Highly accurate and constrained density functional obtained with differentiable programming

Sebastian Dick and Marivi Fernandez-Serra

Phys. Rev. B **104**, L161109 — Published 12 October 2021

DOI: 10.1103/PhysRevB.104.L161109

A highly accurate and constrained density functional obtained with differentiable programming

Sebastian Dick[1,2*] and Marivi Fernandez-Serra[1,2†]
Department of Physics and Astronomy Stony Brook University Stony Brook, NY 11794-3800
[1]*Physics and Astronomy Department, Stony Brook University, Stony Brook, New York 11794-3800, United States and*
[2]*Institute for Advanced Computational Science, Stony Brook University, Stony Brook, New York 11794-3800, United States*

Using an end-to-end differentiable implementation of the Kohn-Sham self-consistent field equations, we obtain a highly accurate neural network-based exchange and correlation (XC) functional of the electronic density. The functional is optimized using information on both energy and density while exact constraints are enforced through an appropriate neural network architecture. We evaluate our model against different families of XC approximations and show that at the meta-GGA level our functional exhibits unprecedented accuracy for both energy and density predictions. for non-empirical functionals, there is a strong linear correlation between energy and density errors. We use this correlation to define a novel XC functional quality metric that includes both energy and density errors, leading to a new, improved way to rank different approximations.

Density functional theory (DFT) serves without doubt as the workhorse method for electronic structure simulations in materials science and physics and has gained popularity within the chemistry community in recent decades. This is in no small part due to its favorable scaling, allowing users to tackle system sizes out of reach for most correlated wavefunction methods. However, inferences made from numerical simulations are only ever as good as their underlying approximations. This remains true for DFT, where these approximations are bundled somewhat opaquely in the elusive exchange-correlation (XC) functional. The Hohenberg-Kohn theorem guarantees that if this functional were known, ground-state properties of any interacting many-electron system could be described exactly [1]. In practice, one needs to pick from a plethora of different approximations, which often boils down to finding the right functional, cost and accuracy-wise, for the problem at hand.

It comes as no surprise that developing new and more accurate density functionals is a field of research on its own. Practitioners of this field generally have worked following two orthogonal approaches. Going back to Perdew and Wang [2], one approach tries to develop functionals from first-principles only, without any empirically-fit parameters. Some of the most notable functionals from this family include PBE [3], TPSS [4], and SCAN [5] which have proven themselves to be both versatile and reliable. Another approach, pioneered by Becke [6], is to fit functionals containing empirical parameters to either experimental or highly accurate simulated data. The size of these datasets can range from a few atoms to thousands of molecules and chemical reactions.

Beyond improving energies, approaching the exact functional should also lead to more accurate densities. However, a recent study suggested that most empirically fitted functionals fall short of this expectation [7]. This is concerning, not only from a theoretical point of view but also for practical reasons. For example, the quality of a functional's electronic density is related to its ability to correctly describe a molecule's response to an external electric field [8].

A guided approach towards empirical functionals that produce better densities is clearly needed. This task poses great challenges, as the Kohn-Sham equations introduce a non-linear relationship between functional form and self-consistent density. A guided optimization of such functionals requires knowledge of the gradients of the functional with respect to changes in the density. Pioneering work by Nagai et al. [9] circumvented the problem of missing gradients by adopting a simulated annealing approach to optimize a functional. DeePKS [10] uses a Coulomb-like term with randomized prefactor in its loss function that drives the functional towards the correct density. A breakthrough solution to this problem was very recently proposed by Li et al. [11], using differentiable programming, a tool that has previously been used to solve related problems in electronic structure [12]. By implementing the solution to the Kohn-Sham equations in JAX, a Python library that supports automatic differentiation on arbitrary operations, they can probe the electron density response to changes in the functional parametrization. The authors further showed that incorporating physical knowledge in the form of Kohn-Sham equations into the optimization algorithm has a regularizing effect making the algorithm more data-efficient. Their work, however, was limited to the study of 1-d model systems.

Here we optimize a density functional using an end-to-end differentiable implementation of the Kohn-Sham equations. In contrast to other approaches that have used machine learning to approximate the exact functional [9, 10, 13, 14], we impose a set of known constraint on the functional form. These include a local Lieb-Oxford bound [15–17] (LOB), which proves to be an important

ingredient to obtaining a more transferable model.

To make our training process computationally feasible, we mostly limit our training set to linear systems. We show that this can be done without loss of generality, meaning that a thus optimized functional is still applicable to more complex molecules. With the goal of obtaining a model with a good balance between computational cost and accuracy, we choose to optimize a neural network-based meta-GGA functional. We demonstrate that it shows optimal accuracy at the meta-GGA level across a diverse selection of datasets both with respect to energy as well as density predictions, in many cases outerperforming the front-runner SCAN [5] by a significant margin. Our analysis of different functionals identifies a high linear correlation between energy and density errors for non-empirical models. Using this relationship we define a novel compound metric which we term energy-density error. Ranked by this metric, our model is competitive even with functionals of the hybrid family.

