JCTC Journal of Chemical Theory and Computation

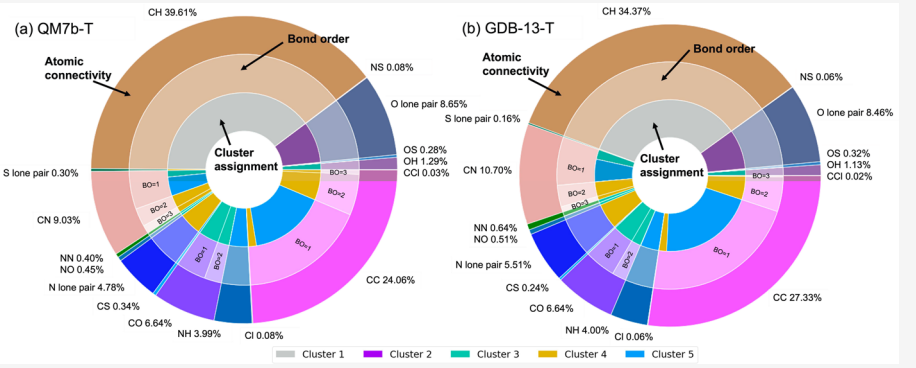
pubs.acs.org/JCTC Article

Accurate Molecular-Orbital-Based Machine Learning Energies via Unsupervised Clustering of Chemical Space

Lixue Cheng, Jiace Sun, and Thomas F. Miller, III*

Cite This: *J. Chem. Theory Comput.* 2022, 18, 4826−4835 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: We introduce an unsupervised clustering algorithm to improve training efficiency and accuracy in predicting energies using molecular-orbital-based machine learning (MOB-ML). This work determines clusters via the Gaussian mixture model (GMM) in an entirely automatic manner and simplifies an earlier supervised clustering approach [*J. Chem. Theory Comput.* **2019**, *15*, 6668] by eliminating both the necessity for user-specified parameters and the training of an additional classifier. Unsupervised clustering results from GMM have the advantages of accurately reproducing chemically intuitive groupings of frontier molecular orbitals and exhibiting improved performance with an increasing number of training examples. The resulting clusters from supervised or unsupervised clustering are further combined with scalable Gaussian process regression (GPR) or linear regression (LR) to learn molecular energies accurately by generating a local regression model in each cluster. Among all four combinations of regressors and clustering methods, GMM combined with scalable exact GPR (GMM/GPR) is the most efficient training protocol for MOB-ML. The numerical tests of molecular energy learning on thermalized data sets of drug-like molecules demonstrate the improved accuracy, transferability, and learning efficiency of GMM/GPR over other training protocols for MOB-ML, i.e., supervised regression clustering combined with GPR (RC/GPR) and GPR without clustering. GMM/GPR also provides the best molecular energy predictions compared with ones from the literature on the same benchmark data sets. With a lower scaling, GMM/GPR has a 10.4-fold speedup in wall-clock training time compared with scalable exact GPR with a training size of 6500 QM7b-T molecules.

1. INTRODUCTION

Machine learning (ML) approaches have attracted considerable interest in the chemical sciences for a variety of applications, including molecular and material design,[1−4,4−8] protein property prediction,[9−11] reaction mechanism discovery,[4,12−16] and analysis and classification tasks for new physical insights.[17−19] As an alternative to physics-based computations, ML has also shown promise for the prediction of molecular energies,[20−34] intermolecular interactions,[31,35] electron densities,[21,24,36−38] and linear response properties.[39−44] ML applications in the chemical sciences often rely on atom- or geometry-specific representations despite the increasing availability and feasibility of wave-function-specific and deep learning representations.[32,33,45−51] Among these recent approaches, molecular-orbital-based machine learning (MOB-ML)[49−53] has been shown to exhibit excellent learning efficiency and transferability for the prediction of energies from post-Hartree−Fock wave function methods.

The previous versions of MOB-ML protocols[49−51] are limited to low training set sizes because of the high computational cost of Gaussian process regression (GPR) training. A local regression with supervised clustering algorithm, termed the regression clustering (RC) algorithm (RC/GPR),[52] has been introduced in our prior work to reduce training costs and enable training on large data sets. RC/GPR

Received: April 20, 2022
Published: July 20, 2022

https://doi.org/10.1021/acs.jctc.2c00396
J. Chem. Theory Comput. 2022, 18, 4826−4835

clusters the training feature space with additional information from the label space of pair energies and classifies the test feature space by an additional classifier trained on the feature space only. The learning efficiency of RC/GPR is mainly affected by the classification errors of the latter. Therefore, it is critical to adapt an efficient unsupervised clustering algorithm to bypass the introduction of this additional classification. Husch et al.[51] proposed an improved MOB feature design, termed a size-consistent feature design, by consistently ordering and numerically adjusting the features. The introduction of this improved MOB feature design not only enhances the prediction accuracy and transferability of MOB-ML but also enables unsupervised clustering in chemical space.

In this work, we apply a more accurate unsupervised clustering method, namely, the Gaussian mixture model (GMM), to cluster systems in the organic chemical space using MOB features. Without any additional information from the label space, the resulting clusters from GMM agree with chemically intuitive groupings of molecular orbital (MO) types. To increase the learning efficiency for molecular energies, we construct the local regression models using a scalable GPR algorithm with exact GP inference but a lower scaling, i.e., the alternative black-box matrix−matrix multiplication (AltBBMM) algorithm, introduced in ref 54. All of the regression with clustering (supervised or unsupervised) methods offer exceptional efficiency and transferability for molecular energy learning. GPR with GMM clustering (GMM/GPR) is the most efficient training protocol for MOB-ML and delivers the best accuracy on QM7b-T and the best transferability on GDB-13-T compared with other models in the literature. It also provides a >10-fold wall-clock training time reduction relative to learning without clustering by AltBBMM.

2. THEORY

2.1. MOB-ML. Molecular energies can be expressed as a sum of Hartree−Fock (HF) energies and correlation energies (E_c). Utilizing Nesbet's theorem,[55,56] which states that E_c for any post-Hartree−Fock wave function theory can be written as a sum over pair energies of occupied MOs, MOB-ML is a strategy that learns pair energies using features comprising elements from the Fock, Coulomb, and exchange matrices.[49] Specifically, Nesbet's theorem is given by eq 1:

$$E_c = \sum_{ij}^{occ} \varepsilon_{ij} \tag{1}$$

where ε_{ij} is the pair energy corresponding to occupied MOs i and j, which is further expressed as a functional of the set of occupied and virtual MOs (eq 2):

$$\varepsilon_{ij} = \varepsilon[\{\phi_p\}^{ij}] \tag{2}$$

MOB features are the unique elements of these matrices between ϕ_i, ϕ_j, and the set of virtual orbitals. For maximum transferability between chemical systems, these properties are computed using localized MOs (LMOs).[49] According to eq 2, ε maps the HF MOs to the pair energies as a universal function, and ε can be approximated with two learned functions, $\varepsilon_d^{ML}[\mathbf{f}_i]$ and $\varepsilon_o^{ML}[\mathbf{f}_{ij}]$, using feature vectors $\mathbf{f}_i$ (corresponding to $\mathbf{f}_{ii}$) and $\mathbf{f}_{ij}$ composed of MOB features, respectively (eq 3):[49−52]

$$\varepsilon_{ij} \approx \begin{cases} \varepsilon_d^{ML}[\mathbf{f}_i] & i = j \\ \varepsilon_o^{ML}[\mathbf{f}_{ij}] & i \neq j \end{cases} \tag{3}$$

In this study, we employ the feature generation and sorting approach described in ref 51.

