The Journal of Chemical Physics

OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features

Cite as: J. Chem. Phys. **153**, 124111 (2020); https://doi.org/10.1063/5.0021955
Submitted: 16 July 2020 . Accepted: 07 September 2020 . Published Online: 25 September 2020

Zhuoran Qiao, Matthew Welborn, Animashree Anandkumar, Frederick R. Manby, and Thomas F. Miller

COLLECTIONS

Note: This paper is part of the JCP Special Topic on Machine Learning Meets Chemical Physics.

EP This paper was selected as an Editor's Pick

View Online Export Citation CrossMark

New
SHFQA
Your Qubits. Measured.
Meet the next generation of quantum analyzers
Readout for up to 64 qubits
Operation at up to 8.5 GHz, mixer-calibration-free
Signal optimization with minimal latency
Find out more
Zurich Instruments

© 2020 Author(s).

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features

Cite as: J. Chem. Phys. **153**, 124111 (2020); doi: 10.1063/5.0021955
Submitted: 16 July 2020 • Accepted: 7 September 2020 •
Published Online: 25 September 2020

Zhuoran Qiao,[1] Matthew Welborn,[2] Animashree Anandkumar,[3] Frederick R. Manby,[2] and Thomas F. Miller III[1,2,a)]

AFFILIATIONS

[1] Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, USA
[2] Entos, Inc., 4470 W Sunset Blvd., Suite 107 PMB 94758, Los Angeles, California 90027, USA
[3] Division of Engineering and Applied Sciences, California Institute of Technology, Pasadena, California 91125, USA

Note: This paper is part of the JCP Special Topic on Machine Learning Meets Chemical Physics.
[a)] **Author to whom correspondence should be addressed:** tfm@caltech.edu and tom@entos.ai

ABSTRACT

We introduce a machine learning method in which energy solutions from the Schrödinger equation are predicted using symmetry adapted atomic orbital features and a graph neural-network architecture. OrbNet is shown to outperform existing methods in terms of learning efficiency and transferability for the prediction of density functional theory results while employing low-cost features that are obtained from semi-empirical electronic structure calculations. For applications to datasets of drug-like molecules, including QM7b-T, QM9, GDB-13-T, DrugBank, and the conformer benchmark dataset of Folmsbee and Hutchison [Int. J. Quantum Chem. (published online) (2020)], OrbNet predicts energies within chemical accuracy of density functional theory at a computational cost that is 1000-fold or more reduced.

Published under license by AIP Publishing. https://doi.org/10.1063/5.0021955

I. INTRODUCTION

The potential energy surface is the central quantity of interest in modeling of molecules and materials. The calculation of these energies with sufficient accuracy in chemical, biological, and materials systems is in many—but not all—cases adequately described at the level of density functional theory (DFT). However, due to its relatively high cost, the applicability of DFT is limited to either relatively small molecules or modest conformational sampling, at least in comparison to force-field and semi-empirical quantum mechanical theories. A major focus of machine learning (ML) for quantum chemistry has, therefore, been to improve the efficiency with which potential energies of molecular and materials systems can be predicted while preserving accuracy.

In the context of quantum chemistry, many applications have focused on the use of atom- or geometry-specific feature representations and kernel-based[1–9] or neural-network (NN) ML architectures.[10–23] Recent studies focus on the featurization of molecules in abstracted representations—such as quantum mechanical properties obtained from low-cost electronic structure calculations[24–28]—and the utilization of novel graph-based neural network[29–35] techniques to improve transferability and learning efficiency.

In this vein, we present a new approach (OrbNet) based on the featurization of molecules in terms of symmetry-adapted atomic orbitals (SAAOs) and the use of graph neural network methods for deep-learning quantum-mechanical properties.

We demonstrate the performance of the new method for the prediction of molecular properties, including the total and relative conformer energies for molecules in a range of datasets of organic and drug-like molecules. The method enables the prediction of molecular potential energy surfaces with full quantum mechanical accuracy while enabling vast reductions in computational cost; moreover, the method outperforms existing methods in terms of its training efficiency and transferable accuracy across diverse molecular systems.

II. METHOD

The target of this work is to machine learn a transferable mapping from input features' values $\{\mathbf{f}\}$ to the regression labels that are quantum mechanical energies,

$$E \approx E^{\mathrm{ML}}[\{\mathbf{f}\}]. \tag{1}$$

The key elements of OrbNet (Fig. 1) include the efficient evaluation of the features in the SAAO basis, the utilization of a graph neural-network architecture with edge and node attributes and message passing layers (MPLs), and a prediction phase that ensures extensivity of the resulting energies. We summarize these elements in the current section and discuss the relationship between OrbNet and other ML approaches. Although results in the current paper are presented for the mapping of features from semi-empirical-quality features to DFT-quality labels, the method is general with respect to the mean-field method used for features [i.e., also allowing for Hartree–Fock (HF) and DFT] and the level of theory used for generating labels (i.e., also allowing for coupled-cluster and other correlated-wavefunction method reference data).

A. SAAO features

Let $\{\phi^A_{n,l,m}\}$ be the set of atomic orbital (AO) basis functions with atom index A and the standard principal and angular momentum quantum numbers, n, l, and m. Let $\mathbf{C}$ be the corresponding molecular orbital coefficient matrix obtained from a mean-field electronic structure calculation, such as HF theory, DFT, or a semi-empirical method. The one-electron density matrix of the molecular system in the AO basis is then

$$P_{\mu\nu} = 2\sum_{i\in\mathrm{occ}} C_{\mu i}C_{\nu i} \tag{2}$$

(for a closed-shell system). We construct a rotationally invariant symmetry-adapted atomic-orbital (SAAO) basis $\{\hat{\phi}^A_{n,l,m}\}$ by diagonalizing diagonal density-matrix blocks associated with indices A, n, and l such that

$$\mathbf{P}^A_{nl}\mathbf{Y}^A_{nl} = \mathbf{Y}^A_{nl}\mathrm{diag}(\lambda^A_{nlm}), \tag{3}$$

where $[\mathbf{P}^A_{nl}]_{mm'} = P^A_{nlm,nlm'}$. For s orbitals ($l = 0$), this symmetrization procedure is obviously trivial and can be skipped. By construction, SAAOs are localized and consistent with respect to geometric perturbations of the molecule, and in contrast with localized molecular orbitals (LMOs) obtained from minimizing a localization objective function (e.g., Pipek–Mezey and Boys), SAAOs are obtained by a series of very small diagonalizations, without the need for an iterative procedure. The SAAO eigenvectors $\mathbf{Y}^A_{nl}$ are aggregated to form a block-diagonal transformation matrix $\mathbf{Y}$ that specifies the full transformation from AOs to SAAOs,

$$|\hat{\phi}_p\rangle = \sum_{\mu} Y_{\mu p}|\phi_\mu\rangle, \tag{4}$$

where μ and p index the AOs and SAAOs, respectively.