At the heart of Kohn-Sham (KS) density functional theory lie the KS-equations

$$\left\{-\frac{1}{2}\nabla^2 + v_s[n](\mathbf{r})\right\}\psi_i(\mathbf{r}) = \epsilon_i\psi_i(\mathbf{r}) \tag{1}$$

In this approach, the electron density $n(\mathbf{r})$ is computed from the occupied one-particle orbitals $n(\mathbf{r}) = \sum_i^N |\psi_i(\mathbf{r})|^2$, and the potential is given as

$$v_s[n](\mathbf{r}) = v_{ext}(\mathbf{r}) + v_H[n](\mathbf{r}) + v_{xc}[n, \boldsymbol{\omega}](\mathbf{r}). \tag{2}$$

Here, $v_{ext}(\mathbf{r})$ is the external potential created by the ion cores, $v_H(\mathbf{r})$ is the Hartree potential capturing the Coulomb interaction of the density with itself and $v_{xc}(\mathbf{r})$ is the functional derivative of the exchange-correlation functional with respect to the electron density $v_{xc}(\mathbf{r}) = \frac{\delta E_{xc}[n,\boldsymbol{\omega}]}{\delta n(\mathbf{r})}$. As all quantities except for the exchange-correlation functional are known, the goal of this work will be to find a parametrization $\boldsymbol{\omega}$ of E_{xc} which accurately reproduces reference energies and electron densities while generalizing well to unseen systems.

As the potential depends on the density and therefore implicitly on the eigenstates ψ_i, the KS equations need to be solved iteratively. A popular Ansatz used in chemistry codes, and the one we choose here due to its efficiency for molecular systems, is to expand the eigenstates in Eq.1 in terms of atom-centered Gaussian orbitals $\psi_i = \sum_\mu C_{i\mu}\phi_\mu$. One advantage of using a Gaussian basis is that integrals can be pre-computed analytically and stored to disk, reducing on-the-fly computations to simple tensor contractions. For this work, we have made use of the open-source python code PySCF [18, 19]. We have re-implemented all routines needed to solve the Kohn-Sham equations to utilize PyTorch [20], making them end-to-end differentiable. One and two-electron integrals were computed with the original version of PySCF as the basis sets can be considered fixed for the purpose of this work.

A fully differentiable implementation of the self-consistent field (SCF) method necessitates that gradients occurring for every mathematical operation, at every SCF iteration, be held in memory until they are used during back-propagation. Especially tensor operations that involve grid points, such as the ones needed to generate the real-space density on which the XC functional is evaluated, contribute a high memory cost. We have chosen to partially circumvent this problem by largely restricting our training set to linear closed-shell molecules during training. We take advantage of their cylindrical symmetry by evaluating grid integrals on a reduced grid, namely a disk in the zx-plane. To obtain the radial part of this grid, we make use of the methods provided by PySCF to generate Treutler-Ahlrichs type grids. For the angular part, we use a simple Legendre-Gauss quadrature. The size of the reduced grid is chosen so that it reproduces the number of electrons, integrated exchange-correlation energy as well as the exchange-correlation potential (in the atomic orbital basis) given by a reference calculation using a converged three-dimensional grid.

We followed the common practice of defining the exchange correlation energy in terms of the energy per unit particle $E_{xc}[n, \boldsymbol{\omega}] = \int \epsilon_{xc}[n, \boldsymbol{\omega}](\mathbf{r})n(\mathbf{r})d\mathbf{r}$. We further decompose this energy density into its exchange and correlation parts $\epsilon_{xc}[n, \boldsymbol{\omega}](\mathbf{r}) = \epsilon_x[n, \boldsymbol{\omega}_x](\mathbf{r}) + \epsilon_c[n, \boldsymbol{\omega}_c](\mathbf{r})$ which are both independently parametrized. This allows us to factorize both functionals into fixed parts describing the behavior of a uniform electron gas (UEG) $e_{x/c}^{\mathrm{UEG}}[n]$ and parametrized enhancement factors $F_{x/c}[n, \boldsymbol{\omega}_{x/c}]$ that take into account effects from inhomogeneities. The exchange energy density of the UEG is given as $\epsilon_x^{\mathrm{UEG}}[n](\mathbf{r}) = -\frac{3}{4}(3/\pi)^{1/3}n^{1/3}(\mathbf{r})$, and the parametrization of $\epsilon_c^{\mathrm{UEG}}$ by Perdew and Wang [21] was used. Rather than having our functionals depend on the electron density and its derivatives directly we define the following commonly used dimensionless quantities which will serve as input to our functionals:

