THE JOURNAL OF PHYSICAL CHEMISTRY A
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCA Article

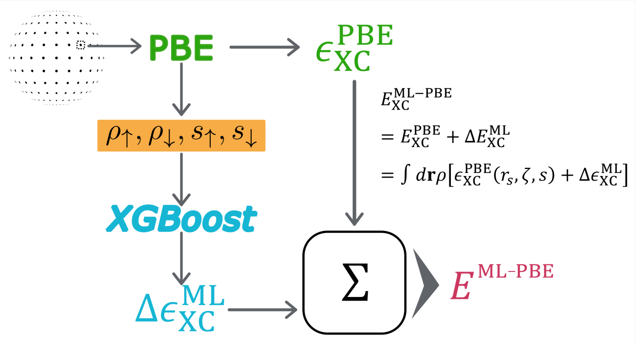
# Improving Density Functional Prediction of Molecular Thermochemical Properties with a Machine-Learning-Corrected Generalized Gradient Approximation

JingChun Wang, DaDi Zhang, Rui-Xue Xu, ChiYung Yam, GuanHua Chen, and Xiao Zheng*

Cite This: J. Phys. Chem. A 2022, 126, 970−978 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** The past decade has seen an increasing interest in designing sophisticated density functional approximations (DFAs) by integrating the power of machine learning (ML) techniques. However, application of the ML-based DFAs is often confined to simple model systems. In this work, we construct an ML correction to the widely used Perdew−Burke−Ernzerhof (PBE) functional by establishing a semilocal mapping from the electron density and reduced gradient to the exchange−correlation energy density. The resulting ML-corrected PBE is immediately applicable to any real molecule and yields significantly improved heats of formation while preserving the accuracy for other thermochemical and kinetic properties. This work highlights the prospect of combining the power of data-driven ML methods with physics-inspired derivations for reaching the heaven of chemical accuracy.


PBE
$\epsilon_{XC}^{PBE}$
$\rho_\uparrow, \rho_\downarrow, s_\uparrow, s_\downarrow$
XGBoost
$\Delta\epsilon_{XC}^{ML}$
$E_{XC}^{ML-PBE}$
$= E_{XC}^{PBE} + \Delta E_{XC}^{ML}$
$= \int d\mathbf{r}\rho[\epsilon_{XC}^{PBE}(r_s, \zeta, s) + \Delta\epsilon_{XC}^{ML}]$
$\Sigma$
$E^{ML-PBE}$

## 1. INTRODUCTION

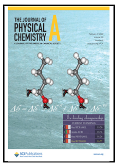
Density functional theory (DFT) has achieved enormous success in physics, chemistry, biology, and materials science as an efficient tool to study electronic structures and reactions in molecules, condensed phases, and extended many-body systems since the 1990s.[1] The success goes to the modern paradigm of DFT which consists of the Hohenberg−Kohn (HK) theorem[2] and the Kohn−Sham (KS) formalism,[3] with the former offering a one-to-one mapping between ground-state electron density and external potential and the later providing a practical approach to finding the ground-state energy by adopting a certain approximation for the exchange−correlation (XC) functional.

A great variety of density functional approximations (DFAs) have been proposed. Diversified strategies have been employed, such as the analytic analysis on simple physical models, the explicit imposition of exact physical constraints or conditions, the adiabatic connection[4] between the reference noninteracting system and real interacting system, and the use of semiempirical or empirical parameters whose values are determined by optimizing the numerical accuracy of density functional calculations.[5−12] Despite the progress made, chemical accuracy has not been achieved universally within the framework of KS-DFT. This is partly because the mathematical representation of the KS mapping, $\rho(\mathbf{r}) \to v_{XC}(\mathbf{r})$, with $\rho(\mathbf{r})$ and $v_{XC}(\mathbf{r})$ being the electron density and XC potential, is not sophisticated enough.

One way to attain more expressive representation of the KS mapping is to introduce more intricate density descriptors. Following this idea, DFAs invoking more complex density related quantities have been constructed, which belong to higher rungs of "Jacob's ladder".[13] It has been demonstrated that the growing complexity of functional form indeed leads to enhanced accuracy.[14] However, practical application of DFT still faces challenges even with the most sophisticated manually designed DFAs.

An alternative approach to enhance the representation of the KS mapping has become increasingly popular, which is to exploit machine learning (ML) techniques.[15−31] As early as in 1996, Tozer et al.[32] have used an artificial neural network (NN) to fit the Zhao−Morrison−Parr XC potential.[33] This pioneering work demonstrated that it is entirely possible to represent the KS mapping by means of ML models. Recently, Zhou et al. have employed a deep learning technique—the convolutional NN—to yield in principle the exact KS mapping.[34] Nagai et al. have constructed an NN-based mapping from a series of density descriptors to the XC energy density $\epsilon_{XC}(\mathbf{r})$, which results in a family of NN-based DFAs.[35]

**Received:** December 12, 2021
**Revised:** January 19, 2022
**Published:** February 3, 2022



https://doi.org/10.1021/acs.jpca.1c10491

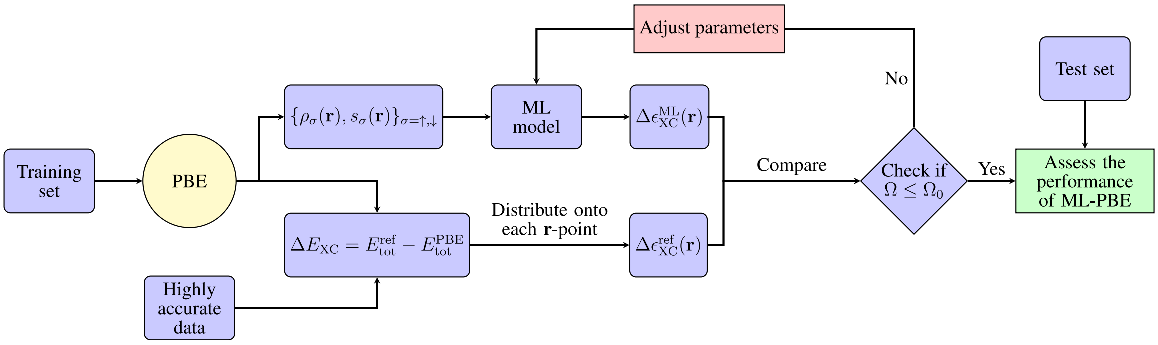

Adjust parameters
No
Test set
$\{\rho_\sigma(\mathbf{r}), s_\sigma(\mathbf{r})\}_{\sigma=\uparrow,\downarrow}$
ML model
$\Delta\epsilon^{\rm ML}_{\rm XC}(\mathbf{r})$
Training set
PBE
Compare
Check if $\Omega \le \Omega_0$
Yes
Assess the performance of ML-PBE
Distribute onto each **r**-point
$\Delta E_{\rm XC} = E^{\rm ref}_{\rm tot} - E^{\rm PBE}_{\rm tot}$
$\Delta\epsilon^{\rm ref}_{\rm XC}(\mathbf{r})$
Highly accurate data

**Figure 1.** Schematic diagram illustrating the workflow for constructing the ML-corrected PBE functional. Details are elaborated in the main text. Here, $\Omega_0$ is a preset threshold for the minimization of loss function $\Omega$.