We employ ML features $\{\mathbf{f}\}$ comprised of tensors obtained by evaluating quantum-chemical operators in the SAAO basis. Hereafter, all quantum mechanical matrices will be assumed to be represented in the SAAO basis, including the density matrix $\mathbf{P}$ and the overlap matrix $\mathbf{S}$. Following our previous work,[24] the features include expectation values of the Fock ($\mathbf{F}$), Coulomb ($\mathbf{J}$), and

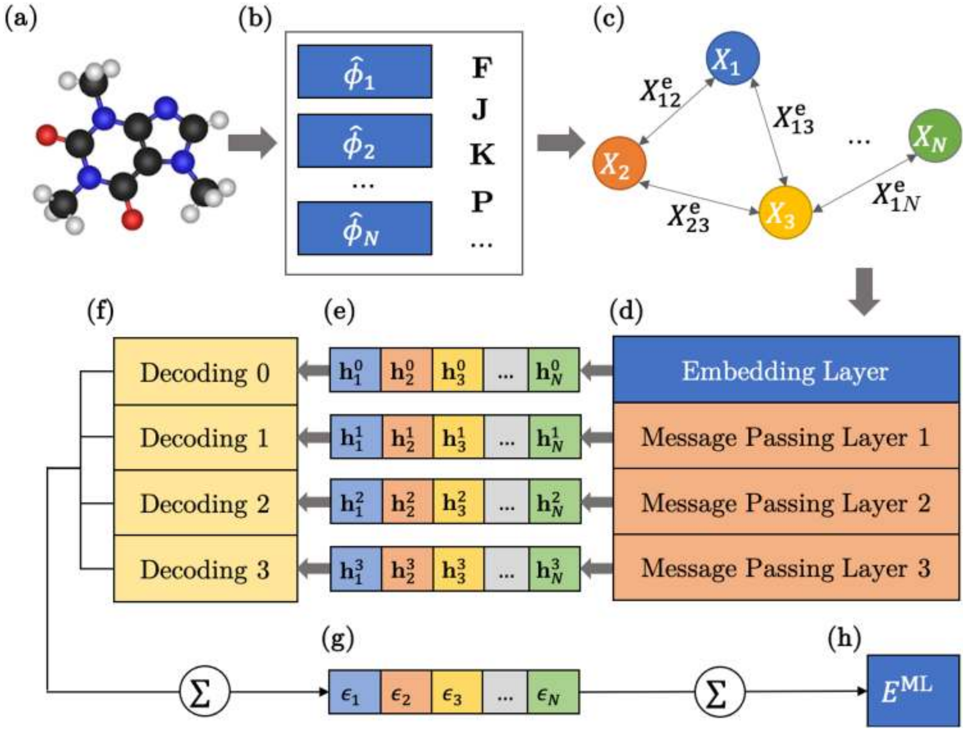
(a)
(b)
$\hat{\phi}_1$
$\hat{\phi}_2$
...
$\hat{\phi}_N$
F
J
K
P
...
(c)
X_1
X_2
X_3
X_N
X^{e}_{12}
X^{e}_{13}
X^{e}_{23}
X^{e}_{1N}
...
(f)
(e)
(d)
Decoding 0
$\mathbf{h}^0_1$ $\mathbf{h}^0_2$ $\mathbf{h}^0_3$ ... $\mathbf{h}^0_N$
Embedding Layer
Decoding 1
$\mathbf{h}^1_1$ $\mathbf{h}^1_2$ $\mathbf{h}^1_3$ ... $\mathbf{h}^1_N$
Message Passing Layer 1
Decoding 2
$\mathbf{h}^2_1$ $\mathbf{h}^2_2$ $\mathbf{h}^2_3$ ... $\mathbf{h}^2_N$
Message Passing Layer 2
Decoding 3
$\mathbf{h}^3_1$ $\mathbf{h}^3_2$ $\mathbf{h}^3_3$ ... $\mathbf{h}^3_N$
Message Passing Layer 3
(g)
(h)
Σ
ϵ_1 ϵ_2 ϵ_3 ... ϵ_N
Σ
E^{ML}

FIG. 1. Summary of the OrbNet workflow. (a) A low-cost mean-field electronic structure calculation is performed for the molecular system, and (b) the resulting SAAOs and the associated quantum operators are constructed. (c) An attributed graph representation is built with node and edge attributes corresponding to the diagonal and off-diagonal elements of the SAAO tensors. (d) The attributed graph is processed by the embedding layer and message passing layers to produce transformed node and edge attributes. (e) The transformed node attributes for the encoding layer and each message passing layer are extracted, and (f) they are passed to MPL-specific decoding networks. (g) The node-resolved energy contributions ε_u are obtained by summing the decoding network outputs node-wise, and (h) the final extensive energy prediction is obtained from a one-body summation over the nodes.

exchange (**K**) operators in the SAAO basis. In this work, we additionally include the SAAO density matrix, **P**, the orbital centroid distance matrix, **D**, the core Hamiltonian matrix, **H**, and the overlap matrix, **S**; other quantum-mechanical matrix elements are also possible for featurization.

B. Approximated Coulomb and exchange SAAO features

When a semi-empirical quantum chemical theory is employed, the computational bottleneck of SAAO feature generation becomes the **J** and **K** terms due to the need to compute four-index electron-repulsion integrals. We address this problem by introducing a generalized form of the Mataga–Nishimoto–Ohno–Klopman formula, as in the sTDA-xTB method,[36,37]

$$(pq|rs)^{\mathrm{MNOK}} = \sum_{A}\sum_{B} Q^{A}_{pq} Q^{B}_{rs} \gamma_{AB}. \tag{5}$$

Here, A and B are atom indices, p, q, r, and s are SAAO indices, and

$$\gamma^{\{\mathbf{J},\mathbf{K}\}}_{AB} = \left(\frac{1}{R_{AB}^{y_{\{\mathbf{J},\mathbf{K}\}}} + \eta^{-y_{\{\mathbf{J},\mathbf{K}\}}}}\right)^{1/y_{\{\mathbf{J},\mathbf{K}\}}}, \tag{6}$$

where R_{AB} is the distance between atoms A and B, η is the average chemical hardness for the atoms A and B, and $y_{\{\mathbf{J},\mathbf{K}\}}$ are the empirical parameters specifying the decay behavior of the damped interaction kernels, $\gamma^{\{\mathbf{J},\mathbf{K}\}}_{AB}$. In this work, we used $y_{\mathbf{J}} = 4$ and $y_{\mathbf{K}} = 10$ similar to which employed in the sTDA-RSH method.[38] The transition density Q^{A}_{pq} is calculated from a Löwdin population analysis,

$$Q^{A}_{pq} = \sum_{\mu\in A} Y'_{\mu p} Y'_{\mu q}, \tag{7}$$

where the pth column of $\mathbf{Y}' = \mathbf{Y}\mathbf{S}^{1/2}$ contains the expansion coefficients for the pth SAAO in the symmetrically orthgonalized AO basis. This yields approximated **J** and **K** matrices for featurization,

$$J^{\mathrm{MNOK}}_{pq} = (pp|qq)^{\mathrm{MNOK}} = \sum_{A,B} Q^{A}_{pp} Q^{B}_{qq} \gamma^{\mathbf{J}}_{AB}, \tag{8}$$

$$K^{\mathrm{MNOK}}_{pq} = (pq|pq)^{\mathrm{MNOK}} = \sum_{A,B} Q^{A}_{pq} Q^{B}_{pq} \gamma^{\mathbf{K}}_{AB}. \tag{9}$$

A naive implementation of Eqs. (8) and (9) is $\mathcal{O}(N^4)$, the leading asymptotic cost. However, this scaling may be reduced to $\mathcal{O}(N^2)$ with negligible loss of accuracy through a tight-binding approximation; for molecules in this study, the computation of $\mathbf{J}^{\mathrm{MNOK}}$ and $\mathbf{K}^{\mathrm{MNOK}}$ is not the leading order cost for feature generation, and such tight-binding approximation is thus not employed.

C. OrbNet

OrbNet encodes the molecular system as graph-structured data and utilizes a graph neural network (GNN) machine-learning architecture. The GNN represents data as an attributed graph $G(\mathbf{V}, \mathbf{E}, \mathbf{X}, \mathbf{X}^{\mathbf{e}})$, with nodes $\mathbf{V}$, edges $\mathbf{E}$, node attributes $\mathbf{X} : \mathbf{V} \to \mathbb{R}^{n\times d}$, and edge attributes $\mathbf{X}^{\mathbf{e}} : \mathbf{E} \to \mathbb{R}^{m\times e}$, where $n = |V|$, $m = |E|$, and d and e are the number of attributes per node and edge, respectively.

Specifically, OrbNet employs a graph representation for a molecular system in which node attributes correspond to diagonal SAAO features $X_u = [F_{uu}, J_{uu}, K_{uu}, P_{uu}, H_{uu}]$ and edge attributes correspond to off-diagonal SAAO features $X^{\mathrm{e}}_{uv} = [F_{uv}, J_{uv}, K_{uv}, D_{uv}, P_{uv}, S_{uv}, H_{uv}]$. By introducing an edge attribute cut-off value for edges to be included, non-interacting molecular systems separated at infinite distance are encoded as disconnected graphs, thereby satisfying size-consistency.

