Transferable and extensible machine learning derived atomic charges for modeling metal-organic frameworks

Vadim Korolev1,2, Artem Mitrofanov1,2, Ekaterina Marchenko3, Nickolay Eremin3, Valery Tkachenko1, Stepan Kalmykov2

1Science Data Software, LLC, 14914 Bradwill Court, Rockville, Maryland 20850, United States
2Lomonosov Moscow State University, Department of Chemistry, Leninskie gory, 1 bld. 3, Moscow 119991, Russia
3Lomonosov Moscow State University, Department of Geology, Leninskie gory, 1 bld. 1, Moscow 119991, Russia

ABSTRACT

In most cases, calculations of properties of metal-organic frameworks (MOFs) based on classical force fields (FFs) are the most suitable in terms of the ratio between accuracy and computational cost, especially in efforts to screen a large number of structures. Such calculations require an initial partial charge assignment to describe the Coulomb contribution. In this study, we would like to present a machine-learning algorithm for MOF partial charge prediction and its verification on experimental data using the FF approach. Proposed ML method offers the accuracy of reference DFT calculations at a fraction of the computational cost with linear scalability.

Keywords metal-organic frameworks · partial charges · machine learning · force field

1 Introduction

Metal-organic frameworks (MOFs) form a relatively new class of crystalline porous materials, which should be considered as a part of “nanoporous materials genome”[1, 2]. Their internal pores comparable to the size of small molecules cause confinement effects that strongly affect adsorbates. Resultant properties may be useful in multiple applications, such as catalysis[3], gas storage/separation[4, 5], and especially — in clean energy-related fields[6]. Due to the enormous structural diversity, the number of synthesized MOFs has now reached several thousand. Computationally developed MOF databases should contain up to several hundreds of thousands of structures[7]; thus, providing synthesis, characterization, and testing all the promising candidates for a desired absorbance-related application is impossible from a practical perspective. Not surprising, in silico design has become the main approach to provide large-scale materials screening studies in the field[1].

Electronic structure calculations, mainly based on density functional theory (DFT), provide an appropriate accuracy in matching experimentally measured data such as adsorbate interaction energies[8]. However, as in the case of experiments, large-scale materials screening studies with DFT calculations are disappointingly time-consuming, and it can be carried out only for a limited set of structures. In most cases, MOF properties calculations based on classical force fields (FFs) provide an optimal trade-off between accuracy and computational cost, especially in screening a large number of structures. A starting point and (usually) a "bottleneck" of all FF calculations with MOFs is a partial charge assignment, required to describe the Coulomb contribution to interatomic potential. There is no unique solution to assign point charges to atoms in MOFs, and numerous methods were implemented for this purpose.

Semiempirical methods such as the charge equilibration[9] (QEq), the extended charge equilibration[10, 11] (EQQeq) and the periodic charge equilibration[12] (PQeq) do not require direct electronic structure calculations, and they are strongly preferable under computational cost aspect. However, these methods enormously overestimate the values of point charges for some elements, which may eventually lead to unreliable computational results[13]. Methods of charge assignment based on partitioning the electron density (ChepG[14], density-derived electrostatic and chemical[15] (DDEC), repeating electrostatic potential extracted atomic[16] (REPEAT)) were previously applied only for relatively small sets (up to three thousand[17]) of structures due to their high computational cost.

Thus, highly parameterized ad hoc methods and ab initio calculations provide two main options in terms of the trade-off between accuracy and time/computational cost restraints[15, 16, 18]. Unfortunately, the substantial issue is common for both of them — every new structure requires distinct calculations and charge assignment. Even with advanced QEq parameters[13], known as MOF electrostatic-potential-optimized (MEPO) QEq parameters, highly parameterized methods do not allow to provide charge assignment for structures distinct from ones considered in the original study.

Xu and Zhong[19] proposed the calculation-free method, contrary to all ones mentioned above. This data-driven approach is based on the assumption that the atoms with the same bonding connectivity have identical charges in

*Email address: korolev@colloid.chem.msu.ru
different MOFs. Unfortunately, the connectivity-based atom contribution (CBAC) method has an inherent limitation — it can be applied only for atoms with the same types of local environments presented in the training set.

