Improved accuracy and transferability of molecular-orbital-based machine learning: Organics, transition-metal complexes, non-covalent interactions, and transition states

Tamara Husch,1 Jiace Sun,1 Lixue Cheng,1 Sebastian J. R. Lee,1 and Thomas F. Miller III1,2
Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125, USA

(Dated: 19 October 2020)

Molecular-orbital-based machine learning (MOB-ML) provides a general framework for the prediction of accurate correlation energies at the cost of obtaining molecular orbitals. The application of Nesbet’s theorem makes it possible to recast a typical extrapolation task, training on correlation energies for small molecules and predicting correlation energies for large molecules, into an interpolation task based on the properties of orbital pairs. We demonstrate the importance of preserving physical constraints, including invariance conditions and size consistency, when generating the input for the machine learning model. Numerical improvements are demonstrated for different data sets covering total and relative energies for thermally accessible organic and transition-metal containing molecules, non-covalent interactions, and transition-state energies. MOB-ML requires training data from only 1% of the QM7b-T data set (i.e., only 70 organic molecules with seven and fewer heavy atoms) to predict the total energy of the remaining 99% of this data set with sub-kcal/mol accuracy. This MOB-ML model is significantly more accurate than other methods when transferred to a data set comprised of thirteen heavy atom molecules, exhibiting no loss of accuracy on a size intensive (i.e., per-electron) basis. It is shown that MOB-ML also works well for extrapolating to transition-state structures, predicting the barrier region for malonaldehyde intramolecular proton-transfer to within 0.35 kcal/mol when only trained on reactant/product-like structures. Finally, the use of the Gaussian process variance enables an active learning strategy for extending MOB-ML model to new regions of chemical space with minimal effort. We demonstrate this active learning strategy by extending a QM7b-T model to describe non-covalent interactions in the protein backbone-backbone interaction data set to an accuracy of 0.28 kcal/mol.

I. INTRODUCTION

The calculation of accurate potential energies of molecules and materials at affordable cost is at the heart of computational chemistry. While state-of-the-art ab initio electronic structure theories can yield highly accurate results, they are computationally too expensive for routine applications. Density functional theory (DFT) is computationally cheaper and has thus enjoyed widespread applicability. However, DFT is hindered by a lack of systematic improbability and from an uncertain quality for many applications.

In recent years, a variety of machine learning approaches has emerged which promise to mitigate the cost of highly accurate electronic structure methods while preserving accuracy.15,19 While these machine learning methods share similar goals, they differ in the representation of the molecules and in the machine learning methodology itself. Here, we will focus on the molecular-orbital-based machine learning (MOB-ML) approach.1,3,10–19 The defining feature of MOB-ML is its framing of learning highly accurate correlation energies as learning a sum of orbital pair correlation energies. These orbital pair correlation energies can be individually regressed with respect to a feature vector representing the interaction of the molecular orbital pairs. Without approximation, it can be shown that such pair correlation energies add up to the correct total correlation energy for single-reference wave function methods. Phrasing the learning problem in this manner has the advantage that a given pair correlation energy, and, hence, a given feature vector, is independent of molecular size (after a certain size threshold has been reached) because of the inherent spatial locality of dynamic electron correlation. Consequently, operating in such an orbital pair interaction framework converts the general extrapolation task of training on small molecules and predicting on large molecule into an interpolation task of training on orbital pairs in a small molecule and predicting on the same orbital pairs in a large molecule.

In this work, we address challenges introduced by operating in a vectorized molecular orbital pair interaction framework (Section II). We show how changes to the feature design affect the performance and transferability of MOB-ML models within the same molecular family (Section IV A) and across molecular families (Sections IV B–IV C). We probe these effects on relative- and total-energy predictions for organic and transition-metal containing molecules, and we investigate the applicability of MOB-ML to transition-state structures and non-covalent interactions.

II. THEORY

MOB-ML predicts correlation energies based on information from the molecular orbitals.1,3,10–19 The correlation energy \( E^{\text{corr}} \) in the current study is defined as the difference between the true total electronic energy and the Hartree–Fock (HF) energy for a given basis set. Without approximation, the correlation energy is expressed as a sum over correlation energy
Electronic structure theories offer different ways of approximating these pair correlation energies. For example, the second-order Möller-Plesset perturbation theory (MP2) correlation energy is

\[ E_{\text{corr}} = \sum_{ij} \varepsilon_{ij}. \]  

\[ \varepsilon_{ij} = \frac{\langle ia||jb \rangle^2}{F_{aa} + F_{bb} - F_{ii} - F_{jj}}, \]  

where \( a, b \) denote virtual orbitals, \( F \) the Fock matrix in the molecular orbital basis, and \( \langle ia||jb \rangle \) the anti-symmetrized exchange integral. We denote a general repulsion integral over the spatial coordinates \( x_1, x_2 \) of molecular orbitals \( p, q, m, n \) following the chemist’s notation as

\[ [K^{pq}]_{mn} = \langle pq|mn \rangle = \int dx_1 dx_2 p(x_1)^* q(x_1) \frac{1}{|x_1 - x_2|} m(x_2)^* n(x_2). \]

