THE JOURNAL OF PHYSICAL CHEMISTRY A

pubs.acs.org/JPCA Article

Ground State Energy Functional with Hartree–Fock Efficiency and Chemical Accuracy

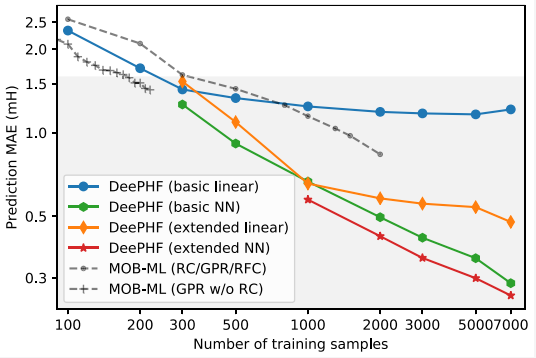
Published as part of The Journal of Physical Chemistry virtual special issue "Machine Learning in Physical Chemistry".

Yixiao Chen, Linfeng Zhang, Han Wang,* and Weinan E*

Cite This: *J. Phys. Chem. A* 2020, 124, 7155–7165 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: We introduce the deep post Hartree–Fock (DeePHF) method, a machine learning-based scheme for constructing accurate and transferable models for the ground-state energy of electronic structure problems. DeePHF predicts the energy difference between results of highly accurate models such as the coupled cluster method and low accuracy models such as the Hartree–Fock (HF) method, using the ground-state electronic orbitals as the input. It preserves all the symmetries of the original high accuracy model. The added computational cost is less than that of the reference HF or DFT and scales linearly with respect to system size. We examine the performance of DeePHF on organic molecular systems using publicly available data sets and obtain the state-of-art performance, particularly on large data sets.



INTRODUCTION

Predicting the ground-state energy of a many-electron system in an environment of clamped ions is one of the most important problems in quantum chemistry. There is a well-known trade-off between efficiency and accuracy. Low level models, such as density functional theory (DFT)[1] and Hartree–Fock (HF),[2] are quite efficient, but their accuracy is often less than adequate. Higher-order schemes based on, e.g., the Møller–Plesset perturbation,[3] coupled cluster,[4] configuration interaction,[5] or adoption of multiple references[6] can generate much more accurate energies, but their much increased computational expense limits their application to no more than dozens of electrons.

In recent years, machine learning (ML) methods have brought some new hope in this difficult area. Significant progress has been made by using ML-based models to represent the ground-state electron energy directly as a function of the positions of the ions and their chemical species.[7−15] Combined with the state-of-art high performance capabilities, molecular dynamics simulations of systems of up to 100 million atoms have been performed, with an accuracy comparable to that of electronic structure models.[16] However, such atom-based ML models usually do not transfer well between different chemical environments. At the same time, the amount of training data required, in terms of the number of atomic configurations and the system size, is beyond the current capability of high level methods such as CCSD(T).

The recently developed molecular-orbital-based machine learning (MOB-ML) method[17−19] has followed a different route. The idea is to fit directly the post-HF correlation energy using information from the electronic orbitals from HF solutions. Since these models do not rely on any atomic-based features, they exhibit much better transferability, with only a small amount of training data.

In this work, we propose an alternative approach in the spirit of the MOB-ML scheme. We suggest a systematic way of devising features from the HF ground state electronic orbitals, and we build our model so that the commonly desired properties including universality, locality, and symmetry are all satisfactorily addressed. We call this methodology DeePHF, standing for the deep post Hartree–Fock method. We also propose an active learning procedure, which enables us to develop a uniformly accurate model with a minimal set of training data. Our model exhibits the up-to-date best performance on the same set of benchmark tests used in previous work.[18−20]

Received: April 30, 2020
Revised: July 28, 2020
Published: August 3, 2020

Downloaded via XI'AN JIAOTONG UNIV on November 6, 2024 at 15:00:08 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ACS Publications © 2020 American Chemical Society

https://dx.doi.org/10.1021/acs.jpca.0c03886
J. Phys. Chem. A 2020, 124, 7155–7165

The scheme proposed here also serves as a preliminary step for our next objective: Developing systematic generalized Kohn−Sham models with uniform chemical accuracy. A key ingredient there is the exchange-correlation (XC) functional. The scheme proposed here is developed for that purpose.

METHODS

Problem Setup. Consider a system containing N electrons indexed by i and M clamped ions indexed by I. Our starting point is the Hartree−Fock (HF) orbitals $\{|\psi_i\rangle\}$, which are solutions of

$$\hat{H}_{\mathrm{HF}}|\psi_i\rangle = \varepsilon_i|\psi_i\rangle;\ \langle\psi_i|\psi_j\rangle = \delta_{ij} \tag{1}$$

with the HF Hamiltonian $\hat{H}_{\mathrm{HF}}$ defined as

$$\hat{H}_{\mathrm{HF}} = \hat{H}_0 + \sum_i^N \hat{J}_i + \sum_i^N \hat{K}_i \tag{2}$$

Here $\hat{H}_0 = -\frac{1}{2}\nabla^2 + \sum_I \frac{Z_I}{|R_I - r|}$ is the single electron non-interacting Hamiltonian, $\hat{J}_i$ is the Coulomb operator, and $\hat{K}_i$ denotes the exchange operator. For simplicity we have also used i to index the orbitals. Note that the full set of solutions $\{|\psi_i\rangle\}$ to eq 1 is a complete basis set and spans the full Hilbert space. The eigenvalues $\{\varepsilon_i\}$ are real and ordered non-decreasingly, so orbitals with $i \le N$ are occupied orbitals, while the unoccupied orbitals, or virtual orbitals, are indexed by $i > N$.

Our goal is to model the energy difference between CCSD(T) and HF models $E_c = E_{\mathrm{EXACT}} - E_{\mathrm{HF}}$ as a function of the HF single-electron orbitals $\{|\psi_i\rangle\}$:

$$E_c \equiv E_c[\{|\psi_i\rangle\}]$$

In the context of HF, E_c is nothing but the correlation energy. The existence of such a functional is an obvious fact for the quantum chemistry community. Our objective is to construct an accurate and transferable approximation to this functional. For this work, we will define E_{EXACT} as $E_{\mathrm{CCSD(T)}}$.

Ideally we would like our model to have the following features.

1. *Universality.* The model should be universally applicable for a large range of systems. It should be noted that, at this point, we are not aiming at true universality. Instead we take a more programmatic approach and aim for models that are applicable for all the systems whose local electronic configurations are well represented by the training data. For this purpose, the input of the model should be purely electronic data; e.g., no information about the chemical species should be used.
2. *Locality.* The model should be relatively local, so that it can potentially be constructed using data from small systems and is generalizable to larger ones.
3. *Symmetry.* The model should respect both physical and gauge symmetries. Here physical symmetry means that E_c should be invariant under translation and rotation of R_I as well as permutation of the ionic indices for the same chemical species. Gauge symmetry means that E_c should be invariant when the occupied orbitals $\{|\psi_i\rangle\}$ undergo a unitary transformation.
4. *Accuracy.* The model should at least achieve chemical accuracy, i.e., a prediction error lower than 1 kcal/mol.
5. *Efficiency.* The cost of solving the model should be comparable to that of HF models.

The Energy Model. We draw inspirations from the density (matrix) functional theory. It was proved decades ago that the ground state energy is a universal functional of the electron density $n(x)$.[21] Therefore, it is also a unique functional of the one particle density matrix $\Gamma(x, x')$.[22] Following the same procedure as for the density functional theory,[1] for the corresponding noninteracting system, we define the density matrix as

$$\Gamma(x, x') = \sum_i^N \langle x|\psi_i\rangle\langle\psi_i|x'\rangle = \sum_i^N \psi_i(x)\psi_i^*(x') \tag{3}$$

It is straightforward to see that $\Gamma(x, x')$ is invariant under gauge transformation. The correlation energy E_c is therefore also a functional of the density matrix $\Gamma(x, x')$.