The similar challenges arise for molecular charge assignment. However, an alternative approach unrelated to “classical” computational techniques has been introduced recently. It has been shown that machine learning (ML) and particularly deep learning techniques can reconcile quantum mechanics accuracy and low computational cost for partial charge calculations[20, 21, 22]. Moreover, ML models can provide transferability and extensibility of charge prediction. In other words, it is possible to assign partial charges in systems differ significantly from systems that were used to train ML models. Extensibility usually relates to system size; transferability can be interpreted in a broader sense, and it usually involves chemical diversity.

Taking into account that the whole number of experimentally/computationally characterized MOFs is several orders of magnitude less than the corresponding number of molecules, transferable and extensible charge assignment for MOFs is a complex task at least because of data scarcity. To overcome this issue, we implement high-interpretable ML models on the most numerous of the available set of DFT-based density derived electrostatic and chemical (DDEC) charges. In order to confirm the viability of this data-driven approach, partial charges derived by ML are used to construct FFs for MOFs that differ significantly from ones used to build ML models.

2 Methods

2.1 Reference DFT simulations

Recent studies have shown that the availability of materials data strongly affects the predictive capability of ML models[23, 24, 25]. Unfortunately, materials datasets are typically far less numerous than molecular ones; the number of samples is in the hundreds in the most complex cases. The high computational cost for generating new data points usually motivates data scarcity; this is particularly true for partial charge assignment with high-precision ab initio calculations. Moreover, the diversity of data can restrict the predictive capability of ML models in unexplored domains; in other words, it can limit the ML model’s transferability.

The most reliable set of partial charges for MOFs using DFT calculations and the DDEC charge partitioning approach was presented by Nazarian et al[17]. Atomic point charge assignment was provided for 2932 experimentally synthesized MOF structures, most of them were collected from Computation-Ready Experimental MOF (CoRE MOF) database[26]. Solvent molecules, typically trapped in experimentally resolved MOFs, and atoms with partial occupancy/symmetry-related copies of atoms were removed, in order to prepare MOFs for molecular simulations. Furthermore, highly disordered structures were also excluded from consideration. Charge assignment was provided using plane-wave DFT calculations and the DDEC charge partitioning method. Optimization of the geometry of structures was not employed due to its negligible influence on resulting point charges. In particular, high-quality charges generated by DDEC method are designed to reproduce electrostatic interactions even outside van der Waals radius of involved atoms; it is especially crucial for adsorption-related simulations.

It should be noted that the total number of distinct materials in CoRE MOF database (3852) is significantly larger than the number of structures for which the electron density was successfully computed (2932). Most calculations that did not converge is due to the large size of a unit cell of corresponding structures and concomitant difficulties (virtual memory exceeding, issues with k-point grid density requirements, etc.). Also, due to computational difficulties, the presence of heavy elements is more prominent in the unlabeled part of the database. Thus, there are now DDEC charges for, e.g., Eu and Tb MOFs. Wherein lanthanide-containing MOFs are highly attractive due to their luminescent properties[27, 28]. Therefore, here we define the ability of modeling lanthanide-containing structures as an essential characteristic of a useful ML model. The mentioned subset was used to probe the extensibility of presented ML models trained on structures from the CoRE MOF database with the relatively small unit cells and lack of heavy-elements.

The total number of samples (lattice sites and corresponding partial atomic charges) in training dataset exceeds four hundred thousand, so it does not seem diverse or sparse. However, the training dataset is significantly diverse in a structural and chemical sense.

2.2 Descriptors

Besides reliable input data, the material’s representation strongly influences the performance of data-driven (machine learning) approach used to materials related tasks — fingerprints (so-called descriptors) serve for this purpose[29, 30, 31, 32]. The main requirement for them is to lose information on the considered material as little as possible. There is no universal method for choosing an appropriate set of descriptors. Numerous types of descriptors differing in levels of complexity have been proposed to represent materials under consideration. They all have in common, however,
some properties. Relevant descriptors should be simple to extract itself and have as low a dimensionality as possible. The choice of descriptors typically proceeds iteratively, and it is based primarily on the domain knowledge of experts. Taking into account the nature of the target property (partial atomic charge), the two types of descriptors are chosen for subsequent implementation/training ML models (the more detailed information on used descriptors is provided in Supporting Information):

- Intrinsic elemental properties of the corresponding site
- Structural descriptors of site characterized its local environment

Thus, using the locality approximation, we determine partial atomic charge $q_i$ exclusively as a function of chemical nature of the corresponding atom (through its physicochemical properties $p_e$) and its local environment $n$ (depends on neighbors atoms $a_j$ with a distance to considered atom less than the cut-off radius: $R_{ij} < R_c$):

$$q_i = f(p_e, \sum_{R_{ij} < R_c} (a_j))$$

This approximation enables one to predict atomic charges for different atoms independently of each other and makes the presented approach linearly scaled with the number of atoms in the unit cell, as opposed to the reference DFT method.