The model capacity is enhanced by introducing nonlinear input-feature transformations to the graph representation via radial basis functions,

$$\mathbf{h}^{\mathrm{RBF}}_{u} = [\phi^{\mathrm{h}}_{1}(\tilde{X}_u), \phi^{\mathrm{h}}_{2}(\tilde{X}_u), \ldots, \phi^{\mathrm{h}}_{n_{\mathrm{r}}}(\tilde{X}_u)], \tag{10}$$

$$\mathbf{e}^{\mathrm{RBF}}_{uv} = [\phi^{\mathrm{e}}_{1}(\tilde{X}^{\mathrm{e}}_{uv}), \phi^{\mathrm{e}}_{2}(\tilde{X}^{\mathrm{e}}_{uv}), \ldots, \phi^{\mathrm{e}}_{m_{\mathrm{r}}}(\tilde{X}^{\mathrm{e}}_{uv})], \tag{11}$$

where $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{X}}^{\mathbf{e}}$ are $n \times d$ and $m \times e$ matrices with pre-normalized attributes, as described in Sec. III. Sine basis functions $\phi^{\mathrm{h}}_{n}(r) = \sin(\pi n r)$ are used for node embedding. Motivated by the embedding approach introduced by a recent atom-based GNN study,[34] we employ 0th order spherical Bessel functions for edge embedding,

$$\phi^{\mathrm{e}}_{m}(r) = j^{m}_{0}(r/c_{\mathrm{X}}) \cdot I_{\mathrm{X}}(r) = \sqrt{\frac{2}{c_{\mathrm{X}}}}\,\frac{\sin(\pi m r/c_{\mathrm{X}})}{r/c_{\mathrm{X}}} \cdot I_{\mathrm{X}}(r), \tag{12}$$

where c_{X} ($\mathbf{X} \in \{\mathbf{F}, \mathbf{J}, \mathbf{K}, \mathbf{D}, \mathbf{P}, \mathbf{S}, \mathbf{H}\}$) is the operator-specific upper cutoff value to $\tilde{X}^{\mathrm{e}}_{uv}$. To ensure that the feature varies smoothly when a node enters the cutoff, we further introduce the mollifier $I_{\mathrm{X}}(r)$,

$$I_{\mathrm{X}}(r) = \begin{cases} \exp\left(-\frac{c_{\mathrm{X}}^2}{(|r|-c_{\mathrm{X}})^2} + 1\right), & \text{if } 0 \le |r| < c_{\mathrm{X}} \\ 0, & \text{if } |r| \ge c_{\mathrm{X}}. \end{cases} \tag{13}$$

Note that $\phi^{\mathrm{e}}_{m}(r)$ decays to zero as an edge approaches the cutoff to ensure size consistency, and the mollifier is infinite order differentiable at the boundaries, which eliminates representation noise that can arise from geometric perturbation of the molecule. To enforce that the output is constant at machine precision when adding arbitrary numbers of zero edge features, which is critical for the extraction of analytical gradients and training potential energy surfaces, we also introduced an "auxiliary edge" scheme to be integrated with the message passing mechanism,

$$\mathbf{e}^{\mathrm{aux}}_{uv} = \mathbf{W}^{\mathrm{aux}} \cdot \mathbf{e}^{\mathrm{RBF}}_{uv}, \tag{14}$$

where $\mathbf{W}^{\mathrm{aux}}$ is a trainable parameter matrix. The radial basis function embeddings are transformed by neural network modules to yield 0th order node and edge attributes,

$$\mathbf{h}^{0}_{u} = \mathrm{Enc}_{\mathrm{h}}(\mathbf{e}^{\mathrm{RBF}}_{uv}), \quad \mathbf{e}^{0}_{uv} = \mathrm{Enc}_{\mathrm{e}}(\mathbf{h}^{\mathrm{RBF}}_{u}), \tag{15}$$

where $\mathrm{Enc}_{\mathrm{h}}$ and $\mathrm{Enc}_{\mathrm{e}}$ are residual blocks[39] comprising three dense NN layers. In contrast to atom-based message passing neural networks, this additional embedding transformation captures the interactions among the physical operators.

The node and edge attributes are updated via the transformer-motivated[40] message passing mechanism in Fig. 2. For a given message passing layer (MPL) $l + 1$, the information carried by each edge is encoded into a message function $\mathbf{m}^{l}_{uv}$ and associated attention weight $\mathbf{m}^{l}_{uv}$ and is accumulated into node features through a graph convolution operation. The overall message passing mechanism is given by

$$\mathbf{h}^{l+1}_{u} = \mathbf{h}^{l}_{u} + \sigma\left(\mathbf{W}^{l}_{\mathrm{h}} \cdot \left[\bigoplus_{i}\left(\sum_{v\in N(u)} w^{l,i}_{uv} \cdot \mathbf{m}^{l}_{uv}\right)\right] + \mathbf{b}^{l}_{\mathrm{h}}\right), \tag{16}$$

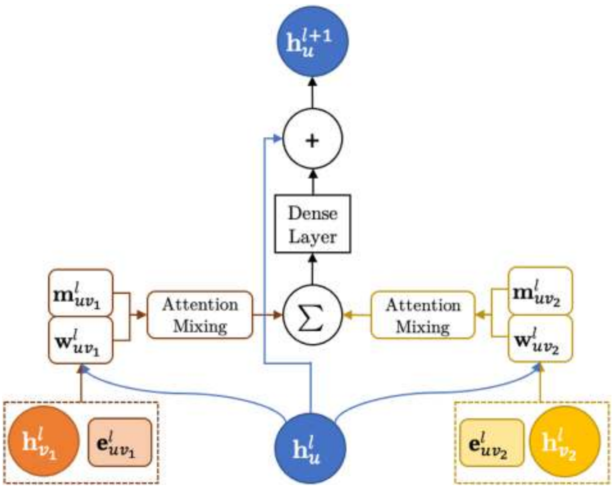

FIG. 2. Summary of the OrbNet MPL update. For the $l + 1$ MPL, the attributes of a given node (blue) are updated due to interactions with nearest-neighbor nodes (red and gold), which depend on both the nearest-neighbor node attributes and the nearest-neighbor edge attributes. The node and edge features (i.e., $\mathbf{h}^l_u$, $\mathbf{h}^l_v$, and $\mathbf{e}^l_{uv}$) combine to produce a message $\mathbf{m}^l_{uv}$ [Eq. (17)] and multi-head attention score $\mathbf{w}^l_{uv}$ [Eq. (18)], which undergo attention mixing. The attention-weighted messages from each nearest-neighbor node and edge are combined and passed into a dense layer, the result of which is added to the original node attributes to perform the update [Eq. (16)].

where $\mathbf{m}^l_{uv}$ is the message function computed on each edge,

$$\mathbf{m}^l_{uv} = \sigma(\mathbf{W}^l_{\mathrm{m}} \cdot [\mathbf{h}^l_u \odot \mathbf{h}^l_v \odot \mathbf{e}^l_{uv}] + \mathbf{b}^l_{\mathrm{m}}), \tag{17}$$

and the convolution kernel weights, $w^{l,i}_{uv}$, are evaluated as (multi-head) attention scores[30] to characterize the relative importance of an orbital pair,

$$w^{l,i}_{uv} = \sigma_{\mathrm{a}}\big(\sum[(\mathbf{W}^{l,i}_{\mathrm{a}} \cdot \mathbf{h}^l_u) \odot (\mathbf{W}^{l,i}_{\mathrm{a}} \cdot \mathbf{h}^l_v) \odot \mathbf{e}^l_{uv} \odot \mathbf{e}^{\mathrm{aux}}_{uv}]/n_{\mathrm{e}}\big), \tag{18}$$

where the summation is applied over the elements of the vector in the summand. Here, the index i specifies the attention head, n_{e} is the dimension of hidden edge features $\mathbf{e}^l_{uv}$, $\oplus$ denotes the vector concatenation operation, $\odot$ denotes the Hadamard product, and $\cdot$ denotes the matrix-vector product.