Aside from the KS mapping, other ML-based mapping schemes have also been proposed to enhance the predictive power of DFT.[36−38] For instance, NN models have been used to optimize the values of semiempirical parameters[39,40] in DFAs such as B3LYP and LC-BLYP. Snyder et al. have adopted an ML approach to construct the mapping $\rho(\mathbf{r}) \to T$, with $T$ being the kinetic energy functional, for one-dimensional (1D) noninteracting models.[41] Brockherde et al. have attempted to bypass the KS equations by directly learning the density-potential and energy-density maps with ML.[42] Dick and Fernandez-Serra[43,44] and Chen et al.[45] have independently constructed an NN-based mapping from the generalized KS single-electron reduced density matrix to an energy correction term.

Despite the above exciting progress, the application of ML-based KS mapping has been confined to simple model systems or specific types of molecules. Moreover, to achieve a satisfactory accuracy, in some ML-based mapping schemes, sophisticated density descriptors corresponding to higher rungs of Jacob's ladder had to be invoked.[35] The goal of this work is to demonstrate that, apart from the design of complex density descriptors, the construction of a highly sophisticated ML-based KS mapping is also important for enhancing the predictive power of DFAs. Specifically, we choose to use simple density descriptors including the electron density function and its first-order derivatives. Thus, the resulting KS mapping is a generalized gradient approximation (GGA) for the XC functional. A GGA-type functional could offer a simple interpretation of the KS mapping[46] and, meanwhile, allows for an efficient treatment of extensive systems such as solids. Furthermore, a semilocal mapping also has the advantages that it is intrinsically universal and transferable, automatically preserves the system's spatial symmetry, and does not require the use of any auxiliary atomic basis functions. At the current stage, the method is potentially useful for systems where hybrid or meta-GGA functionals may be expensive or the correction from the more complicated descriptors is not important.

Regarding the predictive power of the ML-corrected mapping, we shall focus on thermochemical properties for which conventional GGA functionals yield large errors. For such properties, the functional-driven error usually dominates over density-driven error.[47] Therefore, for numerical convenience, in this work an ML model which presents a correction to a parent GGA functional is constructed in a post-self-consistent-field (post-SCF) manner.

The remainder of this paper is organized as follows. Section 2 provides a detailed account of our construct of the ML-corrected GGA functional as well as the data sets adopted to determine and assess the parameters involved in the ML model. In section 3 we demonstrate the numerical performance of the constructed ML-corrected functional and present extensive discussion. Concluding remarks are finally given in section 4.

## 2. METHODOLOGY

We choose the widely used Perdew−Burke−Ernzerhof (PBE) functional[48,49] as the parent DFA. Although the PBE functional yields much improved thermochemical properties over the local density approximation (LDA),[2,50−52] its overall accuracy is still far from satisfactory. This is partly because, while the analytic form of the PBE functional satisfies a number of exact physical constraints, it is not sophisticated enough to account for all the subtle features of electron density as well as their influence on the XC energy.

Instead of constructing a more superior GGA functional completely from scratch, we consider a correction to the PBE energy functional, $E^{\rm PBE}_{\rm XC}$, and an ML model is adopted to represent the energy correction, $\Delta E^{\rm ML}_{\rm XC}$. The ML-corrected PBE functional, abbreviated as ML-PBE hereafter, assumes the following form:

$$\begin{aligned} E^{\rm ML\text{-}PBE}_{\rm XC} &= E^{\rm PBE}_{\rm XC} + \Delta E^{\rm ML}_{\rm XC} \\ &= \int {\rm d}\mathbf{r}\, \rho[\epsilon^{\rm PBE}_{\rm XC}(r_s, \zeta, s) + \Delta\epsilon^{\rm ML}_{\rm XC}] \\ &= \int {\rm d}\mathbf{r}\, \rho\epsilon^{\rm unif}_{\rm X}(\rho)\, F^{\rm PBE}_{\rm XC}(r_s, \zeta, s)\, F^{\rm ML}(\rho, s) \end{aligned} \tag{1}$$

Here, $\epsilon^{\rm unif}_{\rm X}$ is the exchange energy density of a uniform electron gas, and $\epsilon^{\rm PBE}_{\rm XC}$ and $\Delta\epsilon^{\rm ML}_{\rm XC}$ are the XC energy densities corresponding to the PBE functional and the ML correction, respectively. In practice we could take the best possible semilocal DFA as the reference for $\Delta\epsilon^{\rm ML}_{\rm XC}$. Regarding the involving density-related quantities, $\zeta = (\rho_\uparrow - \rho_\downarrow)/\rho$ is the relative spin polarization and $s = |\nabla\rho|/(2k_{\rm F}\rho)$ is the reduced density gradient, with $k_{\rm F} = (3\pi^2\rho)^{1/3}$ and $r_s = (4\pi\rho/3)^{-1/3}$ being the Fermi wavevector and Wigner−Seitz radius, respectively.[48] Upon the last equality of eq 1, the ML-PBE functional is recast into a compact form by referring to the LDA exchange and defining two local enhancement factors, $F^{\rm PBE}_{\rm XC}$ and $F^{\rm ML}$, where $F^{\rm PBE}_{\rm XC}$ enhances the LDA exchange to the PBE functional and $F^{\rm ML}$ enhances the PBE to the ML-corrected PBE.

The challenge then is to construct a universal and accurate semilocal mapping, $\{\rho_\uparrow(\mathbf{r}), \rho_\downarrow(\mathbf{r}), s_\uparrow(\mathbf{r}), s_\downarrow(\mathbf{r})\} \to \Delta\epsilon^{\rm ML}_{\rm XC}(\mathbf{r})$, with $s_\sigma = |\nabla\rho_\sigma|/[2(3\pi^2)^{1/3}\rho_\sigma^{4/3}]$ $(\sigma = \uparrow, \downarrow)$ being the spin-specific reduced density gradient. With the inclusion of an ML correction, the exact physical constraints satisfied by the parent DFA are likely to be compromised. This presents another important challenge for the development of ML-corrected DFAs.

Figure 1 illustrates the workflow for establishing the ML-based KS mapping. To construct a semilocal mapping, an individual point in the **r**-space of any chemical species is taken as a sample. Each sample is associated with a data entry, with the density descriptors $\{\rho_\sigma(\mathbf{r}), s_\sigma(\mathbf{r})\}$ representing the features of the sample and $\Delta\epsilon_{\mathrm{XC}}^{\mathrm{ML}}(\mathbf{r})$ as the label. The key is to acquire a sufficient amount of pointwise data to enable supervised learning of the ML model. To this end, highly accurate reference data $\Delta\epsilon_{\mathrm{XC}}^{\mathrm{ref}}(\mathbf{r})$ are needed to guide the learning process, which aims to minimize the loss function

$$\Omega = \frac{1}{M}\sum_{m=1}^{M}\int \mathrm{d}\mathbf{r}\,|\Delta\epsilon_{\mathrm{XC},m}^{\mathrm{ML}}(\mathbf{r}) - \Delta\epsilon_{\mathrm{XC},m}^{\mathrm{ref}}(\mathbf{r})| \tag{2}$$

where $M$ is the number of species in the training data set.