$$x_0 = n^{1/3} \tag{3}$$

$$x_1 = \frac{1}{2}\left\{(1+\zeta)^{4/3} + (1-\zeta)^{4/3}\right\} \tag{4}$$

$$x_2 = s = \frac{1}{2(3\pi^2)^{1/3}}\frac{|\nabla n|}{n^{4/3}} \tag{5}$$

$$x_3 = \alpha = \frac{\tau - \tau^W}{\tau^{unif}} \tag{6}$$

Where $\tau^W = |\nabla n|^2/8n$, $\tau^{unif} = (3/10)(3\pi^2)^{2/3}n^{5/3}$ and ζ corresponds to the spin polarization.

As neural networks struggle with handling features that range over multiple orders of magnitude, we further

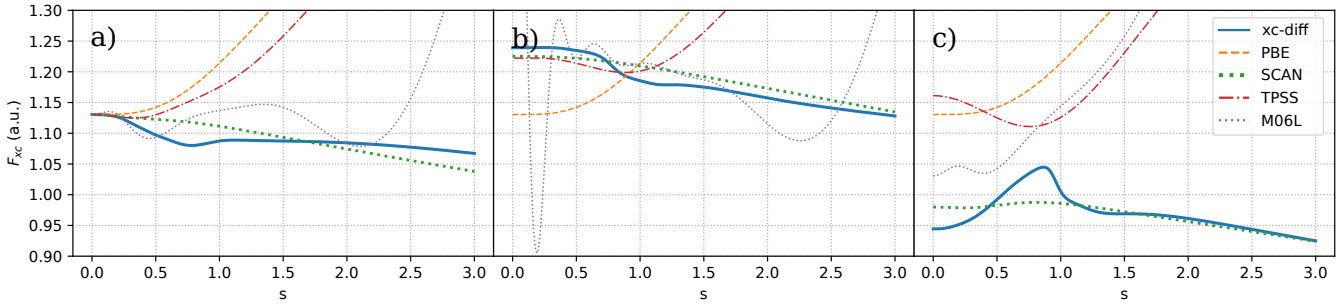

FIG. 1. Exchange-correlation enhancement factors F_{xc} for $r_s = 1$, $\zeta = 0$, and a) $\alpha = 1$, b) $\alpha = 0$, c) $\alpha = 10$

transform our input features x_{0-3} by applying logarithmic transformations

$$\tilde{x}_0 = \log(x_0 + \varepsilon_{\log}) \tag{7}$$
$$\tilde{x}_1 = \log(x_1 + \varepsilon_{\log}) \tag{8}$$
$$\tilde{x}_2 = \left\{1 - \exp(-x_2^2)\right\} \log(x_2 + 1) \tag{9}$$
$$\tilde{x}_3 = \log\left\{(x_3 + 1)/2\right\} \tag{10}$$

with $\epsilon_{\log} = 10^{-5}$. $\tilde{x}_2$ is designed so that its first derivative vanishes at $x_2 = 0$. This poses a soft constraint on the enhancement factors $F_{x/c}$ to have the same property. We have found that doing so greatly improves convergence, especially for periodic systems. Similar reasoning was applied to $\tilde{x}_3$ where the employed transformation lead to better behaved functionals than the more obvious choice $\log(x_3 + \varepsilon_{\log})$.

Both F_x and F_c were parametrized by a fully connected neural network with three hidden layers with 16 nodes each. As activation function, we have used the Gaussian error linear unit (GELU) [22]. We will denote the mapping induced by this neural network as $\mathcal{F}(\cdot)$.

We modify our neural network models to fulfill certain constraints and scaling laws that are known about the exact functional. To make E_x behave correctly under uniform scaling of the electron density and obey the spin-scaling relation, we drop the variables x_0 and x_1 in F_x. We further introduce a transformation $I_a(x)$ that maps its input x to a finite interval $[-1, a-1]$:

$$I_a(x) = \frac{a}{1 + (a-1)\exp(-x)} - 1 \tag{11}$$

while maintaining $I_a(0) = 0$. In the case of F_x, $I_{1.174}(x)$ is used to strictly enforce a rigorous conjectured local LOB[16, 17] $a = 1.174$, following the spirit of SCAN, whereas for F_c we use $I_2(x)$ to ensure non-negativity of the enhancement factor. Collecting all input features into a vector $\tilde{\mathbf{x}}$, the models can be written as:

$$F_x(\tilde{x}_2, \tilde{x}_3) = 1 + I_{1.174}((\tilde{x}_2 + \tanh^2 \tilde{x}_3)\mathcal{F}(\tilde{x}_2, \tilde{x}_3, \boldsymbol{\omega}_x)) \tag{12}$$
$$F_c(\tilde{\mathbf{x}}) = 1 + I_2((\tilde{x}_2 + \tanh^2 \tilde{x}_3)\mathcal{F}(\tilde{\mathbf{x}}, \boldsymbol{\omega}_c)) \tag{13}$$

The factor $(\tilde{x}_2 + \tanh^2 \tilde{x}_3)$ ensures that the UEG limit is recovered for $s = x_2 = 0$ ($\tilde{x}_2 = 0$) and $\alpha = x_3 = 1$ ($\tilde{x}_3 = 0$).