The evaluation of correlation energies with post-HF methods like MP2 or coupled-cluster theory (including CCSD(T)) involves computations that exceed the cost of HF theory by orders of magnitude. By contrast, MOB-ML predicts the correlation energy at negligible cost by machine-learning the map

\[ \varepsilon_{ij} \approx \varepsilon_{\text{ML}}(f_{ij}), \]

where \( f_{ij} \) denotes the feature vector into which we obtain the feature vector elements associated with orbital \( \tilde{i} \) and \( \tilde{j} \). The feature vector \( f_{ij} \) assembles information on the molecular orbital interactions: (i) Orbital energies of the valence-occupied and valence-virtual orbitals \( F_{pp} \), (ii) mean-field interaction energy of valence-occupied and valence-occupied orbitals and of valence-virtual and valence-virtual orbitals \( F_{pq} \), (iii) Coulomb interaction of valence-occupied and valence-occupied orbitals, of valence-occupied and valence-virtual orbitals, and valence-virtual and valence-virtual orbitals \( [K^{pp}]_{pq} \), and (iv) exchange interaction of valence-occupied and valence-occupied orbitals, of valence-occupied and valence-virtual orbitals, and valence-virtual and valence-virtual orbitals \( [K^{pq}]_{pq} \). We note that all of these pieces of information enter either the MP2 or the MP3 correlation energy expressions, which helps to motivate their value within our machine learning framework. We remove repetitive information from the feature vector and separate the learning problem into the cases where (i) \( i \neq j \) where we employ the feature vector as defined in Eq. (5) and (ii) \( i = j \) where we employ the feature vector as defined in Eq. (6).

Here, the index \( k \) denotes an occupied orbital other than \( i \) and \( j \). For blocks in the feature vector that include more than one element, we specify a canonical order of the feature vector elements. In our previous work, this order was given by the sum of the Euclidean distances between the centroids of orbital \( \tilde{i} \) and \( p \) and between the centroids of orbital \( \tilde{j} \) and \( p \). In the current work, we employ a different strategy to sort the feature vector elements (Section II A), we modify the protocol with which we obtain the feature vector elements associated with \( \tilde{i} \) and \( \tilde{j} \) (Section II B), and we revise our feature vector elements to ensure size consistency (Section II C). We provide a conceptual description of the changes to the feature set below and we give the full definition of the feature vector elements and the criteria according to which the feature vector elements are ordered in Tables S3–S6 in the Supporting information.

### A. Defining importance of feature vector elements

Careful ordering of the elements of the feature vector blocks in necessary in the current work because Gaussian process regression (GPR) is sensitive to permutation of the feature vector elements. Furthermore, the application of a Gaussian process requires that the feature vectors be of fixed length.

Given the near-sighted nature of dynamical electron correlation, it is expected that only a limited number of orbital-pair interactions are important to predict the pair correlation energy with MOB-ML. To construct the fixed-length feature vector, a cutoff criterion must be introduced. For some feature vector elements, a robust definition of importance is straightforward. The spatial distance between the orbital centroids \( i \) and \( a \) is, for example, a reliable proxy for the importance of the feature vector elements \( \{[K]_{aa} \} \) of the feature vector \( f_i \). However, the definition of importance is
less straightforward for feature vector elements that involve more than two indices. The most prominent example is the \{\{[K]_{ab}\}_{ab}\} feature vector block of \textbf{f}_{ij}, which contains the exchange integrals between the valence-virtual orbitals \(a\) and \(b\) and which should be sorted with respect to the importance of these integrals for the prediction of the pair correlation energy \(\varepsilon_{ij}\). It is non-trivial to define a spatial metric which defines the importance of the feature vector elements \{\{[K]_{ab}\}_{ab}\} to predict the pair correlation energy \(\varepsilon_{ij}\); instead, we employ the the MP3 approximation for the pair correlation energy,

\[ \varepsilon_{ij}^{MP3} = \frac{1}{8} \sum_{abcd} (t_{ij}^{ab})^* \langle ab|cd \rangle t_{ij}^{cd} + \frac{1}{8} \sum_{klab} (t_{ij}^{ab})^* \langle kl|ij \rangle t_{kl}^{ij} - \sum_{kabc} (t_{ij}^{ab})^* \langle kb|ic \rangle t_{kj}^{bc}, \]  

(7)

where \(t_{ij}^{ab}\) denotes the T-amplitude. Although we operate in a local molecular orbital basis, the canonical formulae are used to define the importance criterion; if we consider orbital localization as a perturbation (as in Kaply–Moller–Plesset theory\),\(^{23}\) the canonical expression is the leading order term. The term we seek to attach an importance to, \{\{[K]_{ab}\}_{ab}\}, appears in the first term of Eq. (7) and all integrals necessary to compute this term are readily available as (a combination of) other feature elements, i.e., we do not incur any additional significant computational cost to obtain the importance of the feature vector elements.