2.3 Model training

Only a few state-of-the-art ML algorithms provide interpretability. To establish relations between suggested descriptors and target property (partial charge) we use gradient boosting decision trees (GBDT) method. In addition to high accuracy, this algorithm provides the feature importance, i.e., it is possible to extract the contribution of individual features to model’s performance. The set of partial charges contains approximately 440 thousand points. Ten-fold cross validation is implemented to evaluate generalizability of models on external test set (ten percent of the initial dataset). A hyperparameter search is performed using Tree-of-Parzen-Estimators (TPE) algorithm implemented in Hyperopt library.

![Figure 1: Distribution of the considered CoRE MOFs by number of atoms in unit cell.](image)

3 Results and discussion

3.1 Partial charge assignment

In most cases, ML models in materials science are applied to predict physicochemical properties characterized compounds from “macroscopical” perspective. In contrast, microscopic features such as atomic energies, forces, and
partial charges are not the ultimate goal of the computational analysis; these quantities are necessary for building semi-empirical potential models. In particular, partial charges generated by the proposed data-driven method can be used for high-throughput screening of MOFs for adsorption-related applications.

To validate the proposed approach for partial charge assignment, force fields were also implemented for several structures from the CoRE MOFs database, for which the calculations of DDEC charges did not converge. Structures were selected according to the following criteria:

- Large unit cells. One of the reasons for the divergence of DFT calculations (the reference method for DDEC charge assignment) is that the maximal virtual memory is exceeded. Fig. 1 shows the distribution of structures from the CoRE MOFs database by the number of atoms in the unit cell. The largest structures with calculated charges contain no more than 584 atoms, while the CoRE MOFs database contains much larger structures (up to 3600 atoms per unit cell). It is these large structures that can be used to test the concept of extensibility, that is, the general possibility of applying (generalizing) the predictive model to larger structures as compared to those used in the training dataset. The extensibility of data-driven models for charge assignment was previously confirmed for the case of molecular systems[20,22]. The use of structures with a relatively small unit cell as training samples is especially critical for the case of MOFs (three-dimensional periodic structures containing up to several thousand atoms) since the reference method (DFT) scales as \( \sim N^3 \). At the same time, our proposed method is based on using only the characteristics of the local environment of the atoms, and, accordingly, the training of the model and the calculation of charges scale as \( \sim N^1 \). The approximate nature of the dependence is because the calculation of some descriptors requires Voronoi tessellation, and this transformation is generally non-linear with the size of the system.

- The presence of rare earth elements. It makes it harder to calculate partial charges. In the case of the equilibration charge methods, it is necessary to have reliable experimental values of the ionization potentials, which are just not available for rare-earth elements (in this case, values calculated using CCSD(T) are usually used). On the other hand, the presence of 4-f highly localized electrons makes DFT calculations challenging. In addition, some rare earth elements are not present in the set DDEC charges (Fig. 2). The use of a predictive model for an extended set of elements is intended to confirm the transferability of our approach.

In order to make a direct comparison of the proposed approach with existing methods for partial charge assignment, we also calculated charges for the above structures using the extended charge equilibration method — a modification of the original algorithm specially designed for calculating the charges on metal nodes of MOFs. The required atomic parameters — the first \( \left( \frac{\partial E}{\partial Q} \right)_{A0} \) and second \( \left( \frac{\partial^2 E}{\partial Q^2} \right)_{A0} \) derivatives (the electronegativity \( \chi_{0}A \) and idempotential \( J_{0}AA \), respectively) were taken from the universal force field implementation in the Open Babel software[33]. The validity of this method was confirmed by screening study for \( \text{CO}_2 \) and \( \text{N}_2 \) adsorption[11,54].

