

Nonlocal machine-learned exchange functional for molecules and solids

Kyle Bystrom^{1,*} and Boris Kozinsky^{1,2,†}¹*Harvard John A. Paulson School of Engineering and Applied Sciences, Cambridge, Massachusetts 02138, USA*²*Robert Bosch LLC Research and Technology Center, Cambridge, Massachusetts 02139, USA*

(Received 27 July 2023; revised 18 June 2024; accepted 12 July 2024; published 15 August 2024)

The design of better exchange-correlation functionals for density functional theory (DFT) is a central challenge of modern electronic structure theory. However, current developments are limited by the mathematical form of the functional, with efficient semilocal functionals being inaccurate for many technologically important systems and the more accurate hybrid functionals being too expensive for large solid-state systems because of the use of the exact exchange operator. In this paper, we use machine learning combined with exact physical constraints to design an exchange functional that is both orbital dependent and nonlocal, but which can be evaluated at roughly the cost of semilocal functionals and is significantly faster than hybrid DFT in plane-wave codes. By training functionals with several different feature sets, we elucidate the roles of orbital-dependent and nonlocal features in learning the exchange energy and determine that both types of features provide vital and independently important information to the model. Having trained our exchange functional with an expressive, nonlocal feature set, we substitute it into existing hybrid functionals to achieve hybrid-DFT accuracy on thermochemical benchmark sets and improve the accuracy of band gap predictions over semilocal DFT. To demonstrate the scalability of our approach as well as the practical benefits of improved band gap prediction, we compute charged defect transition levels in silicon using large supercells. Because of its transferability and computational efficiency for both molecular and extended systems, our model overcomes the cost-accuracy trade-off between semilocal and hybrid DFT, and our general approach provides a feasible path toward a universal exchange-correlation functional with post-hybrid DFT accuracy and semilocal DFT cost.

DOI: [10.1103/PhysRevB.110.075130](https://doi.org/10.1103/PhysRevB.110.075130)

I. INTRODUCTION

One of the central challenges of density functional theory (DFT) research is to design accurate approximations to the exchange-correlation (XC) functional [1]. As the one unknown term in Kohn-Sham DFT [2], errors in the XC functional limit the predictive power of DFT for a wide variety of technological problems, including catalysis [3], battery materials [4], and semiconductor physics [5,6]. Increasing the complexity of the XC functional can improve its accuracy, but often at drastically increased computational cost, which can be prohibitive for large condensed matter simulations. In this paper, we use machine learning (ML) to overcome this cost-accuracy trade-off by training a functional that obeys exact physical constraints and is accurate, transferable across a broad range of chemistries, and scalable to hundreds of atoms.

The key cost-accuracy trade-off we address is that between semilocal and hybrid XC functionals. Semilocal functionals, which include the local density approximation (LDA), generalized gradient approximation (GGA), and meta-generalized gradient approximation (meta-GGA) [7], are computationally efficient but insufficiently accurate for many applications. For example, LDA and GGA functionals are incapable of

correctly predicting band gaps because of their lack of derivative discontinuity [8,9], and they are of limited accuracy for many molecular systems [1,10]. Meta-GGAs introduce derivative discontinuities into the XC functional and can therefore improve band gap predictions, but they typically account for only 20–50% of the GGA band gap error [11]. Meta-GGAs can be designed to further improve band gap prediction [12,13], but modifying a given meta-GGA form to predict larger, more accurate band gaps typically reduces the accuracy of that functional for other chemical and material properties, such as bond dissociation energies [13,14]. On the other hand, hybrid functionals, which mix a fraction of the nonlocal, orbital-dependent exact exchange energy into an XC functional [15], provide significantly improved predictions of band gaps [11,16] and thermochemical properties [1,10] over semilocal DFT, but this accuracy comes at a steep additional computational cost, especially within plane-wave DFT. In spite of the remarkable recent progress of efficient implementations for solids [17–22], hybrid DFT is not as efficient or practical as semilocal DFT for these calculations.

Several approaches to design improved XC functionals have been pursued in the literature. Many popular functionals, like PBE [23] and SCAN [24], are designed primarily or entirely based on mathematical constraints obeyed by the exact functional. Constraint-based functionals are often considered trustworthy because they can predict molecular properties without being explicitly tuned to match them, reducing the

*Contact author: kylebystrom@g.harvard.edu†Contact author: bkoz@g.harvard.edu

risk of bias toward a particular property or set of systems [25,26]. However, some of the most popular and accurate XC functionals for main-group chemistry are empirically fit to chemical data [27,28]. Recently, significant progress has been made toward combining constraint and data-based design approaches using machine learning [29–35], culminating in the development of the DM21 functional [36], which achieves state-of-the-art accuracy on main-group chemistry benchmarks and also improves the description of static correlation by fitting fractional charge constraints.

While promising, these developments have a few key shortcomings, the most notable of which is their computational cost. DM21 uses a local range-separated hybrid formalism, which is both computationally expensive (even more so than global hybrid DFT) and not widely available for the plane-wave DFT codes typically used for investigating large extended systems. As a result, DM21 has not been applied to solids. To the best of our knowledge, no nonlocal machine-learned functional has yet been implemented for both molecular (Gaussian-type orbital) and extended (plane-wave) DFT calculations, with the exception of the DeePKS approach [37,38]. However, DeePKS uses geometry-dependent (rather than universal) functionals and is trained for specific systems, so it does not fulfill the role of a broadly chemically transferable model. All other ML XC functionals that have been applied to solids have been limited to the semilocal form [13,32,33,39–41] (with the exception of minimally parameterized van der Waals functionals and one paper by Riemelmoser *et al.* [42] that introduced a nonlocal functional for C, H, and O-containing systems). This limitation restricts the application space of ML functionals by preventing their use in applications where hybrid DFT accuracy is needed. For example, hybrid DFT has proven useful for studying problems in catalysis within both Gaussian-type orbital and plane-wave codes [43,44] and for addressing the “band-gap problem” in DFT [45]. In addition to this technical limitation, applications of ML functionals to complex and heterogeneous systems are limited by their lack of transferability and universality, as they have previously only been trained on molecular systems.

These limitations raise the question of whether machine learning models can leverage more efficient features to balance accuracy and computational cost. One possible approach in this direction is to generate nonlocal features by taking integrals of the density over some set of kernel functions [30,46]. This way, an ML model of the XC functional can capture nonlocal information about the density without evaluating the computationally demanding exact exchange operator required for hybrid DFT. In line with this strategy, we recently introduced the “compressed scale-invariant density representation” (CIDER) formalism for learning the exchange functional [35], in which the nonlocal features were designed to enforce the uniform scaling rule [47] on the learned exchange functional. We demonstrated that the exchange functional can be accurately learned with the CIDER approach. However, the high computational cost of computing the features, lack of implementation for plane-wave DFT, and need to compute and fit the exchange energy density (rather than the total exchange energy) limited the practicality of our model and restricted its application to small molecules.

In this paper, we overcome all of these limitations by revising the form of the features, implementing efficient algorithms to evaluate the features for both molecular and extended systems, and redesigning our Gaussian process model to fit the total exchange energy. These improvements enable the construction of the first nonlocal, universal ML exchange functional for solids, which can be applied to large systems with similar computational cost to semilocal DFT. We use this exchange functional to implement efficient and accurate hybrid DFT surrogates that can be used in a variety of applications. Because our functional form is both transferable and numerically stable, we believe that our approach will also be useful for learning the full XC functional, allowing us to address problems where DFT accuracy is limited by both exchange and correlation effects.

By training and testing exchange functionals with different feature sets, we find that to learn the exchange functional well enough to reproduce hybrid DFT accuracy, it is necessary to include both a semilocal orbital-dependent feature (i.e., the kinetic energy density) and nonlocal density-dependent features in the model input. Having developed this “nonlocal meta-GGA,” we show that it accurately predicts molecular properties and improves solid-state band gap predictions over standard meta-GGAs, which is a central problem in DFT. The performance for solid-state band gaps is particularly promising because only nine systems in the training set are solids, with the others being isolated molecules and complexes. We also encounter only a few convergence problems in our self-consistent field (SCF) calculations of 2462 molecular and 453 periodic systems, indicating that the functional has good numerical stability. To demonstrate the transferability and efficiency of the model, we simulate systems with over 500 atoms to predict charged defect transition levels in silicon.

The paper is organized as follows. Section II describes the features and models used in this paper. Section III presents our results and discusses the accuracy of the CIDER functionals as well as their computational efficiency within plane-wave DFT. Section IV contains our conclusion. Appendix A provides a detailed description of the computational methods used in this study, and Appendix B provides additional mathematical details for the plane-wave implementation of CIDER.

II. THEORY

In Kohn-Sham DFT [2], it is often convenient to separate the XC energy into exchange and correlation parts, i.e., $E_{xc}[n] = E_x[n] + E_c[n]$. The exchange energy $E_x[n]$ accounts for the antisymmetric nature of the fermionic wave function, while the correlation energy $E_c[n]$ accounts for additional many-body correlation effects between the electrons. The exchange energy is the main focus of this paper and can be trivially generalized from non-spin-polarized DFT to spin-polarized DFT as [48]

$$E_x[n_\uparrow, n_\downarrow] = \frac{1}{2}E_x[2n_\uparrow] + \frac{1}{2}E_x[2n_\downarrow], \quad (1)$$

so all formulas in the theory section are written for the non-spin-polarized case for simplicity.

The exact exchange energy can be expressed in terms of the occupied Kohn-Sham orbitals $\{\phi_i(\mathbf{r})\}$ as

$$E_x^{\text{exact}}[n] = - \sum_{ij} \int d^3\mathbf{r} \int d^3\mathbf{r}' \frac{\phi_i^*(\mathbf{r})\phi_j(\mathbf{r})\phi_i(\mathbf{r}')\phi_j^*(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \quad (2)$$

Evaluating Eq. (2) is computationally expensive. However, the improved accuracy of hybrid DFT arises from mixing a fraction of $E_x^{\text{exact}}[n]$ into an otherwise semilocal XC functional. These considerations explain the cost-accuracy trade-off between semilocal and hybrid DFT. Our goal is to fit Eq. (2) as accurately as possible with a computationally efficient surrogate model.

The remainder of this section sets up our machine learning-based approach to this problem. Specifically, Secs. II A–II C cover the design of the exchange functional form, nonlocal features, and Gaussian process models, and Secs. II D–II F cover the computationally efficient implementation of the nonlocal features. The feature implementation requires different algorithms for Gaussian-type orbital and plane-wave DFT codes owing to the use of different basis sets and integration grids within these codes.

A. Uniform scaling and semilocal exchange functionals

The simplest approximate exchange functional is the local density approximation (LDA), in which the exchange energy is written as an integral over a function of the density $n(\mathbf{r})$ at each point in space

$$E_x[n] = \int d^3\mathbf{r} e_x^{\text{LDA}}(n(\mathbf{r})), \quad (3)$$

where in the Kohn-Sham formalism [2], the density is expressed in terms of the occupied orbitals as

$$n(\mathbf{r}) = \sum_i |\phi_i(\mathbf{r})|^2. \quad (4)$$

Typically, $e_x^{\text{LDA}}(n)$ is defined as the exchange energy density of the uniform electron gas [49]

$$e_x^{\text{LDA}}(n) = -\frac{3}{4} \left(\frac{3}{\pi} \right)^{1/3} n^{4/3}. \quad (5)$$

In addition to making Eq. (3) exact for the uniform electron gas, Eq. (5) also satisfies an important property of the exchange functional known as uniform scaling [47,50], which states that for a scaled density distribution $n_\lambda(\mathbf{r}) = \lambda^3 n(\lambda\mathbf{r})$,

$$E_x[n_\lambda] = \lambda E_x[n], \quad (6)$$

where λ is a positive scalar.

More sophisticated functionals that obey Eq. (6) can be designed by writing the exchange energy density as the product of Eq. (5) and an exchange enhancement factor F_x ,

$$E_x[n] = \int d^3\mathbf{r} e_x^{\text{LDA}}(n(\mathbf{r})) F_x(\mathbf{x}(\mathbf{r})). \quad (7)$$

In the above equation, $\mathbf{x}(\mathbf{r})$ is a vector of scale-invariant features x_i , where scale invariance is defined as

$$x_i[n_\lambda](\mathbf{r}) = x_i[n](\lambda\mathbf{r}). \quad (8)$$

For example, a generalized gradient approximation (GGA) exchange functional is constructed by making F_x a function of the reduced gradient s of the density

$$s = \frac{|\nabla n|}{2(3\pi^2)^{1/3} n^{4/3}}. \quad (9)$$

Because s satisfies Eq. (8), the GGA form

$$E_x[n] = \int d^3\mathbf{r} e_x^{\text{LDA}}(n(\mathbf{r})) F_x(s(\mathbf{r})) \quad (10)$$

satisfies Eq. (6). Likewise, a meta-GGA can be constructed by making the exchange enhancement factor a function of two ingredients $F_x(s, \alpha)$, where

$$\alpha = \frac{\tau - \tau_W}{\tau_0} \quad (11)$$

is a scale-invariant quantity that depends on the kinetic energy density

$$\tau(\mathbf{r}) = \frac{1}{2} \sum_i |\nabla \phi_i(\mathbf{r})|^2, \quad (12)$$

and $\tau_W = \frac{|\nabla n|^2}{8n}$ and $\tau_0 = \frac{3}{10} (3\pi^2)^{2/3} n^{5/3}$ are the single-orbital and uniform electron gas kinetic energies, respectively.

B. Updated nonlocal features for CIDER

In our initial construction of the CIDER formalism [35], the exchange functional is defined as

$$E_x[n] = \int d^3\mathbf{r} e_x^{\text{LDA}}(n(\mathbf{r})) F_x(s(\mathbf{r}), \alpha(\mathbf{r}), \{G_{nlm}(\mathbf{r})\}). \quad (13)$$

In Eq. (13), the functions $\{G_{nlm}(\mathbf{r})\}$ are *nonlocal* features of the density defined by

$$G_{nlm}(\mathbf{r}) = \int d^3\mathbf{r}' g_{nlm}(\mathbf{r}'; \mathbf{r}) n(\mathbf{r} + \mathbf{r}'), \quad (14)$$

where $g_{nlm}(\mathbf{r}'; \mathbf{r})$ is the kernel function

$$g_{nlm}(\mathbf{r}'; \mathbf{r}) = M_l Y_{lm}(\hat{\mathbf{r}}')(a(\mathbf{r})|\mathbf{r}'|^2)^{n+1/2} e^{-a(\mathbf{r})|\mathbf{r}'|^2}, \quad (15)$$

with $M_l = B_0^{3/2} \sqrt{4\pi^{l-1}} \left(\frac{8\pi}{3}\right)^{l/2}$. The position-dependent exponent $a(\mathbf{r})$ is defined as a semilocal quantity

$$a[n](\mathbf{r}) = \pi \left(\frac{n(\mathbf{r})}{2} \right)^{2/3} \left[B_0 + C_0 \left(\frac{\tau(\mathbf{r})}{\tau_0(\mathbf{r})} - 1 \right) \right], \quad (16)$$

where B_0 and C_0 are adjustable constants that control how quickly the width of the Gaussian-type function in Eq. (15) decreases as the density and kinetic energy density increase. Equation (16) behaves quadratically under uniform scaling, i.e.,

$$a[n_\lambda](\mathbf{r}) = \lambda^2 a[n](\lambda\mathbf{r}). \quad (17)$$

This property ensures that $G_{nlm}(\mathbf{r})$ is scale invariant [Eq. (8)] and therefore that Eq. (13) satisfies the uniform scaling rule [Eq. (6)].

Using these scale-invariant nonlocal features, we were able to train an ML exchange functional that accurately reproduced hybrid DFT atomization energies [35], but there is room for improvement in making the features more physically intuitive and easier to implement with production-level performance.

In this paper, we simplify the features by removing the spherical harmonics and instead adding a second position-dependent exponent to the kernel function, resulting in the features

$$G_i(\mathbf{r}_1) = \left(\frac{B_i + B_0}{2}\right)^{3/2} \int d^3\mathbf{r}_2 \Phi(a(\mathbf{r}_2), b_i(\mathbf{r}_1), r_{12})n(\mathbf{r}_2), \quad (18)$$

$$\Phi(a, b, r) = \exp[-(a + b)r^2], \quad (19)$$

where $r_{12} = |\mathbf{r}_1 - \mathbf{r}_2|$. Multiple features G_1, G_2, \dots are generated by evaluating Eq. (18) with different choices of exponent $b_i(\mathbf{r})$, which in turn are tuned by constants B_i and C_i as discussed below. Because the position-dependent exponents increase as the density increases, the a exponent effectively damps the value of the integrand when \mathbf{r}_2 is near the atomic core, which we intend to limit the interaction between the valence region and core region. The prefactor $(\frac{B_i+B_0}{2})^{3/2}$ is chosen so that $G_i = 2$ for the non-spin-polarized uniform electron gas.

To retain the scale invariance of the features, any choice of $a[n](\mathbf{r})$ and $b_i[n](\mathbf{r})$ that satisfies Eq. (17) is sufficient. For example, the exponents can be an orbital-independent function of the density and its gradient

$$a[n](\mathbf{r}) = \pi \left(\frac{n}{2}\right)^{2/3} \left[B_0 + C_0 \left(\frac{|\nabla n|^2}{8n\tau_0} \right) \right], \quad (20)$$

$$b_i[n](\mathbf{r}) = \pi \left(\frac{n}{2}\right)^{2/3} \left[B_i + C_i \left(\frac{|\nabla n|^2}{8n\tau_0} \right) \right], \quad (21)$$

or can be made kinetic energy dependent, as in our previous paper,

$$a[n](\mathbf{r}) = \pi \left(\frac{n}{2}\right)^{2/3} \left[B_0 + C_0 \left(\frac{\tau}{\tau_0} - 1 \right) \right], \quad (22)$$

$$b_i[n](\mathbf{r}) = \pi \left(\frac{n}{2}\right)^{2/3} \left[B_i + C_i \left(\frac{\tau}{\tau_0} - 1 \right) \right]. \quad (23)$$

Equations (20) and (21) are used for the nonlocal GGA presented in this paper (c.f. Sec. II C), while Eqs. (22) and (23) are used for the nonlocal meta-GGA. Throughout Secs. II D–II F, the $(\frac{B_i+B_0}{2})^{3/2}$ factor and the feature index i will be dropped for simplicity, as the inclusion of the scaling factor and the introduction of additional features by introducing new B_i and C_i is trivial. The motivation for the form of Eqs. (20)–(23) is discussed in our previous paper introducing CIDER [35], and the heuristic selection of the B_i and C_i coefficients is discussed in Appendix A 8.

Before continuing, we briefly discuss the construction of the XC potential when nonlocal features are employed. A typical semilocal functional has a multiplicative XC potential

$$\frac{\delta E_{xc}}{\delta n(\mathbf{r})} = \frac{\partial e_{xc}(\mathbf{r})}{\partial n(\mathbf{r})}. \quad (24)$$

A meta-GGA also has a nonmultiplicative potential

$$\frac{\delta E_{xc}}{\delta \tau(\mathbf{r})} = \frac{\partial e_{xc}(\mathbf{r})}{\partial \tau(\mathbf{r})}. \quad (25)$$

This above term is nonmultiplicative because it depends on the orbitals through $\tau(\mathbf{r})$, rather than depending exclusively on the density $n(\mathbf{r})$. For a given basis set $\{\chi_\mu(\mathbf{r})\}$, the contributions of the XC potential to the effective Kohn-Sham Hamiltonian

matrix are

$$V_{\mu\nu}^{xc} = \int d^3\mathbf{r} \chi_\mu^*(\mathbf{r}) \chi_\nu(\mathbf{r}) \frac{\delta E_{xc}}{\delta n(\mathbf{r})} + \frac{1}{2} \int d^3\mathbf{r} \nabla \chi_\mu^*(\mathbf{r}) \cdot \nabla \chi_\nu(\mathbf{r}) \frac{\delta E_{xc}}{\delta \tau(\mathbf{r})}. \quad (26)$$

When nonlocal features are included in the XC energy density, e.g.,

$$E_{xc} = \int d^3\mathbf{r} e_{xc}(n(\mathbf{r}), |\nabla n(\mathbf{r})|, \tau(\mathbf{r}), G(\mathbf{r})), \quad (27)$$

both $\frac{\delta E_{xc}}{\delta n(\mathbf{r})}$ and $\frac{\delta E_{xc}}{\delta \tau(\mathbf{r})}$ include terms that depend on other coordinates \mathbf{r}' ,

$$\frac{\delta E_{xc}}{\delta n(\mathbf{r})} = \frac{\partial e_{xc}(\mathbf{r})}{\partial n(\mathbf{r})} + \frac{\partial e_{xc}(\mathbf{r})}{\partial G(\mathbf{r})} \frac{\partial G(\mathbf{r})}{\partial n(\mathbf{r})} + \int d^3\mathbf{r}' \frac{\partial e_{xc}(\mathbf{r}')}{\partial G(\mathbf{r}')} \frac{\delta G(\mathbf{r}')}{\delta n(\mathbf{r})}, \quad (28)$$

$$\frac{\delta E_{xc}}{\delta \tau(\mathbf{r})} = \frac{\partial e_{xc}(\mathbf{r})}{\partial \tau(\mathbf{r})} + \frac{\partial e_{xc}(\mathbf{r})}{\partial G(\mathbf{r})} \frac{\partial G(\mathbf{r})}{\partial \tau(\mathbf{r})} + \int d^3\mathbf{r}' \frac{\partial e_{xc}(\mathbf{r}')}{\partial G(\mathbf{r}')} \frac{\delta G(\mathbf{r}')}{\delta \tau(\mathbf{r})}. \quad (29)$$

In Eq. (28), $\frac{\partial G(\mathbf{r})}{\partial n(\mathbf{r})}$ describes the local dependence of $G(\mathbf{r})$ on the density, and $\frac{\delta G(\mathbf{r}')}{\delta n(\mathbf{r})}$ describes the nonlocal dependence on the density. While the expressions for $\frac{\delta E_{xc}}{\delta n(\mathbf{r})}$ and $\frac{\delta E_{xc}}{\delta \tau(\mathbf{r})}$ are more complicated than for a meta-GGA, they still contribute to the XC potential matrix exclusively via Eq. (26). Therefore, any DFT code supporting meta-GGAs can evaluate this type of functional if a routine is added to evaluate the nonlocal contributions to Eqs. (28) and (29).

C. Gaussian process models for the exchange functional

1. Overview

To study the impact of different features on the accuracy of an ML exchange functional, we implement four functional types that use different sets of features.

SL-GGA is a standard, semilocal GGA like PBE [23], except trained via machine learning. The exchange enhancement factor $[F_x$ in Eq. (7)] is a function of s only.

NL-GGA is a *nonlocal* GGA. The exchange enhancement factor is a function of s and also three nonlocal features of the form in Eq. (18), with Eqs. (20) and (21) for the exponents.

SL-MGGA is a standard, semilocal meta-GGA like SCAN [24], except trained via machine learning. The exchange enhancement factor is a function of s and α .

NL-MGGA is a *nonlocal* meta-GGA. The exchange enhancement factor is a function of s , α , and three nonlocal features of the form in Eq. (18), with Eqs. (22) and (23) for the exponents.

In short, the SL-GGA and SL-MGGA are semilocal functionals, while the NL-GGA and NL-MGGA augment the semilocal functional form with nonlocal features of the density. As described below, Gaussian process models for the exchange energy are trained using each of these four feature sets. Section II C 2 explains how a Gaussian process can be fit to total exchange energies of systems, Sec. II C 3 lists the specific mathematical form of each feature vector that is used

as input to the Gaussian processes, and Sec. II C 4 lists the covariance kernels for each feature set.

