Article

Message Passing Neural Networks for Partial Charge Assignment to Metal−Organic Frameworks

Ali Raza, Arni Sturluson, Cory M. Simon,* and Xiaoli Fern*

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Virtual screenings can accelerate and reduce the cost of discovering metal−organic frameworks (MOFs) for their applications in gas storage, separation, and sensing. In molecular simulations of gas adsorption/diffusion in MOFs, the adsorbate−MOF electrostatic interaction is typically modeled by placing partial point charges on the atoms of the MOF. For the virtual screening of large libraries of MOFs, it is critical to develop computationally inexpensive methods to assign atomic partial charges to MOFs that accurately reproduce the electrostatic potential in their pores. Herein, we design and train a message passing neural network (MPNN) to predict the atomic partial charges on MOFs under a charge neutral constraint. A set of ca. 2250 MOFs labeled with high-fidelity partial charges, derived from periodic electronic structure calculations, serves as training examples. In an end-to-end manner, from charge-labeled crystal graphs representing MOFs, our MPNN machine-learns features of the local bonding environments of the atoms and, from these features, learns to predict partial atomic charges. Our trained MPNN assigns high-fidelity partial point charges to MOFs with orders of magnitude lower computational cost (runtime: ~3 s) than electronic structure calculations. To enhance the accuracy of virtual screenings of large libraries of MOFs for their adsorption-based applications, we make our trained MPNN model and MPNN-charge-assigned computation-ready, experimental MOF structures publicly available.
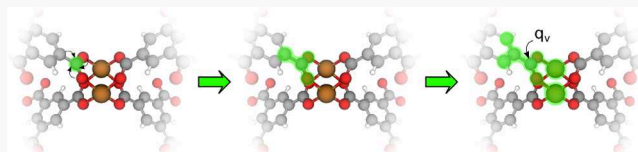
■ INTRODUCTION

Metal−organic frameworks (MOFs) are nanoporous materials that often exhibit large internal surface areas.[1] Because MOFs selectively adsorb gas into their pores/on their internal surface, MOFs can be used to store,[2] separate/purify,[3] and sense[4] gases. Moreover, MOF structures are highly adjustable and therefore can be tuned to optimize a target adsorption property for a given adsorption-based engineering application.[5] The adjustability of MOFs stems from their modular synthesis: metals/metal clusters and organic linker molecules self-assemble into a crystalline structure.[1] By changing the molecular building blocks, many MOFs with diverse pore shapes and internal surface chemistries can be synthesized; on the order of 10 000 porous MOFs[6,7] have been reported to date.

Molecular models and simulations play an important role in the discovery and deployment of MOFs for adsorption-based applications.[8−10] Instead of an Edisonian approach to find MOFs that meet a target adsorption property, classical molecular models and simulations can quickly and cost-effectively predict the adsorption properties of thousands of MOFs and shortlist the most promising subset for experimental investigation. High-throughput computational screenings of MOFs have directly led to the discovery of high-performing MOFs for carbon dioxide capture,[11,12] xenon/krypton separations,[13] oxygen storage,[14] hydrogen storage,[12] and mustard gas capture.[15] In addition to virtual screening, molecular simulations of gas adsorption in MOFs can elucidate the most favorable adsorption sites in a MOF,[16,17] explain

anomalous adsorption phenomena,[18,19] and uncover structure−property relationships.[20,21]

The molecular mechanics description of adsorbate−MOF interactions[86] used in a molecular simulation typically consists of the sum of a van der Waals and an electrostatic contribution.[10] The electrostatic interaction is particularly important for adsorbates with polar bonds, such as CO_2 and H_2O.[22−24] To model the adsorbate−MOF electrostatic interaction, we must model the electrostatic potential in the pores of the MOF, created by the atoms of the MOF. Typically, the MOF-hosted electrostatic potential is described by placing (fixed) partial (i.e., noninteger) point charges at the centers of the atoms of the MOF.[25] Molecular models for adsorbate molecules, such as CO_2 and N_2,[26] also possess partial point charges, which, in a molecular simulation, interact with the point charges on the MOF via Coulomb's law to comprise the adsorbate−MOF electrostatic potential energy of interaction.[27,87] There are several methods to assign partial charges to the atoms of a MOF,[25] which are nonobservable. Choosing a charge assignment method for the virtual screening of a large library of MOFs often involves a trade-off between

computational cost and accuracy of the representations of the electrostatic potential in the pores of the MOFs. Notably, the simulated adsorption properties and thus ranking of MOFs in virtual screenings could be highly dependent on the (accuracy of the) charge assignment method.[22,29−31]

Broadly, methods to assign partial point charges to a MOF[25] take two different approaches: (1) use an electronic structure calculation (e.g., a density functional theory (DFT) calculation) to obtain the electrostatic potential/electron density (a) in the pores of the (periodic) MOF or (b) surrounding a nonperiodic cluster representation of the MOF, then derive charges that are consistent with this electrostatic potential/ electron density; (2) assign charges using a (semi)empirical model whose parameters were fit to experimental data or to charges assigned by approach 1. Approach 1 includes the commonly used methods, Repeating Electrostatic Potential Extracted ATomic (REPEAT)[32] and Density Derived Electrostatic and Chemical (DDEC)[33] charges. In molecular building block-based charge assignment, MOFs inherit the charges of their molecular building blocks,[34] so that molecular charges, derived from electronic structure calculations, on a set of linker molecules and a set of metal clusters provide charges for a combinatorial number of MOFs. Approach 2 includes charge equilibration methods (QEq),[35,36] statistical machine learning models,[37−40] and nearest-neighbor-like approaches based on the chemical element and bonding environment of the atom.[41−44] Generally, approach 1 produces a more accurate electrostatic potential in the pores of the MOF but incurs a computational cost orders of magnitude greater than the cost of approach 2. Often, with the aim of predicting gas adsorption in a given MOF, charge assignment using electronic structure calculations (approach 1) is more time-consuming than the Monte Carlo simulation of adsorption.

Thankfully, Nazarian et al.[45] performed periodic DFT calculations to obtain the electron densities in ca. 2900 experimentally synthesized MOFs[46] and assigned chemically meaningful, high-quality partial point charges to each MOF via the DDEC method.[33] Still, a large number of MOFs lack high-quality charges: (i) the majority of the second version (v2) of the computation-ready, experimental (CoRE) MOF data set[7] of ca. 14 000 structures; (ii) newly synthesized MOFs that are continually reported;[6] and (iii) libraries of hypothetical/ predicted MOF crystal structure models constructed with the aim of discovering new MOFs that have not been synthesized in the laboratory.[47−49] Screening these MOFs for gas storage, separations, and sensing via molecular simulations demands a computationally cheap *and* high-fidelity method for MOF charge assignment.

In this work, we develop and train a message passing neural network (MPNN)[50,51] architecture to assign partial point charges to each atom of a MOF structure under a charge-neutral (hard) constraint. To enable our machine-learning of high-fidelity charges on MOFs, we leverage the database of DFT-derived partial point charges on ca. 2900 experimental MOFs by Nazarian et al.[45] as training examples. Our fundamental hypothesis, supported by refs 42 and 44, is that the charge of any given atom in a MOF is primarily determined by its chemical identity and local bonding environment. As opposed to manually engineering a feature to represent the local bonding environments of atoms,[52,53] our MPNN machine learns vector representations of local bonding environments of atoms within MOFs and then, from these features, predicts their partial point charges. We train the MPNN to do this in an

end-to-end manner, enriching the machine-learned features of the local bonding environments of atoms with information predictive of partial charge. The MOF crystal structures, represented as undirected graphs (nodes: atoms, edges: bonds) with node features encoding their chemical identities, are the direct inputs to the MPNN. Edges (bonds) across the unit cell boundary are included to account for periodicity. Our MPNN begins with a node embedding layer that learns a dense representation of the chemical elements, encoding their typical charge. The MPNN then sequentially passes information along the edges of the graph to learn/construct the vector representations of the local bonding environments. The final layers use these features to assign a charge to each node under a charge neutral constraint. To achieve this, we model the probabilistic distribution of charge on an atom within its local bonding environment and invoke the maximum likelihood principle under the charge neutral constraint. This allows the MPNN to give more slack to the charge of atoms with high variance when enforcing charge neutrality.

Our trained MPNN assigns high-fidelity (treating the DFT-derived DDEC charges[45] as ground truth) charges to MOF atoms (mean absolute deviation on test MOFs, 0.025), outperforming a suite of charge equilibration methods[36] (minimum mean absolute deviation, 0.118, by I-QEq[54]), while incurring orders of magnitude lower computational cost than electronic structure calculations. To enable accurate virtual screenings of large libraries of MOFs for their adsorption-based applications, we make our trained MPNN model available to the molecular simulation community for MOF charge assignment and provide .cif files of MPNN-charge-assigned v2 computation-ready, experimental MOFs.[7]

■ REVIEW OF PREVIOUS WORK

References 25 and 27 review methods for assigning atomic partial charges to MOFs to enable molecular simulation of gas adsorption and diffusion. The most accurate, but computationally costly, approach is to use an electronic structure calculation (e.g., DFT) to obtain the periodic electrostatic potential/electron density in the pores of the MOF and then derive point charges that are consistent with this (e.g., REPEAT[32] and DDEC[33]).[88] The less accurate, but computationally cheap, approach is to use a (semi)empirical model to assign charges to MOF atoms, whose parameters were fit to experimental data or charges assigned with electronic structure calculations as input. Semiempirical charge equilibration (QEq) methods[35] are commonly used to assign point charges to MOFs owing to their low computational cost. Ongari et al.[36] review and compare several QEq variants and assess their correlation with the DFT-derived DDEC charges of ca. 2 900 MOFs by Nazarian et al.[45] The ionizing I-QEq[54] variant produced charges closest to the DDEC charges (mean absolute deviation 0.118), but there were significant deviations, which then propagate onto, e.g., carbon dioxide adsorption in a molecular simulation.