The way in which we determine the importance of the \{\{[K]_{ab}\}_{ab}\} elements here is an example of a more general strategy that we employ, in which the importance is assigned according to the lowest-order perturbation theory in which the features first appear in. Similar considerations have to be made for each feature vector block, all of which are specified in detail in Tables S3 and S4 in the Supporting Information.

B. Orbital-index permutation invariance

The Fock, Coulomb, and exchange matrix elements that comprise MOB features are naturally invariant to rotation and translation of the molecule. However, some care is needed to ensure that these invariances are not lost in the construction of symmetrized MOB features. In particular, rotating the valence-occupied orbitals into gerade and ungerade combinations leads to an orbital-index permutation variance for energetically degenerate orbitals \(i, j\) because the sign of the feature vector elements \(M_{ij}\),

\[ M_{ij} = \frac{1}{\sqrt{2}} M_{ip} - \frac{1}{\sqrt{2}} M_{jp}, \]  

(8)

depends on the arbitrary assignment of the indices \(i\) and \(j\). To rectify this issue, we include the absolute value of the generic feature vector element \(M\) in the feature vector instead of the signed value,

\[ M_{ij} = \left| \frac{1}{\sqrt{2}} M_{ip} - \frac{1}{\sqrt{2}} M_{jp} \right|, \]  

(9)

where \(M_{ip}\) may be \(F_{ip}\), \([K^{ij}]_{pp}\), or \([K^{ij}]_{jp}\). The corresponding equation,

\[ M_{ip} = \frac{1}{\sqrt{2}} M_{ip} + \frac{1}{\sqrt{2}} M_{jp}, \]  

(10)

is already orbital-index permutation invariant because we chose \(M_{pq} (p \neq q)\) to be positive.\(^{33}\)

C. Size consistency

Size consistency is the formal property by which the energy of two isolated molecules equals the sum of their dimer upon infinite separation. In the context of MOB-ML, satisfaction of this property requires that the contributions from the diagonal feature vectors are not affected by distant, non-interacting molecules and that

\[ \varepsilon^{ML}(\textbf{f}_{ij}) = 0 \text{ for } r_{ij} = \infty \]  

(11)

for contributions from the off-diagonal feature vectors. To ensure that MOB-ML exhibits size-consistency without the need for explicit training on the dimeric species, the following modifications to the feature vectors are made.

a. Diagonal feature vector. The feature vector as defined in Eq. (9) contains three blocks whose elements are independent of orbital \(i\), \{\{F\}_{ab}\}, \{\{K^{aa}\}_{ab}\}, and \{\{K^{ab}\}_{ab}\}. The magnitude of these feature vector elements does not decay with an increasing distance between orbital \(i\) localized on molecule \(I\) and an orbital (for example, \(a\)) localized on molecule \(J\). To address this issue, we multiply these feature vector elements by their estimated importance (see Section II A) so that they decay smoothly to zero. The other feature vector elements decay to zero when the involved orbitals are non-interacting albeit at different rates; we take the cube of feature vector elements of the type \{\{K^{pq}\}_{pq}\} to achieve a similar decay rate for all feature vector elements in the short- to medium-range which facilitates machine learning.

b. Off-diagonal feature vector. We modify the off-diagonal feature vector such that \(\textbf{f}_{ij} = 0\) for \(r_{ij} = \infty\) by first applying the newly introduced changes for \(\textbf{f}_{ij}\) also for \(\textbf{f}_{ij}\). Further action is needed for the off-diagonal case because many feature vector elements do not decay to zero when the distance between \(i\) and \(j\) is large due to rotation of the orbitals into a gerade and an ungerade combination, e.g., \(F_{ik} = \frac{1}{\sqrt{2}} F_{ik} + \frac{1}{\sqrt{2}} F_{jk}\) for \(r_{ij} = \infty, r_{jk} = \infty\). As a remedy, we apply a damping function of the form

\[ \frac{1}{1 + \frac{1}{r_{ij}/r_{0}}}, \]  

where the damping radius, \(r_{0}\), needs to be sufficiently large as to not interfere with machine learning at small \(r_{ij}\). If a damping radius close to zero would be chosen, all off-diagonal feature vectors would be zero which nullifies the information content; however, the damping radius \(r_{0}\) also should not be too large as size-consistency has to be fully learned until the off-diagonal feature vector is fully damped to zero. Therefore, we employ
a damping radius in the intermediate-distance regime and we empirically found \( r_0 = 5.0 \) Bohr to work well.