Gaussian process regression is a Bayesian machine-learning method in which the values of the predictive function $f(x)$ are treated as dependent random variables. The key ingredient in a Gaussian process is the covariance between values of the function $\text{Cov}(f(x), f(x')) = k(x, x')$, where the user-defined covariance kernel function $k(x, x')$ determines the relationship between the values of f for different inputs x . Suppose one has training data of the form (x_i, y_i) , where y_i is assumed to be the sum of $f(x_i)$ and Gaussian-distributed random noise of variance σ_{noise}^2 . Then the predictive mean, or expected value, of the function f is given by

$$f(x_*) = \sum_i k(x_*, x_i) \alpha_i, \quad (30)$$

with the weight vector α defined as

$$\alpha = (\mathbf{K} + \sigma_{\text{noise}}^2 \mathbf{I})^{-1} \mathbf{y}. \quad (31)$$

In the above equation $K_{ij} = k(x_i, x_j)$ is called the covariance matrix, and \mathbf{I} is the identity matrix. The reader is referred to the textbook by Rasmussen and Williams [51] for a more thorough introduction to Gaussian processes. The following subsection introduces a modified Gaussian process regression approach for training exchange-correlation functionals.

2. Fitting the total exchange energy

In this section, we introduce a method to train a Gaussian process to the total exchange energy, as opposed to the nonunique exchange energy density used in our previous paper [35]. While we only learn the exchange energy in this paper, the approach in this section also applies to the XC energy, so we write the formulas for E_{xc} for the sake of generality. We index the chemical systems in the training set by i and j ; while the feature implementations are different for Gaussian-type orbital and plane-wave DFT, the role of systems computed within these two frameworks is identical with regard to model training. The XC energy of a system i is given by a numerical integral of the energy density

$$E_{\text{xc}}^i = \sum_{g \in i} w_g^i e_{\text{xc}}(\mathbf{x}_g^i), \quad (32)$$

where g indexes grid points, w_g^i are the quadrature weights, and \mathbf{x}_g^i is the feature vector of molecule i at point g .

Constructing a Gaussian process for E_{xc} requires defining the covariance between the XC energies for two different systems, i.e., $\text{Cov}(E_{\text{xc}}^i, E_{\text{xc}}^j)$ (c.f. Chapter 2.2 of Ref. [51]). The covariance of sums of random variables a, b, c , and d is

$$\begin{aligned} \text{Cov}(a + b, c + d) &= \text{Cov}(a, c) + \text{Cov}(b, c) \\ &\quad + \text{Cov}(a, d) + \text{Cov}(b, d). \end{aligned} \quad (33)$$

Using Eq. (33), $\text{Cov}(E_{\text{xc}}^i, E_{\text{xc}}^j)$ can be expressed in terms of the covariances between the exchange energy densities as

$$\text{Cov}(E_{\text{xc}}^i, E_{\text{xc}}^j) = \sum_{g \in i} \sum_{h \in j} w_g^i w_h^j \text{Cov}(e_{\text{xc}}(\mathbf{x}_g^i), e_{\text{xc}}(\mathbf{x}_h^j)) \quad (34)$$

$$= \sum_{g \in i} \sum_{h \in j} w_g^i w_h^j k_{e_{\text{xc}}}(\mathbf{x}_g^i, \mathbf{x}_h^j), \quad (35)$$

where $k_{e_{\text{xc}}}(\mathbf{x}, \mathbf{x}')$ is the kernel function for the covariance between the exchange energy densities for feature vectors \mathbf{x} and \mathbf{x}' . Unfortunately, this expression can be quite expensive to compute because there are tens of thousands of grid points for each small molecule in the training database, and there are several hundred training molecules. Therefore, even for a database of this size, at least 10^{12} evaluations of the kernel are required. To circumvent this problem, a set of control points $\tilde{\mathbf{x}}_a$ can be sampled from the training database to use as a basis set for the feature space (as described in Appendix A 6), effectively creating a sparse Gaussian process [51]. Equation (35) is then approximated as

$$\text{Cov}(E_{\text{xc}}^i, E_{\text{xc}}^j) \approx (\tilde{\mathbf{k}}_{e_{\text{xc}}}^i)^T \tilde{\mathbf{K}}^{-1} \tilde{\mathbf{k}}_{e_{\text{xc}}}^j, \quad (36)$$

$$(\tilde{\mathbf{K}})_{ab} = k_{e_{\text{xc}}}(\tilde{\mathbf{x}}_a, \tilde{\mathbf{x}}_b), \quad (37)$$

$$(\tilde{\mathbf{k}}_{e_{\text{xc}}}^i)_a = \sum_{g \in i} w_g^i k_{e_{\text{xc}}}(\mathbf{x}_g^i, \tilde{\mathbf{x}}_a). \quad (38)$$

The approximation of Eq. (35) by Eq. (36) is called the Nyström approximation, and it is covered in more detail in Chapter 8.1 of Ref. [51]. The approximation can be thought of as interpolating the kernel function over the control points $\tilde{\mathbf{x}}_a$. Supplemental Material Section S9 investigates the accuracy of this approximation with respect to the number of control points used.

In this paper, the exchange functional alone is trained, and the kernel function is expressed in terms of the exchange enhancement factor $F_x = e_x / e_x^{\text{LDA}}$. This leads to the adjusted expressions

$$\text{Cov}(E_x^i, E_x^j) \approx (\tilde{\mathbf{k}}^i)^T \tilde{\mathbf{K}}^{-1} \tilde{\mathbf{k}}^j = K_{ij}, \quad (39)$$

$$(\tilde{\mathbf{K}})_{ab} = k_{F_x}(\tilde{\mathbf{x}}_a, \tilde{\mathbf{x}}_b), \quad (40)$$

$$(\tilde{\mathbf{k}}^i)_a = \sum_{g \in i} w_g^i (e_x^{\text{LDA}})_g^i k_{F_x}(\mathbf{x}_g^i, \tilde{\mathbf{x}}_a), \quad (41)$$

where $k_{F_x}(\mathbf{x}, \mathbf{x}')$ is the covariance kernel for the exchange enhancement factor and $(e_x^{\text{LDA}})_g^i = e_x^{\text{LDA}}(n_i(\mathbf{r}_g))$ is the LDA exchange energy density at coordinate g for molecule i [see Eq. (5)]. Having obtained these covariance expressions, they can be plugged into the predictive function for a Gaussian process to obtain the predicted exchange energy density [51]

$$e_x^{\text{CIDER}}(\mathbf{x}_*) = e_x^{\text{LDA}}(n_*) \sum_a k_{F_x}(\mathbf{x}_*, \tilde{\mathbf{x}}_a) \alpha_a, \quad (42)$$

$$\alpha = \sum_i \tilde{\mathbf{k}}^i \{[\mathbf{K} + \Sigma_{\text{noise}}]^{-1} \mathbf{y}\}_i, \quad (43)$$

where n_* and \mathbf{x}_* are the density and feature vector at a test point and Σ_{noise} is the noise matrix for the training points.

3. Model types and feature vectors

This section lists the feature vectors for each model type in terms of s [Eq. (9)], α [Eq. (11)], and the nonlocal features G_i [Eq. (18)]. All of the feature vectors below are designed such that for the uniform electron gas, $\mathbf{x} = \mathbf{0}$. Since the exchange enhancement factor is $F_x = 1$ for the uniform electron gas [49], the uniform electron gas constraint is enforced with a single noiseless training point $F_x^{\text{CIDER}}(\mathbf{0}) = 1$. All functionals in this paper satisfy the uniform electron gas constraint.

The primary purpose of the transformations performed on the features below is to constrain the ML feature vector elements to fall in a finite range (e.g., [0,1]), which makes it easier to map the resulting models to cubic splines as discussed in Sec. II C 4. Aside from that, the exact form of the feature transformation is a flexible choice, informed by physical considerations, but should not matter much as long as the target is a reasonably smooth function of the transformed features. However, more systematically fine-tuning these transformations in the future could potentially improve model performance slightly.

For the SL-GGA, the feature vector \mathbf{x} [Eq. (7)] is of dimension 1,

$$x_1 = \frac{\gamma s^2}{1 + \gamma s^2}, \quad (44)$$

where $\gamma = 0.243$ was empirically chosen by Becke [52] to fit the exact exchange energies of noble gas atoms. This value of γ was also used in our previous paper [35] and in the B97-type functionals of the Head-Gordon group [28,53,54]. For the SL-MGGA, the feature vector is of dimension 2,

$$x_1 = \frac{\gamma s^2}{1 + \gamma s^2}, \quad (45)$$

$$x_2 = \frac{2}{1 + \alpha^2} - 1. \quad (46)$$

The NL-GGA feature vector is of dimension 4,

$$x_1 = \frac{\gamma s^2}{1 + \gamma s^2}, \quad (47)$$

$$x_2 = \frac{G_1}{2 + G_1} - \frac{1}{2}, \quad (48)$$

$$x_3 = \frac{G_2}{2 + G_2} - \frac{1}{2}, \quad (49)$$

$$x_4 = \frac{G_3}{2 + G_3} - \frac{1}{2}. \quad (50)$$

Lastly, the NL-MGGA feature vector is of dimension 5,

$$x_1 = \frac{\gamma s^2}{1 + \gamma s^2}, \quad (51)$$

$$x_2 = \frac{2}{1 + \alpha^2} - 1, \quad (52)$$

$$x_3 = \frac{G_1}{2 + G_1} - \frac{1}{2}, \quad (53)$$

$$x_4 = \frac{G_2}{2 + G_2} - \frac{1}{2}, \quad (54)$$

$$x_5 = \frac{G_3}{2 + G_3} - \frac{1}{2}. \quad (55)$$

The three nonlocal features G_1, G_2, G_3 are given by Eq. (18), and the parametrization of the kernel exponents [i.e., the constants B_i and C_i in Eqs. (20)–(23)] is described in detail in Appendix A 8.

4. Kernels

To fully define the Gaussian process model for the exchange energy, we need to specify the kernel function $k_{F_x}(\mathbf{x}, \mathbf{x}')$ and the noise matrix Σ_{noise} . The noise matrix is constructed to heuristically approximate the model uncertainty

for each data point as described in Appendix A 7. To reflect the different number of features in each functional type, a different kernel is used for $k_{F_x}(\mathbf{x}, \mathbf{x}')$ for each, as specified below:

$$k_{\text{SL-GGA}}(\mathbf{x}, \mathbf{x}') = \Sigma_{\text{Cov}} k_1(x_1, x_1'), \quad (56)$$

$$k_{\text{SL-MGGA}}(\mathbf{x}, \mathbf{x}') = \Sigma_{\text{Cov}} k_1(x_1, x_1') k_2(x_2, x_2'). \quad (57)$$

$$k_{\text{NL-GGA}}(\mathbf{x}, \mathbf{x}') = \Sigma_{\text{Cov}} k_1(x_1, x_1') \times \sum_{i=2}^4 \sum_{j=i+1}^4 k_i(x_i, x_i') k_j(x_j, x_j'). \quad (58)$$

$$k_{\text{NL-MGGA}}(\mathbf{x}, \mathbf{x}') = \Sigma_{\text{Cov}} k_1(x_1, x_1') \times \sum_{i=2}^5 \sum_{j=i+1}^5 k_i(x_i, x_i') k_j(x_j, x_j'), \quad (59)$$

where Σ_{Cov} is a covariance hyperparameter that can be tuned for each model, and the base kernel $k_i(x, x')$ for all of the above is the squared-exponential kernel

$$k_i(x, x') = \exp \left[-\frac{1}{2} \left(\frac{x - x'}{l_i} \right)^2 \right]. \quad (60)$$

Equations (58) and (59) use the additive Gaussian process approach of Duvenaud *et al.* [55].

Because the cost of evaluating Gaussian processes scales linearly with the number of training points, we map the models to cubic splines after training so they can be evaluated efficiently. The semilocal kernels have 1 and 2 features, so they can be mapped to one- and two-dimensional cubic splines, respectively. The nonlocal kernels are a sum of three-dimensional kernels, so they can be mapped to a sum of three-dimensional cubic splines [35,56–58]. For a description of the hyperparameter selection process for these models, see Appendix A 8.

D. Auxiliary expansion of the CIDER features

The rest of this section focuses on the efficient implementation of nonlocal features within Gaussian-type orbital and plane-wave DFT, which is vital for the practicality of CIDER as well as any nonlocal density functional. Evaluating Eq. (18) via simple numerical integration is computationally expensive as a result of quadratic scaling with the number of grid points. The key insight to designing a faster approach comes from Román-Pérez and Soler [59], who developed a method to expand van der Waals density functionals as a sum of convolutions that could be computed efficiently using the fast Fourier transform (FFT). They observed that a function of the form in Eq. (19), which depends exclusively on the distance between two points and on a semilocal functional of the density at each point, can be approximately expanded as [59]

$$\Phi(a(\mathbf{r}_2), b(\mathbf{r}_1), r_{12}) \approx \sum_{\alpha\beta} p_\alpha^a(a(\mathbf{r}_2)) p_\beta^b(b(\mathbf{r}_1)) \Phi_{\alpha\beta}(r_{12}). \quad (61)$$

In Eq. (61), α and β are indices, which each correspond to interpolation control points q_α and q_β , respectively, and $p_\alpha^a(a(\mathbf{r}))$ and $p_\beta^b(b(\mathbf{r}))$ are interpolating functions, which take the values $p_\alpha^a(q_\alpha) = \delta_{\alpha\alpha'}$ and $p_\beta^b(q_\beta) = \delta_{\beta\beta'}$ at the interpolation points. The term $\Phi_{\alpha\beta}(r) \equiv \Phi(q_\alpha, q_\beta, r)$.

The Gaussian form of Eq. (19) is particularly conducive to the use of an even-tempered basis for interpolation, in which q_α takes the form

$$q_\alpha = q_0 \lambda^\alpha \quad (62)$$

for some constant $\lambda > 1$. Román-Pérez and Soler [59] used cubic splines for interpolation, and for the solid-state implementation of CIDER we follow this approach, with $p_\alpha^a(a(\mathbf{r}))$ being a cubic spline in the transformed coordinate $\gamma(\mathbf{r}) = \frac{\ln(a(\mathbf{r})/q_0)}{\ln(\lambda)}$. However, for the molecular implementation, we instead treat the control points as a basis set of Gaussians $e^{-q_\alpha r_{12}^2}$ onto which the function $e^{-a(\mathbf{r}_1)r_{12}^2}$ is projected,

$$p_\alpha^a(a(\mathbf{r})) \equiv \sum_\beta (S_a^{-1})_{\alpha\beta} s_\beta(a(\mathbf{r})), \quad (63)$$

$$(S_a)_{\alpha\beta} = \int d^3\mathbf{r} \exp[-(q_\alpha + q_\beta)|\mathbf{r}|^2] \quad (64)$$

$$s_\alpha(a(\mathbf{r})) = \int d^3\mathbf{r}' \exp\{-(a(\mathbf{r}) + q_\alpha)|\mathbf{r}'|^2\}. \quad (65)$$

As $\lambda \rightarrow 1$, the number of control points (for the interpolation scheme) and number of basis functions (for the basis expansion scheme) both become infinite, and the kernel expansion becomes exact for both approaches. We determined that the cubic interpolation yields better numerical stability for the plane-wave implementation, while the basis expansion form is sufficiently stable for the molecular implementation.

The benefit of expanding the kernel in this manner is that it allows the nonlocal features [Eq. (18)] to be expressed as a sum of convolutions

$$G(\mathbf{r}_1) = \sum_\beta p_\beta^b(b(\mathbf{r}_1)) \sum_\alpha \int d^3\mathbf{r}_2 \Phi_{\alpha\beta}(r_{12}) \theta_\alpha(\mathbf{r}_2), \quad (66)$$

with $\theta_\alpha(\mathbf{r}) = p_\alpha^a(a(\mathbf{r}))n(\mathbf{r})$. Within plane-wave DFT, the convolutions can be performed with quasilinear computational complexity by using FFTs [59]. In addition, because CIDER uses a squared exponential for $\Phi_{\alpha\beta}(r_{12})$ [Eq. (19)], the integrals of Eq. (66) can be evaluated analytically if $\theta_\alpha(\mathbf{r})$ is projected onto a Gaussian-type orbital basis. We use this approach to design an efficient implementation of nonlocal features for molecular DFT (Sec. II E).

The Román-Pérez and Soler technique effectively solves the problem of implementing CIDER for an all-electron plane-wave DFT calculation. However, such calculations are infeasible for all but the smallest atoms because a prohibitively large plane-wave basis would be required to describe the core electrons. In addition, pseudopotentials cannot be reliably used with nonlocal features because $G(\mathbf{r})$ calculated on the pseudodensity is not sufficiently similar to the exact, all-electron feature to get accurate results. To solve this problem, in Sec. II F a method is introduced to implement nonlocal features within the projector-augmented wave (PAW) formalism, which requires careful attention to the contributions to $G(\mathbf{r})$ from the core regions around the atoms.

Before proceeding, we note that while we borrow and expand on numerical techniques developed for nonlocal van der Waals functionals, we emphasize that the functionals in this paper are still fit only to the exchange energy and therefore do not account for the correlation effects that give rise to van der Waals interactions. Because of their nonlocality, the nonlocal features in this paper could prove useful for learning van der Waals interactions, but demonstrating this will require further research.

E. Efficient implementation of nonlocal CIDER features for molecular DFT

Computing Eq. (18) using simple numerical integration is prohibitively expensive because the number of operations is proportional to the square of the number of grid points. While van der Waals functionals like VV10 [60] are usually computed this way within Gaussian-type orbital codes, the contribution of such functionals is sufficiently small and smooth that a very sparse grid can be used. The nonlocal features require a grid of the same density as the semilocal XC grid, and the number of operations per pair of grid points is higher for evaluating Eq. (18) than, for example, the VV10 kernel. It is therefore necessary to reduce the computational complexity of evaluating the nonlocal features for molecular systems.

In this section, we present a charge-partitioning and density fitting-based approach to compute the nonlocal features within Gaussian-type orbital DFT codes for molecular systems, which we have implemented for use in the PySCF code [61,62]. Much of this approach is based on the study of Franchini *et al.* [63], who developed a similar method for computing the Coulomb energy. To start, the density is partitioned into atomic contributions using Becke partitioning [64],

$$n_A(\mathbf{r}) = W_A(\mathbf{r})n(\mathbf{r}), \quad (67)$$

$$\sum_A W_A(\mathbf{r}) = 1, \quad (68)$$

where A is the atom index and $W_A(\mathbf{r})$ are the atomic Becke weights. The first step in evaluating $G(\mathbf{r})$ is to project $\theta_\alpha(\mathbf{r})$ onto spherical harmonics for each atom (with L being the combined l and m indices and \mathbf{R}_A being the position of atom A),

$$\theta_{\alpha,L}^A(r) = \int d\Omega_A \theta_\alpha(\mathbf{r}) W_A(\mathbf{r}) Y_L(\Omega_A) \quad (69)$$

$$\mathbf{r} = \mathbf{R}_A + (r \cos \phi \sin \theta, r \sin \phi \sin \theta, r \cos \theta), \quad (70)$$

where $\Omega_A = (\theta, \phi)$ is the solid angle and $Y_L(\Omega)$ are the spherical harmonics. This projection is done numerically using Lebedev grids [65]. Then each L channel is projected onto an even-tempered Gaussian basis $\{\chi_{A,L,\mu}(\mathbf{r})\}$ as follows:

$$\chi_{A,L,\mu}(\mathbf{r}) = Y_L(\hat{\mathbf{r}}) N_{L,\mu} r^L \exp(-\mu r^2), \quad (71)$$

$$\tilde{\theta}_{\alpha,L,\mu}^A = \int dr r^{2+L} N_{L,\mu} \exp(-\mu r^2) \theta_{\alpha,L}^A(r), \quad (72)$$

where $N_{L,\mu}$ is a normalization constant. The exponents μ are constructed using the even-tempered Gaussian approach with

tunable parameter β^{aux} ,

$$\mu_n = \mu_{\min}(\beta^{\text{aux}})^n \quad n = \{0, 1, \dots, N\}. \quad (73)$$

A transformation to $\theta_{\alpha,L,\mu}^A$ is then performed by multiplying $\tilde{\theta}_{\alpha,L,\mu}^A$ by the inverse overlap matrix of the χ basis for each l channel,

$$\theta_{\alpha,L,\mu}^A = \sum_{\nu} [(\mathbf{S}_{\alpha,L}^{\chi})^{-1}]_{\mu\nu} \tilde{\theta}_{\alpha,L,\nu}^A, \quad (74)$$

$$(\mathbf{S}_{\alpha,L}^{\chi})_{\mu\nu} = \langle \chi_{\alpha,L,\mu} | \chi_{\alpha,L,\nu} \rangle. \quad (75)$$

For each atom, a second even-tempered basis $\{\xi_{A,L,\mu}(\mathbf{r})\}$ is introduced to serve as a basis for the convolutions of $\chi_{\alpha,L,\mu}(\mathbf{r})$. The following projections are computed using analytical Gaussian integrals:

$$\tilde{F}_{\beta,L,v}^A = \sum_{\alpha,\mu} \theta_{\alpha,L,\mu}^A \int d^3\mathbf{r}_1 d^3\mathbf{r}_2 \Phi_{\alpha\beta}(r_{12}) \chi_{\alpha,L,\mu}(\mathbf{r}_1) \xi_{A,L,v}(\mathbf{r}_2). \quad (76)$$

The transformation to $F_{\beta,L,v}^A$ is then performed by multiplying $\tilde{F}_{\beta,L,v}^A$ by the inverse overlap matrix of the ξ basis, similarly to Eq. (74).

The second to last step is to transform $F_{\beta,L,v}^A$ to $F_{\beta}(\mathbf{r}_g)$, where \mathbf{r}_g are the DFT integration grid points. Because evaluating every ξ orbital of every atom at every \mathbf{r}_g would be computationally expensive, $F_{\beta,L,v}^A$ is first transformed to a cubic spline $F_{\beta,L,G,p}^A$, where G is a radial coordinate index and p indexes the polynomial coefficients of the spline in the interval between points G and $G+1$. The transformation from $F_{\beta,L,G,p}^A$ to $F_{\beta}(\mathbf{r}_g)$ is then performed as

$$F_{\beta}(\mathbf{r}_g) = \sum_A \sum_L Y_L(\hat{\mathbf{r}}_{gA}) \sum_{p=0}^3 (r_{gA} - r_G)^p F_{\beta,L,G,p}^A, \quad (77)$$

where $\mathbf{r}_{gA} = \mathbf{r}_g - \mathbf{R}_A$. The spline index G is a step function of $|\mathbf{r}_{gA}|$. The computational cost of evaluating the splines depends heavily on the efficiency of memory access patterns on the processor, so the coordinates are sorted by G before evaluating the contributions to $F_{\beta}(\mathbf{r}_g)$ from each atom.

The last and simplest step to compute $G(\mathbf{r})$ is to evaluate the sum

$$G(\mathbf{r}) = \sum_{\beta} p_{\beta}^b(\mathbf{r}_g) F_{\beta}(\mathbf{r}_g). \quad (78)$$

After computing the XC energy from $G(\mathbf{r})$ and the semilocal features, the contributions to the XC potential from the $\theta_{\alpha}(\mathbf{r})$ terms can then be computed by passing backward through the matrix multiplications and spline evaluations detailed above.