Next we introduce a set of atomic bases centered on each atom I, denoted as $\{|\alpha^I_{nlm}\rangle\}$:

$$\langle \mathbf{r}|\alpha^I_{nlm}\rangle = R^I_n(r)Y^I_{lm}(\theta, \phi) \tag{4}$$

where $R_n(r)$ is a radial function indexed by n, Y_{lm} denotes the spherical harmonics of degree l and order m, and we use (r, θ, ϕ) to denote the spherical coordinates relative to atom I. For each atom, we project every orbital onto the basis centered at that atom to obtain a set of overlap coefficients

$$c^I_{inlm} = \langle\alpha^I_{nlm}|\psi_i\rangle \tag{5}$$

Note that in practice we will have $|\psi_i\rangle = \sum_a \lambda_{ia}|\chi_a\rangle$, where $\{|\chi_a\rangle\}$ is the basis set used to perform the HF calculation, $\{\lambda_{ia}\}$ are the corresponding expansion coefficients. Therefore, $c^I_{inlm} = \sum_a \lambda_{ia}\langle\alpha^I_{nlm}|\chi_a\rangle$ (since all basis are real). The density matrix can be represented using the basis set $\{|\alpha^I_{nlm}\rangle\}$ as

$$\begin{aligned}\mathcal{D}^{II'}_{nlm;n'l'm'} &= \sum_i c^I_{inlm}c^{I'}_{in'l'm'} = \sum_i \langle\alpha^I_{nlm}|\psi_i\rangle\langle\psi_i|\alpha^{I'}_{n'l'm'}\rangle \\ &= \sum_{i,a,b} \lambda_{ia}\lambda_{ib}\langle\alpha^I_{nlm}|\chi_a\rangle\langle\chi_b|\alpha^{I'}_{n'l'm'}\rangle\end{aligned} \tag{6}$$

Our task is now reduced to the modeling of $E_c(\{\mathcal{D}^{II'}_{nlm;n'l'm'}\})$. We first discuss the issue of locality and symmetry of $E_c(\mathcal{D}^{II'}_{nlm;n'l'm'})$. To our surprise, for all the cases tested (see next section), it suffices to use the "local density matrix"

$$(\mathcal{D}^I_{nl})_{mm'} = \mathcal{D}^{II}_{nlm;nlm'} = \sum_i c^I_{inlm}c^I_{inlm'} \tag{7}$$

To deal with rotational symmetry of the basis $|\alpha^I_{nlm}\rangle$, we use the eigenvalues of the local density matrix as our descriptor

$$\mathbf{d}^I_{nl} = \mathrm{EigenVals}_{mm'}[(\mathcal{D}^I_{nl})_{mm'}] \tag{8}$$

Note that $\mathbf{d}^I_{nl}$ is a vector with $2l + 1$ components. It is also the square of the singular values of the overlap coefficients

$$\mathbf{d}^I_{nl} = (\mathrm{SingularVals}_{im}[c^I_{inlm}])^2 \tag{9}$$

Finally, our model for E_c takes the form

$$E_c = \sum_I \epsilon^I_c = \sum_I \mathcal{F}(\mathbf{d}^I) \tag{10}$$

where $\mathbf{d}^I$ denotes the flattened vector made from $\{\mathbf{d}^I_{nl}\}$ for all n and l.

https://dx.doi.org/10.1021/acs.jpca.0c03886
J. Phys. Chem. A 2020, 124, 7155−7165

It is straightforward to verify that the resulting E_c model is symmetry preserving. Moreover, since each atomic contribution depends locally on the projection coefficients, the model is, by construction, fairly local and can be trained using data obtained from small systems. Finally, all the operations involved here are less expensive than a typical HF calculation. Therefore, as a tool for postprocessing HF results, the model is very efficient.

In order to approximate the function $\mathcal{F}$, we propose two widely used machine learning ansatz, linear function $\mathcal{F}^{\text{lin}}$ and neural network $\mathcal{F}^{\text{nn}}$, and determine their parameters by fitting the data.

For linear functions, we have

$$\mathcal{F} = \mathcal{F}^{\text{lin}}(\mathbf{d}) = \mathbf{W}\cdot\mathbf{d} + \mathbf{b}$$

where $\mathbf{W}$ and $\mathbf{b}$ are parameters learned from data. we use ordinary least-squares (OLS) regression (or Ridge regression when there are two few training samples) as the training procedure.

For neural network fitting functions $\mathcal{F} = \mathcal{F}^{\text{nn}}$, we use standard fully connected feed forward neural network with skip connection[23] between each layers. The detailed parameters of the neural network model do not influence the results much. In our cases, we use three hidden layers and 100 neurons per hidden layer. We use GELU[24] as the activation function for every layer except the final one. The neural network is trained by the gradient descent method using Adam optimizer[25] (β_1 = 0.9, β_2 = 0.999, $\varepsilon = 1 \times 10^{-8}$) with $\mathcal{L}^2$ loss function and batch size 16. The learning rate is set to 1×10^{-4} at the beginning and decays exponentially with a factor of 0.98 for every 500 epochs. No regularization method is used in training the neural network.

We will see later that for small and simple systems, linear regression is already quite accurate. When the systems get larger and more complex, the neural network model helps to improve the accuracy. We also note that when the number of training samples is very small, the neural network model gives worse results, due to overfitting.

We remark that, unlike conventional post-HF methods, it is not critical to choose HF solutions as the starting point of the DeePHF scheme and to choose the correlation energy as the target. As we will show with numerical examples, this scheme can also be applied when DFT is used as the starting point.

Extensions. We introduce several extensions of the energy model, which can be used to further improve the accuracy of the model.

First, we may insert a Hermitian operator $\hat{O}$ in the overlap coefficients

$$\tilde{c}^I_{inlm} = \langle \alpha^I_{nlm} | \hat{O} | \psi_i \rangle \tag{11}$$

As the most straightforward extension, instead of using the bare "local density matrix" $\mathcal{D}^I_{nl}$, we can use a set of kernel functions of the energy $\{f_k(\varepsilon)\}$ by taking $\hat{O}^2$ to be $f_k(\hat{H}_{HF})$. Now the extended "density matrix" can be defined as

$$(\tilde{\mathcal{D}}^I_{nlk})_{mm'} = \sum_i c^I_{inlm} c^I_{inlm} f_k(\varepsilon_i) \tag{12}$$

There are no specific requirements for the kernel functions $\{f_k\}$ other than that they should be normalized in some way so that shifting the energy will not influence the descriptor. The extended descriptor is defined in the same way:

$$\tilde{\mathbf{d}}^I_{nlk} = \text{EigenVals}_{mm'}[(\tilde{\mathcal{D}}^I_{nlk})_{mm'}] \tag{13}$$

This will be used later.

Second, the choice of the basis set $\langle \mathbf{r}_i | \alpha_{nlm} \rangle = R_n(r_i) Y_{lm}(\theta_i, \phi_i)$, particularly for $R_n(r_i)$ as well as the operator $\hat{O}$, can be made adaptive to the training data. This can be done by introducing some trainable parameters into their expressions. An analogy of such an operation can be found in the construction of the DP model,[15] in which an embedding network is used to define invariant features of the neighboring environment of each atom, and a fitting network is used afterward to map the features onto the final output.

Finally, more nonlocal information can be extracted from the density matrix involving two atoms

$$(\mathcal{D}^{IJ}_{nlk})_{mm'} = w(r^{IJ}) \sum_i c^I_{inlm} c^J_{inlm} f_k(\varepsilon_i) \tag{14}$$

and the descriptor can be calculated accordingly. Here $w(r^{IJ})$ is some weight function (of the distance between the two atoms). To ensure that the eigenvalues stay real, we may use instead:

$$\mathcal{D}^{I;\text{nonlocal}}_{nlk} = \frac{1}{2} \sum_J (\mathcal{D}^{IJ}_{nlk} + \mathcal{D}^{JI}_{nlk}) \tag{15}$$

An Active Learning Procedure. Another important problem in our machine learning method is that the labels (in our case the CCSD(T) energies) are very expensive to obtain. Even a single data point can take hours to calculate. Therefore, it is important to build the model to be as sample-efficient as possible. To address this issue, we borrow ideas from active learning that aims to learn models with a minimum number of labeled data. We need an algorithm that can efficiently go over the unlabeled data and determine which ones should be labeled and put into the training set.