Lastly, we enforce that \( \epsilon^{ML}(0) = 0 \). The MOB features are engineered to respect this limit and would, for example, in a linear regression with a zero intercept trivially predict a zero-valued pair correlation energy without any additional training. However, the Gaussian process regression we apply in this work does not trivially yield a zero-valued pair correlation energy for a zero-valued feature vector. In the case that a training set does not include examples of zero-valued feature vectors, we need to include zero-valued feature vectors and zero-valued pair correlation energies in training to ensure that \( \epsilon^{ML}(0) = 0 \). For no model trained in the current study were more than 5% zero-valued feature vectors included.

The resulting MOB-ML model leads to size consistent energy predictions to the degree to which the underlying MO generation is. It is not required that the dimer is explicitly part of training the MOB-ML model to obtain this result. The detailed definition of each feature vector block is summarized in Tables S5 and S6. We apply the feature set defined in Tables S5 and S6 consistently in this work.

III. COMPUTATIONAL DETAILS

We present results for five different data sets: (i) a series of alkane molecules, (ii) the potential energy surface of the malonaldehyde molecule, (iii) a thermalized version of the QM7b and the GDB13 data set (i.e., QM7b-T and GDB13-T), (iv) a set of backbone-backbone interactions (BBI) and (v) a thermalized version of a subset of mononuclear, octahedral transition metal complexes put forward by Kulik and co-workers. We refer to the Supporting Information Section II for a description how the structures were obtained or generated. All generated structures are available in Ref. [41].

The features for all structures were generated with the ENToS QCORE [23] package. The feature generation is based on a HF calculation applying a cc-pVTZ basis for the elements H, C, N, O, S, and Cl. We apply a def2-TZVP basis set for all transition metals. The HF calculations were accelerated with density fitting for which we applied the corresponding cc-pVTZ-JKFIT and def2-TZVP-JKFIT density fitting bases. Subsequently, we localized the valence-occupied and the valence-virtual molecular orbitals with the Boys–Foster localization scheme [44] or with the intrinsic bond orbital (IBO) localization scheme [45]. We implemented a scheme to localize the valence-virtual orbitals with respect to the Boys–Foster function (for details on this implementation, see Section II in the Supplementary Information). We applied the Boys–Foster localization scheme for the data sets (i), (iii), (iv), and (v) for valence-occupied and valence-virtual molecular orbitals. IBO localization for valence-occupied and valence-virtual molecular orbitals led to better results for data set (ii).

The resulting orbitals are imported into the Molpro 2018.3 package via the matzop functionality to generate the non-canonical MP2 or CCSD(T) pair correlation energies with the same orbitals we applied for the feature generation. These calculations are accelerated with the resolution of the identity approximation. The frozen-core approximation is invoked for all correlated calculations.

We follow the machine learning protocol outlined in previous work [18] to train the MOB-ML models. In a first step, we perform MOB feature selection by evaluating the mean decrease of accuracy in a random forest regression in the scikit-learn v0.22.0 package [56]. We then regress the diagonal and off-diagonal pair correlation energies separately with respect to the selected features in the GPy 1.9.6 software package [57]. We employ the Matérn 5/2 kernel with white noise regularization [22]. We minimize the negative log marginal likelihood objective with respect to the kernel hyperparameters with a scaled conjugate gradient scheme for 100 steps and then apply the BFGS algorithm until full convergence. As indicated in the results, both random-sampling and active-learning strategies were employed for the selection of molecules in the training data sets. In the active-learning strategy, we use a previously trained MOB-ML model to evaluate the Gaussian process variance for each molecule, and then include the points with the highest variance in the training data set, as outlined in Ref. [58]. To estimate the Gaussian process variance for each molecule, it was assumed the variances per molecular orbital pair are mutually independent.

IV. RESULTS

A. Transferability within a molecular family

We first examine the effect of the feature vector generation strategy on the transferability of MOB-ML models within a molecular family. To this end, we revisit our alkane data set [18] which contains 1000 ethane and 1000 propane geometries as well as 100 butane and 100 isobutane geometries. We perform the transferability test outlined in Ref. [18] i.e., training a MOB-ML model on correlation energies for 50 randomly chosen ethane geometries and 20 randomly chosen propane geometries to predict the correlation energies for the 100 butane and 100 isobutane geometries (see Figure 1). This transferability test was repeated with 10000 different training data sets (each consisting of data for 50 ethane molecules and 20 propane molecules) to assess the training set dependence of the MOB-ML models. As suggested in Ref. [25], we consider various performance metrics to assess the prediction accuracy of the MOB-ML models: (i) the mean error (ME, Eq. (S3)), (ii) the mean absolute error (MAE, Eq. (S4)), (iii) the maximum absolute error (MaxAE, Eq. (S5)), and (iv) the mean absolute relative error (MARE, Eq (S6)) which applies a global shift setting the mean error to zero. We report the minimum, peak, and maximum encountered MAREs in Table I alongside literature values obtained in our previous work [18] by Dick et al. [27] and by Chen et al. [25]. The MEs, MAEs, and MaxAEs are reported in Figure S1.