Usually, it is the energy difference between two species that can be obtained accurately from experiment or from high-level quantum chemistry calculation. However, to enable the training of a semilocal ML model, the reference value for the energy of each species, $E_m^{\mathrm{ref}}$, is required. Moreover, the discrepancy between the energy calculated by the PBE functional and the corresponding reference value, $\Delta E_{\mathrm{XC},m} = E_{\mathrm{tot},m}^{\mathrm{ref}} - E_{\mathrm{tot},m}^{\mathrm{PBE}}$, is to be decomposed into pointwise errors, which are then assigned to each individual **r**-point. In this work, we follow an intuitive notion that the XC energy density of a GGA functional on every **r**-point is subject to roughly the same amount of relative error. Thus, the decomposition of error is done by presuming that the correction to energy density is proportional to the XC energy density itself, i.e.

$$\Delta\epsilon_{\mathrm{XC},m}^{\mathrm{ref}}(\mathbf{r}) = \frac{\epsilon_{\mathrm{XC},m}^{\mathrm{DFA}}(\mathbf{r})}{E_{\mathrm{XC},m}^{\mathrm{DFA}}}\Delta E_{\mathrm{XC},m} \tag{3}$$

Apparently, if the training set consisted only of a single species, the ratio $\Delta\epsilon_{\mathrm{XC},m}^{\mathrm{ref}}(\mathbf{r})/\epsilon_{\mathrm{XC},m}^{\mathrm{DFA}}(\mathbf{r})$, which serves as the target for the enhancement factor $F^{\mathrm{ML}}$, would be a constant in the **r**-space. This contradicts the presumption that $F^{\mathrm{ML}}$ is **r**-dependent. To avoid such a conceptual problem, the training set should include an abundant number of species so that the **r**-dependence of $F^{\mathrm{ML}}$ is adequately manifested.

In practice, the **r**-space integral of a function $f(\mathbf{r})$ is often evaluated via a summation over discretized grids, i.e., $\int \mathrm{d}\mathbf{r}\, f(\mathbf{r}) \to \sum_i f(\mathbf{r}_i)\, W(\mathbf{r}_i)$, with $W(\mathbf{r}_i)$ being the weight of the $\mathbf{r}_i$ grid point. Here, $\{\mathbf{r}_i\}$ are just the grid points adopted to evaluate the one-electron integrals for constructing the KS Hamiltonian. Equation 3 is thus rewritten as

$$\Delta\epsilon_{\mathrm{XC},m}^{\mathrm{ref}}(\mathbf{r}_i) = \frac{\epsilon_{\mathrm{XC},m}^{\mathrm{DFA}}(\mathbf{r}_i)}{\sum_j \rho_m(\mathbf{r}_j)\,\epsilon_{\mathrm{XC},m}^{\mathrm{DFA}}(\mathbf{r}_j)\,W(\mathbf{r}_j)}\Delta E_{\mathrm{XC},m} \tag{4}$$

In principle, the DFA in eqs 3 and 4 can be the ML-PBE functional so as to form a self-closed training process, while in practice the DFA is chosen as the original PBE, which is found to yield a more accurate ML model; see the Supporting Information for more details.

At each $\mathbf{r}_i$ grid point, $\Delta\epsilon_{\mathrm{XC}}^{\mathrm{ML}}(\mathbf{r}_i)$ is obtained by feeding the descriptors $\{\rho_\sigma(\mathbf{r}_i), s_\sigma(\mathbf{r}_i)\}$ into the chosen ML model

$$\Delta\epsilon_{\mathrm{XC}}^{\mathrm{ML}}(\mathbf{r}_i) = G^{\mathrm{ML}}[\rho_\uparrow(\mathbf{r}_i), \rho_\downarrow(\mathbf{r}_i), s_\uparrow(\mathbf{r}_i), s_\downarrow(\mathbf{r}_i)] \tag{5}$$

where $G^{\mathrm{ML}}(\boldsymbol{x})$ denotes the explicit function form associated with the ML model. For each species in the training and test sets, calculation with the original PBE functional is performed to generate a set of descriptors $\{\rho_\sigma(\mathbf{r}_i), s_\sigma(\mathbf{r}_i)\}$ and the corresponding PBE energy density $\epsilon_{\mathrm{XC}}^{\mathrm{PBE}}(\mathbf{r}_i)$. $\Delta\epsilon_{\mathrm{XC}}^{\mathrm{ML}}(\mathbf{r}_i)$ are subsequently computed by eq 5, and the corrected total energy of the species is obtained by eq 1.

The data set used in this work consists of two parts: a training set for optimizing the ML model and a test set for evaluating the performance of the ML-corrected functional. The training set contains 166 energetic data, including 148 standard heats of formation (HOF) taken from the G2/97 set[53] (denoted as the G2-HOF set hereafter) and 18 total energies of neutral atoms of the first three periods (from H to Ar). Note that the HOF (or the atomization energy) of a molecular species involves the calculation of the molecular energy as well as the energies of all the constituent atoms. For numerical convenience, the errors associated with the HOF data are assigned completely to the molecular species, i.e., $\Delta E_{\mathrm{XC},m}$ in eqs 3 and 4, while the PBE energies of neutral atoms are taken as constant parameters in the training stage. This is consistent with the presently adopted error assignment scheme where the errors of HOF are attributed only to molecular species and not to the constituent atoms.

The test set contains 75 HOF from the G3-3 subset of the G3/99 set[54] (denoted as the G3-HOF set hereafter), and 88 ionization potentials (IPs) and 58 electron affinities (EAs) from the G2/97 set[55] (denoted as the G2-IP and G2-EA sets), as well as 42 bond dissociation enthalpies (BDEs) of small organic molecules taken from a handbook of BDE experimental data[56] as well as high precision calculations[57] (denoted as the BDE42 set). The G3-HOF set consists mainly of organic molecules containing 2−10 carbon atoms as well as covalent compounds made of elements from the first three rows of the periodic table.

We also assess the predictive power of ML-PBE beyond thermochemical properties by examining energies for reaction kinetics. Specifically, the test set also includes 38 barrier heights for hydrogen-transfer reactions from the HTBH38/08 set[58−60] and 38 barrier heights for non-hydrogen-transfer reactions from the NHTBH38/08 set[58−60] (denoted as the HTBH38 and NHTBH38 sets). There are around 2.4 million **r**-points in the training set and over 8 million in the test set.

In the evaluation stage, the total energies of the species in the test set (including atoms, molecules, ions, and transition state complexes) are calculated with the ML-PBE functional. The only exception is the total atomic energies for the assessment of the G3-HOF set. To be consistent with the training process, the PBE energies of neutral atoms are taken as constant parameters for the computation of HOF.

The density functional calculations are performed by the Gaussian 16 suite of programs.[61] For each species, the atom-centered **r**-points[62,63] are generated by an in-house built quantum chemistry software package, QM$^4$D,[64] with which the density descriptors $\{\rho_\sigma(\mathbf{r}), s_\sigma(\mathbf{r})\}$ on these **r**-points are then evaluated.

The Gaussian basis set 6-311++G(3df,3pd) is employed throughout this work unless specified. Computational details are provided in the Supporting Information.

A number of ML models were tested for constructing the semilocal KS mapping of our interest. These include some well-established deep learning techniques such as the convolutional NN. Among all the candidates, the XGBoost,[73] which implements the gradient boosting decision tree (GBDT) algorithm,[74,75] excels in the overall performance and is finally adopted to construct the ML mapping. A generic function $G^{\mathrm{XGB}}(\boldsymbol{x})$ learned by the XGBoost has the form of

$$G^{\mathrm{XGB}}(\mathbf{x}) = \sum_{t=1}^{T} g_t(\mathbf{x}) \tag{6}$$

where the basis function $g_t(\boldsymbol{x})$ is a simple decision tree. The model is trained additively; i.e., at each epoch a new basis function $g_t(\boldsymbol{x})$ is added to the existing model, followed by optimization of the involving parameters.