2.2. Supervised and Unsupervised Clustering Schemes for Chemical Spaces. A straightforward application of GPR with MOB features encounters a bottleneck due to computational demands since GPR introduces complexities of $O(N^2)$ in memory and $O(N^3)$ in training cost. The property of local linearity for MOB features, which allows pair energies to be fitted as a linear function of MOB features within local clusters, has been investigated previously.[52] Thanks to this property, in ref 52 we proposed a comprehensive framework for local regression with clusters to further scale MOB-ML to the large-data regime with lower training costs. Our previous work applied supervised clustering, i.e., regression clustering (RC), to the training set and then performed GPR or linear regression (LR) as local regressors. A random forest classifier (RFC) was also trained to classify the test data.

However, supervised clustering has its limitations. RC requires a predetermined number of clusters and an additional classifier.[52] The performance of the supervised clustering scheme is also hindered by the classifier, which struggles to classify the results from RC because the pair energy label information is provided only to RC and not to the RFC.[52] Therefore, a more precise and efficient clustering and classification strategy is needed to enhance the performance of the entire framework.

Improved MOB feature engineering results in a continuous MOB feature space[51] and consequently enables an unsupervised clustering scheme for MOB-ML. Points that are close in the feature space have similar chemical groupings, cluster identities, and label values, and therefore, distance is an appropriate measure to cluster the MOB feature space. As a result, any distance-based clustering approach should perform well using the improved MOB features. k-means clustering is the simplest and fastest distance-based unsupervised clustering method, and it can effectively cluster the MOB feature space and produce reliable regression results when used in conjunction with GPR. Unfortunately, the lack of an intrinsic probability measure and the unit ball assumption of k-means clustering makes it less accurate than other distance-based clustering methods, such as DBSCAN,[57] OPTICS,[58] and GMM.

GMM can be treated as a generalized k-means method and was chosen for further investigation in this study. It assumes that all N data points belong to a mixture of a certain number of multivariate Gaussian distributions in the feature space with undetermined means and covariances, with each distribution representing a cluster. For K clusters (or Gaussian distributions) with D feature dimensions, the cluster centers (or means of the distributions) $\{\mu_i \in \mathbb{R}^D,\ i = 1, 2, ..., K\}$ and their corresponding covariance matrices $\{\Sigma_i \in \mathbb{R}^{D\times D},\ i = 1, 2, ..., K\}$ are solved by maximizing the likelihood L using the expectation−maximization (EM) algorithm. The expectation, parameters, and cluster identities are computed and reassigned in the expectation (E) stage, and the parameters to maximize the likelihood are updated in the maximization (M) stage. The two stages are repeated until convergence is reached. For a test point, GMM can not only

https://doi.org/10.1021/acs.jctc.2c00396
J. Chem. Theory Comput. 2022, 18, 4826−4835

provide hard cluster assignments with the maximum posterior probability but also enable soft clustering by computation of the normalized posterior probability of a test point belonging to each cluster.[59] To make the GMM training completely automatic, we also perform model selection using the Bayesian information criterion (BIC) to determine the number of clusters used in GMM via scanning of a reasonable series of candidate cluster sizes based on the training set size N. BIC penalizes the likelihood increase due to inclusion of more clusters and more fitting parameters to avoid overfitting with respect to the number of clusters (eq 4):

$$\mathrm{BIC} = q \ln(N) - 2 \ln(L) \tag{4}$$

where q is the number of parameters in the GMM model.

2.3. Local Regression by the Alternative Blackbox Matrix−Matrix Multiplication Algorithm. While the general framework of regression with clustering considerably improves the efficiency of MOB-ML,[52] local regression with full GPR with a cubic time complexity remains the computational bottleneck for MOB-ML. Recently, AltBBMM has been proposed to speed up and scale the GP training in MOB-ML for molecular energies with exact inferences. AltBBMM reduces the training time complexity to $O(N^2)$, which enables training on 1 million pair energies or, equivalently, 6500 QM7b-T molecules and allows the use of multiple GPUs without sacrificing transferability across chemical systems of different molecular sizes. By application of AltBBMM as the local regressor, the efficiency of MOB-ML training can be significantly increased while incurring lower computational costs. The derivation and implementation details for AltBBMM are discussed in ref 54.

3. COMPUTATIONAL DETAILS

The performance of clustering and subsequent local regression approaches is evaluated on the QM7b-T and GDB-13-T benchmark systems, which comprise molecules with at most seven and only 13 heavy atoms (C, O, N, S, and Cl), respectively. Each molecule in QM7b-T or GDB-13-T has seven or six conformers, respectively, and only one conformer of each randomly selected QM7b-T molecule is used for training. The features are computed at HF/cc-pVTZ level with the Boys−Foster localization scheme[60,61] using ENTOS QCORE,[62] and reference MP2[63,64] pair energy labels with the cc-pVTZ basis set[65] are generated from Molpro 2018.0.[66] All of the features, selected features, and reference pair energies employed in the current work are identical to those reported in ref 51.

3.1. Supervised Clustering with MOB features. RC can cluster the organic chemical space represented by QM7b-T and GDB-13-T[52] by maximizing the local linearity of the MOB feature space. On the MOB feature space, we apply the same standard RC protocol introduced in ref 52, using k-means cluster initialization[52] and ordinary least-squares LR implemented using CUPY.[67] The RC step is fully converged (zero training MAE change between two iterations) to obtain the training clusters. Random forest classification (RFC) with 200 balanced trees implemented in SCIKIT-LEARN[52] is performed to classify the test data. To reduce the cost of training the RFC and local regressors for the off-diagonal clusters with a large number of pairs, we adapt the capping strategy illustrated in ref 52 with a capping size of 10 000 for each training off-diagonal cluster during the training over 2000 QM7b-T molecules. No capping is applied to all diagonal and off-diagonal pairs with training sets of fewer than 2000 molecules.

3.2. Unsupervised Clustering with MOB Features. Following the implementation in SCIKIT-LEARN, we reimplement GMM to enable multiple GPU usage using CUPY, which is initialized by k-means clustering and constructed with a full covariance matrix. The objective function of GMM is to maximize the likelihood, which is solved iteratively by the EM algorithm. A regularization of 1×10^{-6} is added to its diagonal terms to ensure that the covariance matrix of GMM is positive-definite.

The number of clusters K_{best} used in GMM is automatically detected by scanning a series of reasonable cluster sizes and finding the GMM model with the most negative BIC score. According to the previous study in ref 52, the optimal numbers of clusters for the diagonal and off-diagonal models of 1000 training molecules in RC are 20 and 70, respectively. The scanning series of possible K values are $\{5i|i = 1, 2, ..., 10\}$ and $\{5i|i = 7, 8, ..., 32\}$ for diagonal and off-diagonal pairs, respectively. Empirical equations for estimating the scanning range of the number of clusters are also presented. We note that this autodetermination procedure is completely unsupervised and does not require any cross-validation from regression.

Hard clustering from GMM assigns the test point to the cluster with the highest probability, and soft clustering from GMM provides probabilities that the point belongs to each possible cluster. Only a few pairs (under 10%) in QM7b-T can have a second-most-probable cluster with a probability over 1×10^{-4} (Table S3). More details about soft clustering are described in the Supporting Information. The current work presents and analyzes the results from hard clustering without specifying any parameters. To demonstrate the smoothness and continuity of GMM clusters constructed using MOB features on the chemical space, the Euclidean distance between the feature vector of each diagonal pair and the corresponding hard cluster center μ_i is computed and analyzed.