Along the direction of this work, a few authors trained supervised machine learning models to assign partial charges to atoms of molecules (not periodic MOFs), using feature vectors of the local environment of an atom that are either (i) manually engineered[37−39,55] or (ii) learned end-to-end by a message passing neural network.[40,56] An interesting subproblem is that, when a supervised model predicts the charge of each atom in the molecule pseudoindependently, based on its local environment, charge neutrality of the molecule is not

guaranteed. To enforce charge neutrality, directly after the model assigns (before-constraint) charges to the atoms, refs 37 and 38 distributed the negative of the excess charge of the molecule among its atoms, not uniformly, but based on the variances of the charges on the atoms predicted by an ensemble of decision tree regressors. Atoms associated with more (less) variance received more (less) of the excess charge. Reference 44 distributed excess charge based on the magnitude of the predicted charge. Reference 40 enforced charge neutrality by, instead of directly predicting charges, predicting the electronegativity and hardness of each atom in its local bonding environment, then minimizing the potential energy of the atoms of the molecule (dependent the electronegativities and hardnesses of the atoms) under a charge neutrality constraint.

Using the molecular graph as direct input, MPNNs have recently been employed to predict several different properties of molecules[51,57−66] and crystals[67−69] including antibacterial efficacy, electronic properties, solubility, photovoltaic efficiency, odor, and drug efficacy. In addition, MPNNs have been used to predict chemical reactions.[70,71] The key advantage/novelty of MPNNs is that, instead of manually engineering molecular descriptors,[52,53] the MPNN automatically learns a (task-specific) descriptor of the molecule and its set of local bonding environments from the molecular graph in an end-to-end manner, while being trained to perform a prediction task.[57] In contrast to mapping a molecule or crystal (represented as a graph) to a single property, our MPNN architecture is unique in that it maps a crystal (represented as an undirected, node-labeled graph with edges across the periodic boundary included) to targets (partial charges) on each node (atom) under a graph-level (charge neutrality) constraint.

PROBLEM FORMULATION

Here, we mathematically formulate the partial charge assignment problem. The notation is listed in Table 1. We use bold lowercase letters for vectors and bold uppercase letters for matrices.

We represent the crystal structure of each MOF as an undirected graph with node features, $G = (\mathcal{V}, \mathcal{E}, \mathbf{X})$, where $\mathcal{V}$ is the set of $n = |\mathcal{V}|$ nodes or vertices, representing atoms, $\mathcal{E}$ is the set of edges, representing bonds, and $\mathbf{X} \in \mathbb{R}^{d \times n}$ is the node feature matrix. The feature vector of node (atom) v in the graph (MOF), $\mathbf{x}_v \in \mathbb{R}^d$, is a one-hot encoding of its chemical element and is column v of $\mathbf{X}$. Let $e_{uv} = \{u, v\} \in \mathcal{E}$ denote an edge (bond) between nodes (atoms) u and v. Let the adjacency matrix of the graph be $\mathbf{A} \in \mathbb{R}^{n \times n}$, where $A_{uv} = 1$ if nodes u and v are connected by an edge and $A_{uv} = 0$ otherwise. Together, the adjacency matrix $\mathbf{A}$ and node feature matrix $\mathbf{X}$ characterize the crystal graph of a MOF.

Our goal is to learn a function $\mathbf{f}$ that takes the crystal graph G as input and outputs a predicted charge on each node:

$$(\mathbf{X}, \mathbf{A}) \mapsto \mathbf{f}(\mathbf{X}, \mathbf{A}) = \mathbf{q} \tag{1}$$

while satisfying the charge neutrality constraint:

$$\sum_{v=1}^{n} q_v = 0 \tag{2}$$

Here, q_v is the charge on node v and element v of the charge vector $\mathbf{q} \in \mathbb{R}^n$ of the graph. The function $\mathbf{f}$ will be equivariant

Table 1. Notation and Definitions

symbol	description
G	graph
$\mathcal{V}$	set of nodes in graph
$\mathcal{E}$	set of edges in graph
n	number of nodes in graph
e_{ij}	edge between nodes i and j
$\mathbf{A} \in \mathbb{R}^{n \times n}$	adjacency matrix of graph
$\mathbf{x}_v \in \mathbb{R}^d$	feature vector of node v (one-hot encoding of element)
$\mathbf{X} \in \mathbb{R}^{d \times n}$	feature matrix of graph
$\mathbf{x}_v^e \in \mathbb{R}^r$	element embedding of node v
$\mathbf{X}^e \in \mathbb{R}^{r \times n}$	node embedding matrix of graph
t	message passing time
$\mathbf{h}_v^{(t)} \in \mathbb{R}^k$	hidden representation of node v after t messages
$\mathbf{H}^{(t)} \in \mathbb{R}^{k \times n}$	hidden node feature matrix of graph after t messages
q_v	partial charge on node v (units: electron charge)
$\hat{q}_v$	predicted partial charge on node v (units: electron charge)
$\mathbf{q} \in \mathbb{R}^n$	charge vector of graph
$\mathcal{N}(v)$	neighbors of node v
$\boldsymbol{\beta}_\square \ (\mathbf{B}_\square)$	vector (matrix) of weights
$[\mathbf{a}, \mathbf{b}]$	concatenation of vectors $\mathbf{a}$ and $\mathbf{b}$

(i.e., $\mathbf{f}(\mathbf{XP}, \mathbf{PAP}^{\mathrm{T}}) = \mathbf{Pq}$ where $\mathbf{P}$ is a permutation matrix that permutes the nodes) so that the ordering of the atoms is immaterial.

MACHINE-LEARNING PARTIAL CHARGES

Converting a MOF Crystal Structure to a Graph. We first describe how we convert a MOF crystal structure stored in a .cif file into an undirected graph $G = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ (nodes/atoms: $\mathcal{V}$, edges/bonds: $\mathcal{E}$, node features encoding chemical elements: $\mathbf{X}$). A .cif file of a MOF provides its unit cell vectors, a list of its atoms, and the coordinates of its atoms.

Nodes and Edges. We first construct a set of nodes to represent the atoms of the MOF (a bijection). To determine the edges between the nodes (i.e., the adjacency matrix, $\mathbf{A}$), we must automatically infer which atoms are bonded based on their identities and atomic coordinates. We used a bond assignment algorithm from refs 68 and 72 that considers both typical bond lengths and the arrangements of nearby atoms. We assign an edge (bond) between nodes u and v if (1) the periodic Euclidean distance between them is less than the sum of their covalent radii[73] plus a 0.25 Å tolerance and (2) they share a Voronoi face. By applying the minimum image convention when computing the distance, we include edges between atoms bonded across the periodic boundary. To determine which atoms share a Voronoi face with atom u, we used Scipy[74] to compute the Voronoi diagram of all atoms within a 6 Å radius of atom u, periodic images included. To ensure bonds were properly formed with metal atoms commonly found in MOFs, we increased the covalent radius for 10 metals (see section S2).

Node Features. For each node v, we construct its feature vector, $\mathbf{x}_v \in \mathbb{R}^d$, as a one-hot encoding of its chemical element. i.e., entry i of $\mathbf{x}_v$ is one if atom v in the MOF is chemical element i and zero otherwise. Among the charge-labeled MOFs,[45] there were 74 unique chemical elements (see Table
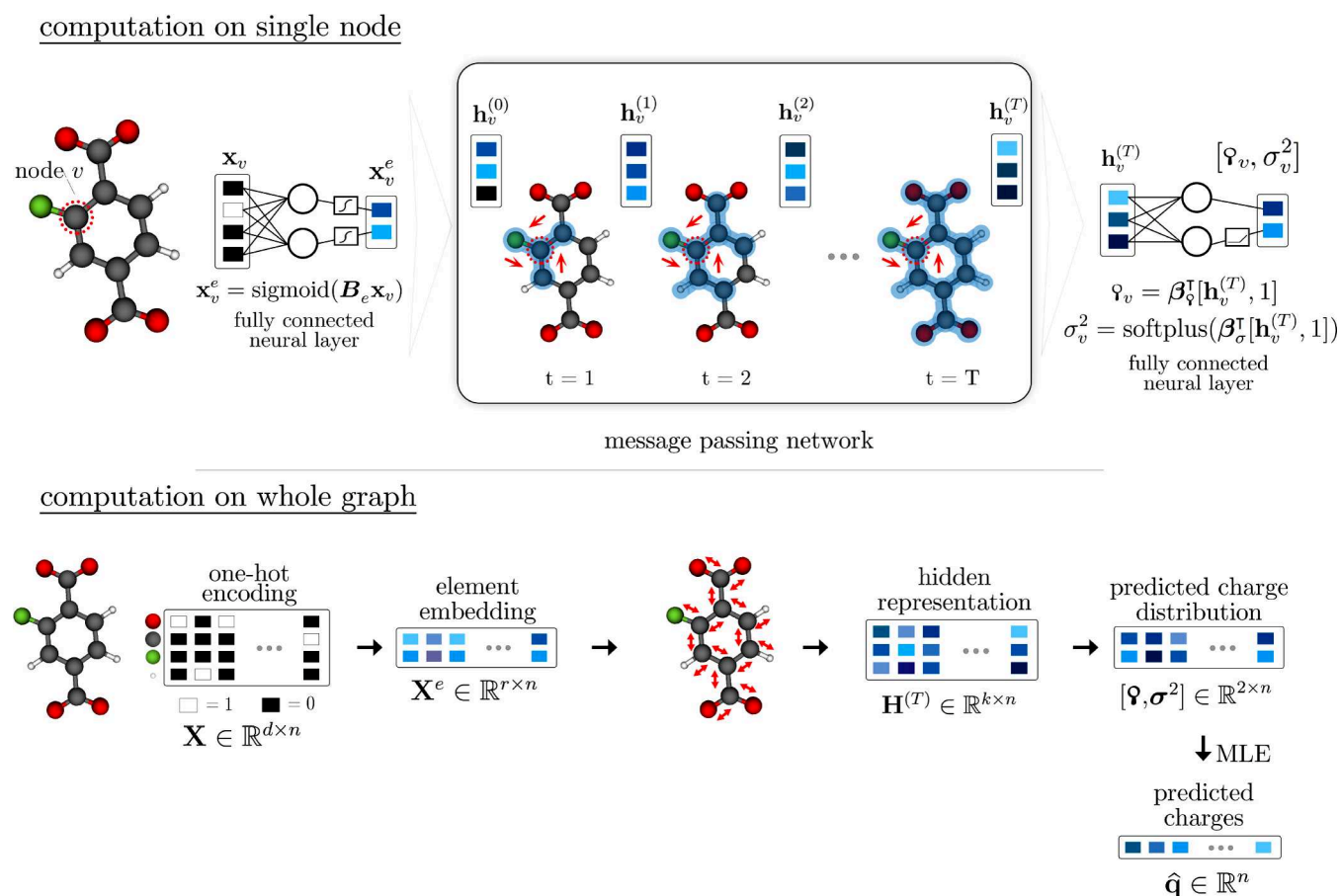
Figure 1. Architecture of our message passing neural network (MPNN) to assign partial charges to the atoms of a MOF atoms under a MOF-level charge neutral constraint. (Top: node-level computations) First, the one-hot encoding of the atomic species of node v, $\mathbf{x}_v$, is mapped to an information-dense chemical element embedding $\mathbf{x}_v^e$ by a fully connected layer. The hidden node feature is initialized as $\mathbf{h}_v^{(0)}$ using $\mathbf{x}_v^e$. Next is the message passing phase, where, in each time step, every node shares information with its neighbors and updates its hidden representation. With more time steps, the hidden representation of the node, $\mathbf{h}_v^{(t)}$, captures a broader view of its local bonding environment. Modeling the probabilistic distribution of the charge on each node (within its local bonding environment) as Gaussian, a fully connected layer outputs the mean φ_v and variance σ_v^2 from the final, learned hidden representation $\mathbf{h}_v^{(T)}$. (Bottom: graph-level computations) The molecular graph G representing the MOF crystal structure is input to the MPNN. Each node is processed independently, as depicted on the top, culminating in the predicted means (bold φ) and variances $\boldsymbol{\sigma}^2$. Subsequently, a maximum likelihood estimation under the charge neutral constraint gives the predicted charges $\hat{\mathbf{q}}$.

