.; Gonzalez, M. I.; Lee, J. S.; Gygi, D.; Howe, J. D.; Lee, K.; Darwish, T. A.; James, M.; Peterson, V. K.; Teat, S. J.; Smit, B.; Neaton, J. B.; Long, J. R.; Brown, C. M. Comprehensive Study of Carbon Dioxide Adsorption in the Metal−Organic Frameworks M₂(dobdc)(M = Mg, Mn, Fe, Co, Ni, Cu, Zn). Chem. Sci. 2014, 5 (12), 4569–4581.
(101) Bloch, E. D.; Queen, W. L.; Krishna, R.; Zadrozy, J. M.; Brown, C. M.; Long, J. R. Hydrocarbon Separations in a Metal−Organic Framework with Open Iron(II) Coordination Sites. Science 2012, 335 (6076), 1606–1610.
(102) Reed, D. A.; Keitz, B. K.; Oktawiec, J.; Mason, J. A.; Runčevski, T.; Xiao, D. J.; Darago, L. E.; Crocellà, V.; Bordiga, S.; Long, J. R. A Spin Transition Mechanism for Cooperative Adsorption in Metal−Organic Frameworks. Nature 2017, 550 (7674), 96–100.
(103) Mason, J. A.; Oktawiec, J.; Taylor, M. K.; Hudson, M. R.; Rodriguez, J.; Bachman, J. E.
Gonzalez, M. I.; Cervellino, A.; Guagliardi, A.; Brown, C. M.; Llewellyn, P. L.; Masciocchi, N.; Long, J. R. Methane Storage in Flexible Metal–Organic Frameworks with Intrinsic Thermal Management. *Nature* 2015, 527 (7578), 357–361.

Verma, P.; Vagiatzis, K. D.; Planas, N.; Borycz, J.; Xiao, D. J.; Long, J. R.; Gagliardi, L.; Truhlar, D. G. Mechanism of Oxidation of Ethane to Ethanol at Iron(IV)-Oxo Sites in Magnesium-Diluted Fe$_2$(DOBDC). *J. Am. Chem. Soc.* 2015, 137 (17), 5770–5781.

Rosen, A. S.; Notestein, J. M.; Snurr, R. Q. High-Valent Metal-Oxo Species at the Nodes of Metal–Triazolate Frameworks: The Effects of Ligand-Exchange and Two-State Reactivity for C–H Bond Activation. *Angew. Chem. Int. Ed.* 2020, 142 (9), 4317–4328.

Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* 2012, 52 (11), 2864–2875.

Li, H.; Eddaoudi, M.; O’Keeffe, M.; Yaghi, O. M. Design and Synthesis of an Exceptionally Stable and Highly Porous Metal–Organic Framework. *Nature* 1999, 402 (6759), 276–279.

Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial Screening for New Materials in Unconstrained Composition Space with Machine Learning. *Phys. Rev. B* 2014, 89 (9), 94104.

Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Crystal Structure Representations for Machine Learning Models of Formation Energies. *Int. J. Quantum Chem.* 2015, 115 (16), 1094–1101.

Lam Pham, T.; Kino, H.; Terakura, K.; Miyake, T.; Tsuda, K.; Takigawa, I.; Chi Dam, H. Machine Learning Reveals Orbital Interaction in Materials. *Sci. Technol. Adv. Mater.* 2017, 18 (1), 756–765.

Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B* 2013, 87 (18), 184115.

De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing Molecules and Solids across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* 2016, 18 (20), 13754–13769.

Pronobis, W.; Müller, K.-R. Kernel Methods for Quantum Chemistry. In *Machine Learning Meets Quantum Physics*; Springer, 2020; pp 25–36.

Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* 2019, 31 (9), 3564–3572.

Olsthoorn, B.; Geihlufe, R. M.; Borysov, S. S.; Balatsky, A. V. Band Gap Prediction for Large Organic Crystal Structures with Machine Learning. *Adv. Quantum Technol.* 2019, 2 (7–8), 1900023.

Gascon, J.; Hernández-Alonso, M. D.; Almeida, A. R.; van Klink, G. P. M.; Kapteijn, F.; Mul, G. Isoreticular MOFs as Efficient Photocatalysts with Tunable Band Gap: An Operando FTIR Study of the Photoinduced Oxidation of Propylene. *ChemSusChem* 2008, 1 (12), 981–983.

Stanzione, D.; Barth, B.; Gaffney, N.; Gaither, K.; Hempel, C.; Minyard, T.; Mehringer, S.; Wernert, E.; Tufo, H.; Panda, D.; Teller, P.; Stampede 2: The Evolution of an XSEDE Supercomputer. In *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*; 2017; pp 1–8.

Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazelwood, V.; Lathrop, S.; Lifka, D.; Peterson, G. D.; Roskies, R.; Scott, J. R.; Wilkins-Diehr, N. XSEDE: Accelerating Scientific Discovery. *Comput. Sci. Eng.* 2014, 16 (5), 62–74.

McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* 2018.

Leland, M.; John, H.; Nathaniel, S.; Lukas, G. UMAP: Uniform Manifold Approximation and
(122) Zeng, X.-Z.; Zhang, A.-Y.; Bu, D.; Li, Y.-W. Hydrothermal Synthesis, Structure and Thermal Properties of a Novel Three-Dimensional La(III)-Sebacate Framework. *Chinese J. Struct. Chem.* 2013, 32 (1), 120–124.

(123) Martínez-Casado, F. J.; Ramos-Riesco, M.; Rodríguez-Cheda, J. A.; Cucinotta, F.; Fernández-Martínez, A.; Garrido, L.; Matiesanz, E.; Marchese, L. Short Lead(II) Soaps: From Weakly Fluorescent Crystals to Strongly Phosphorescent and Structurally Varied Vitreous Phases. A Thermal, Structural and Spectroscopic Study. *J. Mater. Chem. C* 2014, 2 (44), 9489–9496.

(124) Zhang, Z.; Zhao, H.; Matsushita, M. M.; Awaga, K.; Dunbar, K. R. A New Metal–Organic Hybrid Material with Intrinsic Resistance-Based Bistability: Monitoring in Situ Room Temperature Switching Behavior. *J. Mater. Chem. C* 2014, 2 (2), 399–404.

(125) Lopez, N.; Zhao, H.; Ota, A.; Prosvirin, A. V; Reinheimer, E. W.; Dunbar, K. R. Unprecedented Binary Semiconductors Based on TCNQ: Single-Crystal X-Ray Studies and Physical Properties of Cu(TCNQX₂) X = Cl, Br. *Adv. Mater.* 2010, 22 (9), 986–989.

(126) Aubrey, M. L.; Kapelewski, M. T.; Melville, J. F.; Oktawiec, J.; Presti, D.; Gagliardi, L.; Long, J. R. Chemiresistive Detection of Gaseous Hydrocarbons and Interrogation of Charge Transport in Cu[Ni(2,3-Pyrazinedithiolate)₂] by Gas Adsorption. *J. Am. Chem. Soc.* 2019, 141 (12), 5005–5013.

(127) Peng, Y.-L.; Pham, T.; Li, P.; Wang, T.; Chen, Y.; Chen, K.-J.; Forrest, K. A.; Space, B.; Cheng, P.; Zaworotko, M. J.; Zhang, Z. Robust Ultramicroporous Metal–Organic Frameworks with Benchmark Affinity for Acetylene. *Angew. Chem. Int. Ed.* 2018, 57 (34), 10971–10975.

(128) Takaishi, S.; Hosoda, M.; Kajiwara, T.; Miyasaka, H.; Yamashita, M.; Nakanishi, Y.; Kitagawa, Y.; Yamaguchi, K.; Kobayashi, A.; Kitagawa, H. Electroconductive Porous Coordination Polymer Cu[Cu(pdt)₂] Composed of Donor and Acceptor Building Units. *Inorg. Chem.* 2009, 48 (19), 9048–9050.

