CIDER: An Expressive, Nonlocal Feature Set for Machine Learning Density Functionals with Exact Constraints

Kyle Bystrom* and Boris Kozinsky*

Cite This: J. Chem. Theory Comput. 2022, 18, 2180–2192

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Machine learning (ML) has recently gained attention as a means to develop more accurate exchange-correlation (XC) functionals for density functional theory, but functionals developed thus far need to be improved on several metrics, including accuracy, numerical stability, and transferability across chemical space. In this work, we introduce a set of nonlocal features of the density called the CIDER formalism, which we use to train a Gaussian process model for the exchange energy that obeys the critical uniform scaling rule for exchange. The resulting CIDER exchange functional is significantly more accurate than any semilocal functional tested here, and it has good transferability across main-group molecules. This work therefore serves as an initial step toward more accurate exchange functionals, and it also introduces useful techniques for developing robust, physics-informed XC models via ML.

1. INTRODUCTION

Density functional theory (DFT) is an indispensable tool in computational chemistry and materials science due to its combination of efficiency and accuracy. While an exact theory, the exchange-correlation (XC) functional describing quantum mechanical effects must be approximated in practice, and this approximation is the key limiting factor in the accuracy of DFT. Hundreds of approximate XC functionals have been developed in an attempt to calculate accurate chemical data, but significant progress remains to be made in the development of highly accurate XC functionals. Several reviews cover the shortcomings of current XC functionals, such as self-interaction error, overdelocalization, and inconsistent description of static correlation.

The shortcomings of existing approximations have motivated the development of machine learning (ML) models for more accurate functionals. This concept was first introduced by Tozer et al. and more recently pioneered by Burke and coworkers. Some of these models use projections of the electron density or density matrix onto atom-centered basis sets as input to a ML model, resulting in atomic decomposition of the XC energy. These models are highly accurate across a small set of systems similar to those on which the model is trained, but they do not match the universality of most conventional XC functionals. For example, DeePKS is only applicable to closed-shell organic molecules. Another recent approach uses a real-space grid and convolves the density to create features in real space. This approach has good transferability across chemical space and requires less training data than other ML functionals (as shown by Nagai et al., who used a training set of only three molecules), but the accuracy achieved thus far is not consistently better than conventional semiempirical XC functionals. For example, the NN-NRA by Nagai et al. is more accurate than M06-L and M06 for main-group ionization potentials but less accurate for barrier heights.

The latter real-space approach is effectively an extension of the semipirical approach to functional design, in which a parametric functional is fit to experimental or theoretical training data. There is a trade-off between designing functionals that fit experimental data well and functionals that obey analytically known constraints on the exact functional. It has been shown that semipirical functionals have improved accuracy for energetic data, but breaking exact constraints can make the functional less universal and transferable, resulting in poor description of density distributions. Exact constraints for ML functionals have not received much attention thus far. However, one study on the 1D kinetic energy functional showed that the uniform scaling constraint improved the learning curve of a kernel ridge regression-based functional, and several constraints were enforced in a recent ML meta-GGA by Dick and Fernández-Serra. One can also explicitly train a ML model to reproduce the correct density; it has been shown that training

Received: September 8, 2021
Published: March 2, 2022
on both energetic data and density distributions can yield ML functionals that accurately predict both properties.\textsuperscript{14,17,23--25}

Comparing the various approaches above, it seems that the atomic decomposition approach lacks the physical intuition and data efficiency to develop broadly transferable functionals for the entire periodic table, whereas existing real-space approaches lack the flexibility to describe systems to a high degree of accuracy while maintaining sensible physical behavior like accurate density distributions. Therefore, we seek to develop a framework for constructing descriptors of the density distribution that can be used to design accurate and transferable XC functionals with ML. Such features must describe variations in the density that correlate with the target XC energy, but they do not need to have an intuitive physical meaning because the ML algorithm can learn the relationship between the features and the XC energy. Ideally, the features should enable a ML model trained on limited data to generalize across a broad range of chemistries.

For this work, we focus on the Kohn–Sham (KS) exchange functional\textsuperscript{26}

$$E_x^{\text{ex}}[\gamma] = \langle \Phi^{\text{KS}}[\gamma] | \hat{V}_c | \Phi^{\text{KS}}[\gamma] \rangle - U[\gamma]$$

(1)

where $\Phi^{\text{KS}}[\gamma]$ is the KS Slater determinant, $\hat{V}_c = \sum_i e^{-r_i}$ is the Coulomb repulsion operator, and $U[\gamma] = \int d\mathbf{r}_1 d\mathbf{r}_2 n(\mathbf{r}_1)n(\mathbf{r}_2)$ is the classical Coulomb interaction.

There are several reasons for focusing on the exchange functional rather than the full XC functional. First, compared to the full XC energy, it is much easier to collect training data for $E_x^{\text{ex}}[\gamma]$ because it can be computed analytically from a KS density matrix. This makes it ideal for initial studies on real molecular systems. Second, evaluating the exact exchange energy is usually the computational bottleneck of hybrid DFT calculations, in which a fraction of exact exchange is mixed into an otherwise semilocal XC functional. For periodic solids, hybrid DFT can be orders of magnitude more expensive than semilocal DFT and scales poorly with system size, which places limits on its applications.\textsuperscript{27} There have been recent developments in the efficient computation of the exchange energy for periodic systems,\textsuperscript{27--30} but these approaches are still more costly than semilocal DFT. In addition, the linear-scaling approaches are only applicable to systems with band gaps, making them impractical for applications like catalysis on metal surfaces. An efficient ML exchange model could drastically reduce the computational cost of a calculation compared to hybrid DFT while preserving its accuracy. This is important because there are some technologically relevant systems, like semiconductor point defects\textsuperscript{31,32} and battery materials,\textsuperscript{33} for which semilocal DFT is inaccurate but can be corrected by mixing an empirically tuned fraction of exact exchange to form a hybrid functional.\textsuperscript{34--36}

Similar applications might be found in areas like heterogeneous catalysis,\textsuperscript{17} where semilocal DFT is also frequently inadequate to describe the systems of interest. A ML exchange functional could make accurate studies of these systems faster and more practical.

Studying the exchange functional on its own is also important for developing a robust combined XC functional. The exchange and correlation energy errors tend to cancel each other because the exchange-correlation hole is more localized than the exchange or correlation holes on their own. However, this cancellation of error is neither controlled nor universal.\textsuperscript{11} This means that in situations where such cancellation of error does not occur, such as self-interaction-dominated systems, an otherwise accurate XC functional could fail unexpectedly. Having an exchange functional that is accurate on its own provides a baseline for ensuring physical behavior in these systems. Semilocal functionals do rely heavily on cancellation of error effects, so it is an open question what descriptors, if any, are capable of describing the shape of the exchange hole with sufficient accuracy and efficiency to replace the exact exchange operator. Knowing the answer to this question could assist future functional design.

To demonstrate that ML can be used to design transferable and accurate functionals, we introduce the compressed scale-invariant density representation (CIDER)—a set of descriptors of the density distribution that are invariant under uniform scaling of the density (i.e., for $n(\mathbf{r}) = \gamma n(\gamma \mathbf{r})$, invariant with respect to $\gamma$)—and use it to train a Gaussian process for the exchange functional. This scale invariance allows the exchange model to obey the uniform scaling rule $E_x[\gamma n] = \gamma E_x[n]$. The resulting exchange functional requires only 119 training systems to be transferable across the first four rows of the periodic table. It outperforms all semilocal exchange functionals investigated here for predicting the total and atomization exchange energies, and it accurately reproduces main-group atomization energies when replacing the exact exchange contribution in B3LYP.\textsuperscript{38} It has good numerical stability, allowing it to be used in self-consistent field calculations with standard integration grids. The CIDER approach thus provides an initial step toward a robust and efficient alternative to approximating exact exchange.

The rest of the paper is structured as follows. Section 2 introduces the CIDER formalism, and section 3 describes the Gaussian process models used with the CIDER descriptors to train exchange functionals. Section 4 describes the methods for performing the DFT calculations and training the CIDER functional, and section 5 contains the results and a discussion of the new functional’s performance across a diverse set of chemistries. Finally, section 6 concludes with a summary of the findings.