S6); so, $d = 74$. The node feature vectors comprise the columns of the node feature matrix $\mathbf{X}$ of the MOF.

Target Vector. We construct the charge vector $\mathbf{q}$ for the DFT-derived, DDEC charge-labeled MOFs[45] whose .cif files also contain the partial atomic charge on each atom.

Ordering, Equivariance, and Rotation- and Translation-Invariance. Of course, element v of the target (charge) vector $\mathbf{q}$, column v of the node feature matrix $\mathbf{X}$, and row/column v of the adjacency matrix $\mathbf{A}$ all represent the same atom in the MOF. However, the ordering of atoms in the crystal structure file is immaterial, as the function $\mathbf{f}$ in eq 1 learned by the MPNN is equivariant[50] to permutations of the nodes. Notably, the graph representation of the MOF is also rotation- and translation-invariant.

Neural Architecture. Figure 1 shows the architecture of our message passing neural network (MPNN)[50] to assign partial point charges to each node of a graph, representing a MOF crystal structure, under a charge neutral constraint. This MPNN, described in detail below, constitutes the function $\mathbf{f}$ in eq 1 that obeys the constraint in eq 2. Our MPNN architecture is composed of, sequentially, (1) an element embedding layer to map the node features (the one-hot encodings of chemical

elements) into information-dense chemical element representations for initializing hidden node features, (2) a gated graph neural network[75] that passes messages between neighboring nodes, along the edges of the graph, to learn/construct hidden node representations that encode the local bonding environment of each node, and (3) node-level charge prediction under the graph-level charge neutrality constraint.

Chemical Element Embedding. First, we map each node feature vector $\mathbf{x}_v$ to a compressed representation, $\mathbf{x}_v^e \in \mathbb{R}^r$ ($r \ll d$):

$$\mathbf{x}_v^e = \text{sigmoid}.(\boldsymbol{B}_e \mathbf{x}_v) \qquad (3)$$

The learned matrix of weights $\boldsymbol{B}_e$ is shared across all nodes. Because each node feature $\mathbf{x}_v$ is a one-hot encoding of a chemical element, the *embedding* of chemical element i is column i of the matrix $\boldsymbol{B}_e$ passed through a sigmoid activation function (. for element-wise) to limit the range. Thus, the element embedding layer in eq 3 maps each chemical element, one-hot encoded in $\mathbf{x}_v$, to a low-dimensional, dense feature vector that, as we will find, encodes its typical charge. The hidden feature vector of node v is initialized using its element

embedding $\mathbf{x}_v^e$ to facilitate training in the message passing phase, which we describe next.

Message Passing. In the message passing phase,[51] a gated graph neural network (GGNN)[75] iteratively updates the hidden features (representations) of the nodes by passing information between neighboring nodes, along the edges of the graph. The *message* received by a node is a conglomeration of the information received from its neighbors. The GGNN employs a gated recurrent unit (GRU)[76] to, at each time step, update the hidden representation of each node using its current hidden representation and the message from its neighbors. Messages are passed for T time steps, after which each hidden node feature encodes the local bonding environment, which we define more precisely below, of the atom it represents.

The initial hidden representation of node v, $\mathbf{h}_v^{(0)} \in \mathbb{R}^k$, is set as its chemical element embedding concatenated with a zero vector:

$$\mathbf{h}_v^{(0)} = [\mathbf{x}_v^e, \mathbf{0}] \tag{4}$$

We concatenate with the zero vector of dimensionality $k - r$ to allow the hidden representation to be higher-dimensional than the element embedding; conceptually, this is to account for the higher information content in the hidden node representation than in the element embedding of the node, as the former encapsulates both the atomic species of the node (as does the element embedding) *and* the surrounding bonding environment of the node (which the element embedding does not).

In taking a message passing time step from time t to $t + 1$, each node collects, sums, and transforms the hidden representations of its neighbors, summarizing the information received into a message $\mathbf{m}_v^{(t+1)}$. Specifically, the message received by node v is constructed as

$$\mathbf{m}_v^{(t+1)} = B_m \sum_{u \in \mathcal{N}(v)} \mathbf{h}_u^{(t)} \tag{5}$$

where B_m is a learned $k \times k$ matrix shared across all nodes and $\mathcal{N}(v)$ is the neighborhood of node v:

$$\mathcal{N}(v) = \{u \in \mathcal{V} \mid e_{uv} \in \mathcal{E}\}$$

The hidden representation of node v is then updated by a GRU[76] (shared across all nodes) based on its message received and its current hidden representation:

$$\mathbf{h}_v^{(t+1)} = \text{GRU}(\mathbf{h}_v^{(t)}, \mathbf{m}_v^{(t+1)}) \tag{6}$$

See section S4 for GRU details. The message passing phase is comprised of T such time steps.

At the end of the message passing phase, each node has a hidden representation $\mathbf{h}_v^{(T)}$ that encodes both its atomic species and its local bonding environment. Precisely, the *local bonding environment* of node v encoded in the hidden node feature $\mathbf{h}_v^{(T)}$ is the induced subgraph of G containing all nodes with geodesic distance less than or equal to T from node v. See Figure 2.

Charge Prediction. Next, we use the learned representations of the local bonding environments of the nodes to predict their partial charges.

As opposed to directly predicting the charge on node v from $\mathbf{h}_v^{(T)}$, we instead view q_v as a random variable and model its conditional probability density as a Gaussian with mean $\text{\Qoppa}_v = \text{\Qoppa}_v(\mathbf{h}_v^{(T)})$ and variance $\sigma_v^2 = \sigma_v^2(\mathbf{h}_v^{(T)})$:
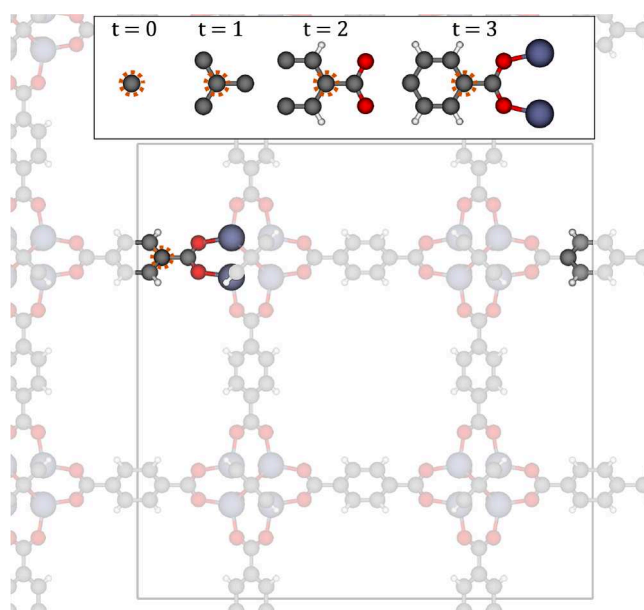
Figure 2. Illustration of the *local bonding environment* of node v after t message passing time steps (top), encoded in its hidden representation, $\mathbf{h}_v^{(t)}$. Focal node v, representing a carbon atom in IRMOF-1, is circled in orange. Edges across the unit cell boundary (gray square) are included to account for periodicity. As t increases, $\mathbf{h}_v^{(t)}$ contains a broader view of the bonding environment of the atom.

$$q_v \mid \mathbf{h}_v^{(T)} \sim N(\text{\Qoppa}_v, \sigma_v^2) \tag{7}$$

i.e., we aim to learn and predict not only the typical charge of an atom within a given local bonding environment, but also its variance. The variance will be useful for adjusting the charges to enforce charge neutrality; charges of atoms with higher variance will be given more slack. We use the archaic, Greek symbol $\text{\Qoppa}$, thought to be the source of the Latin "Q" often used to denote charge, to denote the mean of the Gaussian distribution in eq 7 to encourage interpretation of $\text{\Qoppa}_v$ as a before-constraint charge prediction for atom v.