Figure 2: The frequency of different elements in the set of considered CoRE MOFs.
Figure 3: (left) Predicted ML derived partial charges vs. reference DDEC charges. (right) Normalized histogram of the deviation of predicted ML derived partial charges from reference DDEC charges.

Fig. 3 shows the values of the predicted partial charges vs. the corresponding DDEC charges. Mean absolute error (MAE) and root mean squared error (RMSE) for the training dataset (10% percent of the entire set of DDEC charges) are 0.0096 and 0.0176 e, respectively. As mentioned earlier, partial charges are not the ultimate goal of predictions; they are not physically observable quantities, so there is no universal quantitative metric for assessing the quality of models for their prediction. It should be noted that the value of 0.01e can serve as a quantitative assessment of similar models for molecules [20, 35]. As was shown, such a deviation from the value of the reference method (DFT simulations) provides reliable results for predicting practically important physicochemical properties. It is worth noting that in our case, the training dataset is an order of magnitude smaller than in the study mentioned above, and additional difficulties may be due to the periodicity and large unit cell size of metal-organic frameworks.

Figure 4: Cumulative fraction of samples as a function of MAE (left) for ten most numerous elements and (right) rare earth elements.

Figure 4: Cumulative fraction of samples as a function of MAE (left) for ten most numerous elements and (right) rare earth elements.
Figure 5: Mean absolute errors of partial charge predictions as a function (left) of number of atoms and (right) number of atoms normalized to number of local environment types. Each point corresponds to distinct element.

For a more detailed consideration of the performance of our models, we examine its effectiveness in predicting charges for the case of distinct chemical elements. The corresponding cumulative curves are shown in Fig. 4. The left panel contains information about the ten most numerous elements; the right panel contains curves for REEs. This choice is because these elements significantly complicate DFT calculations; selected MOFs series also contain atoms of REEs. The most general trend that can be identified is the prediction efficiency is higher for the most numerous elements. However, when considering a similar relationship for the whole periodic system of elements (Fig. 5), it is not observed. As it may seem, it goes against the general feature of the data-dependent algorithms “more data is better” which is in particular valid for materials science related applications[25]. It should be noted that some atoms in the dataset have a similar local environment, so the actual size of the dataset is less than the nominal one. However, even normalization to the number of different types of the local environment also did not lead to a pronounced relationship between the size of the dataset for individual chemical elements and the corresponding MAEs. The most obvious explanation is that the model’s performance is only partially determined by the nature of the chemical element and its nearest coordination sphere (the criterion by which structural types were distinguished). Thus, for Zn atoms coordinated to six O atoms the charge value is in the range 1.18—1.85[17], that is, other features influence the partial charge value.

In order to get insights on which of the used features make the most significant contribution to the predictions, we use the SHAP approach implemented for tree ensemble methods[36]. The SHAP summary plots for ten most-valuable features (with the highest global impact $\sum_{i=1}^{N} |\phi_i^j|$ , where $\phi_i^j$ are SHAP values) are presented in Fig. 6. Among these features, one should first highlight the physicochemical elemental properties. Not surprisingly, electronegativity is the most important feature affecting the value of partial charges. Among other elemental properties, one should also highlight the covalent radius, large values of which contribute to a positive charge (e.g., RE elements), Mendeleev number (hydrogen stands out, for which a low value of this value leads to an increase in positive charge).

Nevertheless, 4 out of 10 most valuable features characterize the local environment of the atom from different perspectives:

- The local order parameter is a structural parameter with Voronoi-tessellation based neighbor finding. This particular fingerprint represents how consistent considered local environment is with a coordination number equals to two.
- AGNIFingerprint is fingerprint based on integrating the distances product of the radial distribution function with a Gaussian window function. The parameter $\eta$ governs the extent of coordination around considered atom.
- Gaussian symmetry radial function, where $\eta$ is a radial function parameter.
The above-mentioned structural descriptors can be considered as the development of the connectivity-based atom contribution (CBAC) approach. This method reduces the diversity of local environments to a very limited number (up to a hundred in the case of structurally diverse elements, such as Zn). Thus, it completely ignores the mutual position of atoms in the first coordination sphere, as well as the distance to the central atom. These factors also have a significant impact on the charge, while the structural descriptors make it possible to take into account minor differences between local environments that are identical in terms of bonding connectivity.

