Machine Learning: Science and Technology

Machine learning the derivative discontinuity of density-functional theory

IOP Publishing Mach. Learn.: Sci. Technol. 3 (2022) 015011 https://doi.org/10.1088/2632-2153/ac3149

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
2 September 2021

REVISED
8 October 2021

ACCEPTED FOR PUBLICATION
20 October 2021

PUBLISHED
15 December 2021

PAPER

Machine learning the derivative discontinuity of density-functional theory

Johannes Gedeon[1], Jonathan Schmidt[1], Matthew J P Hodgson[2], Jack Wetherell[3], Carlos L Benavides-Riveros[4,5,*] and Miguel A L Marques[1,*]

[1] Institut für Physik, Martin-Luther-Universität Halle-Wittenberg, 06120 Halle (Saale), Germany
[2] Department of Physics, Durham University, South Road, Durham DH1 3LE, United Kingdom
[3] LSI, École Polytechnique, CNRS, Institut Polytechnique de Paris, F-97728 Palaiseau, France
[4] Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Str. 38, 01187 Dresden, Germany
[5] NR-ISM, Division of Ultrafast Processes in Materials (FLASHit), Area della Ricerca di Roma 1, Via Salaria Km 29.3, I-00016 Monterotondo Scalo, Italy
[*] Authors to whom any correspondence should be addressed.

E-mail: carlosbe@pks.mpg.de **and** miguel.marques@physik.uni-halle.de

Keywords: machine learning, quantum physics, density functional theory, electronic structure, ensemble density functional theory

Abstract

Machine learning is a powerful tool to design accurate, highly non-local, exchange-correlation functionals for density functional theory. So far, most of those machine learned functionals are trained for systems with an integer number of particles. As such, they are unable to reproduce some crucial and fundamental aspects, such as the explicit dependency of the functionals on the particle number or the infamous derivative discontinuity at integer particle numbers. Here we propose a solution to these problems by training a neural network as the universal functional of density-functional theory that (a) depends explicitly on the number of particles with a piece-wise linearity between the integer numbers and (b) reproduces the derivative discontinuity of the exchange-correlation energy. This is achieved by using an ensemble formalism, a training set containing fractional densities, and an explicitly discontinuous formulation.

1. Introduction

In their now famous paper, Hohenberg and Kohn proved that the electron density $\rho(\boldsymbol{r})$ suffices to compute all observables of a system of interacting electrons [1]. Due to a remarkable balance of computational cost and numerical precision, first principles modeling of electronic systems based on this density functional theory (DFT) is nowadays a daily practice, with great impact in material science, quantum chemistry or condensed matter [2]. The success of DFT is to a large extent based on the Kohn–Sham formulation, that utilizes a system of non-interacting electrons that has the same density as the interacting one [3]. The main ingredient of this formulation is $E_{\rm xc}[\rho]$, the universal exchange-correlation (xc) functional, whose functional derivative provides an effective external potential for the non-interacting particles. Yet, while the Hohenberg-Kohn theorem proves the uniqueness of such a functional, it does not give any indication regarding its specific form. To circumvent this issue, a very large number of approximate functionals were developed in the last decades [4, 5], often combining empirical knowledge, exact mathematical conditions, and a great deal of ingenuity.

Inspired by the success of machine learning (ML) in various technological applications, including image and speech recognition [6], the last couple of years have seen the development of several neural-network-based approximations to $E_{\rm xc}[\rho]$. Indeed, machine-learning offers a new generation of accurate, highly non-local, xc functionals [7]. While those functionals are designed to perform tasks of different degree of complexity, all share the aim of learning one of the maps of DFT, namely, the Hohenberg-Kohn map between the external potential $v(\boldsymbol{r})$ and the density $\rho(\boldsymbol{r})$ [8–13], or the Kohn–Sham map between the density $\rho(\boldsymbol{r})$ and the xc functional $E_{\rm xc}[\rho(\boldsymbol{r})]$ and its functional derivative $v_{\rm xc}[\rho(\boldsymbol{r})] = \delta E_{\rm xc}[\rho(\boldsymbol{r})]/\delta\rho(\boldsymbol{r})$ [14–16].