In order to do this, we train multiple neural network models on the same data set with different initialization, and use the standard deviation of the output from the different models

$$\varepsilon = \sqrt{\langle (E_c^{\text{nn}} - \langle E_c^{\text{nn}} \rangle)^2 \rangle} \tag{16}$$

as an indicator for the error in the model. This quantity is called the model deviation in ref 26. Here $\langle \cdots \rangle$ denotes averaging over the different models. The reason that the model deviation can serve as a good error indicator is that the landscape for training neural network models is highly nonconvex and often highly overparametrized. Therefore, different initializations usually lead to different minimizers after the training process. For data points close to the training set, the predictions from different minimizers should all be quite accurate and therefore close to each other, giving rise to small values of the model deviation. But for those that are far from the training set, the different minimizers should give different predictions, resulting in larger values of model deviation.

The idea is to simply choose data points with the largest value of model deviation to be labeled and add them to the training set. This allows us to work with the most representative subsamples of the unlabeled data set, thereby improving the sample efficiency.

https://dx.doi.org/10.1021/acs.jpca.0c03886
J. Phys. Chem. A 2020, 124, 7155−7165

Comparison with State-of-the-Art Methods. The methodology presented here is an alternative version of MOB-ML, although our ultimate goal might be different from that of MOB-ML. MOB-ML uses features that come from the matrix elements of Fock, Coulomb, and exchange operators on a set of localized molecular orbitals. Both occupied and virtual orbitals are involved. In DeePHF, the features come directly from the projected density matrix of the ground state wave function. No gauge transformation for localized molecular orbitals is used. Second, in MOB-ML, the correlation energy is decomposed into contribution from each orbitals, and the per orbital value is directly used as the label. In DeePHF, we only use correlation energy itself as the label. Finally, instead of using kernel-based methods such as the Gaussian process regression, we use linear regression or neural network model as the fitting function.

There are also methods aiming at parametrizing an exchange-correlation functional in Kohn−Sham DFT approach. NeuralXC[20] defines such a functional by neural network which takes as input the projected density onto an atom-centered basis set, and employs a self-consistent-field (SCF) procedure to train it. DeePHF, on the other hand, uses a projected density matrix that contains more information. We also examined the DeePHF method on a large data set containing thousands of different molecules, which is not shown in the NeuralXC article.[20]

RESULTS AND DISCUSSION

To examine the performance of the DeePHF scheme, we follow the testing procedure in ref 18, utilizing their publicly available data set.[27] We also included a larger data set ANI-1ccx[28,29] to further test the transferability of the model. We train and test our model respectively on the following sets of data: (i) single water molecules, (ii) short alkanes, and (iii) small organic molecules. All training and testing data are disjoint. Unless otherwise specified, we use the CCSD(T) correlation energies as labels. In order to obtain the descriptors, we use the geometries provided by the data set to perform the Hartree−Fock calculation. The calculation is done using PySCF[30] with a cc-pVTZ basis.[31] The results for HF energies are very close to the values provided in the data set—the differences are normally smaller than 0.1 mH and should be a result of numerical error. This difference should have no influence on our results.

In order to speed up the calculation, the radial part of the basis function is constructed by linear combinations of Gaussians centered at zero, compatible with the commonly used Gaussian-type orbitals (GTOs) in quantum chemistry. The angular index l is taken to be $l \in \{0, 1, 2\}$, resulting in 1, 3, or 5 values of m respectively. Unlike the traditional GTOs, we use the same set of radial functions for all angular indices. The detailed coefficients can be found in Appendix A. Normally we use 12 different radial functions, resulting in a total of $12 \times (1 + 3 + 5) = 108$ basis functions per atom. However, for the test on short alkanes, due to the limited number of training data, we use 9 radial functions and thus 81 basis functions. It is worth noting that as long as a sufficient number of radial functions are used, the specific number of Gaussians and the associated parameters have little influence on the results.

Simple Systems. We first examine the performance of DeePHF for relatively simple systems, namely (i) single water molecules and (ii) short alkanes. Since these systems are simple enough, we only use our basic descriptor $\{\mathbf{d}^I_{nl}\}$ defined in eq 8 and use a linear regression $\mathcal{F} = \mathcal{F}^{\text{lin}}$ to fit the correlation energy. Extended descriptor and more complex fitting functions can be used, but they do not give significantly better test accuracy.

The first system we tested is single water molecules. We trained our model on different numbers of randomly drawn samples from 1000 water geometries in the data set, and used the rest as the testing set. We examined the mean absolute error (MAE) and maximum absolute error (MaxAE) along the learning curve. When the training set is smaller than the dimension of the descriptors (108), we added a small ridge regularization coefficient ($\alpha = 10^{-9}$). We repeated these steps for 1000 times and report the average MAE.

The resulted learning curve is shown in Figure 1. Generally speaking, the results are comparable with that of MOB-ML.[18]

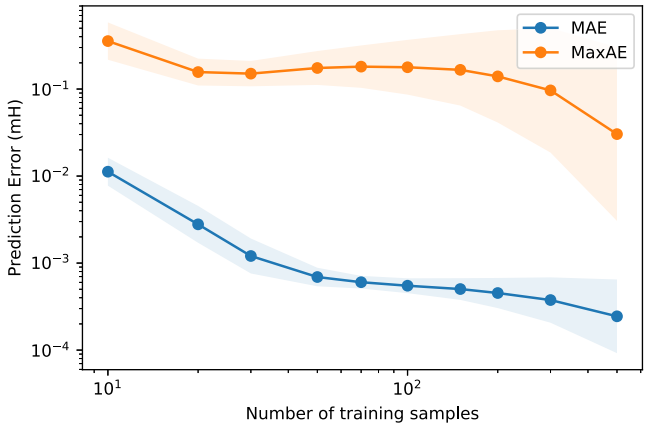


Figure 1. Learning curve of DeePHF for water molecules. Both mean absolute error (MAE) and maximum absolute error are shown (MaxAE). Shaded area shows the size of the standard deviation of 1000 different runs.

With only 10 training samples, we can predict the energy of all the testing data with errors less than 1 mH. Adding more training samples will result in smaller mean absolute error. However, the maximum absolute error behaves differently and has a much larger variance. This is because the maximum absolute error is strongly influenced by the few specific data points whose errors may be hard to reduce by adding training data.

Another relatively simple data set we examined is short alkanes. The data set contains 100 geometries of methane, 1000 of ethane, 1001 of propane, and 101 of both *n*-butane and isobutane. This data set has been used to test the transferability of the model by many authors, e.g., refs 18 and 20.

The same linear fitting function and basic descriptor was used as for the case of water. Following the procedure in ref 18, we randomly chose 50 ethane molecules and 20 propane molecules to train a linear model by OLS with a small ridge regularization coefficient $\alpha = 10^{-7}$ and normalized descriptors. We then used this model to predict the correlation energy of *n*-butane and isobutane molecules. We repeated the training process 10000 times and recorded the resulted mean error (ME), mean absolute error (MAE), maximum absolute error (MaxAE), and mean absolute error after applying a global shift for each type of molecule that sets the mean error to zero (relative error, MARE).

https://dx.doi.org/10.1021/acs.jpca.0c03886
J. Phys. Chem. A 2020, 124, 7155−7165

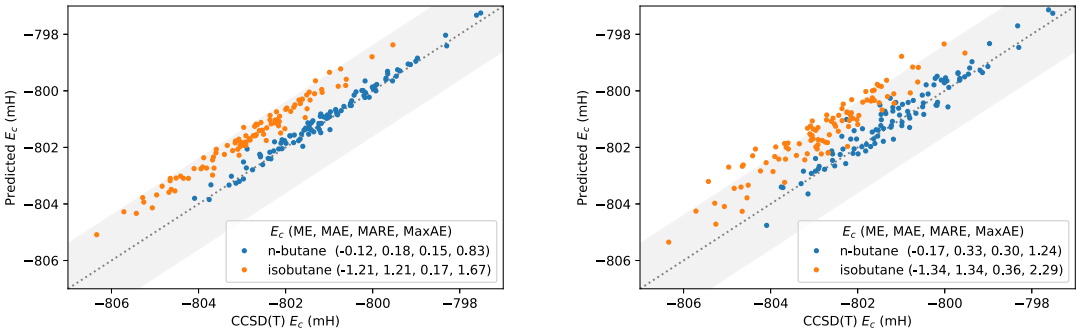


Figure 2. Predicted correlation energy of DeePHF versus the true CCSD(T) energy of *n*-butane and isobutane. Two different results from different training samples are shown. The one on the left is a relatively better result. The one on the right is worse. But these are just results of randomly sampled training data sets. Gray shaded area corresponds to the region where the error is smaller than chemical accuracy (1 kcal/mol). No global shift is applied on these values.