3.3. MO Type Determinations. To compare the cluster compositions and the MO compositions in QM7b-T and GDB-13-T, we also apply an algorithm to determine MO types represented by atomic connectivity and bond order for closed-shell molecules following the octet rule. This procedure requires the coordinates of atoms and the centroids of MOs computed using HF information. More details and the pseudocode for the MO detection algorithm (Algorithm S1) are included in the Supporting Information.

3.4. Regression within Local Clusters. Regressions by GPR or LR on top of RC or GMM clustering are used to predict molecular energies. For LR, we use ordinary least-squares LR with no regularization for diagonal and off-diagonal pairs. To reduce the training cost of local GPRs, AltBBMM[54] is performed with the Matérn 5/2 kernel with white noise regularization of 1×10^{-5} for both diagonal and off-diagonal pair energies. For the clusters with fewer than 10 000 training points, GPR models are directly obtained by minimizing the negative logarithm of the marginal likelihood objective with the BFGS algorithm until full convergence. For the clusters with more than 10 000 training points, the variance and length scale are first optimized using 10 000 randomly selected training points within the cluster, and the Woodbury vector[68] is further solved by the block conjugate gradient method with preconditioner sizes of 10 000 and block sizes of 50.

https://doi.org/10.1021/acs.jctc.2c00396
J. Chem. Theory Comput. 2022, 18, 4826−4835

In order to improve the accuracy and reduce the uncertainty, without specifications, the predicted energies are reported as the averages of 10 independent runs for all MOB-ML with clustering protocols. We abbreviate the RC then RFC classification and GPR regression as RC/GPR since no other classifier is used with RC. Similarly, RC then RFC classification and LR regression, GMM clustering with GPR regression, and GMM clustering with LR regression are abbreviated as RC/LR, GMM/GPR, and GMM/LR, respectively. The entire workflow on the general framework of MOB-ML with clustering is also introduced in ref 52.

4. RESULTS AND DISCUSSION

4.1. Number of Clusters Detected in GMM. Rather than predetermining the number of clusters through pilot experiments,[52] GMM automatically selects the most suitable model among the ones with different numbers of clusters by finding the lowest BIC score, which prevents overfitting due to a large number of clusters and is faithful to the intrinsic feature space structure in the training set.[69] Figure 1 depicts the optimal number of clusters determined by BIC scores as a function of the number of training QM7b-T molecules. The numbers of diagonal and off-diagonal clusters are roughly proportional to the training sizes on a logarithmic scale if the training set is larger than 250 molecules, and the best numbers of clusters can be estimated as functions of the number of training molecules N_{mol} from this set of results as K_d and K_o for diagonal and off-diagonal pairs, respectively:

$$K_d = 0.296 N_{mol}^{0.579} \quad (5)$$

$$K_o = 2.117 N_{mol}^{0.502} \quad (6)$$

These two empirical equations serve as estimation functions to avoid searching an excessive amount of candidate clustering numbers. For future multimolecule data set trainings, it is sufficient to construct the scanning region of possible K values as $[K_{est} - 10, K_{est} - 5, K_{est}, K_{est} + 5, K_{est} + 10]$, where K_{est} is the multiple of 5 closest to the estimated value computed from the above empirical equations.

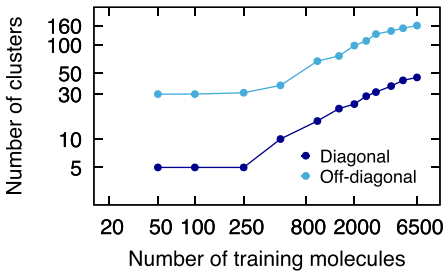


Figure 1. Numbers of clusters in GMMs for diagonal and off-diagonal pairs detected by BIC scores. The average number of clusters over 10 runs is plotted vs the number of training molecules in QM7b-T on a logarithmic scale.

4.2. Unsupervised Clustering Organic Chemical Space. *4.2.1. Chemically Intuitive Clusters from Unsupervised Clustering.* A specific MO can be one-to-one represented by its diagonal feature space, and thus, all of the MO analyses are conducted with the clustering results from GMMs trained on diagonal features using different numbers of QM7b-T molecules. The GMM clustering results and MO types are categorized by multiple pie charts layer by layer for the QM7b-T and GDB-13-T data sets in Figure 2. The first (outermost) layer depicts the atomic connectivity of an MO, which is further classified by the bond order in the second (intermediate) layer. The third (innermost) layer illustrates the classification results for each type of MO obtained from the diagonal GMM model trained on 250 QM7b-T molecules.

MOB-ML has been shown to be transferable in supervised clustering and regression tasks by the creation of an interpolation to weak extrapolation tasks between different chemical systems using the MOB representation.[49−52] When the first and second layers of the two sets of pie charts in Figure 2 are compared, it becomes clear that QM7b-T and GDB-13-T share the same categories of MOs with slightly different abundances. The C−H MO is the most prevalent MO type in QM7b-T, and its popularity declines as the popularity of other less-trained MOs increases in GDB-13-T. This discovery implies that QM7b-T has a majority of the information necessary to predict the properties of molecules in

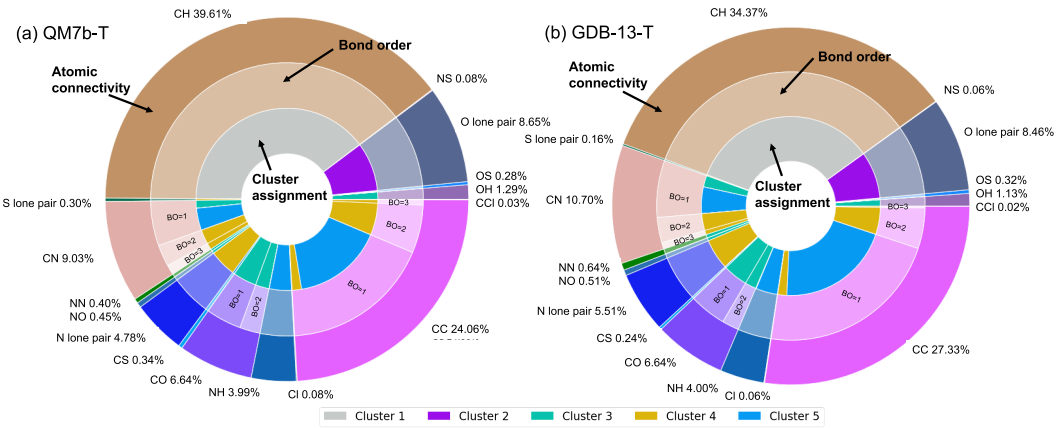


Figure 2. MO types and cluster compositions of (a) QM7b-T and (b) GDB-13-T predicted by a GMM model trained on the diagonal features of 250 QM7b-T molecules (250 GMM model). The layers from outer to inner are the atomic connectivities of MOs, the bond orders (BOs) of the MOs, and the GMM classification results, respectively. The abundance of each type of atomic connectivity in each data set is labeled. The BOs are marked only for cases where one type of atomic connection has more than one possible BO. If the two atoms of an MO can form only a single bond or the MO is a lone pair, the BO is not listed in the figure.

https://doi.org/10.1021/acs.jctc.2c00396
J. Chem. Theory Comput. 2022, 18, 4826−4835