Owing to its prevailing advantages in numerical efficiency and robustness, XGBoost has gained increasing interest from practitioners in various fields of chemistry and materials science, and it is particularly favorable for our task. In contrast, a sophisticated NN model with a large number of hidden neurons, according to our practical experience, is not compatible with a feature vector of only four dimensions in our problem.

Ideally, the explicit form of $\Delta\epsilon_{\mathrm{XC}}^{\mathrm{ML}}$ should allow for SCF calculations with the ML-PBE functional. However, XGBoost is intrinsically nondifferentiable, while a common differentiable NN model yields much inferior training performance than XGBoost (the relative error is at least 50% larger than that of XGBoost). Therefore, at the present stage, the ML correction is implemented in a post-SCF manner.

## 3. RESULTS AND DISCUSSION

Before assessing the performance of the ML-PBE functional, we first explore how the size of the training set affects the training performance. If only a single species in the G2-HOF set is used for training, a mean absolute error (MAE) of 10.8 kcal/mol is yielded for the rest of the G2-HOF set, while if 10 randomly chosen species are taken as the training set, the cross-validation MAE (with the rest of the G2-HOF set) drops to 9.0 kcal/mol. Further enlarging the training set to 100 randomly chosen species results in a smaller cross-validation MAE of 6.9 kcal/mol. Such a trend is consistent with the understanding on the error decomposition scheme of eq 3 that an abundant number of species is required to adequately express the **r**-dependence of $F^{\mathrm{ML}}$.

The MAEs for the various data sets yielded by ML-PBE are exhibited in Figure 2. For the purpose of comparison, the MAEs of the same data sets yielded by the original PBE functional, the strongly constrained and appropriately normed (SCAN) functional[76] which is a meta-GGA on the third rung of Jacob's ladder, and the Becke-3-Lee−Yang−Parr (B3LYP)[77−79] hybrid functional which belongs to the fourth rung of Jacob's ladder are also displayed.

As shown clearly in Figure 2, the original PBE yields appreciable errors for the HOF, with MAEs of 17.1 and 33.7 kcal/mol for the G2-HOF and G3-HOF sets, respectively. By invoking the ML correction to XC energy density, $\Delta\epsilon_{\mathrm{XC}}^{\mathrm{ML}}(\mathbf{r})$, the resulting ML-PBE achieves a much improved accuracy for the calculation of HOF. Specifically, the MAEs are substantially reduced to 5.7 and 8.8 kcal/mol for the G2-HOF and G3-HOF sets, respectively. Such errors are already comparable to those yielded by the meta-GGA SCAN. With both PBE and ML-PBE, the MAE for the G3-HOF set is about twice that for the G2-HOF set, because of the larger sizes of molecules in the former data set. Such error accumulation over the **r**-space will be scrutinized below.

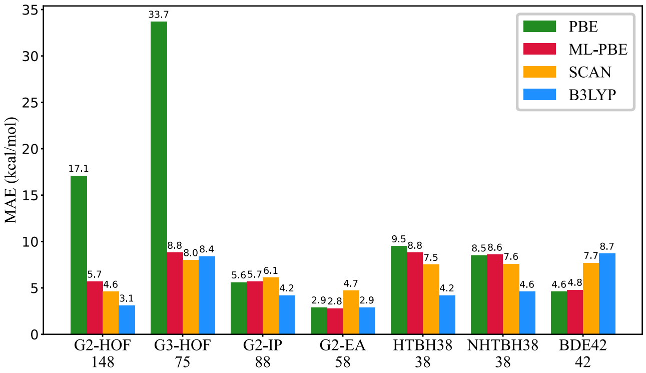


**Figure 2.** Performance of ML-PBE in comparison with the PBE, SCAN, and B3LYP functionals. The number over each bar is the MAE of the corresponding data set yielded by a particular DFA; see the main text for the details about the data sets. The MAEs of the G2-IP and G2-EA sets by SCAN are evaluated via calculations done with the 6-311++G(3df,3pd) basis set, while the other MAEs associated with SCAN are taken from refs 65 and 66. The MAEs of the HTBH38/08 and NHTBH38/08 sets by B3LYP are evaluated via calculations done with the MG3S basis set,[67] while the other MAEs associated with B3LYP are extracted from refs 59 and 68. The calculation of BDEs in the BDE42 set by PBE, SCAN, and B3LYP functionals are based on geometries optimized at the level of B3LYP/6-31G(d) extracted from computational chemistry databases,[69−72] and the corresponding reference values for comparison are adopted from refs 56 and 57.

The ML correction does not improve the prediction of IP and EA, as the MAE yielded by ML-PBE is almost the same as that by the original PBE; see Figure 2. There are two possible reasons. First, the original PBE yields much larger errors for HOF than for IP and EA, and thus the ML model will focus mainly on the former errors to achieve an overall balanced performance across various energetic properties. Second, the training data set only involves neutral species, and thus the ML model might not gain enough knowledge about the charged species. We have attempted to add several ionic species to the training set, but there was not much change in the MAEs. This confirms that the training of the ML model is predominantly driven by the errors of HOF. Nevertheless, both PBE and ML-PBE outperform SCAN in the prediction of IP and EA.

We also extend the assessment of ML-PBE to kinetic properties. As demonstrated in Figure 2, the ML correction leads to a marginal improvement in the prediction of reaction energy barriers. Just like the situation of IP and EA, since the training data set consists mainly of molecular species at their equilibrium geometries, the ML model does not learn much about the transition state species. Moreover, the discrepancy of calculated reaction barriers from the reference values is closely related to the delocalization error of a semilocal DFA, which has been analyzed and understood from the perspective of fractional charges.[80−84] To correct the delocalization error, exact physical constraints such as the Perdew−Parr−Levy−Balduz condition[85] needs to be imposed explicitly,[82,84,86,87] which is of critical significance and to be investigated further. The ML-PBE produces slightly larger MAEs than SCAN for the HTBH38 and NHTBH38 sets.

For the prediction of BDEs, we examine the BDE42 set which is composed of 42 BDEs of organic molecules randomly chosen from refs 56 and 57. The BDE42 set covers three kinds of common chemical bonds, C−C, C−O, and C−N, with half of the data associated with branched C−C bonds. As is demonstrated in Figure 2, the MAEs of ML-PBE and PBE are very similar to each other and are apparently smaller than those of SCAN and B3LYP. The PBE functional gives rise to relatively large errors for the BDEs of branched C−C bonds. Specifically, for the 21 data on branched C−C bonds, the MAEs of the PBE, ML-PBE, SCAN, and B3LYP functionals are 6.6, 6.7, 9.4, and 10.7 kcal/mol, respectively, which are distinctly larger than the overall MAEs of the BDE42 set. Nevertheless, for both the original and ML-corrected PBE functionals, the errors on BDEs (even for branched C−C bonds) are still below the average errors of the entire test set (12.5 and 6.5 kcal/mol for PBE and ML-PBE, respectively), which covers various types of energetic properties.

While the MAE of B3LYP is much larger than that of PBE on the BDE42 set, B3LYP yields much more accurate results than PBE on the G3-HOF set. Such behavior agrees with the previous examination by Xu et al.,[88] who have found that the accuracy of HOF has little correlation with the accuracy of the predicted BDE by the same DFA.

From the above analysis, it is reasonable to conclude that the ML-PBE generally outperforms the original PBE in the prediction of thermochemical and kinetic properties of molecules, and it achieves almost the same level of accuracy as the meta-GGA SCAN. This justifies our proposed strategy of constructing a sophisticated KS mapping by using relatively simple density descriptors. On the other hand, the ML-PBE underperforms the hybrid B3LYP functional for all the energetic properties examined in Figure 2. This accentuates the limitation of a semilocal functional form, as well as the necessity of building nonlocal ingredients into the ML correction to XC energy for further improving its predictive power.