We use a fully connected layer comprised of two neurons, with weights $\boldsymbol{\beta}_{\text{\Qoppa}}$ and $\boldsymbol{\beta}_\sigma$ shared across all nodes, to predict $\text{\Qoppa}_v$ and σ_v^2, respectively, from the learned representation of the local bonding environment of node v, $\mathbf{h}_v^{(T)}$:

$$\text{\Qoppa}_v = \boldsymbol{\beta}_{\text{\Qoppa}}^T [\mathbf{h}_v^{(T)}, 1] \tag{8}$$

$$\sigma_v^2 = \text{softplus}(\boldsymbol{\beta}_\sigma^T [\mathbf{h}_v^{(T)}, 1]) \tag{9}$$

The softplus activation function ensures $\sigma_v^2 > 0$.

Finally, to arrive at the predicted charges $\hat{\mathbf{q}}$ on a given MOF, under the charge neutral constraint in eq 2, we invoke the maximum likelihood principle. Assuming q_v is conditionally independent given its local bonding environment and distributed according to eq 7, the log-likelihood $\mathcal{L}$ of observing charges $\hat{\mathbf{q}}$ on a given MOF is

$$\mathcal{L}(\hat{\mathbf{q}}) = \sum_{v=1}^{n} \left(\log\left(\frac{1}{\sigma_v \sqrt{2\pi}} \right) - \frac{(\text{\Qoppa}_v - \hat{q}_v)^2}{2\sigma_v^2} \right) \tag{10}$$

Maximizing $\mathcal{L}$ under the charge neutral constraint $\sum_{v=1}^{n} \hat{q}_v = 0$, we find (see S5):
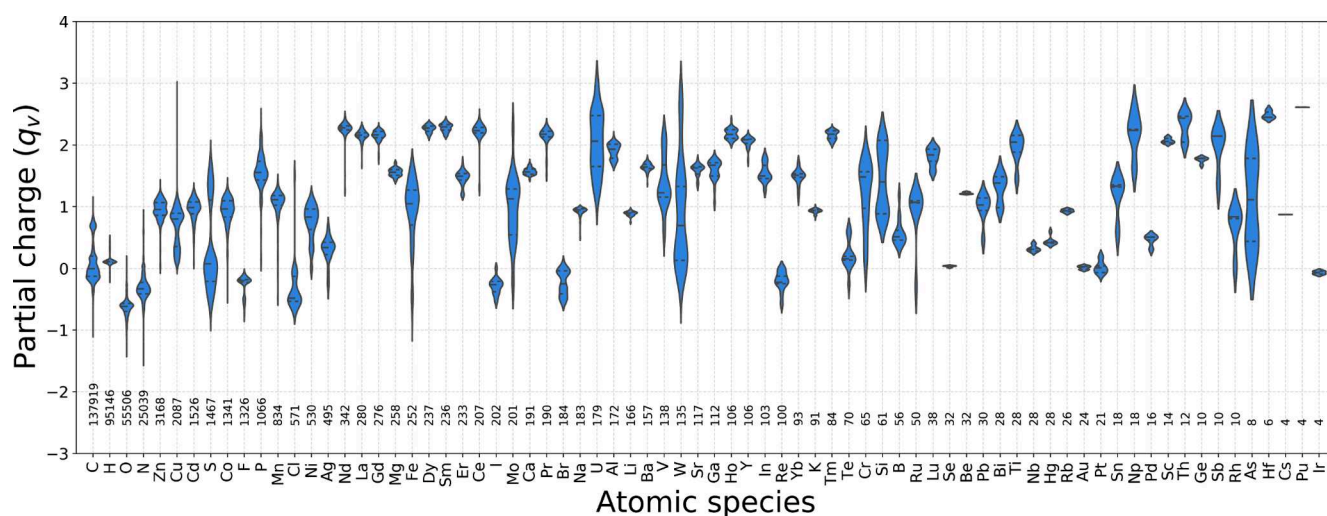
Figure 3. Distribution of partial charges, grouped by element, in the 2266 charge-labeled MOFs[45] comprising our train, validation, and test set, visualized by violins. The number on the bottom is the total number of atoms in the data set, grouped by element.

$$\hat{q}_v = \wp_v - \frac{\sigma_v^2}{\sum_{u=1}^{n} \sigma_u^2} \sum_{u=1}^{n} \wp_u \qquad (11)$$

Interpreting the mean of the Gaussian in eq 7, $\wp_v$, as a before-constraint charge assignment for node v lends a useful interpretation of eq 11. To enforce charge neutrality, eq 11 adjusts the before-constraint charges by distributing the negative of the net before-constraint charge $\sum_{u=1}^{n} \wp_u$ to each atom in proportion to its variance σ_v^2. The idea is that, if the charge of an atom within its local bonding environment exhibits high variance among MOFs, then it should be given more slack when adjusting it to achieve charge neutrality. If all local bonding environments exhibit the same variance, eq 11 reduces to uniformly distributing the before-constraint excess charge among the atoms. Eq 11, the last layer of our MPNN architecture, can be viewed as the *charge-correction layer*. By jointly learning $\boldsymbol{\beta}_\wp$ and $\boldsymbol{\beta}_\sigma$, our network learns not only the typical charge of an atom within its given local bonding environment ($\wp_v$) but also the slack we should give the atom (σ_v^2) when adjusting its charge to achieve graph-level charge neutrality.

Interestingly, there is a direct analogy between the variance σ_v^2 of the charge on a given atom in its given local bonding environment and the hardness (the second derivative of energy with respect to charge) used to enforce charge neutrality in ref 40 (compare eq 14 in ref 40 with eq 11); indeed, a soft atom will be given more slack for adjusting its charge to enforce charge neutrality.

We remark that our MPNN assigns charges $\hat{q}_v$ that are invariant to uniform scaling of the variances σ_v, seen by eq 11; i.e., the variance of the charge on an atom within its local bonding environment *relative to other atoms in the MOF* affects its predicted charge, but the scale of the variances globally does not.

Training of the MPNN. We define the loss function l to train (i.e., identify the parameters of) our network as

$$l = \frac{1}{N} \sum_{m=1}^{M} \|\hat{\mathbf{q}}_m - \mathbf{q}_m\|_1 \qquad (12)$$

where N is the total number of nodes among all of the MOFs, M is the total number of MOFs, $\hat{\mathbf{q}}_m$ is the vector of charges on

atoms of MOF m predicted by the MPNN by eq 11, $\mathbf{q}_m$ is the vector of (taken as ground-truth) DFT-derived, DDEC charges,[45] and $\|\cdot\|_1$ is the L1 norm. The loss l in eq 12 is equivalent to the mean (over all nodes) absolute deviation (MAD) performance metric.

■ RESULTS

Here, we train the MPNN in Figure 1 and evaluate its performance. All computer codes (Python, Julia) to reproduce our work are available on Github at github.com/SimonEnsemble/mpn_charges.

The Train, Test, and Validation Data Sets. Nazarian et al.[45] provide 2932 MOF crystal structures with DFT-derived (PBE functional, DDEC method[33]) partial point charges assigned to each atom. We removed 607 duplicate MOFs (identified in ref 45) and nine invalid MOFs (identified manually). Furthermore, we automatically discarded MOF structures that, via our bonding algorithm, produced invalid bonding motifs (carbon atoms bonded to >4 atoms, hydrogen atoms bonded to >1 atom). See section S3. Remaining are 2266 charge-labeled MOFs. Figure 3 shows the distribution of the partial charges, grouped by chemical element; many elements exhibit a high variance in charge, hinting that assigning charges to each atom solely based on its chemical element, without consideration of its bonding environment, will not give satisfactory charges. Figure S1 shows the prevalence of chemical elements among the MOFs.

We randomly partitioned these 2266 charge-labeled MOFs[45] into training, validation, and test sets (70/10/20%). The training set is used to directly tune the model parameters (weights and biases) by minimizing the loss in eq 12 over training examples via stochastic gradient descent. The validation set is used for hyper-parameter selection to avoid overfitting. The test set provides an unbiased evaluation of the performance of a final model whose parameters were fit using the training data set. We randomly partitioned the set of MOFs into training, validation, and test sets as opposed to the set of atoms (within their local bonding environments) because charge neutrality, a graph-level constraint, is enforced internal to the architecture of the MPNN; consequently, the MPNN must receive the entire crystal graph as input. We also note that chemical elements {Se, Hf, Cs, Pu, Ir} appear in only one

MOF. Instead of discarding these MOFs containing these rare elements, we elected to place them in our training set, with the justification that we can learn about charges on other atoms from these MOFs.

Training and Hyper-parameter Tuning. We used the open source PyTorch[77] machine learning library to construct and train our MPNN.

To minimize the loss l (see eq 12) during training, we use stochastic gradient descent (the Adam optimizer with a learning rate of 0.005) and a batch size of 32 graphs. We use the validation set for early stopping to avoid overtraining. This is achieved by, while training, continuously monitoring the model performance on the validation set. Once no improvement is observed for 100 epochs, we stop the training and output the model with the best performance on the validation set. See Figure S3 for an example learning curve.

The dimension of the element embeddings, r, dimension of the hidden node features, k, and number of message passing time steps, T, are hyper-parameters of our MPNN. The dimension of the message in eq 5 is also a hyper-parameter, but we fixed it to be equal to k by specifying B_m as a square matrix. We explored hyper-parameter space by changing one while holding the others fixed. See section S6. We found the performance of the MPNN to be largely insensitive to k and r for, respectively, $k \geq 30$ and $r \geq 8$ (for fixed $r = 10$ and $k = 30$, respectively, and $T = 4$; see Figure S2). On the other hand, we found the MPNN performance to be sensitive to T, which we discuss later. Based on our empirical hyper-parameter exploration, we select $T = 4$, $r = 10$, and $k = 30$ since these hyperparameters led to the best performance on the validation data set. See Table S1.