Table 1: Calculated enthalpies of structures with the same topology using partially averaged charges and non-averaged charges from present work.

| CSD reference code (composition) | \( \Delta H \), eV/unit cell (partially averaged charges) | \( \Delta H \), eV/unit cell (non-averaged charges) |
|----------------------------------|-------------------------------------------------|--------------------------------------------------|
| AFUPEX (Tm\(_4\)H\(_{12}\)C\(_{36}\)O\(_{24}\)) | -477.82 | -476.63 |
| SEHSUU (Er\(_4\)H\(_{12}\)C\(_{36}\)O\(_{24}\)) | -472.04 | -471.58 |
| SEHTEF (Y\(_4\)H\(_{12}\)C\(_{36}\)O\(_{24}\)) | -468.10 | -466.72 |

### 3.2 Interatomic potentials fitting

For the evaluation of model applicability, we use the predicted results for building a force-field model for MOF structures. As a starting point, we took experimental data on crystallographic parameters of relatively simple structures and the predicted charges to develop a transferable model of interatomic potentials\[^{[37]}\] for C-O, O-O, and H-O interactions. The used AFUPEX, SEHSUU, and SEHTEF compounds from the database are isostructural and contain different lanthanides (Tm, Er, Lu) (Fig. 7). The choice of compounds is based on previously-developed interatomic potentials and similarity of structures that can ease the first steps. All the potential parameters are fitted on geometry data. As a zero-order approximation, we use values of the parameters for potassium biphthalate from \[^{[38]}\]. All the pair interactions (intramolecular and intermolecular) are described with the Morse function. The function is preferable due to a variety of charges in the compounds, including low values. Morse potential is the best variant to describe interactions with low Coulomb contribution, as it is signed alternating function and can be used even in covalent systems with uncharged atoms. We also add three-body potentials to make some intramolecular angles (like angles in aromatic cycles) more rigid as is usually done in common FFs for organic objects. In order to improve the quality of the model, we fitted individual potentials describing the behavior of fundamentally different positions of C atoms (as shown in Fig. 8). REE-O Morse potential parameters for La-Lu and Y were earlier optimized in using MePO\(_4\) xenotimes and monazites structural information from \[^{[39]}\], as the partial charges on metal cations there were quite similar to the investigated
systems. The set of interatomic potentials for REE-O interactions are presented in Table 1 (in ESI), for other interactions in Table 2 (in ESI). All semiempirical calculations were performed using GULP 4.5 software [40].

We compare geometries optimized for AFUPEX, SEHSUU, and SEHTEF structures with the experimental data. The average relative errors of calculated parameters versus experimental data are less than 2%. Moreover, the potential model, developed on relatively simple structures, reproduced much more complicated ones as well: GEDGED (Cu\textsubscript{16}H\textsubscript{80}C\textsubscript{148}O\textsubscript{80}) and WEHHEY (Zn\textsubscript{16}H\textsubscript{80}C\textsubscript{148}O\textsubscript{80}) (Fig. 3 and 4 in Supporting Information), that contain 324 atoms per unit cell have also been successfully relaxed using these potentials, indicating transferability of potentials model to other MOFs. Thus, developed potentials and charges allowed us to reach a good agreement with experimental crystal structures.

Figure 7: AFUPEX (REE=Tm) (a), SEHSUU (REE=Er) (b) and SEHTEF (REE=Y) (c) structures (76 atoms per unit cell) with total formula REE\textsubscript{4}H\textsubscript{12}C\textsubscript{36}O\textsubscript{24}. REE cations Tm, Er, and Y are shown in blue, green and gray balls respectively. Axonometric projection.

Perhaps, the most important metric of quality here is sensitivity to predicted charges. On the one hand, a good model should not be overfitted to a single set of charges, and we would like to average charges on a single sort of atoms to avoid writing individual potential for each atom in a unit cell. On the other, the whole model should reproduce experimental data with appropriate accuracy.

We calculate the enthalpy difference of AFUPEX, SEHSUU and SEHTEF structures using (1) partially averaged ML charges on atoms (average values of charges for Tm, Er, Y, H, C1, C2, C3, and O were used) as well as for (2) non-averaged ML charges. About 1 eV differences (Table 1) mean "uniform" charge distribution between atoms. It means one can use predicted charges with the proposed model without any additional processing. So, the developed model proved to be transferable in terms of both MOF composition and unit cell size.