In general, MOB-ML as well as NeuralXC [25] and DeepHF [25] produce MAREs well below chemical accuracy for correla-
which was also observed for results obtained with DeepHF. This shift is strongly training-set dependent, butane and from previous work, peak MAREs for butane as well as for isobutane which are vector generation strategy for MOB-ML results in the best simulation energies of ethane and propane. Updating the feature translation energies of butane and isobutane when trained on correlation energies for MOB-ML range from −0.92 to 2.70 kcal/mol for butane and from −0.18 to 1.02 kcal/mol for isobutane (see also Figure S1). This shift is strongly training-set dependent, which was also observed for results obtained with DeepHF.

The results highlight that this is an extrapolative transferability test. A considerable advantage of applying GPR in practice is that each prediction is accompanied by a Gaussian process variance which, in this case, indicates that we are in an extrapolative regime (see Figure 1). Extrapolations might be associated with quality degradation which we see, most prominently, for the mean error in butane. By contrast, other machine learning approaches like neural networks are less clear in terms of whether the predictions are in an interpolative or extrapolative regime. By including the butane molecule with the largest variance in the training set (which then consists of 50 ethane, 20 propane, and 1 butane geometries) we reduce the ME from 0.78 to 0.25, MAE from 0.78 to 0.26, MaxAE from 1.11 to 0.51, and the MARE from 0.11 to 0.09 kcal/mol for butane (see Figure S2). These results directly illustrate that MOB-ML can be systematically improved by including training data that is more similar to the test data; the improved confidence of the prediction is then also directly reflected in the associated Gaussian process variances.

As a second example, we examine the transferability of a MOB-ML model trained within a basin of a potential energy surface to the transition-state region of the same potential energy surface. We chose malonaldehyde for this case study as it has also been explored in previous machine learning studies. We train a MOB-ML model on 50 thermalized malonaldehyde structures which all have the property that \( d(\text{O}^1-\text{H}) + d(\text{O}^2-\text{H}) > 0.4 \) Å (where \( d \) denotes the distance between the two nuclei) which ensures that we are sampling from the basins. We then apply this trained model to predict the correlation energies for a potential energy surface mapping out the hydrogen transfer between the two oxygen atoms (see Figure 2). MOB-ML produces an accurate potential energy surface for the hydrogen transfer in malonaldehyde only from information on the basins (compare left and middle left panel of Figure 2). The highest encountered errors on the minimum potential energy path are smaller than 1.0 kcal/mol. Unsurprisingly, the predicted minimum energy structure \( d(\text{O}^1-\text{H}) = 1.00 \) Å, \( d(\text{O}^2-\text{H}) = 1.63 \) Å is very similar to the reference minimum energy structure \( d(\text{O}^1-\text{H}) = 1.00 \) Å, \( d(\text{O}^2-\text{H}) = 1.64 \) Å. Strikingly, the predicted energy of 2.65 kcal/mol at the saddle point at \( d(\text{O}^1-\text{H}) = d(\text{O}^2-\text{H}) = 1.22 \) Å differs from the reference energy by only 0.35 kcal/mol, although the MOB-ML model was not trained on any transition-state like structures. The highest errors are encountered in the high-energy regime and this region is also associated with the highest Gaussian process variance indicating low confidence in the predictions (compare middle right and right panel of Figure 2). The Gaussian process variance reflects the range of structures the MOB-ML model has been trained in and highlights again that we did not include transition-state-like structures in the training.

### B. Transferability across organic chemistry space

The Chemical Space Project computationally enumerated all possible organic molecules up to a certain number of atoms, resulting in the GDB databases. In this work, we examine thermalized subsets of the GDB13 data set to investigate the transferability of MOB-ML models across organic chemistry space. The application of thermalized sets of molecules has the advantage that we can study the transferability of our models for chemical and conformational degrees of freedom at the same time. To test the transferability of MOB-ML across chemical space, we train our models on a thermalized set of seven and fewer heavy-atom molecules.

| Method          | Feature set | Butane MARE | Isobutane MARE |
|-----------------|-------------|-------------|----------------|
| NeuralX[18]     | —           | 0.15        | 0.14           |
| DeepHF[23]      | —           | 0.06        | 0.11 0.43 0.07 0.13 0.53 |
| MOB-ML Ref.[18] | —           | 0.20        | 0.21           |
| this work       | —           | 0.06        | 0.11 0.19 0.06 0.10 0.19 |

TABLE I. Comparison of the minimum, peak, and maximum mean absolute error after global shift (MARE) in kcal/mol for the prediction of CCSD(T) correlation energies for butane and isobutane obtained with different methods.
FIG. 2. Relative energies obtained with MP2/cc-pVTZ (left panel), relative energies predicted with MOB-ML (middle left panel), the difference between the MOB-ML prediction and the reference data (middle right panel), and the Gaussian process variance (right panel) for the proton transfer in malonaldehyde as a function of the distance of the proton from the two oxygen atoms.