## 2. CIDER FORMALISM

To develop a ML model for the exchange energy, an expressive set of nonlocal descriptors of the density must be used as input to the model. One could use a neural network (NN) to learn the features from the raw density distribution in real space, but training features this way is data intensive, with $10^5$–$10^6$ training points used in recent works.\textsuperscript{39,40} In addition, these NNs rely on a specific grid structure over which convolutions are performed, which could impede their use in realistic production calculations. Alternatively, one could project the density or density matrix onto atomic basis sets, as is done in NeuralXC and DeePKS,\textsuperscript{15,14} but these two models do not incorporate any physical constraints into the features, making it infeasible to incorporate exact constraints into the model itself. The difficulty of incorporating physical constraints and intuition into such models could limit their transferability and universality. A compromise between these two approaches is to design features based on nonlocal convolutions of the density, as done by Lei and Medford\textsuperscript{16} and by Nagai et al.,\textsuperscript{17} and then use these features as input to a ML model. We seek to improve on this third approach by designing descriptors that are highly expressive and which also constrain the resulting ML model to known properties of the exact functional.

The most important constraint for the exchange energy is derived from the principle of uniform scaling. Consider a density distribution $n(\mathbf{r})$ and a scaled density

$$n_\gamma(\mathbf{r}) = \gamma^3 n(\gamma \mathbf{r})$$

(2)
Several important exact constraints can be written using this scaled density. They include the uniform scaling rules for the noninteracting kinetic energy \( T[n] \) and exchange energy \( E[n] \) above:

\[
T[n] = \gamma^2 T[n]
\]

\( E[n] = \gamma E[n] \)  

Equation 4 implies that the exchange energy can be written as

\[
E[n] = -\gamma \left( \frac{3}{4} \right)^{1/3} \int d^3 r E_x[n](r) n^{4/3}(r) \]  

which is the form used by most semilocal exchange functionals. The constant in front of the integral is chosen so that \( E_{\text{TIGHT}} = 1 \) for the homogeneous electron gas (HEG). The functional for a spin-unpolarized density \( E_{\text{U}}[n] \) can be extended to the spin-polarized case using the spin-scaling rule:

\[
E[n, n'] = \frac{1}{2} (E_{\text{U}}(2n) + E_{\text{U}}(2n'))
\]

Because eq 6 uniquely and simply defines the spin-polarized exchange energy from the spin-unpolarized exchange energy, the remainder of the discussion in this section refers to the spin-unpolarized case.

The exchange enhancement factor (XEF) \( F_\alpha[n](r) \) in eq 5 is independent of \( \gamma \), a property which will be referred to as scale invariance. Therefore, it is reasonable to predict that a ML model for \( F_\alpha[n] \) will learn more efficiently if the feature vector \( \mathbf{x} \) used as input to the model is scale invariant (\( x[n](r) = x[n](\gamma r) \)). The conventional descriptors of the gradient \( \nabla n \) and kinetic energy density \( \tau = \frac{1}{2} \sum \frac{\nabla^2 \phi_i^2}{V_n} \) satisfy these rules:

\[
s = \frac{2(3\pi)^{1/3} n^{4/3}}{\nabla n}
\]

\[
\alpha = \frac{\tau - \tau_0}{\tau_0}
\]

where \( \tau_0 = |\nabla n|^{1/2}/8n \) is the kinetic energy density of a single-orbital system and \( \tau_0 = (3/10)(3\pi)^{2/3} n^{4/3} \) is the kinetic energy density of the HEG. The descriptor \( \alpha \) was first introduced in by Sun et al. \(^{14}\) While these descriptors are useful, they are semilocal, so they cannot fully encode the complex, nonlocal structure of the exchange functional. The challenge is therefore to construct a set of nonlocal descriptors that are scale invariant, which can be used for accurately training a ML model of the functional \( E[n] \).

Nonlocality can be introduced to the features by performing convolutions on the density with a short-range kernel, as done in the near-region approximation (NRA) by Nagai et al. \(^{17}\) However, these features are not scale invariant. To achieve scale invariance, we use an approach similar to that developed by Janesko and co-workers for "Rung 3.5" semiempirical functionals. \(^{45-47}\) Rung 3.5 functionals use the one-particle density matrix \( n_i(\mathbf{r}, \mathbf{r}') = \sum \phi_i(\mathbf{r}) \phi_i(\mathbf{r}') \)

where \( \phi_i(\mathbf{r}) \) are the Kohn–Sham orbitals and \( f_i \) are the occupation numbers (2 for the occupied orbitals in a spin-unpolarized system and 0 for unoccupied orbitals). The density matrix \( n_i(\mathbf{r}, \mathbf{r}') \) is projected onto a semilocal model for the density matrix at each point \( \mathbf{r} \) to construct the Rung 3.5 energy density \( e_{\text{Rung} 3.5}[n](\mathbf{r}) \):

\[
e_{\text{Rung} 3.5}[n](\mathbf{r}) = \int d^3 r' n_i(\mathbf{r}, \mathbf{r}') n_i(\mathbf{r} - \mathbf{r}' ; \mathbf{r})
\]

Importantly, the model density matrix \( n_i(\mathbf{r} - \mathbf{r} ; \mathbf{r}) \) is position dependent in a way that gives \( e_{\text{Rung} 3.5}[n](\mathbf{r}) \) conventional scaling properties. For example, if \( n_i(\mathbf{r} - \mathbf{r} ; \mathbf{r}) \) is the exchange hole of the HEG then \( e_{\text{Rung} 3.5}[n](\mathbf{r}) = \gamma e_{\text{Rung} 3.5}[n](\gamma \mathbf{r}) \).

Following this approach, we introduce a scale-invariant set of integral descriptors to describe the density distribution around a point

\[
G_{\text{hm}}(\mathbf{r}) = \int d^3 r G_{\text{hm}}(\mathbf{r}' ; \mathbf{r}) n(\mathbf{r} + \mathbf{r}')
\]

\[
g_{\text{hm}}(\mathbf{r}' ; \mathbf{r}) = \frac{3^{2/3}}{4\pi^{1/3}} \frac{8\pi^{2/3}}{3} Y_m^0(\hat{\mathbf{r}}') |\hat{\mathbf{r}}'| e^{-a|\hat{\mathbf{r}}|^2}
\]

In the above equation, the functions \( Y_m^0(\hat{\mathbf{r}}') \) represent the spherical harmonics and the exponent \( a \) is a function of \( \mathbf{r} \) and a semilocal functional of the density

\[
n(\mathbf{r}) \propto e^{-(\sqrt{\alpha} \mathbf{r})}
\]

The \( B_0 = 1 \) case is important because the exchange hole integrates to zero (\( -1 \) if summed over spin for a spin-unpolarized system), so for \( B_0 = 1, g_\text{Rung} \approx 2 \) is approximately shaped like the HEG exchange hole and has the correct norm for the exchange hole for the HEG. This gives the feature a sensible shape and length scale, and it also gives a known HEG reference value for that feature, allowing the HEG constraint to be enforced.

If \( B_0 = C_0 \), \( a \) is finite in atomic core regions but will vanish at the center of single bonds, where \( \tau = 0 \). If \( C_0 = 0, a \) will vanish at the atomic tails as the density goes to zero. In each case, the length scale of the Gaussian distribution becomes infinite, which is incompatible with the goal of a computationally efficient functional and is also poor inductive bias because the exchange hole is localized around the reference point. Therefore, \( B_0 \) must be greater than \( C_0 \) and \( C_0 \) must be positive.

In this work, we use the constant \( A \) to define the length scale, such that

\[
B_0 = A
\]

\[
C_0 = \frac{A}{32 \pi^2 5 \sqrt{5}} \approx 0.18A
\]

This choice was found to yield a smooth length scale in real space. We also define several choices of \( A \) for the model, resulting in different widths of the Gaussian kernels:

- **WIDE**, \( A = 1 \)
- **MEDIUM-WIDE**, \( A = 2 \)
- **MEDIUM-TIGHT**, \( A = 4 \)
- **TIGHT**, \( A = 8 \)
These nonlocal features, along with the semilocal features $s$ and $\alpha$, constitute the compressed scale-invariant density representation (CIDER), so named because the features satisfy $G_{\text{fin}}(n, \gamma) = G_{\text{fin}}[n](\sqrt{r})$ and provide an efficient representation of the density distribution around a point $r$ in real space.

The XC energy is invariant under translation and rotation of the system. The raw CIDER descriptors above are translationally invariant but not rotationally invariant. To create rotationally invariant descriptors, contractions using Clebsch–Gordon coefficients are performed. This is similar to the approaches used in the tensor field network, the moment tensor potential, the RIDR functional, and Lei and Medford’s nonlocal density features, which are recently developed ML methodologies for chemistry and materials science.