(129) Kobayashi, Y.; Jacobs, B.; Allendorf, M. D.; Long, J. R. Conductivity, Doping, and Redox Chemistry of a Microporous Dithiolene-Based Metal–Organic Framework. *Chem. Mater.* 2010, 22 (14), 4120–4122.

(130) Nicholas, T. C.; Goodwin, A.; Deringer, V. L. Understanding the Geometric Diversity of Inorganic and Hybrid Frameworks through Structural Coarse-Graining. *Chem. Sci.* https://doi.org/10.1039/D0SC03287E.

(131) Moosavi, S. M.; Nandy, A.; Jablonka, K. M.; Ongari, D.; Janet, J. P.; Boyd, P. G.; Lee, Y.; Smit, B.; Kulik, H. J. Understanding the Diversity of the Metal–Organic Framework Ecosystem. *Nat. Commun.* 2020, 11, 4068.

(132) Hulvey, Z.; Furman, J. D.; Turner, S. A.; Tang, M.; Cheetham, A. K. Dimensionality Trends in Metal–Organic Frameworks Containing Perfluorinated or Nonfluorinated Benzencarboxylates. *Cryst. Growth Des.* 2010, 10 (5), 2041–2043.

(133) Taylor, M. K.; Runčevski, T.; Oktawiec, J.; Gonzalez, M. I.; Siegelman, R. L.; Mason, J. A.; Ye, J.; Brown, C. M.; Long, J. R. Tuning the Adsorption-Induced Phase Change in the Flexible Metal–Organic Framework Co(bdp). *J. Am. Chem. Soc.* 2016, 138 (45), 15019–15026.

(134) Cui, X.; Chen, K.; Xing, H.; Yang, Q.; Krishna, R.; Bao, Z.; Wu, H.; Zhou, W.; Dong, X.; Han, Y.; Li, B.; Ren, Q.; Zaworotko, M. J.; Chen, B. Pore Chemistry and Size Control in Hybrid Porous Materials for Acetylene Capture from Ethylene. *Science* 2016, 353 (6295), 141–144.

(135) Li, S.; Zhang, L.; Lu, B.; Yan, E.; Wang, T.; Li, L.; Wang, J.; Yu, Y.; Mu, Q. A New Polyoxovanadate-Based Metal–Organic Framework: Synthesis, Structure and Photo-/Electro-Catalytic Properties. *New J. Chem.* 2018, 42 (9), 7247–7253.

(136) Yan, B.; Luo, J.; Dube, P.; Sefat, A. S.; Greedan, J. E.; Maggard, P. A. Spin-Gap Formation and Thermal Structural Studies in Reduced Hybrid Layered Vanadates. *Inorg. Chem.* 2006, 45 (13), 5109–5118.

(137) Lysenko, A. B.; Senchyk, G. A.; Domasevitch, K. V.; Hauser, J.; Fuhrmann, D.; Kobalz, M.; Krautscheid, H.; Neves, P.; Valente, A. A.; Gonçalves, I. S. Synthesis and Structural Elucidation
of Triazolylmolybdenum(VI) Oxide Hybrids and Their Behavior as Oxidation Catalysts. Inorg. Chem. 2015, 54 (17), 8327–8338.

(138) Wang, X.-L.; Zhang, S.; Wang, X.; Zhang, H.-X.; Xu, N.; Li, T.-J. Three β-Octamolybdate-Based Supramolecular Hybrids Constructed from a Bis-Imidazolyl-Bis-Amide Ligand: Fast and Selective Adsorption Activities of Organic Dyes. New J. Chem. 2017, 41 (5), 2178–2185.

(139) Choudhuri, I.; Truhlar, D. G. HLE17: An Efficient Way To Predict Band Gaps of Complex Materials. J. Phys. Chem. C 2019, 123 (28), 17416–17424.

(140) Borlido, P.; Aull, T.; Huran, A. W.; Tran, F.; Marques, M. A. L.; Botti, S. Large-Scale Benchmark of Exchange–Correlation Functionals for the Determination of Electronic Band Gaps of Solids. J. Chem. Theory Comput. 2019, 15 (9), 5069–5079.

(141) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. Hybrid Functionals Based on a Screened Coulomb Potential. J. Chem. Phys. 2003, 118 (14), 7605–7621.

(142) Moellmann, J.; Grimme, S. DFT-D3 Study of Some Molecular Crystals. J. Phys. Chem. C 2014, 118 (14), 7615–7621.

(143) Valenzano, L.; Civalleri, B.; Chavan, S.; Bordiga, S.; Nilsen, M. H.; Jakobsen, S.; Lillerud, K. P.; Lamberti, C. Disclosing the Complex Structure of Uio-66 Metal Organic Framework: A Synergic Combination of Experiment and Theory. Chem. Mater. 2011, 23 (7), 1700–1718.

(144) Saliba, D.; Ammar, M.; Rammal, M.; Al-Ghoul, M.; Hmadeh, M. Crystal Growth of ZIF-8, ZIF-67, and Their Mixed-Metal Derivatives. J. Am. Chem. Soc. 2018, 140 (5), 1812–1823.

(145) Naito, T.; Kakizaki, A.; Inabe, T.; Sakai, R.; Nishibori, E.; Sawa, H. Growth of Nanocrystals in a Single Crystal of Different Materials: A Way of Giving Function to Molecular Crystals. Cryst. Growth Des. 2011, 11 (2), 501–506.

(146) Sekine, Y.; Tonouchi, M.; Yokoyama, T.; Kosaka, W.; Miyasaka, H. Built-in TTF–TCNQ Charge-Transfer Salts in π-Stacked Pillared Layer Frameworks. CrystEngComm 2017, 19 (17), 2300–2304.

(147) Salami, T. O.; Patterson, S. N.; Jones, V. D.; Masello, A.; Abboud, K. A. Synthesis, Characterization, Thermal Study, and Crystal Structure of a New Layered Alkaline Earth Metal Sulfonate: Sr[C\textsubscript{2}H\textsubscript{4}(SO\textsubscript{3})\textsubscript{2}]. Inorg. Chem. Commun. 2009, 12 (11), 1150–1153.

(148) Sun, L.; Hendon, C. H.; Park, S. S.; Tulchinsky, Y.; Wan, R.; Wang, F.; Walsh, A.; Dincă, M. Is Iron Unique in Promoting Electrical Conductivity in MOFs? Chem. Sci. 2017, 8 (6), 4450–4457.

(149) Manna, S. C.; Zanrando, E.; Ribas, J.; Chaudhuri, N. R. Squarato-Bridged Polymeric Networks of Iron(II) with N-Donor Coligands: Syntheses, Crystal Structures and Magnetic Properties. Inorganic Chem. Acta 2005, 358 (15), 4497–4504.

(150) Clements, J. E.; Price, J. R.; Neville, S. M.; Kepert, C. J. Perturbation of Spin Crossover Behavior by Covalent Post-Synthetic Modification of a Porous Metal–Organic Framework. Angew. Chem. Int. Ed. 2014, 126 (38), 10328–10332.

(151) Lee, J.; Asahi, R. Transfer Learning for Materials Informatics Using Crystal Graph Convolutional Neural Network. arXiv. arXiv:2007.09932 2020.

(152) Sanyal, S.; Balachandran, J.; Yadati, N.; Kumar, A.; Rajagopalan, P.; Sanyal, S.; Talukdar, P. MT-CGCNN: Integrating Crystal Graph Convolutional Neural Network with Multitask Learning for Material Property Prediction. arXiv. arXiv:1811.05660 2018.