(also known as QM7b-T) and then we test the prediction accuracy on a QM7b-T test set and on a thermalized set of molecules with thirteen heavy atoms (GDB13-T; see also Section V in the Supporting Information), as also outlined in our previous work.18,19

We first investigate the effect of changing the feature vector generation protocol on the QM7b-T → QM7b-T prediction task (see Figure 3). In Ref. 18, we found that training on about 180 structures is necessary to achieve a model with an MAE below 1 kcal/mol. The FHCL method yields an MAE below 1 kcal/mol when training on about 800 structures and the DeepHF method already exhibits an MAE below 1 kcal/mol when training on their smallest chosen training set which consists of 300 structures (MAE=0.79 kcal/mol).25 The refinements in the current work reduce the number of required training structures to reach chemical accuracy to about 100 structures when sampling randomly. This number is, however, strongly training set dependent. We can remove the training-set dependence by switching to an active learning strategy where we can achieve an MAE below 1 kcal/mol reliably with about 70 structures. In general, the MAE obtained with the active learning strategy is comparable to the smallest MAEs obtained with random sampling strategies. This has the advantage that a small number of reference data can be generated in a targeted manner.

In general, our aim is to obtain a machine learning model which reliably predicts broad swathes of chemical space. For an ML model to be of practical use, it has to be able to describe out-of-set molecules of different sizes to a similar accuracy when accuracy is measured size-intensively.36 We probe the ability of MOB-ML to describe out-of-set molecules with a different number of electron pairs by applying a model trained on correlation energies for QM7b-T molecules to predict correlation energies for GDB13-T. We collect the best results published for this transfer test in the literature in Figure 4. Our previous best single GPR model achieved an MAE of 2.27 kcal/mol when trained on 220 randomly chosen structures.18 The modifications in the current work now yield a single GPR model which achieves an MAE of 1.47–1.62 kcal/mol for GDB13-T when trained on 220 randomly chosen QM7b-T structures. Strikingly, MOB-ML outperforms machine learning models trained on thousands of molecules like our RCR/GPR model and FHCL18.26 The current MOB-ML results are of an accuracy that is similar to the best reported results from DeepHF (an MAE of 1.49 kcal/mol);25 however, MOB-ML only needs to be trained on about 3% of the molecules in the QM7b data set while DeepHF is trained on 42% to obtain comparable results (MAE of 1.52 kcal/mol for 3000 training structures). The best reported result for DeepHF (MAE of 1.49 kcal/mol) was obtained by training on 97% of the molecules in the QM7b data set. We attribute the excellent transferability of MOB-ML to the fact that it focuses on the prediction of orbital-pair contributions, thereby reframing an extrapolation problem into an interpolation problem when training machine learning models on small molecules and testing them on large molecules. The pair correlation energies predicted for QM7b-T and for GDB13-T span a very similar range (0 to −20 kcal/mol), and
FIG. 4. Comparison of the prediction mean absolute errors of total correlation energies for GDB13-T molecules as a function of the number of QM7b-T molecules chosen for model training for different machine learning models: MOB-ML as outlined in this work with random sampling (green circles), MOB-ML with a single GPR\textsuperscript{18} (orange circles), MOB-ML with RCR/GPR\textsuperscript{19} (brown circles), DeepHF\textsuperscript{25} (red squares), FHCL18\textsuperscript{26} (purple squares). The green shaded area corresponds to the 90% confidence interval for the predictions obtained from 50 random samples of the training data.

they are predicted with a similar Gaussian process variance (see Figure S5) which we would expect in an interpolation task. The final errors for GDB13-T are larger than for QM7b-T, because the total correlation energy is size-extensive; however, the size-intensive error per electron pair spans a comparable range for QM7b-T and for GDB13-T (see Figure S4). This presents a significant advantage of MOB-ML over machine learning models which rely on a whole-molecule representation and creates the opportunity to study molecules of a size that are beyond the reach of accurate correlated wave function methods.