The gradient of the total energy with respect to nuclear positions is required for the evaluation of forces. For this implementation of the CIDER functional with nonlocal features, the energy gradient with respect to the nuclear position of atom A is given by

$$\begin{aligned} \nabla_A E_{\text{xc}} &= \nabla_A^{\text{sl}} E_{\text{xc}} + \sum_g \left(\sum_{\alpha} \frac{\delta E_{\text{xc}}}{\delta \theta_{\alpha}(\mathbf{r}_g)} \theta_{\alpha}(\mathbf{r}_g) \right) \nabla_A w_g \\ &+ \sum_g \left(\sum_{\beta} \frac{\delta E_{\text{xc}}}{\delta F_{\beta}(\mathbf{r}_g)} \nabla_A F_{\beta}(\mathbf{r}_g) \right) w_g, \end{aligned} \quad (79)$$

where w_g is the numerical integration quadrature weight corresponding to the coordinate \mathbf{r}_g . The analytical energy gradient for a semilocal functional $\nabla_A^{\text{sl}} E_{\text{xc}}$ is given by Johnson *et al.* [66] Eq. (11).

The computational bottleneck of this algorithm is the spline evaluation in Eq. (77), whose cost scales quadratically with system size. Specifically, the cost scales as $\Theta(N_{\text{grid}} N_{\text{atom}} N_{\beta} l_{\text{max}}^2)$, where N_{grid} is the number of integration grid points, N_{atom} is the number of atoms, N_{β} is the number of kernel interpolation points in Eq. (61), and l_{max} is the maximum spherical harmonic order used in the ξ basis. All other components of the algorithm are performed on individual atoms and therefore scale linearly with system size.

F. Quasilinear-scaling implementation of nonlocal CIDER features for plane-wave DFT with PAW

This section covers our algorithm for evaluating the nonlocal CIDER features within the projector-augmented wave (PAW) method [67], a generalization of the pseudopotential method that can recover the full wave function and electron density. We have implemented this algorithm in the GPAW code [68,69]. As discussed above, Eqs. (61) and (66) enable the evaluation of $G(\mathbf{r})$ in $O(N \log N)$ complexity if the density $n(\mathbf{r})$ is represented on an FFT grid. However, in the PAW formalism of Blöchl [67], the all-electron density $n(\mathbf{r})$ is represented as

$$n(\mathbf{r}) = \tilde{n}(\mathbf{r}) + \sum_A (n_A^1(\mathbf{r}) - \tilde{n}_A^1(\mathbf{r})), \quad (80)$$

where A indexes the atoms, $\tilde{n}(\mathbf{r})$ is the pseudo-density on the FFT grid, and the terms with the 1 superscript are represented on radial support grids centered around each atom and extending to some cutoff radius r_c . Within the cutoff radius of atom A , $\tilde{n}(\mathbf{r}) = \tilde{n}_A^1(\mathbf{r})$ and $n_A^1(\mathbf{r})$ is the all-electron density. Outside the cutoff radius of atom A , $n_A^1(\mathbf{r}) = \tilde{n}_A^1(\mathbf{r})$.

For a semilocal functional, Eq. (80) yields a simple expression for the XC energy $E_{\text{xc}}[n]$,

$$E_{\text{xc}}[n] = E_{\text{xc}}[\tilde{n}] + \sum_A (E_{\text{xc}}[n_A^1] - E_{\text{xc}}[\tilde{n}_A^1]). \quad (81)$$

However, the case is not so simple for a nonlocal functional such as CIDER. If Eq. (18) is evaluated on the pseudodensity $\tilde{n}(\mathbf{r})$, the resulting $G(\mathbf{r})$ is not equal to the all-electron feature even outside the cutoff radius because the convolutions used to compute the feature are nonlocal. For the same reason, evaluating Eq. (18) on the atomic density $n_A^1(\mathbf{r})$ does not yield the all-electron feature inside the cutoff radius. Therefore, an implementation of CIDER (and other nonlocal functionals with similar forms, such as some kinetic energy and van der Waals functionals) for PAW must construct a ‘‘pseudofeature’’ $\tilde{G}(\mathbf{r})$ that is smooth but yields the correct feature outside the cutoff radius, as well as construct the ‘‘all-electron feature’’ $G(\mathbf{r})$ on the radial support grids inside the cutoff radius. Because of this difficulty, current plane-wave implementations of nonlocal functionals like van der Waals correlation [70] use the pseudodensity rather than the full PAW density. When this approach is used, one may need to use PAW datasets with more valence electrons, increasing computational cost [70]. In addition, because CIDER computes the exchange energy,

which is of much larger magnitude than the van der Waals correlation energy, higher precision is required. A similar problem exists for kinetic energy density functionals. There is an implementation of semilocal kinetic energy functionals with PAW [71], and of nonlocal kinetic energy functionals with ultrasoft pseudopotentials [72], but no nonlocal kinetic energy functional implementation within PAW. Therefore, by designing an algorithm for computing nonlocal features for CIDER within the PAW formalism, we also enable the full integration of van der Waals functionals and orbital-free DFT [73] into the PAW method.

Starting from Eq. (80), the all-electron value of $G(\mathbf{r})$ outside the core region is given by

$$\tilde{G}(\mathbf{r}) = \sum_{\beta} p_{\beta}^b(\tilde{\mathbf{b}}(\mathbf{r})) \left[\tilde{F}_{\beta}(\mathbf{r}) + \sum_A \Delta F_{A,\beta}(\mathbf{r}) \right], \quad (82)$$

$$\tilde{F}_{\beta}(\mathbf{r}_1) = \sum_{\alpha} \int d^3\mathbf{r}_2 \Phi_{\alpha\beta}(r_{12}) \tilde{\theta}_{\alpha}(\mathbf{r}_2), \quad (83)$$

$$\Delta F_{A,\beta}(\mathbf{r}_1) = \sum_{\alpha} \int d^3\mathbf{r}_2 \Phi_{\alpha\beta}(r_{12}) (\theta_{A,\alpha}^1(\mathbf{r}_2) - \tilde{\theta}_{A,\alpha}^1(\mathbf{r}_2)). \quad (84)$$

The quantities $\tilde{\mathbf{b}}(\mathbf{r})$ and $\tilde{\theta}_{\alpha}(\mathbf{r})$ are evaluated on the pseudo-density $\tilde{n}(\mathbf{r})$; $\theta_{A,\alpha}^1(\mathbf{r})$ and $\tilde{\theta}_{A,\alpha}^1(\mathbf{r})$ are evaluated on $n_A^1(\mathbf{r})$ and $\tilde{n}_A^1(\mathbf{r})$, respectively. $\tilde{F}_{\beta}(\mathbf{r})$ can be evaluated simply using FFTs. To compute $\Delta F_{A,\beta}(\mathbf{r})$, the term $\Delta\theta_{A,\alpha}^1(\mathbf{r}) = \theta_{A,\alpha}^1(\mathbf{r}) - \tilde{\theta}_{A,\alpha}^1(\mathbf{r})$ is first projected onto spherical harmonics and Gaussians as described in the previous section. Because $\Delta\theta_{A,\alpha}^1(\mathbf{r})$ is localized within the augmentation region and mostly spherical, far fewer spherical harmonics and Gaussians are required per atom than for the Gaussian-type orbital implementation. This projection yields $\Delta\theta_{A,\alpha,L,\mu}^1$. The Gaussians are then analytically projected into reciprocal space to yield $\Delta\theta_{A,\alpha,L}^1(k)$. In reciprocal space,

$$\Delta F_{A,L,\beta}(k) = \sum_{\alpha} \Phi_{\alpha\beta}(k) \Delta\theta_{A,\alpha,L}^1(k). \quad (85)$$

The above equation applies because the convolution $\Phi_{\alpha\beta}(k)$ has angular momentum number $l = 0$. Note that in this section, denoting a symbol as function of k implies that it is the Fourier transform of the real-space function corresponding to that symbol. For example, $\Phi_{\alpha\beta}(k)$ is the Fourier transform of $\Phi_{\alpha\beta}(r)$.

Inspired by the PAW method itself, we perform the following trick to approximate $\Delta F_{A,\beta}(\mathbf{r})$ on the FFT grid. We introduce two sets of functions, $\{g_i^A(\mathbf{r})\}$ and $\{h_j^A(\mathbf{r})\}$. The $h_j^A(\mathbf{r})$ functions form a localized basis (i.e., each function vanishes for $r > r_c$) and are meant to fit the high-frequency components of $\Delta F_{A,\beta}(\mathbf{r})$ introduced by the all-electron density inside the core region. For the current implementation of these functions, we use differences between Gaussians and smooth polynomials that match their derivatives at r_c . The other functions $g_i^A(\mathbf{r})$ are both localized and smooth, and are meant to augment $\Delta\theta_{A,\alpha}^1(\mathbf{r})$ to yield the correct values of $\Delta F_{A,\beta}(\mathbf{r})$ outside of the augmentation region. To do this, least-squares linear regression is used to fit the following function

to $\Delta F_{A,\beta}(\mathbf{r})$,

$$\hat{y}_{A,\beta}(\mathbf{r}) = \hat{y}_{A,\beta}^1(\mathbf{r}) + \hat{y}_{A,\beta}^2(\mathbf{r}), \quad (86)$$

$$\hat{y}_{A,\beta}^1(\mathbf{r}) = \sum_j D_{j\beta} h_j^A(\mathbf{r}), \quad (87)$$

$$\hat{y}_{A,\beta}^2(\mathbf{r}_1) = \sum_{i,\alpha} C_{i\alpha} \int d^3\mathbf{r}_2 \phi_{\alpha\beta}(r_{12}) g_i^A(\mathbf{r}_2), \quad (88)$$

where the coefficients $C_{i\alpha}$ and $D_{j\beta}$ are determined by the regression. The necessary integrals are actually performed in reciprocal space since Eq. (85) yields $\Delta F_{A,L,\beta}(k)$ rather than $\Delta F_{A,\beta}(\mathbf{r})$. At first glance, this appears to be a complicated and expensive fitting problem, but by separating the basis functions into their individual angular momentum channels and doing a separate regression for each channel, it becomes tractable. The details for this procedure are described in Appendix B 1.

Because the functions $h_j^A(\mathbf{r})$ are localized, in the complete basis set limit $\Delta F_{A,\beta}(\mathbf{r}) = \hat{y}_{A,\beta}^2(\mathbf{r})$ outside the augmentation region. Therefore, by defining

$$\tilde{\theta}_{\alpha}^{\text{aug}}(\mathbf{r}) \equiv \tilde{\theta}_{\alpha}(\mathbf{r}) + \sum_A \sum_i C_{i\alpha}^A g_i^A(\mathbf{r}), \quad (89)$$

one can compute

$$\tilde{G}(\mathbf{r}) = \sum_{\beta} p_{\beta}^b(\tilde{\mathbf{b}}(\mathbf{r})) \tilde{F}_{\beta}^{\text{aug}}(\mathbf{r}), \quad (90)$$

$$\tilde{F}_{\beta}^{\text{aug}}(\mathbf{r}) = \sum_{\alpha} \int d^3\mathbf{r}_2 \Phi_{\alpha\beta}(r_{12}) \tilde{\theta}_{\alpha}^{\text{aug}}(\mathbf{r}_2), \quad (91)$$

which satisfies the condition that the pseudofeature $\tilde{G}(\mathbf{r})$ is equal to the all-electron feature $G(\mathbf{r})$ outside the augmentation region (within the limit of complete basis sets).

Within the core regions, $\tilde{F}_{\beta}^{\text{aug}}(\mathbf{r})$ is projected onto the radial support grids using a set of projectors in a very similar formalism to PAW itself. This quantity will be called $\tilde{F}_{A,\beta}^{1,\text{aug}}(\mathbf{r})$, in keeping with the PAW notation, and it takes the form

$$\tilde{F}_{A,\beta}^{1,\text{aug}}(\mathbf{r}) = \sum_j B_{j\beta}^A \tilde{f}_j^A(\mathbf{r}) + \tilde{F}_{A,\beta}^{1,0}(\mathbf{r}) - \sum_j \tilde{f}_j^A(\mathbf{r}) \left[\int_1 d^3\mathbf{r}' \tilde{p}_j^A(\mathbf{r}') \tilde{F}_{A,\beta}^{1,0}(\mathbf{r}') \right], \quad (92)$$

$$B_{j\beta}^A = \text{dv} \sum_{g \in A} \tilde{F}_{\beta}^{\text{aug}}(\mathbf{r}_g) \tilde{p}_j^A(\mathbf{r}_g). \quad (93)$$

In the above equations, $\tilde{f}_i^A(\mathbf{r})$ and $\tilde{p}_j^A(\mathbf{r})$ are sets of atom-centered functions satisfying $\langle \tilde{f}_i^A | \tilde{p}_j^A \rangle = \delta_{ij}$, and $\tilde{p}_j^A(\mathbf{r})$ vanish for $r > r_c$ so that the numerical integral of Eq. (93) can be performed efficiently on the FFT grid points inside the cutoff sphere of atom A . The symbol dv represents the FFT grid volume element. This is directly analogous to the projector-partial wave pairs in PAW [67]. The construction of $\tilde{f}_i^A(\mathbf{r})$ and $\tilde{p}_j^A(\mathbf{r})$ is described in Appendix B 2. $\tilde{F}_{A,\beta}^{1,0}(\mathbf{r})$ is an approximate convolution of $\tilde{\theta}_{A,\alpha}^1(\mathbf{r})$, and while arbitrary, it helps make the $\tilde{f}_i^A(\mathbf{r})$ expansion more efficient. Using this construction, the

all-electron and pseudo-features on the radial support grids are, respectively,

$$G_A^1(\mathbf{r}) = \sum_{\beta} p_{\beta}^b(b_A^1(\mathbf{r}))(\tilde{F}_{A,\beta}^{1,\text{aug}}(\mathbf{r}) + \hat{y}_{A,\beta}^1(\mathbf{r})), \quad (94)$$

$$\tilde{G}_A^1(\mathbf{r}) = \sum_{\beta} p_{\beta}^b(\tilde{b}_A^1(\mathbf{r}))\tilde{F}_{A,\beta}^{1,\text{aug}}(\mathbf{r}). \quad (95)$$

The exchange correlation energy is then

$$E_{\text{xc}}[n] = E_{\text{xc}}[\tilde{n}, \tilde{G}] + \sum_A (E_{\text{xc}}[n_A^1, G_A^1] - E_{\text{xc}}[\tilde{n}_A^1, \tilde{G}_A^1]), \quad (96)$$

where all three terms are semilocal functionals of the density and of the nonlocal features. The first term is evaluated on the

FFT grid and the second two on the radial support grids on the atoms.

The use of nonlocal features introduces a few additional terms to the forces and stresses not present for semilocal DFT. The force terms are

$$\mathbf{F}_A = \mathbf{F}_A^{\text{sl}} + \mathbf{F}_A^{\text{PASDW}}, \quad (97)$$

$$\begin{aligned} \mathbf{F}_A^{\text{PASDW}} = & \sum_{\alpha} \text{d}v \sum_{ig} C_{i\alpha}^A \nabla g_i^A(\mathbf{r}_g) \frac{\partial E_{\text{xc}}}{\partial \theta_{\alpha}(\mathbf{r}_g)} \\ & + \sum_{\beta} \text{d}v \sum_{jg} \frac{\partial E_{\text{xc}}}{\partial B_{j\beta}^A} \nabla \tilde{p}_j^A(\mathbf{r}_g) \tilde{F}_{\beta}(\mathbf{r}_g). \end{aligned} \quad (98)$$

The stress terms are

$$\sigma_{\mu\nu} = \sigma_{\mu\nu}^{\text{sl}} + \sigma_{\mu\nu}^{\text{PASDW}}, \quad (99)$$

$$\begin{aligned} \sigma_{\mu\nu}^{\text{PASDW}} = & \sum_{\alpha} \sum_A \text{d}v \sum_g \left\{ \left[(\mathbf{r}_g - \mathbf{R}_A)_{\mu} \frac{\partial E_{\text{xc}}}{\partial \theta_{\alpha}(\mathbf{r}_g)} \right] \left[\sum_i C_{i\alpha}^A \nabla_v g_i^A(\mathbf{r}_g) \right] \right\} + \delta_{\mu\nu} \sum_{\beta} \sum_A \sum_j \left(\frac{\partial E_{\text{xc}}}{\partial B_{j\beta}^A} B_{j\beta}^A \right) \\ & + \sum_{\beta} \sum_A \text{d}v \sum_g \left\{ [(\mathbf{r}_g - \mathbf{R}_A)_{\mu} F_{\beta}(\mathbf{r}_g)] \left[\sum_j \frac{\partial E_{\text{xc}}}{\partial B_{j\beta}^A} \nabla_v \tilde{p}_j^A(\mathbf{r}_g) \right] \right\} + \sum_k \frac{G_{k\mu} G_{k\nu}}{|\mathbf{G}_k|} \sum_{\beta} \left(\frac{\partial E_{\text{xc}}}{\partial \tilde{F}_{\beta}(\mathbf{G}_k)} \right)^* \sum_{\alpha} \Phi'_{\alpha\beta}(|\mathbf{G}_k|) \theta_{\alpha}(\mathbf{G}_k). \end{aligned} \quad (100)$$

In the above equations, \mathbf{F}_A^{sl} and $\sigma_{\mu\nu}^{\text{sl}}$ refer to the force and stress equations for semilocal functionals within the PAW formalism (see Sec. III D of Ref. [74]).

There are several significant contributions to the computational cost of the above approach for evaluating $G(\mathbf{r})$, which all scale linearly or quasilinearly with system size. The first contribution comes from computing the Fourier transforms of $\tilde{\theta}_{\alpha}^{\text{aug}}(\mathbf{r})$ and $\tilde{F}_{\beta}^{\text{aug}}(\mathbf{r})$ (as well as the functional derivatives with respect to these quantities), which is necessary to compute the convolutions like those in Eq. (91) in reciprocal space. The cost of these operations scales as $\Theta(N_{\alpha} N_{\text{FFT}} \log(N_{\text{FFT}}))$, where N_{α} is the number of kernel interpolation points in Eq. (61) and N_{FFT} is the size of the FFT grid used for integrating the XC energy. The second contribution arises from computing the convolutions in Eq. (91) in reciprocal space, and scales as $\Theta(N_{\alpha}^2 N_{\text{FFT}})$. The third contribution comes from computing the PAW correction to $\tilde{\theta}_{\alpha}^{\text{aug}}(\mathbf{r})$ [the sum in Eq. (89)]. This and related terms scale as $\Theta(N_{\text{FFT,atom}} N_{\text{atom}} N_{\alpha})$, where $N_{\text{FFT,atom}}$ is the number of FFT grid points inside the core region of an atom and N_{atom} is the number of atoms. The last contribution consists of all the other PAW correction-related routines (e.g., computing $C_{i\alpha}^A$), which depend only on the density on the atomic radial grids and therefore scale linearly with N_{atom} . The overall scaling with system size N is then $\Theta(N \log(N))$. Notably, none of the components of the algorithm scale with the number of k points.

III. RESULTS AND DISCUSSION

Using Gaussian process regression, we fit each of the model types in Sec. II C (SL-GGA, NL-GGA, SL-MGGA,

and NL-MGGA) to match the exact exchange energy on subsets of the GMTKN55 [10] and SOL62 [33,75] databases, as well as a few dozen other training points. We provide a summary of our methods in this section and explain them in more technical detail in Appendix A.

To train the ML functionals, ground-state DFT calculations were first performed using the PBE functional [23]. Then, for each chemical system in the training set, the density and orbitals obtained from the PBE calculation were used to compute the features described in Sec. II C 3, as well the exact exchange energy of Eq. (2). The computed features were used as the inputs to the Gaussian process to predict the exact exchange energy. See Appendices A 2 and A 3 for details.

The training and validation systems were chosen as described in Appendix A 4. The SOL62 database [75] contains cohesive energies for a mix of metals and nonmetals, for which nine were used for training, 22 for validation, and 31 for testing. The GMTKN55 database contains 55 subdatabases blocked into five categories: small-molecule properties (“small”), larger molecule reactions and isomerizations (“big”), barrier heights (“barrier”), intermolecular noncovalent interactions (“inter NCI”), and intramolecular noncovalent interactions (“intra NCI”). The training and validation data consist of twelve subdatabases of GMTKN55 and contain some representative systems from each category, but disproportionately taken from the “small” category. All data not used for training or validation is considered the GMTKN55 test set. This GMTKN55 partitioning results in 280 train data, 163 validation data, and 1062 test data. Further technical details for training are provided in Appendices A 6 and A 7.

The SL-GGA, NL-GGA, SL-MGGA, and NL-MGGA functionals used for benchmarking below are the functionals of each type with the best validation set performance, as described in Appendix A 8. One important technical note is that each functional was trained via delta learning on top of some baseline exchange functional. We tried PBE and the Chachiyo GGA [76] as baselines, and Chachiyo proved more accurate for each functional class. However, as described in Sec. III C below, slight numerical stability improvements were found when using the NL-MGGA with a PBE baseline (dubbed NL-MGGA-PBE). As described in more detail in Sec. III C, we also found that retraining NL-MGGA-PBE with a more diverse training set from GMTKN55 (which is described in Appendix A 5) results in improved numerical stability. Therefore, the resulting functional, dubbed NL-MGGA-DTR, was used for the applications in Secs. III D–III G.

When testing and benchmarking the functionals, all calculations (both conventional and CIDER functionals) were performed self-consistently, with the exception of the SOL62 cohesive energies. These were obtained non-self-consistently from PBE orbitals, because of the difficulty of computing the isolated atom energies in plane-wave DFT. This is a small technical barrier that will be addressed in future work. All calculations in the main text used static geometries, with the exception of the defect calculations in Sec. III G; in this section, the silicon lattice parameters and defect structures were optimized with both PBE and CIDER.

In Sec. III A, we demonstrate that the NL-MGGA is capable of learning the exchange functional more accurately than the other feature sets, and in Sec. III B we use one- and two-electron systems to provide intuition for why this is. Next, we fine-tune the accuracy and numerical stability of the NL-MGGA (Sec. III C) and demonstrate that the resulting exchange functional can be used to match the accuracy of hybrid DFT (Sec. III D).

Having designed an accurate exchange functional, we demonstrate that it can be used to improve band gap predictions over standard meta-GGAs in Sec. III E. In Sec. III F, we show that for large systems in plane-wave DFT, our implementation of the NL-MGGA form is similarly efficient to semilocal meta-GGAs and much more efficient than hybrid DFT. Lastly, in Sec. III G, we take advantage of our functional’s efficiency and accurate band gap predictions to compute charge transition levels of point defects in silicon, which is a notoriously challenging task for DFT because of the underestimation of band gaps and the large supercell sizes required to simulate isolated point defects [6].

Taken together, the benchmarks provided below test the capacity of our ML models as well as their transferability across different chemical compositions and structures, including more complicated geometries like defect sites that are not present in the training data. However, our benchmarks mostly lack some system types and properties that play important roles in chemistry and materials science, including out-of-equilibrium geometries (with the exception of the barrier heights in GMTKN55), more complex materials like ternary and quaternary compounds and disordered phases, elastic and vibrational properties, etc. Transferability to these types of systems should not be assumed until future work provides more insight and further improvements to the ML models.

For now, our benchmarks of CIDER functionals for predicting bond lengths and lattice constants in Sec. S6 within the Supplemental Material [77] provide some initial insight into the ability of CIDER functionals to accurately predict structural properties in addition to energetic properties.