The functionals delivered by machine-learning DFT (ML-DFT) are in general non-local, in the sense that they use multiple density points as input, and can be efficiently trained with data from reference methods. Yet, since ML-DFT functionals are mostly trained in Hilbert spaces with an *integer* number of particles, they are still unable to reproduce some critical and fundamental aspects of DFT. For instance, it is known that any satisfactory definition of the energy functional must depend explicitly on the particle number [17–19]. Furthermore, the derivative of the xc functional in terms of the number of particles exhibits a discontinuity that plays a crucial role in the description of electronic bandgaps [20–23], charge-transfer excitations [24, 25], molecular dissociation [26–29], or even Mott insulators [30], to name but a few examples.

Systems with non-integer (fractional) number of electrons $(N+\epsilon)$ are defined as statistical mixtures of systems with integer number of particles [21, 31]. As such, the density $\rho_{N+\epsilon}(\boldsymbol{r})$ and total energy $E(N+\epsilon)$ are piecewise linear functions of ϵ, namely:

$$\rho_{N+\epsilon}(\boldsymbol{r}) = (1-\epsilon)\rho_N(\boldsymbol{r}) + \epsilon\rho_{N+1}(\boldsymbol{r}), \tag{1a}$$

$$E(N+\epsilon) = (1-\epsilon)E(N) + \epsilon E(N+1) \tag{1b}$$

with $0 \leqslant \epsilon \leqslant 1$. At integer N (i.e. when $\epsilon = 0$), the derivatives of the density and the energy exhibit a discontinuity, and the xc potential $v_{\mathrm{xc}}(\boldsymbol{r})$ jumps by a finite value [32]. The difference in the slope on the left/right side of the total energy at integer values is equal to the fundamental gap [21]:

$$I - A = \left.\frac{\partial E}{\partial N}\right|_{+} - \left.\frac{\partial E}{\partial N}\right|_{-}, \tag{2}$$

where I is the ionization energy and A the electron affinity. Yet, in practice, standard approximations to the xc functionals that depend explicitly on the electronic density, such as the local-density (LDA) and generalized-gradient (GGA) approximations, are continuously differentiable functions of N and lack therefore a derivative discontinuity. Meta-GGAs can exhibit a discontinuity due to their dependence on the kinetic-energy density, but it is usually too small or even negative [33]. Due to their dependence on the Kohn–Sham orbitals, orbital functionals are discontinuous [34], but this comes at the price of a much higher computational effort.

In addition to the discontinuity, a universally useful approximation for the xc functional must be 'N-electron self-interaction-free' for *all* positive integer N [35], meaning that the total energy of a system with $N+\epsilon$ electrons in the range $(N, N+1)$ should exhibit a linear variation with respect to ϵ. For attractive interactions the energy is a convex function with straight lines joining subsets of ground-state energies [36]. Yet, approximate functionals deviate from such a correct behavior. It has been shown that semi-local density functionals are in general convex with perhaps small concave pieces [37]. Even the Hartree Fock theory leads to piecewise concave curves between integers [37]. We note that the relatively well-defined curvature of the curves is ultimately the reason for the success of the Slater half-occupation scheme [38] or the LDA-1/2 method [39]. In fact, these schemes use the derivative at the midpoint (i.e. at $N-0.5$), that can be shown to be equal to the slope of the straight line between $E(N-1)$ and $E(N)$ if the curvature is constant, irrespective of its sign [40]. We also note that the role of the basis set in the shape of the predicted energy curves has been recently studied in great detail: For certain atomic systems with diffuse anion states, standard DFT (and Hartree–Fock as well) calculations can predict better the behaviour of the energy as a function of N—and more specifically the sign of the eigenvalues of the highest and lowest occupied molecular orbitals—when highly diffuse basis sets are employed [41, 42].

The centrality of these pressing issues in DFT can be further highlighted by the fact that a rigorous description of the delocalization error can be related with the energy curve of the xc functionals lying below the straight energy lines [37, 43].

In this work, we propose a way to train a neural network as the *ensemble* universal functional of a system of fractional electron numbers that describes correctly the derivative discontinuity and the piecewise linear behavior. The ML functionals we present contain explicitly the physics of the derivative discontinuity of DFT, are highly non-local, and are trained for systems with fractional densities. For this reason, our functionals can potentially address the well known delocalization and static correlation errors of DFT [44–46] simultaneously.

