JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

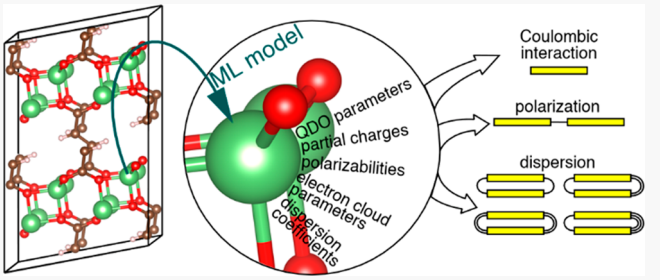
Parametrization of Nonbonded Force Field Terms for Metal−Organic Frameworks Using Machine Learning Approach

Vadim V. Korolev,* Yuriy M. Nevolin, Thomas A. Manz, and Pavel V. Protsenko

Cite This: J. Chem. Inf. Model. 2021, 61, 5774−5784 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The enormous structural and chemical diversity of metal−organic frameworks (MOFs) forces researchers to actively use simulation techniques as often as experiments. MOFs are widely known for their outstanding adsorption properties, so a precise description of the host−guest interactions is essential for high-throughput screening aimed at ranking the most promising candidates. However, highly accurate ab initio calculations cannot be routinely applied to model thousands of structures due to the demanding computational costs. Furthermore, methods based on force field (FF) parametrization suffer from low transferability. To resolve this accuracy−efficiency dilemma, we applied a machine learning (ML) approach: extreme gradient boosting. The trained models reproduced the atom-in-material quantities, including partial charges, polarizabilities, dispersion coefficients, quantum Drude oscillator, and electron cloud parameters, with accuracy similar to the reference data set. The aforementioned FF precursors make it possible to thoroughly describe noncovalent interactions typical for MOF−adsorbate systems: electrostatic, dispersion, polarization, and short-range repulsion. The presented approach can also readily facilitate hybrid atomistic simulation/ML workflows.

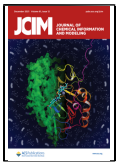


INTRODUCTION

Metal−organic frameworks (MOFs) are soft solids that form the most extensive subspace of the nanoporous materials genome.[1] Their building blocks—metal ions/clusters and organic linkers—are assembled into edge-transitive nets.[2−4] The structural variety of MOFs gives rise to diverse physical behavior.[5] Some structures possess unconventional properties for soft matter, including high electrical conductivity,[6,7] superconductivity,[8,9] and exotic topological bands.[10] However, the keen interest in MOFs is mainly due to the outstanding adsorption properties. In particular, their ultrahigh porosity enables record-breaking volumetric and gravimetric uptakes,[11] whereas specific adsorption sites help to capture the target molecule selectively.[12] MOFs are useful for the storage and separation of a wide range of gases and their mixtures, including hydrogen,[13] methane,[14,15] carbon dioxide,[16,17] and noble gases.[18,19] Unfortunately, complete experimental characterization of a representative candidate set is technically infeasible,[20] since tens of thousands of MOFs have been synthesized to date.[21,22] The hypothetical structures generated in silico are even more numerous.[23,24] For this reason, computational studies have been carried out to reveal the structure−property relationships in MOFs as often as experiments.[1,17,20,25−28]

The accuracy−efficiency dilemma is especially acute for MOFs due to their hybrid organic−inorganic nature and the relatively large sizes of the unit cells (the typical number of atoms is hundreds or even thousands). The level of theory used to describe host−guest interactions depends on the specific task faced by researchers; a broad set of approximations, differing in electronic coarse graining, have been applied.[29,30] Ab initio methods based on Møller−Plesset second-order perturbation theory and coupled cluster approaches provide accurate binding energies.[31−35] Due to the high computational demands, MOF−adsorbate systems are represented as cluster models that contain adsorption sites and gas molecules, resulting in a loss of a reliable description of the dispersion interactions. The hybrid quantum mechanics/molecular mechanics approach has been successfully adopted to solve this issue.[33] Density functional theory (DFT), a workhorse of computational materials science, has been intensively used to model the adsorption properties of MOFs as well. However, most of the popular exchange−correlation functionals based on the generalized gradient approximation do not account for intermolecular interactions properly. Therefore, long-range dispersion correction plays a critical role in the modeling of MOFs within the DFT framework. There are several generations of the empirical scheme

Received: September 13, 2021
Published: November 17, 2021



https://doi.org/10.1021/acs.jcim.1c01124
J. Chem. Inf. Model. 2021, 61, 5774−5784

proposed by Grimme and co-workers, which are usually labeled as D1,[36] D2,[37] D3,[38] and D4.[39] The van der Waals density functional method[40] implemented in the growing set of exchange−correlation functionals[41] captures the van der Waals forces via a nonlocal correlation component. General trends of the adsorption of carbon dioxide,[42−51] methane,[44,47,48,50] water,[47,50,52] and noble gases[53,54] in a series of isostructural MOFs have been revealed by employing DFT-based studies.

Ab initio and DFT methods cannot provide the scalability required for screening large MOF subsets. A few exceptions are related to the calculations of the intrinsic properties of structures,[55−57] regardless of their interactions with adsorbates. Thus, classical simulation techniques, such as the grand canonical Monte Carlo method, provide a theoretical basis for the high-throughput screening of small-molecule adsorption in MOFs.[12,23,24,58−62] In these studies, host−guest interactions are described via the nonbonded terms of force fields (FFs),[63] i.e., interaction potentials. The universal force field[64] and DREIDING[65] are the most popular generic FFs in MOF studies, but they have several well-known drawbacks. In particular, the polarization effects of the adsorbate molecules induced by open metal sites present a significant challenge for conventional FFs.[66−68] Extended versions of the universal force field[69,70] and more specialized FFs[71,72] have also been proposed. Ab initio derived[73−75] and explicitly polarizable[76−78] FFs help to significantly improve the description of intermolecular interactions only for a small series of isoreticular structures, leaving the aforementioned dilemma largely unaddressed. Therefore, to facilitate high-throughput screening adsorption studies, it is necessary to develop a fast automatized procedure for the generation of FF components that will be suitable for the various atomic types present in MOFs. In the rigid framework approximation, only nonbonded FF terms are needed.

Recently, Chen and Manz[79] have presented a collection of FF precursors that can be implemented to fully describe noncovalent interactions that are typical for MOF−adsorbate systems: electrostatic, dispersion, polarization, and short-range repulsion. A FF precursor is a computed quantitative chemical descriptor such as net atomic charge, atom-in-material polarizability, atom-in-material dispersion coefficient, etc. that is useful as a building block to construct FFs. Within this framework, partial charges calculated by the density derived electrostatic and chemical (DDEC)[80−83] approach are used to define Coulombic interactions. Dispersion in the dipole approximation is described via fluctuating polarizabilities and dispersion coefficient C_6.[84] Nondirectionally screened polarizabilities are intended to incorporate interactions between induced dipoles and external electric fields, charged atoms, permanent multipoles, or other induced dipoles into polarizable FFs.[84] In the quantum Drude oscillator (QDO) parametrization scheme,[85−87] (many-body) multipole dispersion and polarization interactions are set through the corresponding QDO parameters: mass, frequency, and charge. The electron cloud parameters fit the electron density tail of an atom to an exponential decay function. They are applicable to describe short-range exchange repulsion.[88]

Several recent studies[89−92] have partially achieved the fast generation of the FF components using machine learning (ML). Particularly, ML algorithms make it possible to predict partial charges in MOFs within the accuracy of the underlying DDEC approach. At the same time, ML techniques are comparable for empirical charge equilibration[93] methods in terms of scalability.

In this study, we applied a data-driven approach to derive a full suite of atom-in-material quantities required for advanced FF parametrization. Taking a collection of high-quality FF precursors extracted for 3056 MOFs as initial data, we implemented gradient boosting models on top of a diverse set of features that described the local site environment. The combination of a state-of-the-art approximation algorithm and a data representation scheme outperformed previous approaches for partial charge assignment. The trained models for other FF precursors demonstrated high accuracy in terms of the correlation coefficients. The relative contributions of these features to the model performance were estimated by means of two methods, including a game-theoretic approach. In addition, we outlined future opportunities for the presented ML approach.