A. Accuracy of different model types for molecular and solid-state benchmarks

We assess the accuracy of our exchange functionals by substituting them for exact exchange in hybrid functionals to create a “surrogate hybrid” functional. For example, one of the simplest and most popular hybrid functionals is PBE0 [78], which mixes a 1/4 fraction of exact (Hartree-Fock, HF) exchange into the PBE functional [23],

$$E_{xc}[n] = \frac{3}{4}E_x^{\text{PBE}}[n] + \frac{1}{4}E_x^{\text{HF}}[n] + E_c^{\text{PBE}}[n]. \quad (101)$$

In this section we perform so-called PBE0/CIDER calculations, where the exchange-correlation energy is

$$E_{xc}[n] = \frac{3}{4}E_x^{\text{PBE}}[n] + \frac{1}{4}E_x^{\text{CIDER}}[n] + E_c^{\text{PBE}}[n]. \quad (102)$$

One can also generalize PBE0/CIDER with different fractions of CIDER exchange, giving the PBE0(α)/CIDER form, which is used in later sections,

$$E_{xc}[n] = (1 - \alpha)E_x^{\text{PBE}}[n] + \alpha E_x^{\text{CIDER}}[n] + E_c^{\text{PBE}}[n]. \quad (103)$$

Throughout the results, we will also use abbreviations like PBE0/SL-GGA and PBE0/NL-MGGA, which indicate that the specific CIDER functionals used in PBE0/CIDER are the SL-GGA and NL-MGGA, respectively.

All ML functional results in this section refer to the PBE0/CIDER surrogate hybrid form. Because our exchange functionals are trained on exact exchange energies, they are “perfect” if they exactly reproduce the predicted reaction energies of PBE0 on the test set. To measure the error of the exchange functionals, we calculate the average error between PBE0/CIDER and PBE0 for each functional type on GMTKN55 and SOL62. All calculations on GMTKN55 are SCF calculations, but the SOL62 cohesive energy CIDER and PBE0 calculations are performed non-self-consistently with PBE orbitals. See Appendices A 2 and A 3 for details.

The average deviations of the machine-learned functionals from PBE0, using the mean of means (MoM) metric (i.e., the mean of the mean absolute deviations, or MADs, for the sub-databases of GMTKN55), are presented in Fig. 1. Results for all functionals include D4 dispersion corrections [79], with the PBE0/CIDER functionals using the PBE0 dispersion parameters. The nonempirical PBE and r^2 SCAN [80] functionals, while not designed to reproduce PBE0, are included in Fig. 1 to show typical deviations between GGAs, meta-GGAs, and hybrids.

While the trends vary slightly in the different subsets of GMTKN55, the key trend is that the quality of the fit to exact exchange improves going from SL-GGA to NL-GGA to SL-MGGA to NL-MGGA. Notably, the SL-GGA and NL-GGA suffer from out-of-distribution transferability issues, resulting in larger deviations from PBE0 even compared to the PBE functional on the test set. On the other hand, the ML meta-GGAs are more transferable outside the training data than the ML GGAs, so they match PBE0 more closely than PBE and

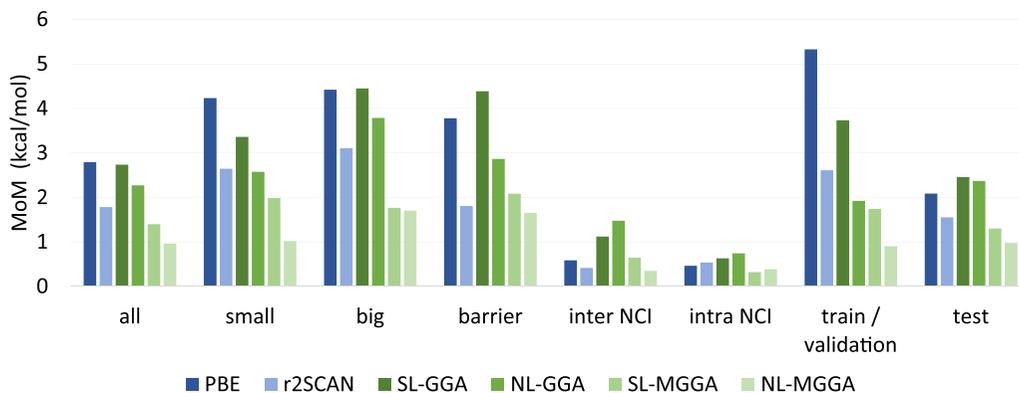


FIG. 1. Accuracy of PBE0/CIDER [Eq. (102)] compared to PBE0 for each functional type on the GMTKN55 database of molecular properties, using the MoM error metric. The database subsets are [10]: “all”, Full database; “small”, Basic properties and reaction energies for small systems; “big”, Reaction energies for large systems and isomerization reactions; “barrier”, Barrier heights; “inter NCI”, Intermolecular noncovalent interactions; “intra NCI”, Intramolecular noncovalent interactions.

r²SCAN on the test set. The NL-MGGA is more accurate than the SL-MGGA, achieving an MoM of less than 1 kcal/mol on the test set. The fact that the NL-MGGA outperforms both the NL-GGA and SL-MGGA suggests that the kinetic energy and nonlocal features provide nonredundant information, meaning that nonlocal density features cannot replace the kinetic energy density nor vice versa. As shown in Fig. 2, the same general trends apply for SOL62; the NL-MGGA is the most accurate and transferable ML functional for reproducing PBE0, followed by the SL-MGGA and then the NL-GGA.

Because GMTKN55 contains a variety of properties and systems, the performance of the functionals is somewhat dependent on the choice of error metric. However, the general trends are the same as for the MoM metric. See Sec. S2A within the Supplemental Material [77] for a discussion of the impact of error metric on the results.

B. Impact of nonlocality and derivative discontinuity on model accuracy

Overall, the SL-MGGA model is more accurate than the NL-GGA model across the databases tested here. However, the fact that NL-GGA improves over SL-GGA and that NL-MGGA improves over SL-MGGA suggests that the nonlocal features contribute useful, nonredundant information about the density for exchange energy prediction. This raises the question: What types of systems require the nonlocal features,

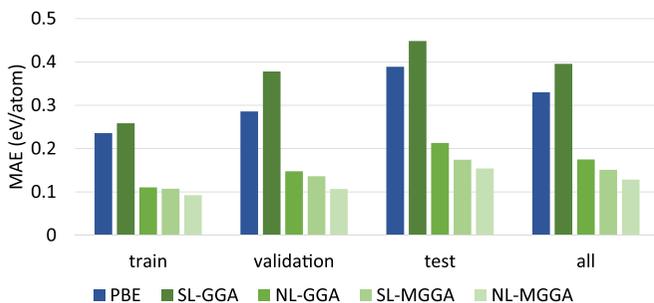


FIG. 2. Average deviation of PBE and the PBE0/CIDER ML functionals from PBE0 on the SOL62 cohesive energies database.

and which require the semilocal orbital dependence of the kinetic energy density? We hypothesize that the derivative discontinuity of the exchange functional [81] plays a key role. Because only the kinetic energy density—and not the nonlocal features of the NL-GGA—includes a derivative discontinuity passing through integer electron number, the NL-GGA (as well as any smooth, pure density functional) cannot be expected to accurately capture energy differences involving a change in electron number, such as ionization energies and electron affinities. On the other hand, existing semilocal meta-GGAs are fundamentally limited by their locality and cannot be expected to improve over GGAs for energy differences between systems of like electron number. This is especially true of one-electron systems, for which the kinetic energy is simply determined by the density gradient.

To explore this hypothesis, we set up two simple sets of systems. The first set is one-electron systems for different arrangements of protons: the hydrogen atom, the H₂⁺ molecular ion, the linear H₃²⁺ molecular ion, and the triangular H₃²⁺ molecular ion, all with bond lengths of 0.7 Å, as illustrated in Fig. 3. Because each system contains only one electron with different degrees of delocalization, we expect that only the functionals with nonlocal features can accurately predict energy differences between these systems. Table I shows the MAD between the ML exchange functionals and Hartree-Fock for the energy differences between the hydrogen atom and the molecular ions. As expected, the SL-GGA and SL-MGGA are both less accurate than NL-GGA and NL-MGGA. This confirms that the nonlocal features successfully

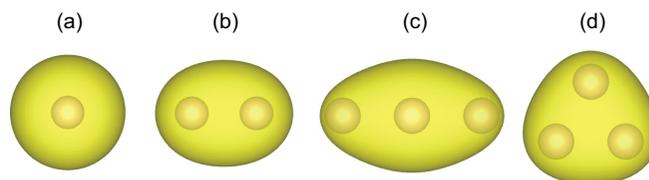


FIG. 3. Geometries and density distributions of the one-electron systems dataset. (a) Hydrogen atom. (b) H₂⁺. (c) H₃²⁺ linear. (d) H₃²⁺ triangular.

TABLE I. Mean absolute deviation (kcal/mol) of the four types of ML functional for simple one- and two-electron systems compared to Hartree-Fock.

Model	1e systems ^a	2e systems ^b
SL-GGA	2.9	8.7
NL-GGA	2.5	5.5
SL-MGGA	4.8	3.9
NL-MGGA	1.1	2.7

^aThe one-electron system dataset consists of the energies of H_2^+ , H_3^2 -linear, and H_3^2 -triangular referenced to the energy of the isolated hydrogen atom. All nearest-neighbor bond lengths are 0.7 Å.

^bThe two-electron system dataset consists of three energy differences between He^+ , He singlet, and He triplet, where He triplet is the neutral helium atom excited state with both electrons having the same spin.

characterize single-electron delocalization in a way that is not feasible with semilocal functionals.

The second set of systems consists of the He^+ ion, the He singlet (closed-shell ground state), and the He triplet (open-shell excited spin state with two same-spin electrons). Because the exchange energy has a discontinuous derivative as the number of electrons of a given spin passes through an integer [81,82], the energy differences between He^+ and He triplet and between He singlet and He triplet can only be described accurately by functionals that capture this derivative discontinuity. As shown in Table I, the absolute deviations from Hartree Fock for the energy differences between these three systems are larger for SL-GGA and NL-GGA than for SL-MGGA and NL-MGGA. This indicates that the derivative discontinuity included in the kinetic energy density is essential for capturing energy differences related to electron addition/removal. Notably, the NL-MGGA is the most accurate of the four functional types for both datasets.

These observations regarding the role of the derivative discontinuity explain why a previous attempt at machine learning a GGA resulted in more accurate atomization energies but less accurate ionization potentials than PBE [31]. It also provides a possible reason for the poor generalization of the SL-GGA and NL-GGA in Fig. 1; training on properties like ionization potentials and electron affinities that are difficult to describe with orbital-free functionals could lead to overfitting and decrease the accuracy of other properties. Likewise, the role of nonlocality explains why Kovács *et al.* [13] found that existing meta-GGAs, as well as 25 meta-GGAs they empirically trained, suffer from a trade-off between cohesive energy accuracy and band gap accuracy. Meta-GGAs are limited in their ability to describe both these properties at once due their semilocality. In addition, while we do not exhaustively try many possible NLDFs to conclusively show that they can only yield accurate models when paired with orbital-dependent features, the NLDFs we have implemented do need to be paired with a meta-GGA formalism to improve their accuracy. These findings provide evidence that to accurately and transferably fit a nonlocal, orbital-dependent quantity like the exchange energy, an ML functional should contain both orbital dependence and nonlocality in its feature set.

TABLE II. The number of systems in the GMTKN55 database for which a given functional did not achieve SCF convergence with thresholds of 10^{-8} Eh for total energy and 10^{-4} Eh for the orbital gradient of the energy. All functionals use the PBE0/CIDER form [Eq. (102)].

Functional	No. unconverged
SL-GGA	10
NL-GGA	16
SL-MGGA	5
NL-MGGA	3
NL-MGGA-PBE	1
NL-MGGA-DTR	0

C. Fine-tuning for numerical stability

Because XC functionals are used within SCF calculations, they must be numerically stable and consistently achieve electronic convergence on real systems to be useful for practical applications. For the GMTKN55 dataset, we found that some systems do not quite converge to our default convergence thresholds, which were 10^{-8} Hartree atomic unit (Eh) for the total energy and 10^{-4} Eh for the orbital gradient of the energy. (See Appendix A 9 for more details on how we remedied convergence problems.) We found that these calculations could only be converged with a 10^{-7} -Eh energy threshold and 2×10^{-3} -Eh gradient threshold. Interestingly, this problem is least severe for the NL-MGGA, as illustrated in Table II, which shows the number of systems for each functional type, which did not converge to the 10^{-8} -Eh energy and 10^{-4} -Eh gradient thresholds. We believe this is because the increased model dimensionality of the NL-MGGA allows it to fit the exchange energy with a smoother function of the input features.

Because the NL-MGGA still sees some convergence problems, we decided to explore two modifications of the NL-MGGA for improved numerical stability. The first is the NL-MGGA-PBE, which has the same hyperparameters and feature shape as the NL-MGGA but uses the PBE functional as a baseline rather than the Chachiyo functional [76]. NL-MGGA-PBE was screened during the validation procedure (Appendix A 8) but was found to be slightly less accurate than the NL-MGGA with the Chachiyo baseline. However, only one system sees a convergence issue in Table II with NL-MGGA-PBE, compared to three with NL-MGGA. We also tried training an NL-MGGA model with the PBE baseline and a more chemically diverse subset of GMTKN55, as described in Appendix A 5. We refer to this new train/validation/test partition as DTR (for “diverse training”), and it yields the NL-MGGA-DTR, for which all GMTKN55 systems achieve an orbital gradient convergence of 10^{-4} Eh or lower. In fact, all but 10 of the 2462 GMTKN55 systems converge to within 10^{-9} Eh for energy and $10^{-4.5}$ Eh for orbital gradient with NL-MGGA-DTR, and the rest to within 10^{-8} and 10^{-4} Eh, respectively.

As shown in Fig. 4, which displays MoM errors of PBE0/CIDER versus PBE0, the accuracy of these three NL-MGGA variants is similar across the GMTKN55 dataset. The three functional variants also have nearly identical performance for the SOL62 cohesive energies, with MADs

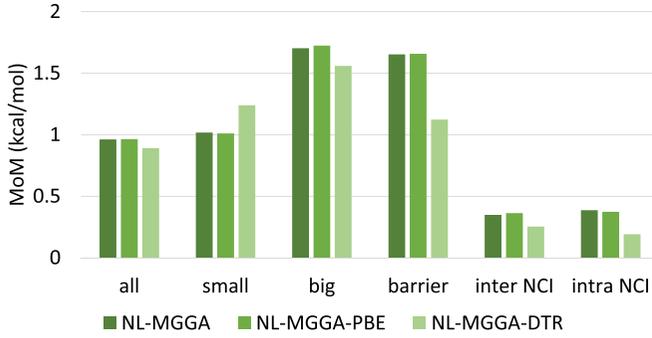


FIG. 4. Accuracy of GMTKN55 predictions for the NL-MGGA and the NL-MGGA-PBE and NL-MGGA-DTR variants described in the text (all using the PBE0/CIDER form), compared to PBE0. Train/test partition results are not shown because the train and test data are different for the NL-MGGA-DTR functional.

compared to PBE0 of 0.13 eV/atom for all three. Therefore, for the remainder of this paper, we adopt the NL-MGGA-DTR model as our primary functional of choice for applications, and we name it CIDER23X-NL-MGGA-DTR. This functional is provided (along with several others used in this paper) with an early release version of the CiderPress package as described in Sec. V, and it is the recommended functional for any researchers interested in running DFT calculations with CIDER.

In addition to the stability of a functional for the purpose of SCF convergence, one might also be interested in other metrics of numerical stability, such as the accuracy of the approximate evaluation of nonlocal density features described in Secs. IIE and IIF and the convergence of the CIDER functional with respect to the size of the grid used to integrate the XC energy. These characteristics are discussed in Sec. S8 within the Supplemental Material [77].

D. Obtaining hybrid DFT accuracy

In this section, we show that by substituting NL-MGGA-DTR for exact exchange in hybrid functionals, accurate predictions of the XC energy can be obtained for molecular systems without additional parameter tuning. In particular, we look at PBE0/CIDER [Eq. (102)] and another type of functional called PW6B95/CIDER. The original PW6B95 functional [83] is

$$E_{xc}^{PW6B95}[n] = 0.28E_x^{HF}[n] + E_{xc,sl}^{PW6B95}[n], \quad (104)$$

where $E_{xc,sl}^{PW6B95}$ is a meta-GGA term. Just like with PBE0/CIDER, we substitute CIDER in for Hartree-Fock exchange to obtain PW6B95/CIDER,

$$E_{xc}^{PW6B95}[n] = 0.28E_x^{CIDER}[n] + E_{xc,sl}^{PW6B95}[n]. \quad (105)$$

We chose to construct PW6B95/CIDER functionals because the original PW6B95 [83] is among the most accurate functionals on the GMTKN55 database [10,36].

For PBE0/NL-MGGA-DTR, PW6B95/NL-MGGA-DTR, and several semilocal [23,53,80,84] and hybrid [27,28,54,78,83] functionals, Fig. 5 shows the MoM error on all of GMTKN55 as well as the mean absolute error (MAE) on the test partition used for NL-MGGA-DTR. The

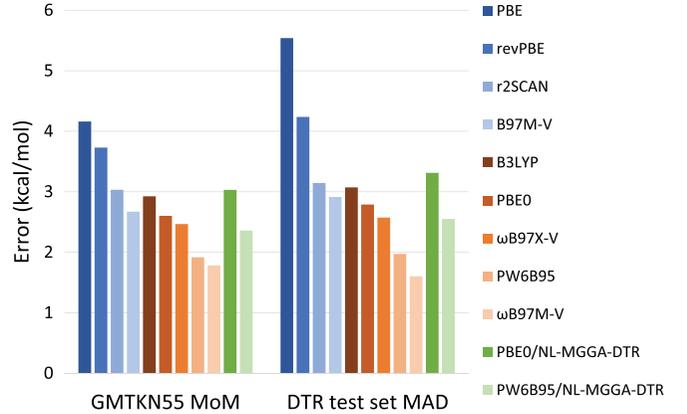


FIG. 5. MoM error for all of GMTKN55 (left) and MAD on the DTR test set (right) relative to GMTKN55 reference values, for semilocal, hybrid, and CIDER functionals.

errors are relative to the high-accuracy benchmark data collected in the GMTKN55 database [10]. All functionals include D4 dispersion corrections [79] except for B97M-V [53], ω B97X-V [28], and ω B97M-V [54], which all use VV10 dispersion [60]. PBE0/CIDER calculations use PBE0 dispersion parameters, and PW6B95/CIDER calculations use PW6B95 dispersion parameters. We show conventional semilocal functionals in shades of blue, conventional hybrid functionals in orange, and CIDER functionals in green.

As Fig. 5 shows, the PBE0/NL-MGGA-DTR functional is comparable in accuracy to r^2 SCAN, but the combination of systematic errors in the NL-MGGA-DTR functional and in PBE0 itself prevent it from outperforming the most accurate meta-GGA (B97M-V) or any of the hybrid functionals. On the other hand, because PW6B95 is much more accurate than PBE0 on the GMTKN55 dataset, PW6B95/NL-MGGA-DTR is more accurate than all semilocal functionals explored here. It is also more accurate than the PBE0 and B3LYP [27] hybrid functionals and equivalent in accuracy to the range-separated hybrid ω B97X-V [28]. This shows that the NL-MGGA-DTR exchange functional can be used to calculate molecular reaction energies with an accuracy that was previously not feasible without exact exchange mixing.

As with Sec. III A, the analysis of the errors on GMTKN55 depends somewhat on the error metric used. These nuances are discussed in further detail in Sec. S2B within the Supplemental Material [77], which also provides a comparison between the nonlocal and semilocal meta-GGA models.

E. Using CIDER for improved band gap prediction

The band gap problem [85,86], in which semilocal DFT tends to drastically underestimate the band gaps of solids, stands in the way of a variety of important applications like high-throughput screening for optical and electronic properties and semiconductor point defect physics [5,6,87]. By introducing a derivative discontinuity into the XC functional, meta-GGAs partly mitigate this problem [88], but they still tend to significantly underestimate band gaps [89]. Exceptions include mBJ [90] and TASK [12], which each have important drawbacks. The mBJ functional provides the exchange

TABLE III. Error statistics for 255 band gaps from the database published by Borlido *et al.* [45], in eV. The ML functionals use the PBE0(α)/CIDER formalism of Eq. (103).

	SL-MGGA ^a		NL-MGGA-DTR ^a		HSE06 ^b	PBE0 ^b	
	PBE ^a	r ² SCAN ^a	$\alpha = 0.25$	$\alpha = 0.35$			
ME ^c (eV)	-1.14	-0.81	-0.74	-0.61	-0.38	-0.17	0.40
MAE ^d (eV)	1.16	0.87	0.82	0.71	0.58	0.53	0.73
RMSE ^e (eV)	1.50	1.18	1.10	1.02	0.86	0.80	0.90
MARE ^f	0.53	0.41	0.41	0.35	0.33	0.33	0.67

^aThis paper.^bBorlido *et al.* [45].^cMean error.^dMean absolute error.^eRoot mean square error.^fMean absolute relative error.

potential only, not the energy [90], and TASK does not predict atomization energies and lattice constants as well as SCAN [12–14]. This reinforces the discussion in Sec. III B regarding the accuracy trade-offs meta-GGAs must make because of their semilocality.

By introducing nonlocality into the exchange energy, PBE0/NL-MGGA-DTR [Eq. (102) or Eq. (103) with $\alpha = 0.25$] improves band gap predictions over PBE and r²SCAN, as shown in Table III. This improvement occurs even though the CIDER functionals are not fit to solid-state band gaps. The PBE0/SL-MGGA also improves band gap predictions over r²SCAN, but significantly less so than PBE0/NL-MGGA-DTR. The database used for testing covers 255 of the 472 band gaps in the database published by Borlido *et al.* [45], with La, Yb, and Th-containing systems being excluded as well as systems with a greater than 0.05 eV difference between the spin-orbit-corrected and non-spin-orbit-corrected band gaps (as computed with PBE by Borlido *et al.* [45]). See Appendix A 11 for details, and Sec. S3 within the Supplemental Material [77] for error metrics for all 453 solids in the database that do not contain La, Yb, or Th.

To provide a clearer picture of the improved band gap prediction provided by the NL-MGGA-DTR functional, Table IV shows the band gap predictions for a selected subset of systems that are commonly studied [11]. In addition, Figs. 6 and 7 contain band structures for silicon (Si) and boron phosphide (BP), respectively, which show that PBE0/NL-MGGA-DTR provides similar band structures to PBE and r²SCAN but with a wider band gap.

There is room for improvement in these band gaps, since PBE0/NL-MGGA-DTR still systematically underestimates band gaps compared to the state-of-the-art HSE06 [91,92] range-separated hybrid functional, as well as the PBE0 functional to which PBE0/CIDER is an approximation. One possible solution is to tune the fraction of CIDER exchange, which is already done with exact exchange within hybrid DFT to optimize functional accuracy for different applications [93]. As shown in Table III, PBE0(α)/NL-MGGA-DTR with $\alpha = 0.35$ has increased band gaps over $\alpha = 0.25$ and similar error statistics to HSE06. However, one drawback of this approach is that because NL-MGGA-DTR increases band gaps less than exact exchange for a given mixing fraction, one might

need a large fraction of NL-MGGA-DTR to fit band gaps. This large fraction might not provide accurate predictions of other material properties. Alternative approaches to improve CIDER band gaps include using physical intuition and exact constraints to tune the derivative discontinuity contribution from the kinetic energy density (as is done in the TASK functional [12]), introducing more nonlocal features to improve the accuracy of the functional in general, and training on fundamental band gaps (i.e., differences between ionization potentials and electron affinities). However, considering the drastic increase in computational cost associated with using HSE06 or PBE0 compared to semilocal DFT, the ability to mitigate the underestimation of band gaps without explicitly fitting them—and while accurately predicting other properties

TABLE IV. Band gaps for a small subset of solids (in eV). CIDER refers to PBE0/NL-MGGA-DTR.