The ML-PBE is designed to improve over PBE in the prediction of all kinds of thermochemical properties in a balanced manner. This means that it is expected to yield roughly the same magnitude of error for different energetic properties. Therefore, the ML model is trained with only the HOF data, since the original PBE performs most unsatisfactorily on HOF among all kinds of energies. Moreover, it is remarkable that the ML also yields a much more balanced error distribution for HOF in the sense that the mean signed error reduces from 16.3 to 0.3 kcal/mol on the G2-HOF training set and from 33.7 to 1.7 kcal/mol on the G3-HOF test set.

Throughout this work, for the computation of HOF, the total energies of the atoms are taken as the PBE values. Such a treatment differs from the common tradition of quantum chemistry but is nevertheless practical and reasonable, as it is consistent with the presently adopted error assignment scheme: the error of HOF is attributed only to molecular species and not to the constituent atoms.

We proceed to examine whether the distribution of pointwise errors in the XC energy density is improved by the ML correction. Parts a and b of Figure 3 depict histograms of pointwise errors, $\Delta\tilde{\epsilon}_{XC}^{PBE}(\mathbf{r}) = \epsilon_{XC}^{ref}(\mathbf{r}) - \epsilon_{XC}^{PBE}(\mathbf{r})$ and $\Delta\tilde{\epsilon}_{XC}^{ML\text{-}PBE}(\mathbf{r}) = \epsilon_{XC}^{ref}(\mathbf{r}) - \epsilon_{XC}^{ML\text{-}PBE}(\mathbf{r})$, assessed for the G2-HOF training set and the G3-HOF test set, respectively. Clearly, the original PBE gives rise to a rather biased error distribution with a long tail of positive values, while with the ML correction the pointwise errors become more centralized and balanced. It is noticed that the calculation result on the G3-HOF set by PBE is so biased that the deviations from the reference values are all positive. Following the decomposition scheme of eq 3, $\Delta\epsilon_{XC}^{ref}(\mathbf{r}_i)$ values are positive for all the species in the G3-HOF set. Nevertheless, ML-PBE counterbalances the biased positive distribution of errors by a relatively larger amount of negative deviations, leading to a drastic drop of the MAE. The improvement is overall consistent for both the training and test sets, though the error distribution is somewhat broader for the latter. We have set the correction to the XC energy density as a linear function of electron density and reduced gradient, and we train the model using a function of the difference in HOF as the cost function. By performing a linear regression for the pointwise errors, the MAE of the resulting corrected PBE is 16.8 kcal/mol, only slightly improved over the original PBE with an MAE of 17.1 kcal/mol. Therefore, a sophisticated ML model is indeed necessary to provide a nontrivial correction to the PBE functional.

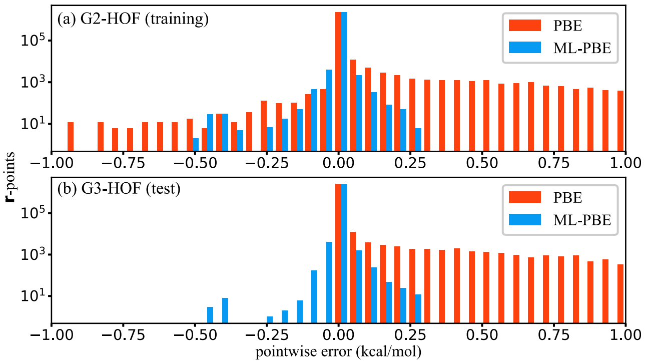


**Figure 3.** Histograms of pointwise errors in XC energy density for PBE and ML-PBE, $\Delta\tilde{\epsilon}_{XC}^{PBE}(\mathbf{r})$ and $\Delta\tilde{\epsilon}_{XC}^{ML\text{-}PBE}(\mathbf{r})$, assessed for the (a) G2-HOF and (b) G3-HOF sets. The horizontal position of a bar represents the magnitude of the pointwise error in units of kilocalories per mole, while the height of a bar stands for the number of **r**-points possessing that particular amount of error. The vertical axis is in a logarithmic scale.

As mentioned earlier, a semilocal KS mapping inevitably leads to the accumulation of pointwise errors; i.e., the error in the total XC energy $\Delta E_{XC}$ increases linearly with the size of the system. As shown in Figure 4, with the original PBE, $\Delta E_{XC}$ of a molecular species grows rapidly with the number of constituent atoms $N_A$, and a linear regression indicates a

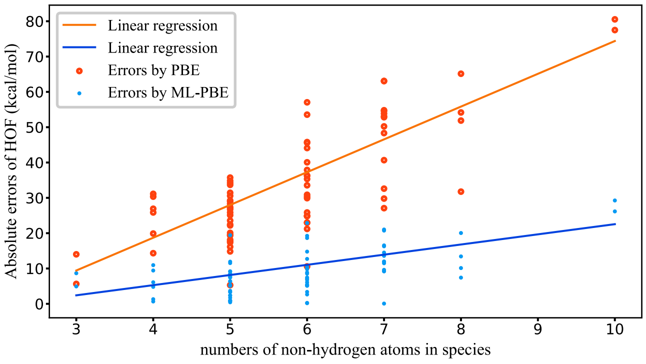


**Figure 4.** Error in XC energy ($\Delta E_{XC}$) versus number of non-hydrogen atoms ($N_A$) for molecular species in the G3-HOF set. The lines are linear regressions of the errors, with the slopes being 9.3 and 2.9 kcal/mol per atom for PBE and ML-PBE, respectively.

slope of 9.3 kcal/mol per atom. Although the linear increase of $\Delta E_{XC}$ still persists after adoption of the ML correction, the rate of growth is greatly suppressed, as verified by a significantly reduced slope of 2.9 kcal/mol per atom. Such a reduction in the slope is consistent with the drop of the MAE associated with the G3-HOF set; cf. Figure 2.

An intriguing question is, how large is the ML correction as compared to the XC energy density of PBE? To answer this question, we examine the magnitude of the enhancement factor corresponding to the ML correction, $F^{\mathrm{ML}} = 1 + \Delta\epsilon^{\mathrm{ML}}_{\mathrm{XC}}/\epsilon^{\mathrm{PBE}}_{\mathrm{XC}}$. Figure 5 exhibits the distribution of $F^{\mathrm{ML}}$ for all the **r**-points associated with the species in the training set. It is found that a predominant majority of $F^{\mathrm{ML}}$ falls within a tiny interval between 0.997 and 1.0, indicating that the ML correction amounts to only about 0.3% of the magnitude of $\epsilon^{\mathrm{PBE}}_{\mathrm{XC}}$. $F^{\mathrm{ML}}$ is also generally smaller than $F^{\mathrm{PBE}}_{\mathrm{XC}}$, with the ratio $|F^{\mathrm{ML}}/F^{\mathrm{PBE}}_{\mathrm{XC}}|$ lying mainly within the range $[0.28, 0.95]$. Therefore, the ML correction scheme proposed in this work is in fact rather conservative, and the physical considerations behind the original design of the PBE functional are largely preserved in the ML-PBE. This is in clear contrast to some other ML-based DFAs which deviate significantly from the traditional DFAs.[35,45] Nagai and co-workers have also constructed ML-based GGA and meta-GGA functionals.[35] Different from our proposed scheme of finding a correction to an existing DFA, they attempted to establish NN-based DFAs completely from scratch, and thus the enhancement factor of a resulting NN-DFA deviates appreciably from those of the existing DFAs in certain regions of density descriptors; cf. Figure 5 in ref 35. In contrast, we always have $F^{\mathrm{ML\text{-}PBE}}_{\mathrm{XC}} \simeq F^{\mathrm{PBE}}_{\mathrm{XC}}$ because of the small magnitudes of the ML correction.