The datasets used in this work for training and validation consist of 21 atomization energies taken from the G2/97 set [23], three barrier heights taken from BH76 by Zhao et al [24] and two reference ionization potentials from IP13 provided in [25]. For the G2/97 dataset, we use atomization energies that were recalculated by Haunshild et al. [26] and are considered more reliable than the enthalpies of formation given in the original version of the dataset.

We augmented the G2/97 dataset with ground-state electron densities that we computed at the CCSD(T) level using the 6-311++G(3df,2pd) basis set, the same basis used for training the functionals. Total atomic energies were taken from Ref. 27 and included in the training set as well. Atomic electron densities were calculated and included for H and Li. For model validation, during training, we used a disjoint subset of the data listed above, consisting of 8 atomization energies and densities from G2/97, and two reference barrier heights from BH76. A detailed list of the structures used for training and validation can be found in the SI.

Models were pre-trained to match SCAN [5] on the 21 molecules contained in the training set by randomly sampling the exchange enhancement factor on molecular grids and fitting to it. The functional parameters are then trained to optimize a compound loss, combining errors in total energies $E^{(i)}_{j;tot}$ and reaction energies (which includes atomization energies and barrier heights) $E^{(i)}_{j;RE}$ at SCF iteration j, as well as electron densities $n^{(i)}$.

$$\mathcal{L} = \lambda_E \mathcal{L}_E + \lambda_{RE} \mathcal{L}_{RE} + \lambda_n \mathcal{L}_n \tag{14}$$
$$\mathcal{L}_E = \mathbb{E}\left[\sum_{j=10}^{25} \left\{w_j (E^{(i)}_{j;tot,ref} - E^{(i)}_{j;tot})\right\}^2\right] \tag{15}$$
$$\mathcal{L}_{RE} = \mathbb{E}\left[\sum_{j=10}^{25} \left\{w_j (E^{(i)}_{j;RE,ref} - E^{(i)}_{j;RE})\right\}^2\right] \tag{16}$$

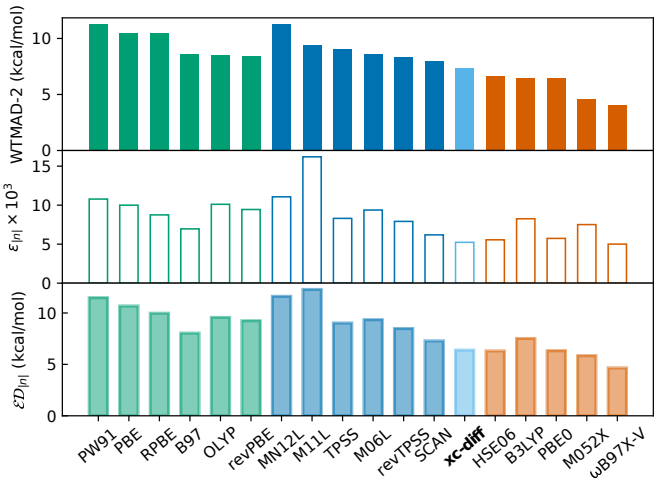

FIG. 2. Weighted mean absolute deviations (WTMAD-2) (top), density errors (center) and energy-density error (bottom) for several functionals including our model, xc-diff. 28.

$$\mathcal{L}_n = \mathbb{E}\left[l_n^{(i)}\right] \tag{17}$$

$$l_n^{(i)} = \frac{1}{N_e^2}\int_{\mathbf{r}} (n^{(i)}(\mathbf{r}) - n_{ref}^{(i)}(\mathbf{r}))^2 \tag{18}$$

with flexible weights λ_E, λ_{RE}, λ_n and expectation values taken over the training set. We set the weights to $\lambda_{RE} = 1$, $\lambda_n = 20$, $\lambda_E = 0.01$. Rather than including only converged energies in our loss function, we follow the approach by Li et. al [11], employing $w_j = \left(\frac{j-10}{15}\right)^2$ that penalize solutions which lead to slowly converging SCF calculations.

The functional parameters are optimized using Adam with an initial learning rate 10^{-4} which is decayed by a factor of 0.1 after every ten consecutive epochs without a decrease in training loss. We employ an l_2-regularization of 10^{-6} and a batch size of one reaction.