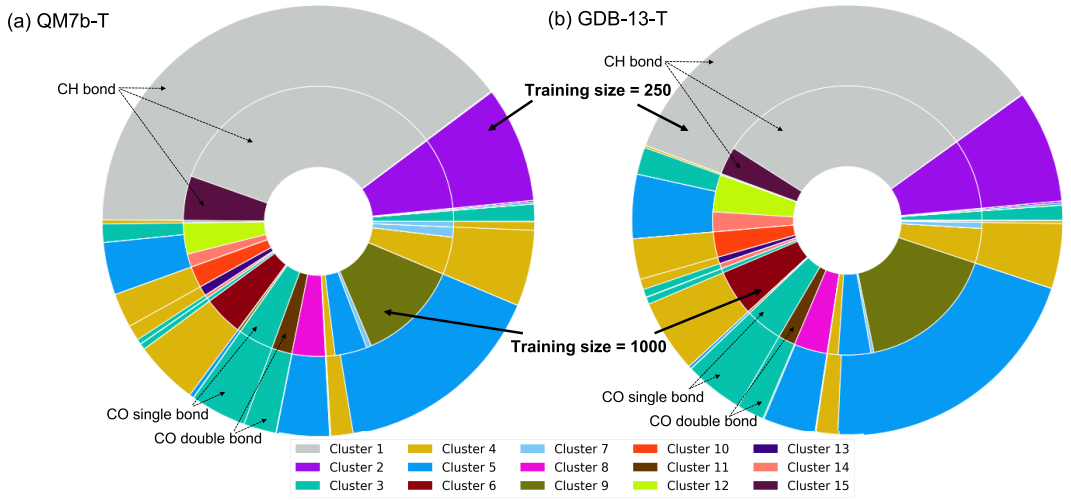


Figure 3. Cluster assignments of (a) QM7b-T and (b) GDB-13-T predicted by the diagonal GMM models trained on 250 (250 GMM model) and 1000 (1000 GMM model) QM7b-T molecules. The outer layers display the same clustering results as the innermost layers in Figure 2 predicted by the GMM model trained on 250 molecules with five detected clusters. The inner layers show the clustering results predicted by the GMM model trained on 1000 molecules with 15 detected clusters. In both panels, the clusters in the inner layers further split up the ones in the outer layers. The MO identities of example clusters analyzed in the main text are labeled in the figure as well.

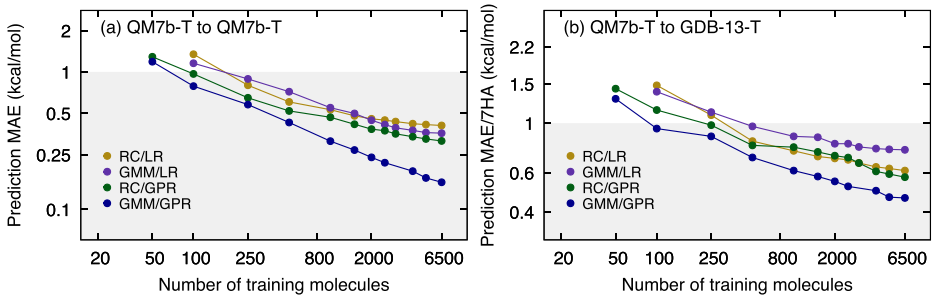


Figure 4. Learning curves for MP2/cc-pVTZ energy predictions with different clustering methods trained on QM7b-T and applied to (a) QM7b-T and (b) GDB-13-T. The models are the same ones trained on QM7b-T for (a) and (b). The prediction performance is reported in terms of (a) MAE and (b) MAE per seven heavy atoms (7HA) by averaging over 10 runs. All of the data are plotted on a logarithmic scale, and the shaded areas correspond to an MAE/7HA of 1 kcal/mol.

GDB-13-T with any MO-based representation and any transferable machine learning approach. The almost identical grouping patterns in the third layers of Figure 2a,b suggest that unsupervised clustering via GMM is transferable, as expected. In addition, the cluster assignments match the chemically intuitive groupings for both QM7b-T and GDB-13-T (Figure 2a, inner layer). Each type of MO is clustered into one type of cluster except for the C−C single bond, while most of the clusters contain more than one type of MO. For example, all C−C double bonds are clustered into Cluster 2, but Cluster 2 contains C−C single, double, and triple bonds, C−N double and triple bonds, and lone pairs on the N atom. More training points are required to capture finer clustering patterns in the chemical space using GMM.

We note that although clustering based on MO types is theoretically feasible, it is not practical as a general approach in MOB-ML to predict molecular properties. It is difficult to intuitively define the types of MOs within chemical systems with complicated electronic structures, such as transition states. On the other hand, MOB features can easily represent these MOs. As demonstrated in the first layers in both charts in Figure 2, the organic chemical space is biased heavily toward C−H and C−C MOs significantly. To avoid overly small clusters and achieve accurate local regression models, careful design of the training sets is also required by inclusion of various MO types for clustering based on MO types. In addition, the local regression models with clusters based on MO types cannot predict the properties of a new type of MO without its explicit inclusion in the training set. Meanwhile, a GMM model trained with MOB features could still classify this MO into a suitable group that has the smallest feature space distances to the MO. Lastly, as an indication of interpretability for the MOB-ML approach, we believe that unsupervised learning with MOB features has great potential to facilitate the exploration of chemical space by inclusion of molecules with MOs close to or far apart from the current space and to propose new molecules by the combination of desired types of MOs.

4.2.2. Resolutions of GMM Clustering with Different Training Set Sizes. As the number of training pairs increases, the number of clusters recognized by GMM for diagonal feature space increases from 5 at 250 training molecules to 15 at 1000 training molecules (Figure 1). Figure 3 compares the clustering patterns predicted by GMMs trained on different training sizes for (a) QM7b-T and (b) GDB-13-T. In both panels, the layers show the cluster compositions determined by

https://doi.org/10.1021/acs.jctc.2c00396
J. Chem. Theory Comput. 2022, 18, 4826−4835

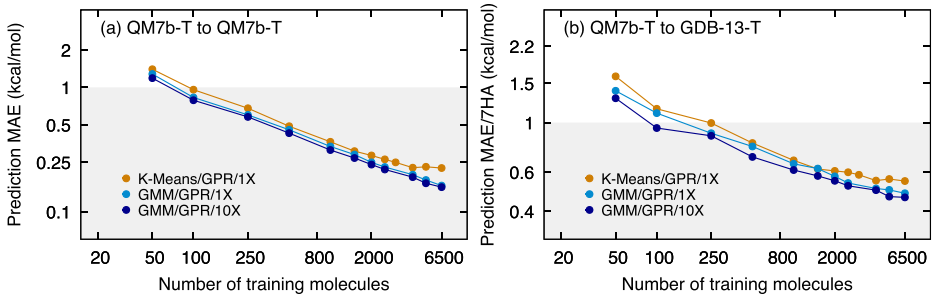


Figure 5. Learning curves for MP2/cc-pVTZ energy predictions with different regression with unsupervised clustering methods trained on QM7b-T and applied to (a) QM7b-T and (b) GDB-13-T. The results for GMM/GPR/10X are the same as the ones for GMM/GPR in Figure 4 and are plotted for comparison. All of the data are plotted on a logarithmic scale, and the shaded areas correspond to an MAE/7HA of 1 kcal/mol.

the GMM trained on 250 molecules (250 GMM model) in the outer layer and 1000 molecules (1000 GMM model) in the inner layer. Training on more molecules not only provides more diverse chemical environments for the same type of MOs but also aids in the resolution of the local structures in the MOB feature space. The MO types with high abundance in QM7b-T and GDB-13-T could be split into multiple clusters. For instance, the one cluster for C−H single bond trained on 250 molecules is split into two clusters trained on 1000 molecules. In addition, the MO types with low abundances in QM7b-T and GDB-13-T could be resolved with more training data rather than mixed into one cluster. For example, C−O single bonds and C−O double bonds are classified into two clusters by the 1000 GMM model instead of one cluster as in the 250 GMM model.