2. Results and discussion

Inspired by the neural network topology proposed in [14], our neural network takes an electronic density as an input and returns the corresponding xc energy, whose functional derivative can in turn be used to solve

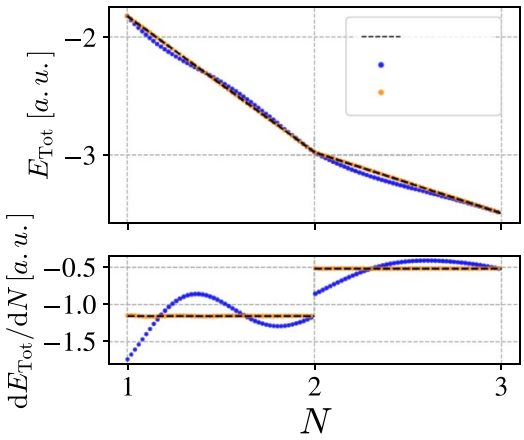

Figure 1. Comparison of the total energy for a certain external potential (top) as well as its derivative with respect to the particle number (bottom) for two different models: model A has been trained with integer densities only. Model B has been trained with fractional densities, the exact Δ_{xc}-shift, and employs the AF (see text).

the Kohn–Sham equations. The network is a sliding window convolution (SWC) network. For a 1D system of discrete spatial points $\{r_1, \ldots, r_W\}$, a *window* with a certain 'kernel size' κ scans each data point $\rho_\sigma(r_j)$ and its $\kappa - 1$ nearest neighbors $\eta_\sigma(r_j, \kappa) = \{\rho_\sigma(r_{j-(\kappa-1)/2}), \ldots, \rho_\sigma(r_{j+(\kappa-1)/2})\}$ with $\sigma = \uparrow, \downarrow$ to calculate a local energy $\epsilon^{\theta}_{\mathrm{loc}}\left[\eta_\uparrow(r_j, \kappa), \eta_\downarrow(r_j, \kappa)\right]$. The total xc energy is calculated by summing over the local energies:

$$E_{\mathrm{xc}}\left[\rho_\uparrow, \rho_\downarrow\right] = \sum_j \rho(r_j)\, \epsilon^{\theta}_{\mathrm{loc}}\left[\eta_\uparrow(r_j, \kappa), \eta_\downarrow(r_j, \kappa)\right]. \tag{3}$$

Here, θ denotes the trainable parameters. The input channels can be the total electronic density $\rho = \rho_\uparrow + \rho_\downarrow$ or the spin densities $\rho_\uparrow$ and $\rho_\downarrow$. The corresponding xc potentials can be computed using automatic differentiation, as shown in [14]. The parameters of the neural network are updated according to the loss function:

$$\mathcal{L}(\theta; \alpha, \beta) = \alpha\, \mathrm{MSE}(v^{\uparrow}_{\mathrm{xc}}[\rho_\uparrow, \rho_\downarrow], v^{\downarrow}_{\mathrm{xc}}[\rho_\uparrow, \rho_\downarrow]) + \beta\, \mathrm{MSE}(E_{\mathrm{xc}}[\rho_\uparrow, \rho_\downarrow]), \tag{4}$$

where α and β are *fixed* weights that can be adjusted to expedite convergence, and MSE is the mean squared error. A more detailed description of our networks can be found in the section 4. We improve the performance of this architecture by (a) training our neural network with non-integer densities, (b) introducing the jump of the xc potential at integer numbers into the loss function, and (c) adding an explicit discontinuity at integer electron numbers. In the following we discuss the details of these new approaches.

As we will also discuss in great detail in the section 4, we have chosen for this work a 1D model system that mimics 3D reality closely enough to render great physical insight. This model is well known in the literature as a quite useful theoretical laboratory for studying strong correlation, developing xc density functionals for DFT or studying the emission of even harmonics [47–50].