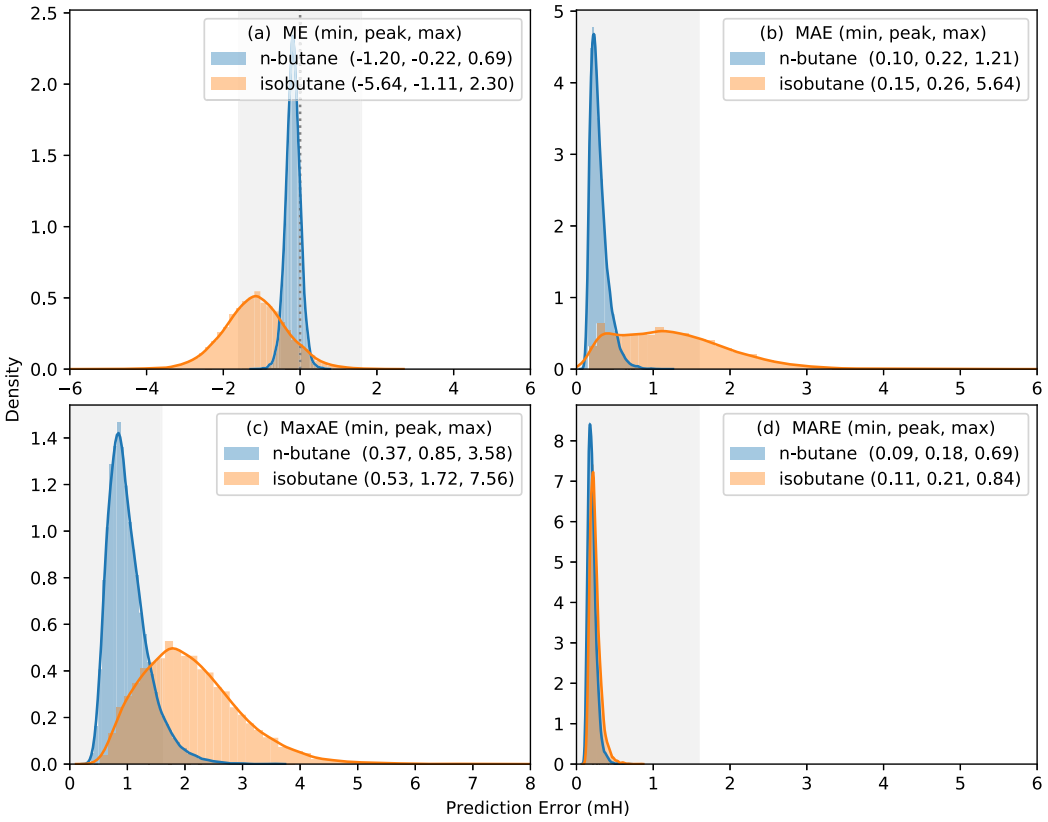


Figure 3. Distributions of prediction error based on different metrics when DeePHF is tested on *n*-butane and isobutane, including (a) mean error (ME), (b) mean absolute error (MAE), (c) maximum absolute error (MaxAE), and (d) mean absolute error after global shift (MARE). The gray shaded area corresponds to the region where the error value is smaller than chemical accuracy (1 kcal/mol). The minimum, the peak, and the maximum values of the distribution are shown in the legend. Note the ranges of X axis are different, except for parts b and d.

The distributions of these errors is shown in Figure 3. If we only look at the error statistics in the better one of Figure 2 or the peak values of the distributions in Figure 3, we might conclude that the model transfers well: almost all the predictions are within chemical accuracy to the ground truth. These results also outperform the current best results reported in refs 18 and 20 (since relative energies were used in these literature, we have to use the MARE metric in order to compare with their results). A detailed comparison can be found in Table 1.

However, the other results shown in Figure 2 are much worse, suggesting that there is a large variance in the test error for these models when different random samples of the training data are selected. This also suggests that the transferability concluded above is not really genuine.

After a closer look we notice that both Figure 3a and Figure 2 exhibit constant bias on the predicted energies. This problem is much more severe for the case of isobutane. One can apply a sample-dependent global shift to reduce the bias, as was done in refs 18 and 20. Indeed, we also found that the relative error with such a global shift applied has much better distribution. However, this is not really practical since the amount of the shift required depends information from the testing set.

One intuitive explanation for the error distribution is to consider the atomic environment from which we build our descriptor. For isobutane, there is a carbon atom connecting

https://dx.doi.org/10.1021/acs.jpca.0c03886
J. Phys. Chem. A 2020, 124, 7155−7165

Table 1. Mean Absolute Error after Global Shift (MARE) for s Small Alkanes System[a]

methods	*n*-butane	isobutane
MOB-ML[18]	0.32	0.33
NeuralXC[20]	0.24	0.22
DeePHF (peak)	0.18	0.21
DeePHF (best)	0.09	0.11

[a]For each model, 50 randomly sampled ethane and 20 propane configurations are used the training set, which was tested on *n*-butane and isobutane. Global shift has been applied for each molecular type so that the mean prediction error is 0. Both the most probable value (peak) and the best result are shown for the proposed model, based on different randomly sampled training sets. Errors are given in mH.

with three other carbons. Such an environment does not occur in neither ethane nor propane. Hence, the model has no clue how much this atom should contribute to the correlation energy. In other words, this is the extrapolation case, so there is usually a systematic error in the predicted energy (the large global shift in isobutane), and the prediction depends largely on how the training data are selected (the wide error distribution). On the other hand, *n*-butane does not contains such an atom. All its atoms have similar ones in ethane and propane, and the prediction is more like interpolation and therefore behaves much better.

This explanation can be partially justified by the observation that the test accuracy was actually *reduced* when more training samples from methane are included. The carbon atom in methane connects to four hydrogen atoms and does not appear in all other molecules. When the amount of training data are very limited and concentrated on a specific type of molecule, adding such data would actually introduce some bias in the model and worsen the results, especially in a linear model.

To see it more clearly, we project the descriptors into two dimension using principal component analysis (PCA) and plot them in Figure 4. Atoms separate clearly into different clusters based on their element type and neighboring atoms. We can see that our training set (ethane and propane) covers the hydrogen cluster and two carbon clusters ((C_1,H_3) and (C_2,H_2)), as expected. The atoms from *n*-butane lie just in these three clusters, hence the interpolation regime of our model, while some atoms from isobutane form another carbon cluster (C_3,H_1) that is outside our training set, corresponding with the extrapolation regime. Moreover, the carbon atoms from methane form yet another cluster that is away from both our training and testing set. This behavior is exactly what we expect from our previous explanation. Therefore, we believe that the apparent "transferability" found above is basically due to data selection and is not an indication of real transferability for larger systems.

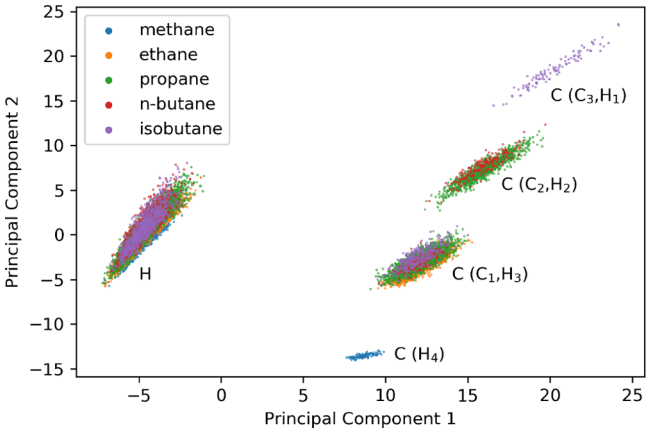


Figure 4. Projected descriptors of small alkane molecules. Each dot on the graph stands for a single atom, with a specific color to identify the molecule it belongs to. Atom type is annotated near the corresponding cluster, including neighboring environments for carbon atoms.