4.3. Molecular Energy Learning by Regression with Clustering. We now present the results of predicting molecular energies utilizing GPR or LR on top of supervised or unsupervised clustering methods in MOB-ML. The RC/LR and GMM/LR training results on 50 molecules are omitted because of the instability of local LR models with 50 training molecules. The prediction accuracy is assessed by the mean absolute error (MAE) of total energies predicted by each MOB-ML model on the test sets, which is plotted as a function of the number of training molecules on a logarithm scale ("learning curve"[70]) in Figure 4. The test sets consist of all remaining QM7b-T thermalized geometries not included in the training sets in Figure 4a and all the GDB-13-T thermalized geometries in Figure 4b. All of the test errors for different training protocols with different training set sizes are reported in Tables S1 and S2.

Among all four training protocols, GMM/GPR provides the best learning accuracy on QM7b-T and transferability on GDB-13-T. By training on 6500 molecules, GMM/GPR can achieve an MAE of 0.157 kcal/mol for QM7b-T and an MAE/7HA of 0.462 kcal/mol for GDB-13-T. The performances of the other three approaches are similar on QM7b-T, but RC/GPR and RC/LR have slightly better performance on GDB-13-T than GMM/LR. For the models clustered by RC, LR provides similar accuracy and transferability compared to GPR since RC maximizes the local linearity for each local cluster. The accuracy loss due to nonlinearity of local regression is more significant with GMM clustering, with an MAE of 0.202 kcal/mol for QM7b-T and an MAE/7HA of 0.298 kcal/mol for GDB-13-T for training on 6500 molecules. Although GMM/LR is not as accurate as GMM/GPR, the reasonably accurate predictions from GMM/LR for both QM7b-T and GDB-13-T infer that GMM still can capture local linearity to some degree, despite the fact that GMM is not trained to maximize local linearity. In comparison with GMM/GPR, the learning efficiency of RC/GPR is harmed by the classification errors from RFC for test points 52, and hence, RC/GPR provides twice as large errors for QM7b-T and 0.111 kcal/mol worse MAE/7HA for GDB-13-T.

With the same clustering method, GPR is a more accurate local regressor compared with LR and generally offers superior accuracy across all training set sizes. GMM/LR has 1.47 to 2.28 times higher MAEs than GMM/GPR, and RC/GPR also marginally outperforms RC/LR. The chemical accuracy of 1 kcal/mol for test QM7b-T molecules can be reached by training on 100 and 250 training molecules using GPR and LR local regressors, respectively, in Figure 4a. In addition, GMM/GPR requires only 100 training molecules to reach the chemical accuracy for GDB-13-T.

4.4. Performances of Alternative Settings of Regression with Unsupervised Clustering. In this section, we show the results of some alternative settings of regression with unsupervised clustering. We compare the accuracy changes with and without averaging over 10 runs and replace the GMM clustering with *k*-means clustering to show the generality of the local GP with unsupervised learning. In Figure 5, the learning curves of molecular energies obtained from different regression with unsupervised clustering learning protocols are plotted and compared with our standard GMM/GPR learning protocol. Although averaging over 10 independent runs offers slightly better prediction accuracy for GMM/GPR, the results from single-run GMM/GPR (GMM/GPR/1X) have only minor accuracy loss. The results obtained from a single run of GMM/GPR (GMM/GPR/1X) are slightly worse than those obtained by averaging over 10 independent runs (GMM/GPR/10X). The best accuracy of GMM/GPR/1X is MAE = 0.162 kcal/mol, which is only 3% worse than that of GMM/GPR/1X for QM7b-T. The transferability of predicting GDB-13-T also remains nearly unchanged. However, RC/GPR/1X is over 10% worse than the corresponding RC/GPR/10X in ref 52. This observation suggests that GMM is not only more accurate but also more stable compared with RC.

This general framework of regression with unsupervised clustering does not restrict the usage of other unsupervised clustering algorithms, such as *k*-means clustering and density-based spatial clustering of applications with noise (DBSCAN). In Figure 5 we include the learning curves for *k*-means with GPR as regressors (*k*-means/GPR/1X). The number of clusters in *k*-means is autodetected by the Davies−Bouldin index,[71] which compares the distance between clusters with the sizes of the clusters themselves,[72] since BIC could not be

https://doi.org/10.1021/acs.jctc.2c00396
J. Chem. Theory Comput. 2022, 18, 4826−4835

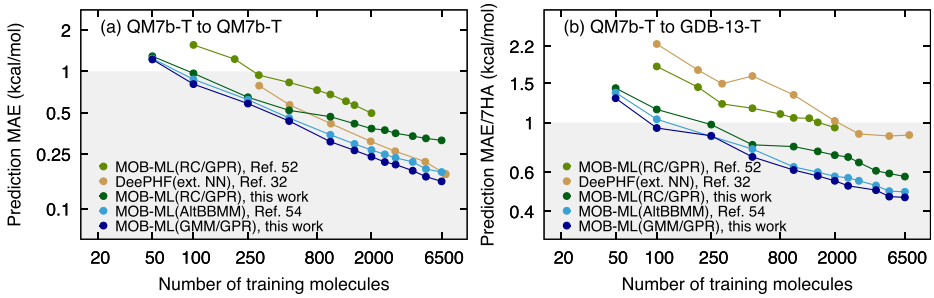


Figure 6. Accuracy comparison between different ML methods trained on QM7b-T and tested on (a) QM7b-T and (b) GDB-13-T. The learning curves of RC/GPR and GMM/GPR are the same ones as shown in Figure 4. The results from RC/GPR in ref 52 were trained on non-size-consistent features and therefore are different from the ones obtained from RC/GPR in this work. In addition, MOB-ML regressed with AltBBMM (MOB-ML (AltBBMM))[54] and DeePHF (ext. NN)[32] are also plotted for comparison. All of the data are digitally extracted from the corresponding studies and plotted on a logarithmic scale. The shaded area corresponds to the chemical accuracy of 1 kcal/mol.

computed without Bayesian inferences. The *k*-means method is found to be a reasonably good choice for clustering and classification. The prediction accuracies of *k*-means/GPR/1X are only at most 36.57% and 10.30% worse for QM7b-T and GDB-13-T than those of GMM/GPR/1X, respectively, which indicates the potential success of using any other unsupervised clustering algorithms in MOB-ML.

4.5. Comparison with Molecular Energy Learning Results from the Literature. In Figure 6, the learning curves for RC/GPR and GMM/GPR in this study are further compared to those of state-of-the-art methods in the literature trained on randomly selected QM7b-T molecules, including MOB-ML regressed with RC/GPR using outdated MOB features from ref 52 (MOB-ML (RC/GPR)), DeePHF trained with an NN regressor[32] (DeePHF (ext. NN)), and MOB-ML regressed with AltBBMM (MOB-ML (AltBBMM)).[54]

The introduction of the most recent improved MOB features[51] considerably enhances the accuracy of MOB-ML, and therefore, RC/GPR from this work is more accurate than the literature RC/GPR with outdated features. Training on the best available MOB features leads to accuracy improvements of over 30% with RC/GPR on both QM7b-T and GDB-13-*t* test molecules. This observation suggests that better feature engineering can not only improve the accuracy for GPR without clustering[51] but also enhance the efficiency of regression with clusters. GMM/GPR achieves slightly higher prediction accuracy than AltBBMM without clustering when the same MOB feature design is used, indicating that an additional GMM clustering step prior to regression benefits the entire training process by replacing the global regression model with more accurate local regression models.

