

Supplementary Information:

High-Throughput Predictions of Metal–Organic Framework Electronic Properties: Theoretical Challenges, Graph Neural Networks, and Data Exploration

Andrew S. Rosen,^{1,2,3,8} Victor Fung,⁴ Patrick Huck,⁵ Cody T. O'Donnell,³ Matthew K. Horton,³
Donald G. Truhlar,⁶ Kristin A. Persson,^{1,7} Justin M. Notestein,⁸ and Randall Q. Snurr⁸

¹Department of Materials Science and Engineering, University of California, Berkeley, California 94720, USA

³Miller Institute for Basic Research in Science, University of California, Berkeley, Berkeley, California 94720, USA

³Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

⁴Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA

⁵Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

⁶Department of Chemistry, Chemical Theory Center, and Minnesota Supercomputing Institute, University of Minnesota, Minnesota 55455, USA

⁷Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States

⁸Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, USA

Contents

Supplementary Methods	2
QMOF Database Update: New Properties at the HLE17, HSE06*, and HSE06 Levels of Theory	2
Additional Methodological Details	4
Accounting for Spin-Forbidden Transitions	4
New Structures in the QMOF Database	5
Supplementary Results	7
Band Gap Analyses	7
Open-Shell Structures	7
Representative Examples with Small and Large Band Gap Differences	8
Partial Charge Analyses	8
Machine Learning	9
Supplementary References	12

Supplementary Methods

QMOF Database Update: New Properties at the HLE17, HSE06, and HSE06 Levels of Theory*

Prior to this study, the QMOF Database consisted only of PBE-D3(BJ) optimized structures with properties provided at the same level of theory.¹ In the present work, we have updated the QMOF Database with properties calculated using the HLE17, HSE06*, and HSE06 density functional approximations using the aforementioned PBE-D3(BJ) geometries. This new data includes but is not limited to: absolute energies, band gaps, total and projected density of states, DDEC6 population analysis (i.e. partial atomic charges, atomic spin densities, bond orders),²⁻⁵ Charge Model 5 (CM5) charges,⁶ magnetic moments, and charge densities. For these calculations, all PBE-D3(BJ) optimized structures were obtained from v4 of the QMOF Database.^{1,7} The DFT settings for the HLE17, HSE06*, and HSE06 calculations are identical to those used in the original construction of the QMOF Database,¹ with the following exceptions:

- A k -point density (kppa) of 500/(number of atoms per cell) was used, as arranged using the automatic k -pointy density generator in Pymatgen.⁸ This results in fewer k -points than the 1000/(number of atoms per cell) density previously used in construction of the QMOF Database in order to decrease the computational cost of the meta-GGA and hybrid-level calculations.
- The preconditioned conjugate gradient algorithm⁹⁻¹¹ (algo=all instead of algo=Fast) was used to converge the self-consistent field (SCF), as this is the recommended algorithm for hybrid- and meta-GGAs.
- An SCF energy convergence criterion of 10^{-5} eV (instead of 10^{-6} eV) was chosen as a reasonable balance between runtime and precision for meta-GGA and hybrid-level calculations.
- A PBE wavefunction with kppa of 500 (“PBE-lowkpt”) was used as the initial guess for the HSE06 and HLE17 calculations. Note that all PBE results reported throughout this work are nonetheless presented using the higher accuracy settings with kppa = 1000 (which we distinguish from “PBE-lowkpt” by referring to the higher accuracy calculations simply as “PBE”).
- For the HSE06* calculations, the wavefunction from the HSE06 calculations was used as the initial guess.

To summarize the most important DFT-related settings not already discussed above, the plane-wave kinetic energy cutoff was set to 520 eV, the “accurate” precision keyword was enabled in VASP, Gaussian smearing of the band occupancies was applied with a smearing width of 0.01 eV, and symmetry considerations were disabled. The VASP v.54 projector-augmented wave (PAW)¹² PBE pseudopotentials consistent with those used in prior work in the QMOF Database¹ were used in all calculations. Although it does not influence the band gaps or partial charges discussed in this work, the D3(BJ) dispersion correction¹³ was enabled for the HSE06 calculations since they were also included during the PBE-D3(BJ) structure relaxations. The same D3(BJ) parameters were used for HSE06* as for HSE06. The dispersion-free electronic energies, energy contribution from dispersion corrections, and total energies are all made available in the QMOF Database.

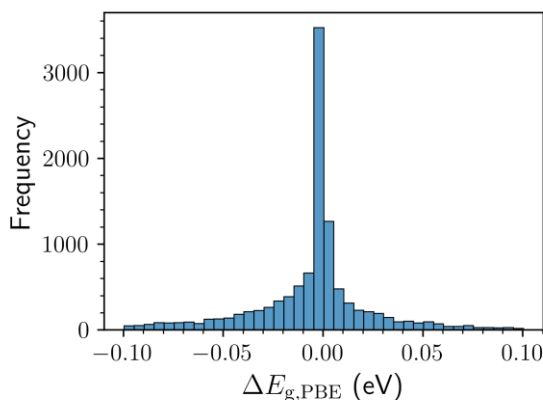
Several checks and filtering steps were performed when carrying out the HLE17, HSE06*, and HSE06 calculations:

- For computational efficiency, we only considered structures with less than or equal to 150 atoms per cell. We also only considered unique structures, as determined by using Pymatgen’s StructureMatcher utility.
- Any calculations with VASP-related errors or SCF convergence issues were discarded.
- We adopted the converged PBE-D3(BJ) magnetic moments from the QMOF Database as the initial guesses in this work, which were based on a high-spin magnetic initialization for d - and f -block

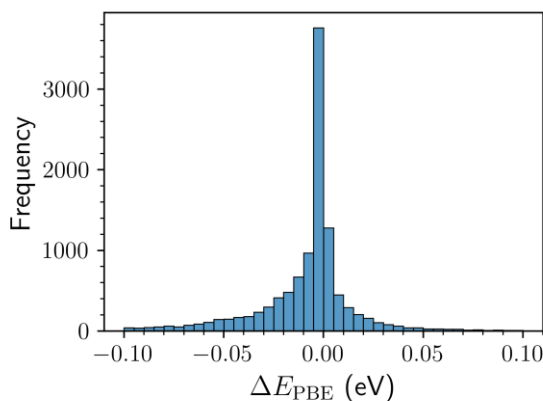
metals.¹ As a matter of internal consistency, we discarded any calculations where the final net magnetic moment converged to one different than the provided initial guess.