We tested our functional on 140 atomization energies contained in the W4-11 [29] dataset, 76 barrier heights from BH76, and 43 decomposition energies of artificial molecules contained in the MB16-43 [28] dataset. To achieve a wider assessment of our functional we further tested it on the diverse diet-GMTKN55 dataset [30]. GMTKN55 consists of 55 subsets that each probe different properties of a given functional. The subsets can be divided into categories by interaction type. These categories comprise reaction energies for small systems, reaction energies for large systems and isomerization reactions, barrier heights, intermolecular noncovalent interactions, and intramolecular noncovalent interactions. Diet-GMTKN55 provides representative sub-samples of GMTKN55 that have been shown to lead to the same ranking of DFs as the full dataset, at a significantly reduced computational cost.

We choose to evaluate our functional on the proposed 150 samples, the largest 'diet' dataset, using a weighted

	AE	BH	DE	WTMAD-2	$\varepsilon_{\lvert n\rvert} \times 10^3$	$\mathcal{ED}_{\lvert n\rvert}$
RPBE [33]	8.3	9.0	50.8	10.5	8.8	10.0
B97 [34]	4.7	7.3	36.1	8.6	7.0	8.0
OLYP [35]	9.9	8.5	29.0	8.5	10.1	9.6
revPBE [36]	7.6	8.3	27.1	8.4	9.4	9.2
M06L	4.4	3.9	63.3	8.6	9.4	9.3
revTPSS	5.7	8.9	36.7	8.4	7.9	8.5
SCAN	4.1	7.8	17.8	8.0	6.2	7.3
xc-diff	3.5	6.5	22.7	7.3	5.2	6.4
PBE0	3.7	5.0	15.9	6.4	5.7	6.3
B3LYP	3.4	5.7	24.8	6.5	8.3	7.5
M05-2X [37]	4.0	1.7	26.3	4.6	7.5	5.8
ωB97X-V [38]	2.8	1.8	32.5	4.1	5.0	4.7

TABLE I. Mean absolute errors in kcal mol^{-1} for atomization energies (AE) over the W4-11 dataset, barrier heights (BH) in BH76 and decomposition energies (DE) for MB16-43. Weighted means WTMAD-2 and Δ are also given in kcal mol^{-1}. Mean density error ε_n is unit-less. A complete list of functionals is provided in the SI. All models include DFT-D3 dispersion corrections. Energy errors for all functionals except xc-diff were taken from Ref. 28.

mean of mean absolute deviations (MAD) across the subsets. The weights are chosen by Gould to reproduce the WTMAD-2 weighted mean of means proposed by Goerig et al., which scales the mean absolute energy deviations MAD_i of a subset i containing N_i reactions by the inverse energy range of a given subset $\overline{|\Delta E|}_i$

$$\text{WTMAD-2} = N^{-1}\sum_i^{55} N_i \cdot \frac{56.84\text{kcal mol}^{-1}}{\overline{|\Delta E|}_i} \cdot \mathrm{MAD}_i, \tag{19}$$

with $N = \sum_i^{55} N_i$. The goal is to give more weight to datasets with little variation in the energy and to scale down systems with large variations.

We conducted all necessary single-point calculations with PySCF using our in-house code libnxc [31, 32] as a plug-in to allow for the use of PyTorch XC models. Libnxc is freely available on Github under the MPL-2.0 License and provides a straightforward way to users to employ our functional in electronic structure calculations. Instructions on how to do so are provided in the documentation accompanying the code. We employed the def2-QZVP basis set and augmented it with diffuse functions for the subsets recommended in Ref. 28. A PySCF grid level of 3 together with an energy convergence tolerance of $10^{-8}E_h$ was chosen.

To ensure the correct treatment of non-covalent interactions, all results reported include the DFT-D3 dispersion correction with Becke-Johnson damping [39, 40]. Parameters for our functional were optimized following the procedure outlined in Ref. 28 and are summarized in the SI.

Fig.1 shows a comparison of XC enhancement factors $F_{\mathrm{xc}} = \varepsilon_{\mathrm{xc}}/\varepsilon_{\mathrm{x}}^{\mathrm{UEG}}$ for a set of density functionals. Despite the small regularization employed, the obtained neural

network-based functional is smooth and no problems regarding convergence during SCF calculations were encountered. We accredit this to the optimization procedure and the weighted loss which penalized parametrizations that would lead to slowly converging calculations. We also show in the supplementary material[41] sec[ix] (see, also, references[42–45] therein) that the convergence of our functional with respect to the real space grid size is as good if not better than that obtained with SCAN.

Comparing the weighted means WTMAD-2 shown in Fig. 2 and Tab. I, we see that xc-diff outperforms SCAN, (rev)TPSS [46] , and the empirically fitted Minnesota functionals M06L, M11L [47] and MN12L [48]. It should be pointed out that the training sets used to optimize the Minnesota functionals were about one order of magnitude larger than the one used in this work.