MATERIALS AND METHODS

Reference Database. We used a collection of 3056 MOFs[79] as a starting data set. Each structure included the atomic coordinates and the corresponding FF precursors. For further consideration, the following nine FF precursors were selected:

1. Atomic partial charge: This parameter quantifies the net charge of an atom in a material. This quantity is a real number.
2. Dispersion coefficient C_6: The long-range London dispersion interaction due to fluctuating dipoles is proportional (to leading order) to $-C_6/r_6$, where r is the distance between two atoms. Each atom in the material had its own C_6 value that was used a FF precursor; this represents the C_6 coefficient for each atom interacting with a like atom. This quantity is a positive real number.
3. Fluctuating polarizability: For each atom, the fluctuating polarizability is the polarizability applicable to fluctuating dipole moments associated with the London dispersion interaction. These fluctuating dipole moments have short-range directional order and long-range directional disorder. This quantity is a positive real number.
4. Nondirectionally screened (aka FF) polarizability: The polarizability of each atom computed with no directional alignment of atomic dipole moments. This polarizability is appropriate for use in classical FFs, because the directional alignment of atomic dipoles occurs during the classical molecular dynamics or Monte Carlo simulation when using the FF. To avoid double counting, the directional alignment of dipoles must not be included in the underlying FF polarizability. This quantity is a positive real number.
5. QDO reduced mass: This is the effective mass of the pseudoelectron in the QDO model. This quantity is a positive real number.
6. QDO effective charge: This is the effective charge of the pseudoelectron in the QDO model. This quantity is a negative real number.
7. QDO effective frequency: This is the effective frequency of the QDO oscillator. This quantity is a positive real number.
8. Electron cloud parameter a: The electron cloud parameters a and b were defined and computed by the least-squares fit of $\ln(\rho_A^{avg}(r_A))$ to the model function $a - br_A$ for the atom's tail region, where r_A is the distance from atom A's nucleus and $\rho_A^{avg}(r_A)$ is the spherically averaged electron density of atom A at distance r_A. This least-squares fit was performed over the region where $10^{-4} \le \rho_A^{avg}(r_A) \le 10^{-2}$ electron/bohr^3.[79]

https://doi.org/10.1021/acs.jcim.1c01124
J. Chem. Inf. Model. 2021, 61, 5774−5784

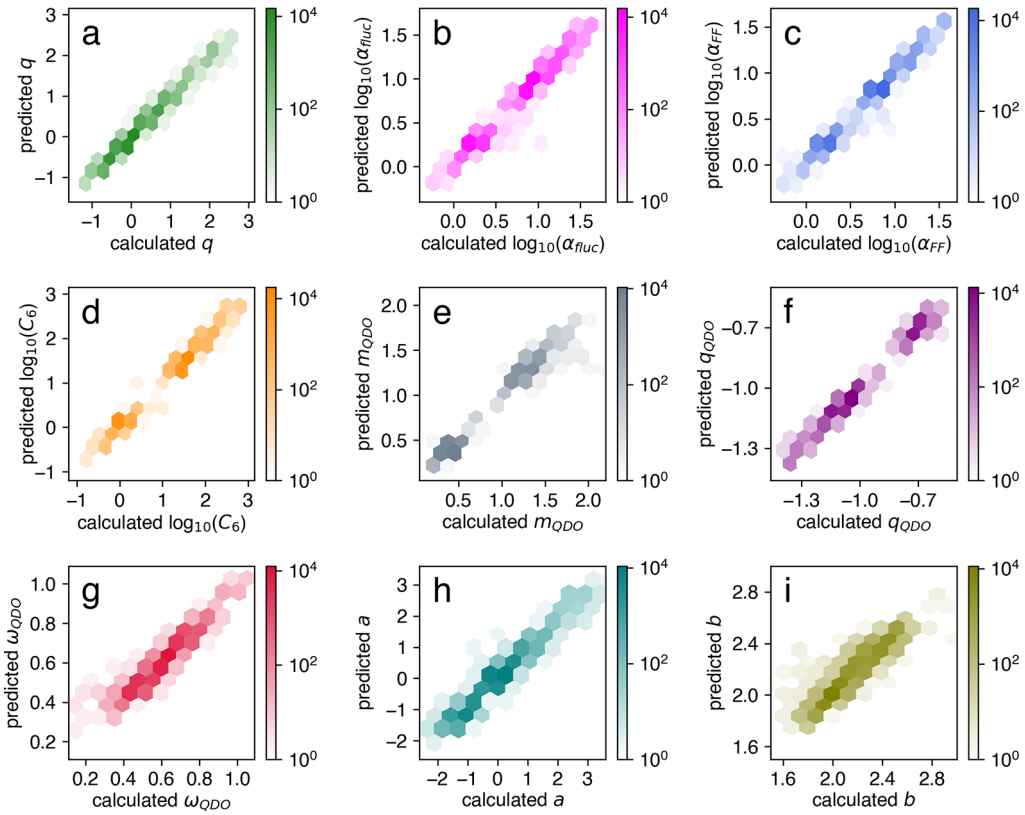


Figure 1. Parity plots of the calculated results of Chen and Manz[79] and ML-predicted FF precursors: (a) partial charge, (b) fluctuating polarizability, (c) FF polarizability, (d) dispersion coefficient C_6, (e) QDO mass, (f) QDO charge, (g) QDO frequency, (h) electron cloud parameter a, and (i) electron cloud parameter b.

Thus, $\rho_A^{avg}(r_A) \approx \exp(a - br_A)$ in the atom's tail region. This quantity is a real number.

9. Electron cloud parameter b: This is the quantity b in the above model. This quantity is a positive real number.

Manz and Chen[79] extracted partial charges and electron cloud parameters using the DDEC6 charge partitioning scheme.[80−83] The dispersion coefficients, polarizabilities, and QDO parameters were calculated with the MCLF method[84,94] applied to the DDEC6 electron density partitions. (MCLF is an acronym from the last names of the authors introducing the method.[84]) The selected FF precursors represent the minimum set required to describe all nonbonded interaction terms thoroughly.

Fingerprints. The properties of an atomic site, beginning with forces[95] and including atom-in-material parameters, are functions of its local environment. In this study, the following diverse set of chemical and structural features was used as input data for the approximation algorithm (ML model):

- The set of descriptors inspired by the electronegativity equalization principle was originally implemented by Kan-charlapalli et al.[92] We used its extended version (referred to as ENFingerprint) that included the electronegativity and first ionization energy of the considered atomic site, averaged electronegativity and averaged first ionization energy of the sites in the first and second coordination spheres, distances between the target atomic site and sites in its first and second coordination spheres, and the corresponding numbers of sites. The first and second coordination shells included sites that formed a bond with the considered site directly and through one of its nearest neighbors, respectively. Two sites were considered to be bonded if the interatomic distance did not exceed the sum of the corresponding Cordero covalent radii,[96] within a penalty distance of 0.5 Å. The thermochemical scale[97] of the dimensionless electronegativities was used.
- The adaptive generalizable neighborhood informed (AGNI)[98,99] fingerprints are integrals of the product of the radial distribution function and the damping function f_d:

$$V_i(\eta) = \sum_{j \neq i} e^{-(r_{ij}/\eta)^2} f_d(r_{ij}) \quad (1)$$

$$f_d(r_{ij}) = 0.5\left[\cos\left(\frac{\pi r_{ij}}{R_c}\right) + 1\right] \quad (2)$$

where r_{ij} is the distance between sites i and j, R_c is the cutoff distance, and η is the Gaussian function width.

- Voronoi-tessellation-based[100−102] fingerprints summarize the features of the Voronoi cells, including the Voronoi indices, the (weighted) i-fold symmetry indices, and the Voronoi volume, area, and nearest-neighbor distance statistics (mean, standard deviation, and minimum and maximum values).
- CrystalNNFingerprint[103,104] and OPSiteFingerprint[103,104] were described as a site environment via the coordination likelihoods and specific local structure order parameters. CrystalNN[104] and the minimum distance[103] neighbor-finding algorithms were used, respectively.

All the aforementioned fingerprints were concatenated into a 109-dimensional vector (the full list of fingerprints is provided in the Supporting Information). The crystal structure processing routines were carried out with the Python Materials Genomics[105] and Atomic Simulation Environment[106] modules.

https://doi.org/10.1021/acs.jcim.1c01124
J. Chem. Inf. Model. 2021, 61, 5774−5784

AGNI, Voronoi, CrystalNN, and OPSite fingerprints were calculated by the matminer[107] library.

Some structures failed to assign one of two (or both) subsets of features during the featurization: Voronoi and ENFingerprint. The issue related to the Voronoi tessellation could be resolved by increasing the cutoff radius when determining the Voronoi neighbors (a default value of 6.5 Å was applied). However, ENFingerprint could not be used for structures containing small (with the longest path in molecular graph no more than three) ions, such as H_3O^+, NH_4^+, and NO_3^-. This is because there was no second coordination sphere for noncentral atoms in such ions. Thus, charged MOFs were naturally excluded from further analysis as well. Finally, unique sites from 2946 structures were used to train/validate ML models.

Machine Learning (ML) Algorithm. Within a local approximation, FF precursors are defined by the site's fingerprints. These generally unknown functions are approximated by the refined implementation of the gradient boosting algorithm, eXtreme Gradient Boosting (XGBoost).[108] The classification and regression tree model, which is a tree ensemble model ϕ as a superposition of K additive functions f, represents the true value of a target property y for the ith site in the following form:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i) \tag{3}$$

where x is the site's representation. The parameters of the classification and regression tree model (tree structure and leaf weights) are fitted during the iterative minimization of regularized objective $\mathcal{L}$:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{4}$$

where l is the differentiable loss function and Ω is the penalty term designed to avoid overfitting via regularization of the model's weights.