- Non-PBE calculations were not run if the material had a net force greater than or equal to 0.1 eV/Å on any atom at the “PBE-lowkpt” level of theory (compared to the expected maximum of ~ 0.03 eV/Å at the “PBE” level of theory with the denser k -point grid) or an energy difference greater than or equal to 0.1 eV.
- Non-PBE calculations were not run if the band gap differed by more than ± 0.1 eV between the original PBE-D3(BJ) calculations and the “PBE-lowkpt” level of theory.
- To simplify the data analysis, we only considered structures where all the investigated levels of theory were successfully completed unless otherwise stated.

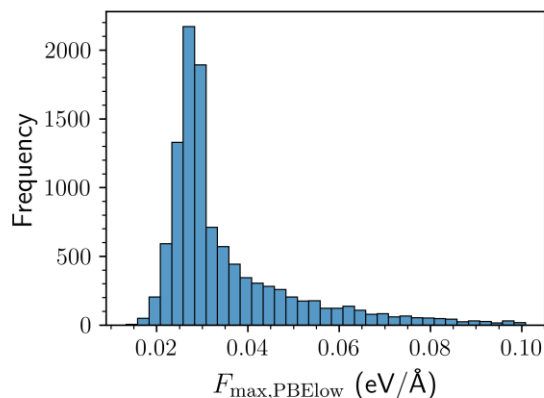
The above process resulted in a dataset of 10,720 structures with computed properties at each of the four levels of theory. For this 10,720-structure dataset, the deviations in computed band gap and total energies between the “PBE” and “PBE-lowkpt” settings are shown in Supplementary Figure 1 and Supplementary Figure 2, respectively. The maximum net force at the “PBE-lowkpt” settings is shown in Supplementary Figure 3.



Supplementary Figure 1: Histogram showing the difference in the band gaps computed using the “PBE” and “PBE-lowkpt” settings, $\Delta E_{g,PBE}$. A positive value indicates the “PBE” band gap is higher and vice versa.



Supplementary Figure 2: Histogram showing the difference in the total energies using the “PBE” and “PBE-lowkpt” settings, ΔE_{PBE} . A positive value indicates the “PBE” energy is higher and vice versa.



Supplementary Figure 3: Histogram showing the maximum net force, $F_{\text{max,PBElow}}$, computed using the “PBE-lowkpt” settings.

Additional Methodological Details

We used Pymatgen’s `eigenvalue_band_properties` function to compute the band gaps from the VASP calculations. When classifying if a MOF has closed-shell or open-shell character, we investigated the net magnetic moment and DDEC6 spin densities at the PBE-D3(BJ) level of theory. If the net magnetic moment was nonzero or the material had an absolute DDEC6 spin density on any atom greater than 0.1, the material was classified as having open-shell character; otherwise, it was classified as having closed-shell character. When classifying if an atom was a metal or nonmetal, we used the following list of nonmetals: H, He, C, N, O, F, Ne, P, S, Cl, Ar, Se, Br, Kr, I, Xe, Rn. All other elements were classified as metals (which also includes metalloids).

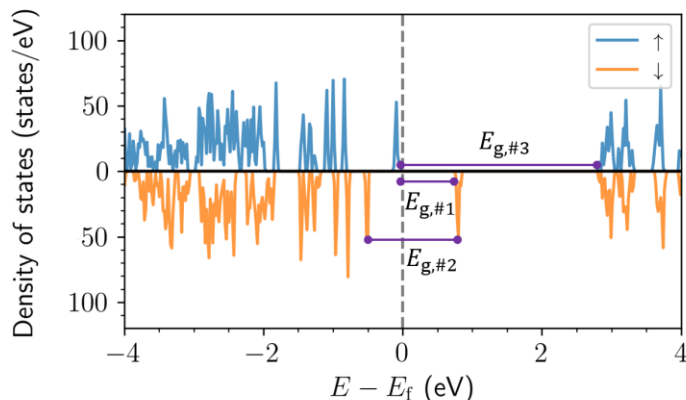
While v4 of the QMOF Database was the most up-to-date version of the database when the HLE17/HSE06*/HSE06 calculations were initiated, substantial additions have been made since then. For instance, the current version of the QMOF Database at the time of writing (v14) contains more PBE-quality structures and computed properties than v4. Therefore, for any figures involving only PBE-quality partial charges meant to compare charge calculation schemes (e.g. Figure 6), we used all the charge data available in the most up-to-date version of the QMOF Database. Because the QMOF Database currently has more DDEC6 and CM5 charge data than Bader charge data, Figure 6a shows the results for 1,429,082 atoms (Bader vs. DDEC6), whereas Figure 6b (CM5 vs. DDEC) shows the results for 2,321,435 atoms.

For the analysis shown in Figure 5d, we searched through every atom in the MOF to identify any metal sites (i.e. those not in the nonmetal list given above). Pymatgen’s `CrystalNN`^{14,15} function was used to identify any nonmetal atoms within the first coordination sphere of each metal center. The difference in partial atomic charge on the metal sites as a function of which density functional approximation is used was then compared to the difference in partial atomic charge on the ligand atoms in the first coordination sphere.