The datasets W4-11, BH76, and MB16-43 illuminate the strengths and weaknesses of the respective functionals. For atomization energies of small systems, xc-diff outperforms SCAN by 0.6 kcal mol^{-1} and is comparable to the global hybrids B3LYP [49] and PBE0 [50]. Being susceptible to delocalization errors, barrier heights pose a challenge to semi-local functionals. Here, xc-diff outperforms SCAN by more than 1 kcal mol^{-1} but is outperformed by about the same amount by PBE0 and B3LYP. Not fully shown in Tab. I due to their large WTMAD-2, the Minnesota functionals provide an excellent treatment of this dataset with MAEs ranging from 3.9 to 1.7 kcal mol^{-1}. However, it is worth noting that barrier heights played a major role in the training sets used to optimize all Minnesota functionals, so their accuracy comes as no surprise. MB14-36 plays a special role as it contains artificial, randomly generated molecules and has proven challenging especially to empirical functionals. Here, xc-diff is less accurate than SCAN but shows reasonable performance compared to all other functionals considered here. Beyond tests in general data, we have also tested xc-diff on a specific data set

Beyond comparing energies, we used the previously calculated CCSD(T) electron densities across the G2/97 dataset to assess the accuracy of our functional regarding densities. Mean errors across the dataset were computed using the metric

$$\varepsilon_{|n|} = \mathbb{E}\left[\frac{1}{N_e}\int_{\mathbf{r}} |n^{(i)}(\mathbf{r}) - n^{(i)}_{ref}(\mathbf{r})|\right] \tag{20}$$

The methods were identical to those used for the diet-GMTKN55 dataset except for the basis set, which was chosen as 6-311++G(3df,2pd) for easier comparison with our coupled-cluster reference densities.

Fig. 2 shows that xc-diff outperforms all other tested meta-GGA functionals by a clear margin. We further included data obtained with global hybrids and GGAs. While most hybrids improve upon traditional meta-GGA functionals, xc-diff is still 9% more accurate regarding the density than PBE0.

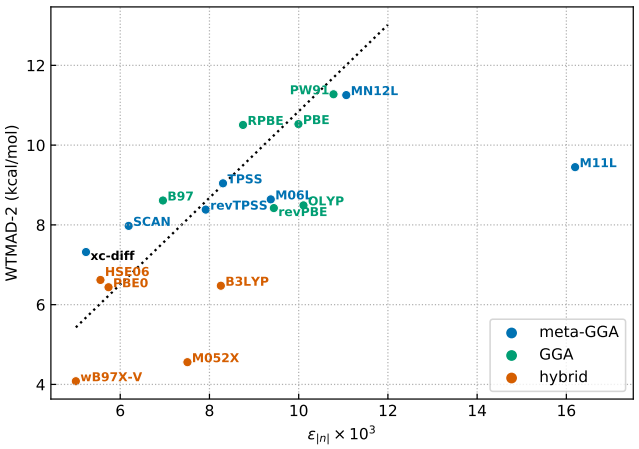

FIG. 3. Correlation plot for density error and total WTMAD-2. Dotted line indicates best linear fit to non-empirical DFs

We believe that a functional should be judged by both its accuracy regarding energies as well as densities. A metric combining both energy and density errors would therefore be useful to score and rank functionals, however finding such a metric is no straightforward task.

An important clue might be provided by the high linear correlation ($R^2 = 0.87$) between WTMAD-2 and density errors for non-empirical DFs (PW91[2], PBE[3], TPSS[4], revTPSS[46], SCAN, PBE0[50]). The best fit of a linear regression model (with zero intercept) is shown in Fig. 3 by a dotted line. Remarkably, regardless of the level of approximation, non-empirical fuctionals closely follow this trend-line, while many empirically fitted DFs seem to deviate significantly from it. We have been able to confirm that this trend persists for other definitions of the density error, such as one based on the Kullback-Leibler divergence (see SI for details). Our functional, xc-diff, shows a density error that is lower than expected from this trend.

