Title

High-throughput predictions of metal–organic framework electronic properties: theoretical challenges, graph neural networks, and data exploration

Journal

npj Computational Materials, 8(1)

ISSN

2057-3960

Authors

Rosen, Andrew S
Fung, Victor
Huck, Patrick
et al.

Publication Date

2022

DOI

10.1038/s41524-022-00796-6

Peer reviewed

npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

High-throughput predictions of metal–organic framework electronic properties: theoretical challenges, graph neural networks, and data exploration

Andrew S. Rosen[1,2,3,4]✉, Victor Fung[5], Patrick Huck[6], Cody T. O'Donnell[3], Matthew K. Horton[3], Donald G. Truhlar[7], Kristin A. Persson[1,8], Justin M. Notestein[4] and Randall Q. Snurr[4]

With the goal of accelerating the design and discovery of metal–organic frameworks (MOFs) for electronic, optoelectronic, and energy storage applications, we present a dataset of predicted electronic structure properties for thousands of MOFs carried out using multiple density functional approximations. Compared to more accurate hybrid functionals, we find that the widely used PBE generalized gradient approximation (GGA) functional severely underpredicts MOF band gaps in a largely systematic manner for semi-conductors and insulators without magnetic character. However, an even larger and less predictable disparity in the band gap prediction is present for MOFs with open-shell 3*d* transition metal cations. With regards to partial atomic charges, we find that different density functional approximations predict similar charges overall, although hybrid functionals tend to shift electron density away from the metal centers and onto the ligand environments compared to the GGA point of reference. Much more significant differences in partial atomic charges are observed when comparing different charge partitioning schemes. We conclude by using the dataset of computed MOF properties to train machine-learning models that can rapidly predict MOF band gaps for all four density functional approximations considered in this work, paving the way for future high-throughput screening studies. To encourage exploration and reuse of the theoretical calculations presented in this work, the curated data is made publicly available via an interactive and user-friendly web application on the Materials Project.

npj Computational Materials (2022)8:112 ; https://doi.org/10.1038/s41524-022-00796-6

INTRODUCTION

Metal–organic frameworks (MOFs) have been extensively studied over the last two decades due to their high degree of synthetic tunability, which makes it possible to tailor their physical and chemical properties for a given application[1,2]. While much attention has been focused on the use of MOFs for industrial gas storage and separations[3,4], the design of MOFs with targeted electronic properties has become a topic of recent interest as well[5–8]. Through a judicious selection of inorganic nodes and organic linkers, MOFs have been proposed for novel electronic and optoelectronic devices, electrocatalysts, photocatalysts, sensors, and energy storage devices, among many other applications[6,9–11]. However, with tens of thousands of MOFs that have been experimentally synthesized[12] and virtually unlimited more that can be proposed[13], it is often difficult to identify promising MOF candidates with the optimal set of electronic properties.

The advent of machine learning (ML) and related big data approaches has made it possible to more efficiently search through MOF chemical space, and high-throughput computational screening can often provide insight into previously unknown structure–function relationships[14–22]. With this goal in mind, a high-throughput density functional theory (DFT) workflow[23] was recently used to construct a publicly accessible dataset of quantum-chemical properties for thousands of MOFs (and coordination polymers), known as the Quantum MOF (QMOF) Database[24]. Like many databases of material properties generated from high-throughput periodic DFT calculations[25,26], the electronic structure properties within the QMOF Database were computed with the relatively inexpensive Perdew–Burke–Ernzerhof (PBE)[27] exchange-correlation functional. While PBE is useful for generating large quantities of material property data that are often needed for ML, the electron self-interaction error[28] of generalized gradient approximation (GGA) functionals like PBE can greatly influence the predicted electronic properties[28,29]. Perhaps most notably, PBE is known to severely underpredict band gaps[30–32], but the degree to which there may be qualitative (as opposed to merely quantitative) errors is not well-established. This inherently limits the practical utility of data-driven, computational screening approaches based on such a functional.

For inorganic solids, several approaches have been taken to increase the accuracy of ML-predicted band gaps trained on high-throughput DFT calculations in a computationally tractable manner. The most straightforward option is to train ML models on experimental band gap data[33] or an ensemble of both theoretical and experimental band gap data[34]. Unfortunately, this approach is challenging to apply to MOFs because there are relatively few reports of experimentally measured MOF band gaps[8]. Furthermore, the reported band gaps of MOFs can vary by several tenths of an eV depending on the synthesis conditions and crystallinity of the material[6]. Another approach is to carry out higher-accuracy DFT calculations on a subset of materials and use

[1]Department of Materials Science and Engineering, University of California, Berkeley, CA 94720, USA. [2]Miller Institute for Basic Research in Science, University of California, Berkeley, Berkeley, CA 94720, USA. [3]Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [4]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208, USA. [5]Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. [6]Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [7]Department of Chemistry, Chemical Theory Center, and Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, MN 55455, USA. [8]Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.
✉email: rosen@u.northwestern.edu

them to train an ML model that can make more reliable predictions. Recently, large datasets of band gaps computed with meta-GGA and hybrid functionals have been published for inorganic solids[35–37], although no such resource currently exists for MOFs.

In the present work, we complement the existing dataset of PBE electronic structure properties in the QMOF Database with analogous data computed using three other functionals: HLE17[38] (a high local-exchange meta-GGA), HSE06[39,40] (a screened-exchange hybrid GGA), and a functional we refer to here as HSE06* in which the amount of screened Hartree–Fock (HF) exchange of HSE06 has been changed from 25% at short interelectronic distances to 10%. By analyzing the electronic structure properties calculated at these levels of theory, we uncover severe theoretical limitations associated with the more computationally efficient (meta-)GGA density functionals that prevent them from achieving quantitatively—and sometimes qualitatively—accurate band gap predictions for MOFs and coordination polymers with respect to hybrid functionals. Since it is known that different density functional approximations (DFAs) can alter the underlying charge density, we also investigated trends related to the computed partial atomic charges. In general, we find that the different levels of theory predict similar partial atomic charges; however, as compared to PBE, the meta-GGA and screened hybrids tend to shift electron density away from the metal centers and onto the ligand environments.

We conclude by using the electronic structure data to train multi-task and multi-fidelity convolutional neural network models that can predict PBE, HLE17, HSE06, and HSE06* band gaps given a graph-based representation of a MOF crystal structure. We anticipate that the computational data, trends, and subsequent deep learning models presented in this work will make it possible to achieve both rapid and accurate predictions of MOF band gaps that can greatly accelerate the materials design and discovery process. To help realize this vision, all the data underlying the QMOF Database is now also made available as a dedicated, interactive application on the widely used Materials Project[41].

RESULTS

Band gap comparison

To develop ML models that can directly guide future experimental efforts, it is essential to first understand the behavior and potential limitations of various levels of theory when predicting MOF electronic structure properties. As such, we begin by comparing the DFT-predicted band gaps for 10,720 structures in the QMOF Database with the PBE (GGA: 0% HF exchange), HLE17 (meta-GGA: 0% HF exchange), HSE06* (screened hybrid: 10% HF exchange at small interelectronic distances decreasing to zero at large distance), and HSE06 (screened hybrid: 25% HF exchange at small interelectronic distances decreasing to zero at large distance) functionals.

As shown in Fig. 1, we observe pronounced differences amongst the predictions of the various DFAs. Starting with the box plots, we find that of the four functionals tested in this work, PBE generally predicts the lowest band gaps. Including HF exchange—as with HSE06* and HSE06—tends to increase the predicted band gap values (as expected[42]), with the relative increase depending on the fraction of HF exchange in the selected functional. Qualitatively, the HSE06* and HSE06 results are more reflective of prior experimental studies[6], which suggest that the majority of MOFs are electronically insulating and that comparatively few exhibit semi-conducting or metallic character. Switching focus to the HLE17 meta-GGA, we find that the median band gap value is within 0.09 eV of the HSE06* calculations, suggesting that the parameterization of this functional can partially improve upon the band gap underprediction problem of PBE despite not incorporating HF exchange.