As another machine learning framework to predict molecular energies at HF cost, DeePHF[32] also achieves accurate predictions for QM7b-T, while its transferability to GDB-13-T is less than that of MOB-ML.[51] GMM/GPR training on 6500 molecules (MAE = 0.157 kcal/mol) outperforms the best DeePHF model training on 7000 molecules (MAE = 0.159 kcal/mol) on the total energies of QM7b-T test molecules. Without sacrificing transferability on GDB-13-T, the best model from GMM/GPR can achieve half of the error from DeePHF on GDB-13-T and become the most accurate model for molecular energies in GDB-13-T. OrbNet[73,74] is another approach to predict molecular energies at DFT accuracy using the AO information from GFN-xTB computations. Although OrbNet is one of the state-of-the-art neural network structures, it is considered most suitable for small-basis inputs, for example, GFN-xTB with small basis,[73] to predict DFT energies. It could not easily work directly with any larger basis or HF inputs to predict post-HF energies. In addition, MOB-ML shows much better accuracy to predict the energies than OrbNet-Equi[74] on QM9 in ref 75. Therefore, we only compare the learning curves from DeePHF in this work.

4.6. Efficient Learning by Local AltBBMM with GMM Clustering. To have a fair comparison, we report the timings of single-run regression with clustering models compared with the ones from AltBBMM models in this section. Figure 7 plots

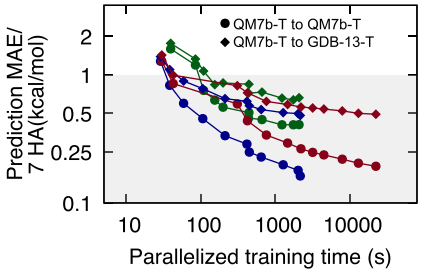


Figure 7. Accuracy and training costs of MP2/cc-pVTZ energy using single-run GPR with RC and GMM clustering (RC/GPR/single-run, green; GMM/GPR/single-run, blue) and AltBBMM without clustering (red, ref 54). Prediction MAEs of test QM7b-T (circles) and GDB-13-T (diamonds) from single runs are plotted as functions of wall-clock training time with parallelization on eight NVIDIA Tesla V100 GPUs on a logarithmic scale. The models are the same as the ones reported in Figure 4, and the corresponding training sizes of QM7b-T are labeled in the figure. The shaded areas correspond to an MAE/7HA of 1 kcal/mol.

the test MAEs of QM7b-T and GDB-13-T from single-run models as a function of parallelized training time on eight NVIDIA Tesla V100-SXM2-32GB GPUs for the three most accurate MOB-ML training protocols. GMM/GPR provides slightly improved accuracy and transferability compared with direct regression by AltBBMM without clustering and significantly reduces the training time of MOB-ML by 10.4-fold with 6500 training molecules. As the most cost-efficient and accurate training protocol for MOB-ML, a single run of GMM/GPR requires only 2170.4 s of wall-clock time to train the best model with 6500 molecules, while AltBBMM needs 22486.4 s to reach a similar accuracy.

We note that the computational costs of GMM and local AltBBMM in GMM/GPR are comparable and lower than that

https://doi.org/10.1021/acs.jctc.2c00396
J. Chem. Theory Comput. 2022, 18, 4826−4835

of AltBBMM without clustering. The complexity analysis is as follows. The training complexity of GMM of each EM iteration is $O(NK)$ with a fixed number of features,[76] where N is the number of training points, which scales linearly with N_{mol}, and K is the number of clusters. Local AltBBMM has a training complexity of $O(KN_{loc}^2)$, where N_{loc} is the number of training points in each local cluster.[68] Since N_{loc} roughly scales as $O(N/K)$, the complexity of local AltBBMM in GMM/GPR can be approximated as $O(N^2/K)$. Therefore, GMM becomes the computational bottleneck in GMM/GPR when K grows faster than $N^{0.5}$; otherwise, local AltBBMM is more expensive than GMM. As discussed in section 4.1, the optimal K_d and K_o for QM7b-T are fitted as functions of N with an approximate scaling of $O(N^{0.579})$ and $O(N^{0.502})$, respectively. GMM and local AltBBMM share similar computational costs in this case, and the overall complexity of GMM/GPR using local AltBBMM is around $O(N^{1.58})$, which is lower than that of AltBBMM without clustering. By the use of this set of complexity analyses, the training size of MOB-ML is no longer limited by GP, and MOB-ML is able to train more than 100 million pair energies (equivalent to 100 000 molecular energies).

5. CONCLUSION

We have extended our previous work on supervised clustering to unsupervised clustering on the organic chemical space with the improved MOB features. An accurate, efficient, and transferable regression with clustering scheme is also introduced to learn the molecular energies of QM7b-T and GDB-13-T. Without specifying the number of clusters, unsupervised clustering via the Gaussian mixture model (GMM) is automatic without human interference and able to cluster the organic chemical space represented by QM7b-T and GDB-13-T in ways consistent with the chemically intuitive groupings of MO types. The finer grouping patterns of MOB feature space are captured as the amount of training data increases, and the resulting clusters are gradually separated following chemical intuition. As the most efficient training protocol for MOB-ML, GMM/GPR surpasses RC/GPR and AltBBM without clustering in prediction accuracy and transferability with a training cost that is a tenth of the cost for AltBBBM without clustering. GMM/GPR not only reaches chemical accuracy for QM7b-T and GDB-13-T by training on only 100 QM7b-T molecules but also offers superior performance to all other state-of-the-art ML methods in the literature, with an MAE of 0.157 kcal/mol for QM7b-T and an MAE/7HA of 0.462 kcal/mol for GDB-13-T. Finally, we have illustrated that the overall complexity of GMM/GPR is lower than that of AltBBMM without clustering and that local AltBBMM regression is no longer the computational bottleneck in GMM/GPR. As a future direction, it is promising to apply GMM/GPR to even larger data sets with more diverse chemistry because of its low computational complexity. The unsupervised nature of GMM also opens an avenue to regress other molecular properties with MOB features by GMM/GPR.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.2c00396.

Detailed descriptions of the algorithm of MO type detection (Algorithm S1), numerical data for Figure 4 (Tables S1 and S2), wall-clock time comparison of RC and GMM on the clustering step only (Figure S1), learning curves of soft clustering from GMM combined with LR regressor (Figure S2), and percentage of pairs whose predicted energies are affected slightly by soft clustering (Table S3) (PDF)

AUTHOR INFORMATION

Corresponding Author

Thomas F. Miller, III − *Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, United States;* orcid.org/0000-0002-1882-5380; Email: tfm@caltech.edu

Authors

Lixue Cheng − *Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, United States;* orcid.org/0000-0002-7329-0585

Jiace Sun − *Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.2c00396

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Dr. Tamara Husch for the guidance on the improved feature generation protocol and Vignesh Bhethanabotla for his help to improve the quality of this manuscript. T.F.M. acknowledges support from the U.S. Army Research Laboratory (W911NF-12-2-0023), the U.S. Department of Energy (DOE) (DE-SC0019390), the Caltech DeLogi Fund, and the Camille and Henry Dreyfus Foundation (Award ML-20-196). 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.