2.1. Fractional particle numbers

In general, neural networks do not extrapolate well outside the distribution of the samples used for their training. Consequently, we can not expect that machines trained solely for integer densities (as usually done) will exhibit the correct linear behaviour of the energy. To illustrate this behavior we plot, in figure 1, the total energy calculated with a neural network trained solely with integer densities (model A in the figure) as a function of the number of particles for the system described in the section 4. Besides the fact that model A is far from linear, we can also note that the sign of the curvature is not constant, with both concave and convex parts. This is somehow to be expected, as the network, in contrast to the usual xc functionals, only incorporates physical knowledge through the training examples with integer densities. As such, we can easily see that approaches such as LDA-1/2 are bound to fail in this case.

As a first strategy to solve this problem, we decided to include samples calculated within ensemble-DFT at fractional densities in our training. To obtain this data we created a set of total energies and electronic densities for a series of 1-dimensional exact calculations. We then constructed ensemble densities and inverted the ensemble Kohn–Sham equations, in order to compute the exact xc energy and potential for these

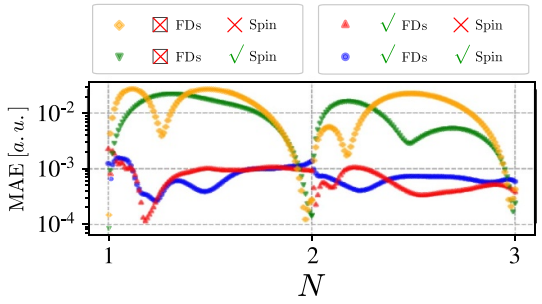

Figure 2. Comparison of the mean absolute error (MAE) of the total energy for four different models (averaged over 100 external potentials). 'Spin' denotes models including spin densities. The models 'FDs' (blue and red) include fractional densities $\epsilon = (0.05, 0.2, 0.5, 0.8, 0.95)$ in the training. The other models (yellow and green) include only integer densities in the training.

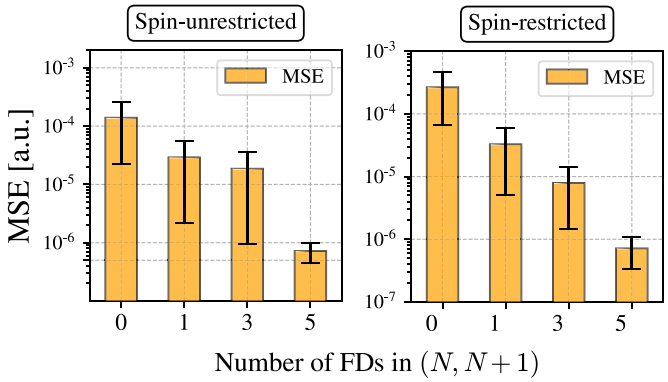

Figure 3. Comparison of the mean squared error (MSE) and its variances for different models (spin-restricted and spin-unrestricted) trained with different numbers of fractional densities. The values corresponding to the number of FDs represent the following fractional densities (ϵ-values) in detail: $0 \equiv$ no fractionals, $1 \equiv (0.5)$, $3 \equiv (0.2, 0.5, 0.8)$ and $5 \equiv (0.05, 0.2, 0.5, 0.8, 0.95)$. The errors are averaged over 100 external potentials. For more information see section 4.

systems. We used an inversion algorithm based on [51] that we extended for both spin-DFT [52] and to ensemble systems. As a result, we created exact training and testing data with particle numbers between 1 and 3 electrons, that we used to train our models. As explained in the Method sections, we use Young diagrams to compute the exact spin densities (i.e. with the correct spin symmetries).

In figure 2 we compare the mean absolute error (MAE) of the total energy for functionals trained with fractional densities (blue and red) and functionals trained only for systems with integer densities (yellow and green) (averaged over a test set of 100 external potentials). As we can see, the models trained at integer densities yields an excellent prediction for the total energy at those integers, but exhibits a considerably larger error at fractional numbers. Remarkably, by simply adding fractional densities to the training set, the error decreases by more than one order of magnitude, and the MAE over the entire $[1, 3]$-range remains below 2×10^{-3} a.u. In addition to that, we study the effect of increasing the number of trained fractional densities in $(N, N+1)$ on the mean squared error (MSE), averaged additionally over the densities, illustrated in figure 3. As expected, the error decreases when adding more fractional densities to the training set. The choice of training with spin does not seem to have any visible benefits in general compared to the training based on spin-restricted calculations. As such, we decided to use the spin-restricted formalism in the following.