(153) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. J. Chem. Theory Comput. 2015, 11 (5), 2087–2096.

(154) Bannwarth, C.; Ehler, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. J. Chem. Theory Comput. 2019, 15 (3), 1652–1671.

(155) Spicher, S.; Grimme, S. Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems. Angew. Chem. Int. Ed. 2020, 59 (36), 15665–15673.
Supporting Information:

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

Andrew S. Rosen,1,* Shaelyn M. Iyer,1 Debmalya Ray,2 Zhenpeng Yao,3 Alán Aspuru-Guzik,3,4,5 Laura Gagliardi,5 Justin M. Notestein,1 Randall Q. Snurr1

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

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

Contents

Publicly Available Data ........................................................................................................................................ 2
Methods ............................................................................................................................................................. 2
Dataset Construction ........................................................................................................................................ 2
Dataset Summary ............................................................................................................................................. 2
The DFT-Ready, Free Solvent Removed CSD-42362 Dataset ........................................................................ 3
Completed Job Statistics to Yield the CSD-14204-SP and CSD-14204-opt Datasets ..................................... 5
De-Duplication to Yield the CSD-13058-SP and CSD-13058-opt Datasets .................................................... 5
High-Throughput Periodic DFT Screening ........................................................................................................ 6
Overview ....................................................................................................................................................... 6
Breakdown of Sequential Steps in Periodic DFT Workflow ........................................................................... 6
Computational Resources for Periodic DFT Screening .................................................................................. 7
Further Investigation of Selected MOFs .......................................................................................................... 8
Dataset Handling for Training Machine Learning Models ............................................................................ 9
Learning Curves ................................................................................................................................................ 9
Kernel Ridge Regression ................................................................................................................................ 9
Featureization Methods for KRR ..................................................................................................................... 10
Description: Stoichiometric-120 Features ........................................................................................................ 10
Description: Stoichiometric-45 Features .......................................................................................................... 10
Description: Sine Coulomb Matrix Eigenspectrum ......................................................................................... 11
Description: Orbital Field Matrix .................................................................................................................. 11
Description: Average SOAP Kernel ................................................................................................................ 11
Hyperparameter Tuning for KRR ..................................................................................................................... 13
Crystal Graph Convolutional Neural Networks .............................................................................................. 13
Dimensionality Reduction ............................................................................................................................... 14
Methodological Comments for Future Benchmarking Studies ....................................................................... 14
Additional Results .......................................................................................................................................... 15
Dataset Overview ........................................................................................................................................... 15
High-Spin Fe MOFs ........................................................................................................................................ 16
Comparing Machine Learning Models for Band Gap Prediction .................................................................. 17
Comparing Against ML Band Gap Models for Other Crystalline Materials .................................................. 18
Additional UMAP Results ................................................................................................................................ 18
Electronic Structure of GUTYAW .................................................................................................................. 21
Limitations of Averaging Schemes ................................................................................................................ 21
References ....................................................................................................................................................... 23
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.

Methods

Dataset Construction

Dataset Summary
A summary of the dataset construction 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 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. Initial, un-relaxed structures are analyzed for the de-duplication process.

| Description                                                                 | Name         | De-duplicated |
|----------------------------------------------------------------------------|--------------|---------------|
| Un-optimized, DFT-ready MOF structures.                                     | CSD-42362    | —             |
| DFT results for all calculations that pass the initial single-point        | CSD-19691-SP | —             |
| calculation.                                                               | CSD-14204-opt|               |
|                                                                           | CSD-11204-opt|               |
|                                                                           | CSD-13058-opt|               |
|                                                                           | CSD-19691-SP |               |
DFT results for all calculations that pass all stages of the workflow. Results are for the single-point calculation of the un-optimized structures.

| CSD-14204-SP | CSD-13058-SP |
|---------------|---------------|

DFT results for all calculations that pass all stages of the workflow. Results are for the fully optimized structures.

| CSD-14204-opt | CSD-13058-opt |
|---------------|---------------|

The DFT-Ready, Free Solvent Removed CSD-42362 Dataset

CSD as the starting point:

In this work, we purposefully start with crystal structures deposited in the Cambridge Structural Database (CSD). As discussed below, existing databases of “cleaned” MOF structures that have been widely used for grand canonical Monte Carlo simulations do not consider several structural fidelity issues that would significantly impact the quality of DFT calculations. Furthermore, starting from the unmodified CSD structures makes it possible to use ConQuest² for more complicated filtering stages that take into account CSD meta-data not necessarily present in the individual CIFs.

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 maximize the possibility 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 errors⁴–⁶ 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 ABETAE⁷, BEFLIJ⁸).

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 entirely unphysical. This is a common 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 species on occasion.⁴ The removal of bound solvent can also lead to undesirable charge-balancing issues. For instance, the structure with refcode ASAHEJ⁹ in the “all solvent removed” subset of the CoRE MOF 2019 database is missing its terminal oxo-based 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 of 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 remove any structures that have an atom with missing 3D coordinates. For instance, the material with refcode ADATAC has terminal oxo ligands that should be water molecules. The water molecules are correctly shown in the CSD entry, but the H atoms of the water ligands are not present in the crystal structure. This framework would be flagged as having missing 3D coordinates but is present in the “free solvent removed” subset of the CoRE MOF 2019 database. Similarly, many Zr-containing frameworks are known to have complicated proton topologies, such that it can be impossible to distinguish between terminal oxo, hydroxo, and water ligands from X-ray diffraction alone. In these cases, the terminal ligands would be included without H atoms in the crystal structure, and removal of bound “solvent” and/or free “solvent” would lead to charge-balancing issues. In this case, omitting the structures with missing 3D coordinates ensures these situations are largely removed.

Missing H atoms:
Any structures without 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.

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 (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 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, which has lone H atoms in the pores of the crystal structure (which should actually be H2).

After all of the above procedures, this resulted in 42,292 cleaned 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. The CoRE MOF structures were not used directly in this work. Rather, the corresponding CSD refcodes were identified and run through the aforementioned cleaning procedure to ensure that all structures were cleaned in the same way. 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 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.

CSD-42362:
After the above procedure, we removed 38 additional structure that had disorder or missing H atoms not flagged via the automated ConQuest search, the majority of which have been mentioned in prior work. Finally, this left us with a suitably DFT-ready dataset containing a grand total of 42,362 structures, which
we refer to as the CSD-42362 dataset. The list of refcodes for the CSD-42362 dataset, the script to remove free solvent, and the intermediate lists of refcodes are available with the supporting dataset.\textsuperscript{10} All the CIFs in the CSD-42362 dataset were converted to their Niggli-reduced primitive unit cells using Pymatgen.\textsuperscript{1}

**Completed Job Statistics to Yield the CSD-14204-SP and CSD-14204-opt Datasets**

From the Niggli-reduced CSD-42362 dataset, we specifically chose MOFs with $\leq 150$ atoms for the DFT calculations in this work to ensure that a large number of calculations could be carried out. A total of 24,019 structures fit this criterion and were considered in the high-throughput periodic DFT workflow. Nonetheless, the 42,362 CIFs in the CSD-42362 dataset are considered suitably “clean” for DFT screening purposes.

Of the 24,019 MOFs with $\leq 150$ atoms per Niggli-reduced unit cell considered for the high-throughput periodic DFT screening, 19,691 successfully completed Stage 1 of the workflow (initial single-point calculation), 16,436 of those also completed Stage 2 (coarse accuracy relaxation of atomic positions), and 15,865 of those also completed Stage 3 (medium accuracy relaxation of volume and atomic position). A total of 14,204 successfully completed every step of the workflow. Approximately, 4,800 structures in total led to errors or wall-time limits at some point of the workflow, and any remaining structures were not carried out to completion due to resource constraints. The majority of the “errors” 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 have the largest requirement in terms of 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 19,691 completed single-point calculations as the CSD-19691-SP dataset. By extension, we refer to the 14,204 completed structural relaxations as the CSD-14204-opt dataset. The corresponding single-point data on the starting structures is referred to as the CSD-14204-SP dataset.