The XGBoost models were trained to predict each FF precursor independently. Preliminarily, the calculated fingerprints were scaled by using MinMaxScaler and were rounded jointly with target values to the fourth decimal; duplicated data were excluded from consideration. We tested models on an external test set (10% of points from the initial set) with 5-fold cross-validation. The optimal values of the hyperparameters (including the number of gradient boosted trees, maximum tree depth, and boosting learning rate) were determined by using the tree-structured Parzen Estimator[109] algorithm implemented in the Hyperopt[110,111] library.

RESULTS AND DISCUSSION

Performance of ML Models. The parity plots of the calculated results of Chen and Manz[79] and the ML-predicted FF precursors are reported in Figure 1. The corresponding histograms of the deviations of the predicted values from the reference values are presented in Figure S1. Table 1 summarizes the performances of the trained ML models intended for FF precursor prediction. The most commonly used regression metrics—mean absolute error (MAE), root-mean-square error (RMSE), and coefficient of determination (R^2)—are shown here. The Pearson and Spearman coefficients, which measure the linear and rank correlations, respectively, are also provided. In general, a high density of points near the diagonal of the parity plots and high values (>0.96) of the three coefficients (R^2, Pearson, and Spearman) indicate the superior predictive capabilities of the presented models. However, these metrics do not provide insights into efficiency for specific modeling tasks per se. In other words, it is unclear whether the presented models simulate the adsorption properties of MOFs with sufficient accuracy.

Table 1. Summary of Performances of Presented Machine Learning (ML) Models

FF precursor	MAE	RMSE	R^2	Pearson	Spearman
partial charge q	0.0113	0.0216	0.9970	0.9985	0.9960
fluctuating polarizability $\log(\alpha_{fluc})$	0.0095	0.0159	0.9977	0.9989	0.9905
FF polarizability $\log(\alpha_{FF})$	0.0070	0.0126	0.9982	0.9991	0.9917
dispersion coefficient $\log(C_6)$	0.0134	0.0217	0.9990	0.9995	0.9923
QDO mass m_{QDO}	0.0090	0.0196	0.9976	0.9988	0.9918
QDO charge q_{QDO}	0.0042	0.0067	0.9985	0.9993	0.9928
QDO frequency ω_{QDO}	0.0081	0.0129	0.9794	0.9897	0.9863
electron cloud parameter a	0.0554	0.0871	0.9816	0.9908	0.9828
electron cloud parameter b	0.0225	0.0358	0.9627	0.9814	0.9785

This difficulty can be partially resolved by comparing our models with those available in the literature. The following ML approaches were used to predict the partial charges in MOFs: the multilayer connectivity-based atom contribution (m-CBAC) method developed by Zou et al.,[90] the message passing neural networks (MPNNs) implemented by Raza et al.,[91] random forest models in conjunction with features inspired by the electronegativity equalization principle (PACMOF[92] package), and our implementation[89] that included local structural features as inputs to the XGBoost models. The direct comparison of the listed approaches is hindered by the differences in the used partitioning scheme (DDEC3[112] versus DDEC6[80−83]) and the sets of MOFs. In addition, although in all these studies the number of structures under consideration was about 3000, the data sizes differed significantly. It is well-known that the availability of materials data has a significant impact on the predictive precision of ML models.[113] Therefore, the following estimates are general and are not true performance benchmarks. In terms of the MAE, the presented partial charge predictor, with an MAE of 0.0113 e^-, slightly outperformed the PACMOF[92] and MPNN,[91] with MAEs of 0.0192 and 0.025 e^-, respectively. The less representative Pearson and Spearman coefficients are given for the m-CBAC[90] approach. Their values (0.997 and 0.984, respectively) were lower than those obtained in this study (0.9985 and 0.9960). The only close competitor was our previous implementation,[89] which showed an even smaller MAE of 0.0096 e^-. The insignificant difference may have been due to the distinction in the featurization schemas and, more importantly, the removal of duplicate data in this study.

The aforementioned approaches[89−92] have been validated by comparing values of the adsorption properties calculated using ML-derived and DDEC charges. These studies used ML-derived and DDEC charges in classical FFs employed in Monte Carlo simulations to compute various gas adsorption properties in MOFs; the ML-derived atomic charges were validated by comparing the resulting gas adsorption properties with

https://doi.org/10.1021/acs.jcim.1c01124
J. Chem. Inf. Model. 2021, 61, 5774−5784

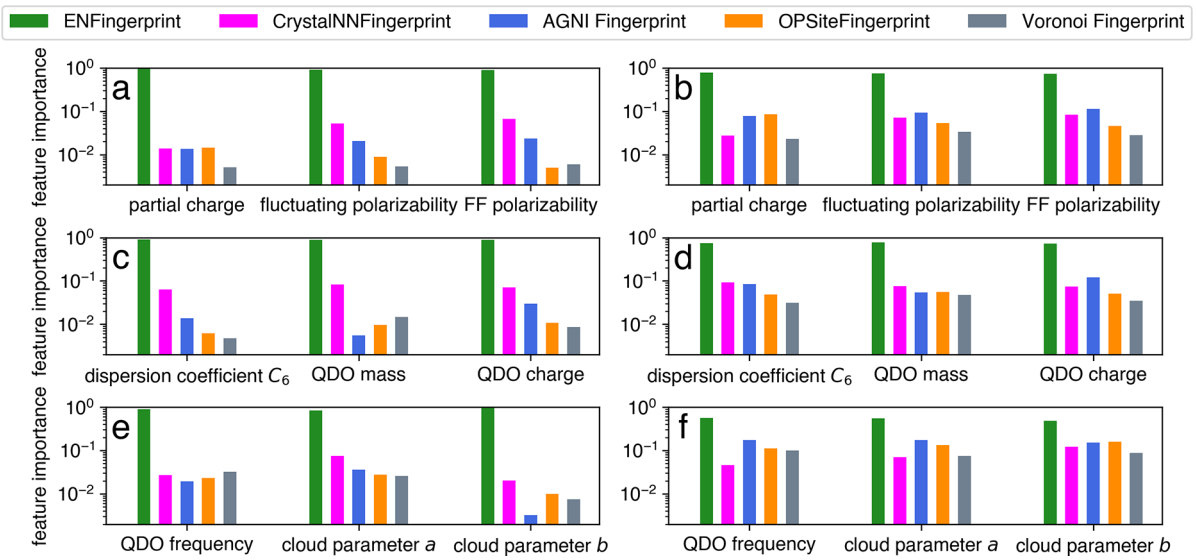


Figure 2. Cumulative feature importances corresponding to fingerprint subsets. The reported values are obtained by (a, c, e) gain-based method and (b, d, f) SHAP analysis.

those obtained when the DDEC charges were used instead. These atomic charges provided an electrostatic interaction model between gas molecules and the MOF. The Spearman rank coefficient between the CO_2 Henry coefficients computed with DDEC and ML-derived charges (obtained by the m-CBAC[90] approach and MPNNs[91]) equaled 0.939 and 0.96, respectively. The Spearman rank coefficient for the CO_2 volumetric uptakes computed with the DDEC and ML-derived charges presented in our previous study[89] was 0.991. PACMOF[92] could reproduce the CO_2 loading, N_2 loading, and CO_2/N_2 selectivity with mean absolute percentage errors of 18.9, 28.3, and 33.9, respectively. Thus, the ML models that yielded MAEs of 0.01−0.02 e^- were sufficiently accurate to reproduce the values of the adsorption properties obtained by using the DDEC charges. In this context, the Spearman rank correlation coefficient is much more representative than in the case of partial charge prediction, since ranking promising candidates for a specific application can be seen as a primary goal of high-throughput screening studies. Similar estimates for other FF precursors are not available.

Another aspect of the ML model's efficiency concerns how its accuracy relates to the reference method. In computational chemistry, the so-called chemical accuracy (∼1 kcal/mol) usually serves as a desirable level of precision for reproducing potential energy surfaces by ab initio methods. Recently, the same can be said about ML models trained on calculated data.[114,115] From a more general perspective, the following guiding principle can be formulated: the accuracy of an approximation model that relies on quantum-chemical inputs should be at least comparable to the accuracy of the underlying computational method relative to the experimentally measured quantities. As for FF precursors, extracting experimental values is quite complex and not straightforward, so such an analysis can be carried out for a very limited set of available data. Thus, the MAE deviations of the DDEC6 charges from those of charges derived via kappa refinement[116−118] for natrolite and formamide were 0.1174 and 0.0570 e^-, respectively.[81] The MCLF method yielded the static polarizability tensor eigenvalues for six small organic molecules and dispersion coefficients C_6 for pairs formed from 49 atoms/molecules within mean absolute relative errors of 8.10 and 4.45%, respectively.[94] The static polarizabilities and dispersion coefficients C_6 for 12 polyacenes defined by this method had mean absolute relative errors within 8.75 and 7.77%, respectively. The values for six fullerenes were 5.92 and 6.84%, respectively.[94] Our XGBoost models reproduced the fluctuating polarizabilities extracted by using the MCLF method and the reference dispersion coefficients C_6 with the following mean absolute relative errors: 2.18 and 3.07%, respectively. We speculate that the accuracies of the presented approximation algorithms are comparable to the precision of the reference methods, DDEC6 and MCLF.