Performance. We evaluate the performance of our MPNN using the mean absolute deviation (MAD) over all nodes, equal to the loss in eq 12. For comparison, we consider the following benchmark models: (i) all charges are zero ($\hat{q}_v = 0$, $\forall v$), (ii) the charge of each atom is equal to the mean charge of atoms of that chemical element, with charge neutrality enforced by distributing excess charge among the atoms (a) uniformly or (b) proportional to the variance of charges of that chemical element in the training set, much like in eq 11, and (iii) the I-QEq[54] charge equilibration method (MAD reported in ref 36). Table 2 summarizes the performance of our trained

Table 2. Performance Benchmarks[a]

method (charge neutrality enforcement)	MAD mean (std)
$\hat{q}_v = 0$, $\forall v \in V$	0.324 (7e−3)
element-mean (uniform distn of excess charge)	0.154 (2e−3)
element-mean (variance-based distn of excess charge)	0.153 (2e−3)
I-QEq[36,54]	0.118[b]
MPNN (uniform distn of excess charge)	0.026 (8e−4)
MPNN (variance-based distn of excess charge)	0.025 (5e−4)

[a]The mean absolute deviation (MAD) on the test set (see eq 12) for different charge assignment models/strategies. [b]214 MOFs were excluded in this analysis by Ongari et al.,[36] and this is for the entire test, train, and validation set.

MPNN and these benchmark models. Results are the average of 10 training/testing sessions with different (random) training, validation, and test splits. Our MPNN outperforms all baseline models, including the charge equilibration variant I-QEq,[54] which was the most consistent with DFT-derived, DDEC charges in the study by Ongari et al.[36] The MAD of our

MPNN-assigned charges from the DFT-derived, DDEC (taken as ground truth) charges[45] is 0.025, a factor of 4 lower than the MAD associated with I-QEq The parity plot in Figure 4
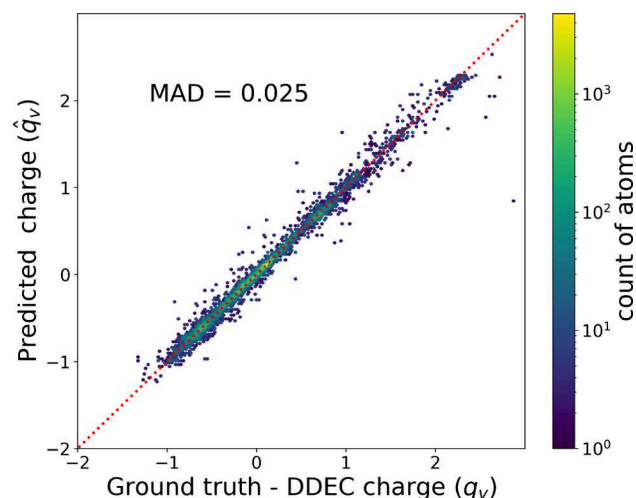
Figure 4. Parity plot showing a 2D histogram of the predicted charge $\hat{q}_v$ against the DFT-derived, DDEC charge q_v (treated as ground truth) for atoms in MOFs belonging to the test set. Color indicates the number of atoms that fall in that bin. Diagonal line shows equality.

visualizes the joint distribution of the predicted charge $\hat{q}_v$ by our MPNN and the DFT-derived DDEC charge q_v; the density hugs the diagonal line. We investigated the outlier in Figure 4 (center right) and found it to be a Cu atom with an outlying charge among Cu atoms in the DFT-derived, DDEC charge-assigned MOFs (see section S11).

Next, we measured the runtime of MOF charge assignment by our MPNN model to place it in the broader scope of partial charge assignment methods. This included reading the .cif file, inferring/constructing the crystal graph, passing the graph through the MPNN, and writing a new .cif file with MPNN-assigned charges—but not including precompilation in Julia—MPNN charge assignment to a single MOF is expected to take ~3 s. The runtime of MPNN-charge assignment is the same order of magnitude as charge equilibration (e.g., EQeq[78]) but orders of magnitude smaller than DFT calculations. See section S12 for details.

We further judge the quality of our MPNN charges by comparing simulated carbon dioxide adsorption in MOFs with (1) MPNN charges, (2) DFT-derived, DDEC charges,[45] (3) I-Qeq charges,[54] and (4) zero charges ($\hat{q}_v = 0$, $\forall v \in \mathcal{V}$). Importantly, (a) the molecular model we used for CO_2 contains point charges[26] (not assigned by our MPNN) so that CO_2-MOF electrostatic interactions contribute to simulated adsorption, and (b) we included in this analysis only MOFs that belong to the test set of the MPNN training session so as to judge generalization error. We found that simulated CO_2 Henry coefficients (298 K) in the MPNN-charge-assigned MOFs are (1) well-correlated (Spearman's rank correlation coefficient, $\rho_r = 0.96$) with the Henry coefficients in the DFT-derived, DDEC charge-assigned MOFs and (2) more correlated with those in the MOFs with DFT-derived, DDEC-charges than when I-QEq is used ($\rho_r = 0.81$). See Figure S10. Moreover, emphasizing the importance of accurate charge assignment for predicting adsorption in MOFs via molecular simulations, the CO_2 Henry coefficients in MOFs without charges were poorly

correlated ($\rho_r = 0.53$) with those in the DFT-derived, DDEC-charge-assigned MOFs. See section S13 for details.

Effect of the Number of Message Passing Time Steps, *T*. We investigate the effect of the number of message passing time steps, T, on the MPNN performance because, as Figure 2 shows, T determines the scope of the local bonding environment of node v encoded in $\mathbf{h}_v^{(T)}$ and used to predict the charge on node v.

Figure 5 shows the performance of our MPNN as T changes (with $r = 10$, $k = 30$ fixed). Without a message passing layer (T
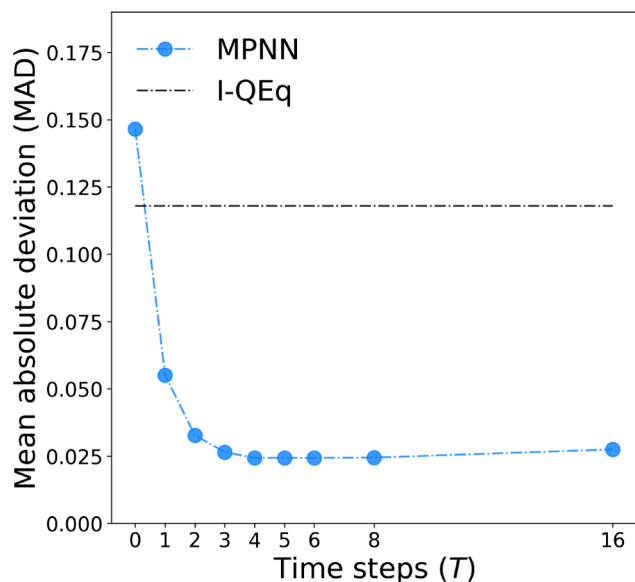

Figure 5. Performance of the MPNN (MAD on test set) as the number of message passing time steps, T, changes ($r = 10$, $k = 30$ fixed). As T increases, a broader view of the local bonding environment of each atom is used to predict its charge.

$= 0$), information is not passed between neighboring nodes, and the neural network learns to assign charge based only on the chemical element of the atom, irrespective of the atoms to which it is bonded. The MAD for $T = 0$ is 0.15, larger than the I-QEq method. Introducing a single message passing iteration ($T = 1$) to bring in information from immediate neighbors significantly improves the prediction ($T = 1$ MAD is 0.06). As we increase T, each node receives information from a longer geodesic distance; $\mathbf{h}_v^{(T)}$ encodes a broader scope of the bonding environment of the node; and, until $T = 4$, the predictions improve, albeit with diminishing returns. Increasing T beyond 4 slightly diminishes the performance, suggesting that the most useful information for charge prediction can be somewhat localized. Specifically, all nodes within a geodesic distance of $T = 4$ of a node appear sufficient for producing a quality prediction of the charge, albeit we fixed r and k for this analysis. Using overly large T can potentially lead to too much focus on the global structure of the MOF, diluting the useful local information. Figure 5 supports our underlying hypothesis that the charge of any given atom in a MOF is largely dictated by its identity and local bonding environment.