The QM7b-T Data Set. To test the transferability of DeePHF in a more realistic setting, we utilized the QM7b-T data set. Again, we randomly select different numbers of samples to train the model and test its accuracy on the rest of the data set.

In this data set, the system is complex enough that linear regression is no longer adequate. Therefore, we introduce a neural network model as an improved fitting function $\mathcal{F} = \mathcal{F}^{\mathrm{nn}}$. Since the number of training samples is large enough, we were able to test the extended descriptor $\{\tilde{\mathbf{d}}^I_{nl}\}$ described in eq 13. We used two simple kernel functions to calculate the extended density matrix, namely identity function $f_1(\varepsilon) = \varepsilon$ and the exponential function $f_2(\varepsilon) = e^{\varepsilon}$. Combined with the basic descriptors which can be viewed as using constant kernel $f_0(\varepsilon) = 1$, the total number of descriptors is three times larger than the case of basic descriptors.

With basic or extended descriptors and linear or neural network fitting functions, we have a total of four scenarios. We examine the learning curve for all the four scenarios and the results are shown in Figure 5. The linear model is only used when the number of samples is roughly no less than the number of descriptors, and the neural network model is used when the number of samples is three times larger. For training samples less than these levels, we should not expect the model to perform well. Due to the cost of training neural network models, we do not test the effect of different selections of the training data, as we did for the case of water. Except for the curve reported in ref 19, all models are trained on the same set of data. When the data set is augmented, existing samples in the data set are kept.

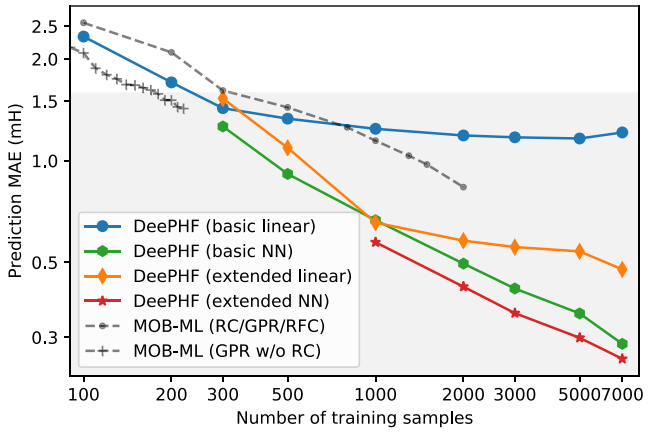


Figure 5. Learning curve of DeePHF on the QM7b-T data set. Depending on whether the extended descriptor and neural network fitting function is used, results for four different constructions are presented. Results from ref 19 are also included for comparison. The gray shaded area corresponds to the region where the error is smaller than chemical accuracy (1 kcal/mol).

https://dx.doi.org/10.1021/acs.jpca.0c03886
J. Phys. Chem. A 2020, 124, 7155−7165

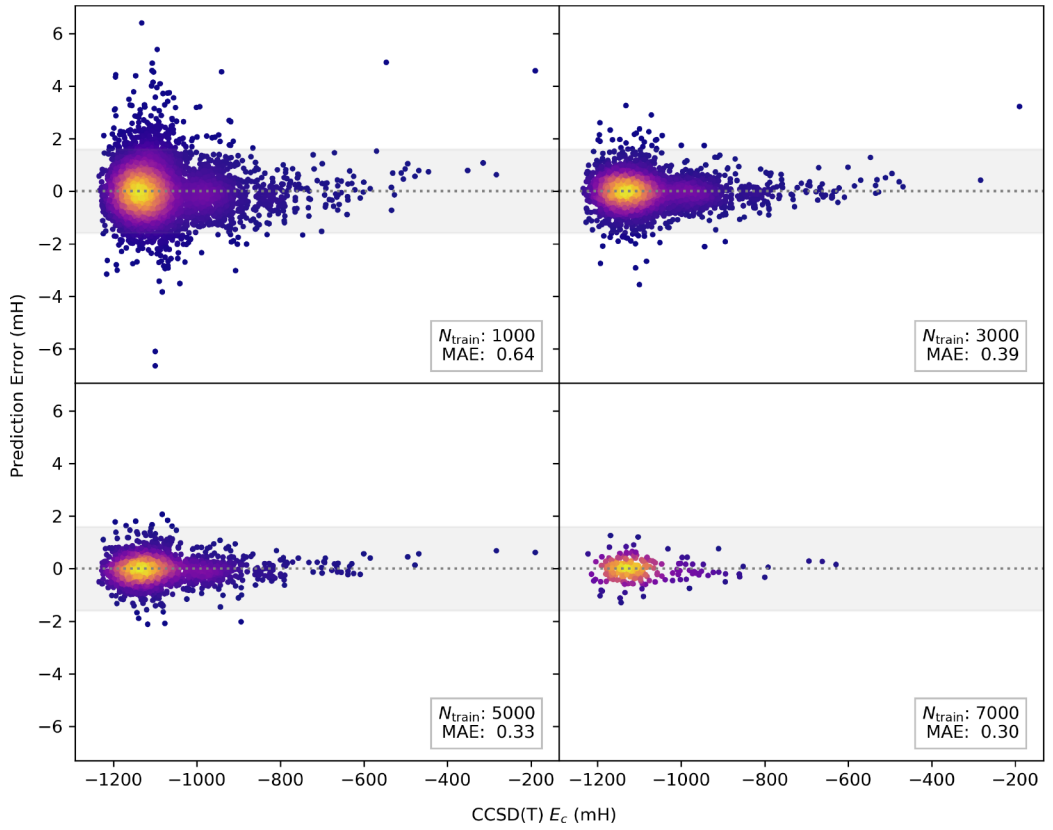


Figure 6. Testing errors of all testing molecules for DeePHF trained on different number of training data. The model is built with extended descriptors and uses neural network as the fitting function. Since we always use the remainder of the data set as the testing set, the number of testing data decreases when we use more training data. The gray shaded area corresponds to the region where the error is smaller than chemical accuracy (1 kcal/mol). Color on the points represents the density of the points.

Not surprisingly, given sufficiently many training samples, the neural net model with extended descriptor performs the best. The performance of the linear model saturates quickly when the number of training samples increases, indicating the need of using more sophisticated models. Neural network models, on the other hand, exhibit the expected power-law behavior,[32] suggesting that further improvement of the accuracy is possible by adding more data. At 7000 training samples, we notice that the model predictions are less consistent, likely due to the small size of the testing set.

We also compare our results with MOB-ML method. Two versions of MOB-ML are included, namely the one with regression clustering, Gaussian process regression and random forest classification (RC/GPR/RFC), and the one without clustering (GPR without RC). The RC/GPR/RFC version reaches the best testing accuracy regardless of number of training samples while the GPR without RC version performs better in terms of sample efficiency. In general, our model is comparable with the MOB-ML method. To be more specific, our model outperforms the RC/GPR/RFC version in all cases as long as we use scenarios that are expressive enough. On the other hand, DeePHF performs slightly worse than GPR without RC version when the size of training set is small. One possible reason is that MOB-ML uses per-orbital pair correlation energies as training labels, which can be dozens of times more than a single correlation energy value. This additional information from the label can be very important when training data are relatively sparse.

To get a more intuitive view, we summarize in Figure 6 all the testing results using the neural network model trained on different numbers of samples. We see that when using only 1000 samples, although the mean absolute energy is 0.64, better than any other model reported, we still get a lot of points with pretty large individual error. When we increase the number of training samples, the testing error decreases as expected and so does the number of points that lie outside the chemical accuracy region. When the size of the training set reaches 5000, the results are very close to chemical accuracy. When the training set increases to 7000 samples, all testing results lie within the chemical accuracy region.

To explore the sample efficiency of our model, we examine the active learning procedure on the QM7b-T data set, using the best performing scenario: neural network fitting function with extended descriptor. At this point, this exercise serves as a proof of concept. We do not calculate new CCSD(T) data. Instead, we iteratively select and add existing data into training set based on the error indicator calculated using model deviation. The detailed steps are as follows: First, we use the same 1000 training samples to train four models independently. Next, we predict the energy using all four models on the remaining data in QM7b-T. We select another 200 molecules with the largest values of model deviation, adding them to the training set and repeat the same procedure. After 20 rounds, we reach a training set of 5000 samples.