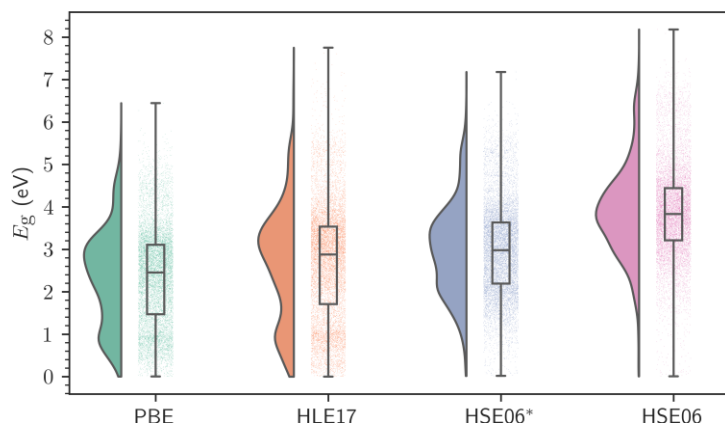
Accounting for Spin-Forbidden Transitions

Supplementary Figure 4 shows the density of states for a bipolar magnetic semiconductor, which can be described by three energy gaps. The $E_{g,\#1}$ gap in Supplementary Figure 4 represents the gap in energy between the highest energy state in the valence band and lowest energy state in the conduction band, although these two states are of opposite spin. Using $\min(\text{CBM}_{\uparrow} - \text{VBM}_{\uparrow}, \text{CBM}_{\downarrow} - \text{VBM}_{\downarrow})$ instead, the band gap is given by $E_{g,\#2}$, which does not involve a spin-flip for electronic excitation to occur. The largest gap, $E_{g,\#3}$, is given by $\max(\text{CBM}_{\uparrow} - \text{VBM}_{\uparrow}, \text{CBM}_{\downarrow} - \text{VBM}_{\downarrow})$ and is the energetic gap associated with the spin channel that is opposite the one used to define $E_{g,\#2}$.

Supplementary Figure 5 is an analogue of Figure 2 in the main text, but now E_g is calculated as $\min(\text{CBM}_\uparrow - \text{VBM}_\uparrow, \text{CBM}_\downarrow - \text{VBM}_\downarrow)$ instead of $\min(\text{CBM}_\uparrow, \text{CBM}_\downarrow) - \max(\text{VBM}_\uparrow, \text{VBM}_\downarrow)$. While there are slight changes for the spin-polarized calculations, the overarching trends between functionals remain unchanged.



Supplementary Figure 4: Density of states (DOS) for an example structure (qmof-ID: qmof-3b46c12) at the PBE level of theory where multiple band gaps can be defined, denoted $E_{g,\#1}$, $E_{g,\#2}$, and $E_{g,\#3}$. All energy values E are with respect to the Fermi level E_f . Color key: Spin-up (blue), Spin-down (orange).

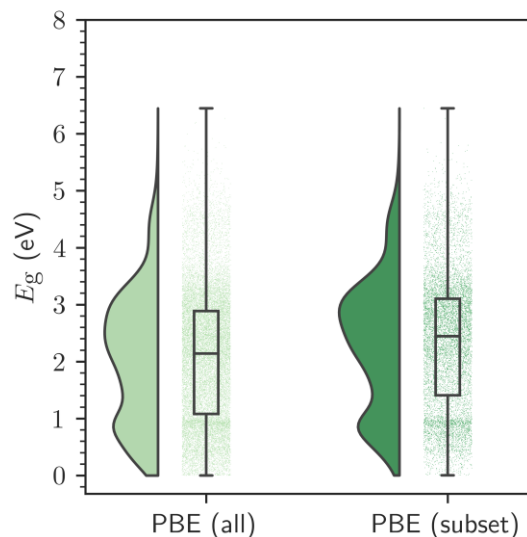


Supplementary Figure 5: Raincloud plots (i.e. combined violin plot, box plot, and strip plot) for the DFT-computed band gaps, E_g , of 10,720 structures in the QMOF Database at the PBE, HLE17, HSE06*, and HSE06 levels of theory. Unlike elsewhere in this work, E_g is defined here as $\min(\text{CBM}_\uparrow - \text{VBM}_\uparrow, \text{CBM}_\downarrow - \text{VBM}_\downarrow)$. 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.

New Structures in the QMOF Database

While updating the QMOF Database and hosting it on the Materials Project, we added over 6000 new DFT-optimized structures to the database with properties computed at the PBE-D3(BJ) level of theory. This brings the total dataset size to more than 20,000 materials at the time of writing, now including hypothetical MOFs in addition to experimentally characterized MOF structures. Unless otherwise stated, the same DFT calculation details and structural fidelity checks were used as in prior work for internal consistency and to reduce the number of MOFs structural fidelity issues (e.g. unresolved disorder, missing atoms).¹ Initial