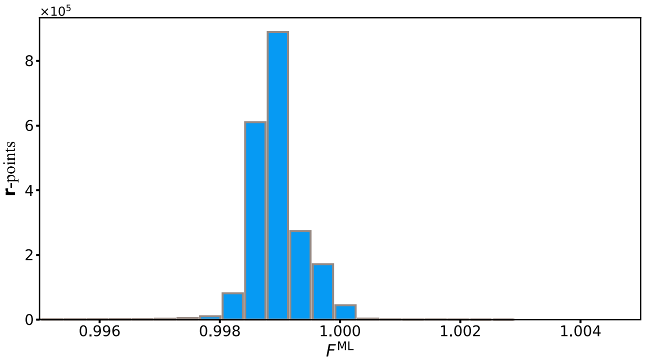


**Figure 5.** Histogram showing the distribution of the pointwise enhancement factor $F^{\mathrm{ML}}$ for all species in the training set. The height of a bar represents the number of **r**-points assuming a particular value of $F^{\mathrm{ML}}$.

In terms of numerical performance, the NN-GGA obtained by Nagai and co-workers yields an MAE of 11.0 kcal/mol for the G2-HOF set and an MAE of 9.6 kcal/mol for the HTBH38 and NHTBH38 sets combined,[35] while the corresponding MAEs are 5.7 and 8.7 kcal/mol respectively by our proposed ML-PBE. In fact, the ML-PBE achieves an MAE of only 8.8 kcal/mol for the G3-HOF set, which consists of molecules considerably larger than those in the G2-HOF set. Therefore, the ML-PBE is expected to outperform the NN-GGA obtained in ref 35 for the prediction of thermochemical energies. Of course, it is worth pointing out that much improved accuracy has been achieved by Nagai and co-workers, with the use of more sophisticated density descriptors.[35]

We have carefully studied the behavior of ML-PBE under the uniform gas limit as well as some other physical constraints with details included in the Supporting Information. The present ML-PBE does not retrieve the exact uniform density limit due to the lack of samples representing that specific limit in the training set. This important aspect calls for more careful exploration[89] in the future.

## 4. CONCLUDING REMARKS

In this work we construct an ML correction to the PBE functional by establishing a semilocal mapping $\{\rho_\sigma(\mathbf{r}), s_\sigma(\mathbf{r})\} \to \Delta\epsilon^{\mathrm{ML}}_{\mathrm{XC}}(\mathbf{r})$. The resulting GGA functional, ML-PBE, yields substantially improved HOF of molecular species over the original PBE, and its overall performance on the prediction of molecular thermochemical and kinetic properties is comparable to that of the widely used meta-GGA SCAN. This indicates that an accurate DFA with a balanced performance can be achieved by combining the power of physically inspired derivation and data-driven ML techniques. Further improvement of the scheme in eq 1 could be realized with the use of more sophisticated density descriptors and more advanced ML models. It is desirable to have an ML-corrected semilocal DFA outperforming the existing meta-GGA functionals. The encouraging performance of ML-PBE confirms that integrating the power of data-driven machine learning with physics-inspired derivation is a promising approach toward the heaven of chemical accuracy.

In principle, the ML-PBE functional constructed in this work is immediately applicable to general electronic systems including molecules, clusters, and bulk solids. However, limited by the computational resources and codes at our disposal, numerical calculations were restricted to simple molecules. Further optimization and more extensive assessment of the ML-corrected DFA are certainly desired.

There is still much room for improvement regarding the design of the ML-based KS mapping. First, the construct of a semilocal mapping requires the error in the total energy of a molecule to be decomposed into pointwise contributions. The present scheme relies on the presumption of eq 3. A more rational way is perhaps to take the XC energy density of a highly accurate hyper-GGA as $\epsilon^{\mathrm{ref}}_{\mathrm{XC}}(\mathbf{r})$. Second, the ML model obtained by this work merely provides a post-SCF correction to the energy, while the electron density is left uncorrected. In principle, the XC potential $v_{\mathrm{XC}}(\mathbf{r})$ can be evaluated by exploiting the analytic gradients of the ML model, which should enable a self-consistent correction to both the energy and the electron density. This is to be pursued in our future work with some differentiable ML models. It is entirely possible to realize a self-consistent implementation of the ML-based correction by utilizing a certain differentiable ML model. Further exploration along this direction is underway.

Finally, to preserve the computational efficiency of a GGA functional, the present work focuses only on a semilocal KS mapping. It has been demonstrated that nonlocal density descriptors may be crucial for addressing long-range electron correlation effects such as the dispersive interaction.[34,35] This thus raises another challenge: incorporating the nonlocal dependence effectively and efficiently into an ML model. Work along this direction is underway.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpca.1c10491.

Details on data sets, numerical performance, and enhancement factor associated with the ML-PBE functional (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Author

**Xiao Zheng** − *Hefei National Laboratory for Physical Sciences at the Microscale & Synergetic Innovation Center of Quantum Information and Quantum Physics & CAS Center for Excellence in Nanoscience, University of Science and Technology of China, Hefei, Anhui 230026, China;* orcid.org/0000-0002-9804-1833; Email: xz58@ustc.edu.cn

### Authors

**JingChun Wang** − *Hefei National Laboratory for Physical Sciences at the Microscale & Synergetic Innovation Center of Quantum Information and Quantum Physics & CAS Center for Excellence in Nanoscience, University of Science and Technology of China, Hefei, Anhui 230026, China*

**DaDi Zhang** − *National Synchrotron Radiation Laboratory, University of Science and Technology of China, Hefei, Anhui 230026, China*

**Rui-Xue Xu** − *Hefei National Laboratory for Physical Sciences at the Microscale & Synergetic Innovation Center of Quantum Information and Quantum Physics & CAS Center for Excellence in Nanoscience, University of Science and Technology of China, Hefei, Anhui 230026, China*

**ChiYung Yam** − *Beijing Computational Science Research Center, Beijing 100193, China;* orcid.org/0000-0002-3860-2934

**GuanHua Chen** − *Department of Chemistry, The University of Hong Kong, Hong Kong, China;* orcid.org/0000-0001-5015-0902

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jpca.1c10491

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

The support from the Ministry of Science and Technology of China (Grant No. 2016YFA0400900) and the National Natural Science Foundation of China (Grant Nos. 21973086 and 21633006) is gratefully acknowledged. G.C. acknowledges the support from the RGC General Research Fund (Grant No. 17309620) and Hong Kong Quantum AI Lab Ltd. Computational resources are provided by the Supercomputing Center of University of Science and Technology of China. J.W. is grateful for the guidance on numerical aspects from Dr. Tian Lu and colleagues at the Beijing Kein Research Center for Natural Sciences.

## ■ REFERENCES

(1) Jones, R. O. Density functional theory: Its origins, rise to prominence, and future. *Rev. Mod. Phys.* **2015**, *87*, 897−923.

(2) Hohenberg, P.; Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **1964**, *136*, B864−B871.

(3) Kohn, W.; Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **1965**, *140*, A1133−A1138.