In its current iteration, the CIDER descriptors are used to define the feature vector $x$, where $G_{nl}$ is a vector containing $G_{\text{fin}}$ for $m \in \{-2l - 1, -2l, ..., 2l + 1\}$

$$
\begin{align*}
x_1 &= s^2 \\
x_2 &= \alpha \\
x_3 &= G_{oo} \\
x_4 &= |G_{o1}|^2 \\
x_5 &= |G_{o2}|^2 / \sqrt{5} \\
x_6 &= s \cdot G_{o1} \\
x_7 &= C(s, G_{o2}, s) \\
x_8 &= C(s, G_{o2}, G_{o1}) \\
x_9 &= G_{oo}(2^{-4/3}A) \\
x_{10} &= G_{oo}(2^{4/3}A)
\end{align*}
$$

Eq 17

$G_{oo}(\lambda A)$ indicates that the length-scale parameter $A$ was changed to $\lambda A$ for this integral. $C(a, b, c)$ contracts the two $l = 1$ terms $a, c$ and the $l = 2$ term $b$ to an $l = 0$ term using Clebsh–Gordon coefficients (see Appendix B for details). For length-scale parameter $A$, eq 17, contains all possible rotationally invariant contractions of features with $l \leq 2$ with the exception of $x_{11} = C(G_{oo}, G_{o2}, G_{oo})$. We found that this set of features provides a good balance between computational efficiency and model expressiveness. In principle, one could expand the feature set by including contractions of features with $l > 2$ and using multiple different length-scale parameters, but to keep the computational cost of feature evaluation tractable, this is not done here. One could also attempt to learn on smaller feature sets, including a semilocal model containing only $x_1$ and $x_2$. However, as discussed in the Supporting Information (section S2.2), this results in insufficiently accurate models. In section S2.2, we also show that adding the feature $x_{11}$ does not improve the accuracy, and excluding the additional $l = 0$ features $x_9$ and $x_{10}$ is detrimental to accuracy. We therefore conclude that eq 17 is a reasonable choice of feature vector.

The above descriptors will be referred to as Version A. We also tried introducing $G_{\text{fin}}$ descriptors for $n \neq 0$, yielding Version C (Compact) descriptors, which are the same as Version A except for the last two

$$
\begin{align*}
x_9 &= G_{10} \\
x_{10} &= G_{oo}(2A)
\end{align*}
$$

While $x_9$ is a “tight” descriptor like in Version A, $x_{10}$ has the same exponent as the other nonlocal features but is multiplied by $\alpha^2$. This version allows us to examine the effects of eliminating the widest feature (thus “Compact”), which is useful because shorter range features might be more amenable to computationally efficient evaluation in optimized implementations of the model.

Normalizing features is generally helpful in machine learning applications. Using the descriptors developed by Becke,50 Becke and Edgecombe,51 and Mardirossian and Head-Gordon52 as guidance, we apply the following transformations to the above descriptors

$$
\begin{align*}
x_1 &\rightarrow \frac{\gamma_x x_1}{1 + \gamma_x x_1} \\
x_2 &\rightarrow \frac{2}{1 + x_2^2} - 1 \\
x_3 &\rightarrow \frac{\gamma_{oo} x_3}{1 + \gamma_{oo} x_3} - c_3 \\
x_4 &\rightarrow \frac{\gamma_x x_4}{1 + \gamma_x x_4} \\
x_5 &\rightarrow \frac{\gamma_{oo} x_5}{1 + \gamma_{oo} x_5} \\
x_6 &\rightarrow x_6 \sqrt{\frac{\gamma_x}{1 + \gamma_x x_1}} \sqrt{\frac{\gamma_1}{1 + \gamma_1 x_4}} \\
x_7 &\rightarrow x_7 \sqrt{\frac{\gamma_x}{1 + \gamma_x x_1}} \sqrt{\frac{\gamma_2}{1 + \gamma_2 x_5}} \\
x_8 &\rightarrow x_8 \sqrt{\frac{\gamma_x}{1 + \gamma_x x_1}} \sqrt{\frac{\gamma_1}{1 + \gamma_1 x_4}} \sqrt{\frac{\gamma_2}{1 + \gamma_2 x_5}} \\
x_9 &\rightarrow \frac{\gamma_{oo} x_9}{1 + \gamma_{oo} x_9} - c_9 \\
x_{10} &\rightarrow \frac{\gamma_{oo} x_{10}}{1 + \gamma_{oo} x_{10}} - c_{10}
\end{align*}
$$

The resulting transformed descriptors all fall in a finite range, making them more convenient for ML models. The hyperparameters in the equations above were selected heuristically as described in the Supporting Information (section S1), but they could also be optimized, if desired, by treating them as hyperparameters of Gaussian process regression models, as explained in section 3. The constants $c_7, c_9$, and $c_{10}$ guarantee that the zero feature vector $x = 0$ corresponds to the HEG. The HEG limit can therefore be enforced by setting $F_x = 1$ for $x = 0$ in eq S.5.

3. GAUSSIAN PROCESS EXCHANGE MODELS

To train ML models of the functionals, we employ Gaussian processes (GPs), which are commonly used for nonparametric regression in Bayesian statistical learning models.53 For a training set size $N$, matrix of inputs $X$ (the set of feature vectors $x^{(i)}$ for the training points $i = 1, ..., N$), vector of outputs $y$, and kernel function $k(x, x')$, the standard GP relations for the predictive mean $f(x)$ and variance $\sigma^2$ are

$$
\begin{align*}
f(x) &= K^{1T} (K + \sigma^2_{\text{noise}} I)^{-1} y \\
\sigma^2(x) &= k(x, x') - K^{1T} (K + \sigma^2_{\text{noise}} I)^{-1} k^{1T}
\end{align*}
$$
In the above equation, \( x \) is the test point, \( k \) is a vector containing \( k(x^{(i)}, x) \) for each \( x^{(i)} \) in \( X \), \( I \) is the identity matrix, and \( K_y = k(x^{(i)}, x^{(j)}) \) for each \( x^{(i)}, x^{(j)} \) in \( X \). \( k(x, x') \) can be any function satisfying the rules for an inner product, and it may contain a set of hyperparameters \( \theta \) that may need to be optimized. These hyperparameters can be chosen by maximizing the marginal likelihood

\[
\ln p(y|X, \theta) = -\frac{1}{2} y^T(K + \sigma^2_{\text{noise}}I)^{-1}y - \frac{1}{2} \ln|K + \sigma^2_{\text{noise}}I| - N \ln 2\pi
\]  

(22)

The noise parameter \( \sigma_{\text{noise}} \) can also be optimized in this manner. One popular covariance kernel is the squared exponential kernel

\[
k(x, x') = \exp\left(-\frac{1}{2}((x - x') \odot a^2)\right)
\]  

(23)

where \( \odot \) represents the element-wise product and \( a \) is a hyperparameter vector containing the inverse standard deviation in each direction in feature space. This kernel is used for some of the preliminary models discussed in the Supporting Information (section S1) and as the base kernel for the additive kernel discussed below. The squared exponential kernel yields highly expressive, nonlinear kernels because it is equivalent to linear regression on an infinite set of nonlinear basis functions.\(^{53,55}\) It is also is smooth and infinitely differentiable, which is important for ML functional models because they must be differentiated to obtain the exchange potential.

One drawback of the GP is that evaluating eq 20 for a single test point scales linearly with the training set size,\(^{54,55}\) which is constant for all observations. However, we expect observations of \( F_x \) to have a larger uncertainty for smaller \( n \). For this reason, each training point has a different kernel. The resulting predictive mean is mapped onto a sum of cubic splines, as described in previous work on molecular dynamics potentials.\(^{54,55}\) Because the maximum number of features per term in eq 25 is 3, the cubic splines are at most three dimensional.

We train our predictive function \( F_x^{\text{pred}}(x^{(n)}(r)) \) to the XEF, \( F_x[n](r) \), and evaluate \( E_{\text{s}}[n](r) \) via eq 5

\[
F_x[n](r) = \frac{e_x^{\text{ex}}[n](r)}{e_x^{\text{LDA}}(n(r))}
\]  

(26)

\[
e_x^{\text{ex}}[n](r) = -\frac{1}{4} \int d^3r \ln(|r - r'|)^2 \frac{\text{ln}(|r - r'|)^2}{|r - r'|}
\]  

(27)

\[
e_x^{\text{LDA}}(n) = -\frac{3}{4} \left(\frac{3}{\pi}\right)^{1/3} n^{4/3}
\]  

(28)

In eq 27, \( e_x^{\text{ex}}[n](r) \) is the exact exchange energy density and \( n_i(r, r') \) is the density matrix (eq 9). Equation 27 is not a unique definition, as any \( e_x^{\text{ex}}[n](r) \) for which

\[
E_{\text{s}}[n] = \int d^3 r e_x^{\text{ex}}[n](r)
\]  

(29)

is equally valid, but eq 27 is the easiest to compute and was found to be the easiest to which to fit the model as well. The Supporting Information (section S4) includes a discussion of alternative exchange energy densities. One of the challenges with the definition in eq 27 is that the XEF increases exponentially at the atomic tails. To make the model easier to train, the Chachiyo GGA exchange functional,\(^{59}\) which reproduces \( F_x \) at atomic tails, was used as a baseline functional, and the difference \( F_x[n](r) - F_x^{\text{Chachiyo}}[n](r) \) was learned.