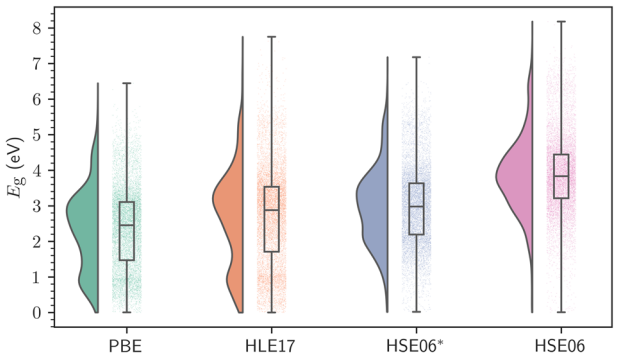

Fig. 1 Distribution of band gap data at four levels of theory. Raincloud plots (i.e., combined violin plot, box plot, and strip plot) for the DFT-computed band gaps, E_g, of 10,720 structures in the QMOF Database at the PBE, HLE17, HSE06*, and HSE06 levels of theory. The strip plots show all the data at that level of theory (jittered horizontally for ease-of-visualization). The box plots show the extrema (whisker tails), interquartile range (box boundaries), and median (horizontal line). The violin plots show the probability density of the data.

When comparing the violin plots in Fig. 1, it is immediately clear that the shape of the band gap distribution can vary significantly depending on the DFA. The PBE-computed band gap data exhibits two distinct distributions with peaks around 0.90 eV and 2.93 eV (Fig. 1), which is observed for the full QMOF Database of ~20,000 structures as well (Supplementary Fig. 6). A qualitatively similar distribution of band gaps is obtained when using the HLE17 functional, which has peaks around 0.86 eV and 3.21 eV. However, the two distributions in the band gap data exhibit much more significant overlap for the HSE06* functional, and for the HSE06 functional there is almost complete overlap such that the overall distribution is virtually unimodal.

The two underlying distributions in the band gap data can be better understood by separating the computed values based on whether the material has closed-shell or open-shell character, the latter of which is associated with lower band gaps on average (Fig. 2a). When including 10% HF exchange with HSE06*, the degree of overlap between the closed-shell and open-shell band gap distributions is partway between that of PBE and HSE06 (Fig. 2a), which illustrates the strong dependence of the trends on the fraction of HF exchange. Taking the hybrid-quality calculations as the more accurate reference point[43], these findings suggest that the PBE functional exhibits severe quantitative and qualitative shortcomings when applied to a wide range of MOF structures and that these shortcomings go beyond a simple underprediction of the band gap. Although HLE17 increases the median band gap of the dataset compared to PBE and decreases the number of structures with a predicted band gap in the low-energy subset, it retains the bimodal nature of the band gap distribution. Nonetheless, HLE17 does significantly increase the band gaps of the closed-shell frameworks, and the distribution of band gaps for the closed-shell MOFs is similar to that of HSE06*.

By directly comparing the predicted band gaps for the PBE, HSE06*, and HSE06 calculations, we find that there is a correlation between the median band gap and the fraction of HF exchange (Fig. 2b), at least within the range of 0–25% HF exchange considered in this work. Assuming linear behavior in this region, it can be concluded that the median band gap across the dataset changes by ~0.05 eV per percent of HF exchange for the closed-shell frameworks and ~0.10 eV per percent of HF exchange for the open-shell frameworks, although we emphasize that these statistics are specific to the QMOF Database and may differ for other datasets of MOFs. Collectively, these results have significant

implications for computational screening studies of MOFs and coordination polymers, as the use of GGA functionals like PBE may lead to incorrect qualitative comparisons between the band gaps of different materials if some have closed-shell character and others have open-shell character.

While Figs. 1 and 2 show how the entire dataset changes with different density functionals, it is also important to investigate the degree of correlation between the various functionals. As shown in Fig. 3, nearly every MOF has a larger predicted band gap with the HSE06* (Fig. 3b) and HSE06 (Fig. 3c) functionals than with PBE. This is also the case for most of the closed-shell MOFs with the HLE17 functional, especially when $E_{g,PBE}$ is above ~1.5 eV (Fig. 3a). For the closed-shell frameworks (Supplementary Fig. 7), there is a linear correlation between the computationally inexpensive PBE-quality band gaps and those calculated with the more accurate HSE06* and HSE06 functionals as well as the HLE17 functional. As

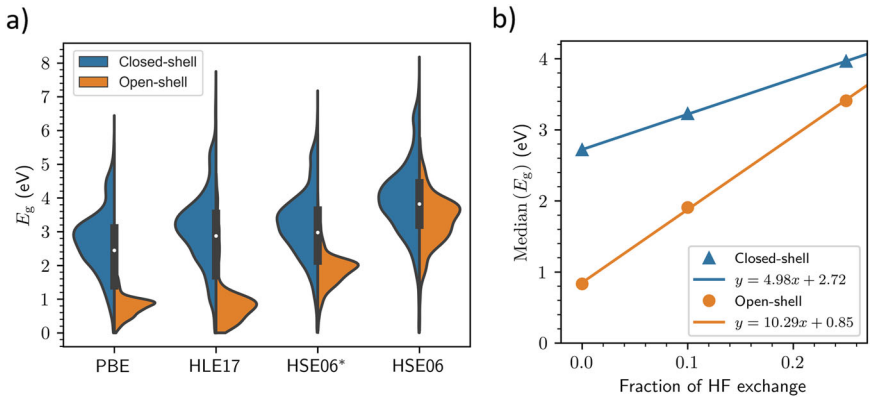

Fig. 2 Difference in band gap distributions for materials with closed- and open-shell character. a Violin plots of the predicted band gaps, E_g, for 10,720 structures in the QMOF Database calculated with PBE, HLE17, HSE06*, and HSE06. The left and right sides of each violin plot include structures with closed-shell (8628 structures) and open-shell (2092 structures) character, respectively. A box plot is included inside each violin, highlighting the extrema (whisker edges), interquartile range (box boundaries), and median (white dot) of the band gap data at the specified level of theory. **b** Median band gap as a function of the fraction of Hartree–Fock (HF) exchange at small interelectronic separation where 0% = PBE, 10% = HSE06*, and 25% = HSE06. The blue triangles and orange circles are the median band gaps for the closed-shell and open-shell structures, respectively. The solid lines display the linear best-fit equations.

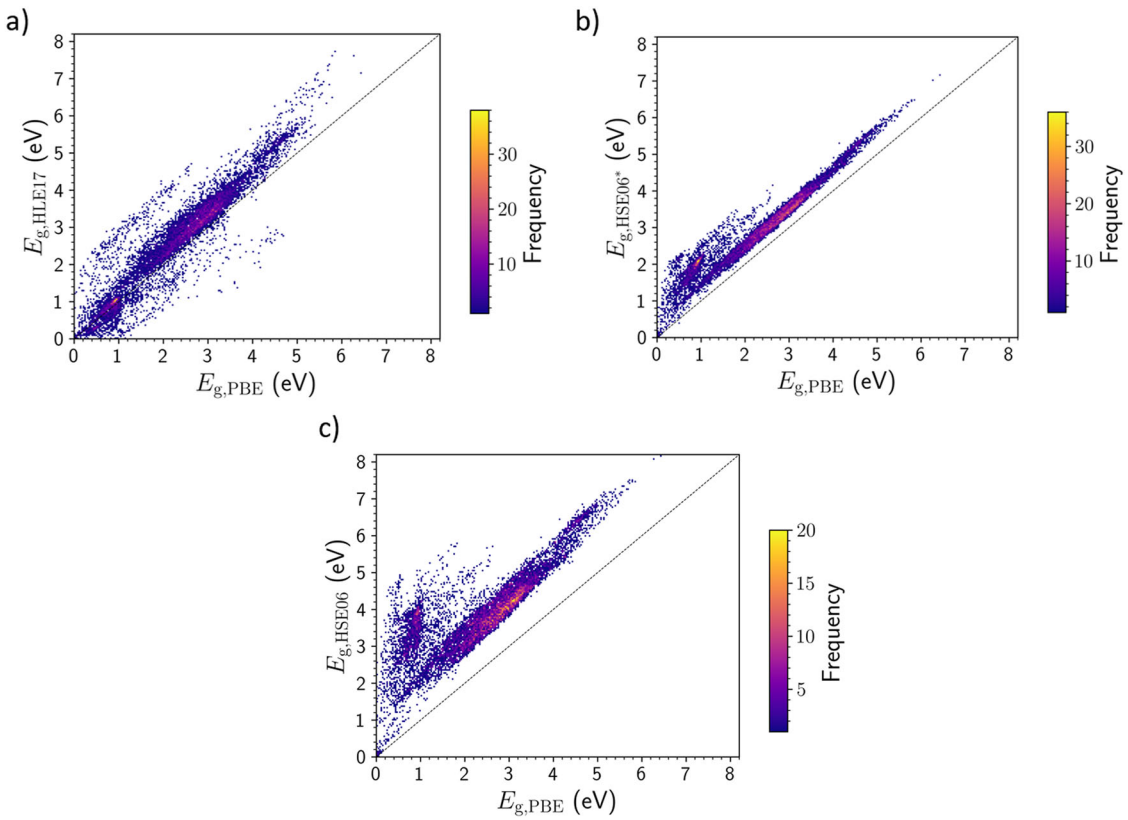

Fig. 3 Correlations between the computed band gaps across multiple levels of theory. Correlation plots of the computed band gaps, E_g, for 10,720 structures in the QMOF Database at various levels of theory. **a** HLE17 vs. PBE; **b** HSE06* vs. PBE; **c** HSE06 vs. PBE. Given the large dataset size, the data is shown as 2D histograms with the color bar reflecting the frequency of points in each bin. The $y = x$ line is shown for reference.