**De-Duplication to Yield the CSD-13058-SP and CSD-13058-opt Datasets**

Prior to this point, duplicate structures were not removed, as 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 highly similar (or in the extreme limit, identical) structures can lead to unrealistic testing statistics. Therefore, we used Pymatgen’s StructureMatcher tool (using the default algorithm) on the 14,204 initial, un-relaxed structures to identify a unique subset of 13,058 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\textsuperscript{17} or a comparison of the underlying crystal graphs, could be used to identify unique MOFs based on their building blocks, here we used a structure matching approach so that identical nodes/linkers but different geometries would still be considered 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 (Figures S5e and S5f).\textsuperscript{18} All three are included in the CSD-13058 subset, as they are expected to have different electronic structure properties (indeed, the PBE-D3(BJ)-quality band gaps are 1.70, 1.95, and 0.31 eV, respectively). We acknowledge that no matter what approach is taken, there will nonetheless be a few MOFs in the dataset with essentially identical structures. For instance, RESWUI01\textsuperscript{19} and RESWUI02\textsuperscript{20} are in the CSD-13058-opt dataset and are the same MOF in all regards, but the initial structures did not meet the StructureMatcher duplicate criteria because the H atoms of the methyl groups and water molecules were oriented differently.

Of the 14,204 structures, 13,058 were classified as unique and used for machine learning. The quantum-chemical properties of these 13,058 is referred to as the CSD-13058-SP and CSD-13058-opt dataset and are inherently subsets of the CSD-14204-SP and CSD-14204-opt datasets, respectively.
High-Throughput Periodic DFT Screening

Overview

Plane-wave, periodic density functional theory calculations were carried out using the Vienna \textit{ab initio} Simulation Package (VASP) v.5.4.4.\textsuperscript{21,22} The widely used and computationally tractable PBE exchange-correlation functional\textsuperscript{23} with Grimme’s D3 dispersion correction\textsuperscript{24} and Becke–Johnson (BJ) damping\textsuperscript{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.\textsuperscript{26,27} Based on prior benchmarking work,\textsuperscript{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 $\sim$1000 per number of atoms, as arranged using Pymatgen 2019.9.16.\textsuperscript{1} The VASP-recommended v5.4 projector-augmented wave (PAW)\textsuperscript{22,29} pseudopotentials were considered for all elements, with the exception of Li (for which we used the standard 140 eV default cutoff potential), Eu (for which we use the Eu$_3$ pseudopotential rather than the Eu$_2$ pseudopotential), Yb (for which we use the Yb$_3$ pseudopotential rather than the Yb$_2$ pseudopotential), and W (for which we use the \_sv pseudopotential since the \_pv pseudopotential is not included in the v.5.4 PAW set). All elements have a default cutoff of $\leq$ 520 eV when multiplied by 1.3 (to prevent Pulay stresses upon volume relaxation\textsuperscript{30}).

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.\textsuperscript{31} If the SCF did not converge to $10^{-6}$ eV within 150 iterations, the calculation was aborted and the results not considered for the machine learning process. 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.\textsuperscript{32,33} Here, any $d$-block metal (excluding Zn, Cd, and Hg) were initialized with a magnetic moment of 5 $\mu_B$. All $f$-block elements (excluding Lu and Lr) were initialized with 7 $\mu_B$. 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 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.\textsuperscript{28} These stages include: 1) An initial, high-accuracy single-point calculation; 2) A (coarse accuracy) relaxation of the atomic positions (default plane-wave kinetic energy cutoff and $\sim$100 $k$-points per number of atoms); 3) A medium-accuracy relaxation of the cell volume and atomic positions (520 eV cutoff and $\sim$100 $k$-points per number of atoms); 4) A high-accuracy relaxation of the cell volume and atomic positions (520 eV cutoff, $\sim$1000 $k$-points per number of atoms, and the aforementioned settings); 5) A final 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,\textsuperscript{28} but if for any reason the job crashed and could not be successfully continued, that MOF was also not considered further. The input files for every calculation are provided with the supporting dataset.\textsuperscript{16} The VASP input parameters are also summarized in Table S2. All VASP calculations were carried out using the Atomic Simulation Environment (ASE) 3.19.0b1.\textsuperscript{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\textsuperscript{35–38} as implemented in Chargemol 09-26-2017.\textsuperscript{39} PyMOSFScreen commit #e9768a5 was used to manage and carry out the automated DFT calculations.\textsuperscript{40}
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     | 400     | 520     | 520     | 520     |
| kppa   | 1000    | 100     | 100     | 1000    | 1000    |
| isif   | ---     | 2       | 3       | 3       | 3       |
| ibrion | ---     | 2       | 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     |
| nelm   | 3       | 3       | 3       | 3       | 3       |
| lreal  | False   | False   | False   | False   | False   |
| nsf    | 0       | 250     | 30     | 30     | 0       |
| ediff  | ---     | -0.05   | -0.03   | -0.03   | ---     |
| lorbit | 11      | 11      | 11      | 11      | 11      |
| ismym  | 0       | 0       | 0       | 0       | 0       |
| symrec | 1E-8    | 1E-8    | 1E-8    | 1E-8    | 1E-8    |

*a 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.

*b Switches to lreal='Auto' if the VASP output file suggests doing so due to a large unit cell (only for Stages 2 – 3).

*c The max 250-cycle relaxations of atomic positions were sequentially repeated until the force tolerance given by ediffg (0.05 eV/Å) was achieved.

*d 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.

*e 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) (ibrion=3, iopt=7, potim=0). All input parameters are freely accessible in the provided input files.

** The coarse-accuracy update of the atomic positions is preceded by an initial relaxation using the BFGSLinSearch 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.

Computational Resources for Periodic DFT Screening

High-throughput VASP calculations were carried out on the Stampede2 supercomputer, dedicating 1,530,000 CPU hours (equivalent to ~175 years in wall-time) to this project. While some of these resources (< 5%) were used for debugging, development, and post-processing, the majority was used for the screening study, and calculations were carried out until all resources were consumed. For context, the CPUs used on Stampede2 were Intel Xeon Platinum 8160 (“Skylake”) processors with 48 cores per node. All calculations were carried out with one node per job.

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 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\(^\text{45-47}\) (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 carried out using a smaller \(-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.

| 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.

| Refcode   | \(a\) (Å) | \(b\) (Å) | \(c\) (Å) | \(\alpha\) (°) | \(\beta\) (°) | \(\gamma\) (°) |
|-----------|-----------|-----------|-----------|---------------|---------------|---------------|
| 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 complex that has a low-to-high spin transition with increasing temperature.\(^\text{48}\) 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.\(^\text{49}\) 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
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(C₂H₄(SO₃)₂): This material has Sr(II) cations, and the framework is modeled as spin-restricted based on its structure.