Inspired by this finding, we propose a new metric $\mathcal{ED}$ that allows us to combine density with energy errors:

$$\mathcal{ED}_{|n|} = 2\left(\frac{1}{\text{WTMAD-2}} + \frac{1}{f_E(\varepsilon_{|n|})}\right)^{-1}. \tag{21}$$

$f_E(\varepsilon_{|n|}) = \gamma \cdot \varepsilon_{|n|}$ with $\gamma = 1084.87$ kcal mol^{-1} corresponds to the linear regression model used in Fig. 3, and can be interpreted as the energy error (WTMAD-2) a fictional non-empirical functional with density error ε_n would exhibit according to our model. Fig. 2 shows $\mathcal{ED}_{|n|}$ across density functionals. We see that within meta-GGAs, the order of functionals remains largely unchanged but due to xc-diff's accuracy for densities, it now outperforms B3LYP and matches the accuracy range of other popular hybrids such as PBE0. It is out of the scope of this manuscript to study how xc-diff performs for systems and problems that SCAN has difficulty with such

the the self interaction error in water clusters.[51, 52]. We expect such study to provide results similar to SCAN. Here, we have computed the optimized geometry of the water molecule (supplementary material[41] Sec[vii]). Xc-diff improves the $\widehat{HOH}_{xc-diff} = 104.5^o$ (same as experimental[51]) over SCAN (104.3^o[51]), while for the OH-bond length we obtain $r_{OH}^{xc-diff} = 0.964$ Å, as compared to $r_{OH}^{exp} = 0.958$ Å, and $r_{OH}^{SCAN} = 0.960$ Å. Additional results, showing the similarity to SCAN regarding the self-interaction error in the ionized water dimer [52] are provided in Supplementary material[41] sec[viii].

Using an end-to-end differentiable implementation of the Kohn-Sham equations we have successfully optimized an accurate meta-GGA XC functional. Our results indicate that a highly constrained functional like SCAN has already almost exhausted the accuracy limit that a meta-GGA functional can achieve. Nevertheless, within this narrow window, our method was able to improve upon SCAN regarding both a diverse set of reaction energies and electron densities. It has been argued that such improvement should be achieved in a non-empirical approach imposing physically motivated exact constraints with a minimal number of free parameters [53]. We have shown that a data-driven search using machine learning combined with an adherence to constraints can provide an equally valid path. A crucial ingredient of our method is given by automatic differentiation, which allows the optimization algorithm to make use of valuable information contained in the electron density, effectively enlarging the training set size. It remains to be tested how a thus optimized functional performs for solid systems; work that will be the subject of future research. While we believe that our functional could be further improved by fitting to larger training sets, its accuracy is inherently limited by the functional form of meta-GGAs. This issue particularly emerges when trying to address systems for which self-interaction errors play a significant role. We predict that advances in hardware development along with more efficient implementations of our code will soon allow us to apply our method to much larger training sets and higher rungs of DFT's Jacob's ladder [53], opening the path towards functionals of optimal accuracy, within their rungs of approximation.

This work was supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, under Awards DE-SC0001137 and DE-SC0019394, as part of the CCS and CTC Programs. Sebastian Dick was supported by a fellowship from The Molecular Sciences Software Institute under NSF grant ACI-1547580. We would like to thank Stony Brook Research Computing and Cyberinfrastructure, and the Institute for Advanced Computational Science at Stony Brook University for access to the high-performance SeaWulf computing system, which was made possible by a $1.4M National Science Foundation grant (#1531492).

* sebastian.dick@stonybrook.edu
† maria.fernandez-serra@stonybrook.edu

[1] P. Hohenberg and W. Kohn, Physical review **136**, B864 (1964).
[2] J. P. Perdew, J. A. Chevary, S. H. Vosko, K. A. Jackson, M. R. Pederson, D. J. Singh, and C. Fiolhais, Physical review B **46**, 6671 (1992).
[3] J. P. Perdew, K. Burke, and M. Ernzerhof, Physical review letters **77**, 3865 (1996).
[4] J. Tao, J. P. Perdew, V. N. Staroverov, and G. E. Scuseria, Physical Review Letters **91**, 146401 (2003).
[5] J. Sun, A. Ruzsinszky, and J. P. Perdew, Physical review letters **115**, 036402 (2015).
[6] A. D. Becke, Physical review A **38**, 3098 (1988).
[7] M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, and K. A. Lyssenko, Science **355**, 49 (2017).
[8] D. Hait, Y. H. Liang, and M. Head-Gordon, The Journal of chemical physics **154**, 074109 (2021).
[9] R. Nagai, R. Akashi, and O. Sugino, npj Computational Materials **6**, 1 (2020).
[10] Y. Chen, L. Zhang, H. Wang, and E. Weinan, Journal of Chemical Theory and Computation **17**, 170 (2021), arXiv:2008.00167.
[11] L. Li, S. Hoyer, R. Pederson, R. Sun, E. D. Cubuk, P. Riley, and K. Burke, Physical Review Letters **126**, 1 (2021), arXiv:2009.08551.
[12] T. Tamayo-Mendoza, C. Kreisbeck, R. Lindh, and A. Aspuru-Guzik, ACS central science **4**, 559 (2018).
[13] S. Dick and M. Fernandez-Serra, The Journal of chemical physics **151**, 144102 (2019).
[14] S. Dick and M. Fernandez-Serra, Nature communications **11**, 1 (2020).
[15] E. H. Lieb and S. Oxford, International Journal of Quantum Chemistry **19**, 427 (1981).
[16] J. P. Perdew, A. Ruzsinszky, J. Sun, and K. Burke, The Journal of chemical physics **140**, 18A533 (2014).
[17] J. Sun, J. P. Perdew, and A. Ruzsinszky, Proceedings of the National Academy of Sciences **112**, 685 (2015).
[18] 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.*, Wiley Interdisciplinary Reviews: Computational Molecular Science **8**, e1340 (2018).
[19] 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.*, The Journal of chemical physics **153**, 024109 (2020).
[20] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, arXiv preprint arXiv:1912.01703 (2019).
[21] J. P. Perdew and Y. Wang, Physical review B **45**, 13244 (1992).
[22] D. Hendrycks and K. Gimpel, arXiv preprint arXiv:1606.08415 (2016).
[23] L. A. Curtiss, K. Raghavachari, P. C. Redfern, and J. A. Pople, The Journal of Chemical Physics **106**, 1063 (1997).
[24] Y. Zhao, N. González-García, and D. G. Truhlar, The Journal of Physical Chemistry A **109**, 2012 (2005).