ML Model Interpretability. The selection of a reliable material representation is an essential step in constructing a precise predictive model.[119−121] The initial choice of descriptors is usually based on domain expertise. For instance, previous studies that aimed at partial charge assignment using ML techniques used a small set of physically meaningful parameters[92] or a collection of atomic-connectivity-based patterns.[90] In this study, we used a different approach, applying a diverse suite of heterogeneous fingerprints. The validity of this approach, i.e., nonredundancy of chosen features, can be confirmed by conducting feature importance analysis. As a result, the revealed input parameters that do not contribute to the model's output can be reasonably excluded from consideration. First, we calculated the analogue of one of the most popular feature importance measures for ensemble learners, also known as Gini importance ("gain" in XGBoost implementation).[122] This quantity is defined as the mean decrease in impurity, which is the sum of all decreases in the Gini impurity over all trees in the ensemble. The normalized values of the gain-based importance for five feature subsets are shown in Figure 2a,c,e. ENFingerprint made a decisive contribution: its importance varied from 83.5 to 95.9% for the electron cloud parameters *a* and *b*, respectively. On the basis of these data, it can be concluded that all other subsets had a negligible impact on the performances of the ML models. To confirm this, we retrained models to predict the partial charges using only the ENFingerprint as an input (importance of 95.2%). Surprisingly, the MAE increased from

https://doi.org/10.1021/acs.jcim.1c01124
J. Chem. Inf. Model. 2021, 61, 5774−5784

0.0113 e^- (see Table 1) to 0.0185 e^-; i.e., it grew by 63%. Such a dramatic decrease in the model accuracy was inconsistent with the minor importance of the four other fingerprint subsets determined by a gain-based method and was likely due to its intrinsic shortcomings, including sampling bias.[123]

We then calculated the SHAP (SHapley Additive exPlanations) values using the TreeExplainer[124] algorithm to obtain a more reliable estimate of the feature importance. These quantities represent an extended version of the classical Shapley values,[125] which originate from game theory. Explanations of the predictions expressed using the SHAP values are guaranteed to satisfy the following properties: local accuracy, monotonicity, and missingness. Global feature attribution was carried out by averaging the magnitudes of the SHAP values over all testing set points, since TreeExplainer provides local objectwise explanations. Mean SHAP values normalized to unity over all features are provided in Figure 2b,d,f. The impact of ENFingerprint was significantly decreased compared to gain-based feature importance values and varied from 48.0% (electron cloud parameter b) to 78.6% (partial charge). The cumulative importance for each of the other fingerprint subsets reached 10% for at least one FF precursor. Thus, the used suite of features should be considered nonredundant and reasonable.

Future Opportunities. In addition to feature representation and approximation algorithm, training data availability also determines the accuracy of ML predictors.[113] We trained a series of models on data sets that differed in size (from 1000 to 300 000 atomic sites) to extract this dependency for three low-correlated targets: partial charge, fluctuating polarizability, and QDO frequency. The corresponding data set sizes versus scaled error dependencies are presented in Figure 3. The

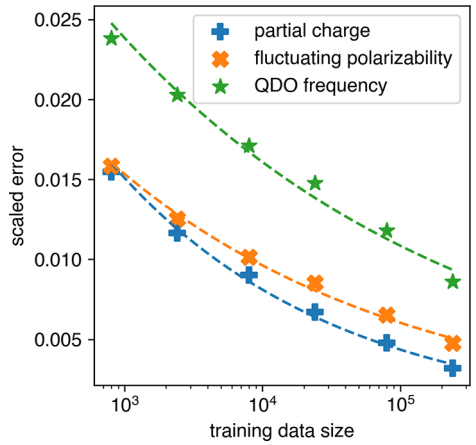


Figure 3. Scaled error as a function of training data size.

power law SE = mDS^{-n} almost perfectly describes all three sets of points, where SE is the scaled error (MAE normalized by the range of the corresponding FF precursor), DS is the training data size, and m and n are empirical parameters. It should be noted that coefficient n (the slope of a line in logarithmic coordinates) differed in each case: 0.27 ± 0.03, 0.20 ± 0.03, and 0.17 ± 0.04 for the partial charge, fluctuating polarizability, and QDO frequency, respectively. The given values are significantly lower than those obtained for a diverse set of properties[113] (0.372) and the formation energy of perovskites[126] (0.297). Therefore, the universal dependency derived by Zhang and Ling[113] was not observed, at least for the atom-in-material quantities considered in this study. Nevertheless, the revealed power law suggests that the FF precursors' accuracy can be improved extensively by increasing the training data size.

As indicated under Materials and Methods, ML models for each FF precursor were trained independently. However, due to the use of a common representation of the atomic site, a multitask learning[127] framework can be efficiently applied here. The performance of the primary method for multitask learning, deep neural networks, improves in the presence of highly correlated targets.[128] To assess the potential of multitask learning for FF precursor predictions, we calculated Pearson coefficients for all pairs of the considered atom-in-material quantities. The correlation matrix in the form of a heatmap is shown in Figure 4. Two groups of highly correlated FF precursors can be distinguished. The first group includes the fluctuating polarizability, FF polarizability, dispersion coefficient C_6, and electron cloud parameter a. The second group contains the QDO mass, QDO charge, and electron cloud parameter b. Therefore, multitask learning predictors for the listed end points can potentially outperform the corresponding single-task models.

The presented models can also be helpful to facilitate the ML prediction of adsorption properties. Data-driven approximations are at best able to reproduce the quantitative structure−property relationships hidden in the input data but still inheriting errors specific to the underlying computational approach. Thus, most ML models (as opposed to atomwise predictors) that were aimed at predicting macroscopic adsorption properties[25,27] were trained on results of grand canonical Monte Carlo simulations, for which the universal force field is almost no alternative choice for describing noncovalent interactions. Therefore, the outputs of those ML models suffered from the issues mentioned in the Introduction, such as lacking a description of the polarization effects. The following hybrid workflow can improve the reliability of the predicted targets without losing scalability: advanced parametrization using a full suite of FF precursors (main scope of this study) → high-throughput screening adsorption modeling in rigid framework approximation → construction of ML predictors of macroscopic properties.

CONCLUSIONS

In summary, we present the ML workflow to reproduce atom-in-material quantities useful for the parametrization of FFs. Extreme gradient boosting models trained on a diverse set of descriptors have reached state-of-the-art performance in partial charge prediction (mean absolute error about 0.01 e^-), as follows from comparative analysis with prior studies. Our choice of fingerprints has been confirmed by feature importances extracted using the game-theoretic approach. Each of five subsets that describe local atomic environments from different perspectives has a nonnegligible cumulative contribution to model outputs, at least for one of the considered FF precursors. We also show that several types of FF precursors (partial charges, fluctuating polarizabilities, and dispersion coefficients) are reproduced by presented models within the accuracy comparable to the accuracy of the applied computational method relative to available experimental measurements. The impressive performance of trained models (coefficient of determination, Pearson and Spearman coefficients take values more than 0.96 for all FF precursors) can be improved extensively by increasing the training data size,

https://doi.org/10.1021/acs.jcim.1c01124
J. Chem. Inf. Model. 2021, 61, 5774−5784

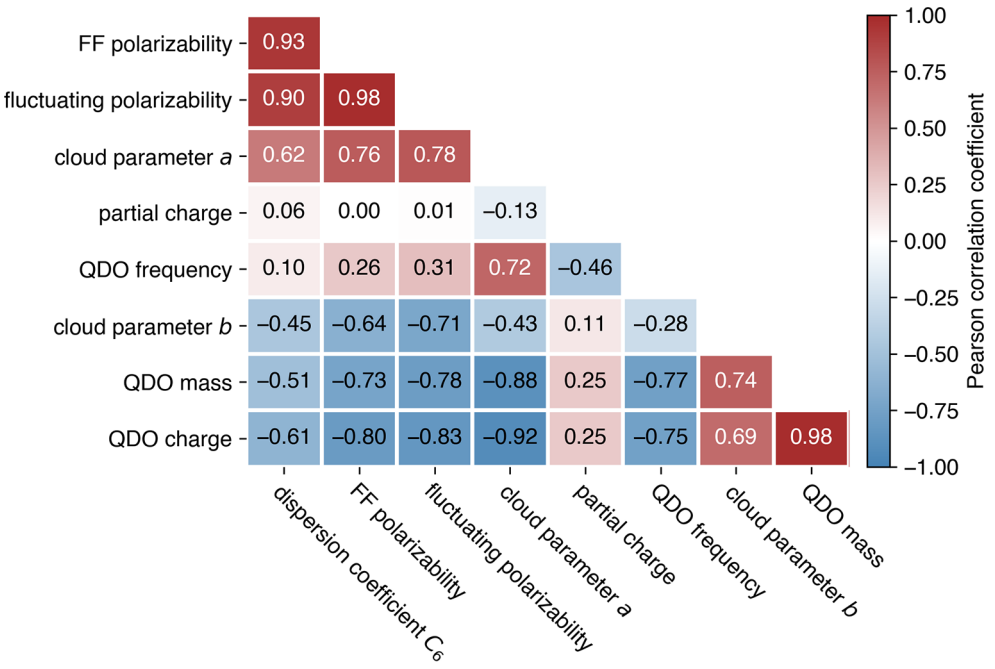


Figure 4. Correlation matrix for considered FF precursors.

i.e., by considering new types of local environments. Another avenue for improving the performance of ML predictors is multitask learning due to the presence of two highly correlated subsets of targets.