The standard GP model uses a single noise parameter \( \sigma_{\text{noise}} \) which is constant for all observations. However, we expect observations of \( F_x \) to have a larger uncertainty for smaller densities because the exchange energy density is divided by \( n(r)^{4/3} \). Therefore, we use a heteroskedastic noise model in which each training point has a different noise parameter. This noise parameter is a function of the density \( n(r) \) and takes the form

\[
\sigma_{\text{noise}}^2(n) = v_1 + \frac{v_2}{1 + t_2 n} + \frac{v_3}{1 + t_3 n}
\]  

(30)

where \( t_2 = 50, t_3 = 10^6, \) and \( v_1, v_2, \) and \( v_3 \) are treated as hyperparameters and optimized by marginal likelihood maximization.

For numerical stability, the ML part of the functional is cut off at low densities. This is done using the function

\[
F_x = F_x^{\text{Chachiyo}} + F_x^{\text{ML}}
\]  

(31)
where

\[
J_{\text{cut}} = \begin{cases} 0 & \text{if } n < n_{\text{cut}}^\text{min} \\ \frac{1}{2} \left(1 - \cos\left(\frac{\pi n}{n_{\text{cut}}^\text{max}}\right)\right) & \text{if } n_{\text{cut}}^\text{min} \leq n \leq n_{\text{cut}}^\text{max} \\ 1 & \text{if } n > n_{\text{cut}}^\text{max} \end{cases}
\]

\[
\alpha_{\text{cut}} = \frac{\ln\left(n/n_{\text{cut}}^\text{min}\right)}{\ln(n_{\text{cut}}^\text{max}/n_{\text{cut}}^\text{min})}
\]

\[
n_{\text{cut}} = \max\left(n, n_{\text{cut}}^\text{max}\right)
\]

where \(n_{\text{cut}}^\text{max} = 10^{-3}\) and \(n_{\text{cut}}^\text{min} = 10^{-6}\) Bohr\(^{-3}\).

Also, for numerical stability, we define a value \(a_{\text{cut}}\) below which the exponent \(a(r)\) in eq 13 is exponentially damped so that it cannot go to zero

\[
a \rightarrow \begin{cases} a & a \geq a_{\text{cut}} \\ a_{\text{cut}} \exp\left(a/a_{\text{cut}} - 1\right) & a < a_{\text{cut}} \end{cases}
\]

\(a_{\text{cut}}\) is set to \(A/16\) for \(A\) in eq 15. This damping of the exponent violates the uniform scaling rule but only at low densities for very diffuse orbitals.

4. METHODS

4.1. Computational Details. The CIDER implementation was available in the CiderPress repository.\(^61\) To compute the CIDER descriptors, we first projected the density matrix onto the def2-SCF Coulomb fitting auxiliary basis.\(^62\) Then, the overlaps of the descriptor functions \(G_\text{best}(r; \mathbf{r})\) with the auxiliary basis were computed using the libcint\(^63\) library as interfaced through PySCF. These overlaps were contracted with the density to give the descriptors \(G_\text{best}(\mathbf{r})\), which were then used to compute the XEF. Then, the functional derivatives were computed and used to evaluate the exchange matrix elements. The functional was treated self-consistently except for the \(J_{\text{cut}}\) function in eq 32, which was not differentiated with respect to the density. The formulas for the functional derivatives and matrix elements are written in Appendix D. The fireworks\(^64\) package was used to automate calculation workflows.

4.2. Training and Testing the Gaussian Process. The data set developed by Jensen et al.\(^65\) was used to train, validate, and test the exchange model. This data set is an expansion of the G2/97 test set\(^66\) that includes elements that are under-represented in the G2/97 set as well as six nonbonded systems.\(^67\) The molecules were categorized by the elements they contained (Al, Be, B, Li, Mg, Na, P, Si, S, F, Cl), with the earlier elements in the list taking precedence (e.g., SF\(_6\) was categorized under S). The last set contained mostly organic molecules with none of these elements. Each subgroup was duplicated, and then for each data set of size \(N\), \(N_{\text{train}} = \text{floor}\left(0.4N\right)\), \(N_{\text{val}} = \text{round}\left(0.2N\right)\), and \(N_{\text{test}} = N - N_{\text{train}} - N_{\text{val}}\) molecules were placed in the training, validation, and test sets, respectively. He, Be, Ne, and Ar\(_2\) were placed in the training, validation, and test sets, respectively. Later it was realized that the data set contained a duplicated P\(_2\) molecule. Both duplicates were placed in the test set, so one was simply removed. The final partitions contained 79 training set molecules (55 closed shell, 24 open shell), 42 validation set molecules (29 closed shell, 13 open shell), and 90 test set molecules (64 closed shell, 26 open shell). In addition to these molecules, the training set was augmented with isolated atoms H–Kr as well as the excited spin states of Sc, Ti, V, and Cr with \(2S = 3, 4, 5,\) and 4, respectively. The inclusion of isolated atoms helped better describe the shapes of the tails of atoms, and it also introduced some transition-metal atoms to the training set. There were no transition-metal-containing systems in the Jensen data set.

The density matrices and energies for each system were evaluated self-consistently using PBE\(^68\) in the def2-QZVPPD basis set. From the PBE density matrix, the conventional exchange energy density (eq 27) was calculated in real space for each molecule on the level-3 grids implemented in PySCF. The same level-3 grids were used for all SCF calculations. The \(S = 0\) systems were performed in the Restricted Kohn–Sham (RKS) formalism, and the rest were performed in the unrestricted Kohn–Sham (UKS) formalism. From the real-space grids, 14,695 points were sampled from which to train the exchange enhancement factor model

- 481 points from closed-shell atoms,
- 4471 points from open-shell atoms,
- 6903 points from closed-shell molecules in the Jensen train partition,
- 2840 points from open-shell molecules in the Jensen train partition.

The training set size of 14,695 was chosen to maximize the training set size while keeping the computational cost tractable. For comparison, we benchmarked the train and validation performance of the CIDER functional for different training set sizes in Supporting Information section S2.1 and found only a weak dependence of the accuracy on training set size. The points were selected randomly from the set of points for which the electron density was greater than \(10^{-8}\) Bohr\(^{-3}\). Finally, the HEG limit was appended as an additional training point with the density set to \(10^8\) to minimize the uncertainty in eq 30. In summary, the CIDER model was trained to the exact exchange energy densities of density matrices obtained from self-consistent PBE calculations.

Gaussian process models with the kernel specified in eq 25 were trained to the training set described above. The parameters for the kernels were selected as described in detail in the Supporting Information (section S1). Several combinations of exact constraints, descriptor types, and descriptor length scales were tested, but we focused on three for the main body of this work.

- CIDER-X-AHW: Version A descriptors, Homogeneous Electron Gas (HEG) constraint, WIDE descriptors (\(A = 1\) in eq 15).
- CIDER-X-CHW: Same as CIDER-X-AHW but with Version C descriptors.
- CIDER-X-CHMT: Same as CIDER-X-CHW but with MEDIUM-TIGHT descriptors (\(A = 4\) in eq 15).

All models were implemented in the scikit-learn package,\(^69\) the additive RBF was implemented as a custom Kernel object. The additive kernel models were mapped to cubic splines using the interpolation.py package.\(^70\) The best-performing model on the validation set (CIDER-X-AHW) was selected as the final model.

To evaluate the accuracy and transferability of CIDER-X-AHW on static densities, the exchange energies and atomization energies were evaluated on the test set PBE density matrices using CIDER-X-AHW. To evaluate the self-consistent performance of the functional, the atomization energies of the Jensen set were computed self-consistently using two functionals: CIDER-X-AHW alone (i.e., replacing exact exchange with CIDER in an
HF calculation) and B3LYP-CIDER, in which the 20% HF exchange contribution of the B3LYP hybrid functional\(^{38}\) was replaced with 20% CIDER-X-AHW.

4.3. Benchmarking CIDER on the Minnesota Database. To evaluate the applicability of the CIDER functional to different properties, the B3LYP-CIDER functional was applied to the BH76 (barrier heights), IP23 (ionization potentials), and TMBE33 (transition-metal bond energies) data sets from the Minnesota 2015B Database\(^{71}\) as distributed in the ACCDB collection of databases.