[25] B. J. Lynch, Y. Zhao, and D. G. Truhlar, The Journal of Physical Chemistry A **107**, 1384 (2003).

[26] R. Haunschild and W. Klopper, The Journal of chemical physics **136**, 164102 (2012).

[27] S. J. Chakravorty, S. R. Gwaltney, E. R. Davidson, F. A. Parpia, and C. F. p Fischer, Physical Review A **47**, 3649 (1993).

[28] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, and S. Grimme, Physical Chemistry Chemical Physics **19**, 32184 (2017).

[29] A. Karton, S. Daon, and J. M. Martin, Chemical Physics Letters **510**, 165 (2011).

[30] T. Gould, Physical Chemistry Chemical Physics **20**, 27735 (2018).

[31] S. Dick, libnxc, `https://github.com/semodi/libnxc` (2021).

[32] M. Fernandez-Serra and S. Dick, A highly accurate and constrained density functional obtained with differentiable programming, `https://doi.org/10.5281/zenodo.5516522` (2021).

[33] B. Hammer, L. B. Hansen, and J. K. Nørskov, Physical review B **59**, 7413 (1999).

[34] A. D. Becke, The Journal of chemical physics **107**, 8554 (1997).

[35] N. C. Handy and A. J. Cohen, Molecular Physics **99**, 403 (2001).

[36] Y. Zhang and W. Yang, Physical Review Letters **80**, 890 (1998).

[37] Y. Zhao, N. E. Schultz, and D. G. Truhlar, Journal of chemical theory and computation **2**, 364 (2006).

[38] N. Mardirossian and M. Head-Gordon, Physical Chemistry Chemical Physics **16**, 9904 (2014).

[39] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, The Journal of chemical physics **132**, 154104 (2010).

[40] S. Grimme, S. Ehrlich, and L. Goerigk, Journal of computational chemistry **32**, 1456 (2011).

[41] See Supplemental Material at [URL needed] for additional results and a detailed information of the implementation of the method.

[42] S. Kullback and R. A. Leibler, The annals of mathematical statistics **22**, 79 (1951).

[43] J. Brandenburg, J. Bates, J. Sun, and J. Perdew, Physical Review B **94**, 115144 (2016).

[44] A. P. Bartk and J. R. Yates, The Journal of Chemical Physics **150**, 161101 (2019).

[45] D. Meja-Rodrguez and S. B. Trickey, The Journal of Chemical Physics **151**, 207101 (2019).

[46] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, L. A. Constantin, and J. Sun, Physical Review Letters **103**, 026403 (2009).

[47] R. Peverati and D. G. Truhlar, The Journal of Physical Chemistry Letters **3**, 117 (2012).

[48] R. Peverati and D. G. Truhlar, Physical Chemistry Chemical Physics **14**, 13171 (2012).

[49] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, The Journal of physical chemistry **98**, 11623 (1994).

[50] C. Adamo and V. Barone, The Journal of chemical physics **110**, 6158 (1999).

[51] K. Sharkas, K. Wagle, B. Santra, S. Akter, R. R. Zope, T. Baruah, K. A. Jackson, J. P. Perdew, and J. E. Peralta, Proceedings of the National Academy of Sciences **117**, 11283 (2020), https://www.pnas.org/content/117/21/11283.full.pdf.

[52] V. Sharma and M. Fernández-Serra, Phys. Rev. Research **2**, 043082 (2020).

[53] J. P. Perdew, A. Ruzsinszky, J. Tao, V. N. Staroverov, G. E. Scuseria, and G. I. Csonka, The Journal of chemical physics **123**, 062201 (2005).