The resulting learning curve is shown in Figure 7. A comparison of the testing error between actively learned model and model trained on randomly sampled data is also shown in

https://dx.doi.org/10.1021/acs.jpca.0c03886
J. Phys. Chem. A 2020, 124, 7155−7165

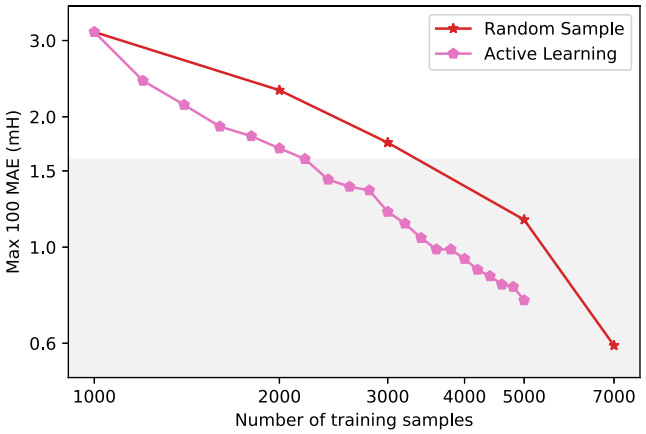


Figure 7. Learning curve of DeePHF with active learning on the QM7b-T data set. Extended descriptor with neural network fitting function is used. The curve for randomly selected samples is same as the one shown in Figure 5 and is used here for comparison. The gray shaded area indicates the region where the error is smaller than chemical accuracy (1 kcal/mol).

Figure 8. For the model with active learning, we used the mean absolute error of the largest 100 points over the entire data set (Max 100 MAE) as our error metric. We see that the error is significantly reduced by active learning on the whole learning curve. The number of training samples needed to reach chemical accuracy is also reduced. Note that in the case of 7000 training samples, the number of data outside the training set is only 211. Hence the Max 100 MAE exhibits a large reduction due to the lack of data. This is not expected to happen if we used a larger data set.

A more concrete illustration can be found in Figure 8. Active learning also largely reduces the maximum absolute error and the number of samples that lie outside the chemical accuracy region. This points to the possibility of using active learning to continuously improve our model by adaptively choosing and labeling new samples on a much larger unlabeled data set. We intend to pursue this in the future, and we invite interested readers to collaborate on this project.

To further examine the transferability of our DeePHF model, we also test the (QM7b-T trained) model on the GDB-13-T data set. We follow a similar approach in refs 18 and 19 by training the model on randomly sampled data from QM7b-T and testing the model accuracy on GDB-13-T. We use the same set of data as that in the study of the learning curve (Figure 5), only that the training and testing labels are MP2 correlation energy instead of CCSD(T), since we do not have CCSD(T) labels for the GDB-13-T data set. We examine two

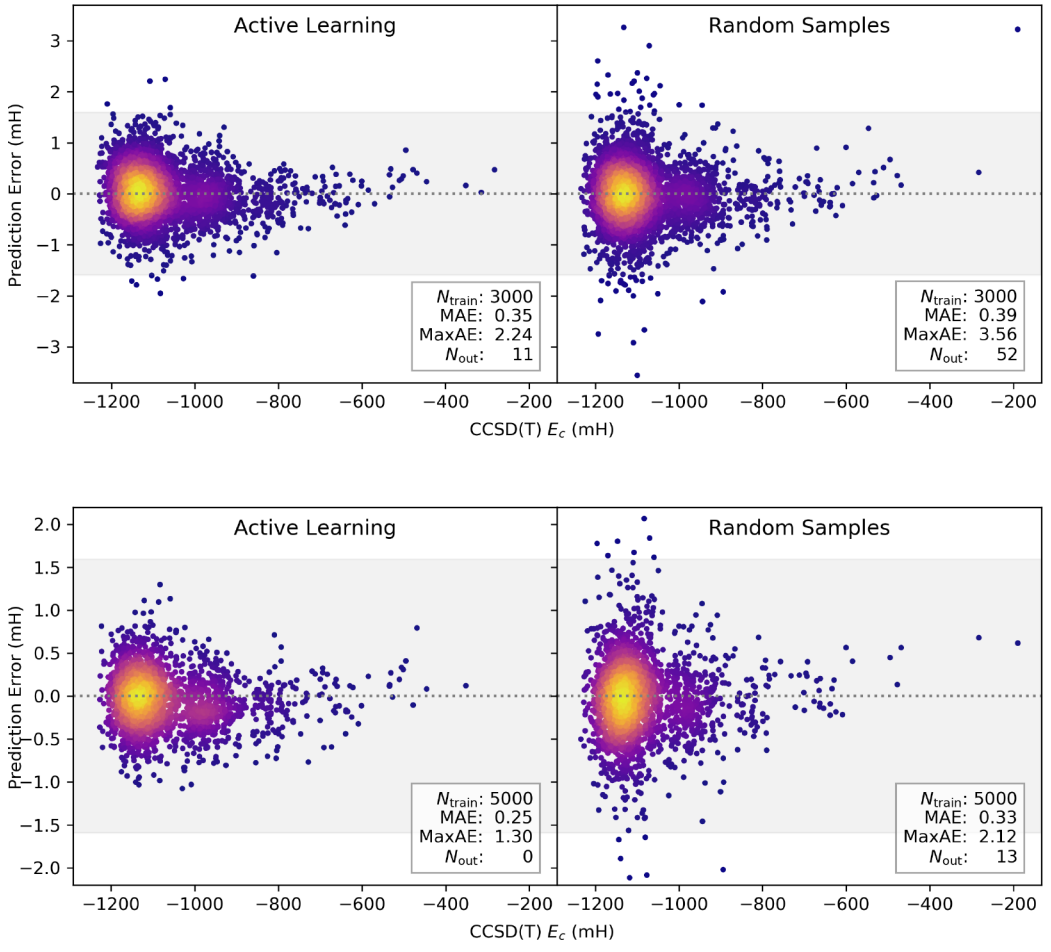


Figure 8. Scatter plots of the error for the testing samples with and without active learning. The size of the training set is 3000 and 5000, respectively, for the upper and lower figure. Error statistics include the mean absolute error (MAE), maximum absolute error (MaxAE) and the number of points with error larger than chemical accuracy (1 kcal/mol) (N_{out}). The gray shaded area indicates the region where the error is smaller than chemical accuracy. Color on points represents the density of the points.

https://dx.doi.org/10.1021/acs.jpca.0c03886
J. Phys. Chem. A 2020, 124, 7155−7165

scenarios, neural network model with basic or extended descriptor. No active learning procedure is used in this test.

The results can be found in Figure 9. In general, it is comparable with that of the MOB-ML (RC/GPR/RFC)

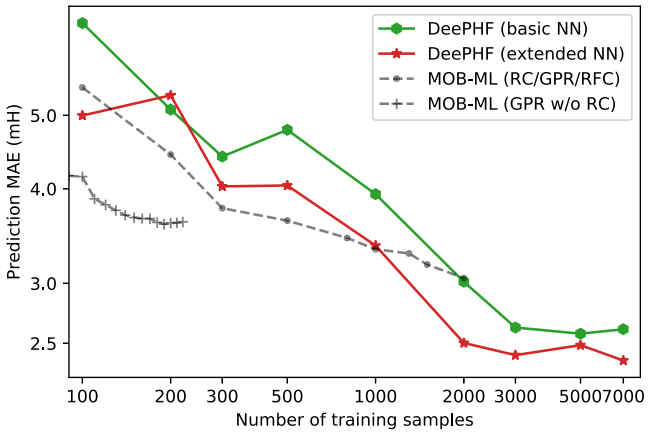


Figure 9. Learning curve of DeePHF by training on the QM7b-T data set and testing on GDB-13-T data set. The extended descriptor and neural network fitting function is used. Results from ref 19 are also included for comparison.

method. With a sufficiently large training set, our model performs relatively better. Using an extended descriptor outperforms MOB-ML (RC/GPR/RFC) in the large training set regime. When the number of the training samples is less than 1000, the error of DeePHF is relatively large, possibly due to the sparsity of the data.