Additional Software and Hardware Details
Regression-based machine learning model development was carried out using scikit-learn v.0.22.1 and the standard SciPy stack with NumPy v.1.18.1, pandas 1.0.3, Matplotlib 3.2.2, and Seaborn v.0.10.0. Pymatgen v.2020.3.13¹ was used throughout this work to analyze structures and generate descriptors. The smooth overlap of atomic positions (SOAP) features were computed using DScribe v.0.4.0. The sine Coulomb matrix and Meredig and Agrawal et al. features were generated using Matminer v.0.6.2. Crystal graph convolutional neural networks (CGCNNs) were based on the work of Xie and Grossman and used PyTorch v.1.4.0 for constructing and evaluating the neural networks. Specifically, our CGCNN code is built upon commit #d612a69 of the CGCNN code with slight variations, as reflected in our fork of the CGCNN code. This fork saves the crystal graphs to .pkl files so they can be read in as-needed instead of 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 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. This is potentially because there are many crystal structures in the CSD-13058 database that are connected in 1D or 2D, such that there are disconnected regions of the Pymatgen-generated crystal graph. PTable Trends v.2.0⁵ 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. PyProcar v.5.5.2 was used to parse the DOS data. Timing data for the machine learning models are reported using Python 3.7 on a laptop with 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 Machine Learning Models
Unless otherwise stated 80% of the 13,058 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, total dataset sizes of 2¹⁰, 2¹¹, 2¹², 2¹³, and the full dataset of 13,058 data points were investigated. Powers of 2 were chosen to allow for equidistant spacing on a logarithmic grid. For the KRR models, 80% of the data was used for training and 20% for testing. For all models involving CGCNN, 80% of the data was used for training, 10% for validation, and 10% for testing.

Kernel Ridge Regression
KRR combines the kernel trick with ridge regression. Like all regression methods, the goal of KRR is to predict a response variable y from a set of individual input vectors x (which, when combined, form a feature matrix 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

\[
\mathbf{w} = (\mathbf{K}_{\text{train}} + \lambda \mathbf{I})^{-1} \mathbf{y}_{\text{train}}
\]

where \( \mathbf{w} \) is the vector of model weights, \( \mathbf{K}_{\text{train}} \) is the training set kernel matrix, \( \lambda \) 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 \( \alpha \) is supplied, which is defined as \( \alpha \equiv \lambda / 2 \).

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

\[
\mathbf{y}_{\text{ML}} = \mathbf{K}_{\text{test}} \mathbf{w}
\]

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

\[
k(x_i, x_j) = \exp \left( -\gamma \|x_i - x_j\|_1 \right)
\]

where \( \gamma \) serves as an adjustable KRR model hyperparameter. In the case of SOAP, a similarity kernel \( \mathbf{K} \) is directly generated as the output 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 \( 0 \rightarrow 1 \).

**Description: Stoichiometric-120 Features**

The Meredig and Agrawal et al.\(^{58}\) feature set (“Stoichiometric-120”) is a composition-based descriptor that was originally developed for formation energy predictions of inorganic solids in the 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 dataset. 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.\(^{69}\) 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.
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 402 out of the 13058 MOF structures in the CSD-13058 set 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\(^{70}\) is a structure-based featurization method where a pair-wise interaction matrix \(M_{ij}\) is generated by the following formula:

\[
M_{ij} = \begin{cases} 
0.5Z_i^2, & i = j \\
\left| B \cdot \sum_{k=x,y,z} \hat{e}_k \sin^2 \left( \pi B^{-1} \cdot (R_i - R_j) \right) \right|, & i \neq j 
\end{cases}
\]

where \(i\) and \(j\) are two atoms in the structure, \(Z_i\) is the atomic number of \(i\), \(B\) is a matrix formed by the lattice vectors, \(\hat{e}_k\) are the Cartesian unit vectors, and \(R_i - R_j\) 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. 150 atoms in the case of the CSD-13058 set). 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 \(n_{\text{max atoms}}\) (i.e. 150 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 \(n_{\text{max atoms}} \times n_{\text{max atoms}}\) (i.e. 22,500 total entries upon flattening).

**Description: Orbital Field Matrix**

The orbital field matrix\(^{71}\) 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. \(s^1, s^2, p^1, p^2, \ldots, p^6, d^1, d^2, \ldots, d^{10}, f^1, f^2, \ldots, f^{14}\)). This is a 32-entry one-hot encoding. As implemented in matminer,\(^{59}\) we supplement 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 \(39 \times 39\) 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 \(39 \times 39\) matrix, which is flattened to a 1521-length encoding. Additional details can be found in the original work by Pham and coworkers.\(^{71}\)

**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\(^{72}\) and the brief summary and implementation of SOAP
in the original DScribe paper,\textsuperscript{57} which we summarize below. We note that we have adopted much of the nomenclature from Musil and coworkers\textsuperscript{73} to clarify the description of the SOAP kernel.

We start by representing a given structure using local atomic densities $\rho$, separately defined for each atomic element $Z$. The local density of atoms within a chemical environment $\chi_i$ (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 $r_{\text{cut}}$. Mathematically, this is expressed as

$$
\rho_{\chi_i}^Z(r) = \sum_k \exp \left( -\frac{|r - R_k|^2}{2\sigma^2} \right)
$$

where $\sigma$ is the standard deviation of the Gaussians, and $|r - R_k|$ describes the distance between atom $k$, $R_k$, and position vector $r$. The origin, $r = 0$, 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 $r_{\text{cut}}$ from atom $i$.

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

$$
\bar{k}(\chi_i^A, \chi_j^B) = \int_{SO(3)} \left| \sum_Z \int_{\mathbb{R}^3} \rho_{\chi_i}^Z(r)\rho_{\chi_j}^Z(r)\,dr \right|^2 \,d\bar{R}
$$

where $SO(3)$ and $\bar{R}$ 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 $\bar{k}$ is carried out by expanding $\rho(r)$ using $n_{\text{max}}$ real spherical harmonic and $\ell_{\text{max}}$ radial basis functions.\textsuperscript{57,72} We use spherical Gaussian type orbitals (GTOs) for the radial basis function in this work. The expression for $\bar{k}$ can be normalized via

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

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 as a matter of convention.

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

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

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

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

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 \( \xi \) 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 \( K \) 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 \( K \) is normalized such that self-similarity is unity (i.e. \( K(A, A) = K(B, B) = 1 \)).

![Image of IRMOF-1, IRMOF-2, and ZIF-8 structures](image)

*Figure S2.* Average (normalized) SOAP similarity kernel for IRMOF-1, IRMOF-2, and ZIF-8. Here, \( r_{cut} = 4 \) 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, and eventually decided upon the use of 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 also chosen.

For the SOAP-based KRR model, there are two KRR parameters to tune: \( \alpha \) and \( \xi \). 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 (\( \ell_{max} \)), and the standard deviation of the Gaussians used to expand the atomic density (\( \sigma \)). 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 \( r_{cut} = 4 \) Å, \( \sigma = 0.1 \) Å, and \( n_{max} = \ell_{max} = 9 \). All other SOAP hyperparameters and settings were set to the default values in DScribe.

**Crystal Graph Convolutional Neural Networks**

A 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 by Xie and Grossman.\(^6\)
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 and iCGCNN, did not show notable improvements over the original CGCNN implementation, although a detailed exploration of the hyperparameter space was not carried out.

In the original CGCNN work, 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. 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; 5) removed the atomic volume feature; 6) removed the electron affinity feature. Tabulated values were taken from mendeleev v.0.5.2. 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.

**Dimensionality Reduction**

Dimensionality reduction was carried out via the uniform manifold approximation and projection (UMAP) algorithm as implemented in umap v.0.4.6. 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). 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. 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 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 $D$ by invoking the following metric:

$$D_{ij} = \sqrt{K_{ii} + K_{jj} - 2K_{ij}}$$

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, $\text{max}(Z)$, is not a feature in Stoichiometric-120, it can be directly related to the range($Z$) feature since $\text{min}(Z) = 1$ in every MOF. For this reason, we use the more intuitive $\text{max}(Z)$ feature in Figure 6a.

**Methodological Comments for Future Benchmarking Studies**

One of the motivations for the dataset in this work 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 goal 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 can influence the machine learning process. 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 dataset 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. The most common scenario is likely omitted/additional H atoms, which are particularly difficult to flag if not already specified in the CSD entry. While many of these instances are likely to be filtered out either via the workflow in Figure S1 or due to SCF convergence issues when the DFT calculations were performed, additional cleansing steps are always possible and are encouraged for new applications of interest.