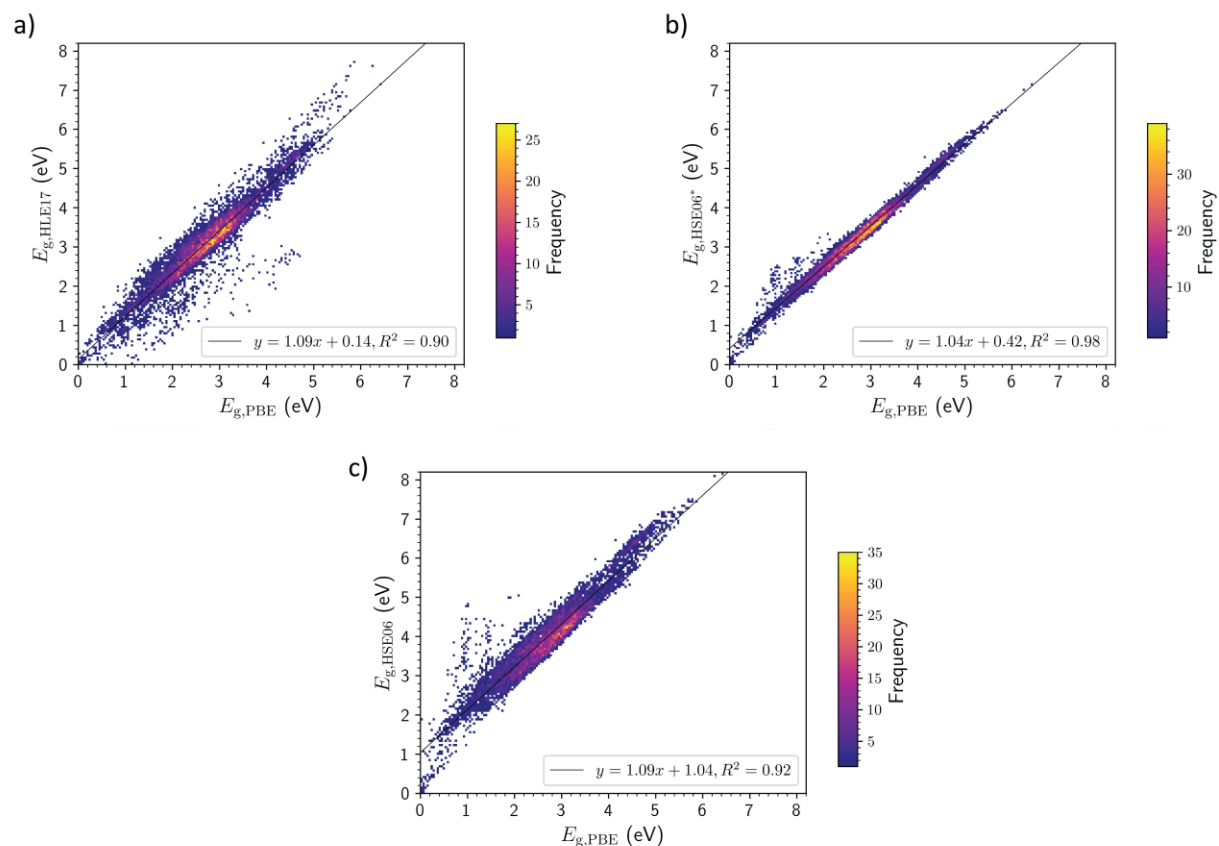
MOF structures were taken from the Cambridge Structural Database (CSD) MOF subset,¹⁶ the Boyd et al. set of hypothetical MOFs,^{17,18} Haranczyk and coworkers' hypothetical MOF-5 and MOF-74 analogues,^{19,20} Anderson and Gómez-Gualdrón's hypothetical MOF dataset,²¹ experimentally characterized pyrene-based MOFs from Smit and coworkers,²² experimental MOFs within the 2019 Computation-Ready, Experimental (CoRE) MOF Database^{23,24} based on the curated set by Chan and Manz²⁵ as well as Kancharlapalli et al.,²⁶ hypothetical MOFs in the genomic MOF (GMOF) Database,^{27,28} and hypothetical MOFs from the Topologically Based Crystal Constructor (ToBaCCo) dataset of Colón and coworkers.²⁹ The full list of structures can be found on the QMOF Database.⁷ Supplementary Figure 6 shows the distribution of PBE band gaps in the full QMOF Database compared to the subset for which all four levels of theory are available.



Supplementary Figure 6: Raincloud plots (i.e. combined violin plot, box plot, and strip plot) for the DFT-computed band gaps, E_g , at the PBE level of theory for the full QMOF Database (“PBE (all)” – 20,423 structures) and the subset of the QMOF Database for which there are band gaps at all four levels of theory (“PBE (subset)” – 10,720). 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.

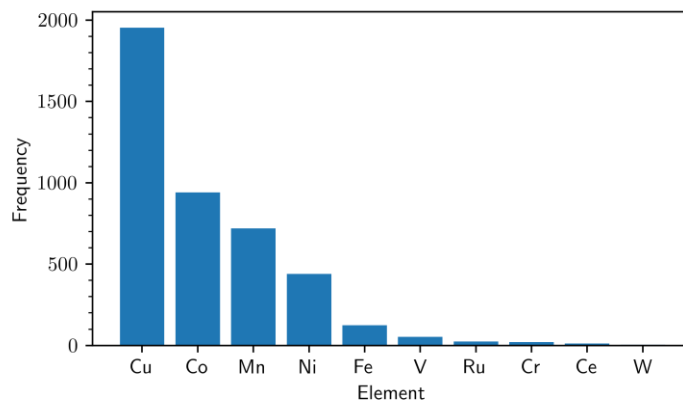
Supplementary Results

Band Gap Analyses



Supplementary Figure 7: Comparison of the DFT-computed band gaps, E_g , for 8628 closed-shell 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 size of the dataset, the data are plotted with hexagonal bins with the color bar reflecting the frequency of points in each bin. Best-fit lines (generated for the un-binned data) are shown as solid lines.

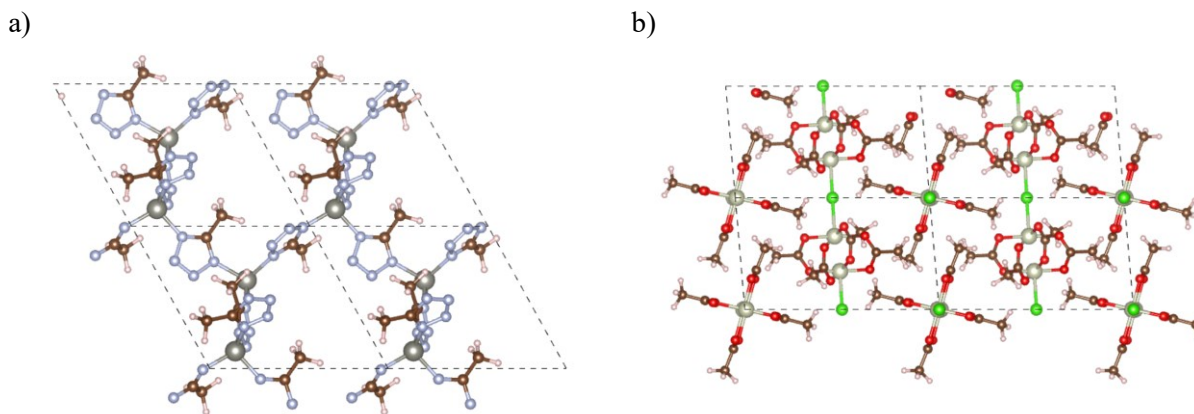
Open-Shell Structures



Supplementary Figure 8: Ten most frequent metal elements with open-shell character in the full QMOF Database at the PBE-D3(BJ) level of theory. Metal elements are only counted once per MOF.