Latent Space of Chemical Elements. The chemical element embedding in eq 3 maps each chemical element into a low-dimensional, dense, information-rich representation of the chemical elements for initializing the hidden node features. To verify the MPNN has learned a meaningful element embedding, we visualize these $r = 10$-dimensional element

embeddings $\mathbf{x}_v^e$ via the Uniform Manifold Approximation and Projection (UMAP)[79] (hyper-parameters; number of neighbors, 8; minimum distance, 0.05). UMAP is a dimension reduction technique that aims to keep local and global structures exhibited by the data (in the high-dimensional space) intact in the low-dimensional representation. Figure 6
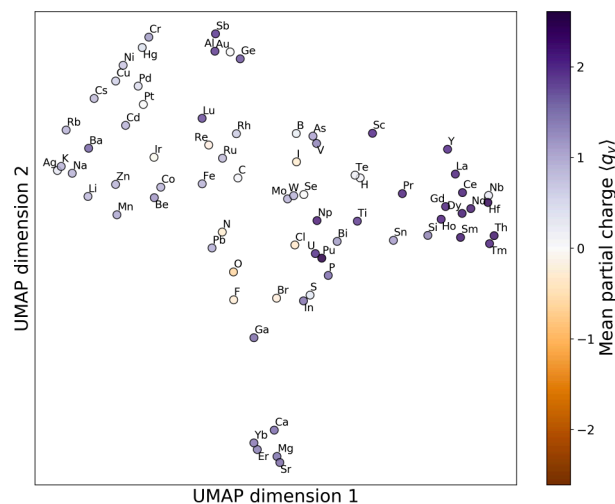
Figure 6. Visualization of the embeddings of the chemical elements learned by our MPNN and how they encode the mean charge of the element. We used UMAP[79] to reduce the dimension of the $r = 10$-dimensional embeddings $\mathbf{x}_v^e$ to the two dimensions shown here. Each point, representing the embedding of a chemical element, is colored according to the mean partial charge of that element in the training set. Nearby chemical elements tend to have similar charges.

visualizes the 2D embedding of each of the 74 chemical elements in the MOFs, colored by the average charge on that element in the DFT-derived, DDEC charge-assigned MOFs.[45] Judging from how nearby chemical elements tend to have a similar mean charge, it appears that the learned element embeddings indeed are encoding information predictive of partial charge. Interestingly, although the clustering according to the family in the periodic table to which the elements belong (see Figure S5) is not as prominent, as according to the mean partial charge, some clusters are recovered. For example, the alkali earth metals {Mg, Ca, Sr}, the alkali metals {Li, Na, K, Rb, Cs}, halogens {Br, Cl, F, I}, and many lanthanoids are clustered together, while the other periodic table families are more scattered.

Enforcement of Charge Neutrality. We enforced charge neutrality on a given MOF through the design of our MPNN, by treating the charge on each atom within its local bonding environment as a conditionally independent random variable (eq 7), predicting the mean and variance of this distribution (eqs 8 and 9), and using maximum likelihood to estimate the charges on the atoms of the MOF while satisfying charge neutrality (eq 11). A simpler strategy to enforce charge neutrality is to use a single (shared) neuron to directly predict a before-constraint charge on each atom from the learned node representation $\mathbf{h}_v^{(T)}$, then uniformly distribute the excess before-constraint charge on the MOF (see section S7). The strategy to uniformly distribute excess before-constraint charge does not account for the tendency of some atoms within certain local bonding environments to vary in their charge more than others; still, interestingly, this crude method of

enforcing charge neutrality in the MPNN suffered in performance only marginally, with a MAD of 0.026.

Figure S4 shows the distribution of the excess preliminary charge per atom, $\frac{1}{n}\sum_v \varphi_v$, among the MOFs for both strategies of enforcing charge neutrality. The standard deviation of $\frac{1}{n}\sum_v \varphi_v$ among the MOFs, with φ_v computed using eq 8, is only ≈ 0.02. This shows that relatively little charge neutrality correction of the before-constraint charges φ_v is needed (compare 0.02 with the MADs in Table 2); i.e., despite predicting charges based on the *local* bonding environment of the atoms, the MPNN outputs before-constraint charges φ_v that are "close" to satisfying the graph-level (global) charge neutrality constraint.

Deployment. For deployment, we retrained our MPNN (with fixed hyper-parameters) using more training examples (2040 MOFs plus 226 MOFs used for validation) to maximize its accuracy (setting aside a test set was necessary only for unbiased performance evaluation against different models). Note our MPNN (i) cannot predict charges for MOFs that contain chemical elements outside the set of 74 elements included in the DFT-derived, DDEC-charge assigned MOFs[45] we used for training and (ii) refrains from predicting charges on the elements {Se, Hf, Cs, Pu, Ir} since training examples were too scarce to have confidence in predictions for these elements. See Table S6 for the list of viable chemical elements.

Public Availability. Our deployment-ready MPNN model and our code to convert MOF crystal structures to graphs are available on Github (github.com/SimonEnsemble/mpn_charges) so the computational MOF community can assign high-quality charges to MOFs without performing computationally expensive electronic structure calculations. Moreover, our MPNN can easily handle MOFs with a large number of atoms, in contrast to periodic electronic structure calculations. To streamline and facilitate MPNN-charge assignment to a new MOF, we provide: (i) a single script that (a) reads in a .cif file of a MOF, (b) constructs the crystal graph (i.e., infers the bonds between atoms) and node feature matrix representing the MOF, (c) passes the crystal graph and node feature matrix to our trained MPNN model, which then assigns charges to the atoms of the MOF, and finally, (d) writes a MPNN-charge-assigned .cif file of the MOF for input to molecular simulation software and (ii) a Docker image with the required source code, libraries, and dependencies to execute this script in an isolated Docker container without the hassle of installing many libraries under dependency constraints.

MPNN-Charge-Assigned CoRE MOFs. The updated computation-ready, experimental (CoRE) MOF data set[7] contains ca. 14 000 structures, the majority of which are not present in the DDEC-charge-assigned set of Nazarian et al.[45] We used our deployment-ready MPNN to assign charges to each MOF in the v2 CoRE MOF database. To facilitate the use of these MPNN-assigned charges in molecular simulation studies, we provide .cif files on Github of the MPNN-charge-assigned v2 CoRE MOF structures.

As a caveat, the v2 CoRE MOF data set is partitioned into two separate subsets based on the extent of solvent removal: (1) both bound and free solvent molecules removed and (2) only free solvent removed. The charge-assigned MOFs of Nazarian et al.,[45] upon which we trained our MPNN, are based on structures in the v1 CoRE MOF data set,[46] where both free and bound solvent molecules were removed. Consequently, the charge predictions by the MPNN may be less accurate on

the subset of the v2 CoRE MOFs where only free solvent molecules were removed.

■ DISCUSSION

Remarks on the MPNN Charge Assignment Model. We emphasize that our MPNN is trained to assign a Gaussian distribution (with a mean and variance) to the charge of an atom given its atomic species *and* its specific local bonding environment. Notably, two atoms of the same species could have very different charge distributions if they appear in substantially different local bonding environments. Conversely, atoms of different species, if placed in similar local environments, could have a similar charge distribution. For this reason, knowledge could be transferred from one atomic species to another if their local environments are similar.

We remark that, given two atoms in the unit cell of a given MOF that are (i) the same chemical element and (ii) symmetrically equivalent, our MPNN will assign identical charges to them, as desired. The reason is that these two atoms will possess identical local bonding environments, and thus, our MPNN will output the same (φ_v, σ_v^2) and then the same charge $\hat{q}_v$ on each.

Charge Neutrality as a Soft Constraint. We enforced charge neutrality as a graph-level, *hard* constraint through the design of our MPNN architecture in Figure 1. Alternatively, charge-neutrality of a MOF could be encouraged as a *soft* constraint by adding a penalty term to the loss function in eq 12 (inside the sum) that increases monotonically with the absolute value of the net predicted charge on MOF m. A weight on the penalty term then balances the contributions to the loss from a constraint violation and from the mean absolute deviation. If the weight is too small, the constraint is ignored. If the weight is too large, the constraint is (more) satisfied, at the expense of the mean absolute deviation. However, such a soft constraint does not guarantee exact charge neutrality of each MOF, where as our hard constraint does (well, to numerical precision).

Limitations and Future Directions. Notably, machine learning models can perform differently when employed on data drawn from a different distribution than the training data set. The training MOFs used for this MPNN are experimentally synthesized MOFs.[45,46] Consequently, we are confident that our MPNN will perform well on experimentally synthesized MOFs. However, caution is warranted when using the MPNN on hypothetical MOFs sampled from a dramatically different distribution over MOF-space. For example, if hypothetical MOFs are constructed from elements in atomic environments that are rare in our training set of MOFs, then the accuracy of our MPNN could be reduced from what we report here.

Because we convert each MOF crystal structure to an undirected graph (with node features but not edge features), our MPNN will assign the same charges to (a) all conformations of the same MOF and (b) all interpenetrated isomers[80] of a MOF. To expand on part a, consider MOFs whose structures are flexible and adopt different conformations depending on the temperature, imposed mechanical stress, and presence of adsorbed molecules.[81−83] To expand on part b, some MOFs form interpenetrated networks, and the level of interpenetration can be controlled.[80,84] Conceivably, the partial point charges that reproduce the electrostatic potential in the pores could differ depending on the conformation that the MOF adopts and its level of interpenetration. Our MPNN,

however, would assign the same charges to the atoms of a MOF regardless of its conformation or level of inter-penetration, since the graph representations of local bonding environments are invariant to flexing and interpenetration. To instead learn charges dependent on the conformation of and level of interpenetration in a MOF, we can represent the MOF as a complete graph with pairwise atomic (periodic) distances as edge features, as described further below. Information about the atomic coordinates of the atoms in the MOF, up to a translation and rotation, would then be input to the MPNN.

We mention a few ways to include more information about the structure of a MOF in its representation that is input to the MPNN in order to, potentially, improve the predictions. First, while we elected to label each node of the crystal graph representing the MOF with the atomic species it represents, instead, we could supply more detailed node labels that, e.g., distinguish between aliphatic and aromatic carbon atoms. However, given T is sufficiently large, the MPNN can infer from the crystal graph, e.g., if a carbon atom is a member of an aromatic ring. Second, we could include edge features in our representation of the MOF. MPNNs can use edge features in the message passing phase.[51] As examples, the edge feature for edge e_{uv} could include (i) the order of the bond between atom u and v[58] and/or (ii) the periodic distance between atom u and v. Expanding on the latter, we could encode the 3D coordinates of the MOF atoms, up to a rigid transformation, by including the pairwise atomic distance matrix in the representation of the MOF.[85] In this view, the MOF is a *complete* graph with each edge e_{uv} labeled with the periodic distance between atom u and v. Perhaps, there could be two classes of edges: one for bonded atoms; another for nonbonded atoms.