**Additional Results**

**Dataset Overview**

The average band gaps associated with each element are shown in Figure S3. The partial charges before and after optimization are shown in Figure S4a, showing that the values do not substantially change upon structure relaxation. We also plot the cumulative frequency of absolute deviations in Figure S4b.

**Figure S3.** Average DFT-computed band gap, $\tilde{E}_g$, for MOFs containing a given metal element in the CSD-13058-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 S4.** (a) Parity plot comparing the DDEC6 partial atomic charges for the CSD-14204-opt ($q_{opt}$) and CSD-14204-SP ($q_{SP}$) datasets. (b) The cumulative fraction of DDEC6 partial atomic charges in the CSD-14204-opt dataset that are within some tolerance, given by $|q_{opt} - q_{SP}|$, of the CSD-14204-SP dataset.
High-Spin Fe MOFs

Figure S5. MOFs in the CSD-13058-opt dataset with high-spin Fe sites and a pore-limiting diameter greater than 3.6 Å. a) Fe(bpz) (refcode: ACODAA); b) Fe$_2$(dobdc) (refcode: COKNOH); c) Fe$_2$(dobdc) with bound propylene (refcode: SARHAW); d) Fe$_2$Cl$_2$(bbta) (refode: HAYYUE); e) Fe(bdp) (refcode: QUPZIM01); f) Fe(bdp) – distorted (refcode: QUPZIM02). Color key: Fe (orange), N (blue), O (red), Cl (green), C (gray), H (white).
Comparing Machine Learning Models for Band Gap Prediction

Figure S6. 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, $E_{g, DFT}$. 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 $E_{g, ML}$ and $E_{g, DFT}$ 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. 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. 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. Particularly relevant for the present study, Olsthoorn et al. used a weighted average of SOAP- and SchNet-based ML regression models trained on 10,000 band gaps of organic crystals in the Organic Materials Database (OMDB) to achieve a testing MAE of 0.39 eV. The band gaps in the organic crystal 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. The CGCNN model in the present study has a testing set MAE of 0.27 eV when trained on 10,446 MOFs, suggesting that it performs slightly better than the aforementioned models for other crystalline materials.

Additional UMAP Results

It is also worth investigating the degree of overlap in feature space between the CSD-13058-SP dataset and the parent CSD-42362 dataset that the former was drawn from. To carry out this analysis, we used UMAP to project the feature space of the CSD-42362 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 CSD-13058-SP dataset. As shown in Figure S7, there is significant overlap between the two datasets, such that we can expect ML models trained on the CSD-13058-SP dataset to be applicable to other MOFs deposited in the CSD. We carried out a similar analysis to compare the CSD-13058-SP dataset with the CoRE MOF 2019 (v.1.1.3) database. 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 S8).

Figures S9 and S10 show Stoichiometric-120- and SOAP-based UMAPs generated for the CSD-13058-SP 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 order of increasing max(Z) (Figure 6a, Figure S9). The edge connections for the SOAP-based UMAP also makes the local regions, and their connectivity, clearer.
Figure S7. 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 CSD-42362 dataset. The CSD-13058-SP subset is overlaid onto the projection.

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 CSD-13058-SP and CoRE MOF 2019 (free solvent removed) databases.
**Figure S9.** 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 CSD-13058-SP dataset. The connectivity between points are shown. Brighter colors indicate a greater density of connections.

**Figure S10.** Structural dimensionality reduction performed using UMAP, with a distance matrix obtained from the average SOAP similarity kernel of the structures in the CSD-13058-SP dataset. The connectivity between points are 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[C₂H₄(SO₃)₂] (refcode: GUTYAW87) is shown in Figure S11.

Figure S11. a) Structure of Sr[C₂H₄(SO₃)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 are encoded for machine learning. In the case of an average 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 Zn₄O(bdc)₃ (bdc = benzene-1,4-dicarboxylate), IRMOF-2 with the formula Zn₄O(bdc-Br)₃, IRMOF-10 with the formula Zn₄O(bpdc) (bpdc = 4,4′-biphenyldicarboxylate), Zn₂(dobdc) (dobdc = 2,5-dihydroxybenzene-1,4-dicarboxylate), and MFU-4l (MFU = Metal-Organic Framework Ulm University, \( l \) = large) with the formula Zn₅Cl₄(btdd)₃ (btdd = bis(1,2,3-triazolato-[4,5-b],[4′,5′-i])dibenzo-[1,4]-dioxin) (Figure S12). 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 S5). In fact, IRMOF-1 and Zn₂(dobdc) are more similar than IRMOF-1 and IRMOF-2 based on the average SOAP kernel (Table S5). This can likely be traced back to the similarity of the linkers (bdc vs. dobdc) in IRMOF-1 and Zn₂(dobdc) despite their very
different inorganic nodes (Zn₄O 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. 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.

![Crystal structures of IRMOF-1, IRMOF-2, IRMOF-10, Zn₂(dobdc) (also known as Zn-MOF-74 and Zn-CPO-27), and MFU-4l.](image)

**Figure S12.** Crystal structures of IRMOF-1, IRMOF-2, IRMOF-10, Zn₂(dobdc) (also known as Zn-MOF-74 and Zn-CPO-27), and MFU-4l.

**Table S5.** Average (normalized) SOAP similarity kernel for IRMOF-1, IRMOF-2, IRMOF-10, Zn₂(dobdc), and MFU-4l. Here, \( r_{\text{cut}} = 4 \text{ Å}, \sigma = 0.1 \text{ Å}, \xi = 2, \) and \( n_{\text{max}} = \ell_{\text{max}} = 9. \)

|          | IRMOF-1 | IRMOF-2 | IRMOF-10 | Zn₂(dobdc) | MFU-4l |
|----------|---------|---------|----------|------------|--------|
| 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   |
| Zn₂(dobdc)| 0.92    | 0.94    | 0.29     | 1.00       | 0.64   |
| MFU-4l   | 0.73    | 0.72    | 0.35     | 0.64       | 1.00   |
References

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

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

(3) Moghadam, P. Z.; Li, A.; Wiggin, S. B.; Tao, A.; Maloney, A. G. P.; Wood, P. A.; Ward, S. C.; Fairen-Jimenez, D. Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future. *Chem. Mater.* **2017**, *29* (7), 2618–2625.

(4) Altintas, C.; Avci, G.; Daglar, H.; Azar, A. N. V.; Erucar, I.; Velioglu, S.; Keskin, S. An Extensive Comparative Analysis of Two MOF Databases: High-Throughput Screening of Computation-Ready MOFs for CH$_4$ and H$_2$ Adsorption. *J. Mater. Chem. A* **2019**, *7* (16), 9593–9608.

(5) Velioglu, S.; Keskin, S. Revealing the Effect of Structure Curation on the Simulated CO$_2$ Separation Performances of MOFs. *Mater. Adv.* **2020**, *1* (3), 341–353.

(6) Chen, T.; Manz, T. A. Identifying Misbonded Atoms in the 2019 CoRE Metal–Organic Framework Database. *RSC Adv.* **2020**, *10* (45), 26944–26951.

(7) Jouffret, L.; Rivenet, M.; Abraham, F. Linear Alkyl Diamine-Uranium-Phosphate Systems: U(VI) to U(IV) Reduction with Ethylenediamine. *Inorg. Chem.* **2011**, *50* (10), 4619–4626.

(8) Feng, P.; Bu, X.; Stucky, G. D. Hydrothermal Syntheses and Structural Characterization of Zeolite Analogue Compounds Based on Cobalt Phosphate. *Nature* **1997**, *388* (6644), 735–741.