Table 2: Calculated enthalpies of MOFs structures (total formula: REE\textsubscript{16}Co\textsubscript{8}H\textsubscript{192}C\textsubscript{384}N\textsubscript{64}O\textsubscript{128}) with the same topology using ML derived and EQEq charges.

| CSD reference code | REE cation | \(\Delta H\), eV/formula (with ML charges) | \(\Delta H\), eV/formula (with EQEq charges) |
|--------------------|------------|------------------------------------------|-------------------------------------------|
| MIFROK             | La         | -380.3                                   | -409.5                                    |
| MIFSAX             | Pr         | -383.2                                   | -422.8                                    |
| MIFSUR             | Nd         | -392.1                                   | -420.1                                    |
| MIFTEC             | Sm         | -395.6                                   | -425.1                                    |
| MIFTIG             | Eu         | -397.8                                   | -436.7                                    |
| MIPTUS             | Gd         | -399.2                                   | -414.8                                    |
| MIFVEE             | Tb         | -403.9                                   | -722.8                                    |
| MIFNUM             | Dy         | -405.9                                   | -442.4                                    |

The final model testing was performed on a series of isostructural MOFs [41] with general composition REE\textsubscript{16}Co\textsubscript{8}H\textsubscript{192}C\textsubscript{384}N\textsubscript{64}O\textsubscript{128} (792 atoms per unit cell, REE: La, Pr, Nd, Sm, Eu, Gd, Tb, Dy). We compare enthalpies of formation calculated based on the predicted charges and the charges, obtained from EQEq model (Table 2).
It is worth to note, the EQEq model predicts +13e relative charge for Tb that is far from the expected one. The predicted enthalpies showed an expected decrease from La to Dy for our model and no evidence of any ordered behavior in a case of EQEq charges. Again, we should note, the linear decrease is not the only possible behavior of REE-containing structures. In some cases, we may see the inverse situation or dome-like dependence that is usually an evidence of a competing process. However, an unordered list of thermodynamic properties in the isostructural REE-containing compounds is a very unlikely situation. Thus, it approves the reliability of the ML and FF models.

4 Conclusions

We present computational approach to generate partial charges with DDEC level accuracy. The main difference between the proposed approach and the reference DFT method is scalability — the use of local features as descriptors results in an approximately linear scale for calculation of charge. It makes it possible to assign charges for structures previously unavailable for this procedure, both in terms of the size of the unit cell and elemental composition.

The developed FF model proved to be transferable and, together with the calculated charges, reproduce structural properties or trends in thermodynamic parameters of metal-organic frameworks, including REE-containing ones. We have also showed a way of simplifying the model by averaging charges on equivalent positions. The model can be used, for example, for molecular dynamics simulations of MOFs, including those, containing rare-earth elements. All the potentials details are presented in the Supporting Information.

The graphical interface to a developed model will be soon available on [https://api.scidatasoft.com/mof](https://api.scidatasoft.com/mof).

References

[1] Peter G. Boyd, Yongjin Lee, and Berend Smit. Computational development of the nanoporous materials genome. *Nature Reviews Materials*, 2:17037, 2017.

[2] Song Dang, Qi Long Zhu, and Qiang Xu. Nanomaterials derived from metal-organic frameworks. *Nature Reviews Materials*, 3(1):17075, 2017.

[3] Jiewei Liu, Lianfen Chen, Hao Cui, Jianyong Zhang, Li Zhang, and Cheng Yong Su. Applications of metal-organic frameworks in heterogeneous supramolecular catalysis. *Chemical Society Reviews*, 43(16):6011–6061, 2014.

[4] Jian-rong Li, Julian Sculley, and Hong-cai Zhou. Metal–Organic Frameworks for Separations. *Chemical Reviews*, 112(2):869–932, 2012.

[5] Shengqian Ma and Hong-cai Zhou. Gas storage in porous metal – organic frameworks for clean energy applications. *Chemical Communications*, 46(1):44–53, 2010.