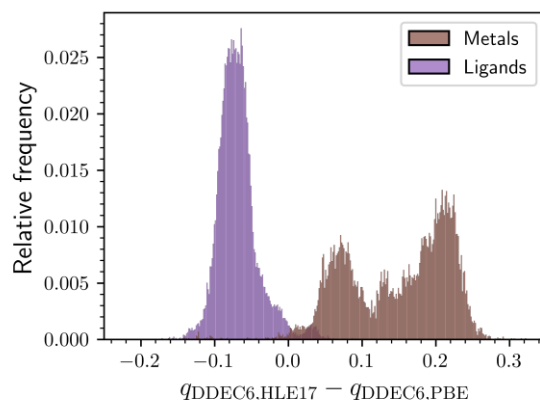
Representative Examples with Small and Large Band Gap Differences

Here, we briefly highlight two representative examples of materials in the QMOF Database with small and large band gap differences between PBE and HSE06. Towards the extreme end, the Zn-tetrazolate MOF qmof-574737f (CSD refcode: SIVGIP)³⁰ shown in Supplementary Figure 9a has a PBE-computed band gap of 5.31 eV and HSE06-computed band gap of 6.95 eV, amounting to a 1.64 eV difference. In contrast, the conductive Rh paddlewheel coordination polymer qmof-4ce077a (CSD refcode: ILICOW)³¹ shown in Supplementary Figure 9b has a PBE band gap of 0.20 eV and an HSE06 band gap of 0.50 eV, amounting to only a 0.30 eV difference. While these are merely two representative examples, the MOF Explorer web application makes it easy to carry out queries of the data underlying the QMOF Database to identify materials of interest for further computational and/or experimental investigations.

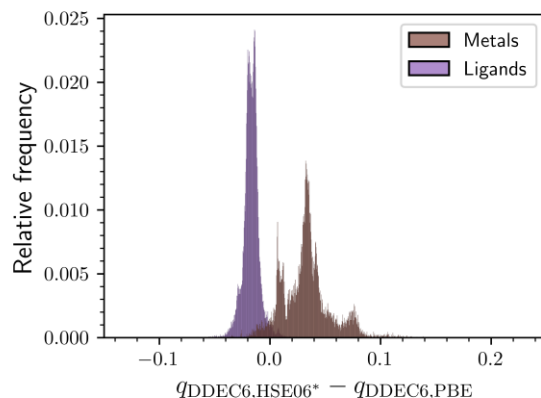


Supplementary Figure 9: a) qmof-574737f. Color key: Zn (gray), N (blue), C (brown), H (white). b) qmof-4ce077a. Color key: Rh (silver), Cl (green), O (red), C (brown), H (white).

Partial Charge Analyses

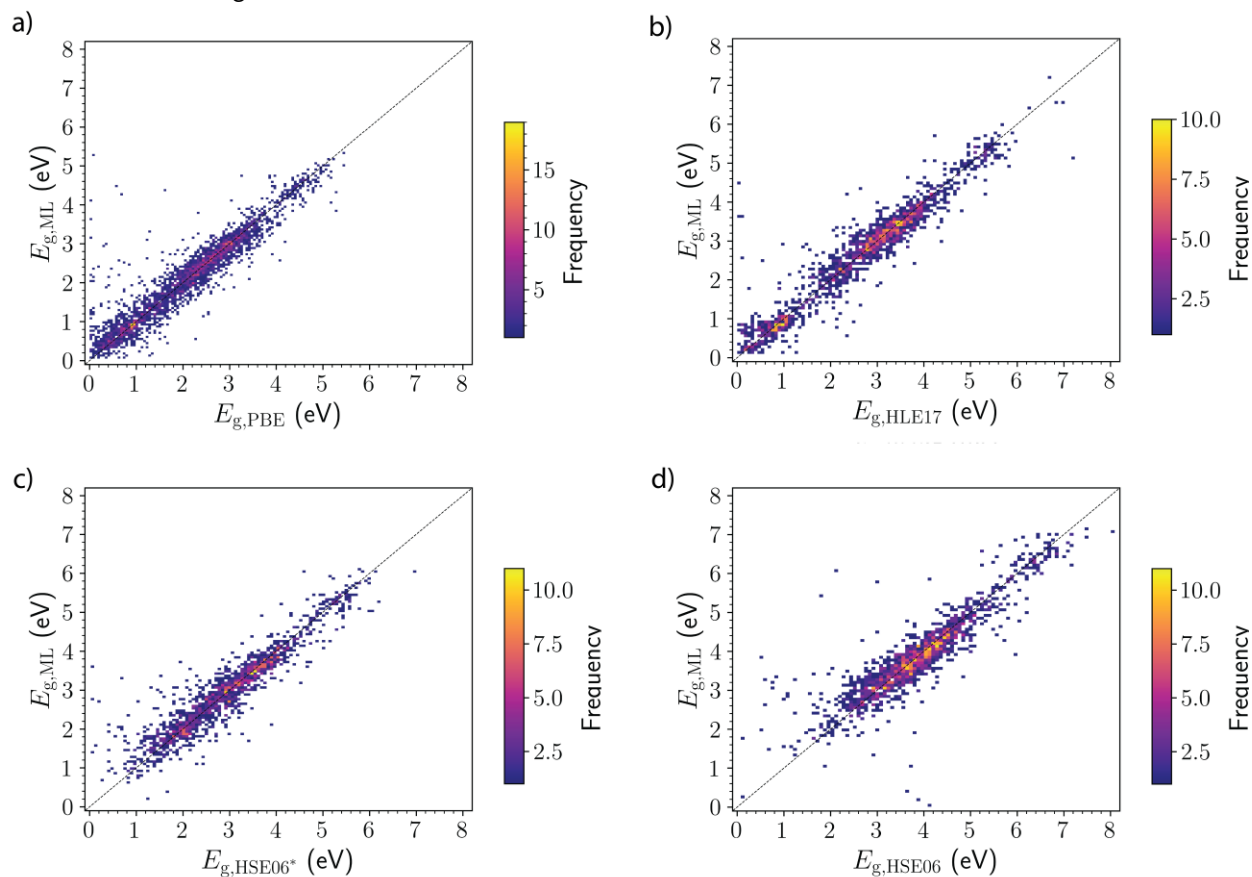


Supplementary Figure 10: Histogram of the change in DDEC6 charges between the PBE and HLE17 levels of theory for the metal sites and ligand atoms within the first coordination sphere.