The Minnesota Database contains some systems, especially metal-containing systems, which can settle into metastable densities due to the presence of symmetry-breaking ground states. To avoid these stable minima, all Minnesota Database calculations were performed in the UKS formalism. For these calculations, each system was initially converged with a PBE calculation, and its internal stability was then tested using the uhf\_internal tool in PySCF.\(^{60}\) If the test found an instability and returned new orbitals, these orbitals were used to initialize another UKS calculation. This was repeated until a stable ground state was found. Finally, an SCF calculation was performed with the functional of interest starting from the stable ground state of the PBE calculation. All calculations were performed in the def2-QZVPPD basis,\(^{73}\) with the def2 effective core potential (ECP)\(^ {76}\) used for elements of period 5 and above. Convergence of calculations was attempted to as low a threshold as 10\(^{-7}\) hartree (Ha), but calculations were accepted if they converged to 10\(^{-6}\) Ha.

To compare the nonrelativistic DFT calculations performed here with the relativistic reference values in the Minnesota Database, the spin–orbit corrections (SOC) from Supporting Information Table S19 of the revM06 paper\(^ {77}\) were applied.

5. RESULTS AND DISCUSSION

5.1. Static Density Benchmark. As shown in Figure 1, three different versions of the CIDER functional all outperform existing semilocal functionals for predicting the exact exchange energy of PBE density matrices. We chose to compare with other semilocal functionals,\(^ {80–84}\) because of their popularity, with the Minnesota Database containing some correlation-like character.\(^ {79}\) Due to the cancellation of error between the exchange and the correlation parts, however, it makes sense that the nonseparable XC part of MN15-L might achieve a lower error for exact exchange than other semilocal functionals. The Supporting Information (Table S1) contains the validation set errors for several other semilocal functionals,\(^ {80–84}\) which perform similarly to the functionals discussed here.

The best-performing CIDER functional in Figure 1 is CIDER-X-AHW, with an RMSE on the test set of 12.0 kcal/mol for exchange energy (EX) and 12.5 kcal/mol for atomization exchange energy (AEX), which is defined as the difference between EX and the value of EX for the molecule’s constituent atoms. These errors are only 18% (EX) and 27% (AEX) of the RMSE of the most accurate semilocal functional, the Chachiyo GGA.\(^ {59}\) Notably, the performance of all CIDER functionals is nearly identical on the train, validation, and test sets, indicating good transferability.

The other notable finding in Table 1 is that longer length-scale descriptors (CIDER-X-AHW) do not improve the description of EX over shorter length-scale descriptors (CIDER-X-CHMT), but they do give an improvement for AEX. In particular, the accuracy of AEX degrades as the length scale gets shorter, while the accuracy of EX stays about the same. This suggests that a descriptor with a short length scale cannot describe the localized atomic density and the more delocalized molecular density simultaneously. On the other hand, descriptors with longer length scales can accurately describe both single-center and multicenter exchange holes. Because of its accuracy on the
validation set for both EX and AEX, CIDER-X-AHW was used for the self-consistent field calculations below. The adjustable parameters for eqs 19 and 30 for CIDER-X-AHW are γ₀ = γ₀(0) = γ₀(1/2), γ₁ = 0.025, γ₂ = 0.015, v₁ = 10⁻⁶, v₂ = 0.000 503, and v₃ = 0.391.

5.2. Self-Consistent Field Calculations with B3LYP-CIDER. To be useful, the CIDER exchange functional must not only outperform semilocal exchange functionals but also accurately match the results of calculations performed using HF or hybrid functionals. To test this, SCF calculations were performed using CIDER-X-AHW and B3LYP-CIDER. Table 1 gives the mean absolute error (MAE) and RMSE of these functionals compared to HF and B3LYP, respectively.

To be useful, the CIDER exchange functional must not only outperform semilocal exchange functionals but also accurately match the results of calculations performed using HF or hybrid functionals. To test this, SCF calculations were performed using CIDER-X-AHW and B3LYP-CIDER. Table 1 gives the mean absolute error (MAE) and RMSE of these functionals compared to HF and B3LYP, respectively.

Before examining these results, it is worth noting that KS exact exchange (against which CIDER is trained) and HF exact exchange (against which CIDER is tested for SCF calculations) are different quantities yielding different effective potentials. To illustrate why this is, consider that in mean-field theory the Hamiltonian matrix elements (μHFμ) must be evaluated for some basis set (χμ,r) (assumed to be real for simplicity). The KS and HF exchange potential matrix elements are, respectively

\[
\langle \mu_{k}^{KS} | \nu \rangle = \int d^{3} r \mu_{k}(r) \chi_{\nu}(r) \frac{\partial \nu_k[n]}{\partial n(r)}
\]

(36)

\[
\langle \mu_{k}^{HF} | \nu \rangle = -\frac{1}{2} \int d^{3} r d^{3} r' \mu_{k}(r) \chi_{\nu}(r') \frac{n_1(r, r')}{|r - r'|}
\]

(37)

These two forms of the exchange matrix elements are distinct. Therefore, the HF and KS exchange energies, potentials, and densities are different, though this difference tends to be small (about 0.03–0.04% of the total exchange energy for isolated atoms). Because exact computation of \( \frac{\partial \nu_k[n]}{\partial n(r)} \) is complicated and computationally expensive, most modern hybrid DFT calculations (including those performed in this work) use the Generalized Kohn–Sham (GKS) scheme, in which the HF matrix elements are used for the exact exchange potential instead of the pure KS matrix elements. Because the difference between HF and KS exchange is small and because evaluating \( \frac{\partial \nu_k[n]}{\partial n(r)} \) is difficult, it is reasonable to compare CIDER exchange to HF exchange. A more detailed explanation of exchange functionals in the GKS scheme is provided in Supporting Information section S5 using the Levy–Lieb constrained search formalism.

As shown in Table 1, CIDER-X-AHW has an RMSE of 12 kcal/mol, which is fully explained by the RMSE of the CIDER-X-AHW predictions for static densities of 14 kcal/mol (Figure 1). It is notable that the SCF atomization energy errors slightly improve on the atomization exchange energy errors for static densities, even though the functional was only trained on static densities. This suggests that the CIDER exchange potential is sensible.

The 14 kcal/mol error of the CIDER-X-AHW functional with respect to exact exchange is small compared to that of the semilocal exchange functionals investigated here (Figure 1) but large compared to the desired chemical accuracy of 1 kcal/mol. However, hybrid functionals often use a small fraction of exact exchange, e.g., 20% for B3LYP. Because of this, B3LYP-CIDER reproduces B3LYP with an RMSE of only 2.3 kcal/mol on the test set (Table 1). While greater than 1 kcal/mol, this error is promisingly small considering that many of the atomization energies in the Jensen database involve breaking several chemical bonds at once.

To test the accuracy of the CIDER functional for both properties and systems significantly different than those contained in the training set, B3LYP and B3LYP-CIDER were used to compute the barriers heights of the BH76 database, the ionization potentials of the IP23 database, and the transition-metal bond energies of the TMBE33 database. These databases are subsets of the Minnesota 2015 Database. The MAEs for the transition-metal bond energies in Figure 2 are divided by the average number of bonds broken per data point, which matches the presentation in the original work.

As shown in Figure 2, B3LYP-CIDER exacerbates the systematic underestimation of barrier heights of B3LYP, leading to an increase in MAE from 4.1 to 6.8 kcal/mol compared to the reference values. The MAE between B3LYP-CIDER and B3LYP is 2.8 kcal/mol. The deviation from B3LYP could potentially be improved by including the density distributions of transition states in the CIDER training set, as currently all training set systems are isolated atoms or molecules at equilibrium geometry.

For ionization potentials, B3LYP-CIDER performs well; the MAE compared to B3LYP is 1.7 kcal/mol, and the error compared to reference values is worse by only 0.5 kcal/mol. Most of the error compared to B3LYP arises from the transition-metal systems. For the main-group IP13-MG data set, B3LYP-CIDER reproduces B3LYP to chemical accuracy (0.8 kcal/mol), while the functionals deviate by 3.0 kcal/mol for the transition-metal IP10-TM subset. The chemically accurate reproduction of main-group IPs is notable because B3LYP-CIDER was not trained on any ionic systems. It might be that the uniform scaling factor of CIDER is not effective in ionic systems.

For transition-metal bond energies, B3LYP-CIDER has an MAE of 8.7 kcal/mol relative to B3LYP. However, B3LYP-CIDER is more accurate than B3LYP compared to the reference values; the error of the CIDER functional seems to cancel some systematic error of B3LYP for metal bond energies. It is known to be challenging to describe transition-metal bonds with hybrid exchange...
functionals because the semilocal functional’s cancellation of error is lost. While ideally the CIDER functional would describe the exchange energy accurately, it is reassuring to see that it behaves like a semilocal functional when its accuracy breaks down, leading to sensible and explainable behavior.