The edge attributes are updated according to

$$\mathbf{e}^{l+1}_{uv} = \sigma(\mathbf{W}^l_{\mathrm{e}} \cdot \mathbf{m}^l_{uv} + \mathbf{b}^l_{\mathrm{e}}). \tag{19}$$

$\mathbf{W}^l_{\mathrm{m}}$, $\mathbf{W}^l_{\mathrm{h}}$, $\mathbf{W}^l_{\mathrm{e}}$, $\mathbf{b}^l_{\mathrm{m}}$, $\mathbf{b}^l_{\mathrm{h}}$, $\mathbf{b}^l_{\mathrm{e}}$, and $\mathbf{a}^l$ are MPL-specific trainable parameter matrices, $\mathbf{W}^{l,i}_{\mathrm{a}}$ are MPL- and attention-head-specific trainable parameter matrices, $\sigma(\cdot)$ is an activation function with a normalization layer, and $\sigma_{\mathrm{a}}(\cdot)$ is the activation function used for generating attention scores.

The decoding phase of OrbNet [Figs. 1(f)–1(h)] is designed to ensure the size-extensivity of energy predictions. The employed mechanism outputs node-resolved energy contributions for the embedding layer ($l = 0$) and all MPLs ($l = 1, 2, \ldots, L$) to predict the energy components associated with all nodes and MPLs. The final energy prediction E^{ML} is obtained by first summing over l [Fig. 1(g)] for each node u and then performing a one-body sum over nodes (i.e., orbitals) [Fig. 1(h)] such that

$$E^{\mathrm{ML}} = \sum_{u\in\mathbf{V}} \varepsilon_u = \sum_{u\in\mathbf{V}} \sum_{l=0}^{L} \mathrm{Dec}^l(\mathbf{h}^l_u), \tag{20}$$

where the decoding networks Dec^l are multilayer perceptrons.

D. COMPARISON WITH OTHER METHODS THAT USE QUANTUM MECHANICAL FEATURES

Several ML methods have been developed for the prediction of high-level (i.e., coupled-cluster) correlation energies based on quantum mechanical features from a mean-field-level (i.e., HF theory or DFT) electronic structure calculation.[24,28,41,42] An example from our own work includes the molecular-orbital-based machine-learning (MOB-ML) approach to predict molecular properties using localized molecular orbitals for input feature generation.[24–26] Localized molecular orbitals are obtained via an orbital localization procedure (Boys and IBO) with the orbitals obtained from a mean-field electronic structure calculation. Feature vectors are then calculated for diagonal and off-diagonal molecular orbital pairs from matrix elements of the molecular orbitals with respect to various operators (i.e., Fock, Coulomb, and exchange operators) within the basis and using a feature sorting scheme. The Gaussian-process or clustering-based regressors are trained for the pair correlation energy labels associated with the MOB feature vectors.

Closer in spirit to OrbNet are NeuralXC[27] and DeePHF[28] that employ AO-based features obtained from electronic structure calculations to perform the regression and prediction of molecular energies. Both NeuralXC and DeePHF utilize the electronic density and orbitals obtained from either a Hartree–Fock (HF) (in DeePHF) or low-level density functional theory (DFT) (in NeuralXC) calculation using cc-pVDZ or larger atomic-orbital basis sets. However, these methods typically require a mean-field calculation in the same-sized atomic orbital basis set as that of the high-level correlation method (i.e., they do not directly make predictions on the basis of features that are obtained in a minimal basis), and they have not been applied for the prediction of DFT-quality results on the basis of lower-level semi-empirical methods, such as GFN-xTB, as is done here.

In terms of featurization methods, OrbNet differs from NeuralXC and DeePHF by providing a more information-rich quantum mechanical representation. Unlike NeuralXC, OrbNet avoids shell-averaging of the AOs, and unlike both NeuralXC and DeePHF, OrbNet includes all off-diagonal operator matrix elements (including both intra- and inter-atom elements, as well as intra- and inter-shell elements) within the features, thereby preserving information content while also enabling the description of long-range contributions. Unlike DeePHF, OrbNet includes interactions between different shells on the same atom and avoids the need for a pre-determined weighting function based on inter-atomic distances. OrbNet additionally includes quantum-chemical matrices including **F**, **J**, and **K** that are valuable components for energy prediction tasks. Other differences arise in the way in which rotational invariance is enforced within the features. In NeuralXC, the rotational invariance of the features is guaranteed by summing all sub-shell components of the AO-projected density $d^{nl} = \sum_{m=-l}^{l} c^2_{nlm}$ (i.e., the trace of the local density matrix) such that the information content is not preserved. In DeePHF, the rotational invariance of the features is enforced by using the eigenvalues of the local density matrix instead of the trace to build the feature vector for each shell. By contrast, OrbNet achieves the rotational invariance of features through the use of SAAOs, which involve no loss of information content.

In terms of ML regression methods, OrbNet also differs from NeuralXC and DeePHF. For NeuralXC, the ML regression is performed using a Behler–Parrinello[43] type dense neural network. Similarly, for DeePHF, the ML regression is performed using a dense neural network, with the labels associated with a one-body summation over the atoms to yield the total correlation energy. In contrast, OrbNet uses a GNN for the ML regression. Specifically, we report results using a multi-head graph attention mechanism and residual blocks to improve the representation capacity of the model in order to learn complex chemical environments. Unlike the pre-tuned aggregation coefficients in DeePHF, OrbNet also offers a flexible framework for learning orbital interactions and could be naturally transferred to downstream tasks.

III. COMPUTATIONAL DETAILS

Results are presented for the QM7b-T dataset[25,44] (which has seven conformations for each of 7211 molecules[13] with up to seven heavy atoms of type C, O, N, S, and Cl), the QM9 dataset[45] (which has locally optimized geometries for 133 885 molecules with up to nine heavy atoms of type C, O, N, and F), the GDB-13-T dataset[25,44] (which has six conformations for each of 1000 molecules from the GDB-13 dataset[46] with 13 heavy atoms of type C, O, N, S, and Cl), DrugBank-T (which has six conformations for each of 168 molecules from the DrugBank database[47] with 14 to 30 heavy atoms of type C, O, N, S, and Cl), and the Hutchison conformer dataset from Ref. 48 (which has up to ten conformations for each of 622 molecules with 9 to 50 heavy atoms of type C, O, N, F, P, S, Cl, Br, and I). Except for DrugBank-T, all of these datasets have been described previously; thermalized geometries from the DrugBank dataset are sampled at 50 fs intervals from *ab initio* molecular dynamics trajectories performed using the B3LYP[49–52]/6-31g*[53] level of theory and a Langevin thermostat[54] at 350 K. The structures for the DrugBank-T dataset are provided in the supplementary material, and all other employed datasets are already available online.[44,45,48] For results reported in Sec. IV A, the pre-computed DFT labels from Ref. 45 were employed. For results reported in Sec. IV B, all DFT labels were computed using the ωB97X-D functional[55] with a Def2-TZVP AO basis set[56] and using density fitting[57] for both the Coulomb and exchange integrals using the Def2-Universal-JKFIT basis set;[58] these calculations are performed using PSI4.[59] Semi-empirical calculations are performed using the GFN1-xTB method[60] using the ENTOS QCORE[61] package, which is also employed for the SAAO feature generation.