	PBE	r ² SCAN	CIDER	HSE06	Expt. [45]
Si	0.57	0.76	1.05	1.15	1.17
Ge	0.12	0.48	0.71	0.81	0.74
InP	0.69	1.12	1.47	1.52	1.42
GaAs	0.59	1.08	1.36	1.43	1.52
CdSe	0.71	1.19	1.48	1.69	1.74
BP	1.25	1.42	1.71	1.98	2.10
GaP	1.61	1.88	2.09	2.29	2.35
CdS	1.20	1.65	2.02	2.27	2.48
GaN	1.88	2.33	2.65	3.18	3.50
ZnS	2.13	2.69	3.04	3.37	3.72
C	4.13	4.34	4.70	5.33	5.50
BN	4.21	4.79	5.08	5.63	5.96
CaO	3.68	4.23	4.53	5.32	6.88
MgO	4.73	5.63	6.11	6.42	7.67
NaCl	5.11	5.90	6.32	6.45	8.75
LiF	9.07	10.05	10.40	11.39	13.60
Ar	8.71	9.61	9.71	10.36	14.15
ME	-1.93	-1.42	-1.11	-0.74	
MAE	1.93	1.42	1.11	0.76	
RMSE	2.41	1.88	1.66	1.31	
MARE	0.45	0.29	0.18	0.11	

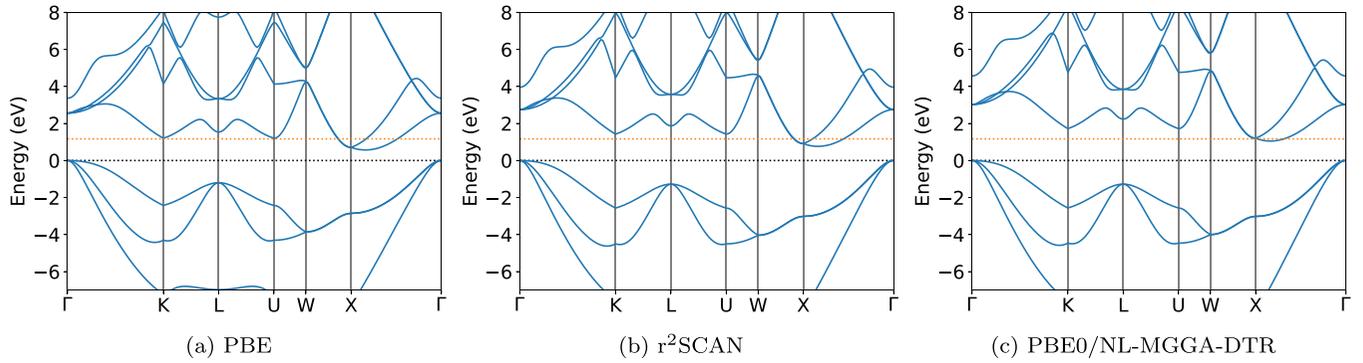


FIG. 6. Band structures of silicon with three different functionals. The black dotted line is positioned at the valence band maximum, and the orange dotted line is located at the experimental band gap of 1.17 eV. (a) PBE. (b) r^2 SCAN. (c) PBE0/NL-MGGA-DTR.

like bond energies—already serves as a major step toward solving the band gap problem.

F. Benchmarking performance on large condensed matter systems

The key benefit of the nonlocal features used to learn the exchange functional is that the quasilinear-scaling algorithm used to compute them within plane-wave DFT (Secs. II D and II F) is much faster than the evaluation of exact exchange for large systems and does not depend on the number of k-points. To illustrate this efficiency, Fig. 8(a) shows the wall time to perform a ground-state SCF calculation for a 216-atom diamond supercell with semilocal and PBE0/CIDER functionals in GPAW [68,69]. These wall times are compared to those of the hybrid DFT implementation in Quantum ESPRESSO [94,95], which uses the adaptively compressed exchange (ACE) algorithm [17] to compute the exact exchange energy as efficiently as possible. A Γ point only k-point mesh was used for these calculations. The CIDER functionals, including those with nonlocal features, all take less than 170 seconds (compared to 60 seconds for PBE and 79 seconds for r^2 SCAN). The total wall time for PBE0 is roughly 6.5 times slower than the NL-MGGA-DTR functional, in spite of the use of the ACE algorithm to accelerate the exact exchange calculation. Also, Quantum ESPRESSO has a general performance advantage over the plane-wave version of GPAW, with the 216-atom PBE calculation in Fig. 8(a) requiring 60 seconds for GPAW and only 21 seconds for Quantum

ESPRESSO. Therefore, more performance optimization in both the GPAW code and the CIDER feature evaluation might further increase the computational advantage of CIDER over hybrid DFT.

The efficiency of CIDER functionals for large systems makes it practical to perform more complicated calculations like defect formation energies. Figure 8(b) shows the wall time for computing the neutral vacancy formation energies in diamond and silicon using 216-atom supercells for semilocal and PBE0/CIDER functionals. These calculations used a $2 \times 2 \times 2$ Monkhorst-Pack k-point mesh [96]. All of the functionals are of comparable computational cost, with the NL-MGGA-DTR model being only about 25% slower than r^2 SCAN for diamond and 50% slower for silicon. This small decrease in efficiency comes with a significant benefit: As shown in Table V, of the functionals tested, the NL-MGGA model and its variants give the closest match to PBE0 for predicting these vacancy formation energies, with PBE0/NL-MGGA-DTR giving deviations of 0.18 eV and 0.23 eV for diamond and silicon, respectively. The HSE06 and PBE0 formation energies were computed using VASP [97–99]. See Appendix A 12 for further calculation details.

G. Charged defect transition levels in silicon

One of the consequences of the band gap problem is that transition levels in charged defects are not accurately described by semilocal DFT, so hybrid DFT is typically used when accurate transition level predictions are needed.

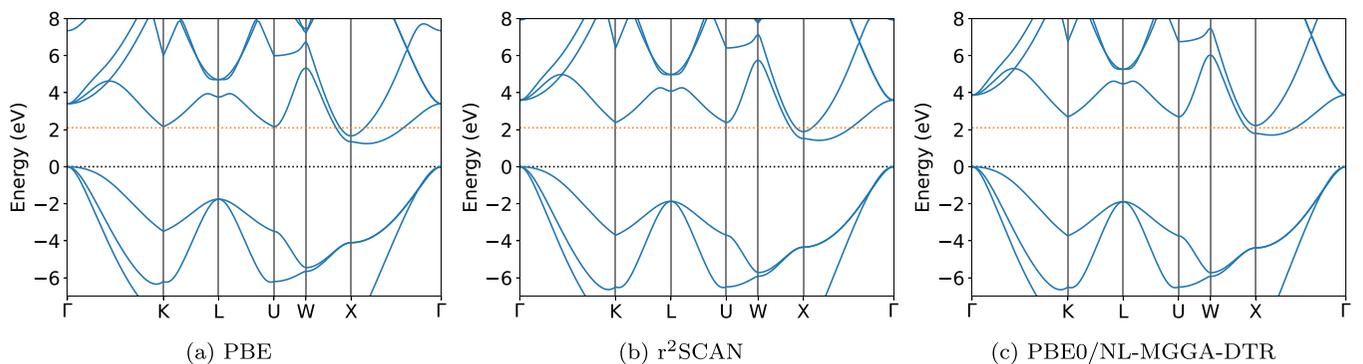


FIG. 7. Band structures of boron phosphide with three different functionals. The black dotted line is positioned at the valence band maximum, and the orange dotted line is located at the experimental band gap of 2.10 eV. (a) PBE. (b) r^2 SCAN. (c) PBE0/NL-MGGA-DTR.

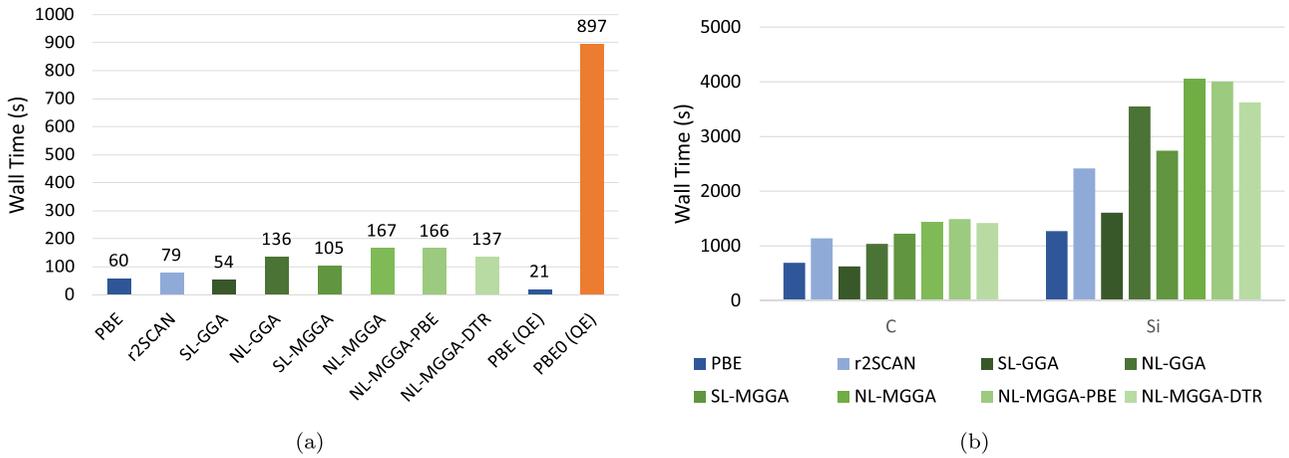


FIG. 8. Computational cost for plane-wave DFT SCF calculations of large systems with different functionals. The semilocal (blue) and CIDER (green) calculations were performed with GPAW on one 64-core Intel Ice Lake node for diamond and two 64-core nodes for silicon. The PBE0 calculation (orange) was performed on one 64-core node with the Quantum ESPRESSO (QE) code [94,95]. An additional PBE calculation was also performed with QE to compare the computational cost of GPAW and QE. (a) Total SCF calculation wall time for diamond 216-atom bulk supercell with Γ -point sampling only. (b) Total SCF calculation wall times for diamond and silicon neutral vacancy formation energies in a 216-atom supercell with $2 \times 2 \times 2$ Monkhorst-Pack k-point mesh [96].

In this section, we show that by improving band gap prediction, CIDER is able to accurately predict defect transition levels in silicon. The transition levels associated with the single vacancy [101], as well as the boron [102], phosphorus [103], copper [104], and sulfur [105,106] substitutionals, were computed as described in Appendix A 13, using a 512-atom supercell, $2 \times 2 \times 2$ Γ -centered k-point mesh, and full structural relaxation for each defect and functional.

Figure 9 shows the computed transition levels compared to experiment with two different finite-size correction schemes: the Freysoldt finite-size corrections [100] [Fig. 9(a)] and the potential alignment correction [Fig. 9(b)]. We computed the transition levels with PBE and PBE0(0.3)/NL-MGGA-DTR [Eq. (103) with $\alpha = 0.3$]. The latter functional was chosen because its band gap matches the 1.17 eV experimental gap; however, PBE0/NL-MGGA-DTR yields similar results (see Sec. S7 within the Supplemental Material [77]). Table S13 within the Supplemental Material provides a more detailed list

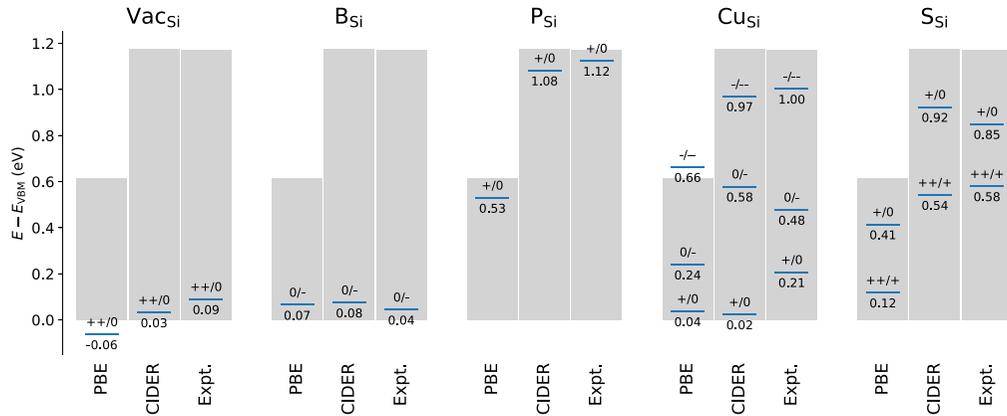
TABLE V. The formation energies of the neutral vacancies in diamond and silicon (in eV), computed with 216-atom supercells and a $2 \times 2 \times 2$ Monkhorst-Pack k-point mesh [96]. All CIDER functionals use the PBE0/CIDER surrogate hybrid form.

Functional	C	Si
PBE	6.47	3.61
r ² SCAN	6.36	4.19
SL-GGA	6.11	3.48
NL-GGA	6.73	3.50
SL-MGGA	6.66	4.16
NL-MGGA	6.77	4.58
NL-MGGA-PBE	6.77	4.58
NL-MGGA-DTR	6.79	4.42
HSE06	6.96	4.54
PBE0	6.97	4.65

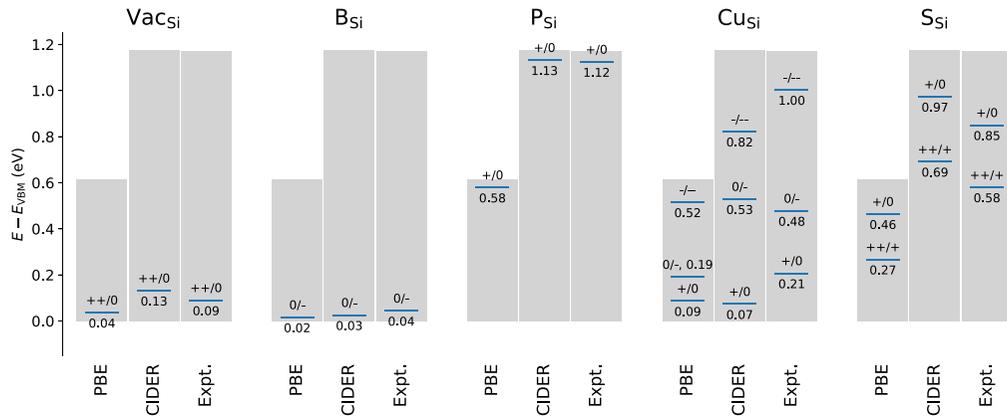
of all computed transition levels. In the rest of this section, PBE0(0.3)/NL-MGGA-DTR is referred to simply as CIDER for brevity.

PBE and CIDER both give reasonable descriptions of the $++/0$ transition level in the silicon vacancy and the boron acceptor level. However, the phosphorus donor level is not well described by PBE because of its severe underestimation of the band gap; CIDER remedies the band gap underestimation and therefore predicts a transition level in good agreement with experiment. To correct for band gap underestimation, one could predict defect levels from PBE by scaling the gap (and with it, the transition levels) to the experimental gap, but this makes it difficult to distinguish true shallow levels from deep levels that happen to sit near the (too-low) PBE conduction band edge [5]. For example, PBE predicts a transition level of 0.66 eV for the $-/-$ transition of Cu_{Si} [Fig. 9(a)], 0.05 eV above the conduction band minimum. Without further information, it is unclear whether this state is resonant in the conduction band or localized in the gap. CIDER, on the other hand, places the $-/-$ level inside the gap at 0.97 eV, in good agreement with the 1.00 eV level from experiment.

More generally, Fig. 9(a) shows that paired with the Freysoldt correction, CIDER matches each level in the figure with an error of 0.10 eV or less compared to experiment, with the exception of the $+/0$ level in Cu_{Si} . For this transition, the experimental level is 0.20 eV [104], while the CIDER level is 0.02 eV with the Freysoldt correction and 0.07 eV with the potential alignment correction only. The HSE06 level was previously reported to be 0.21 eV by Sharan *et al.* [107], in good agreement with experiment. However, Sharan *et al.* used a 64-atom supercell with a $2 \times 2 \times 2$ Monkhorst-Pack k-point mesh [96]. This supercell size is relatively small, and the k-point mesh is both small and fails to sample the band edges, which can interfere with the description of relatively delocalized defect states. When we recalculated the CIDER transition level using Sharan *et al.*'s supercell size and k-point



(a) Freysoldt finite-size corrections [99].



(b) Potential alignment correction only.

FIG. 9. Electronic transition levels for the single vacancy and several substitutional defects in silicon computed using different functionals and finite-size correction schemes. The grey bars represent the band gap for the given functional (or experiment). The experimental references are from the following sources: vacancy [101], boron [102], phosphorus [103], copper [104], and sulfur [105,106]. CIDER refers to PBE0(0.3)/NL-MGGA-DTR. (a) Transition levels computed using Freysoldt finite-size corrections [100]. (b) Transition levels computed using potential alignment correction only.

mesh, we obtained 0.18 eV with Freysoldt corrections, in good agreement with their result and with experiment. In Sec. S7A within the Supplemental Material [77], we provide further evidence that the HSE06 level itself is not converged with respect to k-point mesh. This finding illustrates the importance of using well-converged supercell sizes and k-point meshes in point defect calculations, a requirement that is made much easier by the efficiency of the CIDER functionals compared to hybrid DFT. Why this particular transition level is underestimated by both hybrid DFT and CIDER is an interesting question warranting future investigation.

In addition to yielding more accurate transition levels, CIDER predicts somewhat different charge distributions than PBE as well. Figure 10 shows the charge density differences between charge states for several of the transition levels in Fig. 9. There are a few notable takeaways from these plots. First, as expected, the charge density differences are larger for transitions that result in large ionic displacements (as indicated by the MaxDisp in the bottom of each plot). Second, these ionic displacements and resulting charge distortions are noticeable more than 5 Å away from the defect site for some transitions, indicating the importance of large

supercells. Lastly, a slightly larger fraction of the charge associated with the transition is contained within a 5-Å sphere around the defect site for CIDER than for PBE (as indicated by the Loc. metric in the bottom of each plot), suggesting that CIDER localizes the defect charge more effectively than PBE. This is reassuring because unphysical charge delocalization is one of the key driving factors of erroneous charge transition level predictions in semiconductor point defects [5].

In summary, our results indicate that the CIDER approach could be a route toward high-accuracy defect calculations at similar cost to semilocal DFT, which would significantly broaden the potential scope of point defects research. In particular, it could be used to improve the accuracy of high-throughput point defects studies [108], which are currently only computationally feasible with semilocal functionals like PBE.

IV. CONCLUSIONS

We have demonstrated that by combining the kinetic energy density (a semilocal, orbital-dependent quantity) with nonlocal, density-dependent features, it is possible to design functionals with comparable accuracy to hybrid DFT at a

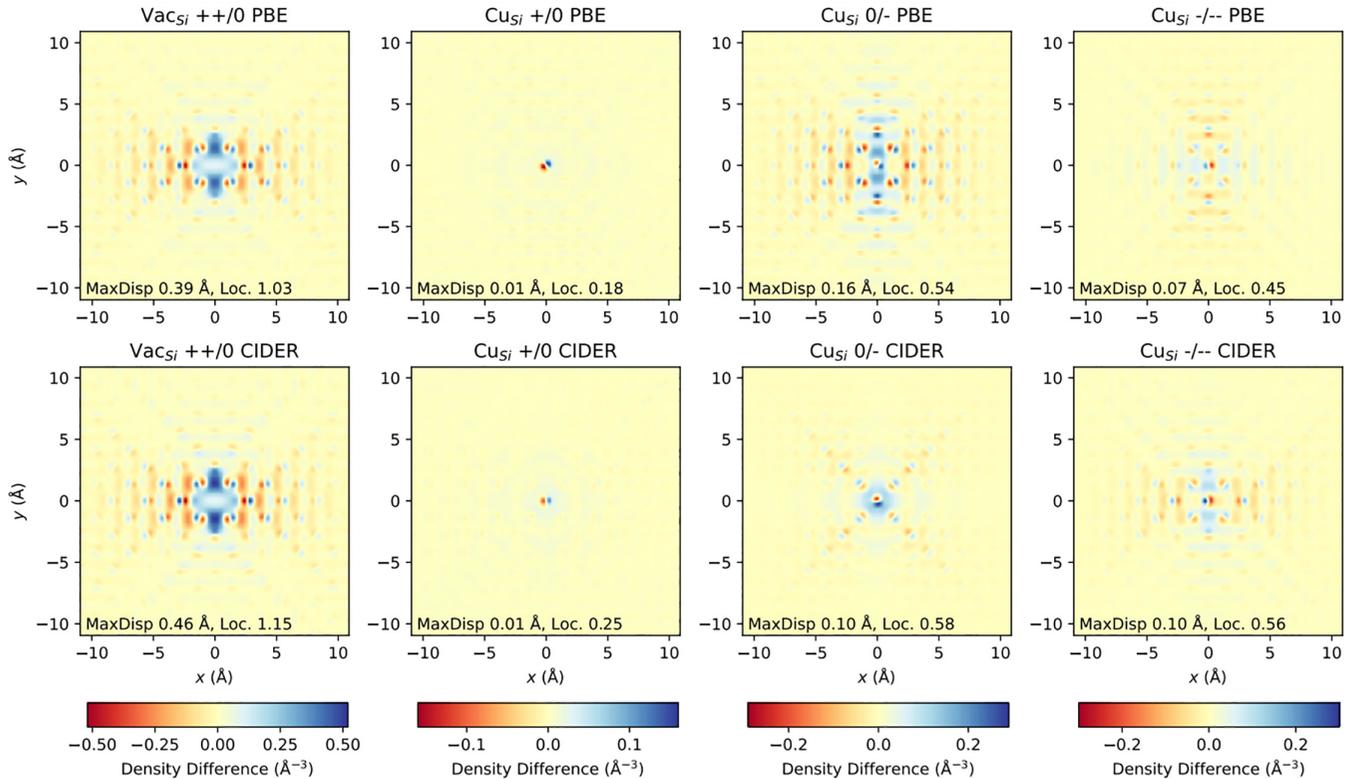


FIG. 10. Change in charge distribution in the (001) plane (integrated over the z axis) for several charge transition levels in point defects of silicon. Blue shades represent an increase in electron density, while red shades indicate a decrease in electron density. “MaxDisp” is the maximum ionic displacement between the charge states, and “Loc.” is the fraction of the charge difference contained within a 5 Å sphere around the defect site. Loc. is greater than one for the vacancy ++/0 transition because more than two electrons are drawn in to the defect region upon addition of an electron pair, depleting electron density in the surrounding region.

drastically reduced computational cost for plane-wave DFT calculations. We can do this by learning exact exchange energies, without the use of any high-accuracy reference data from experiments or wave function theory. We have shown that our nonlocal meta-GGA model has higher accuracy than semilocal functionals on both molecular benchmarks and solid-state band gap prediction, and that these improvements can assist with applications such as semiconductor point-defect calculations.