The modular structure of presented workflow, typical for data-driven approaches, rests on three pillars: input data, feature representation, and approximation algorithm. Each of the parts can be modified depending on the specific task. In principle, our approach is also applicable for other subclasses of nanoporous materials, including covalent organic frameworks and hydrogen-bonded organic frameworks. Since the transferability of the presented models to structures beyond MOFs is in question, reliable results can be obtained using reference DDEC and MCLF data derived for a specific subclass of materials under consideration. The set of local features, i.e., the input for the ML algorithm, is modifiable as well. However, it is highly desirable to confirm the validity of the new set based on feature importance, as was demonstrated in this study. Finally, a reasonable choice of the approximation algorithm requires full-fledged benchmarking that takes into account accuracy and time efficiency.

DATA AND SOFTWARE AVAILABILITY

All MOF crystal structures and corresponding FF precursors used to train XGBoost models are available as Supporting Information at https://doi.org/10.1039/C9RA07327B. The full pipeline, including featurization and FF precursor prediction, is shared through GitHub as an open-source python library, FFP4MOF (https://github.com/korolewadim/ffp4mof). The trained XGBoost models are available on Zenodo at https://doi.org/10.5281/zenodo.5500641.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c01124.

Histograms of the deviations of the predicted FF precursor values from the reference FF precursor values; list of used fingerprints (PDF)

AUTHOR INFORMATION

Corresponding Author

Vadim V. Korolev − *Department of Chemistry, Lomonosov Moscow State University, Moscow 119991, Russia;* orcid.org/0000-0001-6117-5662; Email: korolev@colloid.chem.msu.ru

Authors

Yuriy M. Nevolin − *Frumkin Institute of Physical Chemistry and Electrochemistry, Russian Academy of Sciences, Moscow 119071, Russia*

Thomas A. Manz − *Department of Chemical & Materials Engineering, New Mexico State University, Las Cruces, New Mexico 88003-8001, United States;* orcid.org/0000-0002-4033-9864

Pavel V. Protsenko − *Department of Chemistry, Lomonosov Moscow State University, Moscow 119991, Russia;* orcid.org/0000-0002-1503-3679

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c01124

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Prof. Vladimir G. Sergeyev (Lomonosov Moscow State University) for helpful comments. This study was funded by the Ministry of Science and Higher Education of Russian Federation, Project No. 121031300084-1. T.A.M.

received support from a Career Award from the National Science Foundation.

REFERENCES

(1) Boyd, P. G.; Lee, Y.; Smit, B. Computational Development of the Nanoporous Materials Genome. *Nat. Rev. Mater.* **2017**, *2* (8), 17037.

(2) Yaghi, O. M.; O'Keeffe, M.; Ockwig, N. W.; Chae, H. K.; Eddaoudi, M.; Kim, J. Reticular Synthesis and the Design of New Materials. *Nature* **2003**, *423* (6941), 705−714.

(3) Yaghi, O. M.; Kalmutzki, M. J.; Diercks, C. S. *Introduction to Reticular Chemistry*; Wiley: 2019. DOI: 10.1002/9783527821099.

(4) Jiang, H.; Alezi, D.; Eddaoudi, M. A Reticular Chemistry Guide for the Design of Periodic Solids. *Nat. Rev. Mater.* **2021**, *6* (6), 466−487.

(5) Mezenov, Y. A.; Krasilin, A. A.; Dzyuba, V. P.; Nominé, A.; Milichko, V. A. Metal-Organic Frameworks in Modern Physics: Highlights and Perspectives. *Adv. Sci.* **2019**, *6* (17), 1900506.

(6) Ko, M.; Mendecki, L.; Mirica, K. A. Conductive Two-Dimensional Metal-Organic Frameworks as Multifunctional Materials. *Chem. Commun.* **2018**, *54* (57), 7873−7891.

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

(8) Zhang, X.; Zhou, Y.; Cui, B.; Zhao, M.; Liu, F. Theoretical Discovery of a Superconducting Two-Dimensional Metal-Organic Framework. *Nano Lett.* **2017**, *17* (10), 6166−6170.

(9) Huang, X.; Zhang, S.; Liu, L.; Yu, L.; Chen, G.; Xu, W.; Zhu, D. Superconductivity in a Copper(II)-Based Coordination Polymer with Perfect Kagome Structure. *Angew. Chem., Int. Ed.* **2018**, *57* (1), 146−150.

(10) Jiang, W.; Ni, X.; Liu, F. Exotic Topological Bands and Quantum States in Metal-Organic and Covalent-Organic Frameworks. *Acc. Chem. Res.* **2021**, *54* (2), 416−426.

(11) Chen, Z.; Li, P.; Anderson, R.; Wang, X.; Zhang, X.; Robison, L.; Redfern, L. R.; Moribe, S.; Islamoglu, T.; Gómez-Gualdrón, D. A.; Yildirim, T.; Stoddart, J. F.; Farha, O. K. Balancing Volumetric and Gravimetric Uptake in Highly Porous Materials for Clean Energy. *Science* **2020**, *368* (6488), 297−303.

(12) Boyd, P. G.; Chidambaram, A.; García-Díez, E.; Ireland, C. P.; Daff, T. D.; Bounds, R.; Gładysiak, A.; Schouwink, P.; Moosavi, S. M.; Maroto-Valer, M. M.; Reimer, J. A.; Navarro, J. A. R.; Woo, T. K.; Garcia, S.; Stylianou, K. C.; Smit, B. Data-Driven Design of Metal-Organic Frameworks for Wet Flue Gas CO2 Capture. *Nature* **2019**, *576* (7786), 253−256.

(13) Suh, M. P.; Park, H. J.; Prasad, T. K.; Lim, D.-W. Hydrogen Storage in Metal-Organic Frameworks. *Chem. Rev.* **2012**, *112* (2), 782−835.

(14) He, Y.; Zhou, W.; Qian, G.; Chen, B. Methane Storage in Metal-Organic Frameworks. *Chem. Soc. Rev.* **2014**, *43* (16), 5657−5678.

(15) Mason, J. A.; Veenstra, M.; Long, J. R. Evaluating Metal-Organic Frameworks for Natural Gas Storage. *Chem. Sci.* **2014**, *5* (1), 32−51.

(16) Sumida, K.; Rogow, D. L.; Mason, J. A.; McDonald, T. M.; Bloch, E. D.; Herm, Z. R.; Bae, T.-H.; Long, J. R. Carbon Dioxide Capture in Metal-Organic Frameworks. *Chem. Rev.* **2012**, *112* (2), 724−781.

(17) Yu, J.; Xie, L.-H.; Li, J.-R.; Ma, Y.; Seminario, J. M.; Balbuena, P. B. CO2 Capture and Separations Using MOFs: Computational and Experimental Studies. *Chem. Rev.* **2017**, *117* (14), 9674−9754.

(18) Banerjee, D.; Cairns, A. J.; Liu, J.; Motkuri, R. K.; Nune, S. K.; Fernandez, C. A.; Krishna, R.; Strachan, D. M.; Thallapally, P. K. Potential of Metal-Organic Frameworks for Separation of Xenon and Krypton. *Acc. Chem. Res.* **2015**, *48* (2), 211−219.

(19) Banerjee, D.; Simon, C. M.; Elsaidi, S. K.; Haranczyk, M.; Thallapally, P. K. Xenon Gas Separation and Storage Using Metal-Organic Frameworks. *Chem.* **2018**, *4* (3), 466−494.

(20) Ongari, D.; Talirz, L.; Smit, B. Too Many Materials and Too Many Applications: An Experimental Problem Waiting for a Computational Solution. *ACS Cent. Sci.* **2020**, *6* (11), 1890−1900.

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

(22) Moosavi, S. M.; Nandy, A.; Jablonka, K. M.; Ongari, D.; Janet, J. P.; Boyd, P. G.; Lee, Y.; Smit, B.; Kulik, H. J. Understanding the Diversity of the Metal-Organic Framework Ecosystem. *Nat. Commun.* **2020**, *11* (1), 4068.

(23) Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. Large-Scale Screening of Hypothetical Metal-Organic Frameworks. *Nat. Chem.* **2012**, *4* (2), 83.