REFERENCES

(1) Gawehn, E.; Hiss, J. A.; Schneider, G. Deep learning in drug discovery. *Mol. Inf.* **2016**, *35*, 3−14.
(2) Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **2018**, *4*, No. eaap7885.
(3) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.* **2016**, *30*, 595.
(4) Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59*, 2545−2559.
(5) Kim, E.; Huang, K.; Jegelka, S.; Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Comput. Mater.* **2017**, *3*, 53.
(6) Ren, F.; Ward, L.; Williams, T.; Laws, K. J.; Wolverton, C.; Hattrick-Simpers, J.; Mehta, A. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci. Adv.* **2018**, *4*, No. eaaq1566.
(7) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.

(8) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360−365.
(9) Yang, K. K.; Wu, Z.; Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **2019**, *16*, 687−694.
(10) Casalino, L.; Dommer, A. C.; Gaieb, Z.; Barros, E. P.; Sztain, T.; Ahn, S.-H.; Trifan, A.; Brace, A.; Bogetti, A. T.; Clyde, A.; Ma, H.; Lee, H.; Turilli, M.; Khalid, S.; Chong, L. T.; Simmerling, C.; Hardy, D. J.; Maia, J. D.; Phillips, J. C.; Kurth, T.; Stern, A. C.; Huang, L.; McCalpin, J. D.; Tatineni, M.; Gibbs, T.; Stone, J. E.; Jha, S.; Ramanathan, A.; Amaro, R. E. AI-driven multiscale simulations illuminate mechanisms of SARS-CoV-2 spike dynamics. *Int. J. High Perform. Comput. Appl.* **2021**, *35*, 432−451.
(11) Gussow, A. B.; Park, A. E.; Borges, A. L.; Shmakov, S. A.; Makarova, K. S.; Wolf, Y. I.; Bondy-Denomy, J.; Koonin, E. V. Machine-learning approach expands the repertoire of anti-CRISPR protein families. *Nat. Commun.* **2020**, *11*, 3784.
(12) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural networks for the prediction of organic chemistry reactions. *ACS Cent. Sci.* **2016**, *2*, 725−732.
(13) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73−76.
(14) Ulissi, Z. W.; Medford, A. J.; Bligaard, T.; Nørskov, J. K. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat. Commun.* **2017**, *8*, 14621.
(15) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using machine learning to predict suitable conditions for organic reactions. *ACS Cent. Sci.* **2018**, *4*, 1465−1476.
(16) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604−610.
(17) Aarva, A.; Deringer, V. L.; Sainio, S.; Laurila, T.; Caro, M. A. Understanding X-ray spectroscopy of carbonaceous materials by combining experiments, density functional theory, and machine learning. Part I: Fingerprint spectra. *Chem. Mater.* **2019**, *31*, 9243−9255.
(18) Zhang, Y.; Tang, Q.; Zhang, Y.; Wang, J.; Stimming, U.; Lee, A. A. Identifying degradation patterns of lithium ion batteries from impedance spectroscopy using machine learning. *Nat. Commun.* **2020**, *11*, 1706.
(19) Magdau, I.-B.; Miller, T. F., III. Machine Learning Solvation Environments in Conductive Polymers: Application to ProDOT-2Hex with Solvent Swelling. *Macromolecules* **2021**, *54*, 3377−3387.
(20) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192−3203.
(21) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **2019**, *10*, 2903.
(22) Lubbers, N.; Smith, J. S.; Barros, K. Hierarchical modeling of molecular energies using a deep neural network. *J. Chem. Phys.* **2018**, *148*, 241715.
(23) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K.-R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.
(24) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513.
(25) Nguyen, T. T.; Székely, E.; Imbalzano, G.; Behler, J.; Csányi, G.; Ceriotti, M.; Götz, A. W.; Paesani, F. Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions. *J. Chem. Phys.* **2018**, *148*, 241725.
(26) Fujikake, S.; Deringer, V. L.; Lee, T. H.; Krynski, M.; Elliott, S. R.; Csányi, G. Gaussian approximation potential modeling of lithium intercalation in carbon nanostructures. *J. Chem. Phys.* **2018**, *148*, 241714.
(27) Li, H.; Collins, C.; Tanha, M.; Gordon, G. J.; Yaron, D. J. A density functional tight binding layer for deep learning of chemical Hamiltonians. *J. Chem. Theory Comput.* **2018**, *14*, 5764−5776.
(28) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **2018**, *120*, 143001.
(29) Nandy, A.; Duan, C.; Janet, J. P.; Gugler, S.; Kulik, H. J. Strategies and Software for Machine Learning Accelerated Discovery in Transition Metal Chemistry. *Ind. Eng. Chem. Res.* **2018**, *57*, 13973−13986.
(30) Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.-R.; Burke, K. Quantum chemical accuracy from density functional approximations via machine learning. *Nat. Commun.* **2020**, *11*, 5223.
(31) Glick, Z. L.; Metcalf, D. P.; Koutsoukas, A.; Spronk, S. A.; Cheney, D. L.; Sherrill, C. D. AP-Net: An atomic-pairwise neural network for smooth and transferable interaction potentials. *J. Chem. Phys.* **2020**, *153*, 044112.
(32) Chen, Y.; Zhang, L.; Wang, H.; E, W. Ground State Energy Functional with Hartree−Fock Efficiency and Chemical Accuracy. *J. Phys. Chem. A* **2020**, *124*, 7155−7165.
(33) Dick, S.; Fernandez-Serra, M. Machine learning accurate exchange and correlation functionals of the electronic density. *Nat. Commun.* **2020**, *11*, 3509.
(34) Christensen, A. S.; von Lilienfeld, O. A. On the role of gradients for machine learning of molecular energies and forces. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045018.
(35) Mezei, P. D.; von Lilienfeld, O. A. Noncovalent Quantum Machine Learning Corrections to Density Functionals. *J. Chem. Theory Comput* **2020**, *16*, 2647−2653.
(36) Grisafi, A.; Wilkins, D. M.; Willatt, M. J.; Ceriotti, M. *Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions*; American Chemical Society, 2019; pp 1−21.
(37) Pereira, F.; Aires-de Sousa, J. Machine learning for the prediction of molecular dipole moments obtained by density functional theory. *J. Cheminf.* **2018**, *10*, 43.
(38) Fabrizio, A.; Grisafi, A.; Meyer, B.; Ceriotti, M.; Corminboeuf, C. Electron density learning of non-covalent systems. *Chem. Sci.* **2019**, *10*, 9424−9432.
(39) Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; von Lilienfeld, O. A. Electronic spectra from TDDFT and machine learning in chemical space. *J. Chem. Phys.* **2015**, *143*, 084111.
(40) Gastegger, M.; Behler, J.; Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **2017**, *8*, 6924−6935.
(41) Yao, K.; Herr, J. E.; Toth, D. W.; McKintyre, R.; Parkhill, J. The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chem. Sci.* **2018**, *9*, 2261−2269.
(42) Christensen, A. S.; Faber, F. A.; von Lilienfeld, O. A. Operators in quantum machine learning: Response properties in chemical space. *J. Chem. Phys.* **2019**, *150*, 064105.
(43) Ghosh, K.; Stuke, A.; Todorović, M.; Jørgensen, P. B.; Schmidt, M. N.; Vehtari, A.; Rinke, P. Deep learning spectroscopy: neural networks for molecular excitation spectra. *Adv. Sci.* **2019**, *6*, 1801367.
(44) Veit, M.; Wilkins, D. M.; Yang, Y.; DiStasio, R. A.; Ceriotti, M. Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles. *J. Chem. Phys.* **2020**, *153*, 024113.
(45) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. Bypassing the Kohn−Sham equations with machine learning. *Nat. Commun.* **2017**, *8*, 872.
(46) McGibbon, R. T.; Taube, A. G.; Donchev, A. G.; Siva, K.; Hernández, F.; Hargus, C.; Law, K.-H.; Klepeis, J. L.; Shaw, D. E. Improving the accuracy of Møller-Plesset perturbation theory with neural networks. *J. Chem. Phys.* **2017**, *147*, 161725.
(47) Nudejima, T.; Ikabata, Y.; Seino, J.; Yoshikawa, T.; Nakai, H. Machine-learned electron correlation model based on correlation