On the other hand, in all cases, the errors are larger than chemical accuracy. The anomalous zigzag behavior in the learning curve (the *decrease* of accuracy when adding more training data) implies that the testing performance is largely decided by the selection of data. Furthermore, the testing error exhibits a large constant shift (∼1.5mH) on the whole GDB-13-T data set. This behavior indicates that the test on the GDB-13-T data set is likely in the extrapolation regime. There should be some information not captured by the training set (QM7b-T). In other words, the data from QM7b-T are insufficient for getting a robust model on the GDB-13-T data set that contains larger molecules.

As we have mentioned earlier, the proposed scheme is not limited to starting with the HF solution and fitting the correlation energy $E_c = E_{\mathrm{CCSD(T)}} - E_{\mathrm{HF}}$. Other self-consistent models such as KS-DFT can also be used as the starting point and similar "DeePHF" models can be trained to predict the modified "correlation" energy $E_c' = E_{\mathrm{CCSD(T)}} - E_{\mathrm{KS}}$. Here we demonstrate an example of using KS-DFT with the PBE functional as the starting point. The resulted learning curve can be found in Figure 10. Although the linear fitting results are not as good, the neural network results are comparable and even slightly better than that of the HF model.

Finally, we would like to point out the eigenvalue construction in DeePHF is rather important especially when we deal with large data sets. As an example, let us examine another commonly used way of imposing rotational symmetry, summing over the angular indices—the trace construction, using $\mathrm{Tr}[(\mathcal{D}^I_{nl})_{mm'}] = \sum_m (\mathcal{D}^I_{nl})_{mm}$ as descriptors. As shown in Figure 11, such a construction performs much worse than the eigenvalue construction using the same density matrix.

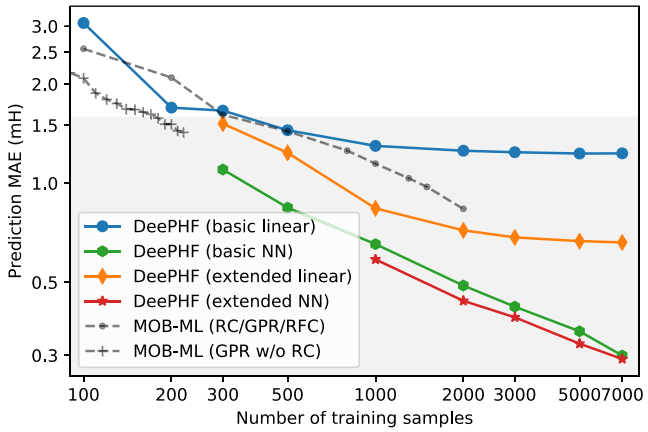


Figure 10. Learning curve of the DeePHF method on QM7b-T data set using DFT with PBE functional as the starting point. Depending on whether the extended descriptors and the neural network model fitting functions are used, results from four different constructions are presented. Results from ref 19 are also included for comparison. The gray shaded area indicates the region where the error is smaller than chemical accuracy (1 kcal/mol).

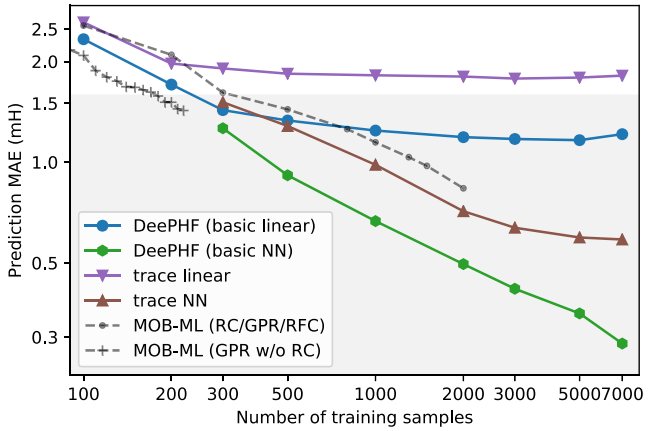


Figure 11. Learning curve of the proposed method on the QM7b-T data set, using trace as descriptor. Results of using basic eigenvalue descriptor as well as results from ref 19 are also included for comparison. The gray shaded area indicates the region where the error is smaller than chemical accuracy (1 kcal/mol).

Moreover, the testing error of the trace construction saturates when the number of training samples increases over 3000, even if we use neural network fitting functions. Such a behavior suggests that the trace construction is reaching its limit of expressive power. By using eigenvalues as descriptors, DeePHF has a more faithful representation of the ground state density matrix and is able to capture the functional dependence of the correlation energies even for large data sets.

The ANI-1ccx Data Set. As mentioned previously, we find training on QM7b-T may be inadequate for us to acquire a model that can transfer well to larger molecules like those in GDB-13-T. Therefore, we take a step further to include the newly published ANI-1ccx data set[29] as our training set. The ANI-1ccx data set includes about 500k configurations of diverse molecule that are consisted by C, H, N and O. These configurations are intelligently selected from a larger set of data, ANI-1x,[33] with around 5 M configurations. The energies of these molecule configurations are calculated by the CCSD(T)*/CBS method,[28] an accurate approximation of

https://dx.doi.org/10.1021/acs.jpca.0c03886
J. Phys. Chem. A 2020, 124, 7155−7165

CCSD(T) extrapolating to the complete basis limit. This data set is originally constructed and used to train a atom-based neural network potential model, also named ANI-1ccx,[28] and hence is designed to be diverse and unduplicated.

Given this diverse nature of the data set, we do not test the transferability within the data set. Instead, we perform the same test used to examine the ANI-1ccx model.[28] Due to the lack of testing data, we only measure the accuracy of hydrocarbon reaction and isomerization energies using HC7/11[34] and ISOL6[35] benchmarks. We examine the neural network model with basic descriptors. Detailed results can be find in Table 2.

Table 2. Mean Absolute Error of Reaction and Isomerization Energies Comparing to the Best Estimation Reference[a]

methods	HC7/11	ISOL6
ωB97X/6-31g*	26.1	6.1
CCSD(T)*/CBS	2.5	0.8
ANI-1ccx	4.0	2.4
DeePHF	1.1	1.4

[a]Results of methods other than DeePHF come from ref 28. Errors are given in mH.

On both benchmarks, DeePHF performs pretty good and the predictions are within chemical accuracy. We note here the ANI-1cxx model is obtained by first training one a even larger data set ANI-1x and adapting to ANI-1ccx data through transfer learning. Hence it in fact utilizes more information than DeePHF do. Also note that the CCSD(T)*/CBS row corresponds to the method that provides labels of DeePHF training. It might looks mysterious in the HC7/11 case that the DeePHF model even outperforms its "ground truth". This may be understood by the fact that the DeePHF prediction happens to lies in the middle between CCSD(T)*/CBS method and the best estimation reference. Since the benchmark only contains six reaction, such behavior can happen by chance.

CONCLUSIONS

We have developed the DeePHF scheme, a machine learning-based method for predicting the ground-state energy of electronic structures with an accuracy comparable to post Hartree−Fock methods like CCSD(T) and a computational cost comparable to that of HF and DFT. In this paper, we focused on predicting the correlation energies using the ground state density matrix calculated by HF or DFT. The models are designed to satisfy several requirements, namely universality, locality, symmetry, accuracy, and efficiency.

The procedure of the DeePHF scheme consists of two parts: The selection of the descriptors and the construction of the fitting function using the descriptors as the input. The descriptors are constructed by projecting the density matrix onto some judiciously chosen local basis. The fitting function is constructed with either a linear model or a neural network model, depending on the required accuracy. An active learning procedure is introduced to minimize the number of data for labeling from a given unlabeled data set.

We examined the performance of DeePHF on organic molecular systems using data sets provided in ref 27. In all the tests, the accuracy of our model is either comparable to or better than previously reported results.[19,20] For simple data sets, we demonstrated that simple linear regression gives satisfactory results already. For the larger data set QM7b-T, DeePHF performs better than ref 19 in terms of accuracy and sample efficiency.

APPENDIX A

Projection Basis Coefficients

The basis coefficients are given in NWChem format that could be found in Basis Set Exchange website.[36] Here we do not include the element name since it does not depend on the element type.