(24) Aghaji, M. Z.; Fernandez, M.; Boyd, P. G.; Daff, T. D.; Woo, T. K. Quantitative Structure-Property Relationship Models for Recognizing Metal Organic Frameworks (MOFs) with High CO2 Working Capacity and CO2/CH4 Selectivity for Methane Purification. *Eur. J. Inorg. Chem.* **2016**, *2016* (27), 4505−4511.

(25) Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B. Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chem. Rev.* **2020**, *120* (16), 8066−8129.

(26) Sturluson, A.; Huynh, M. T.; Kaija, A. R.; Laird, C.; Yoon, S.; Hou, F.; Feng, Z.; Wilmer, C. E.; Colón, Y. J.; Chung, Y. G.; Siderius, D. W.; Simon, C. M. The Role of Molecular Modelling and Simulation in the Discovery and Deployment of Metal-Organic Frameworks for Gas Storage and Separation. *Mol. Simul.* **2019**, *45* (14−15), 1082−1121.

(27) 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.* **2021**, *61* (5), 2131−2146.

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

(29) Odoh, S. O.; Cramer, C. J.; Truhlar, D. G.; Gagliardi, L. Quantum-Chemical Characterization of the Properties and Reactivities of Metal-Organic Frameworks. *Chem. Rev.* **2015**, *115* (12), 6051−6111.

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

(31) Grajciar, L.; Bludsky, O.; Nachtigall, P. Water Adsorption on Coordinatively Unsaturated Sites in CuBTC MOF. *J. Phys. Chem. Lett.* **2010**, *1* (23), 3354−3359.

(32) Dzubak, A. L.; Lin, L.-C.; Kim, J.; Swisher, J. A.; Poloni, R.; Maximoff, S. N.; Smit, B.; Gagliardi, L. Ab Initio Carbon Capture in Open-Site Metal-Organic Frameworks. *Nat. Chem.* **2012**, *4* (10), 810−816.

(33) Yu, D.; Yazaydin, A. O.; Lane, J. R.; Dietzel, P. D. C.; Snurr, R. Q. A Combined Experimental and Quantum Chemical Study of CO2 Adsorption in the Metal-Organic Framework CPO-27 with Different Metals. *Chem. Sci.* **2013**, *4* (9), 3544−3556.

(34) Howe, J. D.; Liu, Y.; Flores, L.; Dixon, D. A.; Sholl, D. S. Acid Gas Adsorption on Metal-Organic Framework Nanosheets as a Model of an "All-Surface" Material. *J. Chem. Theory Comput.* **2017**, *13* (3), 1341−1350.

(35) Barnes, A. L.; Bykov, D.; Lyakh, D. I.; Straatsma, T. P. Multilayer Divide-Expand-Consolidate Coupled-Cluster Method: Demonstrative Calculations of the Adsorption Energy of Carbon Dioxide in the Mg-MOF-74 Metal-Organic Framework. *J. Phys. Chem. A* **2019**, *123* (40), 8734−8743.

(36) Grimme, S. Accurate Description of van Der Waals Complexes by Density Functional Theory Including Empirical Corrections. *J. Comput. Chem.* **2004**, *25* (12), 1463−1473.

(37) Grimme, S. Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comput. Chem.* **2006**, *27* (15), 1787−1799.

https://doi.org/10.1021/acs.jcim.1c01124
J. Chem. Inf. Model. 2021, 61, 5774−5784

(38) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104.
(39) Caldeweyher, E.; Bannwarth, C.; Grimme, S. Extension of the D3 Dispersion Coefficient Model. *J. Chem. Phys.* **2017**, *147* (3), 034112.
(40) Berland, K.; Cooper, V. R.; Lee, K.; Schröder, E.; Thonhauser, T.; Hyldgaard, P.; Lundqvist, B. I. Van Der Waals Forces in Density Functional Theory: A Review of the VdW-DF Method. *Rep. Prog. Phys.* **2015**, *78* (6), 066501.
(41) Larsen, A. H.; Kuisma, M.; Löfgren, J.; Pouillon, Y.; Erhart, P.; Hyldgaard, P. Libvdwxc: A Library for Exchange-Correlation Functionals in the VdW-DF Family. *Modell. Simul. Mater. Sci. Eng.* **2017**, *25* (6), 065004.
(42) Rana, M. K.; Koh, H. S.; Hwang, J.; Siegel, D. J. Comparing van Der Waals Density Functionals for CO2 Adsorption in Metal Organic Frameworks. *J. Phys. Chem. C* **2012**, *116* (32), 16957−16968.
(43) Poloni, R.; Smit, B.; Neaton, J. B. Ligand-Assisted Enhancement of CO2 Capture in Metal-Organic Frameworks. *J. Am. Chem. Soc.* **2012**, *134* (15), 6714−6719.
(44) Hou, X.-J.; He, P.; Li, H.; Wang, X. Understanding the Adsorption Mechanism of C2H2, CO2, and CH4 in Isostructural Metal-Organic Frameworks with Coordinatively Unsaturated Metal Sites. *J. Phys. Chem. C* **2013**, *117* (6), 2824−2834.
(45) Queen, W. L.; Hudson, M. R.; Bloch, E. D.; Mason, J. A.; Gonzalez, M. I.; Lee, J. S.; Gygi, D.; Howe, J. D.; Lee, K.; Darwish, T. A.; James, M.; Peterson, V. K.; Teat, S. J.; Smit, B.; Neaton, J. B.; Long, J. R.; Brown, C. M. Comprehensive Study of Carbon Dioxide Adsorption in the Metal-Organic Frameworks M2(Dobdc) (M= Mg, Mn, Fe, Co, Ni, Cu, Zn). *Chem. Sci.* **2014**, *5* (12), 4569−4581.
(46) Poloni, R.; Lee, K.; Berger, R. F.; Smit, B.; Neaton, J. B. Understanding Trends in CO2 Adsorption in Metal-Organic Frameworks with Open-Metal Sites. *J. Phys. Chem. Lett.* **2014**, *5* (5), 861−865.
(47) Tan, K.; Zuluaga, S.; Gong, Q.; Gao, Y.; Nijem, N.; Li, J.; Thonhauser, T.; Chabal, Y. J. Competitive Coadsorption of CO2 with H2O, NH3, SO2, NO, NO2, N2, O2, and CH4 in M-MOF-74 (M= Mg, Co, Ni): The Role of Hydrogen Bonding. *Chem. Mater.* **2015**, *27* (6), 2203−2217.
(48) Lee, K.; Howe, J. D.; Lin, L.-C.; Smit, B.; Neaton, J. B. Small-Molecule Adsorption in Open-Site Metal-Organic Frameworks: A Systematic Density Functional Theory Study for Rational Design. *Chem. Mater.* **2015**, *27* (3), 668−678.
(49) 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.* **2016**, *144* (17), 174104.