CONCLUSIONS

We developed and trained a message passing neural network (MPNN)[51] to, in an end-to-end manner, learn representations of the local bonding environments of atoms within MOFs and, from these representations, predict the partial charges on the atoms of a MOF under a charge neutral constraint. The crystalline structure of the MOF, represented as an undirected graph with node features encoding the chemical elements, is directly input to the MPNN. The MPNN constructs features of the local bonding environments by sequentially passing information between bonded atoms. We trained and evaluated the performance of our MPNN by leveraging 2266 DFT-derived DDEC charge-labeled MOFs.[45] Our MPNN accurately predicts the partial charges on MOFs (mean absolute deviation from DDEC charges on test set, 0.025) while incurring orders of magnitude lower computational cost (runtime of MPNN charge assignment: ∼3 s) than performing electronic structure calculations and deriving charges from the electron density/electrostatic potential. We make our code and trained MPNN model openly available to enable more accurate virtual screenings of thousands of MOFs, via molecular simulations using atomistic force fields,[8] for their adsorption-based applications in gas storage, separation/purification, and sensing. For convenience, we provide MPNN-charge-labeled v2 computation-ready, experimental MOF structures[7] in the widely used .cif format.

ASSOCIATED CONTENT

ⓈI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcc.0c04903.

> Prevalence of elements in the training, testing, and validation set, modication of covalent radii for graph creation, manual removal of MOFs, gated recurrent unit, correction for charge neutrality via maximum likelihood estimation, hyperparameters and training, uniform distribution of excess charge, interpreting the elemental embedding, element types in deployment, distribution of predicted variance grouped by element, investigation into outlier in parity plot, runtime performance bench-mark, and molecular simulations of carbon dioxide adsorption (PDF)

AUTHOR INFORMATION

Corresponding Authors

Cory M. Simon − *School of Chemical, Biological, and Environmental Engineering, Oregon State University, Corvallis, Oregon 97331, United States;* ⓞ orcid.org/0000-0002-8181-9178; Email: Cory.Simon@oregonstate.edu

Xiaoli Fern − *School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, Oregon 97331, United States;* Email: xfern@eecs.oregonstate.edu

Authors

Ali Raza − *School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, Oregon 97331, United States*

Arni Sturluson − *School of Chemical, Biological, and Environmental Engineering, Oregon State University, Corvallis, Oregon 97331, United States*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jpcc.0c04903

Notes

The authors declare no competing financial interest.
Our code, MPNN model, and MPNN-charge-assigned v2 CoRE MOFs are available at github.com/SimonEnsemble/mpn_charges.

REFERENCES

(1) Furukawa, H.; Cordova, K. E.; O'Keeffe, M.; Yaghi, O. M. The chemistry and applications of metal-organic frameworks. *Science* **2013**, *341*, 1230444.

(2) Schoedel, A.; Ji, Z.; Yaghi, O. M. The role of metal−organic frameworks in a carbon-neutral energy cycle. *Nature Energy* **2016**, *1*, 1−13.

(3) Mueller, U.; Schubert, M.; Teich, F.; Puetter, H.; Schierle-Arndt, K.; Pastre, J. Metal−organic frameworks—prospective industrial applications. *J. Mater. Chem.* **2006**, *16*, 626−636.

(4) Kreno, L. E.; Leong, K.; Farha, O. K.; Allendorf, M.; Van Duyne, R. P.; Hupp, J. T. Metal−organic framework materials as chemical sensors. *Chem. Rev.* **2012**, *112*, 1105−1125.

(5) Wang, C.; Liu, D.; Lin, W. Metal−organic frameworks as a tunable platform for designing functional molecular materials. *J. Am. Chem. Soc.* **2013**, *135*, 13222−13234.

(6) Moghadam, P. Z.; Li, A.; Wiggin, S. B.; Tao, A.; Maloney, A. G.; 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*, 2618−2625.

(7) Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S.; et al. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal−Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data* **2019**, *64*, 5985−5998.

(8) 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.; et al. 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*, 1082−1121.

(9) Boyd, P. G.; Lee, Y.; Smit, B. Computational development of the nanoporous materials genome. *Nature Reviews Materials* **2017**, *2*, 1−15.

(10) Cho, E. H.; Lyu, Q.; Lin, L.-C. Computational discovery of nanoporous materials for energy-and environment-related applications. *Mol. Simul.* **2019**, *45*, 1122−1147.

(11) 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.; et al. Data-driven design of metal−organic frameworks for wet flue gas CO2 capture. *Nature* **2019**, *576*, 253−256.

(12) Bucior, B. J.; Bobbitt, N. S.; Islamoglu, T.; Goswami, S.; Gopalan, A.; Yildirim, T.; Farha, O. K.; Bagheri, N.; Snurr, R. Q. Energy-based descriptors to rapidly predict hydrogen storage in metal−organic frameworks. *Molecular Systems Design & Engineering* **2019**, *4*, 162−174.

(13) 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*, ncomms11831.

(14) Moghadam, P. Z.; Islamoglu, T.; Goswami, S.; Exley, J.; Fantham, M.; Kaminski, C. F.; Snurr, R. Q.; Farha, O. K.; Fairen-Jimenez, D. Computer-aided discovery of a metal−organic framework with superior oxygen uptake. *Nat. Commun.* **2018**, *9*, 1−8.

(15) Matito-Martos, I.; Moghadam, P. Z.; Li, A.; Colombo, V.; Navarro, J. A.; Calero, S.; Fairen-Jimenez, D. Discovery of an optimal porous crystalline material for the capture of chemical warfare agents. *Chem. Mater.* **2018**, *30*, 4571−4579.

(16) Hulvey, Z.; Lawler, K. V.; Qiao, Z.; Zhou, J.; Fairen-Jimenez, D.; Snurr, R. Q.; Ushakov, S. V.; Navrotsky, A.; Brown, C. M.; Forster, P. M. Noble gas adsorption in copper trimesate, HKUST-1: an experimental and computational study. *J. Phys. Chem. C* **2013**, *117*, 20116−20126.

(17) Dubbeldam, D.; Frost, H.; Walton, K. S.; Snurr, R. Q. Molecular simulation of adsorption sites of light gases in the metal-organic framework IRMOF-1. *Fluid Phase Equilib.* **2007**, *261*, 152−161.

(18) Krause, S.; Bon, V.; Senkovska, I.; Stoeck, U.; Wallacher, D.; Többens, D. M.; Zander, S.; Pillai, R. S.; Maurin, G.; Coudert, F.-X.; et al. A pressure-amplifying framework material with negative gas adsorption transitions. *Nature* **2016**, *532*, 348−352.

(19) Elsaidi, S. K.; Mohamed, M. H.; Simon, C. M.; Braun, E.; Pham, T.; Forrest, K. A.; Xu, W.; Banerjee, D.; Space, B.; Zaworotko, M. J.; et al. Effect of ring rotation upon gas adsorption in SIFSIX-3-M (M= Fe, Ni) pillared square grid networks. *Chemical Science* **2017**, *8*, 2373−2380.

(20) Wilmer, C. E.; Farha, O. K.; Bae, Y.-S.; Hupp, J. T.; Snurr, R. Q. Structure−property relationships of porous materials for carbon dioxide separation and capture. *Energy Environ. Sci.* **2012**, *5*, 9849−9856.

(21) Sikora, B. J.; Wilmer, C. E.; Greenfield, M. L.; Snurr, R. Q. Thermodynamic analysis of Xe/Kr selectivity in over 137000 hypothetical metal−organic frameworks. *Chemical Science* **2012**, *3*, 2217−2223.

(22) Sladekova, K.; Campbell, C.; Grant, C.; Fletcher, A. J.; Gomes, J. R.; Jorge, M. The effect of atomic point charges on adsorption isotherms of CO2 and water in metal organic frameworks. *Adsorption* **2020**, *26*, 663−685.

(23) Zheng, C.; Liu, D.; Yang, Q.; Zhong, C.; Mi, J. Computational study on the influences of framework charges on CO2 uptake in metal-organic frameworks. *Ind. Eng. Chem. Res.* **2009**, *48*, 10479−10484.

(24) Walton, K. S.; Millward, A. R.; Dubbeldam, D.; Frost, H.; Low, J. J.; Yaghi, O. M.; Snurr, R. Q. Understanding inflections and steps in carbon dioxide adsorption isotherms in metal-organic frameworks. *J. Am. Chem. Soc.* **2008**, *130*, 406−407.

(25) Hamad, S.; Balestra, S. R.; Bueno-Perez, R.; Calero, S.; Ruiz-Salvador, A. R. Atomic charges for modeling metal−organic frameworks: Why and how. *J. Solid State Chem.* **2015**, *223*, 144−151.

(26) Potoff, J. J.; Siepmann, J. I. Vapor−liquid equilibria of mixtures containing alkanes, carbon dioxide, and nitrogen. *AIChE J.* **2001**, *47*, 1676−1682.

(27) Dubbeldam, D.; Walton, K. S.; Vlugt, T. J.; Calero, S. Design, parameterization, and implementation of atomic force fields for adsorption in nanoporous materials. *Advanced Theory and Simulations* **2019**, *2*, 1900135.

(28) Dubbeldam, D.; Torres-Knoop, A.; Walton, K. S. On the inner workings of Monte Carlo codes. *Mol. Simul.* **2013**, *39*, 1253−1292.

(29) Watanabe, T.; Manz, T. A.; Sholl, D. S. Accurate treatment of electrostatics during molecular adsorption in nanoporous crystals without assigning point charges to framework atoms. *J. Phys. Chem. C* **2011**, *115*, 4824−4836.

(30) Li, W.; Rao, Z.; Chung, Y. G.; Li, S. The Role of Partial Atomic Charge Assignment Methods on the Computational Screening of Metal-Organic Frameworks for CO2 Capture under Humid Conditions. *ChemistrySelect* **2017**, *2*, 9458−9465.

(31) Altintas, C.; Keskin, S. Role of partial charge assignment methods in high-throughput screening of MOF adsorbents and membranes for CO2/CH4 separation. *Molecular Systems Design & Engineering* **2020**, *5*, 532−543.