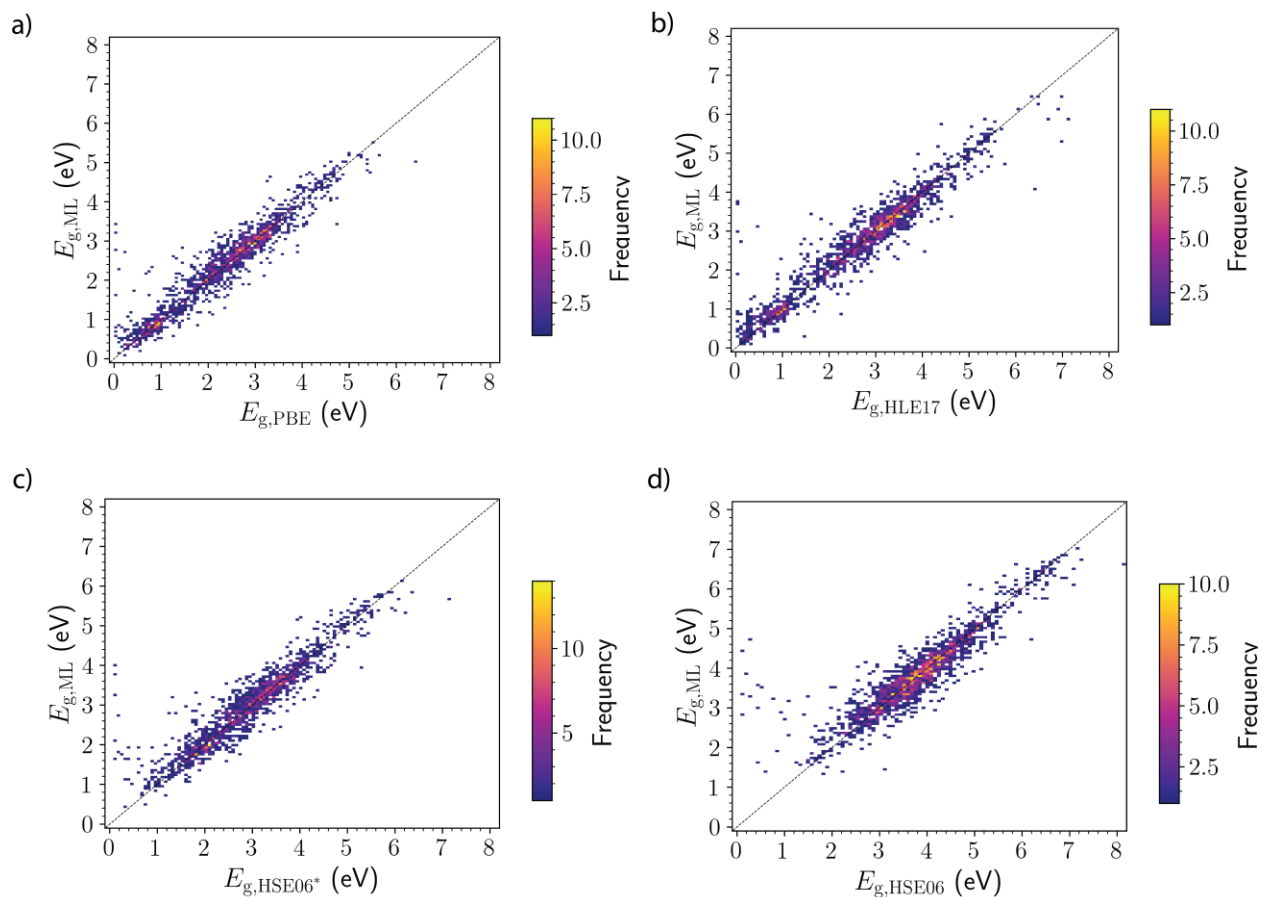


Supplementary Figure 11: Histogram of the change in DDEC6 charges between the PBE and HSE06* levels of theory for the metal sites and ligand atoms within the first coordination sphere.