The developments presented here serve a variety of purposes. As is, the exchange functional we designed can be used as an exact exchange surrogate in scenarios where hybrid DFT is impractical, such as large extended systems. This expands the scope of calculations that can be performed at hybrid DFT accuracy. In addition, the techniques we have introduced to learn smooth, physically constrained functionals and implement efficient nonlocal features could be used in other contexts, such as orbital-free DFT [73] and van der Waals functional design [59,70,109].

Importantly, the ability to systematically vary the descriptor complexity in our machine learning approach allows us to extract physically nontrivial information about the nature of the exchange interactions. Specifically, we find that both nonlocality and derivative discontinuity are independently important for the structure of the exchange energy density functional. This clarifies a key limitation of existing semilocal functionals.

Lastly, our approach can be generalized to learn the full XC functional, providing a path to highly transferable and universal functionals capable of describing challenging and heterogeneous chemistry. We anticipate that these developments will contribute to significantly improving the cost-accuracy trade-off of DFT, accelerating the pace and efficacy of materials and chemical research by providing a stronger theoretical backing in various subfields.

The code to run CIDER calculations in both the PySCF and GPAW codes is available at the following Github repository [110]. The SL-GGA, NL-GGA, SL-MGGA, NL-MGGA, NL-MGGA-PBE, and NL-MGGA-DTR functionals are available for download using the instructions in the Github repository.

ACKNOWLEDGMENTS

The authors thank Jennifer Coulter and Stefano Falletta for help with the hybrid DFT benchmark calculation. The authors also thank Stefan Riemelmoser and Georg Kresse for useful discussions. This work was supported by the National Defense Science and Engineering Graduate (NDSEG) Fellowship Program under Contract No. FA9550-21-F-0003, the Camille and Henry Dreyfus Foundation Grant No. ML-22-075, and the Department of Navy Award No. N00014-20-1-2418 issued by the Office of Naval Research. Computational resources were provided by the Harvard University FAS Division of Science Research Computing Group.

APPENDIX A: COMPUTATIONAL DETAILS

1. Overview

To explore the effectiveness of different features for learning the exchange energy, four types of functional were trained, with the model and feature vector details provided in Sec. IIC:

(i) SL-GGA: A semilocal GGA model based on the reduced squared gradient s^2 [Eq. (9)].

(ii) NL-GGA: A nonlocal model based on s^2 and three nonlocal features with Eqs. (20) and (21) for the exponents.

(iii) SL-MGGA: A semilocal meta-GGA model based on s^2 and the iso-orbital indicator α [Eq. (11)].

(iv) NL-MGGA: A nonlocal model based on s^2 , α , and three nonlocal features with Eqs. (22) and (23) for the exponents.

The purpose of training semilocal GGA and meta-GGA models is to test the expressive power of the nonlocal descriptors. In our previous paper (see Table S1 of Ref. [35]), we found that a standard, semilocal meta-GGA could not effectively learn the exchange energy density, but in this paper, we determined that a semilocal meta-GGA can learn the *total* exchange energy much more effectively than it can learn the exchange energy density.

The two primary datasets used in this paper were the GMTKN55 [10] and SOL62 [33,75] databases. The GMTKN55 database contains 1505 molecular reaction energies based on 2462 single-point calculations, covering five categories: small-molecule properties, larger molecule reactions and isomerization, barrier heights, intermolecular noncovalent interactions, and intramolecular noncovalent interactions. The SOL62 dataset consists of 42 nonmetals and 20 metals as categorized by Zhang *et al.* [75]. We follow Trepte and Voss [33] in excluding Pb and Th from Zhang *et al.*'s initial set of 64 solids. Twelve subsets of GMTKN55 and half of SOL62, along with a few other systems, were used for training and validation, as detailed in Appendix A4.

Throughout this paper, we used pymatgen [111] and ase [112] for structure manipulation and calculation setup, fireworks [113] for automated workflow management, and Scikit-learn [114] for the Gaussian process models. The computational details for the PySCF and GPAW calculations are provided in Appendices A2 and A3, respectively. Appendices A4 and A5 describe the training and validation set selection, and Appendix A6 covers the selection of control points for the sparse Gaussian processes. Appendix A7 describes the selection of noise hyperparameters and training/validation loss functions, and Appendix A8 describes the list of hyperparameters tested for each model and the model selection procedure. Appendix A9 describes how we handled convergence issues for CIDER models on GMTKN55, and Appendix A10 describes the PySCF and GPAW settings that are specific to the CIDER functionals. Appendix A11 describes the band gap benchmark methods. Appendix A12 describes the computational cost benchmarking of CIDER for plane-wave DFT calculations. Finally, Appendix A13 describes our methodology for computing charged defect transition levels.

2. Molecular calculations and reference data generation

Molecular systems, including the nanoclusters discussed in Appendix A4, were computed using the PySCF code [61,62]. All calculations were performed using the def2-QZVPPD basis set (with effective core potentials for atoms larger than Kr) [115,116] for the atomic orbitals and the def2-universal-jkfit auxiliary basis set [117] for the classical Coulomb energy. The seminumerical exchange (SGX) module in PySCF, which is similar to the computationally efficient chain-of-spheres exchange algorithm [118], was used to compute the exchange energy and potential for hybrid functionals. For SGX, level 1 PySCF grids were used with P-junction screening turned off. For a given subdatabase in GMTKN55, restricted DFT was used if all systems in that subdatabase had $S_z = 0$; otherwise, unrestricted DFT was used. The exception was the PCl_3 transition state in the INV24 database, for which spin symmetry breaking occurs for some functionals. This system was calculated with unrestricted DFT as well. For SCF calculations, the nonlocal features and XC energy were computed on level 3 PySCF grids.

For benchmarking comparison, the GMTKN55 database was evaluated with the above settings using the PBE [23], revPBE [84], $r^2\text{SCAN}$ [80], PBE0 [78], B3LYP [27], and PW6B95 [83] functionals, all with D4 dispersion corrections [79], as well as with the B97M-V [53], $\omega\text{B97X-V}$ [28], and $\omega\text{B97M-V}$ [54] functionals.

To collect the exact exchange energy training data, the ground-state orbitals from the PBE calculations were used to compute the exact exchange energy E_x^{exact} via Eq. (2), and the feature vectors were computed from the corresponding PBE densities. The features were computed on level 1 PySCF grids. For training set feature generation only (not the SCF calculations), the integral in Eq. (18) was computed numerically with the \mathbf{r}_2 coordinate evaluated on level 3 PySCF grids without pruning. This dense grid is necessary for numerical stability and precision in the core region.

3. Solid-state calculations and reference data generation

The electron densities and ground-state energies of the SOL62 dataset were computed in GPAW [68,69] using the PBE geometries obtained by Trepte and Voss [33], with an energy cutoff of 1000 eV and the same k-point meshes as Trepte and Voss. The h parameter in GPAW was set to the minimum value required to avoid aliasing of the density, i.e., 0.096 Å. Fermi-Dirac smearing with a width of 0.01 eV was used for periodic systems. For isolated atoms, fixed occupations were used unless numerical instabilities occurred as a result, which was the case for six atoms (Pb-2, Pd-2, Pt-2, Sn-2, Ta-5, and V-5, where the number is the ground-state magnetic moment $2S$). In these cases, Fermi-Dirac smearing with a width of 0.002 eV was used. The default Davidson solver was used for all solids except the alkaline earth metals, which were solved with the conjugate gradient (CG) method. The CG method was used for all isolated atomic systems. To compute cohesive energies for training, the experimental ground state magnetic moments were used for the isolated atom references. The convergence threshold was set to 10^{-6} eV per valence electron for both isolated atoms and periodic systems. The density convergence threshold was set

to 10^{-4} per electron for periodic systems and 10^{-3} per electron for isolated atoms. The eigenstates convergence threshold was set to 4×10^{-8} eV²/electron for periodic systems and 10^{-5} eV²/electron for isolated atoms. The cell size for the isolated atoms was chosen to be about 11.7 Å across, with perturbations included to break symmetry. Compared to a larger box size of 17.5 Å, the PBE atomic energies had a root mean square deviation of 0.012 eV and maximum deviation of 0.080 eV (for Zr). Because of the small average deviation, and the fact that the train and test reference values were computed on the atom-in-a-box systems rather than determined by experiment, the smaller boxes were used to lower the computational cost and memory requirements of training and testing. Symmetry constraints were also turned off for the isolated atom calculations.

Training the CIDER models to the solid-state data required computing the exact exchange energy and the nonlocal features for each system. To obtain exact exchange energies, the hybrid module of GPAW was used to compute the non-self-consistent exact exchange energy (EXX) from the PBE orbitals obtained in the above calculations. However, because the k-point meshes above are too dense for evaluation of EXX to be computationally feasible, the periodic systems were first recomputed with a coarser k-point mesh (for each system, the smallest even, Γ -centered mesh with a linear k-point density of at least 4.5 k-points per Å⁻¹), and the EXX was computed on this mesh. All nonlocal density features were evaluated on the density and orbitals obtained from the denser k-point mesh. They were evaluated with the internal CIDER settings $q_{\max} = 300$ and $\lambda = 1.8$ (see Appendix A 10 for details).

The PBE0 cohesive energies were computed using the same procedure above, with the PBE0 energies being defined as

$$E^{\text{PBE0}} = E^{\text{PBE,dense}} - E^{\text{PBE,coarse}} + E^{\text{PBE0,coarse}}, \quad (\text{A1})$$

with *dense* denoting the calculation done with the dense k-point mesh and *coarse* denoting the coarser k-point mesh discussed above.

4. Construction of training and validation sets

The total exchange energy of a molecule is highly basis set and code dependent, and it is dominated by the core contributions for large atoms. In practice, only relative exchange energies between systems is relevant for most chemistry and materials science applications. Therefore, rather than training the model to match total exchange energies, we trained to match the exchange energy differences between systems. For example, the target and predicted values for the atomization energy of water would be

$$E_x^{\text{target}} = E_x^{\text{ex}}[n_{\text{O}}] + 2E_x^{\text{ex}}[n_{\text{H}}] - E_x^{\text{ex}}[n_{\text{H}_2\text{O}}], \quad (\text{A2})$$

$$E_x^{\text{pred}} = \hat{E}_x[n_{\text{O}}] + 2\hat{E}_x[n_{\text{H}}] - \hat{E}_x[n_{\text{H}_2\text{O}}], \quad (\text{A3})$$

where $E_x^{\text{ex}}[n]$ and $\hat{E}_x[n]$ are the exact and model exchange energies for density distribution $n(\mathbf{r})$, respectively. The training and validation sets were constructed from four sources:

Molecules. The W4-11 (atomization energies), BH76 (barrier heights), BH76RC (reaction energies for BH76 reactions),

MB16-43 (“mindless benchmarking” of reaction energies), S22 (noncovalently bound dimers), G21IP (ionization potentials), PA26 (proton affinities), RG18 (noble gas dimers), ACONF (alkane conformers), ALKBDE10 (alkali/alkaline earth diatomic bond energies), HEAVYSB11 (dissociation energies for molecules with heavy atoms), and SIE4x4 (self-interaction problems) subsets from the GMTKN55 database were used for training and validation. See the original GMTKN55 paper [10] and references therein for more details on the datasets. Each set was randomly split into two-thirds training and one-third validation, except for the BH76 and BH76RC subsets. Because the data points in these subsets are based on the same systems, and also because they include the forward and backward reactions for the same chemical equation, a different splitting procedure was chosen to maintain train-validation independence. First, the 34 unique transition states in BH76 were identified and split half-and-half into the training and validation sets. All 40 reactions with the training transition states were used for training, and all 36 reactions with the validation transition states were used for validation. All 17 BH76RC reactions with systems in the BH76 training partition were used for training, and the other 13 reactions were used for validation. All data not in the training or validation set was classified as the GMTKN55 test set. This data partitioning scheme resulted in 280 train data, 163 validation data, and 1062 test data in GMTKN55.

Bulk solids. Half of the SOL62 cohesive energy database (11 randomly selected elemental metals and 20 randomly selected nonmetals) was used for training and validation. Nine systems (C-diamond, Li, GaAs, MgO, LiF, ZrC, TiC, Pd, and Pt) were hand-selected for the training set to cover a range of chemical space, while the other 22 systems were used for validation. All data not in the training or validation set was classified as the SOL62 test set.

Nanoclusters. Six nanocluster atomization energies were added to the training set as a “bridge” between the isolated and extended systems in the study. These were Ca₁₃ and Zn₂₆, with the initial structure obtained based on the Materials Project [119] bulk lattice constant and structure optimization at the PBE/def2-TZVP level in Orca [120–122]; C₃₅H₃₆ (hydrogen-capped diamond), obtained from Refs. [123,124] and also structurally optimized in Orca at the PBE/def2-TZVP level; and Na₄C₁₄, [Na₁₃Cl₁₄]⁻, and [Na₁₄Cl₁₃]⁺, all using the Materials Project bulk lattice constant without further structure optimization.

Atoms. Absolute and relative energies of the isolated atoms H-Kr and a few excited spin states were included in the training set. Specifically, the training points consisted of the absolute energies of the H atom and the noble gases He, Ne, Ar, and Kr; the relative energies of the elements with nuclear charge Z and $Z + 1$ for $Z = \{1, \dots, 35\}$, and the relative energies of the following ground state/excited state spin pairs: Sc 1/3, Ti 2/4, V 3/5, Cr 6/4, and Ni 2/0. The preceding numbers correspond to the magnetic moment $2S$.

In addition to fitting to the above training data, the uniform electron gas constraint was enforced as described in Sec. II C 3, and the uniform scaling constraint was enforced by the scale-invariance of the input features. The ground-state total energies of the SOL62 solids and isolated atoms were computed with GPAW as described in Appendix A 3.

The ground-state energies for all other training data were evaluated with PySCF as described in Appendix A 2. The selection of train and validation databases from GMTKN55 includes a wide variety of chemistries and properties. However, because training data is sampled disproportionately from small-molecule properties, there is a large distribution shift between the train/validation and test sets, with the test sets containing more large molecules, isomerization reactions, and barrier heights. The use of this train/test partition therefore serves as a strong test of transferability. The accuracy on SOL62 also strongly reflects the transferability of a model because only nine solid-state systems were included in the training set, and SOL62 covers a wide variety of chemistries.

5. A more diverse training set for numerical stability

To design a more diverse training set for the NL-MGGA-DTR functional, a kernel function was defined for the “covariance” between two training points in GMTKN55,

$$k_{\text{GMTKN}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{|\mathbf{x} - \mathbf{x}'|^2}{328}\right), \quad (\text{A4})$$

where \mathbf{x} is the vector of errors for 82 functionals benchmarked on the GMTKN55 database by Goerigk *et al.* [10] with D3(BJ) dispersion corrections [125–127], normalized so that the mean square error across all functionals for that individual data point is 1. This kernel essentially provides a larger similarity score between data points for which the (normalized) errors tend to be similar for a given functional. The kernel matrix was constructed for the entire GMTKN55 database and then factorized using a method called Cholesky decomposition with complete pivoting [128,129], which factors a given matrix \mathbf{A} as follows:

$$\mathbf{P}^T \mathbf{A} \mathbf{P} = \mathbf{L} \mathbf{L}^T. \quad (\text{A5})$$

In Eq. (A5), \mathbf{P} is a permutation matrix that orders the vectors \mathbf{x} in Eq. (A4) in decreasing order of their variance, conditioned on the preceding feature vectors, *in the space implicitly spanned by the kernel function*. Intuitively, this means that later vectors, as ordered by \mathbf{P} , are likely to be nearly linearly dependent (in the kernel space) on the earlier vectors. Therefore, by picking the first N training points as ordered by \mathbf{P} and using the remaining data for testing, we can get a training set whose errors correlate as much as possible with all the test data, effectively making the training set more diverse and potentially preventing out-of-distribution error of the ML model.

We follow this procedure and select the first 225 pivots for the training set, the next 75 pivots for the validation set, and the remaining 1205 pivots for the test set. With this dataset partitioning, more than 99% of the GMTKN55 test set data points contain at least one system not present in the train or validation sets, so the test set is still sufficiently independent to assess the transferability of the functionals to new systems. All training points not contained in GMTKN55 were kept the same as in Appendix A 4.

6. Control point selection for sparse Gaussian processes

In Eqs. (37) and (38), a set of control points $\{\tilde{\mathbf{x}}_a\}$ is required to evaluate the Gaussian process covariance kernel in the sparse approximation. These control points were initially selected by randomly sampling feature vectors from the PySCF grids for the training set (Appendix A 2). All training data was sampled except the SOL62 data, since the features for these systems were evaluated in GPAW.

Because of the size of the numerical grids from which the features were sampled, there were a large number of potential control points for the sparse Gaussian process, many of which were nearly linearly dependent in the implicit feature space of the covariance kernel function. To reduce the size of the control point set, a Cholesky decomposition with pivoting [Eq. (A5)] was performed on an initial sample size of roughly 10^4 feature vector samples with a tolerance of 10^{-5} , and only control points corresponding to significant pivot indices were kept. This reduced the number of control points from roughly 10^4 to 10^2 .

7. Selection of noise hyperparameters, training loss, and validation loss

Chemical and material properties have a wide range of magnitudes, from less than 1 kcal/mol for some van der Waals binding energies to greater than 100 kcal/mol for many covalent bond dissociation energies. Correspondingly, different magnitudes of XC errors are expected and acceptable for these properties; an error of 1 kcal/mol might make a van der Waals binding energy prediction useless but be acceptable for an atomization energy. To train the exchange functionals, it was therefore necessary to weight the noise hyperparameter of each training point to reflect the feasible and desired precision for the property of interest. To do so, we used the data provided with GMTKN55 [10] to compute the MAD of four hybrid functionals [PBE0-D3(BJ), B3LYP-D3(BJ), PW6B95-D3(BJ), and ω B97X-V] compared to the reference data on each subdatabase (denoted $\text{MAD}_{\text{DB}}^{\text{hybrid}}$, where DB is the sub-database of interest), and then created characteristic uncertainties for each sub-database as follows (in Hartree atomic units):

$$\tilde{\sigma}_{\text{DB}} = \text{MAD}_{\text{DB}}^{\text{hybrid}} \left(\frac{\sigma_0}{\text{MAD}_{\text{W4-11}}^{\text{hybrid}}} \right), \quad (\text{A6})$$

where W4-11 is a small-molecule atomization energy subdatabase in GMTKN55 and $\sigma_0 = 0.03$ Eh. One drawback of the above approach is that it favors high accuracy on systems that are already described well by conventional, dispersion-corrected DFT (like simple van der Waals dimers) and assumes a large uncertainty on systems like self-interaction-dominated bond energies, for which even hybrid functionals with small exchange mixing perform poorly. To remedy this, the noise parameter was augmented with a constant weighting for every system,

$$\sigma_{\text{DB}} = \sqrt{\frac{2}{1/\sigma_0^2 + 1/\tilde{\sigma}_{\text{DB}}^2}}. \quad (\text{A7})$$

With this definition, the matrix Σ_{noise} in Eq. (43) is a diagonal matrix with $(\Sigma_{\text{noise}})_{ii}$ being the σ_{DB}^2 for the database to which

training point i belongs. Constructing the Gaussian process for this noise parameter is equivalent to regression with the loss function

$$\mathcal{L} = \frac{1}{2} \sum_{\text{DB}} N_{\text{DB}} \text{MSD}_{\text{DB}} \left(\frac{1}{\sigma_0^2} + \frac{1}{\tilde{\sigma}_{\text{DB}}^2} \right), \quad (\text{A8})$$

where MSD_{DB} stands for the mean squared deviation of a given GMTKN55 subdatabase labeled by ‘‘DB,’’ and N_{DB} is the number of points in the subdatabase. This weighting ensures that training points with small expected errors are appropriately weighted while those with larger expected errors for conventional functionals do not become underweighted. For systems outside the GMTKN55 database that were in the training set, $\tilde{\sigma}_{\text{DB}}$ was set by hand.

8. Hyperparameter selection and validation procedure

The models tested in this paper have several discrete and continuous hyperparameters to be optimized, which are listed below.

- (i) The covariance prefactor hyperparameter Σ_{Cov} in Eqs. (56)–(59).
- (ii) The kernel length scale (l_i , where i is the feature index) in Eq. (60).
- (iii) For the nonlocal functionals only, the feature exponent parameters B_i and C_i in Eqs. (20)–(23).
- (iv) The baseline exchange functional on top of which the CIDER model is trained via Δ learning. We tested PBE [23] and the Chachiyo GGA [76].

It would be quite challenging and expensive to implement differentiation of the marginal likelihood with respect to the second parameter, and nearly impossible for the third since backpropagation would need to be performed through the feature evaluation step. Therefore, we used various combinations of hyperparameters to train the models, and then for each functional type, we selected the best-performing model on the validation set as the final candidate functional.

To obtain a discrete set of exponent hyperparameters to test for the nonlocal functionals, we developed a reduced number of parameters to specify the feature exponents. Similarly to the previous CIDER paper [35], a control parameter A tunes B_2 and C_2 simultaneously, in two different schemes. The first scheme (S1) is

$$B_2 = A, \quad C_2 = \frac{A}{32} \frac{6}{5\pi} (6\pi^2)^{2/3} \approx 0.18A. \quad (\text{A9})$$

The second scheme (S2) provides a different mixture of density-based and gradient/kinetic-energy-based terms,

$$B_2 = \frac{A}{16} \frac{6}{5\pi} (6\pi^2)^{2/3} \approx 0.36A, \quad C_2 = B_2. \quad (\text{A10})$$

In both schemes, the other constants are determined by B_2 and C_2 ,

$$B_1 = \frac{1}{2}B_2, \quad B_3 = 2B_2, \quad B_0 = \frac{D}{A}B_2, \quad (\text{A11})$$

$$C_1 = \frac{1}{2}C_2, \quad C_3 = 2C_2, \quad C_0 = \frac{D}{A}C_2, \quad (\text{A12})$$

where D is an additional parameter to tune the B_0 and C_0 coefficients separately from the other coefficients. With

TABLE VI. R_1 values for each functional type and baseline functional.

Type/baseline	R_1
NL-GGA/PBE	1.0
NL-GGA/Chachiyo	20.0
NL-MGGA/PBE	0.05
NL-MGGA/Chachiyo	1.0

these schemes, all eight parameters $\{B_i\}, \{C_i\}$ that define the three-feature CIDER model were reduced to two parameters per scheme. For each scheme, we performed a hyperparameter search over the grid $A = \{1/2, 1, 2, 4, 8\}$ and $D = \{1/2, 1, 2, 4, 8\}$, resulting in a total of 50 feature constructions to test.

Because there were already 50 features from which to select the nonlocal functionals, we set the covariance and feature length-scales for these models heuristically,

$$\Sigma_{\text{Cov}} = R_1 \langle \Delta F_x^2 \rangle_{\text{train}}, \quad (\text{A13})$$

$$l_i = R_2 \sqrt{\langle x_i^2 \rangle_{\text{train}}}, \quad (\text{A14})$$

where $\langle \rangle_{\text{train}}$ is an expectation value over all the grid points in the exchange energy training set, ΔF_x is the deviation of the exact exchange enhancement factor at a given point from the baseline functional enhancement factor, x_i is the i th regularized feature (see Sec. II C 3), and R_1 and R_2 are constants. R_1 and R_2 were set based on the criteria that the resulting functionals converge consistently in both GTO and plane-wave calculations and exhibit a small total energy deviation between the PAW and all-electron GTO atomization energies for a few small molecules. $R_2 = 1$ worked best for both the nonlocal GGA and meta-GGA, while the most practical values of R_1 depended on the baseline functional and model type, as shown in Table VI. The meta-GGA models required a smaller covariance prefactor to achieve good stability than the GGA models. We set R_1 to be larger by a factor of 20 for functionals with the Chachiyo baseline compared to the PBE baseline because $\langle \Delta F_x^2 \rangle_{\text{train}}$ was roughly 20 times larger for PBE than Chachiyo.