[6] Shun Li Li and Qiang Xu. Metal-organic frameworks as platforms for clean energy. *Energy and Environmental Science*, 6(6):1656–1683, 2013.

[7] Christopher E. Wilmer, Michael Leaf, Chang Yeon Lee, Omar K. Farha, Brad G. Hauser, Joseph T. Hupp, and Randall Q. Snurr. Large-scale screening of hypothetical metal-organic frameworks. *Nature Chemistry*, 4(2):83–89, 2012.
[8] Lukáš Grajciar, Andrew D. Wiersum, Philip L. Llewellyn, Jung San Chang, and Petr Nachtigall. Understanding CO2 adsorption in CuBTC MOF: Comparing combined DFT-ab initio calculations with microcalorimetry experiments. *Journal of Physical Chemistry C*, 115(36):17925–17933, 2011.

[9] Anthony K. Rappé and William A. Goddard. Charge equilibration for molecular dynamics simulations. *Journal of Physical Chemistry*, 95(8):3358–3363, 1991.

[10] Christopher E. Wilmer and Randall Q. Snurr. Towards rapid computational screening of metal-organic frameworks for carbon dioxide capture: Calculation of framework charges via charge equilibration. *Chemical Engineering Journal*, 171(3):775–781, 2011.

[11] Christopher E. Wilmer, Ki Chul Kim, and Randall Q. Snurr. An extended charge equilibration method. *Journal of Physical Chemistry Letters*, 3(17):2506–2511, 2012.

[12] Sunder Ramachandran, T. G. Lenz, W. M. Skiff, and A. K. Rappé. Toward an understanding of zeolite Y as a cracking catalyst with the use of periodic charge equilibration. *Journal of Physical Chemistry*, 100(14):5898–5907, 1996.

[13] Eugene S. Kadantsev, Peter G. Boyd, Thomas D. Daff, and Tom K. Woo. Fast and accurate electrostatics in metal organic frameworks with a robust charge equilibration parameterization for high-throughput virtual screening of gas adsorption. *Journal of Physical Chemistry Letters*, 4(18):3056–3061, 2013.

[14] Curt M. Breneman and Kenneth B. Wiberg. Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *Journal of Computational Chemistry*, 11(3):361–373, 1990.

[15] Thomas A. Manz and David S. Sholl. Chemically meaningful atomic charges that reproduce the electrostatic potential in periodic and nonperiodic materials. *Journal of Chemical Theory and Computation*, 6(8):2455–2468, 2010.

[16] Carlos Campañá, Bastien Mussard, and Tom K. Woo. Electrostatic potential derived atomic charges for periodic systems using a modified error functional. *Journal of Chemical Theory and Computation*, 5(10):2866–2878, 2009.

[17] Dalar Nazarian, Jeffrey S. Camp, and David S Sholl. A Comprehensive Set of High-Quality Point Charges for Simulations of Metal-Organic Frameworks. *Chemistry of Materials*, 28(3):785–793, 2016.

[18] Thomas A. Manz and David S. Sholl. Improved atoms-in-molecule charge partitioning functional for simultaneously reproducing the electrostatic potential and chemical states in periodic and nonperiodic materials. *Journal of Chemical Theory and Computation*, 8(8):2844–2867, 2012.

[19] Qing Xu and Chongli Zhong. A general approach for estimating framework charges in metal-organic frameworks. *Journal of Physical Chemistry C*, 114(11):5035–5042, 2010.

[20] Benjamin Nebgen, Nicholas Lubbers, Justin S Smith, Andrew E Sifain, Andrey Lokhov, Olexandr Isayev, Adrian E Roitberg, Kipton Barros, and Sergei Tretiak. Transferable Dynamic Molecular Charge Assignment Using Deep Neural Networks. *Journal of Chemical Theory and Computation*, 14(9):4687–4698, 2018.

[21] Kay Schaller and Sereina Riniker. Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations. *Journal of chemical information and modeling*, 58(3):579–590, 2018.

[22] Andrew E Sifain, Nicholas Lubbers, Benjamin T Nebgen, Justin S Smith, Andrey Y. Lokhov, Olexandr Isayev, Adrian E Roitberg, Kipton Barros, and Sergei Tretiak. Discovering a Transferable Charge Assignment Model Using Machine Learning. *Journal of Physical Chemistry Letters*, 9(16):4495–4501, 2018.