(32) Campañá, C.; Mussard, B.; Woo, T. K. Electrostatic potential derived atomic charges for periodic systems using a modified error functional. *J. Chem. Theory Comput.* **2009**, *5*, 2866−2878.

(33) Manz, T. A.; Sholl, D. S. Chemically meaningful atomic charges that reproduce the electrostatic potential in periodic and nonperiodic materials. *J. Chem. Theory Comput.* **2010**, *6*, 2455−2468.

(34) Argueta, E.; Shaji, J.; Gopalan, A.; Liao, P.; Snurr, R. Q.; Gómez-Gualdrón, D. A. Molecular Building Block-Based Electronic Charges for High-Throughput Screening of Metal−Organic Frameworks for Adsorption Applications. *J. Chem. Theory Comput.* **2018**, *14*, 365−376.

(35) Rappe, A. K.; Goddard, W. A., III Charge equilibration for molecular dynamics simulations. *J. Phys. Chem.* **1991**, *95*, 3358−3363.

(36) 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*, 382−401.

(37) Bleiziffer, P.; Schaller, K.; Riniker, S. Machine learning of partial charges derived from high-quality quantum-mechanical calculations. *J. Chem. Inf. Model.* **2018**, *58*, 579−590.

(38) Rai, B. K.; Bakken, G. A. Fast and accurate generation of ab initio quality atomic charges using nonparametric statistical regression. *J. Comput. Chem.* **2013**, *34*, 1661−1671.

(39) Martin, R.; Heider, D. ContraDRG: Automatic partial charge prediction by Machine Learning. *Front. Genet.* **2019**, *10*, 990.

(40) Wang, Y.; Fass, J.; Stern, C. D.; Luo, K.; Chodera, J. Graph Nets for Partial Charge Prediction. *arXiv* **2019**, 1909.07903.

(41) Xu, Q.; Zhong, C. A general approach for estimating framework charges in metal-organic frameworks. *J. Phys. Chem. C* **2010**, *114*, 5035−5042.

(42) Zheng, C.; Zhong, C. Estimation of framework charges in covalent organic frameworks using connectivity-based atom contribution method. *J. Phys. Chem. C* **2010**, *114*, 9945−9951.

K

(43) Engler, M. S.; Caron, B.; Veen, L.; Geerke, D. P.; Mark, A. E.; Klau, G. W. Automated partial atomic charge assignment for drug-like molecules: a fast knapsack approach. *Algorithms Mol. Biol.* **2019**, *14*, 1.

(44) Zou, C.; Penley, D. R.; Cho, E. H.; Lin, L.-C. Efficient and Accurate Charge Assignments via Multi-Layer Connectivity-based Atom Contribution (m-CBAC) Approach. *J. Phys. Chem. C* **2020**, *124*, 11428.

(45) 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*, 785−793.

(46) 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*, 6185−6192.

(47) 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*, 83.

(48) Colón, Y. J.; Gómez-Gualdrón, D. A.; Snurr, R. Q. Topologically guided, automated construction of metal−organic frameworks and their evaluation for energy-related applications. *Cryst. Growth Des.* **2017**, *17*, 5801−5810.

(49) Boyd, P. G.; Woo, T. K. A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. *CrystEngComm* **2016**, *18*, 3777−3792.

(50) Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. S. A comprehensive survey on graph neural networks. *arXiv* **2019**, 1901.00596.

(51) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for quantum chemistry. *Proceedings of the 34th International Conference on Machine Learning* **2017**, *70*, 1263−1272.

(52) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.

(53) Collins, C. R.; Gordon, G. J.; Von Lilienfeld, O. A.; Yaron, D. J. Constant size descriptors for accurate machine learning models of molecular properties. *J. Chem. Phys.* **2018**, *148*, 241718.

(54) Wells, B. A.; De Bruin-Dickason, C.; Chaffee, A. L. Charge equilibration based on atomic ionization in metal−organic frameworks. *J. Phys. Chem. C* **2015**, *119*, 456−466.

(55) Nebgen, B.; Lubbers, N.; Smith, J. S.; Sifain, A. E.; Lokhov, A.; Isayev, O.; Roitberg, A. E.; Barros, K.; Tretiak, S. Transferable dynamic molecular charge assignment using deep neural networks. *J. Chem. Theory Comput.* **2018**, *14*, 4687−4698.

(56) Unke, O. T.; Meuwly, M. PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678−3693.

(57) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.

(58) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595−608.

(59) Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackerman, Z.; et al. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180*, 688−702.

(60) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. *Advances in Neural Information Processing Systems* **2015**, 2224−2232.

(61) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional embedding of attributed molecular graphs for physical property prediction. *J. Chem. Inf. Model.* **2017**, *57*, 1757−1772.

(62) Jørgensen, P. B.; Jacobsen, K. W.; Schmidt, M. N. Neural message passing with edge updates for predicting properties of molecules and materials. *arXiv* **2018**, 1806.03146.

(63) Tang, B.; Kramer, S. T.; Fang, M.; Qiu, Y.; Wu, Z.; Xu, D. A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. *J. Cheminf.* **2020**, *12*, 1−9.

(64) Sanchez-Lengeling, B.; Wei, J. N.; Lee, B. K.; Gerkin, R. C.; Aspuru-Guzik, A.; Wiltschko, A. B. Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules. *arXiv* **2019**, 1910.10685.

(65) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **2019**, *31*, 3564−3572.

(66) St. John, P. C.; Phillips, C.; Kemper, T. W.; Wilson, A. N.; Guan, Y.; Crowley, M. F.; Nimlos, M. R.; Larsen, R. E. Message-passing neural networks for high-throughput polymer screening. *J. Chem. Phys.* **2019**, *150*, 234111.

(67) Xie, T.; Grossman, J. C. Hierarchical visualization of materials space with graph convolutional neural networks. *J. Chem. Phys.* **2018**, *149*, 174111.

(68) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.

(69) Park, C. W.; Wolverton, C. Developing an improved Crystal Graph Convolutional Neural Network framework for accelerated materials discovery. *arXiv* **2019**, 1906.05267.

(70) Jin, W.; Coley, C.; Barzilay, R.; Jaakkola, T. Predicting organic reaction outcomes with Weisfeiler-Lehman network. *Advances in Neural Information Processing Systems* **2017**, 2607−2616.

(71) Do, K.; Tran, T.; Venkatesh, S. Graph transformation policy network for chemical reaction prediction. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2019**, 750−760.

(72) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **2017**, *8*, 1−12.

(73) Cordero, B.; Gómez, V.; Platero-Prats, A. E.; Revés, M.; Echeverría, J.; Cremades, E.; Barragán, F.; Alvarez, S. Covalent radii revisited. *Dalton Transactions* **2008**, 2832−2838.

(74) Virtanen, P.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261−272.

(75) Li, Y.; Tarlow, D.; Brockschmidt, M.; Zemel, R. Gated graph sequence neural networks. *arXiv* **2015**, 1511.05493.

(76) Cho, K.; Van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv* **2014**, 1409.1259.

(77) Paszke, A.; Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché Buc, F.; Fox, E.; Garnett, R. *Advances in Neural Information Processing Systems* **2019**, *32*, 8024−8035.

(78) Wilmer, C. E.; Kim, K. C.; Snurr, R. Q. An extended charge equilibration method. *J. Phys. Chem. Lett.* **2012**, *3*, 2506−2511.

(79) McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv* **2018**, 1802.03426.

(80) Jiang, H.-L.; Makal, T. A.; Zhou, H.-C. Interpenetration control in metal−organic frameworks for functional applications. *Coord. Chem. Rev.* **2013**, *257*, 2232−2249.

(81) Coudert, F.-X. Responsive metal−organic frameworks and framework materials: under pressure, taking the heat, in the spotlight, with friends. *Chem. Mater.* **2015**, *27*, 1905−1916.

(82) Ortiz, A. U.; Boutin, A.; Fuchs, A. H.; Coudert, F.-X. Metal−organic frameworks with wine-rack motif: What determines their flexibility and elastic properties? *J. Chem. Phys.* **2013**, *138*, 174703.

(83) Boutin, A.; Springuel-Huet, M.-A.; Nossov, A.; Gedeon, A.; Loiseau, T.; Volkringer, C.; Férey, G.; Coudert, F.-X.; Fuchs, A. H. Breathing Transitions in MIL-53 (Al) Metal−Organic Framework Upon Xenon Adsorption. *Angew. Chem., Int. Ed.* **2009**, *48*, 8314−8317.

(84) Eddaoudi, M.; Kim, J.; Rosi, N.; Vodak, D.; Wachter, J.; O'Keeffe, M.; Yaghi, O. M. Systematic design of pore size and functionality in isoreticular MOFs and their application in methane storage. *Science* **2002**, *295*, 469−472.

(85) Dokmanic, I.; Parhizkar, R.; Ranieri, J.; Vetterli, M. Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Processing Magazine* **2015**, *32*, 12−30.

(86) We focus on classical as opposed to quantum methods to describe adsorbate−MOF interactions because quantum methods are too computationally costly to be used for molecular simulations in thousands of MOFs.

(87) This electrostatic potential energy of interaction is usually computed via the Ewald summation[27,28] given that MOFs are modeled as periodic systems.

(88) Consider the case where we use an electronic structure calculation to obtain the periodic electrostatic potential at a 3D grid of points superimposing the unit cell of the MOF. Instead of translating this grid into a set of partial point charges that can reproduce it, directly interpolating this grid during a molecular simulation of adsorption[29] confers both (i) higher accuracy, as there may be model error in the translation of the grid into a set of partial point charges, and (ii) greater speed, as using Ewald summations[27,28] to compute the electrostatic potential created by the point charges is likely more computationally expensive than grid interpolation. Counter arguments are that (i) storing the 3D electrostatic potential grid will consume more disk space and memory during the simulation and (ii) if the MOF is flexible during the simulation, the (fixed) partial point charges that follow the atoms can still be used, whereas the grid cannot.