Most studies in computational chemistry require accurate relative energies rather than accurate total energies. Therefore, we also assess the errors in the relative energies for the sets of conformers for each molecule in the QM7b-T and in the GDB13-T data sets obtained with MOB-ML with respect to the reference energies (see Figure 5). We emphasize that MOB-ML is not explicitly trained to predict conformer energies, and we include at most one conformer for each molecule in the training set. Nevertheless, MOB-ML produces on average chemically accurate relative conformer energies for QM7b-T when trained on correlation energies for only 30 randomly chosen molecules (or 0.4% of the molecules) in the QM7b set. We obtain chemically accurate relative energies for the GDB13-T data set when training on about 100 QM7b-T molecules. The prediction accuracy improves steadily when training on more QM7b-T molecules reaching a mean MAE of 0.98 kcal/mol for the BBI data set (see Figure 6). However, these predictions are uncertain as indicated by the large Gaussian process variances associated with these data points which strongly suggested that we are now, as expected, in an extrapolative regime. We further improve the predictive capability of MOB-ML by augmenting the MOB-ML model with data from the BBI set. Specifically, we can draw on an active learning strategy and consecutively include data points until all uncertainties are below 1 kcal/mol which in this case corresponds to only two data points. This reduces the MAE to 0.28 kcal/mol for the remaining 98 data points in the BBI set. Including more reference data points would further improve the performance for this specific data set. However, this is not the focus of this work. Instead, we simply emphasize that MOB-ML is a clearly extensible strategy to accurately predict energies for large molecules and non-covalent intermolecular interactions while providing a useful estimation of confidence.

C. Transition-metal complexes

We finally present the first application of MOB-ML to transition-metal complexes. To this end, we train a MOB-ML model on a thermalized subset of mononuclear, octahedral transition-metal complexes introduced by Kulik and co-workers\textsuperscript{40} which we denote as TM-T. The chosen closed-shell transition-metal complexes feature different transition metals (Fe, Co, Ni) and ligands. The ligands span the spectrochemical series from weak-field (e.g., thiocyanate) over to strong-field (e.g., carbonyl) ligands. We see in Figure 7 that the learn-
FIG. 6. Top panel: Errors in predictions were made with a MOB-ML model trained on 20 randomly selected QM7b-T molecules with FS 3 with respect to reference MP2/cc-pVTZ interaction energies for the BBI data set. Bottom panel: Errors in predictions were made with a MOB-ML model trained on 20 randomly selected QM7b-T molecules and augmented with the 2 BBI data points with the largest variance (orange circles) with respect to reference MP2/cc-pVTZ interaction energies. The bar attached to each prediction error indicates the associated Gaussian process variance. The gray shaded area corresponds to the region where the error is smaller than chemical accuracy (1 kcal/mol).

V. CONCLUSIONS

Molecular-orbital-based machine learning (MOB-ML) provides a general framework to learn correlation energies at the cost of molecular orbital generation. In this work, we demonstrate that preservation of physical symmetries and constraints leads to machine-learning methods with greater learning efficiency and transferability. Exploiting physical principles like size consistency and energy invariances not only leads to a conceptually more satisfying method, but it also leads to substantial improvements in prediction errors for different data sets covering total and relative energies for thermally accessible organic and transition-metal containing molecules, non-covalent interactions, and transition-state energies. With the modifications presented in the current work, MOB-ML is shown to be highly data efficient, which is important due to the high computational cost of generating reference correlation energies. Only 1% of the QM7b-T data set (containing organic molecules with seven and fewer heavy atoms) needs to be drawn on to train a MOB-ML model which produces on average chemically accurate total energies for the remaining 99% of the data set. Without ever being trained to predict relative energies, MOB-ML provides chemically accurate relative energies for QM7b-T when training on only 0.4% of the QM7b-T molecules. Furthermore, we have demonstrated that MOB-ML is not restricted to the organic chemistry space and that we are able to apply our framework out-of-the-box to describe a diverse set transition-metal complexes when training on correlation energies for tens of molecules.

Beyond data efficiency, MOB-ML models are shown to be very transferable across chemical space. We demonstrate this transferability by training a MOB-ML model on QM7b-T and predicting energies for a set of molecules with thirteen heavy atoms (GDB13-T). We obtain the best result for GDB13-T reported to date despite only training on 3% of QM7b-T. The successful transferability of MOB-ML is shown to result from its recasting of a typical extrapolation task (i.e., larger molecules) into an interpolation task (i.e., by predicting on the basis of size-intensive orbital-pair contributions). Even when MOB-ML enters an extrapolative regime as identified by a large Gaussian process variance, accurate results can be obtained; for example, we predict the transition-state energy for the proton transfer in malonaldehyde and interaction energies in the protein backbone-backbone interaction data set to chemical accuracy without training on transition-state-like data or non-covalent interactions, respectively. In this case,
the uncertainty estimates also offer a clear avenue for active learning strategies which can further improve the model performance. Active learning offers an attractive way to reduce the number of expensive reference calculations further by picking the most informative molecules to be included in the training set. This provides a general recipe how to evolve a MOB-ML model to describe new regions of chemical space with minimal effort.

Future work will focus on the expansion of MOB-ML to cover more of chemical space. Specifically, particular areas of focus include open-shell systems and electronically excited states. Physical insight from exact conditions in electronic structure theory \(^2\) will continue to guide the development of the method, with the aim of providing a machine-learning approach for energies and properties of arbitrary molecules with controlled error.