[23] Felix A. Faber, Alexander Lindmaa, O. Anatole Von Lilienfeld, and Rickard Armineto. Machine Learning Energies of 2 Million Elpasolite (ABC2D6) Crystals. *Physical Review Letters*, 117(13):2–7, 2016.

[24] Jonathan Schmidt, Jingming Shi, Pedro Borlido, Liming Chen, Silvana Botti, and Miguel A. L. Marques. Predicting the Thermodynamic Stability of Solids Combining Density Functional Theory and Machine Learning. *Chemistry of Materials*, 29(12):5090–5103, 2017.

[25] Ying Zhang and Chen Ling. A strategy to apply machine learning to small datasets in materials science. *npj Computational Materials*, 4(1):28–33, 2018.

[26] Yongchul G Chung, Jeffrey Camp, Maciej Haranczyk, Benjamin J Sikora, Wojciech Bury, Vaiva Krungleviciute, Omar K Farha, David S Sholl, and Randall Q Snurr. Computation-Ready, Experimental Metal-Organic Frameworks: A Tool To Enable High-Throughput Screening of Nanoporous Crystals. *Chemistry of Materials*, 26:6185–6192, 2014.

[27] João Rocha, Luís D. Carlos, Filipe A. Almeida Paz, and Duarte Ananias. Luminescent multifunctional lanthanides-based metal-organic frameworks. *Chemical Society Reviews*, 40(2):926–940, 2011.
[28] Yuanjing Cui, Banglin Chen, and Guodong Qian. Lanthanide metal-organic frameworks for luminescent sensing and light-emitting applications. *Coordination Chemistry Reviews*, 273-274:76–86, 2014.

[29] John B O Mitchell. Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(5):468–481, 2014.

[30] Luca M. Ghiringhelli, Jan Vybiral, Sergey V. Levchenko, Claudia Draxl, and Matthias Scheffler. Big data of materials science: Critical role of the descriptor. *Physical Review Letters*, 114(10):1–5, 2015.

[31] Atsuto Seko, Hiroyuki Hayashi, Keita Nakayama, Akira Takahashi, and Isao Tanaka. Representation of compounds for machine-learning prediction of physical properties. *Physical Review B*, 95(14):1–11, 2017.

[32] Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, and Chiho Kim. Machine learning in materials informatics: Recent applications and prospects. *npj Computational Materials*, 3(1), 2017.

[33] Noel M O Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, 2011.

[34] Christopher E. Wilmer, Omar K. Farha, Youn Sang Bae, Joseph T. Hupp, and Randall Q. Snurr. Structure-property relationships of porous materials for carbon dioxide separation and capture. *Energy and Environmental Science*, 5(12):9849–9856, 2012.

[35] Toon Verstraelen, Veronique Van Speybroeck, and Michel Waroquier. The electronegativity equalization method and the split charge equilibration applied to organic systems: Parametrization, validation, and comparison. *Journal of Chemical Physics*, 131(4), 2009.

[36] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv:1802.03888*, 2018.

[37] Nickolay N Eremin, Ekaterina I Marchenko, Vladimir G Petrov, Artem A Mitrofanov, and Amina S Ulanova. Solid solutions of monazites and xenotimes of lanthanides and plutonium: Atomistic model of crystal structures, point defects and mixing properties. *Computational Materials Science*, 157:43–50, 2019.

[38] Vadim S Urusov and Nickolay N Eremin. Atomistic Computer Simulation of the Structures and Properties of Inorganic Crystals and Minerals, their Defects, and Solid Solutions. *GEOS, Moscow*, 2012.

[39] Yunxiang Ni, John M Hughes, and Anthony N Mariano. Crystal chemistry of the monazite and xenotime structures. *American Mineralogist*, 80:21–26, 1995.

[40] Julian D. Gale and Andrew L. Rohl. The General Utility Lattice Program (GULP). *Molecular Simulation*, 29(5):291–341, 2003.

[41] Jing Min Zhou, Wei Shi, Na Xu, and Peng Cheng. A new family of 4f-3d heterometallic metal-organic frameworks with 2,2’-bipyridine-3,3’-dicarboxylic acid: Syntheses, structures and magnetic properties. *Crystal Growth and Design*, 13(3):1218–1225, 2013.