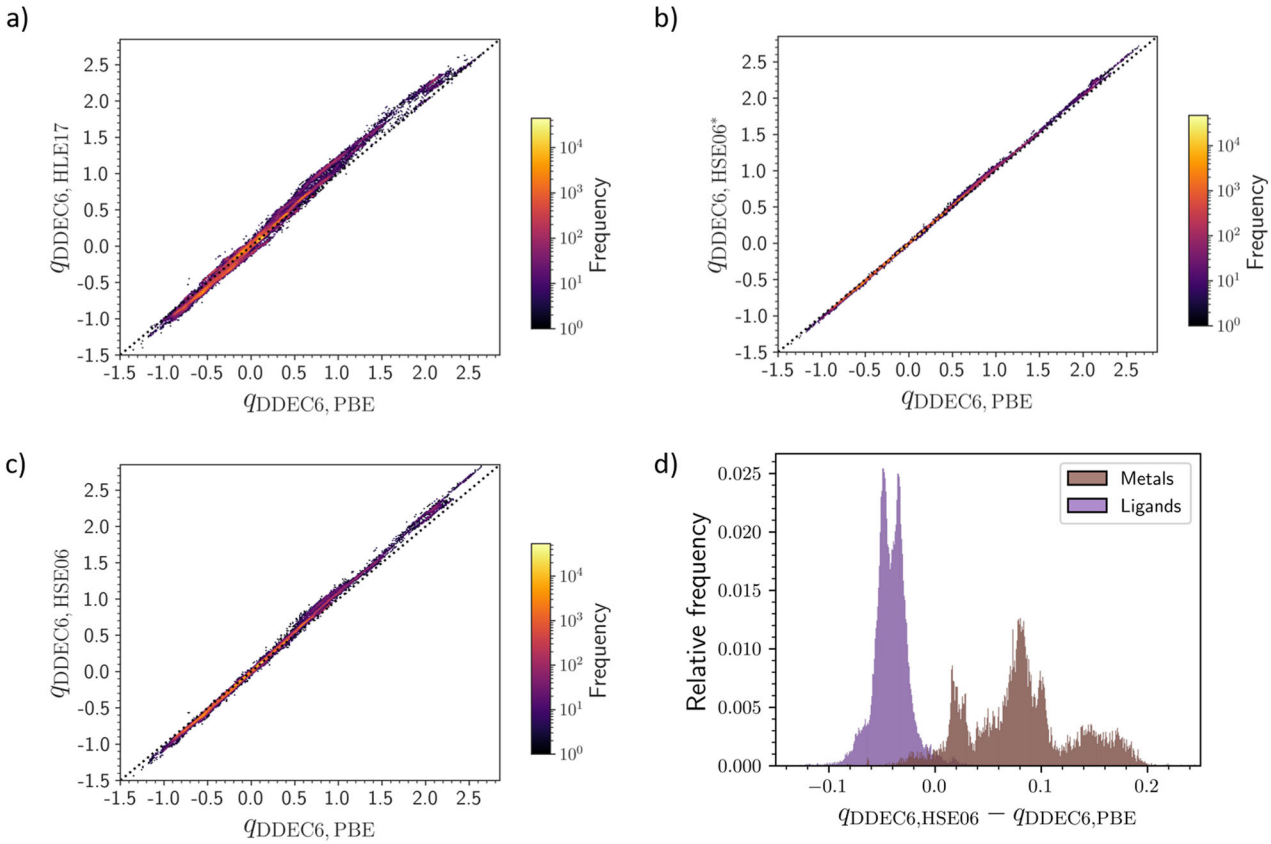

Fig. 4 Correlations between the computed partial atomic charge across multiple levels of theory. Comparison of DDEC6 partial atomic charges, q_{DDEC6}, for 922,879 atoms based on charge densities at various levels of theory: (**a**) HLE17 vs. PBE; **b** HSE06* vs. PBE; **c** HSE06 vs. PBE. Given the large dataset size, the data is shown as a 2D histogram with the logarithmic color bar reflecting the frequency of points in each bin. The $y = x$ line is shown for reference. **d** A histogram of the change in DDEC6 charges between the PBE and HSE06 levels of theory for the metal sites and ligand atoms within the first coordination sphere.

shown in Supplementary Fig. 7c, a simple linear equation of the form $1.09E_{g,PBE} + 1.04$ eV can predict HSE06 band gaps with an R^2 value of 0.92, provided the frameworks are closed-shell systems and have HSE06 band gaps above ~1.0 eV. Similar linear equations can be obtained for HLE17 and HSE06* for the closed-shell structures (Supplementary Fig. 7a and Supplementary Fig. 7b). The correlation between PBE and the hybrid functionals is weaker for MOFs with open-shell character, hence the larger degree of scatter in the low $E_{g,PBE}$ range of Fig. 3b and c.

As might be anticipated based on trends in crystal-field splitting parameters and spin-pairing energies[44], most open-shell materials in the QMOF Database contain 3*d* transition metal cations (particularly Cu, Co, Mn, Ni, Fe, V, and Cr in decreasing frequency of occurrence) (Supplementary Fig. 8). Previous theoretical work on transition metal complexes and gas-phase molecules containing transition metal cations has implicated large self-interaction errors (a consequence of each electron interacting with the total electron density, including its own[28]) as a major source of errors in systems with 3*d* transition metal cations that have open-shell character[45,46]. More generally, self-interaction error is usually considered to be responsible for many of the deficiencies of DFT across virtually all properties and material classes, often due to the associated delocalization error[47,48]. Since self-interaction error is partially decreased by the inclusion of HF exchange, this is a major reason that the hybrid functionals give different results than the local functionals for the band gap predictions in this work.

Partial charge comparison

Beyond band gaps, it is well-established that different DFAs can change how the charge density is distributed in a given material[49–53]. Furthermore, partial atomic charges (which can be computed directly from the underlying charge density) are commonly used in molecular simulations of MOFs and can be used to interpret trends when modeling redox processes and chemical reactions[54,55]. One such method to compute partial atomic charges, the sixth-generation Density Derived Electrostatic and Chemical (DDEC6) partitioning scheme[56–58], has found widespread use in molecular simulations of MOFs[54] (e.g., for gas storage and separations) and has performed well in tests of reproducing the electrostatic potential[59]. To explore the sensitivity of partial atomic charges to different DFAs, we compared over 900,000 partial charges calculated from the DDEC6 method using charge densities at the PBE, HLE17, HSE06*, and HSE06 levels of theory.

As shown in Fig. 4a, the DDEC6 partial atomic charges calculated by PBE and HLE17 are highly correlated across the entire dataset, with most points falling within 0.04 charge units from the $y = x$ line. When investigating the computed partial charges by HSE06*, we find that the HSE06* partial charges are even closer to the PBE reference than the HLE17 partial charges are (Fig. 4b), indicating that 10% HF exchange at small interelectronic distances does not substantially change the first moment of the charge density. However, when increasing the HF exchange at small interelectronic distances to 25% with HSE06, a slightly larger difference can be observed (Fig. 4c).