For the 108 basis functions, the coefficients are as follows.

```
SPD
9.8526125336e+02  1.0 -1.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
1.9461950684e+02  0.0  1.0 -1.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
5.7665039062e+01  0.0  0.0  1.0 -1.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
1.7085937500e+01  0.0  0.0  0.0  1.0 -1.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
7.5937500000e+00  0.0  0.0  0.0  0.0  1.0 -1.0  0.0  0.0  0.0  0.0  0.0  0.0
3.3750000000e+00  0.0  0.0  0.0  0.0  0.0  1.0 -1.0  0.0  0.0  0.0  0.0  0.0
2.2500000000e+00  0.0  0.0  0.0  0.0  0.0  0.0  1.0 -1.0  0.0  0.0  0.0  0.0
1.5000000000e+00  0.0  0.0  0.0  0.0  0.0  0.0  0.0  1.0 -1.0  0.0  0.0  0.0
1.0000000000e+00  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  1.0 -1.0  0.0  0.0
6.6666666667e-01  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  1.0 -1.0  0.0
4.4444444444e-01  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  1.0 -1.0
2.9629629630e-01  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  1.0
```

For the 81 basis functions, the coefficients are as follows.

```
SPD
6.400000e+01  1.0 -1.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
3.200000e+01  0.0  1.0 -1.0  0.0  0.0  0.0  0.0  0.0  0.0
1.600000e+01  0.0  0.0  1.0 -1.0  0.0  0.0  0.0  0.0  0.0
8.000000e+00  0.0  0.0  0.0  1.0 -1.0  0.0  0.0  0.0  0.0
4.000000e+00  0.0  0.0  0.0  0.0  1.0 -1.0  0.0  0.0  0.0
2.000000e+00  0.0  0.0  0.0  0.0  0.0  1.0 -1.0  0.0  0.0
1.000000e+00  0.0  0.0  0.0  0.0  0.0  0.0  1.0 -1.0  0.0
5.000000e-01  0.0  0.0  0.0  0.0  0.0  0.0  0.0  1.0 -1.0
2.500000e-01  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  1.0
```

AUTHOR INFORMATION

Corresponding Authors

Han Wang − *Laboratory of Computational Physics, Institute of Applied Physics and Computational Mathematics, Beijing 100088, People's Republic of China*; orcid.org/0000-0001-5623-1148; Email: wang_han@iapcm.ac.cn

Weinan E − *Program in Applied and Computational Mathematics and Department of Mathematics, Princeton, New Jersey 08544, United States*; Email: weinan@math.princeton.edu

Authors

Yixiao Chen − *Program in Applied and Computational Mathematics, Princeton, New Jersey 08544, United States*; orcid.org/0000-0001-8201-5887

Linfeng Zhang − *Program in Applied and Computational Mathematics, Princeton, New Jersey 08544, United States*; orcid.org/0000-0002-8470-5846

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpca.0c03886

https://dx.doi.org/10.1021/acs.jpca.0c03886
J. Phys. Chem. A 2020, 124, 7155−7165

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The work of Y.C., L.Z. and W.E was supported in part by a gift from iFlytek to Princeton University, ONR Grant N00014-13-1-0338, and the Center Chemistry in Solution and at Interfaces (CSI) funded by DOE Award DE-SC0019394. The work of H.W. is supported by the National Science Foundation of China under Grant No. 11871110, the National Key Research and Development Program of China under Grant Nos. 2016YFB0201200 and 2016YFB0201203, and Beijing Academy of Artificial Intelligence (BAAI).

■ REFERENCES

(1) Kohn, W.; Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **1965**, *140*, A1133.
(2) Hartree, D. R.; Hartree, W. Self-consistent field, with exchange, for beryllium. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences* **1935**, *150*, 9−33.
(3) Møller, C.; Plesset, M. S. Note on an approximation treatment for many-electron systems. *Phys. Rev.* **1934**, *46*, 618.
(4) Čížek, J. On the correlation problem in atomic and molecular systems. Calculation of wavefunction components in Ursell-type expansion using quantum-field theoretical methods. *J. Chem. Phys.* **1966**, *45*, 4256−4266.
(5) Pople, J. A.; Head-Gordon, M.; Raghavachari, K. Quadratic configuration interaction. A general technique for determining electron correlation energies. *J. Chem. Phys.* **1987**, *87*, 5968−5975.
(6) Jeziorski, B.; Monkhorst, H. J. Coupled-cluster method for multideterminantal reference states. *Phys. Rev. A: At., Mol., Opt. Phys.* **1981**, *24*, 1668.
(7) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(8) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.
(9) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; VonLilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(10) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Science Advances* **2017**, *3*, No. e1603015.
(11) Schütt, K.; Kindermans, P.-J.; Felix, H. E. S.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems* **2017**, 992−1002.
(12) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science* **2017**, *8*, 3192−3203.
(13) Han, J.; Zhang, L.; Car, R.; E, W. Deep Potential: a general representation of a many-body potential energy surface. *Commun. Comput. Phys.* **2018**, *23*, 629−639.
(14) 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.
(15) Zhang, L.; Han, J.; Wang, H.; Saidi, W.; Car, R.; E, W. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. *Advances in Neural Information Processing Systems* **2018**, 4436−4446.
(16) Lu, D.; Wang, H.; Chen, M.; Liu, J.; Lin, L.; Car, R.; E, W.; Jia, W.; Zhang, L. 86 PFLOPS Deep Potential Molecular Dynamics simulation of 100 million atoms with ab initio accuracy. *arXiv* **2020**, 2004.11658.
(17) 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.
(18) 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.
(19) 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.
(20) Dick, S.; Fernandez-Serra, M. Machine learning accurate exchange and correlation functionals of the electronic density. *Nat. Commun.* **2020**, *11*, 1−10.
(21) Hohenberg, P.; Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **1964**, *136*, B864.
(22) Gilbert, T. L. Hohenberg-Kohn theorem for nonlocal external potentials. *Phys. Rev. B* **1975**, *12*, 2111.
(23) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition* **2016**, 770−778.
(24) Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, 1606.08415.
(25) Kingma, D. P.; Ba, J. A method for stochastic optimization. *arXiv* **2014**, 1412.6980.
(26) Zhang, L.; Lin, D.-Y.; Wang, H.; Car, R.; E, W. Active learning of uniformly accurate interatomic potentials for materials simulation. *Physical Review Materials* **2019**, *3*, 023804.
(27) Cheng, L.; Welborn, M.; Christensen, A. S.; Miller, T. F., III Thermalized (350K) QM7b, GDB-13, water, and short alkane quantum chemistry dataset including MOB-ML features. https://data.caltech.edu/records/1177 (accessed July 7, 2020).
(28) 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*, 1−8.
(29) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **2020**, *7*, 1−10.
(30) Sun, Q.; Berkelbach, T. C.; Blunt, N. S.; Booth, G. H.; Guo, S.; Li, Z.; Liu, J.; McClain, J. D.; Sayfutyarova, E. R.; Sharma, S.; et al. PySCF: the Python-based simulations of chemistry framework. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018**, *8*, No. e1340.
(31) Woon, D. E.; Dunning, T. H., Jr Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. *J. Chem. Phys.* **1993**, *98*, 1358−1371.
(32) Hestness, J.; Narang, S.; Ardalani, N.; Diamos, G.; Jun, H.; Kianinejad, H.; Patwary, M.; Ali, M.; Yang, Y.; Zhou, Y. Deep learning scaling is predictable, empirically. *arXiv* **2017**, 1712.00409.
(33) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733.
(34) Peverati, R.; Zhao, Y.; Truhlar, D. G. Generalized gradient approximation that recovers the second-order density-gradient expansion with optimized across-the-board performance. *J. Phys. Chem. Lett.* **2011**, *2*, 1991−1997.
(35) Luo, S.; Zhao, Y.; Truhlar, D. G. Validation of electronic structure methods for isomerization reactions of large organic molecules. *Phys. Chem. Chem. Phys.* **2011**, *13*, 13683−13689.
(36) Pritchard, B. P.; Altarawy, D.; Didier, B.; Gibson, T. D.; Windus, T. L. New Basis Set Exchange: An Open, Up-to-Date Resource for the Molecular Sciences Community. *J. Chem. Inf. Model.* **2019**, *59*, 4814−4820.

https://dx.doi.org/10.1021/acs.jpca.0c03886
J. Phys. Chem. A 2020, 124, 7155−7165