(50) Vlaisavljevich, B.; Huck, J.; Hulvey, Z.; Lee, K.; Mason, J. A.; Neaton, J. B.; Long, J. R.; Brown, C. M.; Alfè, D.; Michaelides, A.; Smit, B. Performance of van Der Waals Corrected Functionals for Guest Adsorption in the M2(Dobdc) Metal-Organic Frameworks. *J. Phys. Chem. A* **2017**, *121* (21), 4139−4151.
(51) Asgari, M.; Jawahery, S.; Bloch, E. D.; Hudson, M. R.; Flacau, R.; Vlaisavljevich, B.; Long, J. R.; Brown, C. M.; Queen, W. L. An Experimental and Computational Study of CO2 Adsorption in the Sodalite-Type M-BTT (M= Cr, Mn, Fe, Cu) Metal-Organic Frameworks Featuring Open Metal Sites. *Chem. Sci.* **2018**, *9* (20), 4579−4588.
(52) You, W.; Liu, Y.; Howe, J. D.; Tang, D.; Sholl, D. S. Tuning Binding Tendencies of Small Molecules in Metal-Organic Frameworks with Open Metal Sites by Metal Substitution and Linker Functionalization. *J. Phys. Chem. C* **2018**, *122* (48), 27486−27494.
(53) Vazhappilly, T.; Ghanty, T. K.; Jagatap, B. N. Computational Modeling of Adsorption of Xe and Kr in M-MOF-74 Metal Organic Frame Works with Different Metal Atoms. *J. Phys. Chem. C* **2016**, *120* (20), 10968−10974.
(54) Kancharlapalli, S.; Natarajan, S.; Ghanty, T. K. Confinement-Directed Adsorption of Noble Gases (Xe/Kr) in MFM-300 (M)-Based Metal-Organic Framework Materials. *J. Phys. Chem. C* **2019**, *123* (45), 27531−27541.
(55) Chung, Y. G.; Camp, J.; Haranczyk, M.; Sikora, B. J.; Bury, W.; Krungleviciute, V.; Yildirim, T.; Farha, O. K.; Sholl, D. S.; Snurr, R. Q. Computation-Ready, Experimental Metal-Organic Frameworks: A Tool to Enable High-Throughput Screening of Nanoporous Crystals. *Chem. Mater.* **2014**, *26* (21), 6185−6192.
(56) Nazarian, D.; Camp, J. S.; Sholl, D. S. A Comprehensive Set of High-Quality Point Charges for Simulations of Metal-Organic Frameworks. *Chem. Mater.* **2016**, *28* (3), 785−793.
(57) Rosen, A. S.; Iyer, S. M.; Ray, D.; Yao, Z.; Aspuru-Guzik, A.; Gagliardi, L.; Notestein, J. M.; Snurr, R. Q. Machine Learning the Quantum-Chemical Properties of Metal-Organic Frameworks for Accelerated Materials Discovery. *Matter* **2021**, *4* (5), 1578−1597.
(58) Simon, C. M.; Kim, J.; Gomez-Gualdron, D. A.; Camp, J. S.; Chung, Y. G.; Martin, R. L.; Mercado, R.; Deem, M. W.; Gunter, D.; Haranczyk, M.; Sholl, D. S.; Snurr, R. Q.; Smit, B. The Materials Genome in Action: Identifying the Performance Limits for Methane Storage. *Energy Environ. Sci.* **2015**, *8* (4), 1190−1199.
(59) Simon, C. M.; Mercado, R.; Schnell, S. K.; Smit, B.; Haranczyk, M. What Are the Best Materials to Separate a Xenon/Krypton Mixture? *Chem. Mater.* **2015**, *27* (12), 4459−4475.
(60) Banerjee, D.; Simon, C. M.; Plonka, A. M.; Motkuri, R. K.; Liu, J.; Chen, X.; Smit, B.; Parise, J. B.; Haranczyk, M.; Thallapally, P. K. Metal-Organic Framework with Optimally Selective Xenon Adsorption and Separation. *Nat. Commun.* **2016**, *7* (1), 11831.
(61) Thornton, A. W.; Simon, C. M.; Kim, J.; Kwon, O.; Deeg, K. S.; Konstas, K.; Pas, S. J.; Hill, M. R.; Winkler, D. A.; Haranczyk, M.; Smit, B. Materials Genome in Action: Identifying the Performance Limits of Physical Hydrogen Storage. *Chem. Mater.* **2017**, *29* (7), 2844−2854.
(62) Ahmed, A.; Seth, S.; Purewal, J.; Wong-Foy, A. G.; Veenstra, M.; Matzger, A. J.; Siegel, D. J. Exceptional Hydrogen Storage Achieved by Screening Nearly Half a Million Metal-Organic Frameworks. *Nat. Commun.* **2019**, *10* (1), 1568.
(63) Dubbeldam, D.; Walton, K. S.; Vlugt, T. J. H.; Calero, S. Design, Parameterization, and Implementation of Atomic Force Fields for Adsorption in Nanoporous Materials. *Adv. Theory Simulations* **2019**, *2* (11), 1900135.
(64) Rappé, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A., III; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114* (25), 10024−10035.
(65) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: A Generic Force Field for Molecular Simulations. *J. Phys. Chem.* **1990**, *94* (26), 8897−8909.
(66) Pham, T.; Forrest, K. A.; McLaughlin, K.; Eckert, J.; Space, B. Capturing the H2-Metal Interaction in Mg-MOF-74 Using Classical Polarization. *J. Phys. Chem. C* **2014**, *118* (39), 22683−22690.
(67) Pham, T.; Forrest, K. A.; Franz, D. M.; Guo, Z.; Chen, B.; Space, B. Predictive Models of Gas Sorption in a Metal-Organic Framework with Open-Metal Sites and Small Pore Sizes. *Phys. Chem. Chem. Phys.* **2017**, *19* (28), 18587−18602.
(68) Becker, T. M.; Luna-Triguero, A.; Vicent-Luna, J. M.; Lin, L.-C.; Dubbeldam, D.; Calero, S.; Vlugt, T. J. H. Potential of Polarizable Force Fields for Predicting the Separation Performance of Small Hydrocarbons in M-MOF-74. *Phys. Chem. Chem. Phys.* **2018**, *20* (45), 28848−28859.
(69) Addicoat, M. A.; Vankova, N.; Akter, I. F.; Heine, T. Extension of the Universal Force Field to Metal-Organic Frameworks. *J. Chem. Theory Comput.* **2014**, *10* (2), 880−891.
(70) Coupry, D. E.; Addicoat, M. A.; Heine, T. Extension of the Universal Force Field for Metal-Organic Frameworks. *J. Chem. Theory Comput.* **2016**, *12* (10), 5215−5225.
(71) Bureekaew, S.; Amirjalayer, S.; Tafipolsky, M.; Spickermann, C.; Roy, T. K.; Schmid, R. MOF-FF − A Flexible First-Principles Derived Force Field for Metal-Organic Frameworks. *Phys. Status Solidi B* **2013**, *250* (6), 1128−1141.