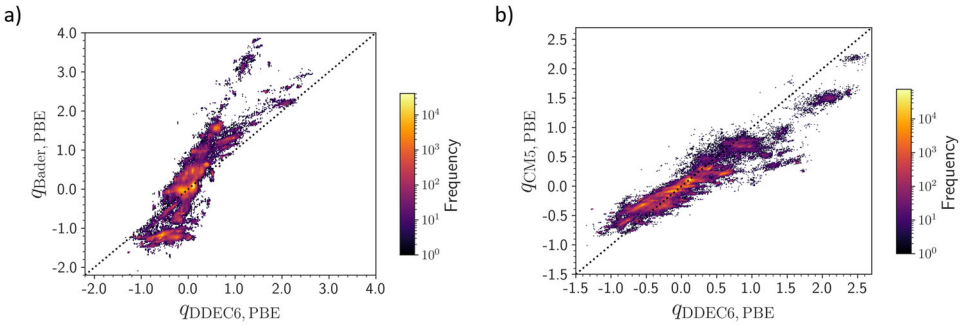

Fig. 5 Correlation between partial atomic charges with different charge partitioning schemes. a Comparison of the partial atomic charges, q, for 1,429,082 atoms computed using the Bader and DDEC6 charge partitioning schemes at the PBE level of theory. **b** Comparison of the partial atomic charges, q, for 2,321,435 atoms computed using the CM5 and DDEC6 charge partitioning schemes at the PBE level of theory. Given the large dataset size, the data is shown as 2D histograms with the logarithmic color bar reflecting the frequency of points in each bin. The $y = x$ line is shown as for reference.

By focusing solely on the metal elements and the ligand atoms within their first coordination spheres (as determined using the CrystalNN near-neighbor finding algorithm[60,61]), we find that—compared to the PBE reference—there is often a loss of electron density (i.e., increased partial atomic charge) at the metal and corresponding gain of electron density (i.e., decreased partial atomic charge) on the surrounding ligands when using the HSE06 functional (Fig. 4d). These trends are consistent with previous partial charge analyses carried out on transition metal complexes and open-framework solids[46,52,62]. Given the large partial charge dataset in the present work, we can conclude that this shifting of electron density occurs for an enormously diverse range of metal–ligand environments and can be taken as a rule-of-thumb in most cases. While there are differences in the partial atomic charges between the various levels of theory, they are generally relatively minor. The overall strong agreement suggests that the less expensive PBE-quality charges, which are available for thousands of MOFs[24,54], are likely suitable when carrying out high-throughput computational screening studies.

Since no single charge partitioning scheme is expected to be ideal for all applications, we also compared the effect of different charge partitioning schemes for a given DFA. As shown in Fig. 5, the differences between Bader[63,64], DDEC6[56,57,65], and Charge Model 5 (CM5)[66] partial atomic charges (as computed with the PBE functional) tend to be far larger than any differences observed when changing the DFA, similar to what has been observed for several inorganic solids[67]. This is especially the case when directly comparing the Bader and DDEC6 methods. As one example of many, large deviations are often observed for the S and P atoms of SO_4^{2-} and PO_4^{2-} groups, which have partial atomic charges upwards of ~2.4 charge units higher with the Bader method than the DDEC6 method. In addition, there can be qualitative differences between Bader and DDEC6 charges, such as atoms that have a partial positive charge with the Bader method but a partial negative charge with the DDEC6 method. While there are also clear differences between the DDEC6 and CM5 methods (Fig. 5b), the agreement between these two charge partitioning approaches is generally greater than that between DDEC6 and Bader. For applications involving systems quite different from those in available benchmarks[55,56,66], it might be advisable to compare multiple partial charge schemes and further investigate any substantial differences[68].

Machine learning

With the goal of reducing the number of DFT calculations needed in future high-throughput computational screening studies, we have evaluated the performance of several ML models that can predict MOF band gaps from graph representations of their three-dimensional structures (for the prediction of partial atomic charges, we refer the reader to several ML models[69–71] that have been shown to accurately predict PBE-quality DDEC6 and CM5 charges for MOFs). Using MatDeepLearn[72], we first trained individual graph neural networks for each DFA and found that they performed well at predicting DFT-computed band gaps compared to a baseline model that simply predicts the mean of the dataset for each entry (Table 1). Prior work[24,72] on the QMOF Database showed that a crystal graph convolutional neural network model[73] could predict PBE band gaps with a comparable accuracy, and it is reassuring that relatively low testing-set MAEs on the order of 0.24–0.29 eV can be obtained for the more accurate DFAs (i.e., HLE17, HSE06*, HSE06). Overall, the graph neural network trained on PBE band gap data performs better than the graph neural networks trained on the HLE17, HSE06*, or HSE06 datasets, which can likely be attributed to the greater number of data points available for training with PBE. Despite similar training set sizes for the HLE17, HSE06*, and HSE06 levels of theory, the model based on HSE06 data has the largest testing set MAE of 0.29 eV, which may be attributed in part to a wider range of possible band gap values and a greater overlap in the band gap distributions for the closed- and open-shell frameworks.

Next, we considered various approaches that could make more efficient use of the available band gap data obtained with different functionals. Starting with a multi-task learning approach that predicts band gaps for all four DFAs simultaneously using a single model architecture, perceptible but minor improvements to the model performance are obtained (Table 1). While more convenient to use than multiple individual models if multiple band gap estimates are desired, an inherent drawback of the multi-task learning method is that the training process requires structures that have band gaps computed for all DFAs of interest, which limits the amount of data that can be used.

An alternate way to efficiently leverage data at multiple levels of theory is to construct a multi-fidelity model, which treats each level of theory as a unique sample[74,75]. With a substantially expanded dataset size of up to 52,806 samples, we find that the multi-fidelity MEGNet model architecture of Chen et al[75]. achieves significantly lower MAEs than the individual and multi-task models for the 3-fi (i.e., PBE, HLE17, and HSE06*) and 4-fi (i.e., PBE, HLE17, HSE06*, and HSE06) models (Table 1). These results demonstrate that data at multiple levels of theory can be used to improve the overall model performance, which is especially important for the prediction of band gaps from hybrid functionals that are more computationally demanding to calculate. However, we note that the 2-fi model (i.e., PBE + HSE06) does not outperform the multi-task model. In future studies, it may be worthwhile to consider additional approaches (e.g., Δ-learning)[76] if only two fidelities are available, especially given the correlation between the PBE and HSE06 functionals (Fig. 4c). The testing set parity plots for each

Table 1. Graph neural network performance for predicting band gaps at multiple levels of theory.

Level of theory	Constant mean	Individual		Multi-task		Multi-fidelity (2-fi)		Multi-fidelity (3-fi)		Multi-fidelity (4-fi)	
	Test MAE (eV)	Test MAE (eV)	Dataset size	Test MAE (eV)	Dataset size	Test MAE (eV)	Dataset size	Test MAE (eV)	Dataset size	Test MAE (eV)	Dataset size
PBE	0.940	0.228	20,423	0.217	10,720	0.214	31,235	0.209	41,993	0.175	52,806
HLE17	1.076	0.242	10,758	0.239	10,720	—	—	0.145	41,993	0.119	52,806
HSE06[a]	0.858	0.257	10,813	0.236	10,720	—	—	—	—	0.094	52,806
HSE06	0.802	0.289	10,812	0.267	10,720	0.276	31,235	0.179	41,993	0.119	52,806

Individual, multi-task, and multi-fidelity model performance. The individual models represent four separate models that are each trained on band gaps at a single level of theory. The multi-task model is a single model that is trained on and predicts band gaps at all four levels of theory simultaneously. The multi-fidelity models combine data from different levels of theory without all samples needing to have band gaps at each level of theory. The 2-fi, 3-fi, and 4-fi models are trained/tested on PBE + HSE06, PBE + HLE17 + HSE06, and PBE + HLE17 + HSE06[a] + HSE06 data, respectively. A baseline model that simply predicts the mean value of the dataset is shown for reference. The dataset sizes refer to the entire available dataset, which is split 80:5:15 train:validation:test. The mean absolute errors (MAEs) are shown for the testing set.

model are presented in Supplementary Figs. S12–S16, which show that the predictive accuracy generally holds over the range of band gaps, albeit with an increase in scatter toward the low band gap region (e.g., $E_{g,DFT} < 0.5$ eV). The increased error in the low band gap region can likely be traced back to several factors, such as a smaller number of MOFs to train on in this range and a higher fraction of open-shell MOFs whose properties are likely more difficult to predict with ML models. Collectively, we anticipate that the multi-task and multi-fidelity ML models will be a valuable resource for future high-throughput screening studies by minimizing the need to carry out computationally demanding hybrid DFT calculations, particularly if low-fidelity PBE band gap data is readily available (as is the case with the QMOF Database). Given the promising nature of the multi-fidelity ML models, incorporating experimentally determined band gaps[6,8] during the training process would likely be worth pursuing in future work.