ACKNOWLEDGMENTS

This work is supported in part by the U.S. Army Research Laboratory (W911NF-12-2-0023), the U.S. Department of Energy (DE-SC0019390), the Caltech DeLogi Fund, and the Camille and Henry Dreyfus Foundation (Award ML-20-196). T.H. acknowledges funding through an Early PostDoc Mobility Fellowship by the Swiss National Science Foundation (Award P2EZP2_184234). S.J.R.L. thanks the Molecular Software Sciences Institute (MoSSI) for a MoSSI investment fellowship. Computational resources were provided by the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science User Facility supported by the DOE Office of Science under contract DE-AC02-05CH11231.

SUPPORTING INFORMATION

Details on feature generation for all data sets used in this work, definition of error metrics, expanded results for the alkane transferability test, expanded results for the transferability within the organic chemistry space. Features and labels for all data sets used in this work.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available within the article and its supplementary material. Additional data that support the findings of this study are openly available in Caltech Data Repository.\(^3\)

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, 058301 (2012).

3. 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).

4. 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, 2087–2096 (2015).

5. J. Behler, “Perspective: Machine learning potentials for atomistic simulations,” J. Chem. Phys. 145, 170901 (2016).

6. F. Brockerhede, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller, “Bypassing the Kohn-Sham equations with machine learning,” Nat. Comm. 8, 872 (2017).

7. K. Schütt, P.-J. Kindermans, H. E. Sauceda, 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 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017) pp. 991–1001.

8. 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).

9. 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).

10. R. T. McGibbon, A. G. Taube, A. G. Donchev, K. Siva, F. Hernández, C. Harmes, K.-H. Law, J. L. Klepeis, and D. E. Shaw, “Improving the accuracy of moller-plettes perturbation theory with neural networks,” J. Chem. Phys. 147, 161725 (2017).

11. C. R. Collins, G. J. Gordon, O. A. von Lilienfeld, and D. J. Yaron, “Constant size descriptors for accurate machine learning models of molecular properties,” J. Chem. Phys. 148, 241718 (2018).

12. 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).

13. N. Lubbers, J. S. Smith, and K. Barros, “Hierarchical modeling of molecular energies using a deep neural network,” J. Chem. Phys. 148, 241715 (2018).

14. T. T. Nguyen, E. Székey, G. Imbalzano, Ö. Behler, G. Csányi, M. Ceriotti, A. W. Götze, and F. Faesani, “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).

15. 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).

16. Z. Wu, B. Ramusundar, 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, 513–530 (2018).

17. K. Yao, J. E. Herr, D. W. Toth, R. McKinnire, and J. Parkhill, “The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics,” Chem. Sci. 9, 2261–2269 (2018).

18. 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).

19. 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).

20. A. S. Christensen, F. A. Faber, and O. A. von Lilienfeld, “Operators in quantum machine learning: Response properties in chemical space,” J. Chem. Phys. 150, 064105 (2019).

21. A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corinmboeuf, and M. Ceriotti, “Transferable Machine-Learning Model of the Electron Density,” ACS Cent. Sci. 5, 57–64 (2019).

22. J. S. Smith, B. T. Nebgen, R. Zubiakuy, 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, 2903 (2019).
24 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).

25 A. Fabrizio, B. Meyer, and C. Corminboeuf, “Machine learning models of the energy curvature vs particle number for optimal tuning of long-range corrected functionals,” J. Chem. Phys. 152, 154103 (2020).

26 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, 7155–7165 (2020).

27 A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. Anatole von Lilienfeld, “FCHL revisited: Faster and more accurate quantum machine learning,” J. Chem. Phys. 152, 044107 (2020).

28 Z. Liu, L. Lin, Q. Jia, Z. Cheng, Y. Jiang, Y. Guo, and J. Ma, “Transferable Multi-level Attention Neural Network for Accurate Prediction of Quantum Chemistry Properties via Multi-task Learning,” arXiv:2012.15288 [physics].

29 Z. Qiao, M. Welborn, A. Anandkumar, F. R. Manby, and T. F. Miller III, “OrbNet: Deep Learning for Quantum Chemistry Using Symmetry-Adapted Atomic-Orbital Features,” arXiv:2007.08026 [physics].

30 S. Hanhóz, “Machine learning for the solution of the Schrödinger equation,” Mach. Learn.: Sci. Technol. 1, 013002 (2020).

31 R. K. Nesbet, “Brueckner’s Theory and the Method of Superposition of Configurations,” Phys. Rev. 109, 1632–1638 (1958).

32 C. Møller and M. S. Plesset, “Note on an Approximation Treatment for Many-Electron Systems,” Phys. Rev. 46, 145–147 (1934).

33 C. E. Rasmussen and C. K. I. Williams, Gaussian processes for machine learning (MIT Press, Cambridge, MA, 2006) publication Title: Gaussian processes for machine learning.