(4) Becke, A. D. A new mixing of Hartree−Fock and local density-functional theories. *J. Chem. Phys.* **1993**, *98*, 1372−1377.

(5) Grimme, S. Semiempirical hybrid density functional with perturbative second-order correlation. *J. Chem. Phys.* **2006**, *124*, 034108.

(6) Zhang, Y.; Xu, X.; Goddard, W. A. Doubly hybrid density functional for accurate descriptions of nonbond interactions, thermochemistry, and thermochemical kinetics. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 4963−4968.

(7) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. Doubly hybrid meta DFT: New multi-coefficient correlation and density functional methods for thermochemistry and thermochemical kinetics. *J. Phys. Chem. A* **2004**, *108*, 4786−4791.

(8) Leininger, T.; Stoll, H.; Werner, H.-J.; Savin, A. Combining long-range configuration interaction with short-range density functionals. *Chem. Phys. Lett.* **1997**, *275*, 151−160.

(9) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **2003**, *118*, 8207−8215.

(10) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. A long-range correction scheme for generalized-gradient-approximation exchange functionals. *J. Chem. Phys.* **2001**, *115*, 3540−3544.

(11) Yanai, T.; Tew, D. P.; Handy, N. C. A new hybrid exchange−correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.* **2004**, *393*, 51−57.

(12) Verma, P.; Wang, Y.; Ghosh, S.; He, X.; Truhlar, D. G. Revised M11 exchange-correlation functional for electronic excitation energies and ground-state properties. *J. Phys. Chem. A* **2019**, *123*, 2966−2990.

(13) Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits. *J. Chem. Phys.* **2005**, *123*, 062201.

(14) Hait, D.; Head-Gordon, M. How accurate is density functional theory at predicting dipole moments? An assessment using a new database of 200 benchmark values. *J. Chem. Theory Comput.* **2018**, *14*, 1969−1981.

(15) Ryczko, K.; Strubbe, D. A.; Tamblyn, I. Deep learning and density-functional theory. *Phys. Rev. A* **2019**, *100*, 022512.

(16) Denner, M. M.; Fischer, M. H.; Neupert, T. Efficient learning of a one-dimensional density functional theory. *Phys. Rev. Res.* **2020**, *2*, 033388.

(17) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theory Comput.* **2017**, *13*, 5255−5264.

(18) Margraf, J. T.; Reuter, K. Pure non-local machine-learned density functional theory for electron correlation. *Nat. Commun.* **2021**, *12*, 344.

(19) Chandrasekaran, A.; Kamal, D.; Batra, R.; Kim, C.; Chen, L.; Ramprasad, R. Solving the electronic structure problem with machine learning. *npj Comput. Mater.* **2019**, *5*, 22.

(20) Fowler, A. T.; Pickard, C. J.; Elliott, J. A. Managing uncertainty in data-derived densities to accelerate density functional theory. *J. Phys.: Mater.* **2019**, *2*, 034001.

(21) Wellendorff, J.; Lundgaard, K. T.; Møgelhøj, A.; Petzold, V.; Landis, D. D.; Nørskov, J. K.; Bligaard, T.; Jacobsen, K. W. Density functionals for surface science: Exchange-correlation model development with Bayesian error estimation. *Phys. Rev. B* **2012**, *85*, 235149.

(22) Lei, X.; Medford, A. J. Design and analysis of machine learning exchange-correlation functionals via rotationally invariant convolutional descriptors. *Phys. Rev. Mater.* **2019**, *3*, 063801.

(23) Aldegunde, M.; Kermode, J. R.; Zabaras, N. Development of an exchange−correlation functional with uncertainty quantification capabilities for density functional theory. *J. Comput. Phys.* **2016**, *311*, 173−195.

(24) Nelson, J.; Tiwari, R.; Sanvito, S. Machine learning density functional theory for the Hubbard model. *Phys. Rev. B* **2019**, *99*, 075132.

(25) Schmidt, J.; Benavides-Riveros, C. L.; Marques, M. A. L. Machine learning the physical nonlocal exchange−correlation functional of density-functional theory. *J. Phys. Chem. Lett.* **2019**, *10*, 6425−6431.

(26) Wellendorff, J.; Lundgaard, K. T.; Jacobsen, K. W.; Bligaard, T. mBEEF: An accurate semi-local Bayesian error estimation density functional. *J. Chem. Phys.* **2014**, *140*, 144107.

(27) Nagai, R.; Akashi, R.; Sasaki, S.; Tsuneyuki, S. Neural-network Kohn-Sham exchange-correlation potential and its out-of-training transferability. *J. Chem. Phys.* **2018**, *148*, 241737.

(28) Fritz, M.; Fernández-Serra, M.; Soler, J. M. Optimization of an exchange-correlation density functional for water. *J. Chem. Phys.* **2016**, *144*, 224101.

(29) Tsubaki, M.; Mizoguchi, T. Quantum deep field: Data-driven wave function, electron density generation, and atomization energy prediction and extrapolation with machine learning. *Phys. Rev. Lett.* **2020**, *125*, 206401.

(30) Seino, J.; Kageyama, R.; Fujinami, M.; Ikabata, Y.; Nakai, H. Semi-local machine-learned kinetic energy density functional with third-order gradients of electron density. *J. Chem. Phys.* **2018**, *148*, 241705.

(31) Suzuki, Y.; Nagai, R.; Haruyama, J. Machine learning exchange-correlation potential in time-dependent density-functional theory. *Phys. Rev. A* **2020**, *101*, 050501.

(32) Tozer, D. J.; Ingamells, V. E.; Handy, N. C. Exchange-correlation potentials. *J. Chem. Phys.* **1996**, *105*, 9200−9213.

(33) Zhao, Q.; Morrison, R. C.; Parr, R. G. From electron densities to Kohn-Sham kinetic energies, orbital energies, exchange-correlation potentials, and exchange-correlation energies. *Phys. Rev. A* **1994**, *50*, 2138−2142.

(34) Zhou, Y.; Wu, J.; Chen, S.; Chen, G. Toward the exact exchange−correlation potential: A three-dimensional convolutional neural network construct. *J. Phys. Chem. Lett.* **2019**, *10*, 7264−7269.

(35) Nagai, R.; Akashi, R.; Sugino, O. Completing density functional theory by machine learning hidden messages from molecules. *npj Comput. Mater.* **2020**, *6*, 43.

(36) Vargas-Hernández, R. A. Bayesian optimization for calibrating and selecting hybrid-density functional models. *J. Phys. Chem. A* **2020**, *124*, 4053−4061.

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

(38) Yu, M.; Yang, S.; Wu, C.; Marom, N. Machine learning the Hubbard U parameter in DFT+U using Bayesian optimization. *npj Comput. Mater.* **2020**, *6*, 180.

(39) Zheng, X.; Hu, L.; Wang, X.; Chen, G. A generalized exchange-correlation functional: the Neural-Networks approach. *Chem. Phys. Lett.* **2004**, *390*, 186−192.

(40) Liu, Q.; Wang, J.; Du, P.; Hu, L.; Zheng, X.; Chen, G. Improving the performance of long-range-corrected exchange-correlation functional with an embedded neural network. *J. Phys. Chem. A* **2017**, *121*, 7273−7281.

(41) Snyder, J. C.; Rupp, M.; Hansen, K.; Müller, K.-R.; Burke, K. Finding density functionals with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 253002.

(42) 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.

(43) Dick, S.; Fernandez-Serra, M. Learning from the density to correct total energy and forces in first principle simulations. *J. Chem. Phys.* **2019**, *151*, 144102.