QMOF database on the materials project

With DFT-computed properties at multiple levels of theory, we aimed to make the QMOF Database align with the findable, accessible, interoperable, and reusable (FAIR) guiding principles[77,78]. Therefore, we conclude by showcasing an interactive web application hosted on the Materials Project[41,79], which can be accessed at the following webpage: https://materialsproject.org/mofs. Known as the Materials Project MOF Explorer, the web application makes it possible to investigate the computed properties in the QMOF Database through a user-friendly, search-based interface. The data driving the MOF Explorer is made available to the public through the Material Project's contribution platform MPContribs[80,81]. The MPContribs application programming interface and its accompanying Python client[82] provide a unified mechanism for contributors to submit a dataset and for the community at large to programmatically retrieve, download, and query the contributed materials data. Here, contributions containing materials data are linked to a given MOF via a dedicated, unique identifier (QMOF ID) and are organized in components of queryable dictionary data, Pymatgen[83] structure objects, and binary data files.

As shown in Fig. 6, the Materials Project-hosted MOF Explorer allows users to sort and filter materials in the QMOF Database by numerous geometric, compositional, textural, topological, magnetic, and electronic properties. Selecting a single material on the MOF Explorer leads to a detailed calculation summary page, which lists various tabulated properties for that material and an interactive visualization of the DFT-optimized crystal structure. In addition to DFT-computed properties, each material has an associated MOFid/MOFkey[84] (where computable) to support substructure searches as well as cross-referencing with other MOF databases. As the QMOF Database continues to evolve, we plan to incorporate additional computed properties and visualizations on the Materials Project to enable further data exploration.

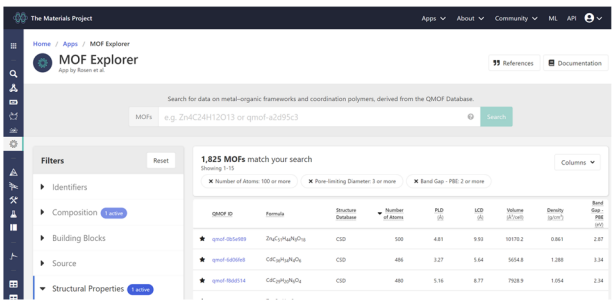

Fig. 6 Screenshot of the Materials Project MOF Explorer interface to the QMOF Database. Representative snapshot of the current search interface to the MOF Explorer application on the Materials Project with an example multi-query search applied.

DISCUSSION

With a generated dataset of electronic structure properties for a subset of ~10,700 MOFs (and coordination polymers) in the QMOF Database[24], we compare the performance of different DFAs for the prediction of band gaps and partial atomic charges. When comparing DFT-computed band gaps with the commonly used PBE functional against those that incorporate some fraction of HF exchange, we observe that PBE almost universally results in a lower band gap prediction, as might be expected from prior work. Notably, this difference is largely systematic for MOFs with closed-shell electronic configurations and can be empirically corrected through a simple linear relationship for structures that are semi-conductors or insulators. For MOFs with open-shell electronic configurations (in particular, those containing 3*d* transition metals), an even larger—and less predictable—disparity between band gap predictions is observed as a function of the fraction of HF exchange. As compared to the PBE results, the meta-GGA HLE17 is found to increase the computed band gaps for the closed-shell MOFs such that they are similar to values predicted using the HSE06 screened hybrid functional with 10% HF exchange at small interelectronic distances (denoted here as HSE06*).

However, compared to the hybrid functionals, HLE17 does not as significantly increase the band gaps of the open-shell MOFs.

When investigating partial atomic charges, which are reflective of the underlying charge density for a given density functional approximation, we find that there are slight systematic differences amongst the predictions of the different functionals. For both the HLE17 meta-GGA and the screened hybrid functionals, electron density localized on the metals is lower than with PBE, and the opposite is true for the coordinating ligand atoms. Nonetheless, these changes in the partial atomic charges are relatively minor compared to the differences that arise from using different charge partitioning schemes.

Finally, we used the electronic structure data generated in this work to train multiple ML models that can predict MOF band gaps at various levels of theory from graphs of the underlying crystal structures. We find that individual graph neural network models can predict PBE, HLE17, HSE06* or HSE06 band gaps from the QMOF Database with a testing-set MAE of 0.23–0.29 eV. A multi-task graph neural network model capable of simultaneously predicting MOF band gaps for all four functionals performs slightly better than the individual models, but with three or more functionals to train on, a multi-fidelity model achieves the best performance of the models tested in this work.

High-throughput computational screening approaches have historically been devoted to the discovery of MOFs tailored for gas storage and separations. With the dataset and ML models presented in this work—coupled with an increased understanding of the behavior of common DFAs for predicting electronic properties—we anticipate that a computational materials design perspective can be brought to countless application areas for MOFs. Now hosted on the widely used Materials Project platform (https://materialsproject.org/mofs), theorists and experimentalists alike can leverage the data from tens of thousands of quantum-mechanical calculations to accelerate the discovery of promising MOFs for electronic and optoelectronic applications.

METHODS

Density functional theory calculations

Plane-wave, periodic DFT calculations were carried out using the Vienna ab initio Simulation Package (VASP)[85,86] version 5.4.4 and the Atomic Simulation Environment (ASE)[87] version 3.20.0b1. All structures were adopted from the QMOF Database[24]. We consider properties calculated with four exchange-correlation functionals: PBE-D3(BJ)[27,88,89], HLE17[38], HSE06[39,40], and HSE06* (i.e., HSE06 with reduced HF exchange). The PBE-D3(BJ) calculations were obtained from the QMOF Database, as previously reported[24]. The HLE17, HSE06, and HSE06* calculations are carried out in this work using structures from the QMOF Database[24] that were previously optimized with the PBE-D3(BJ) exchange-correlation functional. In commonly accepted notation, these levels of theory would generally be referred to as PBE-D3(BJ), HLE17//PBE-D3(BJ), HSE06//PBE-D3(BJ), and HSE06*//PBE-D3(BJ), indicating that the functional to the left of the double-slash is a single-point (i.e., static) calculation carried out on the geometry obtained using the functional to the right of the double-slash. For brevity, we will simply refer to these levels of theory as PBE, HLE17, HSE06, and HSE06*, respectively. Of the 20,000+ structures in the QMOF Database with properties computed using PBE, ~10,700 have computed properties at the HLE17, HSE06, and HSE06* levels of theory based on the calculations in this work.

The HSE06 functional is a screened-exchange functional built upon PBE and replaces a portion of PBE's local exchange with 25% HF exchange at small interelectronic distances, decreasing continuously to zero at large interelectronic distances[39,40]. HSE06 was selected in this work because it is currently the most widely used functional for predicting the band gaps of solid-state materials when high accuracy is required, including for MOFs[43,90]. Other functionals may have comparable or slightly better performance for certain systems[37,91–93] but are less widely used and tested. In addition to HSE06, we considered the hybrid functional defined here as HSE06*, which has 10% HF exchange at small interelectronic distances and decreases to zero at large interelectronic distances. HSE06* was considered because the standard HSE06 functional can overcorrect the band gap underprediction problem of PBE for some materials[94], as is the case with MOF-5[95,96]. Considering a functional with an intermediate fraction of HF exchange between that of PBE and HSE06 also makes it easier to discern the impact of HF exchange. The HSE06 and HSE06* calculations are considerably more expensive than the PBE calculations because of the nonzero fraction of HF exchange. With this in mind, we included the HLE17 meta-GGA functional as well because prior benchmarking studies[38,43] suggest that it can greatly improve the prediction of semiconductor band gaps without the need for computationally expensive HF exchange. While one could also consider the GGA+U approach[97], relatively little is currently known about selecting empirically ideal U values for MOFs[90,98,99] despite its widespread use in correcting the predicted energetic and electronic properties of inorganic solids in high-throughput DFT databases [100–102].