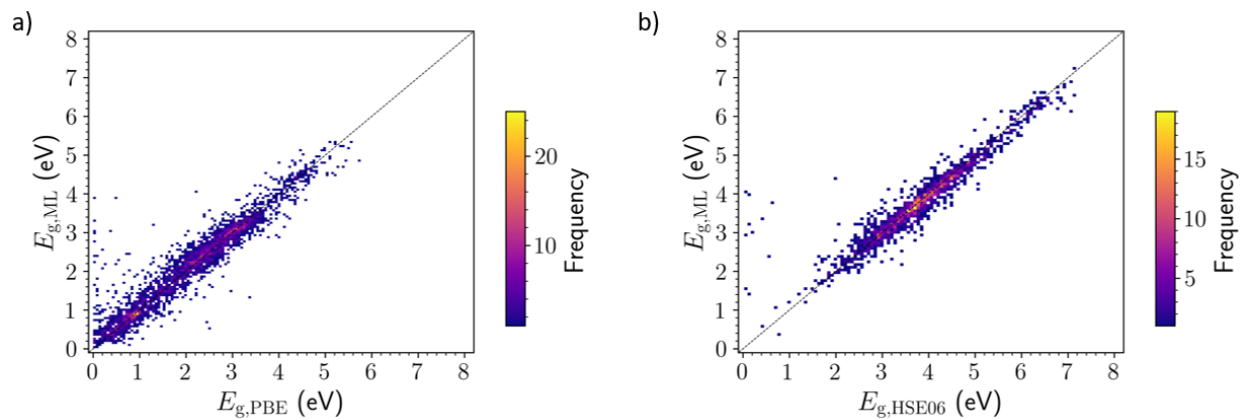
Machine Learning



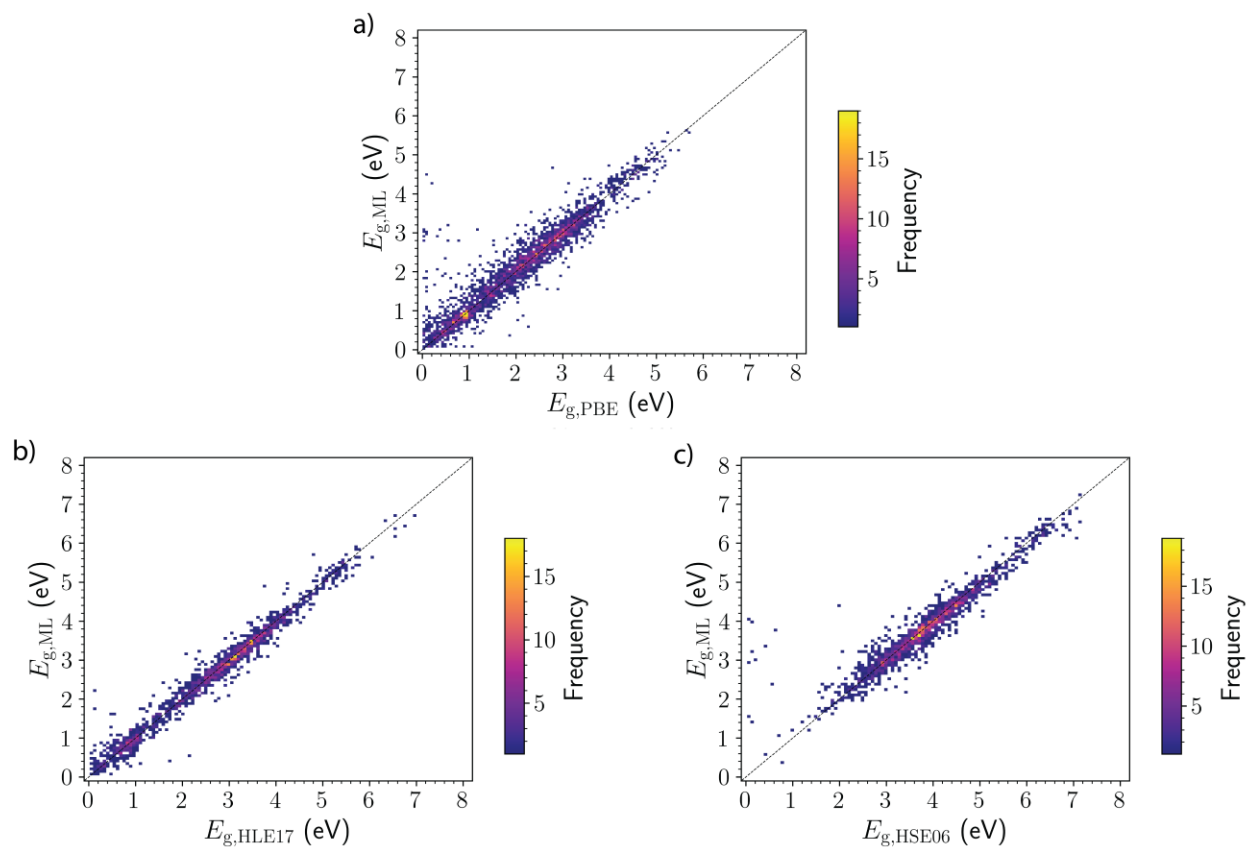
Supplementary Figure 12: Testing set parity plots between the DFT-computed band gaps and machine learned band gaps for the “individual” graph neural network models. a) ML vs. PBE. b) ML vs. HLE17. c) ML vs. HSE06*. d) ML vs. HSE06.



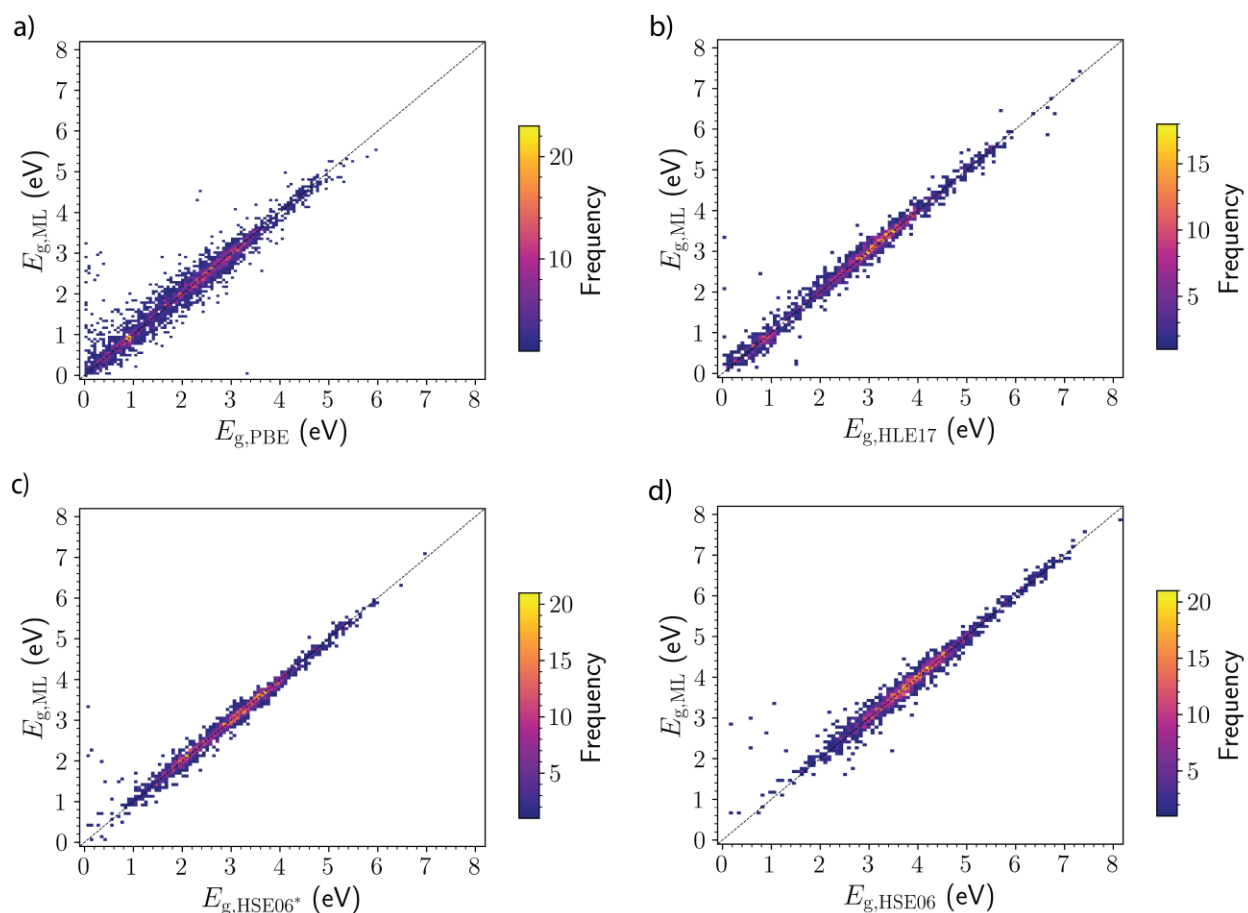
Supplementary Figure 13: Testing set parity plots between the DFT-computed band gaps and machine learned band gaps for the multi-task graph neural network model. a) ML vs. PBE. b) ML vs. HLE17. c) ML vs. HSE06*. d) ML vs. HSE06.