Of note, the IP10-TM and TMBE33 data sets include ions, bonded transition-metal systems, and fifth-period elements treated with effective core potentials, none of which are present in the training set for CIDER. While CIDER is clearly a less faithful reproduction of HF for these systems than for systems similar to the training set, the resulting errors compared to reference values are similar. In addition, other than the isolated Fe atom and Pd’ ion, which had to be treated with level shifting and a high damping factor, calculations using the CIDER exchange functional converged successfully, as shown in Table 2. This suggests that the structure of the CIDER functional encourages transferability and stability.

Table 2. Threshold to Which B3LYP-CIDER SCF Calculations Converged (in Ha) for the 451 Systems Involved in This Study

| convergence threshold | no. of systems |
|-----------------------|---------------|
| 10^-8                 | 425           |
| 10^-7                 | 18            |
| 10^-6                 | 3             |
| unconverged           | 2             |
| total                 | 451           |

“The two “unconverged” calculations were the Pd’ ion and Fe atom, which would only converge to 10^-7 Ha when level shifting was applied.

Lastly, it is worth noting that the CIDER-X-AHW functional was not trained to any total energies, only exchange energy densities. It is possible that retraining the coefficients of hybrid functionals specifically for use with CIDER could significantly improve their accuracy, especially when the functional includes a higher fraction of exact exchange or already has a systematic error for some properties or systems (like B3LYP for barrier heights). An additional consequence of training to the exchange energy density is that it is nontrivial to extend the methodology presented here to the correlation functional. This is because the correlation energy density would be much harder to compute and more ambiguously defined than the exchange energy density. However, this is a limitation of the current Gaussian process model and training procedure, not of the CIDER features themselves. To train a correlation functional, one could either develop an approach to train to total correlation energies with a Gaussian process or use a model for which it is easier to train to total energies, such as a neural network.

5.3. Comment on Computational Cost. The cost of evaluating the ML model is insignificant because it is mapped to a cubic spline, and the computational bottleneck is the evaluation of the features. Because this initial CIDER model has a large feature set, uses a relatively dense integration grid, and lacks matrix element screening for scalability, feature evaluation is slow for practical applications. However, all of the nonlocal features used in the model are orbital independent and have a finite length scale, suggesting that linear scaling and efficient implementations are possible. The challenge of optimizing CIDER functionals to have near-semilocal DFT cost will be the subject of future work. As a first step toward improving the computational efficiency, the Supporting Information (section S3) introduces a prospective algorithm for evaluating CIDER functionals in a linear-scaling fashion. This algorithm will be implemented in a future work.

6. CONCLUSION

In this work, we presented the CIDER formalism, which is based on a set of nonlocal features to describe the density distribution in a scale-invariant manner. This feature set was used to train a Gaussian process regression model to accurately describe the Kohn–Sham exchange functional $E_{\alpha}[n]$ to a level of precision previously only attained by exact evaluation of the functional. The CIDER functional can replace a small fraction of HF exchange in hybrid functionals to accurately reproduce atomization energies, and it has excellent numerical stability, which has previously been a challenge for ML functionals.

As it stands, CIDER demonstrates that smooth, numerically stable exchange functionals can be learned that satisfy known exact constraints, accurately reproduce the target energy, and can be applied across a broad range of the periodic table. This demonstration is a first step toward a functional that could help bridge the gap between efficient semilocal functionals and accurate hybrid functionals as well as provide a groundwork for developing XC functionals with posthybrid DFT accuracy.

APPENDIX A: RELATIONSHIP BETWEEN CIDER LENGTH SCALE AND SLATER ORBITAL DENSITY

Consider eq 13 in the case that $B_0 = C_0 = \frac{6}{5\pi}(6\pi^2)^{1/3}$. Then, noting that $\tau_0 = \frac{3}{10}(3\pi^2)^{1/3}n^{5/3}$, eq 13 becomes

$$a[n](r) = \alpha \left(\frac{n}{\tau_0} \right)^{2/3} C_0 \tau \tau_0 (38)$$

$$= \frac{4\tau}{n} (39)$$

Now suppose that the density distribution is a spin-unpolarized, two-electron system, in which case $\tau = \frac{\sqrt{\pi} \alpha}{8n}$. Then

$$a[n](r) = \frac{1}{2} \left(\frac{\sqrt{n}}{n} \right)^2 (40)$$

If the electron pair occupies a Slater-type orbital then

$$n(x) = \frac{\alpha^3}{4\pi} e^{-\alpha r} (41)$$

for some $\alpha$. This orbital has $|\nabla n|/n = \alpha$, so

$$a[n](r) = \frac{\alpha^2}{2} (42)$$

Then, it holds from eqs 41 and 42 that $n(x) \propto e^{-(\sqrt{2\alpha} r)}$, as asserted in eq 14.

APPENDIX B: CONTRACTING TWO $L = 1$ FEATURES AND ONE $L = 2$ FEATURE INTO AN $L = 0$ FEATURE

Take tensor features $a$, $b$, and $c$, with $l = 1$, 2, and 1, respectively, defined with real spherical harmonics ($x$, $y$, $z$ for $l = 1$ and $x^2$, $y^2$, $z^2$, $x^2y^2$ for $l = 2$). We can convert these features to and from the space of complex spherical harmonics using the following rules, with $Y_{\ell m}$ for the real spherical harmonics and $Y^{*}_{\ell m}$ the complex ones.
evaluation easy and brings the computational cost of evaluation down to \(O(1)\) per test point.

\section*{APPENDIX D: FUNCTIONAL DERIVATIVES FOR CIDER NONLOCAL FEATURES}

The Gaussian process is a function of a set of features

\[
e_{c}[n](r) = -\frac{3}{4} \frac{1}{\pi} \left( n(r) \right)^{4/3} \mathcal{E}_c(\mathbf{x}[n](r))
\]

with the total exchange energy being

\[
E_{\text{ex}}[n] = \int d^{3}r e_{\text{ex}}[n](r)
\]

For the remainder of this section, we denote terms like \(\mathbf{x}[n]\) as \(\mathbf{x}\) for brevity. Calculating the exchange potential requires functional derivatives with respect to \(n(r)\), \(\partial_{\alpha} n(r)\) for \(\alpha = x, y, z\), and \(r(r)\). Existing routines in PySCF\textsuperscript{60} and other DFT codes can compute the Generalized Kohn–Sham potential from the functional derivatives with respect to these quantities. These functional derivatives can be written as

\[
\frac{\delta E_{\text{ex}}}{\delta n(r)} = -\frac{3}{4} \frac{1}{\pi} n(r)^{4/3} \mathcal{E}_c(\mathbf{x}(r))
\]

\[
- \frac{3}{4} \frac{1}{\pi} \int d^{3}r n(r)^{4/3} \left( \sum_{i} \frac{\partial \mathcal{E}_c}{\partial \mathcal{X}_i} \frac{\delta \mathcal{X}_i(r)}{\delta n(r)} \right)
\]

\[
\frac{\delta E_{\text{ex}}}{\delta \partial_{\alpha} n(r)} = -\frac{3}{4} \frac{1}{\pi} \int d^{3}r n(r)^{4/3} \left( \sum_{i} \frac{\partial \mathcal{E}_c}{\partial \mathcal{X}_i} \frac{\delta \mathcal{X}_i(r)}{\delta \partial_{\alpha} n(r)} \right)
\]

The terms \(\frac{\delta \mathcal{E}_c}{\delta \partial_{\alpha} n(r)}\) are provided by the Gaussian process or cubic spline. If the index \(i\) corresponds to a semilocal descriptor then

\[
\left. \frac{\delta \mathcal{X}_i(r)}{\delta n(r)} \right|_{n(r)} = \delta(r - r')
\]

and the integral over \(r'\) reduces to evaluating the derivatives at \(r\).