For materials that are closed-shell (i.e., without magnetic character), the band gap is defined as the energy difference between the conduction band minimum (CBM) and valence band maximum (VBM). For materials with open-shell character, there can be more than one way to characterize the band gap[103]. Except where otherwise stated, we define the band gap for spin-polarized systems as $\min(\mathrm{CBM}_{\uparrow}, \mathrm{CBM}_{\downarrow}) - \max(\mathrm{VBM}_{\uparrow}, \mathrm{VBM}_{\downarrow})$, where ↑ and ↓ refer to the spin-up and spin-down spin-orbital manifolds, respectively. Nonetheless, we note that this definition can occasionally result in a band gap that is associated with a formally spin-forbidden electronic excitation, as depicted in Supplementary Fig. 4. Using the band gap instead defined as $\min(\mathrm{CBM}_{\uparrow} - \mathrm{VBM}_{\uparrow}, \mathrm{CBM}_{\downarrow} - \mathrm{VBM}_{\downarrow})$ does not involve a spin-flip. Regardless of which band gap definition is employed, the trends and conclusions reported throughout this work remain unchanged (Supplementary Fig. 5). We also note that the computed band gaps refer to electronic band gaps and are not directly comparable to experimentally measured optical gaps (e.g., via UV-Vis spectroscopy)[104,105], particularly when the exciton binding energies are non-negligible, as has been observed for some MOFs[106].

The following software packages were used to analyze the DFT data in this work this work: Chargemol v. 09-26-2017 (DDEC6 and CM5 calculations)[107], ASE v. 3.20.0b1 (orchestrate the VASP calculations)[87], Pymatgen v. 2020.12.3 (electronic structure analysis)[83], Bader v. 1.04 (Bader analysis)[64], NumPy/Pandas/SciPy/matplotlib/seaborn (data analysis and visualization)[108–112], and PtitPrince v.0.2.5 (for raincloud plots[113]). Additional methodological details regarding the DFT calculations, dataset curation, updates to the QMOF Database, and data analysis can be found in the Supplementary Information.

Machine learning

Graph neural network architectures, which take graphs representing the crystal structures as inputs, were used for the ML models. The graph representations contain atoms as nodes and interatomic distances as edges. Here, the atoms are represented with a one-hot encoding of the element with a vector length of 100 within the node attributes. The edge attributes contain interatomic distances within a cutoff of 8 Å and up to 12 neighbors per node, where the distances were then expanded by a Gaussian basis[114] to a length of 50. In this work, an additional state attribute is included, representing the level of theory used (i.e., fidelity) as an integer. The graph neural network itself adopts the MatErials Graph Network (MEGNet) architecture[115] where the node, edge, and state attributes are propagated sequentially in the stated order during the graph convolutional steps. The overall model contains one pre-processing layer, four graph convolutional layers, one pooling layer using the Set2Set function, and finally two post-processing layers. The pre-processing, post-processing, and graph convolutional update functions are all fully-connected layers with Rectified Linear Unit activation functions and with dimensions of 128, 128 and (128, 128), respectively. The models were trained with the AdamW optimizer[116,117] using an initial learning rate of 0.0005 and a batch size of 128 for a total of 250 epochs. The model state with the lowest validation MAE is saved and used for testing. The training:validation:testing ratio used is 80:5:15, and the samples were randomly split across the training, validation, and testing sets. For all cases in this work, the same hyperparameters were used in the models. For the individual models, the models were trained separately. In multi-task learning, the output dimension was expanded to four, and the predictions were performed simultaneously with a single model for all fidelities (i.e., levels of theory). For multi-fidelity learning, we adopt the approach used by Chen et al[75]. where each fidelity is considered a unique data sample and structures with different fidelities can appear in both training and testing data splits. The model training and testing was set up and performed using the MatDeepLearn framework[72], which is implemented using the PyTorch[118]

and PyTorch geometric[119] libraries. The training and evaluation were conducted on four NVIDIA Tesla V100 ('Volta') graphics processing units.

DATA AVAILABILITY

With the release of the Materials Project-hosted MOF Explorer interface to the QMOF Database, all data in this work can be accessed at the following webpage: https://materialsproject.org/mofs. Each version of the QMOF Database made available on the Materials Project is permanently archived on Figshare at the following DOI: 10.6084/m9.figshare.13147324. The VASP input and output files are made available via the Novel Materials Discovery (NOMAD) platform[120,121] with the following dataset names and DOIs: QMOD Database - PBE (10.17172/NOMAD/2021.10.10-1), QMOF Database—HLE17 (10.17172/NOMAD/2021.11.17-3), QMOF Database—HSE06* (10.17172/NOMAD/2021.11.17-2), and QMOF Database—HSE06 (10.17172/NOMAD/2021.11.17-1).

CODE AVAILABILITY

The codes used to carry out this work are described and referenced in the Methods section and are available free-of-charge with the exception of VASP.

Received: 11 December 2021; Accepted: 21 April 2022;
Published online: 17 May 2022