(9) Liang, L.; Zhang, R.; Zhao, J.; Liu, C.; Weng, N. S. Two Actinide-Organic Frameworks Constructed by a Tripodal Flexible Ligand: Occurrence of Infinite \{(UO$_2$)$_3$(OH)$_3$\}$^4_\text{n}$ and Hexanuclear \{Th$_6$O$_4$(OH)$_4$\} Motifs. *J. Solid State Chem.* **2016**, *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.; Eddaoudi, M. On Demand: The Singular rht Net, an Ideal Blueprint for the Construction of a Metal–Organic Framework (MOF) Platform. *Angew. Chem.* **2012**, *124* (40), 10246–10250.

(11) Planas, N.; Mondloch, J. E.; Tussupbayev, S.; Borycz, J.; Gagliardi, L.; Hupp, J. T.; Farha, O. K.; Cramer, C. J. Defining the Proton Topology of the Zr$_6$-Based Metal–Organic Framework NU-1000. *J. Phys. Chem. Lett.* **2014**, *5* (21), 3716–3723.

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

(13) 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.; Long, J. R. Hydrogen Adsorption in the Metal–Organic Frameworks Fe$_2$(dobdc) and Fe$_2$(O$_2$(dobdc). *Dalt. Trans.* **2012**, *41* (14), 4180–4187.

(14) Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S.; Slater, B.; Siepmann, J. I.; Sholl, D. S.; Snurr, R. Q. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data* **2019**, *64* (12), 5985–5998.

(15) Li, A.; Bueno-Perez, R.; Wiggins, S.; Fairen-Jimenez, D. Enabling Efficient Exploration of Metal–Organic Frameworks in the Cambridge Structural Database. *CrystEngComm* **2020**, https://doi.org/10.1039/D0CE00299B.

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

(17) Bucior, B. J.; Rosen, A. S.; Haranczyk, M.; Yao, Z.; Ziebel, M. E.; Farha, O. K.; Hupp, J. T.; Siepmann, J. I.; Aspuru-Guzik, A.; Snurr, R. Q. Identification Schemes for Metal–Organic Frameworks to Enable Rapid Search and Cheminformatics Analysis. *Cryst. Growth Des.* **2019**, *19* (11), 6682–6697.
(18) Mason, J. A.; Oktawiec, J.; Taylor, M. K.; Hudson, M. R.; Rodriguez, J.; Bachman, J. E.; Gonzalez, M. I.; Cervellino, A.; Guagliardi, A.; Brown, C. M.; Llewellyn, P. L.; Masciocchi, N.; Long, J. R. Methane Storage in Flexible Metal–Organic Frameworks with Intrinsic Thermal Management. *Nature* 2015, 527 (7578), 357–361.

(19) He, H.; Yuan, D.; Ma, H.; Sun, D.; Zhang, G.; Zhou, H.-C. Control over Interpenetration in Lanthanide–Organic Frameworks: Synthetic Strategy and Gas-Adsorption Properties. *Inorg. Chem.* 2010, 49 (17), 7605–7607.

(20) He, H.; Ma, H.; Sun, D.; Zhang, L.; Wang, R.; Sun, D. Porous Lanthanide–Organic Frameworks: Control over Interpenetration, Gas Adsorption, and Catalyst Properties. *Cryst. Growth Des.* 2013, 13 (7), 3154–3161.

(21) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* 1996, 54 (16), 11169–11186.

(22) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B* 1999, 59 (3), 1758–1775.

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

(24) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H–Pu. *J. Chem. Phys.* 2010, 132 (15), 154104.

(25) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* 2011, 32 (7), 1456–1465.

(26) Nazarian, D.; Ganesh, P.; Sholl, D. S. Benchmarking Density Functional Theory Predictions of Framework Structures and Properties in a Chemically Diverse Test Set of Metal–Organic Frameworks. *J. Mater. Chem. A* 2015, 3 (4), 22432–22440.

(27) Formalik, F.; Fischer, M.; Rogacka, J.; Firlej, L.; Kuchta, B. Benchmarking of GGA Density Functionals for Modeling Structures of Nanoporous, Rigid and Flexible MOFs. *J. Chem. Phys.* 2018, 149, 064110.

(28) Rosen, A. S.; Notestein, J. M.; Snurr, R. Q. Identifying Promising Metal–Organic Frameworks for Heterogeneous Catalysis via High-Throughput Periodic Density Functional Theory. *J. Comput. Chem.* 2019, 40 (12), 1305–1318.

(29) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B* 1994, 50 (24), 17953–17979. https://doi.org/10.1103/PhysRevB.50.17953.

(30) Pulay, P. Convergence Acceleration of Iterative Sequences. The Case of SCF Iteration. *Chem. Phys. Lett.* 1980, 73 (2), 393–398.

(31) Kresse, G.; Furthmüller, J. Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Comput. Mater. Sci.* 1996, 6 (1), 15–50.

(32) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aylol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): Assessing the Accuracy of DFT Formation Energies. *npj Comput. Mater.* 2015, 1 (1), 1–15.

(33) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* 2013, 1 (1), 11002.

(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.; Maxton, T.; Olsen, T.; Pastewka, L.; Peterson, A.; Rostgaard, C.; Schütz, J.; Schütz, O.; Strange, M.; Thygesen, K.; Vegge, T.; Vilhelm, S.; Walter, M.; Zeng, Z.; Jacobsen, K. The Atomic Simulation Environment—A Python Library for Working with Atoms. *J. Phys. Condens. Matter* 2017, 29, 273002.

(35) Manz, T. A.; Limas, N. G. Introducing DDEC6 Atomic Population Analysis: Part 1. Charge Partitioning Theory and Methodology. *RSC Adv.* 2016, 6 (53), 47771–47801.
(36) Limas, N. G.; Manz, T. A. Introducing DDEC6 Atomic Population Analysis: Part 2. Computed Results for a Wide Range of Periodic and Nonperiodic Materials. *RSC Adv.* **2016**, *6* (51), 45727–45747.

(37) Manz, T. A. Introducing DDEC6 Atomic Population Analysis: Part 3. Comprehensive Method to Compute Bond Orders. *RSC Adv.* **2017**, *7* (72), 45552–45581.

(38) Limas, N. G.; Manz, T. A. Introducing DDEC6 Atomic Population Analysis: Part 4. Efficient Parallel Computation of Net Atomic Charges, Atomic Spin Moments, Bond Orders, and More. *RSC Adv.* **2018**, *8* (5), 2678–2707.

(39) Manz, T. A.; Gabaldon Limas, N. Chargemol program for performing DDEC analysis http://ddec.sourceforge.net/.

(40) Rosen, A. S.; Notestein, J. M.; Snurr, R. Q. PyMOFScreen https://doi.org/10.5281/zenodo.1451873.

(41) Bitzek, E.; Koskinen, P.; Gähler, F.; Moseler, M.; Gumbsch, P. Structural Relaxation Made Simple. *Phys. Rev. Lett.* **2006**, *97* (17), 170201.

(42) Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazlewood, V.; Lathrop, S.; Lifka, D.; Peterson, G. D.; Roskies, R.; Scott, J. R.; Wilkins-Diehr, N. XSEDE: Accelerating Scientific Discovery. *Comput. Sci. Eng.* **2014**, *16* (5), 62–74.

(43) Krukau, A. V; Vydrov, O. A.; Izmaylov, A. F.; Scuseria, G. E. Influence of the Exchange Screening Parameter on the Performance of Screened Hybrid Functionals. *J. Chem. Phys.* **2006**, *125* (22), 224106.

(44) Moellmann, J.; Grimme, S. DFT-D3 Study of Some Molecular Crystals. *J. Phys. Chem. C* **2014**, *118* (14), 7615–7621.