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

(45) Chen, Y.; Zhang, L.; Wang, H.; E, W. DeePKS: A comprehensive data-driven approach toward chemically accurate density functional theory. *J. Chem. Theory Comput.* **2021**, *17*, 170−181.

(46) Handy, N. C. The importance of Colle−Salvetti for computational density functional theory. *Theor. Chem. Acc.* **2009**, *123*, 165−169.

(47) Kim, M.-C.; Sim, E.; Burke, K. Understanding and reducing errors in density functional calculations. *Phys. Rev. Lett.* **2013**, *111*, 073003.

(48) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.

(49) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Constantin, L. A.; Zhou, X.; Burke, K. Restoring the density-gradient expansion for exchange in solids and surfaces. *Phys. Rev. Lett.* **2008**, *100*, 136406.

(50) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **1980**, *58*, 1200−1211.

(51) Perdew, J. P.; Wang, Y. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B* **1992**, *45*, 13244−13249.

(52) Functionals were obtained from the Density Functional Repository as developed and distributed by the Quantum Chemistry Group, CCLRC Daresbury Laboratory, Daresbury, Cheshire, WA4 4AD U.K. Contact Huub van Dam (h.j.j.vandam@dl.ac.uk) or Paul Sherwood for further information.

(53) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-2 and density functional theories for the computation of enthalpies of formation. *J. Chem. Phys.* **1997**, *106*, 1063−1079.

(54) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-3 and density functional theories for a larger experimental test set. *J. Chem. Phys.* **2000**, *112*, 7374−7383.

(55) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Pople, J. A. Assessment of Gaussian-2 and density functional theories for the computation of ionization potentials and electron affinities. *J. Chem. Phys.* **1998**, *109*, 42−55.

(56) Luo, Y.-R. *Handbook of Bond Dissociation Energies in Organic Compounds*; CRC Press: Boca Raton, FL, 2003.

(57) Prasad, V. K.; Khalilian, M. H.; Otero-de-la Roza, A.; DiLabio, G. A. BSE49, a diverse, high-quality benchmark dataset of separation energies of chemical bonds. *Sci. Data* **2021**, *8*, 300.

(58) Zheng, J.; Zhao, Y.; Truhlar, D. G. The DBH24/08 database and its use to assess electronic structure model chemistries for chemical reaction barrier heights. *J. Chem. Theory Comput.* **2009**, *5*, 808−821.

(59) Peverati, R.; Truhlar, D. G. Quest for a universal density functional: the accuracy of density functionals across a broad spectrum of databases in chemistry and physics. *Philos. Trans. R. Soc. London, Ser. A* **2014**, *372*, 20120476.

(60) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. Design of density functionals by combining the method of constraint satisfaction with parametrization for thermochemistry, thermochemical kinetics, and noncovalent interactions. *J. Chem. Theory Comput.* **2006**, *2*, 364−382.

(61) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; et al. *Gaussian 16*, rev. C.01; Gaussian Inc: Wallingford, CT, 2016.

(62) Mura, M. E.; Knowles, P. J. Improved radial grids for quadrature in molecular density-functional calculations. *J. Chem. Phys.* **1996**, *104*, 9848−9858.

(63) Lebedev, V. I.; Laikov, D. N. A quadrature formula for the sphere of the 131st algebraic order of accuracy. *Dokl. Math.* **1999**, *59*, 477−481.

(64) Hu, X.; Hu, H.; Zheng, X.; Zeng, X.; Wu, P.; Peng, D.; Jin, Y.; Yu, L.; Yang, W. *An In-House Program for QM/MM Simulations*. http://www.qm4d.info (accessed 2019-01-01).

(65) Hui, K.; Chai, J.-D. SCAN-based hybrid and double-hybrid density functionals from models without fitted parameters. *J. Chem. Phys.* **2016**, *144*, 044114.

(66) Mo, Y.; Tian, G.; Tao, J. Performance of a nonempirical exchange functional from density matrix expansion: comparative study with different correlations. *Phys. Chem. Chem. Phys.* **2017**, *19*, 21707−21713.

(67) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. Effectiveness of diffuse basis functions for calculating relative energies by density functional theory. *J. Phys. Chem. A* **2003**, *107*, 1384−1388.

(68) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes. *J. Chem. Phys.* **2003**, *119*, 12129−12137.

(69) *NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101*, 2020. Johnson, R. D., III, Ed. https://cccbdb.nist.gov/ (accessed 2021-04-01).

(70) Maho, N. The PubChemQC project: A large chemical database from the first principle calculations. *AIP Conf. Proc.* **2015**, *1702*, 090058.

(71) Nakata, M.; Shimazaki, T. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *J. Chem. Inf. Model.* **2017**, *57*, 1300−1308.

(72) Nakata, M.; Shimazaki, T.; Hashimoto, M.; Maeda, T. PubChemQC PM6: Data Sets of 221 Million Molecules with Optimized Molecular Geometries and Electronic Properties. *J. Chem. Inf. Model.* **2020**, *60*, 5891−5899.

(73) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceeedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, 2016; pp 785−794.

(74) Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **2001**, *29*, 1189−1232.

(75) Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367−378.

(76) Sun, J.; Ruzsinszky, A.; Perdew, J. P. Strongly constrained and appropriately normed semilocal density functional. *Phys. Rev. Lett.* **2015**, *115*, 036402.

(77) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098−3100.

(78) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785−789.

(79) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(80) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Fractional charge perspective on the band gap in density-functional theory. *Phys. Rev. B* **2008**, *77*, 115123.

(81) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Insights into current limitations of density functional theory. *Science* **2008**, *321*, 792−794.

(82) Li, C.; Zheng, X.; Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Local scaling correction for reducing delocalization error in density functional approximations. *Phys. Rev. Lett.* **2015**, *114*, 053001.

(83) Zheng, X.; Liu, M.; Johnson, E. R.; Contreras-García, J.; Yang, W. Delocalization error of density-functional approximations: A distinct manifestation in hydrogen molecular chains. *J. Chem. Phys.* **2012**, *137*, 214106.

(84) Li, C.; Zheng, X.; Su, N. Q.; Yang, W. Localized orbital scaling correction for systematic elimination of delocalization error in density functional approximations. *Natl. Sci. Rev.* **2018**, *5*, 203−215.

(85) Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz, J. L. Density-Functional Theory for fractional particle number: Derivative discontinuities of the energy. *Phys. Rev. Lett.* **1982**, *49*, 1691−1694.

(86) Zheng, X.; Cohen, A. J.; Mori-Sánchez, P.; Hu, X.; Yang, W. Improving band gap prediction in density functional theory from molecules to solids. *Phys. Rev. Lett.* **2011**, *107*, 026403.

(87) Zheng, X.; Li, C.; Zhang, D.; Yang, W. Scaling correction approaches for reducing delocalization error in density functional approximations. *Sci. China Chem.* **2015**, *58*, 1825−1844.

(88) Zhang, I. Y.; Wu, J.; Xu, X. Extending the reliability and applicability of B3LYP. *Chem. Commun.* **2010**, *46*, 3057−3070.

(89) Kirkpatrick, J.; McMorrow, B.; Turban, D. H. P.; Gaunt, A. L.; Spencer, J. S.; Matthews, A. G. D. G.; Obika, A.; Thiry, L.; Fortunato, M.; Pfau, D.; et al. Pushing the frontiers of density functionals by solving the fractional electron problem. *Science* **2021**, *374*, 1385−1389.