REFERENCES

1. Yaghi, O. M. et al. Reticular synthesis and the design of new materials. *Nature* **423**, 705–714 (2003).
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.* **4**, eaat9180 (2018).
3. Yaghi, O. M., Kalmutzki, M. J. & Diercks, C. S. *Introduction to Reticular Chemistry: Metal-Organic Frameworks and Covalent Organic Frameworks*. 1st edn (John Wiley & Sons, 2019).
4. Chen, Z. et al. The state of the field: from inception to commercialization of metal–organic frameworks. *Faraday Discuss.* **225**, 9–69 (2021).
5. Stavila, V., Talin, A. A. & Allendorf, M. D. MOF-based electronic and opto-electronic devices. *Chem. Soc. Rev.* **43**, 5994–6010 (2014).
6. Xie, L. S., Skorupskii, G. & Dincă, M. Electrically Conductive Metal–Organic Frameworks. *Chem. Rev.* **120**, 8536–8580 (2020).
7. Johnson, E. M., Ilic, S. & Morris, A. J. Design Strategies for Enhanced Conductivity in Metal–Organic Frameworks. *ACS Cent. Sci.* **7**, 445–453 (2021).
8. Zanca, F. et al. Computational Techniques for Characterisation of Electrically Conductive MOFs: Quantum Calculations and Machine Learning Approaches. *J. Mater. Chem. C.* **9**, 13584–13599 (2021).
9. Zhang, H., Nai, J., Yu, L. & Lou, X. W. D. Metal-organic-framework-based materials as platforms for renewable energy and environmental applications. *Joule* **1**, 77–107 (2017).
10. Wu, X.-P., Choudhuri, I. & Truhlar, D. G. Computational studies of photocatalysis with metal–organic frameworks. *Energy Environ. Mater.* **2**, 251–263 (2019).
11. Tajik, S. et al. Recent electrochemical applications of metal–Organic framework-based materials. *Cryst. Growth Des.* **20**, 7034–7064 (2020).
12. Moghadam, P. Z. et al. Development of a Cambridge Structural Database Subset: A Collection of Metal–Organic Frameworks for Past, Present, and Future. *Chem. Mater.* **29**, 2618–2625 (2017).
13. Wilmer, C. E. et al. Large-scale screening of hypothetical Metal−Organic frameworks. *Nat. Chem.* **4**, 83–89 (2012).
14. Colón, Y. J. & Snurr, R. Q. High-throughput computational screening of metal −organic frameworks. *Chem. Soc. Rev.* **43**, 5735–5749 (2014).
15. Borboudakis, G. et al. Chemically intuited, large-scale screening of MOFs by machine learning techniques. *npj Comput. Mater.* **3**, 40 (2017).
16. Jablonka, K. M., Ongari, D., Moosavi, S. M. & Smit, B. Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chem. Rev.* **120**, 8066–8129 (2020).
17. Shi, Z. et al. Machine-learning-assisted high-throughput computational screening of high performance metal–Organic frameworks. *Mol. Syst. Des. Eng.* **5**, 725–742 (2020).
18. Chong, S., Lee, S., Kim, B. & Kim, J. Applications of machine learning in metal-organic frameworks. *Coord. Chem. Rev.* **423**, 213487 (2020).
19. Altintas, C., Altundal, O. F., Keskin, S. & Yildirim, R. Machine Learning Meets with Metal Organic Frameworks for Gas Storage and Separation. *J. Chem. Inf. Model.* **61**, 2131–2146 (2021).
20. Mukherjee, K. & Colón, Y. J. Machine learning and descriptor selection for the computational discovery of metal-organic frameworks. *Mol. Simul.* **47**, 857–877 (2021).
21. Moosavi, S. M., Jablonka, K. M. & Smit, B. The Role of Machine Learning in the Understanding and Design of Materials. *J. Am. Chem. Soc.* **142**, 20273–20287 (2020).
22. Rosen, A. S., Notestein, J. M. & Snurr, R. Q. Realizing the Data-Driven, Computational Discovery of Metal-Organic Framework Catalysts. *Curr. Opin. Chem. Eng.* **35**, 100760 (2022).
23. 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.* **40**, 1305–1318 (2019).
24. Rosen, A. S. et al. Machine Learning the Quantum-Chemical Properties of Metal–Organic Frameworks for Accelerated Materials Discovery. *Matter* **4**, 1578–1597 (2021).
25. Hill, J., Mannodi-Kanakkithodi, A., Ramprasad, R. & Meredig, B. Materials Data Infrastructure and Materials Informatics, in *Computational Materials System Design* 2017 193–225 (Springer International Publishing, 2017).
26. Schleder, G. R., Padilha, A. C. M., Acosta, C. M., Costa, M. & Fazzio, A. From DFT to machine learning: recent approaches to materials science–A review. *J. Phys. Mater.* **2**, 32001 (2019).
27. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
28. Mori-Sánchez, P., Cohen, A. J. & Yang, W. Many-electron self-interaction error in approximate density functionals. *J. Chem. Phys.* **125**, 201102 (2006).
29. Mori-Sánchez, P., Cohen, A. J. & Yang, W. Localization and delocalization errors in density functional theory and implications for band-gap prediction. *Phys. Rev. Lett.* **100**, 146401 (2008).
30. Borlido, P. et al. Large-scale benchmark of exchange–correlation functionals for the determination of electronic band gaps of solids. *J. Chem. Theory Comput.* **15**, 5069–5079 (2019).
31. Filippi, C., Singh, D. J. & Umrigar, C. J. All-electron local-density and generalized-gradient calculations of the structural properties of semiconductors. *Phys. Rev. B* **50**, 14947 (1994).
32. Zhao, Y. & Truhlar, D. G. Calculation of semiconductor band gaps with the M06-L density functional. *J. Chem. Phys.* **130**, 74103 (2009).
33. Zhuo, Y., Mansouri Tehrani, A. & Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *J. Phys. Chem. Lett.* **9**, 1668–1673 (2018).
34. Kauwe, S. K., Welker, T. & Sparks, T. D. Extracting Knowledge from DFT: experimental Band Gap Predictions Through Ensemble Learning. *Integr. Mater. Manuf. Innov.* **9**, 213–220 (2020).
35. Kingsbury, R. et al. Performance comparison of r^{2}SCAN and SCAN metaGGA density functionals for solid materials via an automated, high-throughput computational workflow. *Phys. Rev. Mater.* **6**, 013801 (2022).
36. Kim, S. et al. A band-gap database for semiconducting inorganic materials calculated with hybrid functional. *Sci. Data* **7**, 387 (2020).
37. Borlido, P. et al. Exchange-correlation functionals for band gaps of solids: benchmark, reparametrization and machine learning. *npj Comput. Mater.* **6**, 96 (2020).
38. Verma, P. & Truhlar, D. G. HLE17: an improved local exchange–correlation functional for computing semiconductor band gaps and molecular excitation energies. *J. Phys. Chem. C.* **121**, 7144–7154 (2017).
39. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **118**, 8207–8215 (2003).
40. 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.* **125**, 224106 (2006).
41. Jain, A. et al. The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 11002 (2013).
42. Janesko, B. G., Henderson, T. M. & Scuseria, G. E. Screened hybrid density functionals for solid-state chemistry and physics. *Phys. Chem. Chem. Phys.* **11**, 443–454 (2009).
43. Choudhuri, I. & Truhlar, D. G. HLE17: an Efficient Way To Predict Band Gaps of Complex Materials. *J. Phys. Chem. C.* **123**, 17416–17424 (2019).
44. Atkins P., Overton T., Rourke J., Weller M., Armstrong F., M. H. *Shriver & Atkins' Inorganic Chemistry*. (Oxford University Press, 2009).
45. Liu, F. & Kulik, H. J. Impact of Approximate DFT Density Delocalization Error on Potential Energy Surfaces in Transition Metal Chemistry. *J. Chem. Theory Comput.* **16**, 264–277 (2019).
46. Ioannidis, E. I. & Kulik, H. J. Towards quantifying the role of exact exchange in predictions of transition metal complex properties. *J. Chem. Phys.* **143**, 34104 (2015).
47. Wasserman, A. et al. The importance of being inconsistent. *Annu. Rev. Phys. Chem.* **68**, 555–581 (2017).
48. Janesko, B. G. Replacing hybrid density functional theory: motivation and recent advances. *Chem. Soc. Rev.* **50**, 8470–8495 (2021).