For the CIDER model, the \(G_{\text{nlm}}\) descriptors (eq 11) only have nonlocal dependence on the density \(n(r)\); the dependence on \(r(r)\) is local, and \(G_{\text{nlm}}\) does not depend on \(\partial_{\alpha} n(r)\). Therefore, eq 59 applies with \(r(r)\) in place of \(n(r)\), even if \(x_i\) is nonlocal. The kinetic term \(\frac{\partial \mathcal{E}_c}{\partial \partial_{\alpha} n(r)}\), needed to evaluate eq 58, arises solely from the derivative of the exponent

\[
\frac{\partial G_{\text{nlm}}}{\partial \partial_{\alpha} n(r)} = \left. \frac{1}{2a} G_{\text{nlm}}(r) - H_{\text{nlm}}(r) \right|_{r(r)}
\]
\[ \frac{\partial}{\partial r} = C_0 \left( \frac{n}{2} \right)^{2/3} \frac{1}{\tau_0} \]  

\[ H_{nlm}(r) = \int d^3r' \, r' \cdot r_l^2 g_{nlm}(r' - r, r)n(r') \]  

The density derivatives are similar but with an additional nonlocal term \( g_{nlm}(r - r'; r') \)

\[ \frac{\partial G_{nlm}(r'}{\partial n} \bigg|_{n(r)} = \frac{\partial G_{nlm}}{\partial n} \bigg|_{n(r)} + g_{nlm}(r - r'; r') \]  

\[ \frac{\partial G_{nlm}}{\partial n} \bigg|_{n(r)} = \frac{\partial G_{nlm}}{\partial n} \bigg|_{n(r)} + \frac{\partial}{\partial n} \]  

\[ \frac{\partial}{\partial n} = \rho \left( \frac{1}{4\pi} \right)^{1/3} \frac{2}{3} \left( B_0 - C_0 \right) \frac{r}{\tau_0} \]  

The second term on the right-hand side of eq 64 introduces a nonlocal term \( v_{nlm}(r) \) in eq 56 of the form

\[ v_{nlm}(r) = \int d^3r' f_{nlm}(r') g_{nlm}(r - r'; r') \]  

\[ f_{nlm}(r) = -\frac{3}{4 \pi} \left( \frac{3}{4} \right)^{1/3} n(r)^{1/3} \frac{\partial}{\partial G_{nlm}} \bigg|_{n(r)} \]  

To obtain the Generalized Kohn–Sham potential, one must compute the matrix elements for an atomic orbital basis \( \{ \mu \} \) (such that \( \langle \mu | r \rangle = \chi_{\mu}(r) \))

\[ v_{\mu} = \int d^3r \chi_{\mu}(r) \frac{\delta E}{\delta n(r)} \chi_{\mu}(r) \]  

\[ + \sum_{a=x,y,z} \int d^3r \chi_{\mu}(r) \frac{\delta E}{\delta \partial_n(r)} \delta \chi_{\mu}(r) \]  

\[ + \sum_{a=x,y,z} \int d^3r \chi_{\mu}(r) \frac{\delta E}{\delta \partial_{\delta_n}(r)} \partial \chi_{\mu}(r) \]  

\[ + \frac{1}{2} \int d^3r \, (\nabla \chi_{\mu}(r) \cdot \nabla \chi_{\mu}(r)) \frac{\delta E}{\delta n(r)} \]  

The above equation assumes real orbitals. Typically, the integrals over \( r \) are evaluated numerically on a real-space grid. Because their contributions are local, all terms except for eq 67 can be evaluated with this standard numerical integration approach. Equation 67 gives rise to a double integration in eq 69

\[ v_{\mu}^{nlm} = \int d^3r \chi_{\mu}(r) \chi_{\mu}(r) \]  

\[ \times \int d^3r' f_{nlm}(r') g_{nlm}(r - r'; r') \]  

This analysis leaves three nonlocal terms that must be evaluated at each iteration: \( G_{nlm}(r) \) (eq 11), \( H_{nlm}(r) \) (eq 63), and \( v_{\mu}^{nlm} \) (eq 70). This is done using a density fitting (DF) auxiliary basis \( \{ \Theta_p(r) \} \)

\[ \chi_{\mu}(r) \chi_{\mu}(r) = \sum_p C_p^{\mu} \Theta_p(r) \]  

Then, the nonlocal terms can be evaluated in a computationally efficient manner

\[ G_{nlm}(r) = \sum_p n_p (g_{nlm}(r)|\Theta_p) \]  

\[ H_{nlm}(r) = \sum_p n_p (g_{nlm}(r)|\Theta_p) \]  

\[ v_{\mu}^{nlm} = \sum_p C_p^{\mu} v_{\mu}^{nlm} \]  

\[ v_{\mu}^{nlm} = \int d^3r f_{nlm}(r) (g_{nlm}(r)|\Theta_p) \]  

The integral over \( r \) in the last equation is evaluated numerically; all bracket notation terms are evaluated analytically. The density \( n_p \) in the auxiliary basis is computed from the atomic orbital density matrix \( P_{\mu\nu} \) which in turn is evaluated from the KS orbitals \( \phi_i \) and occupations \( f_i \)

\[ n_p = \sum_{\mu} C_{\mu}^{P_{\mu\nu}} \]  

\[ P_{\mu\nu} = \sum_i f_i c_{\mu\nu} c_i \]  

\[ \phi(r) = \sum_{\mu} c_{\mu} \chi_{\mu}(r) \]  

The above equations assume real coefficients \( c_{\mu} \).

### ASSOCIATED CONTENT

#### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.1c00904.

Training and validation detail for the CIDER functionals; additional benchmarking details for the Gaussian process models; linear-scaling algorithm for CIDER in Gaussian-type orbital basis sets; discussion of different possible choices for the exchange energy density; discussion of the difference between KS and HF exchange energies (PDF)

### AUTHOR INFORMATION

**Corresponding Authors**

Kyle Bystrom — John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, United States; orcid.org/0000-0003-1342-4972; Email: kylebystrom@g.harvard.edu

Boris Kozinsky — John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, United States; orcid.org/0000-0002-0638-539X; Email: bkoz@seas.harvard.edu

Complete contact information is available at: https://pubs.acs.org/doi/10.1021/acs.jctc.1c00904

**Notes**

The authors declare no competing financial interest.

### ACKNOWLEDGMENTS

The authors thank Prof. Donald Truhlar for helpful discussions regarding use of the Minnesota 2015 Database. This work was supported by the US Department of Defense MURI under Award No. N00014-20-1-2418 and by the STC Center for Integrated Quantum Materials, NSF Grant No. DMR-1231319. K.B. was supported by the Department of Defense (DoD)
through the National Defense Science & Engineering Graduate (NDSEG) Fellowship Program.