For the semilocal functionals, there was no need to validate over the possible feature constructions, leaving more flexibility to test R_1 and R_2 systematically. For the functionals with Chachiyo baseline, we tested $R_1 = \{1, 2, 4, 8, 16, 32, 64, 128\}$, while for the PBE baseline each R_1 was divided by 20 to give a roughly equivalent Σ_{Cov} . For the semilocal functionals, $R_2 = \{1, \frac{1}{2}, \frac{1}{4}\}$ were tested, resulting in 24 functionals to validate. Note that for NL-MGGA-DTR, we determined from validation set SCF calculations that $R_1 = 0.1$ with the PBE baseline resulted in sufficiently stable models, so this larger value was used. In addition, we did not search over feature length-scale for NL-MGGA-DTR, and instead used the features determined to be most accurate for NL-MGGA and NL-MGGA-PBE.

After the functionals were trained, we selected the top 10 functionals of each type and baseline functional, as classified by non-self-consistent performance on the validation set via Eq. (A8). To validate these functionals, the

GMTKN55 train and validation set reaction energies were predicted using self-consistent calculations with a convergence threshold of 10^{-6} Eh and a gradient threshold of 2×10^{-3} Eh. For a given functional, if any calculation failed to converge when starting from both the default initial guess of PySCF and from the PBE orbitals, the functional was deemed insufficiently numerically stable and disqualified from further consideration. The non-self-consistent cohesive energies of the SOL62 validation set were also computed. Of the functionals for which all calculations converged, the validation loss consisted of

$$\mathcal{L}_{\text{val}} = \mathcal{L}_{\text{MOL}} + \mathcal{L}_{\text{SOL}}, \quad (\text{A15})$$

$$\mathcal{L}_{\text{MOL}} = \frac{1}{2} \sum_{\text{DB}} N_{\text{DB/VAL}} \times \text{MSD}_{\text{DB/VAL}}^{\text{SCF}} \times \left(1 + \frac{\sigma_0^2}{\bar{\sigma}_{\text{DB}}^2} \right), \quad (\text{A16})$$

$$\mathcal{L}_{\text{SOL}} = \text{MSD}_{\text{SOL62/VAL22}}^{\text{NSCF}}, \quad (\text{A17})$$

where $\text{MSD}_{\text{DB/VAL}}^{\text{SCF}}$ is the self-consistent mean square deviation on the validation partition of GMTKN55 subset DB, $N_{\text{DB/VAL}}$ is the number of validation data in DB, and $\text{MSD}_{\text{SOL62/VAL22}}^{\text{NSCF}}$ is the non-self-consistent mean square deviation of the 22 validation-set cohesive energies per atom. The final functional of each type was selected based on which candidate had the lowest loss for Eq. (A15). The validation results and selected hyperparameters for the functionals are provided in Sec. S4 within the Supplemental Material [77].

9. Difficult convergence cases

The default convergence tolerance for the calculations performed with CIDER functionals was 10^{-8} Eh, except for the NL-MGGA-PBE and NL-MGGA-DTR calculations, which were converged to 10^{-9} Eh when possible. Some CIDER calculations did not converge to this tolerance within the default PySCF settings. There were two different classes of convergence problems. First, for a few systems, some CIDER calculations did not converge from the PySCF starting guess, but the convergence issues were resolved by starting from PBE ground-state orbitals. Second, for some systems (especially isolated atoms and ions, as well as alkali and alkaline earth-containing systems), a numerical stability issue was observed for some CIDER calculations, in which the energy converged to a reasonable tolerance (less than 10^{-7} Eh), but the orbital gradients oscillated around 10^{-3} Eh without converging fully. In these cases, the energy tolerance was set to 10^{-7} Eh, and the gradient tolerance was set to the (rather large) 2×10^{-3} Eh. All CIDER calculations converged to these thresholds. PBE starting orbitals were also used for these loose-threshold calculations. Even with these looser thresholds, the achieved energy and orbital gradient convergence is still sufficient to compare thermochemical data in the GMTKN55 database.

10. CIDER functional settings

The CIDER functionals require a few additional settings to specify the precision of the accelerated feature calculations. For the molecular version, these settings are l_{max} , the maximum spherical harmonic order for auxiliary basis expansions

in Sec. II E; λ , in Eq. (62); β^{aux} , in Eq. (73); and q_{max} , the maximum value of q_{α} [Eq. (62)]. The ξ basis in Eq. (76) uses the same parameter λ for the even-tempered basis as Eq. (62). We chose $l_{\text{max}} = 10$, $\lambda = \beta^{\text{aux}} = 1.6$, and $q_{\text{max}} = \frac{1000}{36} Z_{\text{max}}^2$, where Z_{max} is the minimum of 36 and the charge of the largest nucleus in the given GMTKN55 subdatabase. These settings were found to be in good agreement with numerical integration of the features. The minimum control point q_0 was set automatically based on the minimum allowed value of the kernel exponent $a(\mathbf{r})$.

For the plane-wave implementation, only $\lambda = 1.8$ and $q_{\text{max}} = 300$ needed to be specified. A larger $q_{\text{max,atom}} = 10Z^2$ was used for each atom on the radial support grids, with Z being the atomic number. In both the Gaussian-type orbital and plane-wave implementations, all other CIDER-specific parameters (like the minimum exponents for the various auxiliary bases) were set automatically based on the chosen basis set and nonlocal feature parameters (i.e., the parameters in Appendix A 8).

11. Band gap calculations

Band gaps for the database of Borlido *et al.* [45] were computed in GPAW [68,69] as the difference between the conduction band minimum and valence band maximum for a ground-state SCF calculation. The energy cutoff for these calculations was 520 eV, and the grid spacing was set to $h = 0.134$ Å to avoid any aliasing of the density. Fermi-Dirac k-point smearing was used with a width of 0.01 eV, except for the noble gas systems, for which no k-point smearing was used. The k-point mesh for each system was selected to be the smallest even, Γ -centered mesh with a linear k-point density of at least six k points per Å⁻¹. For each system, if the PBE band gap we obtained was more than 0.1 eV larger than the PBE band gap obtained by Borlido *et al.* [45], we inferred that the k-point mesh was too coarse and reran the band gap calculation with a linear k-point density of 10 k points per Å⁻¹. Using this methodology, we reproduced the PBE band gaps of Borlido *et al.* with a mean, mean absolute, and maximum error of -0.01, 0.03, and 0.17 eV, respectively, and the SCAN results with a mean, mean absolute, and maximum error of 0.00, 0.03, and 0.15 eV, respectively.

For PBE0/NL-MGGA-DTR and PBE0(0.35)/NL-MGGA-DTR, a few systems did not converge with the default CIDER PAW corrections, and a denser core integration grid and smaller set of feature projectors (i.e., the g_i^A and p_i^A functions of Sec. II F) were used. Even with these settings, one system (ReSe₂) did not converge with PBE0(0.35)/NL-MGGA-DTR, and it had to be run with SG15 norm-conserving pseudopotentials (NCP) [130]. For the 255 calculations in the main text, the mean, mean absolute, and maximum deviations between PBE0/NL-MGGA-DTR with NCP and PAW were 0.03, 0.12, and 1.12 eV, respectively. In addition, the band gap for ReSe₂ with PBE0/NL-MGGA-DTR was 1.29 eV with PAW and 1.30 eV with NCP, so the use of pseudopotentials appears reasonable for this particular case. We think that the convergence issues are caused by numerical problems with the PAW corrections for nonlocal features, and improving the numerical stability of these corrections will be a subject of future work.

For the band gaps in Table IV, the same procedure was used, except with denser k-point meshes containing 12 points per \AA^{-3} . The HSE06 band gaps were computed using VASP [97], with the k-point mesh set to contain 4000 k points divided by the number of atoms in the unit cell. This is a coarser mesh than was used for the semilocal and CIDER calculations, but we found that the band gap difference compared to larger k-point densities was less than 0.01 eV.

12. Supercell and neutral defect performance benchmark

For the neutral defect calculations, the atomic structure of the 215-atom single vacancy cells of diamond and silicon were obtained using VASP [97–99] with PAW datasets [67,74], an energy cutoff of 520 eV, the “Accurate” precision setting, and a $2 \times 2 \times 2$ Monkhorst-Pack k-point mesh [96], starting with the PBE lattice constant obtained in VASP using the same settings with a $14 \times 14 \times 14$ Γ -centered k-point mesh. Using these structures, SCF ground-state calculations with static atomic positions were performed in GPAW [68,69] for the 216-atom bulk and 215-atom vacancy cells for each functional. The formation energy of the neutral defect was computed as

$$E^{\text{formen}} = E^{\text{defect}} - \frac{215}{216} E^{\text{bulk}}, \quad (\text{A18})$$

where the bulk and defect energies were computed with 216 and 215-atom supercells, respectively, with a $2 \times 2 \times 2$ Monkhorst-Pack k-point mesh (resulting in one reduced k-point for the bulk supercell and four reduced k-points for the defect supercell), 520 eV plane-wave cutoff, 0.134 \AA grid spacing, and 10^{-5} eV convergence threshold. 523 bands were computed for the bulk, and 520 bands were computed for the defect cell. The time to evaluate the formation energy for each functional was computed as the sum of the time taken for the bulk and vacancy cell calculations, as obtained from the GPAW internal calculation timer. The HSE06 and PBE0 formation energies were obtained using VASP with the same structures, k-point grid, and energy cutoff.

For the Γ -point calculations used in Fig. 8(a), GPAW was also used for the semilocal and ML functionals. For the hybrid DFT calculations, Quantum ESPRESSO [94,95] was used with the adaptively compressed exchange [17] algorithm for exchange, a 38-Rydberg plane-wave cutoff for the wave functions, a 152-Rydberg cutoff for the charge density and potential, a 10^{-6} Rydberg convergence threshold, and SG15 norm-conserving pseudopotentials [130]. The use of a 38-Rydberg (517 eV) plane-wave cutoff might be too small when using norm-conserving pseudopotentials, but for performance benchmarks, it provides the most fair comparison to the 520 eV cutoff used with the PAW setups in GPAW. The PBE calculation with Quantum ESPRESSO used the same settings but with PBE in place of PBE0. The GPAW calculations used 523 bands, and the Quantum ESPRESSO calculations used 432 bands.

13. Charged defect calculations

The charged defect levels were computed with PBE, PBE0/NL-MGGA-DTR, and PBE0(0.3)/NL-MGGA-DTR. The mixing parameter $\alpha = 0.3$ of the last functional was

selected to match the zero-temperature experimental band gap of silicon, 1.17 eV. For each functional, the lattice constant, band edges, and band gap were obtained from a unit-cell relaxation with 520 eV cutoff, $h = 0.134$ \AA , and a $24 \times 24 \times 24$ k-point mesh in GPAW [68,69]. Then, a 512-atom supercell was created for each defect, with perturbed structures used for the vacancy to break symmetry. For each charge state and defect, the geometry was relaxed in GPAW using the BFGS with Line Search algorithm to a tolerance of 0.05 eV/ \AA with a Γ -point only calculation. This geometry was then used as the starting point for the final geometry relaxation, which was performed with a $2 \times 2 \times 2$ Γ -centered k-point mesh and a force convergence tolerance of 0.01 eV/ \AA . Because PBE0(0.3)/NL-MGGA-DTR is so similar to PBE0/NL-MGGA-DTR, and because the lattice constants are similar, we saved time for this functional by starting the final relaxation from the fractional coordinates of the PBE0/NL-MGGA-DTR calculations.

The formation energies of the defects were calculated as [6]

$$E^f[X^q] = E_{\text{tot}}[X^q] - E_{\text{tot}}[\text{bulk}] - \sum_i n_i \mu_i + qE_F + E_{\text{corr}}[X^q], \quad (\text{A19})$$

where E_{tot} is the total ground-state energy of the supercell DFT calculation, X^q is defect X in charge state q , E_F is the Fermi level, and $E_{\text{corr}}[X^q]$ is a finite-size energy correction for the defect. The stoichiometry changes and chemical potentials n_i and μ_i were neglected because they do not impact electronic transition levels. We used the Freysoldt finite-size correction-scheme for $E_{\text{corr}}[X^q]$ [100], which combines an electrostatic and potential alignment correction. This correction requires an estimate of the dielectric constant ϵ , for which we used the experimental value of 11.9 [131].

Because some of the defects studied in this paper are shallow donors and acceptors, and because electrostatic corrections like those developed by Freysoldt *et al.* [100] and Kumagai *et al.* [132] assume localized charges, these corrections might be inaccurate. Therefore, for comparison, we also computed the transition levels using the potential alignment correction only, i.e.,

$$E_{\text{corr}}[X^q] = q(\bar{V}_{\text{esp}}^\infty[X^q] - \bar{V}_{\text{esp}}^\infty[\text{bulk}]), \quad (\text{A20})$$

where $\bar{V}_{\text{esp}}^\infty$ is the average electrostatic potential in a region far from the defect site. This is equivalent to the full Freysoldt correction with $\epsilon \rightarrow \infty$. We used pymatgen [111] to perform the corrections.

APPENDIX B: CONSTRUCTION OF FEATURE PROJECTOR FUNCTIONS FOR CIDER WITHIN PAW

1. Construction of feature smoothing projectors

We solve for the coefficients $D_{j\beta}$ and $C_{i\alpha}$ in Eqs. (87) and (88) by minimizing the loss function

$$\mathcal{L} = \sum_{\beta} w_{\beta} \int d^3\mathbf{r} [\hat{y}_{A,\beta}(\mathbf{r}) - \Delta F_{A,\beta}(\mathbf{r})]^2. \quad (\text{B1})$$

For the current implementation, $w_\beta = 1$, but this could be changed if needed to encourage better smoothness or numerical precision. We start with the following set of functions:

$$g_{i \leftarrow nlm}(\mathbf{r}) = \begin{cases} Y_{lm}(\hat{\mathbf{r}})P(2n+l, r/r_c) & r < r_c \\ 0 & r \geq r_c \end{cases}, \quad (\text{B2})$$

$$P(n, x) = \begin{cases} \frac{1}{2}(1 + \cos \pi x) & n = 0 \\ Q(n-1, x) & n \in \mathbb{O}, \\ R(n-2, x) & n \in \mathbb{E} \end{cases}, \quad (\text{B3})$$

$$Q(m, x) = 5 \left(\sin \frac{\pi x}{2} - \frac{1}{2} + \frac{1}{2} \cos \pi x \right) x^m e^{-mx^2}, \quad (\text{B4})$$

$$R(m, x) = \frac{1}{2} (1 - \cos(2\pi x)) x^m e^{-mx^2}, \quad (\text{B5})$$

where \mathbb{O} and \mathbb{E} are the sets of positive odd and even integers, respectively. Then we introduce the matrix \mathbf{P}

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}^{11} & \mathbf{P}^{12} \\ (\mathbf{P}^{12})^\top & \mathbf{P}^{22} \end{pmatrix}. \quad (\text{B6})$$

In the above equation, \mathbf{P}^{11} is an $N_h N_a \times N_h N_a$ matrix,

$$P_{j\beta', j'\beta'}^{11} = \delta_{\beta\beta'} \langle h_j^A | h_{j'} \rangle + \delta_{jj'} \delta_{\beta\beta'} \sigma_{11}, \quad (\text{B7})$$

where N_h is the number of h_j^A functions and N_a is the number of control points for kernel interpolation. Similarly, \mathbf{P}^{22} is an $N_g N_a \times N_g N_a$ matrix,

$$P_{i\alpha', i'\alpha'}^{22} = \sum_\beta w_\beta I_{\beta, i\alpha', i'\alpha'} + \delta_{ii'} \delta_{\alpha\alpha'} \sigma'_{22}, \quad (\text{B8})$$

$$I_{\beta, i\alpha', i'\alpha'} = \int d^3 \mathbf{r} \phi_{i\alpha\beta}(\mathbf{r}) \phi_{i'\alpha'\beta}(\mathbf{r}) (1 + \sigma_{22} f_{\text{cut}}(\mathbf{r})) \quad (\text{B9})$$

$$\phi_{i\alpha\beta}(\mathbf{r}) = \int d^3 \mathbf{r}' \Phi_{\alpha\beta}(\mathbf{r}' - \mathbf{r}) g_i^A(\mathbf{r}'), \quad (\text{B10})$$

where N_g is the number of g_i^A functions. The constants σ_{11} , σ_{22} , and σ'_{22} are regularization constants to reduce the magnitude of the feature augmentation correction in the core region, which helps with numerical stability by preventing the pseudo-feature from being unrealistically large in the core.

Lastly, \mathbf{P}^{12} is an $N_h N_a \times N_g N_a$ matrix,

$$P_{j\beta, i\alpha}^{12} = w_\beta \int d^3 \mathbf{r} h_j(\mathbf{r}) \phi_{i\alpha\beta}(\mathbf{r}). \quad (\text{B11})$$

The coefficients can be solved by minimizing the loss Eq. (B1), which results in the linear system

$$\begin{pmatrix} \mathbf{D} \\ \mathbf{C} \end{pmatrix} = \begin{pmatrix} \mathbf{P}^{11} & \mathbf{P}^{12} \\ (\mathbf{P}^{12})^\top & \mathbf{P}^{22} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{b}^1 \\ \mathbf{b}^2 \end{pmatrix}, \quad (\text{B12})$$

where

$$b_{j\beta}^1 = w_\beta \int d^3 \mathbf{k} \Delta F_{A,\beta}(\mathbf{k}) h_j^A(\mathbf{k}), \quad (\text{B13})$$

$$b_{i\alpha}^2 = \sum_\beta w_\beta \int d^3 \mathbf{k} \Delta F_{A,\beta}(\mathbf{k}) \phi_{i\alpha\beta}(\mathbf{k}). \quad (\text{B14})$$

The matrix \mathbf{P} is block diagonal in the angular momentum channels because of the orthogonality of spherical harmonics, so each angular momentum channel can be solved separately. This makes the whole routine fairly efficient, and the number of operations per SCF cycle is linear-scaling in the number of atoms. The Fourier transforms of the g_i^A and h_j^A functions needed for Eqs. (B13) and (B14) are obtained using the NUMSBT algorithm [133], and the $\Delta F_{A,\beta}(\mathbf{k})$ is decomposed into angular momentum channels via the $\Delta F_{A,L,\beta}(k)$ term obtained in Eq. (85). Also, note that there are a larger number of α and β indices in the atomic augmentation regions than on the FFT grid. The above fitting procedure is only used for indices going up to the maximum α index on the FFT grid. The higher indices (corresponding to larger values of q_α that are only relevant in the core regions) are solved by simply projecting onto the $h_j^A(\mathbf{r})$ basis, i.e., setting $\mathbf{D} = (\mathbf{P}^{11})^{-1} \mathbf{b}^1$ in the subspace of core region-only α indices. This procedure is equivalent to assuming that the larger q_α contributions to θ_α and F_β are localized in the core region, which is reasonable because these contributions correspond to high densities and short length-scales for $\Phi_{\alpha\beta}(r)$ [Eq. (19)].

2. Construction of the feature augmentation projectors

The feature augmentation projectors used in Eq. (92) are constructed as

$$p_{i \leftarrow nlm}(\mathbf{r}) = Y_{lm}(\hat{\mathbf{r}}) P(2n+l, r/r_0), \quad (\text{B15})$$

$$P(n, x) = \begin{cases} x^n (x^2 - 1)^2 & x < 1 \\ 0 & x \geq 1 \end{cases}. \quad (\text{B16})$$

The on-site function sets $\tilde{f}_i^A(\mathbf{r})$ and $\tilde{p}_j^A(\mathbf{r})$ are then computed as

$$\tilde{p}_i(\mathbf{r}) = p_i(\mathbf{r}) f_{\text{cut}}(\mathbf{r}), \quad (\text{B17})$$

$$\tilde{f}_i(\mathbf{r}) = \sum_j (\mathbf{S}^{-1})_{ij} p_j(\mathbf{r}), \quad (\text{B18})$$

$$S_{ij} = \int d^3 \mathbf{r} p_i(\mathbf{r}) p_j(\mathbf{r}) f_{\text{cut}}(\mathbf{r}), \quad (\text{B19})$$

$$f_{\text{cut}}(\mathbf{r}) = \begin{cases} \frac{1}{2} (1 + \cos \frac{\pi r}{r_c}) & r < r_c \\ 0 & r \geq r_c \end{cases}. \quad (\text{B20})$$

The construction localizes $\tilde{p}_j^A(\mathbf{r})$ inside the core region and satisfies the condition that $\langle \tilde{f}_i^A | \tilde{p}_j^A \rangle = \delta_{ij}$.

- [1] N. Mardirossian and M. Head-Gordon, Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals, *Mol. Phys.* **115**, 2315 (2017).
 [2] W. Kohn and L. J. Sham, Self-consistent equations including exchange and correlation effects, *Phys. Rev.* **140**, A1133 (1965).

- [3] J. Wellendorff, T. L. Silbaugh, D. Garcia-Pintos, J. K. Nørskov, T. Bligaard, F. Studt, and C. T. Campbell, A benchmark database for adsorption bond energies to transition metal surfaces and comparison to selected DFT functionals, *Surf. Sci.* **640**, 36 (2015).
 [4] Q. He, B. Yu, Z. Li, and Y. Zhao, Density functional theory for battery materials, *Energy Environ. Mater.* **2**, 264 (2019).