While this strategy resolved, to a large extent, the many-electron self-interaction error, a problem still remains in the vicinity of the integer particle numbers. In fact, our network is fully differentiable, and does not (in fact can not) exhibit a true derivative discontinuity (in a mathematical sense) as a function of N.

2.2. Jump in the xc potential

We can easily relate the discontinuity of the total energy at integer particle numbers with an uniform shift in the potential. Indeed, the exact uniform shift Δ_{xc}^N of v_{xc} at integer particle number N obeys the relation [53]:

$$\left.\frac{\partial E}{\partial N}\right|_+ - \left.\frac{\partial E}{\partial N}\right|_- = \varepsilon_{\mathrm{s}}^N + \Delta_{\mathrm{xc}}^N, \tag{5}$$

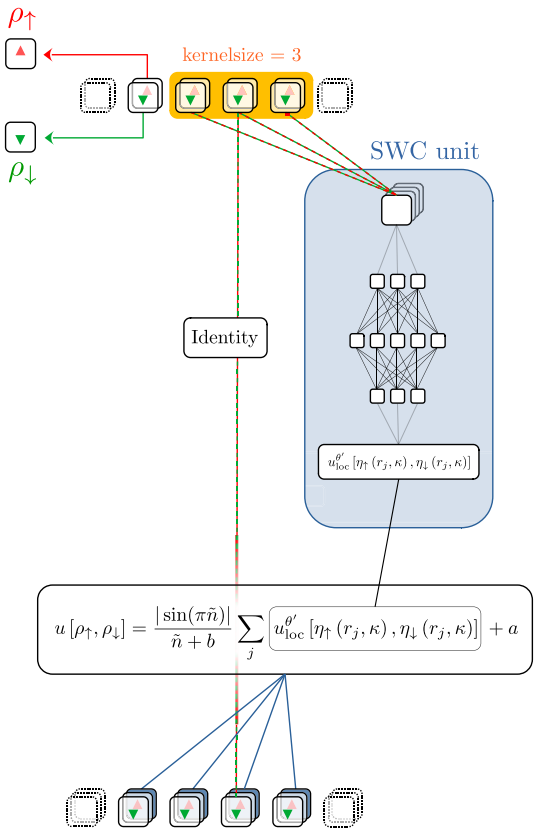

Figure 4. Integration of the non-differentiable auxiliary function $u\left[\rho_\uparrow,\rho_\downarrow\right]$ as a third channel to the basic SWC unit. The input data contains two channels for both spin-densities $\rho_\uparrow$ and $\rho_\downarrow$ respectively.

where ε_s^N is the Kohn–Sham gap, i.e. the difference between the lowest unoccupied (LUMO) and the highest occupied (HOMO) molecular orbital energies. Noticeably, $E(N)$ and $E(N\pm1)$, as well as the eigenvalues corresponding to the LUMO and HOMO can be computed while creating the training sets.

Our second strategy consists of computing, in our learning process, both $\rho_{N+\epsilon}$, and the exact shift $v_{\mathrm{xc}}(N_+) - v_{\mathrm{xc}}(N_-)$. The corresponding mean squared error is then used to extend the loss function in equation (4)

$$\mathcal{L} \to \mathcal{L} + \lambda \mathrm{MSE}(v_{\mathrm{xc}}(N+\epsilon) - v_{\mathrm{xc}}(N))\,, \tag{6}$$

where λ is an additional hyperparameter.

We expect that this extended loss function can help the functional to learn the correct shift in the derivative. Training our basic SWC network with only this loss function failed for any learning rate tested. This can be understood from the fact that our network is still fully differentiable, and can not be forced to learn a discontinuous function. It is clear that to resolve this issue we have to allow explicitly for a discontinuous behaviour in the neural network topology.