REFERENCES

(1) Hohenberg, P.; Kohn, W. Inhomogeneous electron gas. Phys. Rev. 1964, 136, B864–B871.
(2) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. Phys. Rev. 1965, 140, A1133–A1138.
(3) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals. Mol. Phys. 2017, 115, 2315–2372.
(4) Verma, P.; Truhlar, D. G. Status and Challenges of Density Functional Theory. Trends Chem. 2020, 2, 302–318.
(5) Su, N. Q.; Xu, X. Development of New Density Functional Approximations. Annu. Rev. Phys. Chem. 2017, 68, 155–182.
(6) Yu, H. S.; Li, S. L.; Truhlar, D. G. Perspective: Kohn-Sham density functional theory descending a staircase. J. Chem. Phys. 2016, 145, 130901.
(7) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for density functional theory. Chem. Rev. 2012, 112, 289–320.
(8) Kalita, B.; Li, L.; McCarty, R. J.; Burke, K. Learning to Approximate Density Functionals. Acc. Chem. Res. 2021, 54, 818–826.
(9) Tozer, D. J.; Ingamells, V. E.; Handy, N. C. Exchange-correlation potentials. J. Chem. Phys. 1996, 105, 9200–9213.
(10) Snyder, J. C.; Rupp, M.; Hansen, K.; Müller, K. R.; Burke, K. Finding density functionals with machine learning. Phys. Rev. Lett. 2012, 108, 253002.
(11) Snyder, J. C.; Rupp, M.; Hansen, K.; Bloostom, L.; Müller, K.-R.; Burke, K. Orbital-free bond breaking via machine learning. J. Chem. Phys. 2013, 139, 224104.
(12) Li, L.; Snyder, J. C.; Pelaschier, I. M.; Huang, J.; Niranjan, U.-N.; Duncan, P.; Rupp, M.; Müller, K.-R.; Burke, K. Understanding machine-learned density functionals. Int. J. Quantum Chem. 2016, 116, 819–833.
(13) Dick, S.; Fernandez-Serra, M. Machine learning accurate exchange and correlation functionals of the electronic density. Nat. Commun. 2020, 11, 3509.
(14) Chen, Y.; Zhang, L.; Wang, H.; Weinan, E. DeePKS: A Comprehensive Data-Driven Approach toward Chemically Accurate Density Functional Theory. J. Chem. Theory Comput. 2021, 17, 170–181.
(15) Margraf, J. T.; Reuter, K. Pure non-local machine-learned density functional theory for electron correlation. Nat. Commun. 2021, 12, 344.
(16) Lei, X.; Medford, A. J. Design and analysis of machine learning exchange-correlation functionals via rotationally invariant convolutional descriptors. Phys. Rev. Mater. 2019, 3, 63801.
(17) Nagai, R.; Akashi, R.; Sugino, O. Completing density functional theory by machine learning hidden messages from molecules. npj Comput. Mater. 2020, 6, 43.
(18) Zhao, Y.; Truhlar, D. G. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. J. Chem. Phys. 2006, 125, 194101.
(19) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other function. Theor. Chem. Acc. 2008, 120, 215–241.
(20) 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, 062021.
(21) Medvedev, M. G.; Bucharinov, I. S.; Sun, J.; Perdew, J. P.; Lyssenko, K. A. Density functional theory is straying from the path toward the exact functional. Science 2017, 355, 49–52.
(22) Hollingsworth, J.; Baker, T. E.; Burke, K. Can exact conditions improve machine-learned density functionals? J. Chem. Phys. 2018, 148, 241743.
(45) Janesko, B. G. Rung 3.5 density functionals. J. Chem. Phys. 2010, 133, 104103.
(46) Janesko, B. G. Rung 3.5 density functionals: Another step on Jacob’s ladder. Int. J. Quantum Chem. 2013, 113, 83–88.
(47) Janesko, B. G.; Proynov, E.; Scalmani, G.; Frisch, M. J. Long-range-corrected Rung 3.5 density functional approximations. J. Chem. Phys. 2018, 148, 104112.
(48) Thomas, N.; Smidt, T.; Kearnes, S.; Yang, L.; Li, L.; Kohlhoff, K.; Riley, P. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. ArXiv abs/1802.08219 2018.
(49) Shapere, A. V. Moment tensor potentials: A class of systematically improvable interatomic potentials. Multiscale Model. Simul. 2016, 14, 1153–1173.
(50) Becke, A. D. Density-functional thermochemistry. V. Systematic optimization of exchange-correlation functionals. J. Chem. Phys. 1997, 107, 8554–8560.
(51) Becke, A. D.; Edgecombe, K. E. A simple measure of electron localization in atomic and molecular systems. J. Chem. Phys. 1990, 92, 5397–5403.
(52) Mardirossian, N.; Head-Gordon, M. Mapping the genome of meta-generalized gradient approximation density functionals: The search for B97M-V. J. Chem. Phys. 2015, 142, 074111.
(53) Rasmussen, C. E.; Williams, C. K. I. Gaussian Processes for Machine Learning; The MIT Press, 2005.
(54) Glielmo, A.; Zeni, C.; De Vita, A. Efficient nonparametric n-body force fields from machine learning. Phys. Rev. B 2018, 97, 184307.
(55) Xie, Y.; Vandermause, J.; Sun, L.; Cepellotti, A.; Kozinsky, B. Bayesian force fields from active learning for simulation of interdimensional transformation of stanene. npj Comput. Mater. 2021, 7, 40.
(56) Vandermause, J.; Torrisi, S. B.; Batzner, S.; Xie, Y.; Sun, L.; Kolpak, A. M.; Kozinsky, B. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. npj Comput. Mater. 2020, 6, 120.
(57) Vandermause, J.; Xie, Y.; Lim, J. S.; Owen, C. J.; Kozinsky, B. Active learning of reactive Bayesian force fields: Application to heterogeneous hydrogen-platinum catalysts. arXiv 2106.04481 2021 DOI: 10.21203/rs.3.rs.1178160/v1.
(58) Duvenaud, D.; Nickisch, H.; Rasmussen, C. E. Additive Gaussian processes. Advances in Neural Information Processing Systems 24 (NIPS 2011), 2011.
(59) Chachiyo, T.; Chachiyo, H. Simple and accurate exchange energy for density functional theory. Molecules 2020, 25, 3485.
(60) 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.; Wouters, S.; Chan, G. K. L. PySCF: the Python-based simulations of chemistry framework. WIREs Comput. Mol. Sci. 2018, 8, e1340.
(61) CiderPress Github Repository; https://github.com/mir-group/CiderPress.
(62) Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. Phys. Chem. Chem. Phys. 2006, 8, 1057.
(63) Sun, Q. Libcint: An efficient general integral library for Gaussian basis functions. J. Comput. Chem. 2015, 36, 1664–1671.
(64) Jain, A.; Ong, S. P.; Chen, W.; Medasani, B.; Qu, X.; Kocher, M.; Brafull, M.; Petretto, G.; Rignanese, G. M.; Hautier, G.; Gunter, D.; Persson, K. A. Fireworks: A dynamic workflow system designed for highthroughput applications. Concurre. Comput. 2015, 27, 5037–5059.
(65) Jensen, S. R.; Saha, S.; Flores-livas, A.; Huhn, W.; Blum, V.; Goedecker, S.; Frediani, L. GGA-PBE and hybrid-PBE0 energies and dipole moments with MRChem, NWChem and ELK. DataverseNO 2017, 1–8.
(66) Schmider, H. L.; Becke, A. D. Optimized density functionals from the extended G2 test set. J. Chem. Phys. 1998, 108, 9624–9631.
(67) Jensen, S. R.; Saha, S.; Flores-Livas, J. A.; Huhn, W.; Blum, V.; Goedecker, S.; Frediani, L. The Elephant in the Room of Density Functional Theory Calculations. J. Phys. Chem. Lett. 2017, 8, 1449–1457.
(68) Perdew, J.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. Phys. Rev. Lett. 1996, 77, 3865–3868.
(69) Pedregosa, F.; et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 2011, 12, 2825–2830.
(70) interpolation.py Python Package; https://github.com/EconForge/interpolation.py.
(71) Yu, H. S.; He, X.; Li, S. L.; Truhlar, D. G. MN15: A Kohn-Sham global-hybrid exchange-correlation density functional with broad accuracy for multi-reference and single-reference systems and noncovalent interactions. Chem. Sci. 2016, 7, 5032–5051.
(72) Morgante, P.; Peverati, R. Statistically representative databases for density functional theory: Via data science. Phys. Chem. Chem. Phys. 2019, 21, 19092–19103.
(73) Weigend, F.; Furche, F.; Ahlrichs, R. Gaussian basis sets of quadruple zeta valence quality for atoms H-Kr. J. Chem. Phys. 2005, 119, 12753–12762.
(74) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. Phys. Chem. Chem. Phys. 2005, 7, 3297.
(75) Rapporport, D.; Furche, F. Property-optimized Gaussian basis sets for molecular response calculations. J. Chem. Phys. 2010, 133, 134105.
(76) Andrae, D.; Häubermann, U.; Dolg, M.; Stoll, H.; Preub, H. Energy-adjusted ab initio pseudopotentials for the second and third row transition elements. Theor. Chem. Acta 1990, 77, 123–141.
(77) Wang, Y.; Verma, P.; Jin, X.; Truhlar, D. G.; He, X. Revised M06 density functional for main-group and transition-metal chemistry. Proc. Natl. Acad. Sci. U. S. A. 2018, 115, 10257–10262.
(78) Sun, J.; Ruzsinszky, A.; Perdew, J. Strongly Constrained and Appropriately Normed Semilocal Density Functional. Phys. Rev. Lett. 2015, 115, 036402.
(79) Yu, H. S.; He, X.; Truhlar, D. G. MN15-L: A New Local Exchange-Correlation Functional for Kohn-Sham Density Functional Theory with Broad Accuracy for Atoms, Molecules, and Solids. J. Chem. Theory Comput. 2016, 12, 1280–1293.
(80) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. Phys. Rev. A 1988, 38, 3098–3100.
(81) Becke, A. D.; Rousell, M. R. Exchange holes in inhomogeneous systems: A coordinate-space model. Phys. Rev. A 1989, 39, 3761–3767.
(82) Proynov, E.; Gan, Z.; Kong, J. Analytical representation of the Becke-Roussel exchange functional. Chem. Phys. Lett. 2008, 455, 103–109.
(83) Tao, J.; Mo, Y. Accurate Semilocal Density Functional for Condensed-Matter Physics and Quantum Chemistry. Phys. Rev. Lett. 2016, 117, 073001.
(84) Wellendorf, 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.
(85) Kümmel, S.; Kronik, L. Orbital-dependent density functionals: Theory and applications. Rev. Mod. Phys. 2008, 80, 3–60.
(86) Seidl, A.; Görling, A.; Vogl, P.; Majewski, J. A.; Levy, M. Generalized Kohn-Sham schemes and the band-gap problem. Phys. Rev. B 1996, 53, 3764–3774.
(87) Görling, A.; Levy, M. Hybrid schemes combining the Hartree-Fock method and density-functional theory: Underlying formalism and properties of correlation functionals. J. Chem. Phys. 1997, 106, 2675–2680.
(88) Levy, M. Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the v-representability problem. Proc. Natl. Acad. Sci. U. S. A. 1979, 76, 6062–6065.
(89) Levy, M. Electron densities in search of Hamiltonians. Phys. Rev. A 1982, 26, 1200–1208.
(90) Lieb, E. H. Density functionals for coulomb systems. Int. J. Quantum Chem. 1983, 24, 243–277.