For the results presented in this work, we train OrbNet models using the following training-test splits of the datasets. For results on the QM9 dataset, we removed 3054 molecules due to a failed a geometric consistency check, as recommended in Ref. 45; we then randomly sampled 110 000 molecules for training and used 10 831 molecules for testing. The training sets of 25 000 and 50 000 molecules in Sec. IV A are subsampled from the 110 000-molecule dataset. For the QM7b-T dataset, two sets of training-test splits are generated; for the model trained on the QM7b-T dataset only (Model 1 in Sec. IV B), we randomly selected 6500 different molecules (with seven geometries for each) from the total 7211 molecules for training, holding out 500 molecules (with seven geometries for each) for testing; for Models 2–4 in Sec. IV B, we used a 361-molecule subset of this 500-molecule set for testing, and we used the remaining 6850 molecules of QM7b-T for training. For the GDB13-T dataset, we randomly sampled 948 different molecules (with six geometries for each) for training, holding out 48 molecules (with six geometries for each) for testing. For the DrugBank-T dataset, we randomly sampled 158 different molecules (with six geometries for each) for training, holding out ten molecules (with six geometries for each) for testing. No training on the Hutchison conformer dataset was performed, as it was only used for transferability testing. Since none of the training datasets for OrbNet included molecules with elements of type P, Br, and I, we excluded the molecules in the Hutchison dataset that included elements of these types for the reported tests (as was also done in Ref. 48 and Fig. 4 for the ANI methods). Moreover, following Ref. 48, we excluded sixteen molecules due to missing DLPNO-LCCSD(T) reference data; additional eight molecules were excluded on the basis of DFT convergence issues for at least one conformer using PSI4. The specific molecules that appear in all training-test splits are listed in the supplementary material.

Table I summarizes the hyperparameters used for training OrbNet for the reported results. We perform a pre-transformation on the input features from **F**, **J**, **K**, **D**, **P**, **H** and **S** to obtain $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{X}}^{\mathrm{e}}$: We normalize all diagonal SAAO tensor values X_{uu} to range [0, 1) for each operator type to obtain $\tilde{X}_u$; for off-diagonal SAAO tensor values, we take $\tilde{X}_{uv} = -\ln(|X_{uv}|)$ for $\mathbf{X} \in \{\mathbf{F}, \mathbf{J}, \mathbf{K}, \mathbf{P}, \mathbf{S}, \mathbf{H}\}$ and $\tilde{D}_{uv} = D_{uv}$. The model hyperparameters are selected within a limited search space; the cut-off hyperparameters $c_{\mathbf{X}}$ are obtained by examining the overlap between feature element distributions between the QM7b-T and GDB13-T datasets. The same set of hyperparameters is used throughout this work, thereby providing a universal model.

To provide additional regularization for predicting energy variations from the configurational degree of freedom, we performed training on the loss function of the form

$$\mathscr{L}(\hat{\mathbf{E}}, \mathbf{E}) = (1-\alpha)\sum_i \mathscr{L}_2(\hat{E}_i, E_i) + \alpha \sum_i \mathscr{L}_2\big(\hat{E}_i - \hat{E}_{t(i)}, E_i - E_{t(i)}\big). \tag{21}$$

For a conformer i in a minibatch, we randomly sample another conformer $t(i)$ of the same molecule to be paired with i to evaluate the relative conformer loss $\mathscr{L}_2(\hat{E}_i - \hat{E}_{t(i)}, E_i - E_{t(i)})$, putting an additional penalty on the prediction errors for configurational energy variations, where **E** denotes the ground truth energy values of the minibatch, $\hat{\mathbf{E}}$ denotes the model prediction values of the minibatch, and $\mathscr{L}_2$ denotes the L2 loss function $\mathscr{L}_2(\hat{y}, y) = \|\hat{y} - y\|_2^2$. For all models in Sec. IV A, we choose $\alpha = 0$ as only the optimized geometries are available; for models in Sec. IV B, we choose $\alpha = 0.9$ for all training setups.

All models are trained on a single NVIDIA Tesla V100-SXM2-32GB GPU using the Adam optimizer.[62] For all training runs, we set the minibatch size to 64 and use a cyclical learning rate schedule[63]

TABLE I. Model hyperparameters employed in OrbNet. All cut-off values are in atomic units.

Hyperparameter	Meaning	Value or name
n_{r}	Number of basis functions for node embedding	8
m_{r}	Number of basis functions for edge embedding	8
n_{h}	Dimension of hidden node attributes	256
n_{e}	Dimension of hidden edge attributes	64
n_{a}	Number of attention heads	4
L	Number of message passing layers	3
L_{enc}	Number of dense layers in $\mathrm{Enc_h}$ and $\mathrm{Enc_e}$	3
L_{dec}	Number of dense layers in a decoding network	4
	Hidden dimensions of a decoding network	128, 64, 32, 16
Σ	Activation function	Swish
σ_{a}	Activation function for attention generation	TanhShrink
Γ	Batch normalization momentum	0.4
$c_{\mathbf{F}}$	Cut-off value for $\tilde{F}_{uv}$	8.0
$c_{\mathbf{J}}$	Cut-off value for $\tilde{J}_{uv}$	1.6
$c_{\mathbf{K}}$	Cut-off value for $\tilde{K}_{uv}$	20.0
$c_{\mathbf{D}}$	Cut-off value for $\tilde{D}_{uv}$	9.45
$c_{\mathbf{P}}$	Cut-off value for $\tilde{P}_{uv}$	14.0
$c_{\mathbf{S}}$	Cut-off value for $\tilde{S}_{uv}$	8.0
$c_{\mathbf{H}}$	Cut-off value for $\tilde{H}_{uv}$	8.0

that performs a linear learning rate increase from 3×10^{-5} to 3×10^{-3} for the initial 100 epochs, a linear decay from 3×10^{-3} to 3×10^{-5} for the next 100 epochs, and an exponential decay with a factor of 0.9 every epoch for the final 100 epochs. Batch normalization[64] is employed before every activation function σ except for that used in the attention heads, σ_{a}.

IV. RESULTS

We present results that focus on the prediction of accurate DFT energies using input features obtained from the GFN1-xTB method.[60] The GFN family of methods[60,65,66] have proven to be extremely useful for the simulation of large molecular systems (1000s of atoms or more) with time-to-solution for energies and forces on the order of seconds. However, this applicability can be limited by the accuracy of the semi-empirical method,[48,67] thus creating a natural opportunity for "delta-learning" the difference between the GFN1 and DFT energies on the basis of the GFN1 features. Specifically, we consider regression labels associated with the difference between high-level DFT and the GFN1-xTB total atomization energies,

$$E^{\mathrm{ML}} \approx E^{\mathrm{DFT}} - E^{\mathrm{GFN1}} - \Delta E^{\mathrm{fit}}_{\mathrm{atoms}}, \tag{22}$$

where the last term is the sum of differences for the isolated-atom energies between DFT and GFN1 as determined by a linear model. This approach yields the direct ML prediction of total DFT energies, given the results of a GFN1-xTB calculation.

A. The QM9 dataset

We begin with a broad comparison of recently introduced ML methods for the total energy task, U_0, from the widely studied QM9 dataset.[45] QM9 is composed of organic molecules with up to nine heavy atoms at locally optimized geometries, so this test (Table II) examines the expressive power of the ML models for systems in similar chemical environments. Results for OrbNet are presented both without ensemble averaging of independently trained models (i.e., predicting only on the basis of the first of trained model) and with ensemble averaging the results of five independently trained models (OrbNet-ens5). As observed previously,[33] ensembling helps in this

TABLE II. MAEs (reported in meV) for predicting the QM9 dataset of total energies at the B3LYP/6-31G(2df,p) level of theory. Results from the current work are reported for a single model (OrbNet) and with ensembling over 5 models (OrbNet-ens5). Boldface indicates the best model for each training set size and for each model class, i.e., with and without ensembling.

Training size	SchNet[32]	PhysNet[33]	PhysNet-ens5[33]	DimeNet[34]	DeepMoleNet[35]	OrbNet	OrbNet-ens5
25 000	...	...	...	...	...	**11.6**	**10.4**
50 000	15	13	10	...	...	**8.22**	**6.80**
110 000	14	8.2	6.1	8.02	6.1	**5.01**	**3.92**

and other learning tasks, reducing the OrbNet prediction error by approximately 10%–20%.