49. Wang, J., Johnson, B. G., Boyd, R. J. & Eriksson, L. A. Electron densities of several small molecules as calculated from density functional theory. *J. Phys. Chem.* **100**, 6317–6324 (1996).
50. Schwerdtfeger, P., Pernpointner, M. & Laerdahl, J. K. The accuracy of current density functionals for the calculation of electric field gradients: A comparison with ab initio methods for HCl and CuCl. *J. Chem. Phys.* **111**, 3357–3364 (1999).
51. Schultz, N. E., Gherman, B. F., Cramer, C. J. & Truhlar, D. G. Pd_nCO ($n = 1,2$): Accurate ab initio bond energies, geometries, and dipole moments and the applicability of density functional theory for fuel cell modeling. *J. Phys. Chem. B* **110**, 24030–24046 (2006).
52. Zhao, Q. & Kulik, H. J. Where Does the Density Localize in the Solid State? Divergent Behavior for Hybrids and DFT+*U*. *J. Chem. Theory Comput.* **14**, 670–683 (2018).
53. Grotjahn, R., Lauter, G. J., Haasler, M. & Kaupp, M. Evaluation of Local Hybrid Functionals for Electric Properties: Dipole Moments and Static and Dynamic Polarizabilities. *J. Phys. Chem. A* **124**, 8346–8358 (2020).
54. Nazarian, D., Camp, J. S. & Sholl, D. S. A comprehensive set of high-quality point charges for simulations of metal−Organic frameworks. *Chem. Mater.* **28**, 785–793 (2016).
55. Wang, B., Li, S. L. & Truhlar, D. G. Modeling the partial atomic charges in inorganometallic molecules and solids and charge redistribution in lithium-ion cathodes. *J. Chem. Theory Comput.* **10**, 5640–5650 (2014).
56. Manz, T. A. & Limas, N. G. Introducing DDEC6 atomic population analysis: part 1. Charge partitioning theory and methodology. *RSC Adv.* **6**, 47771–47801 (2016).
57. 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.* **6**, 45727–45747 (2016).
58. Manz, T. A. Introducing DDEC6 atomic population analysis: part 3. Comprehensive method to compute bond orders. *RSC Adv.* **7**, 45552–45581 (2017).
59. Manz, T. A. & Sholl, D. S. Chemically meaningful atomic charges that reproduce the electrostatic potential in periodic and nonperiodic materials. *J. Chem. Theory Comput.* **6**, 2455–2468 (2010).
60. Zimmermann, N. E. R. & Jain, A. Local Structure Order Parameters and Site Fingerprints for Quantification of Coordination Environment and Crystal Structure Similarity. *RSC Adv.* **10**, 6063–6081 (2019).
61. Pan, H. et al. Benchmarking Coordination Number Prediction Algorithms on Inorganic Crystal Structures. *Inorg. Chem.* **60**, 1590–1603 (2020).
62. Gani, T. Z. H. & Kulik, H. J. Where does the density localize? Convergent behavior for global hybrids, range separation, and DFT+*U*. *J. Chem. Theory Comput.* **12**, 5931–5945 (2016).
63. Bader, R. F. W. & Matta, C. F. Atomic charges are measurable quantum expectation values: a rebuttal of criticisms of QTAIM charges. *J. Phys. Chem. A* **108**, 8385–8394 (2004).
64. Tang, W., Sanville, E. & Henkelman, G. A grid-based Bader analysis algorithm without lattice bias. *J. Phys. Condens. Matter* **21**, 84204 (2009).
65. 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.* **8**, 2678–2707 (2018).
66. Marenich, A. V., Jerome, S. V., Cramer, C. J. & Truhlar, D. G. Charge model 5: an extension of hirshfeld population analysis for the accurate description of molecular interactions in gaseous and condensed phases. *J. Chem. Theory Comput.* **8**, 527–541 (2012).
67. Choudhuri, I. & Truhlar, D. G. Calculating and Characterizing the Charge Distributions in Solids. *J. Chem. Theory Comput.* **16**, 5884–5892 (2020).
68. Manz, T. A. Seven confluence principles: a case study of standardized statistical analysis for 26 methods that assign net atomic charges in molecules. *RSC Adv.* **10**, 44121–44148 (2020).
69. Raza, A., Sturluson, A., Simon, C. & Fern, X. Message Passing Neural Networks for Partial Charge Assignment to Metal-Organic Frameworks. *J. Phys. Chem. C.* **124**, 19070–19082 (2020).
70. Kancharlapalli, S., Gopalan, A., Haranczyk, M. & Snurr, R. Q. Fast and Accurate Machine Learning Strategy for Calculating Partial Atomic Charges in Metal–Organic Frameworks. *J. Chem. Theory Comput.* **17**, 3052–3064 (2021).
71. Korolev, V. V. et al. Transferable and extensible machine learning derived atomic charges for modeling hybrid nanoporous materials. *Chem. Mater.* **32**, 7822–7831 (2020).
72. Fung, V., Zhang, J., Juarez, E. & Sumpter, B. Benchmarking graph neural networks for materials chemistry. *npj Comput. Mater.* **7**, 84 (2021).
73. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
74. Pilania, G., Gubernatis, J. E. & Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **129**, 156–163 (2017).
75. Chen, C., Zuo, Y., Ye, W., Li, X. & Ong, S. P. Learning properties of ordered and disordered materials from multi-fidelity data. *Nat. Comput. Sci.* **1**, 46–53 (2021).
76. 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.* **11**, 2087–2096 (2015).
77. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
78. Coudert, F.-X. Materials databases: the need for open, interoperable databases with standardized data and rich Metadata. *Adv. Theory Simul.* **2**, 1900131 (2019).
79. Jain, A. et al. The materials project: accelerating materials design through theory-driven data and tools. In *Handbook of Materials Modeling. Methods: Theory and Modeling* (eds. Andreoni, W. & Yip, S.) 1751–1784 (Springer Chem, 2020).
80. Huck, P. et al. User applications driven by the community contribution framework MPContribs in the Materials Project. *Concurr. Comput. Pract. Exp.* **28**, 1982–1993 (2016).
81. MPContribs. https://mpcontribs.org.
82. MPContribs-Client. https://pypi.org/project/mpcontribs-client.
83. Ong, S. P. et al. Python Materials Genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
84. Bucior, B. J. et al. Identification Schemes for Metal–Organic Frameworks to Enable Rapid Search and Cheminformatics Analysis. *Cryst. Growth Des.* **19**, 6682–6697 (2019).
85. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
86. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
87. Larsen, A. et al. The Atomic Simulation Environment—A Python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002 (2017).
88. 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.* **132**, 154104 (2010).
89. Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32**, 1456–1465 (2011).
90. Mancuso, J. L., Mroz, A. M., Le, K. N. & Hendon, C. H. Electronic Structure Modeling of Metal–Organic Frameworks. *Chem. Rev.* **120**, 8641–8715 (2020).
91. Garza, A. J. & Scuseria, G. E. Predicting band gaps with hybrid density functionals. *J. Phys. Chem. Lett.* **7**, 4165–4170 (2016).
92. Moussa, J. E., Schultz, P. A. & Chelikowsky, J. R. Analysis of the Heyd-Scuseria-Ernzerhof density functional parameter space. *J. Chem. Phys.* **136**, 204117 (2012).
93. Wang, Y. et al. M06-SX screened-exchange density functional for chemistry and solid-state physics. *Proc. Natl Acad. Sci.* **117**, 2294–2301 (2020).
94. Meng, Y. et al. When density functional approximations meet iron oxides. *J. Chem. Theory Comput.* **12**, 5132–5144 (2016).
95. Yang, L.-M., Fang, G.-Y., Ma, J., Ganz, E. & Han, S. S. Band gap engineering of paradigm MOF-5. *Cryst. Growth Des.* **14**, 2532–2541 (2014).
96. Butler, K. T., Hendon, C. H. & Walsh, A. Electronic structure modulation of metal–organic frameworks for hybrid devices. *ACS Appl. Mater. Interfaces* **6**, 22044–22050 (2014).
97. Kulik, H. J. Perspective: treating electron Over-Delocalization with the DFT+U method. *J. Chem. Phys.* **142**, 240901 (2015).
98. Mann, G. W., Lee, K., Cococcioni, M., Smit, B. & Neaton, J. B. First-principles Hubbard U approach for small molecule binding in metal-organic frameworks. *J. Chem. Phys.* **144**, 174104 (2016).
99. Rosen, A. S., Notestein, J. M. & Snurr, R. Q. Comparing GGA, GGA+*U*, and Meta-GGA Functionals for Redox-Dependent Binding at Open Metal Sites in Metal −Organic Frameworks. *J. Chem. Phys.* **152**, 224101 (2020).
100. Wang, L., Maxisch, T. & Ceder, G. Oxidation energies of transition metal oxides within the GGA+U framework. *Phys. Rev. B* **73**, 195107 (2006).
101. Jain, A. et al. Formation enthalpies by mixing GGA and GGA+*U* calculations. *Phys. Rev. B* **84**, 45115 (2011).
102. Kirklin, S. et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
103. Li, X. & Yang, J. First-principles design of spintronics materials. *Natl Sci. Rev.* **3**, 365–381 (2016).
104. Shu, Y. & Truhlar, D. G. Relationships between Orbital Energies, Optical and Fundamental Gaps, and Exciton Shifts in Approximate Density Functional Theory and Quasiparticle Theory. *J. Chem. Theory Comput.* **16**, 4337–4350 (2020).
105. Baerends, E. J., Gritsenko, O. V. & Van Meer, R. The Kohn–Sham gap, the fundamental gap and the optical gap: the physical meaning of occupied and virtual Kohn–Sham orbital energies. *Phys. Chem. Chem. Phys.* **15**, 16408–16425 (2013).
106. Kshirsagar, A. R., Blase, X., Attaccalite, C. & Poloni, R. Strongly Bound Excitons in Metal–Organic Framework MOF-5: A Many-Body Perturbation Theory Study. *J. Phys. Chem. Lett.* **12**, 4045–4051 (2021).
107. Manz, T. A. & Gabaldon Limas, N. Chargemol program for performing DDEC analysis. http://ddec.sourceforge.net/.

108. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
109. McKinney, W. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* vol. 445, 51–56 (2010).
110. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
111. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
112. Seaborn. https://doi.org/10.5281/zenodo.592845.
113. Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R. & Kievit, R. A. Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Res.* **4**, 63 (2019).
114. 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.* **148**, 241722 (2018).
115. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).
116. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
117. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. Preprint at https://arxiv.org/abs/1711.05101 (2017).
118. Paszke, A. et al. PyTorch: An imperative style, high-performance deep learning library. in *Advances in Neural Information Processing Systems* 8024–8035 (2019).
119. Fey, M. & Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. Preprint at https://arxiv.org/abs/1903.02428 (2019).
120. Draxl, C. & Scheffler, M. NOMAD: the FAIR concept for big data-driven materials science. *MRS Bull.* **43**, 676–682 (2018).
121. Draxl, C. & Scheffler, M. The NOMAD laboratory: from data sharing to artificial intelligence. *J. Phys. Mater.* **2**, 36001 (2019).

ACKNOWLEDGEMENTS

A.S.R. acknowledges support via a Miller Research Fellowship from the Miller Institute for Basic Research in Science, University of California, Berkeley. P.H., C.T.O., M.K.H., and K.A.P. acknowledge support by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under Contract No. DE-AC02-05-CH11231 (Materials Project program KC23MP). D.G.T. and R. Q.S. acknowledge support from 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. A. S.R. acknowledges computing support from the Department of Defense High Performance Computing (HPC) Modernization Program via the Mustang HPC environment at the Air Force Research Laboratory and the Onyx HPC environment at the U.S. Army Engineer Research and Development Center.

AUTHOR CONTRIBUTIONS

A.S.R. conceived and designed the project, led the collaboration, carried out the DFT calculations, analyzed the results, and wrote the manuscript. V.F. constructed and carried out the machine learning analyses. A.S.R., P.H., M.K.H., and C.T.O. created the interactive interface to the QMOF Database on the Materials Project. All authors (A.S.R., V.F., P.H., C.T.O., M.K.H., D.G.T., K.A.P., J.M.N., and R.Q.S.) reviewed and edited the manuscript.

COMPETING INTERESTS

R.Q.S. has a financial interest in the start-up company NuMat Technologies, which is seeking to commercialize metal−organic frameworks. The remaining authors declare no competing financial interests, and all authors declare no competing non-financial interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-022-00796-6.

Correspondence and requests for materials should be addressed to Andrew S. Rosen.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.