energy density at complete basis set limit. *J. Chem. Phys.* **2019**, *151*, 024104.
(48) Townsend, J.; Vogiatzis, K. D. Data-Driven acceleration of the coupled-cluster singles and doubles iterative solver. *J. Phys. Chem. Lett.* **2019**, *10*, 4129−4135.
(49) Welborn, M.; Cheng, L.; Miller, T. F., III. Transferability in machine learning for electronic structure via the molecular orbital basis. *J. Chem. Theory Comput.* **2018**, *14*, 4772−4779.
(50) Cheng, L.; Welborn, M.; Christensen, A. S.; Miller, T. F., III. A universal density matrix functional from molecular orbital-based machine learning: Transferability across organic molecules. *J. Chem. Phys.* **2019**, *150*, 131103.
(51) Husch, T.; Sun, J.; Cheng, L.; Lee, S. J.; Miller, T. F., III. Improved accuracy and transferability of molecular-orbital-based machine learning: Organics, transition-metal complexes, non-covalent interactions, and transition states. *J. Chem. Phys.* **2021**, *154*, 064108.
(52) Cheng, L.; Kovachki, N. B.; Welborn, M.; Miller, T. F., III. Regression Clustering for Improved Accuracy and Training Costs with Molecular-Orbital-Based Machine Learning. *J. Chem. Theory Comput.* **2019**, *15*, 6668−6677.
(53) Lee, S. J. R.; Husch, T.; Ding, F.; Miller, T. F., III. Analytical gradients for molecular-orbital-based machine learning. *J. Chem. Phys.* **2021**, *154*, 124120.
(54) Sun, J.; Cheng, L.; Miller, T. F., III. Molecular Energy Learning Using Alternative Blackbox Matrix−Matrix Multiplication Algorithm for Exact Gaussian Process. Presented at the NeurIPS 2021 AI for Science Workshop, 2021.
(55) Nesbet, R. K. Brueckner's theory and the method of superposition of configurations. *Phys. Rev.* **1958**, *109*, 1632.
(56) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry*; Dover: Mineola, NY, 1996; pp 231−239.
(57) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996; pp 226−231.
(58) Ankerst, M.; Breunig, M. M.; Kriegel, H.-P.; Sander, J. OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod Rec.* **1999**, *28*, 49−60.
(59) Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer, 2006.
(60) Boys, S. F. Construction of some molecular orbitals to be approximately invariant for changes from one molecule to another. *Rev. Mod. Phys.* **1960**, *32*, 296−299.
(61) Foster, J. M.; Boys, S. F. Canonical Configurational Interaction Procedure. *Rev. Mod. Phys.* **1960**, *32*, 300−302.
(62) Manby, F. R.; Miller, T. F.; Bygrave, P. J.; Ding, F.; Dresselhaus, T.; Batista-Romero, F. A.; Buccheri, A.; Bungey, C.; Lee, S. J. R.; Meli, R.; Miyamoto, K.; Steinmann, C.; Tsuchiya, T.; Welborn, M.; Wiles, T.; Williams, Z. entos: A Quantum Molecular Simulation Package. *ChemRxiv* **2019**, DOI: 10.26434/chemrxiv.7762646.v2.
(63) Møller, C.; Plesset, M. S. Note on an approximation treatment for many-electron systems. *Phys. Rev.* **1934**, *46*, 618.
(64) Saebo, S.; Pulay, P. Local treatment of electron correlation. *Annu. Rev. Phys. Chem.* **1993**, *44*, 213−236.
(65) Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007.
(66) Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M.; Celani, P.; Györffy, W.; Kats, D.; Korona, T.; Lindh, R.; Mitrushenkov, A.; Rauhut, G.; Shamasundar, K. R.; Adler, T. B.; Amos, R. D.; Bennie, S. J.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hesselmann, A.; Hetzer, G.; Hrenar, T.; Jansen, G.; Köppl, C.; Lee, S. J. R.; Liu, Y.; Lloyd, A. W.; Ma, Q.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Miller, T. F., III; Mura, M. E.; Nicklass, A.; O'Neill, D. P.; Palmieri, P.; Peng, D.; Pflüger, K.; Pitzer, R.; Reiher, M.; Shiozaki, T.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Welborn, M. *MOLPRO, a Package of Ab Initio Programs*, ver. 2018.3, 2018; see http://www.molpro.net (accessed 2022-03-04).
(67) Okuta, R.; Unno, Y.; Nishino, D.; Hido, S.; Loomis, C. CuPy: A NumPy-Compatible Library for NVIDIA GPU Calculations. In *Proceedings of the Workshop on Machine Learning Systems (LearningSys) in The 31st Annual Conference on Neural Information Processing Systems (NIPS)*, 2017.
(68) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, 2006.
(69) Findley, D. F. Counterexamples to parsimony and BIC. *Ann. Inst. Stat. Math* **1991**, *43*, 505−514.
(70) Cortes, C.; Jackel, L. D.; Solla, S. A.; Vapnik, V.; Denker, J. S. In *Advances in Neural Information Processing Systems 6*; Cowan, J. D., Tesauro, G., Alspector, J., Eds.; Morgan-Kaufmann, 1994; pp 327−334.
(71) Davies, D. L.; Bouldin, D. W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1*, 224−227.
(72) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn Res.* **2011**, *12*, 2825.
(73) Qiao, Z.; Welborn, M.; Anandkumar, A.; Manby, F. R.; Miller, T. F., III. OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *J. Chem. Phys.* **2020**, *153*, 124111.
(74) Qiao, Z.; Christensen, A. S.; Welborn, M.; Manby, F. R.; Anandkumar, A.; Miller, T. F., III. Informing Geometric Deep Learning with Electronic Interactions to Accelerate Quantum Chemistry. *arXiv (Computer Science.Machine Learning)*, April 1, 2022, 2105.14655, ver. 4. https://arxiv.org/abs/2105.14655 (accessed 2022-06-10).
(75) Sun, J.; Cheng, L.; Miller, T. F., III. Molecular Dipole Moment Learning via Rotationally Equivariant Gaussian Process Regression with Derivatives in Molecular-Orbital-Based Machine Learning. *arXiv (Physics.Chemical Physics)*, May 21, 2022, 2205.15510, ver. 1. https://arxiv.org/abs/2205.15510 (accessed 2022-06-10).
(76) Pinto, R. C.; Engel, P. M. A fast incremental Gaussian mixture model. *PLoS One* **2015**, *10*, No. e0139931.