In Table II, previously published methods utilizing graph representations of atom-based features are also included, which are SchNet,[32] PhysNet,[33] DimeNet,[34] and DeepMoleNet.[35] We note that DimeNet employs a directional message passing mechanism, and PhysNet and DeepMoleNet employ supervision based on prior physical information to improve the model transferability, which could also be employed within OrbNet; it is clear that without these additional strategies and even without model ensembling, OrbNet provides greater accuracy and learning efficiency than all previous deep-learning methods.

B. Transferability and conformer energy predictions

A more realistic and demanding test of ML methods is to train them on datasets of relatively small molecules (for which high-accuracy data are more readily available) and then to test on datasets of larger and more diverse molecules. This provides useful insight into the transferability of the ML methods and the general applicability of the trained models.

To this end, we investigate the performance of OrbNet on a series of datasets containing organic and drug-like molecules. Figure 3 presents results in which OrbNet models are trained with increasing amounts of data. Using the training-test splits described in Sec. III, Model 1 is trained using data from only the QM7b-T dataset; Model 2 is trained using data from the QM7b-T, GDB13-T, and DrugBank-T datasets; Model 3 is trained using data from the QM7b-T, QM9, GDB13-T, and DrugBank-T datasets; and Model 4 is obtained by ensembling five independent training runs with the same data as used for Model 3. Predictions are made for total energies [Fig. 3(a)] and relative conformer energies [Fig. 3(b)] for held-out molecules from each of these datasets, as well as for the Hutchison conformer dataset.

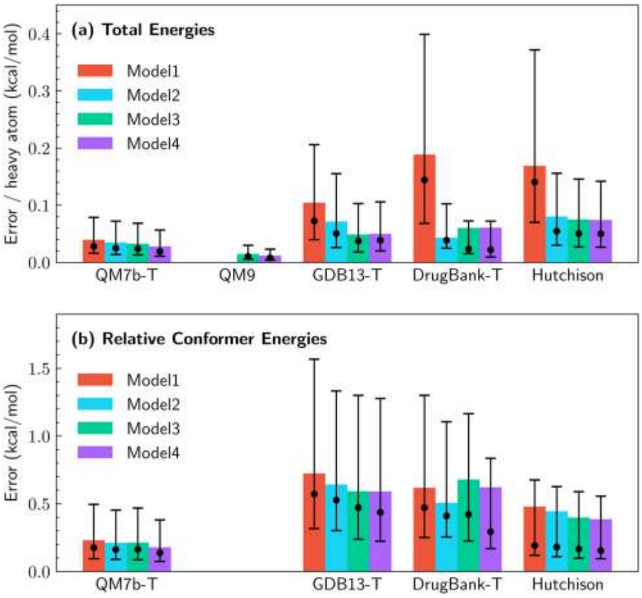

FIG. 3. Prediction errors for (a) molecule total energies and (b) relative conformer energies performed using OrbNet models trained using various datasets. The mean absolute error (MAE) is indicated by the bar height, the median of the absolute error is indicated by a black dot, and the first and third quantiles for the absolute error are indicated as the lower and upper bars. Model 1 uses training data from QM7b-T; Model 2 additionally includes training data from GDB13-T and DrugBank-T; Model 3 additionally includes training data from QM9; and Model 4 additionally includes ensemble averaging over five independent training runs. Testing is performed on data that are held-out from training in all cases. Training and prediction employs energies at the ωB97X-D/Def2-TZVP level of theory. All energies in kcal/mol.

As expected, it is seen from Fig. 3 that the OrbNet predictions improve with additional data and ensemble modeling. The median and mean of the absolute errors consistently decrease from Model 1 to Model 4 except for a non-monotonicity in the DrugBank-T mean absolute error (MAE), likely due to the relatively small size of that dataset. It is nonetheless striking that Model 1 that includes only data from QM7b-T yields relative conformer energy predictions on the DrugBank-T and Hutchison datasets (which include molecules with up to 50 heavy atoms) with an accuracy that is comparable to the models trained on more and larger molecules. Note that all of the OrbNet models predict relative conformer energies with MAE and median prediction errors that are well within the 1 kcal/mol threshold of chemical accuracy, across all four test datasets. Predictions for QM9 using Models 1 and 2 are not included since QM9 includes F atoms, whereas the training data in those models do not; relative conformer energies are not predicted for QM9 since they are not available in this dataset. Although the total energy prediction error for OrbNet is slightly larger per heavy atom on the Hutchison dataset than for the other datasets, the relative conformer energy prediction error for the Hutchison dataset is slightly smaller than for GDB13-T and DrugBank-T; this is due to the fact that the Hutchison dataset involves locally minimized conformers that are less spread in energy per heavy atom than the conformers of the thermalized datasets. This relatively small energy spread among conformers in the Hutchison dataset is a realistic and challenging aspect of drug-molecule conformer-ranking prediction, which we next consider.

Figure 4 presents a direct comparison of the accuracy and computational cost of OrbNet in comparison to a variety of other force-field, semiempirical, machine-learning, DFT, and wavefunction methods, as compiled in Ref. 48. For the Hutchison conformer dataset of drug-like molecules that range in size from 9 to 50 heavy atoms, the accuracy of the various methods was evaluated using the median R^2 of the predicted conformer energies in comparison to DLPNO-CCSD(T) reference data and with the computation time evaluated on a single central processing unit (CPU) core.[48]

The OrbNet conformer energy predictions (Fig. 4, black) are reported using Model 4 (i.e., with training data from QM7b-T, GDB13-T, DrugBank-T, and QM9 and with ensemble averaging over five independent training runs). The solid black circle indicates the median R^2 value (0.81) of the OrbNet predictions relative to the DLPNO-CCSD(T) reference data, as for the other methods; this point provides the most direct comparison to the accuracy of the other methods. The open black circle indicates the median R^2 value (0.90) of the OrbNet predictions relative to the ωB97X-D/Def2-TZVP reference data against which the model was trained; this point indicates the accuracy that would be expected for the Model 4

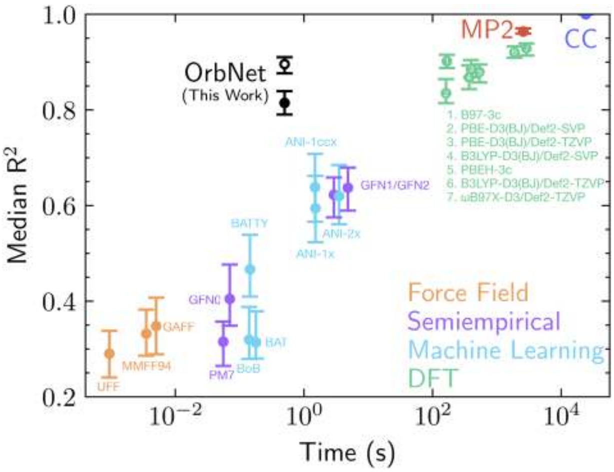

FIG. 4. Comparison of the accuracy/computational-cost trade-off for a range of potential energy methods for the Hutchison conformer benchmark dataset. Aside from the OrbNet results (black), all data were previously reported in Ref. 48, with median R^2 values for the predicted conformer energies computed relative to DLPNO-CCSD(T) reference data and with the computation time evaluated on a single CPU core. The OrbNet results (black) are obtained using Model 4 (i.e., with training data from QM7b-T, GDB13-T, DrugBank-T, and QM9 and with ensemble averaging over five independent training runs). The solid black circle plots the median R^2 value from the OrbNet predictions relative to DLPNO-CCSD(T) reference data, as for the other methods. The open black circle plots the median R^2 value from the OrbNet predictions relative to the ωB97X-D/Def2-TZVP reference data against which the OrbNet model was trained. Error bars correspond to the 95% confidence interval, determined by statistical bootstrapping.

implementation of OrbNet if it had employed coupled-cluster training data rather than DFT training data. We calculated timings for OrbNet on a single core of an Intel Core i5-1038NG7 CPU at 2.00GHz, finding that the OrbNet computational cost is dominated by the GFN1-xTB calculation for the feature generation. In contrast to Ref. 48 that used the XTB code of Grimme and co-workers,[68] we used ENTOS QCORE for the GFN1-xTB calculations. We find the reported timings for GFN1-xTB to be surprisingly slow in Ref. 48, particularly in comparison to the GFN0-xTB timings. For GFN0-xTB, our timings with ENTOS QCORE are very similar to those reported in Ref. 48, which is sensible given that the method involves no self-consistent field (SCF) iteration. However, whereas Ref. 48 indicates GFN1-xTB timings that are 43-fold slower than GFN0-xTB, we find this ratio to be only 4.5 with ENTOS QCORE, perhaps due to differences of SCF convergence. To account for the issue of code efficiency in the GFN1-xTB implementation and to control for the details of the single CPU core used in the timings for this work vs in Ref. 48, we normalize the OrbNet timing reported in Fig. 4 with respect to the GFN0-xTB timing from Ref. 48. The CPU neural-network inference costs for OrbNet are negligible contribution to this timing.