- [5] S. Lany and A. Zunger, Assessment of correction methods for the band-gap problem and for finite-size effects in supercell defect calculations: Case studies for ZnO and GaAs, *Phys. Rev. B* **78**, 235104 (2008).
- [6] C. Freysoldt, B. Grabowski, T. Hickel, J. Neugebauer, G. Kresse, A. Janotti, and C. G. Van de Walle, First-principles calculations for point defects in solids, *Rev. Mod. Phys.* **86**, 253 (2014).
- [7] J. P. Perdew, Jacob's ladder of density functional approximations for the exchange-correlation energy, *AIP Conf. Proc.* **577**, 1 (2001).
- [8] J. P. Perdew and M. Levy, Physical content of the exact Kohn-Sham orbital energies: Band gaps and derivative discontinuities, *Phys. Rev. Lett.* **51**, 1884 (1983).
- [9] R. W. Godby, M. Schlüter, and L. J. Sham, Accurate exchange-correlation potential for silicon and its discontinuity on addition of an electron, *Phys. Rev. Lett.* **56**, 2415 (1986).
- [10] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, and S. Grimme, A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions, *Phys. Chem. Chem. Phys.* **19**, 32184 (2017).
- [11] Z.-h. Yang, H. Peng, J. Sun, and J. P. Perdew, More realistic band gaps from meta-generalized gradient approximations: Only in a generalized Kohn-Sham scheme, *Phys. Rev. B* **93**, 205205 (2016).
- [12] T. Aschebrock and S. Kümmel, Ultranonlocality and accurate band gaps from a meta-generalized gradient approximation, *Phys. Rev. Res.* **1**, 033082 (2019).
- [13] P. Kovács, F. Tran, P. Blaha, and G. K. H. Madsen, What is the optimal mGGA exchange functional for solids? *J. Chem. Phys.* **157**, 094110 (2022).
- [14] T. Lebeda, T. Aschebrock, and S. Kümmel, First steps towards achieving both ultranonlocality and a reliable description of electronic binding in a meta-generalized gradient approximation, *Phys. Rev. Res.* **4**, 023061 (2022).
- [15] A. D. Becke, Density-functional thermochemistry. III. The role of exact exchange, *J. Chem. Phys.* **98**, 5648 (1993).
- [16] B. G. Janesko, T. M. Henderson, and G. E. Scuseria, Screened hybrid density functionals for solid-state chemistry and physics, *Phys. Chem. Chem. Phys.* **11**, 443 (2009).
- [17] L. Lin, Adaptively compressed exchange operator, *J. Chem. Theory Comput.* **12**, 2242 (2016).
- [18] W. Hu, L. Lin, A. S. Banerjee, E. Vecharynski, and C. Yang, Adaptively compressed exchange operator for large-scale hybrid density functional calculations with applications to the adsorption of water on silicene, *J. Chem. Theory Comput.* **13**, 1188 (2017).
- [19] I. Carnimeo, S. Baroni, and P. Giannozzi, Fast hybrid density-functional computations using plane-wave basis sets, *Electron. Struct.* **1**, 015009 (2019).
- [20] J. Vinson, Faster exact exchange in periodic systems using single-precision arithmetic, *J. Chem. Phys.* **153**, 204106 (2020).
- [21] H.-Y. Ko, J. Jia, B. Santra, X. Wu, R. Car, and R. A. DiStasio Jr., Enabling large-scale condensed-phase hybrid density functional theory based *ab initio* molecular dynamics. 1. Theory, algorithm, and performance, *J. Chem. Theory Comput.* **16**, 3757 (2020).
- [22] J. Lee, A. Rettig, X. Feng, E. Epifanovsky, and M. Head-Gordon, Faster exact exchange for solids via occ-RI-K: Application to combinatorially optimized range-separated hybrid functionals for simple solids with pseudopotentials near the basis set limit, *J. Chem. Theory Comput.* **18**, 7336 (2022).
- [23] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized gradient approximation made simple, *Phys. Rev. Lett.* **77**, 3865 (1996).
- [24] J. Sun, A. Ruzsinszky, and J. P. Perdew, Strongly constrained and appropriately normed semilocal density functional, *Phys. Rev. Lett.* **115**, 036402 (2015).
- [25] J. P. Perdew, A. Ruzsinszky, J. Tao, V. N. Staroverov, G. E. Scuseria, and G. I. Csonka, Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits, *J. Chem. Phys.* **123**, 062201 (2005).
- [26] A. D. Kaplan, M. Levy, and J. P. Perdew, The Predictive power of exact constraints and appropriate norms in density functional theory, *Annu. Rev. Phys. Chem.* **74**, 193 (2023).
- [27] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, *Ab initio* calculation of vibrational absorption and circular dichroism spectra using density functional force fields, *J. Phys. Chem.* **98**, 11623 (1994).
- [28] N. Mardirossian and M. Head-Gordon, ω B97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy, *Phys. Chem. Chem. Phys.* **16**, 9904 (2014).
- [29] S. Dick and M. Fernandez-Serra, Highly accurate and constrained density functional obtained with differentiable programming, *Phys. Rev. B* **104**, L161109 (2021).
- [30] R. Nagai, R. Akashi, and O. Sugino, Completing density functional theory by machine learning hidden messages from molecules, *npj Comput. Mater.* **6**, 43 (2020).
- [31] M. F. Kasim and S. M. Vinko, Learning the exchange-correlation functional from nature with fully differentiable density functional theory, *Phys. Rev. Lett.* **127**, 126403 (2021).
- [32] R. Nagai, R. Akashi, and O. Sugino, Machine-learning-based exchange correlation functional with physical asymptotic constraints, *Phys. Rev. Res.* **4**, 013106 (2022).
- [33] K. Trepte and J. Voss, Data-driven and constrained optimization of semi-local exchange and nonlocal correlation functionals for materials and surface chemistry, *J. Comput. Chem.* **43**, 1104 (2022).
- [34] Z. M. Sparrow, B. G. Ernst, T. K. Quady, and R. A. DiStasio, Uniting nonempirical and empirical density functional approximation strategies using constraint-based regularization, *J. Phys. Chem. Lett.* **13**, 6896 (2022).
- [35] K. Bystrom and B. Kozinsky, CIDER: An expressive, nonlocal feature set for machine learning density functionals with exact constraints, *J. Chem. Theory Comput.* **18**, 2180 (2022).
- [36] J. Kirkpatrick, B. McMorro, D. H. P. Turban, A. L. Gaunt, J. S. Spencer, A. G. D. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau *et al.*, Pushing the frontiers of density functionals by solving the fractional electron problem, *Science* **374**, 1385 (2021).
- [37] Y. Chen, L. Zhang, H. Wang, and W. Weinan, DeePKS: A comprehensive data-driven approach toward chemically

- accurate density functional theory, *J. Chem. Theory Comput.* **17**, 170 (2021).
- [38] W. Li, Q. Ou, Y. Chen, Y. Cao, R. Liu, C. Zhang, D. Zheng, C. Cai, X. Wu, H. Wang *et al.*, DeePKS + ABACUS as a bridge between expensive quantum mechanical models and machine learning potentials, *J. Phys. Chem. A* **126**, 2c05000 (2022).
- [39] J. Wellendorff, K. T. Lundgaard, A. Møgelhøj, V. Petzold, D. D. Landis, J. K. Nørskov, T. Bligaard, and K. W. Jacobsen, Density functionals for surface science: Exchange-correlation model development with Bayesian error estimation, *Phys. Rev. B* **85**, 235149 (2012).
- [40] J. Wellendorff, K. T. Lundgaard, K. W. Jacobsen, and T. Bligaard, mBEEF: An accurate semi-local Bayesian error estimation density functional, *J. Chem. Phys.* **140**, 144107 (2014).
- [41] K. T. Lundgaard, J. Wellendorff, J. Voss, K. W. Jacobsen, and T. Bligaard, mBEEF-vdW: Robust fitting of error estimation density functionals, *Phys. Rev. B* **93**, 235162 (2016).
- [42] S. Riemelmoser, C. Verdi, M. Kaltak, and G. Kresse, Machine learning density functionals from the random-phase approximation, *J. Chem. Theory Comput.* **19**, 7287 (2023).
- [43] M. G. Mavros, T. Tsuchimochi, T. Kowalczyk, A. McIsaac, L.-P. Wang, and T. V. Voorhis, What can density functional theory tell us about artificial catalytic water splitting? *Inorg. Chem.* **53**, 6386 (2014).
- [44] V. Shukla, Y. Jiao, J. H. Lee, E. Schröder, J. B. Neaton, and P. Hyldgaard, Accurate nonempirical range-separated hybrid van der Waals density functional for complex molecular problems, solids, and surfaces, *Phys. Rev. X* **12**, 041003 (2022).
- [45] P. Borlido, T. Aull, A. W. Huran, F. Tran, M. A. L. Marques, and S. Botti, Large-scale benchmark of exchange–correlation functionals for the determination of electronic band gaps of solids, *J. Chem. Theory Comput.* **15**, 5069 (2019).
- [46] X. Lei and A. J. Medford, Design and analysis of machine learning exchange-correlation functionals via rotationally invariant convolutional descriptors, *Phys. Rev. Mater.* **3**, 063801 (2019).
- [47] M. Levy and J. P. Perdew, Hellmann-Feynman, virial, and scaling requisites for the exact universal density functionals. Shape of the correlation potential and diamagnetic susceptibility for atoms, *Phys. Rev. A* **32**, 2010 (1985).
- [48] G. L. Oliver and J. P. Perdew, Spin-density gradient expansion for the kinetic energy, *Phys. Rev. A* **20**, 397 (1979).
- [49] P. A. M. Dirac, Note on exchange phenomena in the Thomas atom, *Math. Proc. Camb. Phil. Soc.* **26**, 376 (1930).
- [50] R. M. Dreizler and E. K. U. Gross, The Kohn-Sham scheme, *Density Functional Theory an Approach to Quantum Many-Body Problem* (Springer, Berlin, 1990) pp. 43–74.
- [51] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, MA, 2006).
- [52] A. D. Becke, On the large-gradient behavior of the density functional exchange energy, *J. Chem. Phys.* **85**, 7184 (1986).
- [53] N. Mardirossian and M. Head-Gordon, Mapping the genome of meta-generalized gradient approximation density functionals: The search for B97M-V, *J. Chem. Phys.* **142**, 074111 (2015).
- [54] N. Mardirossian and M. Head-Gordon, ω B97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation, *J. Chem. Phys.* **144**, 214110 (2016).
- [55] D. Duvenaud, H. Nickisch, and C. E. Rasmussen, Additive Gaussian processes, *Adv. Neural Inf. Process. Syst.* **24**, 226 (2011).
- [56] A. Glielmo, C. Zeni, and A. De Vita, Efficient nonparametric n -body force fields from machine learning, *Phys. Rev. B* **97**, 184307 (2018).
- [57] J. Vandermause, S. B. Torrisi, S. Batzner, Y. Xie, L. Sun, A. M. Kolpak, and B. Kozinsky, On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events, *npj Comput. Mater.* **6**, 20 (2020).
- [58] Y. Xie, J. Vandermause, L. Sun, A. Cepellotti, and B. Kozinsky, Bayesian force fields from active learning for simulation of inter-dimensional transformation of stanene, *npj Comput. Mater.* **7**, 40 (2021).
- [59] G. Román-Pérez and J. M. Soler, Efficient implementation of a van der Waals density functional: Application to double-wall carbon nanotubes, *Phys. Rev. Lett.* **103**, 096102 (2009).
- [60] O. A. Vydrov and T. Van Voorhis, Nonlocal van der Waals density functional: The simpler the better, *J. Chem. Phys.* **133**, 244103 (2010).
- [61] Q. Sun, T. C. Berkelbach, N. S. Blunt, G. H. Booth, S. Guo, Z. Li, J. Liu, J. D. McClain, E. R. Sayfutyarova, S. Sharma *et al.*, PySCF: the Python-based simulations of chemistry framework, *WIREs Comput. Mol. Sci.* **8**, e1340 (2018).
- [62] Q. Sun, X. Zhang, S. Banerjee, P. Bao, M. Barbry, N. S. Blunt, N. A. Bogdanov, G. H. Booth, J. Chen, Z.-H. Cui *et al.*, Recent developments in the PySCF program package, *J. Chem. Phys.* **153**, 024109 (2020).
- [63] M. Franchini, P. H. T. Philipsen, E. van Lenthe, and L. Visscher, Accurate coulomb potentials for periodic and molecular systems through density fitting, *J. Chem. Theory Comput.* **10**, 1994 (2014).
- [64] A. D. Becke, A multicenter numerical integration scheme for polyatomic molecules, *J. Chem. Phys.* **88**, 2547 (1988).
- [65] V. I. Lebedev and D. N. Laikov, A quadrature formula for the sphere of 131st algebraic order of accuracy, *Dokl. Akad. Nauk* **59**, 477 (1999).
- [66] B. G. Johnson, P. M. W. Gill, and J. A. Pople, The performance of a family of density functional methods, *J. Chem. Phys.* **98**, 5612 (1993).
- [67] P. E. Blöchl, Projector augmented-wave method, *Phys. Rev. B* **50**, 17953 (1994).
- [68] J. J. Mortensen, L. B. Hansen, and K. W. Jacobsen, Real-space grid implementation of the projector augmented wave method, *Phys. Rev. B* **71**, 035109 (2005).
- [69] J. Enkovaara, C. Rostgaard, J. J. Mortensen, J. Chen, M. Dułak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. A. Hansen *et al.*, Electronic structure calculations with GPAW: A real-space implementation of the projector augmented-wave method, *J. Phys.: Condens. Matter* **22**, 253202 (2010).
- [70] J. Klimeš, D. R. Bowler, and A. Michaelides, van der Waals density functionals applied to solids, *Phys. Rev. B* **83**, 195131 (2011).
- [71] J. Lehtomäki, I. Makkonen, M. A. Caro, A. Harju, and O. Lopez-Acevedo, Orbital-free density functional theory implementation with the projector augmented-wave method, *J. Chem. Phys.* **141**, 234102 (2014).
- [72] Q. Xu, C. Ma, W. Mi, Y. Wang, and Y. Ma, Nonlocal pseudopotential energy density functional for orbital-free density functional theory, *Nat. Commun.* **13**, 1385 (2022).

- [73] W. C. Witt, B. G. Del Rio, J. M. Dieterich, and E. A. Carter, Orbital-free density functional theory for materials research, *J. Mater. Res.* **33**, 777 (2018).
- [74] G. Kresse and D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, *Phys. Rev. B* **59**, 1758 (1999).
- [75] G.-X. Zhang, A. M. Reilly, A. Tkatchenko, and M. Scheffler, Performance of various density-functional approximations for cohesive properties of 64 bulk solids, *New J. Phys.* **20**, 063020 (2018).
- [76] T. Chachiyo and H. Chachiyo, Simple and accurate exchange energy for density functional theory, *Molecules* **25**, 3485 (2020).
- [77] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevB.110.075130> for benchmarks of additional properties and more details on the main text results and on the training and validation procedure.
- [78] C. Adamo and V. Barone, Toward reliable density functional methods without adjustable parameters: The PBE0 model, *J. Chem. Phys.* **110**, 6158 (1999).
- [79] E. Caldeweyher, S. Ehlert, A. Hansen, H. Neugebauer, S. Spicher, C. Bannwarth, and S. Grimme, A generally applicable atomic-charge dependent London dispersion correction, *J. Chem. Phys.* **150**, 154122 (2019).
- [80] J. W. Furness, A. D. Kaplan, J. Ning, J. P. Perdew, and J. Sun, Accurate and numerically efficient r^2 SCAN meta-generalized gradient approximation, *J. Phys. Chem. Lett.* **11**, 8208 (2020).
- [81] P. Mori-Sánchez and A. J. Cohen, The derivative discontinuity of the exchange–correlation functional, *Phys. Chem. Chem. Phys.* **16**, 14378 (2014).
- [82] A. J. Cohen, P. Mori-Sánchez, and W. Yang, Insights into current limitations of density functional theory, *Science* **321**, 792 (2008).
- [83] Y. Zhao and D. G. Truhlar, Design of density functionals that are broadly accurate for thermochemistry, thermochemical kinetics, and nonbonded interactions, *J. Phys. Chem. A* **109**, 5656 (2005).
- [84] Y. Zhang and W. Yang, Comment on “Generalized gradient approximation made simple”, *Phys. Rev. Lett.* **80**, 890 (1998).
- [85] L. J. Sham and M. Schlüter, Density-functional theory of the energy gap, *Phys. Rev. Lett.* **51**, 1888 (1983).
- [86] J. P. Perdew, Density functional theory and the band gap problem, *Int. J. Quantum Chem.* **30**, 451 (1986).
- [87] P. A. Schultz, Theory of defect levels and the “band gap problem” in silicon, *Phys. Rev. Lett.* **96**, 246401 (2006).
- [88] J. P. Perdew, W. Yang, K. Burke, Z. Yang, E. K. U. Gross, M. Scheffler, G. E. Scuseria, T. M. Henderson, I. Y. Zhang, A. Ruzsinszky *et al.*, Understanding band gaps of solids in generalized Kohn–Sham theory, *Proc. Natl. Acad. Sci. USA* **114**, 2801 (2017).
- [89] P. Borlido, J. Schmidt, A. W. Huran, F. Tran, M. A. L. Marques, and S. Botti, Exchange-correlation functionals for band gaps of solids: Benchmark, reparametrization and machine learning, *npj Comput. Mater.* **6**, 96 (2020).
- [90] F. Tran and P. Blaha, Accurate band gaps of semiconductors and insulators with a semilocal exchange–correlation potential, *Phys. Rev. Lett.* **102**, 226401 (2009).
- [91] J. Heyd, G. E. Scuseria, and M. Ernzerhof, Hybrid functionals based on a screened Coulomb potential, *J. Chem. Phys.* **118**, 8207 (2003).
- [92] A. V. Kruckau, O. A. Vydrov, A. F. Izmaylov, and G. E. Scuseria, Influence of the exchange screening parameter on the performance of screened hybrid functionals, *J. Chem. Phys.* **125**, 224106 (2006).
- [93] J. E. Moussa, P. A. Schultz, and J. R. Chelikowsky, Analysis of the Heyd-Scuseria-Ernzerhof density functional parameter space, *J. Chem. Phys.* **136**, 204117 (2012).
- [94] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo *et al.*, Quantum espresso: A modular and open-source software project for quantum simulations of materials, *J. Phys.: Condens. Matter* **21**, 395502 (2009).
- [95] P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. B. Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni *et al.*, Advanced capabilities for materials modelling with quantum espresso, *J. Phys.: Condens. Matter* **29**, 465901 (2017).
- [96] H. J. Monkhorst and J. D. Pack, Special points for Brillouin-zone integrations, *Phys. Rev. B* **13**, 5188 (1976).
- [97] G. Kresse and J. Hafner, *Ab initio* molecular dynamics for liquid metals, *Phys. Rev. B* **47**, 558 (1993).
- [98] G. Kresse and J. Furthmüller, Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set, *Phys. Rev. B* **54**, 11169 (1996).
- [99] G. Kresse and J. Furthmüller, Efficiency of *ab-initio* total energy calculations for metals and semiconductors using a plane-wave basis set, *Comput. Mater. Sci.* **6**, 15 (1996).
- [100] C. Freysoldt, J. Neugebauer, and C. G. Van de Walle, Fully *ab initio* finite-size corrections for charged-defect supercell calculations, *Phys. Rev. Lett.* **102**, 016402 (2009).
- [101] G. D. Watkins and J. R. Troxell, Negative-U properties for point defects in silicon, *Phys. Rev. Lett.* **44**, 593 (1980).
- [102] D. W. Fischer and J. J. Rome, Additional structure in infrared excitation spectra of group-III acceptors in silicon, *Phys. Rev. B* **27**, 4826 (1983).
- [103] C. Jagannath, Z. W. Grabowski, and A. K. Ramdas, Linewidths of the electronic excitation spectra of donors in silicon, *Phys. Rev. B* **23**, 2082 (1981).
- [104] S. Knack, J. Weber, H. Lemke, and H. Riemann, Copper-hydrogen complexes in silicon, *Phys. Rev. B* **65**, 165203 (2002).
- [105] M. Schulz, T. Dalibor, W. Martienssen, H. Landolt, and R. Börnstein, *Impurities Defects Gr. IV Elem. IV-IV III-V Compd. Part a Gr. IV Elem.*, edited by O. Madelung, U. Rössler, and M. Schulz, Landolt-Börnstein - Group III Condensed Matter Vol. 41A2a (Springer-Verlag, Berlin, 2002).
- [106] P. Deák, B. Aradi, T. Frauenheim, E. Jánzén, and A. Gali, Accurate defect levels obtained from the HSE06 range-separated hybrid functional, *Phys. Rev. B* **81**, 153203 (2010).
- [107] A. Sharan, Z. Gui, and A. Janotti, Hybrid-functional calculations of the copper impurity in silicon, *Phys. Rev. Appl.* **8**, 024023 (2017).
- [108] D. Broberg, K. Bystrom, S. Srivastava, D. Dahliah, B. A. D. Williamson, L. Weston, D. O. Scanlon, G.-M. Rignanese, S. Dwaraknath, J. Varley *et al.*, High-throughput calculations of charged point defect properties with semi-local

- density functional theory–performance benchmarks for materials screening applications, *npj Comput. Mater.* **9**, 72 (2023).
- [109] K. Berland, V. R. Cooper, K. Lee, E. Schröder, T. Thonhauser, P. Hyldgaard, and B. I. Lundqvist, van der Waals forces in density functional theory: A review of the vdW-DF method, *Rep. Prog. Phys.* **78**, 066501 (2015).
- [110] <https://github.com/mir-group/CiderPressLite>.
- [111] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis, *Comput. Mater. Sci.* **68**, 314 (2013).
- [112] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus *et al.*, The atomic simulation environment—a Python library for working with atoms, *J. Phys.: Condens. Matter* **29**, 273002 (2017).
- [113] A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G. M. Rignanese, G. Hautier *et al.*, Fireworks: A dynamic workflow system designed for high-throughput applications, *Concurr. Comput.* **27**, 5037 (2015).
- [114] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.* **12**, 2825 (2011).
- [115] F. Weigend, F. Furche, and R. Ahlrichs, Gaussian basis sets of quadruple zeta valence quality for atoms H–Kr, *J. Chem. Phys.* **119**, 12753 (2003).
- [116] D. Rappoport and F. Furche, Property-optimized Gaussian basis sets for molecular response calculations, *J. Chem. Phys.* **133**, 134105 (2010).
- [117] F. Weigend, Hartree–Fock exchange fitting basis sets for H to Rn, *J. Comput. Chem.* **29**, 167 (2008).
- [118] F. Neese, F. Wennmohs, A. Hansen, and U. Becker, Efficient, approximate and parallel Hartree–Fock and hybrid DFT calculations. A ‘chain-of-spheres’ algorithm for the Hartree–Fock exchange, *Chem. Phys.* **356**, 98 (2009).
- [119] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, *APL Mater.* **1**, 011002 (2013).
- [120] F. Neese, The ORCA program system, *WIREs Comput. Mol. Sci.* **2**, 73 (2012).
- [121] F. Neese, Software update: The ORCA program system, version 4.0, *WIREs Comput. Mol. Sci.* **8**, e1327 (2018).
- [122] F. Neese, F. Wennmohs, U. Becker, and C. Riplinger, The ORCA quantum chemistry program package, *J. Chem. Phys.* **152**, 224108 (2020).
- [123] A. S. Barnard, Modeling polydisperse ensembles of diamond nanoparticles, *Nanotechnology* **24**, 085703 (2013).
- [124] A. Barnard, Diamond Nanoparticle Structure Set (2014), doi:10.4225/08/546A9F79EC99C.
- [125] A. D. Becke and E. R. Johnson, A density-functional model of the dispersion interaction, *J. Chem. Phys.* **123**, 154101 (2005).
- [126] E. R. Johnson and A. D. Becke, A post-Hartree-Fock model of intermolecular interactions: Inclusion of higher-order corrections, *J. Chem. Phys.* **124**, 174104 (2006).
- [127] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H–Pu, *J. Chem. Phys.* **132**, 154104 (2010).
- [128] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed. (Society for Industrial and Applied Mathematics, Philadelphia, 2002).
- [129] H. Harbrecht, M. Peters, and R. Schneider, On the low-rank approximation by the pivoted Cholesky decomposition, *Appl. Numer. Math.* **62**, 428 (2012).
- [130] D. R. Hamann, Optimized norm-conserving Vanderbilt pseudopotentials, *Phys. Rev. B* **88**, 085117 (2013).
- [131] S. Sze and K. K. Ng, *Physics of Semiconductor Devices* (John Wiley, Hoboken, NJ, 2006).
- [132] Y. Kumagai and F. Oba, Electrostatics-based finite-size corrections for first-principles point defect calculations, *Phys. Rev. B* **89**, 195205 (2014).
- [133] J. D. Talman, NumSBT: A subroutine for calculating spherical Bessel transforms numerically, *Comput. Phys. Commun.* **180**, 332 (2009).