https://doi.org/10.1021/acs.jcim.1c01124
J. Chem. Inf. Model. 2021, 61, 5774−5784

(72) Bristow, J. K.; Tiana, D.; Walsh, A. Transferable Force Field for Metal-Organic Frameworks from First-Principles: BTW-FF. *J. Chem. Theory Comput.* **2014**, *10* (10), 4644−4652.
(73) Vanduyfhuys, L.; Vandenbrande, S.; Verstraelen, T.; Schmid, R.; Waroquier, M.; Van Speybroeck, V. QuickFF: A Program for a Quick and Easy Derivation of Force Fields for Metal-Organic Frameworks from ab Initio Input. *J. Comput. Chem.* **2015**, *36* (13), 1015−1027.
(74) Haldoupis, E.; Borycz, J.; Shi, H.; Vogiatzis, K. D.; Bai, P.; Queen, W. L.; Gagliardi, L.; Siepmann, J. I. Ab Initio Derived Force Fields for Predicting CO2 Adsorption and Accessibility of Metal Sites in the Metal-Organic Frameworks M-MOF-74 (M= Mn, Co, Ni, Cu). *J. Phys. Chem. C* **2015**, *119* (28), 16058−16071.
(75) Jawahery, S.; Rampal, N.; Moosavi, S. M.; Witman, M.; Smit, B. Ab Initio Flexible Force Field for Metal-Organic Frameworks Using Dummy Model Coordination Bonds. *J. Chem. Theory Comput.* **2019**, *15* (6), 3666−3677.
(76) Pham, T.; Forrest, K. A.; Banerjee, R.; Orcajo, G.; Eckert, J.; Space, B. Understanding the H2 Sorption Trends in the M-MOF-74 Series (M= Mg, Ni, Co, Zn). *J. Phys. Chem. C* **2015**, *119* (2), 1078−1090.
(77) Becker, T. M.; Heinen, J.; Dubbeldam, D.; Lin, L.-C.; Vlugt, T. J. H. Polarizable Force Fields for CO2 and CH4 Adsorption in M-MOF-74. *J. Phys. Chem. C* **2017**, *121* (8), 4659−4673.
(78) Becker, T. M.; Lin, L.-C.; Dubbeldam, D.; Vlugt, T. J. H. Polarizable Force Field for CO2 in M-MOF-74 Derived from Quantum Mechanics. *J. Phys. Chem. C* **2018**, *122* (42), 24488−24498.
(79) Chen, T.; Manz, T. A. A Collection of Forcefield Precursors for Metal-Organic Frameworks. *RSC Adv.* **2019**, *9* (63), 36492−36507.
(80) Manz, T. A.; Limas, N. G. Introducing DDEC6 Atomic Population Analysis: Part 1. Charge Partitioning Theory and Methodology. *RSC Adv.* **2016**, *6* (53), 47771−47801.
(81) Limas, N. G.; Manz, T. A. Introducing DDEC6 Atomic Population Analysis: Part 2. Computed Results for a Wide Range of Periodic and Nonperiodic Materials. *RSC Adv.* **2016**, *6* (51), 45727−45747.
(82) Manz, T. A. Introducing DDEC6 Atomic Population Analysis: Part 3. Comprehensive Method to Compute Bond Orders. *RSC Adv.* **2017**, *7* (72), 45552−45581.
(83) Limas, N. G.; Manz, T. A. Introducing DDEC6 Atomic Population Analysis: Part 4. Efficient Parallel Computation of Net Atomic Charges, Atomic Spin Moments, Bond Orders, and More. *RSC Adv.* **2018**, *8* (5), 2678−2707.
(84) Manz, T. A.; Chen, T.; Cole, D. J.; Limas, N. G.; Fiszbein, B. New Scaling Relations to Compute Atom-in-Material Polarizabilities and Dispersion Coefficients: Part 1. Theory and Accuracy. *RSC Adv.* **2019**, *9* (34), 19297−19324.
(85) Jones, A. P.; Crain, J.; Sokhan, V. P.; Whitfield, T. W.; Martyna, G. J. Quantum Drude Oscillator Model of Atoms and Molecules: Many-Body Polarization and Dispersion Interactions for Atomistic Simulation. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87* (14), 144103.
(86) Sadhukhan, M.; Manby, F. R. Quantum Mechanics of Drude Oscillators with Full Coulomb Interaction. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2016**, *94* (11), 115106.
(87) Cipcigan, F. S.; Crain, J.; Sokhan, V. P.; Martyna, G. J. Electronic Coarse Graining: Predictive Atomistic Modeling of Condensed Matter. *Rev. Mod. Phys.* **2019**, *91* (2), 025003.
(88) Van Vleet, M. J.; Misquitta, A. J.; Stone, A. J.; Schmidt, J. R. Beyond Born−Mayer: Improved Models for Short-Range Repulsion in Ab Initio Force Fields. *J. Chem. Theory Comput.* **2016**, *12* (8), 3851−3870.
(89) Korolev, V. V.; Mitrofanov, A.; Marchenko, E. I.; Eremin, N. N.; Tkachenko, V.; Kalmykov, S. N. Transferable and Extensible Machine Learning-Derived Atomic Charges for Modeling Hybrid Nanoporous Materials. *Chem. Mater.* **2020**, *32* (18), 7822−7831.
(90) Zou, C.; Penley, D. R.; Cho, E. H.; Lin, L.-C. Efficient and Accurate Charge Assignments via a Multilayer Connectivity-Based Atom Contribution (m-CBAC) Approach. *J. Phys. Chem. C* **2020**, *124* (21), 11428−11437.
(91) Raza, A.; Sturluson, A.; Simon, C. M.; Fern, X. Message Passing Neural Networks for Partial Charge Assignment to Metal-Organic Frameworks. *J. Phys. Chem. C* **2020**, *124* (35), 19070−19082.
(92) 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.* **2021**, *17* (5), 3052−3064.
(93) Ongari, D.; Boyd, P. G.; Kadioglu, O.; Mace, A. K.; Keskin, S.; Smit, B. Evaluating Charge Equilibration Methods to Generate Electrostatic Fields in Nanoporous Materials. *J. Chem. Theory Comput.* **2019**, *15* (1), 382−401.
(94) Manz, T. A.; Chen, T. New Scaling Relations to Compute Atom-in-Material Polarizabilities and Dispersion Coefficients: Part 2. Linear-Scaling Computational Algorithms and Parallelization. *RSC Adv.* **2019**, *9* (57), 33310−33336.
(95) Feynman, R. P. Forces in Molecules. *Phys. Rev.* **1939**, *56* (4), 340.
(96) Cordero, B.; Gómez, V.; Platero-Prats, A. E.; Revés, M.; Echeverría, J.; Cremades, E.; Barragán, F.; Alvarez, S. Covalent Radii Revisited. *Dalt. Trans.* **2008**, *21* (No), 2832−2838.
(97) Tantardini, C.; Oganov, A. R. Thermochemical Electronegativities of the Elements. *Nat. Commun.* **2021**, *12* (1), 2087.
(98) Botu, V.; Ramprasad, R. Adaptive Machine Learning Framework to Accelerate Ab Initio Molecular Dynamics. *Int. J. Quantum Chem.* **2015**, *115* (16), 1074−1083.
(99) Botu, V.; Ramprasad, R. Learning Scheme to Predict Atomic Forces and Accelerate Materials Simulations. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2015**, *92* (9), 094306.
(100) Okabe, A.; Boots, B.; Sugihara, K.; Chiu, S. N. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, 2nd ed.; Wiley Online Library: 2000.
(101) Peng, H. L.; Li, M. Z.; Wang, W. H. Structural Signature of Plastic Deformation in Metallic Glasses. *Phys. Rev. Lett.* **2011**, *106* (13), 135503.
(102) Wang, Q.; Jain, A. A Transferable Machine-Learning Framework Linking Interstice Distribution and Plastic Heterogeneity in Metallic Glasses. *Nat. Commun.* **2019**, *10* (1), 5537.
(103) Zimmermann, N. E. R.; Horton, M. K.; Jain, A.; Haranczyk, M. Assessing Local Structure Motifs Using Order Parameters for Motif Recognition, Interstitial Identification, and Diffusion Path Characterization. *Front. Mater.* **2017**, *4*, 34.
(104) Zimmermann, N. E. R.; Jain, A. Local Structure Order Parameters and Site Fingerprints for Quantification of Coordination Environment and Crystal Structure Similarity. *RSC Adv.* **2020**, *10* (10), 6063−6081.
(105) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (Pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Comput. Mater. Sci.* **2013**, *68*, 314−319.
(106) Larsen, A. H.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; Hermes, E. D.; Jennings, P. C.; Jensen, P. B.; Kermode, J.; Kitchin, J. R.; Kolsbjerg, E. L.; Kubal, J.; Kaasbjerg, K.; Lysgaard, S.; Maronsson, J. B.; Maxson, T.; Olsen, T.; Pastewka, L.; Peterson, A.; Rostgaard, C.; Schiøtz, J.; Schütt, O.; Strange, M.; Thygesen, K. S.; Vegge, T.; Vilhelmsen, L.; Walter, M.; Zeng, Z.; Jacobsen, K. W. The Atomic Simulation Environment—a Python Library for Working with Atoms. *J. Phys.: Condens. Matter* **2017**, *29* (27), 273002.
(107) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E. R.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; Chard, K.; Asta, M.; Persson, K. A.; Snyder, G. J.; Foster, I.; Jain, A. Matminer: An Open Source Toolkit for Materials Data Mining. *Comput. Mater. Sci.* **2018**, *152*, 60−69.
(108) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International*

https://doi.org/10.1021/acs.jcim.1c01124
J. Chem. Inf. Model. 2021, 61, 5774−5784

Conference on Knowledge Discovery and Data Mining, KDD '16; ACM: New York, NY, 2016; pp 785−794.
(109) Bergstra, J. S.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for Hyper-Parameter Optimization. *Advances in Neural Information Processing Systems 24 (NIPS 2011)*; Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K. Q., Eds.; Neural Information Processing Systems Foundation, Inc.: 2011; pp 2546−2554.
(110) Bergstra, J.; Yamins, D.; Cox, D. D. Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. *Proceedings of the 12th Python in science conference* **2013**, 13−20.
(111) Bergstra, J.; Komer, B.; Eliasmith, C.; Yamins, D.; Cox, D. D. Hyperopt: A Python Library for Model Selection and Hyperparameter Optimization. *Comput. Sci. Discovery* **2015**, *8* (1), 014008.
(112) Manz, T. A.; Sholl, D. S. Improved Atoms-in-Molecule Charge Partitioning Functional for Simultaneously Reproducing the Electrostatic Potential and Chemical States in Periodic and Nonperiodic Materials. *J. Chem. Theory Comput.* **2012**, *8* (8), 2844−2867.
(113) Zhang, Y.; Ling, C. A Strategy to Apply Machine Learning to Small Datasets in Materials Science. *npj Comput. Mater.* **2018**, *4*, 25.
(114) Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.-R.; Burke, K. Quantum Chemical Accuracy from Density Functional Approximations via Machine Learning. *Nat. Commun.* **2020**, *11* (1), 5223.
(115) Kalita, B.; Li, L.; McCarty, R. J.; Burke, K. Learning to Approximate Density Functionals. *Acc. Chem. Res.* **2021**, *54* (4), 818−826.
(116) Hansen, N. K.; Coppens, P. Testing Aspherical Atom Refinements on Small-Molecule Data Sets. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1978**, *34* (6), 909−921.
(117) Coppens, P.; Guru Row, T. N.; Leung, P.; Stevens, E. D.; Becker, P. J. t; Yang, Y. W. Net Atomic Charges and Molecular Dipole Moments from Spherical-Atom X-Ray Refinements, and the Relation between Atomic Charge and Shape. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1979**, *35* (1), 63−72.
(118) Zuo, J. M. Measurements of Electron Densities in Solids: A Real-Space View of Electronic Structure and Bonding in Inorganic Crystals. *Rep. Prog. Phys.* **2004**, *67* (11), 2053.
(119) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015**, *114* (10), 105503.
(120) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *npj Comput. Mater.* **2017**, *3* (1), 54.
(121) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559* (7715), 547−555.
(122) Breiman, L.; Friedman, J.; Stone, C. J.; Olshen, R. A. *Classification and Regression Trees*; CRC Press: 1984.
(123) Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinf.* **2007**, *8* (1), 25.
(124) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2* (1), 56−67.
(125) Shapley, L. S. A Value for N-Person Games. *Contrib. to Theory Games* **1953**, *2* (28), 307−317.
(126) Schmidt, J.; Shi, J.; Borlido, P.; Chen, L.; Botti, S.; Marques, M. A. L. Predicting the Thermodynamic Stability of Solids Combining Density Functional Theory and Machine Learning. *Chem. Mater.* **2017**, *29* (12), 5090−5103.
(127) Caruana, R. Multitask Learning. *Mach. Learn.* **1997**, *28* (1), 41−75.
(128) Xu, Y.; Ma, J.; Liaw, A.; Sheridan, R. P.; Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2017**, *57* (10), 2490−2504.

https://doi.org/10.1021/acs.jcim.1c01124
J. Chem. Inf. Model. 2021, 61, 5774−5784