The results in Fig. 4 make clear that OrbNet enables the prediction of relative conformer energies for drug-like molecules with an accuracy that is comparable to DFT but with a computational cost that is 1000-fold reduced from DFT to realm of semiempirical methods. Alternatively viewed, the results indicate that OrbNet provides dramatic improvements in prediction accuracy over currently available ML and semiempirical methods for realistic applications, without significant increases in computational cost.

V. CONCLUSIONS

Electronic structure methods typically face a punishing trade-off between the prediction accuracy of the method and its computational cost, across all areas of the chemical, biological, and materials science. We present a new machine-learning method with the potential to substantially shift that trade-off in favor of *ab initio*-quality accuracy at a low computational cost. OrbNet utilizes a graph neural network architecture to predict high-quality electronic-structure energies on the basis of features obtained from low-cost/minimal-basis mean-field electronic structure methods. The method is demonstrated for the case of predicting ωB97X-D/Def2-TZVP energies using GFN1-xTB input features, although it is completely general with respect to both the choice of high-level (including correlated wavefunction) methods used for generating reference data and the choice of mean-field methods used for feature generation. In comparison to state-of-the-art GNN methods for the prediction of total molecule energies for the QM9 dataset, it is shown that OrbNet provides a 33% improvement in prediction accuracy with the same amount of data relative to the next-most accurate method (DeepMoleNet).[35] Additionally, in comparison to the wide array of methods used for predicting relative conformer energies in a realistic and diverse dataset of drug-like molecules, as compiled by Folmsbee and Hutchison,[48] it is shown that OrbNet provides a prediction accuracy that is similar to DFT and much improved over existing ML methods, but at a computational cost that is reduced by at least three orders of magnitude relative to DFT. Natural future directions for development will include the expansion of OrbNet to a broader set of chemical elements, incorporation of directional message-passing and model supervision using prior physical information,[33–35] and end-to-end refitting of the semi-empirical method used for feature generation.[22,69]

SUPPLEMENTAL MATERIAL

The supplementary material includes the structures for the DrugBank-T dataset, as well as specification of molecules that appear in all training-test splits for the trained models.

ACKNOWLEDGMENTS

The authors thank Lixue Sherry Cheng for providing geometries for the DrugBank-T dataset and Anders Christensen for helpful comments on the manuscript. Z.Q. acknowledges the graduate research funding from Caltech. T.F.M. and A.A. acknowledge partial support from the Caltech DeLogi fund, and A.A. acknowledges support from a Caltech Bren professorship.

DATA AVAILABILITY

The data that support the findings of this study are available within the article and its supplementary material.

REFERENCES

[1] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," Phys. Rev. Lett. **104**, 136403 (2010).

[2] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," Phys. Rev. Lett. **108**, 58301 (2012).

[3] A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. A. von Lilienfeld, "FCHL revisited: Faster and more accurate quantum machine learning," J. Chem. Phys. **152**, 044107 (2020).

[4] A. S. Christensen and O. A. von Lilienfeld, "Operator quantum machine learning: Navigating the chemical space of response properties," CHIMIA **73**, 1028–1031 (2019).

[5] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Big data meets quantum chemistry approximations: The Δ-machine learning approach," J. Chem. Theory Comput. **11**, 2087 (2015).

[6] T. T. Nguyen, E. Székely, G. Imbalzano, J. Behler, G. Csányi, M. Ceriotti, A. W. Götz, and F. Paesani, "Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions," J. Chem. Phys. **148**, 241725 (2018).

[7] S. Fujikake, V. L. Deringer, T. H. Lee, M. Krynski, S. R. Elliott, and G. Csányi, "Gaussian approximation potential modeling of lithium intercalation in carbon nanostructures," J. Chem. Phys. **148**, 241714 (2018).

[8] A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, and M. Ceriotti, "Transferable machine-learning model of the electron density," ACS Cent. Sci. **5**, 57–64 (2019).

[9] Y. Zhai, A. Caruso, S. Gao, and F. Paesani, "Active learning of many-body configuration space: Application to the Cs^+–water MB-nrg potential energy function as a case study," J. Chem. Phys. **152**, 144103 (2020).

[10] J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost," Chem. Sci. **8**, 3192–3203 (2017).

[11] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, and A. E. Roitberg, "Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning," Nat. Commun. **10**, 1–8 (2019).

[12] N. Lubbers, J. S. Smith, and K. Barros, "Hierarchical modeling of molecular energies using a deep neural network," J. Chem. Phys. **148**, 241715 (2018).

[13] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. Anatole von Lilienfeld, "Machine learning of molecular electronic properties in chemical compound space," New J. Phys. **15**, 095003 (2013).

[14] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller, "Assessment and validation of machine learning methods for predicting molecular atomization energies," J. Chem. Theory Comput. **9**, 3404 (2013).

[15] P. Gasparotto and M. Ceriotti, "Recognizing molecular patterns by machine learning: An agnostic structural definition of the hydrogen bond," J. Chem. Phys. **141**, 174110 (2014).

[16] J. Behler, "Perspective: Machine learning potentials for atomistic simulations," J. Chem. Phys. **145**, 170901 (2016).

[17] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: Moving beyond fingerprints," J. Comput. Aided Mol. Des. **30**, 595 (2016).

[18] K. T. Schütt, F. Arbabzadah, S. Chmiela, K.-R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," Nat. Commun. **8**, 13890 (2017).

[19] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller, "Bypassing the Kohn-Sham equations with machine learning," Nat. Commun. **8**, 872 (2017).

[20] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "MoleculeNet: A benchmark for molecular machine learning," Chem. Sci. **9**, 513 (2018).

[21] K. Yao, J. E. Herr, D. W. Toth, R. McKintyre, and J. Parkhill, "The TensorMol-0.1 model chemistry: A neural network augmented with long-range physics," Chem. Sci. **9**, 2261–2269 (2018).

[22] H. Li, C. Collins, M. Tanha, G. J. Gordon, and D. J. Yaron, "A density functional tight binding layer for deep learning of chemical Hamiltonians," J. Chem. Theory Comput. **14**, 5764–5776 (2018).

[23] L. Zhang, J. Han, H. Wang, R. Car, and W. E, "Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics," Phys. Rev. Lett. **120**, 143001 (2018).

[24] M. Welborn, L. Cheng, and T. F. Miller III, "Transferability in machine learning for electronic structure via the molecular orbital basis," J. Chem. Theory Comput. **14**, 4772–4779 (2018).

[25] L. Cheng, M. Welborn, A. S. Christensen, and T. F. Miller III, "A universal density matrix functional from molecular orbital-based machine learning: Transferability across organic molecules," J. Chem. Phys. **150**, 131103 (2019).

[26] L. Cheng, N. B. Kovachki, M. Welborn, and T. F. Miller III, "Regression clustering for improved accuracy and training costs with molecular-orbital-based machine learning," J. Chem. Theory Comput. **15**, 6668–6677 (2019).