Supplementary Figure 14: Testing set parity plots between the DFT-computed band gaps and machine learned band gaps for the 2-fi multi-fidelity graph neural network model. a) ML vs. PBE. b) ML vs. HSE06.



Supplementary Figure 15: Testing set parity plots between the DFT-computed band gaps and machine learned band gaps for the 3-fi multi-fidelity graph neural network model. a) ML vs. PBE. b) ML vs. HLE17. c) ML vs. HSE06.



Supplementary Figure 16. Testing set parity plots between the DFT-computed band gaps and machine learned band gaps for the 4-fi multi-fidelity graph neural network model. a) ML vs. PBE. b) ML vs. HLE17. c) ML vs. HSE06*. d) ML vs. HSE06.

Supplementary References

1. Rosen, A. S. *et al.* Machine Learning the Quantum-Chemical Properties of Metal–Organic Frameworks for Accelerated Materials Discovery. *Matter* **4**, 1578–1597 (2021).
2. Manz, T. A. & Limas, N. G. Introducing DDEC6 atomic population analysis: part 1. Charge partitioning theory and methodology. *RSC Adv.* **6**, 47771–47801 (2016).
3. 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).
4. Manz, T. A. Introducing DDEC6 atomic population analysis: part 3. Comprehensive method to compute bond orders. *RSC Adv.* **7**, 45552–45581 (2017).
5. 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).
6. 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).

7. QMOF Database. <https://www.materialsproject.org/mofs>, <https://dx.doi.org/10.6084/m9.figshare.13147324>.
8. 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).
9. Teter, M. P., Payne, M. C. & Allan, D. C. Solution of Schrödinger's equation for large systems. *Phys. Rev. B* **40**, 12255–12263 (1989).
10. Bylander, D. M., Kleinman, L. & Lee, S. Self-consistent calculations of the energy bands and bonding properties of B12C3. *Phys. Rev. B* **42**, 1394–1403 (1990).
11. Freysoldt, C., Boeck, S. & Neugebauer, J. Direct minimization technique for metals in density functional theory. *Phys. Rev. B* **79**, 241103 (2009).
12. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
13. Moellmann, J. & Grimme, S. DFT-D3 study of some molecular crystals. *J. Phys. Chem. C* **118**, 7615–7621 (2014).
14. 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).
15. Pan, H. *et al.* Benchmarking Coordination Number Prediction Algorithms on Inorganic Crystal Structures. *Inorg. Chem.* **60**, 1590–1603 (2020).
16. 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).
17. Boyd, P. G. & Woo, T. K. A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. *CrystEngComm* **18**, 3777–3792 (2016).
18. Boyd, P. G. *et al.* Data-driven design of metal–Organic frameworks for wet flue gas CO₂ capture. *Nature* **576**, 253–256 (2019).
19. Martin, R. L., Lin, L.-C., Jariwala, K., Smit, B. & Haranczyk, M. Mail-order metal–Organic frameworks (MOFs): designing isorecticular MOF-5 analogues comprising commercially available organic molecules. *J. Phys. Chem. C* **117**, 12159–12167 (2013).
20. Witman, M. *et al.* In silico design and screening of hypothetical MOF-74 analogs and their experimental synthesis. *Chem. Sci.* **7**, 6263–6272 (2016).
21. Anderson, R. & Gómez-Gualdrón, D. Deep learning combined with IAST to screen thermodynamically feasible MOFs for adsorption-based separation of multiple binary mixtures. *J. Chem. Phys.* **154**, 234102 (2021).
22. Kinik, F. P., Ortega-Guerrero, A., Ongari, D., Ireland, C. P. & Smit, B. Pyrene-based metal organic frameworks: from synthesis to applications. *Chem. Soc. Rev.* **50**, 3143–3177 (2021).
23. Chung, Y. G. *et al.* Computation-ready, experimental Metal–Organic frameworks: A tool to enable high-throughput screening of nanoporous crystals. *Chem. Mater.* **26**, 6185–6192 (2014).
24. Chung, Y. G. *et al.* Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data* **64**, 5985–5998 (2019).

25. Chen, T. & Manz, T. A. Identifying misbonded atoms in the 2019 CoRE metal–organic framework database. *RSC Adv.* **10**, 26944–26951 (2020).
26. 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).
27. Lan, Y., Yan, T., Tong, M. & Zhong, C. Large-scale computational assembly of ionic liquid/MOF composites: synergistic effect in the wire-tube conformation for efficient CO₂/CH₄ separation. *J. Mater. Chem. A* **7**, 12556–12564 (2019).
28. Lan, Y. *et al.* Materials genomics methods for high-throughput construction of COFs and targeted synthesis. *Nat. Commun.* **9**, 5274 (2018).
29. Colón, Y. J., Gómez-Gualdrón, D. A. & Snurr, R. Q. Topologically Guided, Automated Construction of Metal–Organic Frameworks and Their Evaluation for Energy-Related Applications. *Cryst. Growth Des.* **17**, 5801–5810 (2017).
30. Wang, F., Hou, D.-C., Yang, H., Kang, Y. & Zhang, J. Tetrahedral tetrazolate frameworks for high CO₂ and H₂ uptake. *Dalt. Trans.* **43**, 3210–3214 (2014).
31. Amo-Ochoa, P. *et al.* MMX Chains and Molecular Species Containing Rh₂ⁿ⁺ (*n* = 4, 5, and 6) Units: Electrical Conductivity in Crystal Phase of MMX Polymers. *Eur. J. Inorg. Chem.* **31**, 4924–4932 (2010).