(45) Freysoldt, C.; Boeck, S.; Neugebauer, J. Direct Minimization Technique for Metals in Density Functional Theory. *Phys. Rev. B* **2009**, *79* (24), 241103.

(46) Teter, M. P.; Payne, M. C.; Allan, D. C. Solution of Schrödinger’s Equation for Large Systems. *Phys. Rev. B* **1989**, *40* (18), 12255–12263.

(47) Bylander, D. M.; Kleinman, L.; Lee, S. Self-Consistent Calculations of the Energy Bands and Bonding Properties of B12C3. *Phys. Rev. B* **1990**, *42* (2), 1394–1403.

(48) Clements, J. E.; Price, J. R.; Neville, S. M.; Kepert, C. J. Perturbation of Spin Crossover Behavior by Covalent Post-Synthetic Modification of a Porous Metal–Organic Framework. *Angew. Chem. Int. Ed.* **2014**, *126* (38), 10328–10332.

(49) Manna, S. C.; Zangrado, E.; Ribas, J.; Chaudhuri, N. R. Squarato-Bridged Polymeric Networks of Iron(II) with N-Donor Coligands: Syntheses, Crystal Structures and Magnetic Properties. *Inorganica Chim. Acta* **2005**, *358* (15), 4497–4504.

(50) Sekine, Y.; Tonouchi, M.; Yokoyama, T.; Kosaka, W.; Miyasaka, H. Built-in TTF–TCNQ Charge-Transfer Salts in π-Stacked Pillared Layer Frameworks. *CrystEngComm 2017*, *19* (17), 2300–2304.

(51) 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.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(52) 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.; Illian, 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 Mulbregt, P.; Contributors, S. 1. . SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272.
Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E. Array Programming with NumPy. *Nature* **2020**, *585* (7825), 357–362.

(54) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9* (3), 90–95.

(55) McKinney, W. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*; 2010; Vol. 445, pp 51–56.

(56) Seaborn http://doi.org/10.5281/zenodo.592845.

(57) Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of Descriptors for Machine Learning in Materials Science. *Comput. Phys. Commun.* **2020**, 247, 106949.

(58) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial Screening for New Materials in Unconstrained Composition Space with Machine Learning. *Phys. Rev. B* **2014**, *89* (9), 94104.

(59) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E. R.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystron, K.; Dylla, M.; Chard, K.; Asta, M.; Persson, K. A.; Snyder, J.; Foster, I.; Jain, A. Matminer: An Open Source Toolkit for Materials Data Mining. *Comput. Mater. Sci.* **2018**, *152*, 60–69.

(60) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120* (14), 145301.

(61) 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.; Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*; 2019; pp 8024–8035.

(62) Xie, T.; Grossman, J. Crystal Graph Convolutional Neural Networks https://github.com/txie-93/cgcnn.

(63) Rosen, A. S.; Notestein, J. M.; Snurr, R. Q. Crystal Graph Convolutional Neural Networks https://github.com/snurr-group/cgcn.

(64) Rosen, A. S.; Notestein, J. M.; Snurr, R. Q. Crystal Graph Convolutional Neural Networks https://github.com/snurr-group/cgcn/tree/scatter.

(65) Rosen, A. S. PTable Trends. https://doi.org/10.5281/zenodo.1451879.

(66) Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M. Algorithms and Tools for High-Throughput Geometry-Based Analysis of Crystalline Porous Materials. *Microporous Mesoporous Mater.* **2012**, *149* (1), 134–141.

(67) Herath, U.; Tavadze, P.; He, X.; Bousquet, E.; Singh, S.; Munoz, F.; Romero, A. H. PyProcar: A Python Library for Electronic Structure Pre/Post-Processing. *Comput. Phys. Commun.* **2020**, *251*, 107080.

(68) Murphy, K. P. *Machine Learning: A Probabilistic Perspective*; MIT press, 2012.

(69) He, Y.; Cubuk, E. D.; Allendorf, M. D.; Reed, E. J. Metallic Metal–Organic Frameworks Predicted by the Combination of Machine Learning Methods and Ab Initio Calculations. *J. Phys. Chem. Lett.* **2018**, *9* (6), 4562–4569.

(70) Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Crystal Structure Representations for Machine Learning Models of Formation Energies. *Int. J. Quantum Chem.* **2015**, *115* (16), 1094–1101.

(71) Lam Pham, T.; Kino, H.; Terakura, K.; Miyake, T.; Tsuda, K.; Takigawa, I.; Chi Dam, H. Machine Learning Reveals Orbital Interaction in Materials. *Sci. Technol. Adv. Mater.* **2017**, *18* (1), 756–765.

(72) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing Molecules and Solids across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* **2016**, *18* (20), 13754–13769.

(73) Musil, F.; De, S.; Yang, J.; Campbell, J. E.; Day, G. M.; Ceriotti, M. Machine Learning for the Structure–Energy–Property Landscapes of Molecular Crystals. *Chem. Sci.* **2018**, *9* (5), 1289–1300.

(74) Noh, J.; Gu, G. H.; Kim, S.; Jung, Y. Uncertainty-Quantified Hybrid Machine Learning/Density
Functional Theory High Throughput Screening Method for Crystals. *J. Chem. Inf. Model.* 2020.

(75) Park, C. W.; Wolverton, C. Developing an Improved Crystal Graph Convolutional Neural Network Framework for Accelerated Materials Discovery. *arXiv.* *arXiv:1906.05267* 2019.

(76) Issue #2: atom_init.json https://github.com/txie-93/cgcnn/issues/2.

(77) Cordero, B.; Gómez, V.; Platero-Prats, A. E.; Revés, M.; Echeverría, J.; Cremades, E.; Barragán, F.; Alvarez, S. Covalent Radii Revisited. *Dalton Trans.* 2008, No. 21, 2832–2838.

(78) Mentel, L. Mendeleev - A Python Resource for Properties of Chemical Elements, Ions and Isotopes. 2014.

(79) McInnes, L.; Healy, J.; Melville, J. Umap: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv.* *arXiv:1802.03426* 2018.

(80) Leland, M.; John, H.; Nathaniel, S.; Lukas, G. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* 2018, 3 (29), 861.

(81) Hurter, C.; Ersoy, O.; Telea, A. Graph Bundling by Kernel Density Estimation. In *Computer Graphics Forum*; 2012; Vol. 31, pp 865–874.

(82) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* 2019, 31 (9), 3564–3572.

(83) Louis, S.-Y. M.; Zhao, Y.; Nasiri, A.; Wang, X.; Song, Y.; Liu, F.; Hu, J. Graph Convolutional Neural Networks with Global Attention for Improved Materials Property Prediction. *Phys. Chem. Chem. Phys.* 2020, 22, 18141–18148.

(84) Olsthoorn, B.; Geilhufe, R. M.; Borysov, S. S.; Balatsky, A. V. Band Gap Prediction for Large Organic Crystal Structures with Machine Learning. *Adv. Quantum Technol.* 2019, 2 (7–8), 1900023.

(85) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet – A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* 2018, 148 (24), 241722.

(86) Borysov, S. S.; Geilhufe, R. M.; Balatsky, A. V. Organic Materials Database: An Open-Access Online Database for Data Mining. *PLoS One* 2017, 12 (2).

(87) Salami, T. O.; Patterson, S. N.; Jones, V. D.; Masello, A.; Abboud, K. A. Synthesis, Characterization, Thermal Study, and Crystal Structure of a New Layered Alkaline Earth Metal Sulfonate: Sr[C_2H_4(SO_3)_2]. *Inorg. Chem. Commun.* 2009, 12 (11), 1150–1153.
Database of 14,000+ MOFs

Quantum-chemical calculations

Machine learning

EXPLORATION OF MOF CHEMICAL SPACE