[27] S. Dick and M. Fernandez-Serra, "Machine learning accurate exchange and correlation functionals of the electronic density," Nat. Commun. **11**, 3509 (2020).

[28] Y. Chen, L. Zhang, H. Wang, and W. E, "Ground state energy functional with Hartree-Fock efficiency and chemical accuracy," J. Phys. Chem. A **124**(35), 7155–7165 (2020).

[29] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in International Conference on Learning Representations, 2017.

[30] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in International Conference on Learning Representations, 2018.

[31] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, and R. Barzilay, "Analyzing learned molecular representations for property prediction," J. Chem. Inf. Model. **59**, 3370–3388 (2019).

[32] K. Schütt, P.-J. Kindermans, H. E. S. Felix, S. Chmiela, A. Tkatchenko, and K.-R. Müller, "Schnet: A continuous-filter convolutional neural network for modeling quantum interactions," in *Advances in Neural Information Processing Systems* (Neural Information Processing Systems Foundation, Inc., 2017), pp. 991–1001.

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

[34] J. Klicpera, J. Groß, and S. Günnemann, "Directional message passing for molecular graphs," in International Conference on Learning Representations, 2019.

[35] Z. Liu, L. Lin, Q. Jia, Z. Cheng, Y. Jiang, Y. Guo, and J. Ma, "Transferable multilevel attention neural network for accurate prediction of quantum chemistry properties via multi-task learning," ChemRxiv:12588170.v1 (2020).

[36] S. Grimme, "A simplified Tamm-Dancoff density functional approach for the electronic excitation spectra of very large molecules," J. Chem. Phys. **138**, 244104 (2013).

[37] S. Grimme and C. Bannwarth, "Ultra-fast computation of electronic spectra for large systems by tight-binding based simplified Tamm-Dancoff approximation (sTDA-xTB)," J. Chem. Phys. **145**, 054103 (2016).

[38] T. Risthaus, A. Hansen, and S. Grimme, "Excited states using the simplified Tamm–Dancoff-approach for range-separated hybrid density functionals: Development and application," Phys. Chem. Chem. Phys. **16**, 14408–14419 (2014).

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 770–778.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (Neural Information Processing Systems Foundation, Inc., 2017), pp. 5998–6008.

[41] R. T. McGibbon, A. G. Taube, A. G. Donchev, K. Siva, F. Hernández, C. Hargus, K.-H. Law, J. L. Klepeis, and D. E. Shaw, "Improving the accuracy of Møller-Plesset perturbation theory with neural networks," J. Chem. Phys. **147**, 161725 (2017).

[42] J. T. Margraf and K. Reuter, "Making the coupled cluster correlation energy machine-learnable," J. Phys. Chem. A **122**, 6343–6348 (2018).

[43] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett. **98**, 146401 (2007).

[44] L. Cheng, M. Welborn, A. S. Christensen, and T. F. Miller, Thermalized (350k) QM7b, GDB-13, water, and short alkane quantum chemistry dataset including MOB-ML features, 2019, https://data.caltech.edu/records/1177.

[45] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," Sci. Data **1**, 1–7 (2014).

[46] L. C. Blum and J.-L. Reymond, "970 million druglike small molecules for virtual screening in the chemical universe database GDB-13," J. Am. Chem. Soc. **131**, 8732 (2009).

[47] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart, "DrugBank 4.0: Shedding new light on drug metabolism," Nucl. Acids Res. **42**, D1091–D1097 (2014).

[48] D. Folmsbee and G. Hutchison, "Assessing conformer energies using electronic structure and machine learning methods," Int. J. Quantum Chem. (published online).

[49] S. H. Vosko, L. Wilk, and M. Nusair, "Accurate spin-dependent electron liquid correlation energies for local spin density calculations: A critical analysis," Can. J. Phys. **58**, 1200 (1980).

[50] C. Lee, W. Yang, and R. G. Parr, "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density," Phys. Rev. B **37**, 785 (1988).

[51] A. D. Becke, "Density-functional thermochemistry. III. The role of exact exchange," J. Chem. Phys. **98**, 5648 (1993).

[52] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, "*Ab initio* calculation of vibrational absorption and circular dichroism spectra using density functional force fields," J. Phys. Chem. **98**, 11623 (1994).

[53] P. C. Hariharan and J. A. Pople, "The influence of polarization functions on molecular orbital hydrogenation energies," Theor. Chim. Acta **28**, 213 (1973).

[54] G. Bussi and M. Parrinello, "Accurate sampling using Langevin dynamics," Phys. Rev. E **75**, 056707 (2007).

[55] Y.-S. Lin, G.-D. Li, S.-P. Mao, and J.-D. Chai, "Long-range corrected hybrid density functionals with improved dispersion corrections," J. Chem. Theory Comput. **9**, 263–272 (2013).

[56] F. Weigend and R. Ahlrichs, "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy," Phys. Chem. Chem. Phys. **7**, 3297 (2005).

[57] R. Polly, H.-J. Werner, F. R. Manby, and P. J. Knowles, "Fast Hartree-Fock theory using local density fitting approximations," Mol. Phys. **102**, 2311–2321 (2004).

[58] F. Weigend, "Hartree-Fock exchange fitting basis sets for H to Rn," J. Comput. Chem. **29**, 167–175 (2008).

[59] D. G. A. Smith, L. A. Burns, A. C. Simmonett, R. M. Parrish, M. C. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan, A. M. James, S. Lehtola, J. P. Misiewicz, M. Scheurer, R. A. Shaw, J. B. Schriber, Y. Xie, Z. L. Glick, D. A. Sirianni, J. S. O'Brien, J. M. Waldrop, A. Kumar, E. G. Hohenstein, B. P. Pritchard, B. R. Brooks, H. F. Schaefer, A. Y. Sokolov, K. Patkowski, A. E. DePrince, U. Bozkaya, R. A. King, F. A. Evangelista, J. M. Turney, T. D. Crawford, and C. D. Sherrill, "Psi4 1.4: Open-source software for high-throughput quantum chemistry," J. Chem. Phys. **152**, 184108 (2020).

[60] S. Grimme, C. Bannwarth, and P. Shushkov, "A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements ($Z = 1$–86)," J. Chem. Theory Comput. **13**, 1989–2009 (2017).

[61] F. Manby, T. Miller, P. Bygrave, F. Ding, T. Dresselhaus, F. Batista-Romero, A. Buccheri, C. Bungey, S. Lee, R. Meli, K. Miyamoto, C. Steinmann, T. Tsuchiya, M. Welborn, T. Wiles, and Z. Williams, "Entos: A quantum molecular simulation package," ChemRxiv:7762646.v2 (2019).

[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in 3rd International Conference for Learning Representations, San Diego, 2015.

[63] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications* (International Society for Optics and Photonics, 2019), Vol. 11006, p. 1100612.

[64] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning* (International Machine Learning Society, 2015), pp. 448–456.

[65] P. Pracht, E. Caldeweyher, S. Ehlert, and S. Grimme, "A robust non-self-consistent tight-binding quantum chemistry method for large molecules," ChemRxiv:8326202.v1 (2019).

[66] C. Bannwarth, S. Ehlert, and S. Grimme, "GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions," J. Chem. Theory Comput. **15**, 1652–1671 (2019).

[67] H. Jiang, X. Tao, M. Kammler, F. Ding, A. M. Wodtke, A. Kandratsenka, T. F. Miller III, and O. Bünermann, "Nuclear quantum effects in scattering of H and D from graphene," arXiv:2007.03372 (2020).

[68] Semiempirical extended tight-binding program package 2020, accessed 14 July 2020, https://github.com/grimme-lab/xtb.

[69] G. Zhou, B. Nebgen, N. Lubbers, W. F. Malone, A. M. Niklasson, and S. Tretiak, "Graphics processing unit-accelerated semiempirical Born Oppenheimer molecular dynamics using PyTorch," J. Chem. Theory Comput. **16**, 4951–4962 (2020).