2.3. Incorporating the discontinuity

An intuitive way to introduce a discontinuity in the derivatives of the neural network is to use non-differentiable activation functions (e.g. the rectified linear unit [54]). But there is no obvious reason why a non-differentiability at integer particle numbers will appear—and these networks will most likely become non-differentiable with respect to the density ρ. To overcome this problem we take an alternative route: We define an 'auxiliary function' (AF), which we force to be non-differentiable at all integer particle numbers:

$$u[\rho_\uparrow,\rho_\downarrow] = a + \frac{|\sin(\pi\tilde{n})|}{\tilde{n}+b}\sum_j u_{\mathrm{loc}}^{\theta'}\left[\eta_\uparrow(r_j,\kappa),\eta_\downarrow(r_j,\kappa)\right],$$

where $\tilde{n} = \int(\rho_\uparrow+\rho_\downarrow)\,\mathrm{d}r$ is the (fractional) number of particles and a and b are arbitrary positive constants. Notice that the non-differentiability of the AF comes from the non-differentiability of the function $|\sin(\pi\tilde{n})|$. The local functions $u_{\mathrm{loc}}^{\theta'}$ are obtained by using another SWC neural network. We then replace the functional

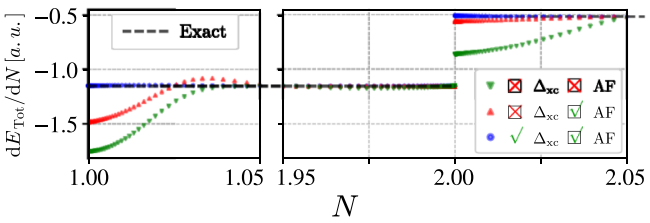

Figure 5. Comparison of the derivatives of the energy with respect to the particle number N for three different models evaluated at a randomly chosen external potential: the first model uses only fractional densities, the second incorporates the exact xc shift in the loss function, and the third one uses a non-differentiable AF, in addition to the exact xc shift and the fractional densities.

Table 1. Mean and median percentage errors gapE between the estimated and the actual gap for each model presented in this section (see equation (17)). The corresponding figure for each model is indicated in the last column.

FDs	AF	Δ_{xc}	$\text{gapE}_{\text{mean}}$	$\text{gapE}_{\text{median}}$	Figure
0	—	—	57.43%	55.61%	Figures 1 and 6(A)
5	—	—	57.46%	53.32%	Figure 5
5	√	—	14.35%	9.69%	Figure 5
5	√	√	7.24%	3.18%	Figures 1, 5 and 6(B)

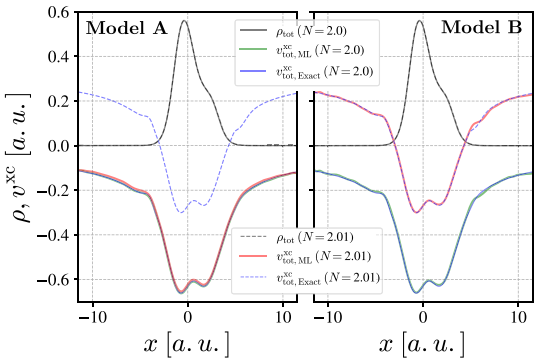

Figure 6. Comparison between the predicted jump of the xc potential at $N = 2$ for Model A and Model B (as explained in the main text). The results correspond to the same external potential used in figure 1.

$E_{xc}[\rho_\uparrow, \rho_\downarrow]$ in equation (3) by $E'_{xc}[\rho_\uparrow, \rho_\downarrow] \equiv E_{xc}[\rho_\uparrow, \rho_\downarrow, u[\rho_\uparrow, \rho_\downarrow]]$, where the additional channel has been appended to the spin channels. Here each point carries the (same) information of the value of the non-differentiable AF as illustrated in figure 4. As we show below, this procedure ensures that the machine can learn the non-differentiable function in an efficient manner.

To study the performance of our approach, figure 5 presents the predicted derivative of the total energy for three different training models: (1) a model trained only with fractional densities, (2) a model using the AF in the network architecture and trained with fractional densities, and (3) a model using the AF, the shift in the loss function and fractional densities. As expected, the first model does not predict the correct derivative discontinuity, despite yielding very good errors at fractional particle numbers. While the second model already demonstrates major improvements, especially at $N = 2$, our most important finding is that the third model does exhibit a remarkable agreement with the exact results. Note that the results shown here correspond to a certain, randomly picked external potential, but it can be seen as representative for a large number of systems as we shall see by analyzing the test set errors presented in table 1. We also trained models without an AF, but with the Δ_{xc}-shift. Those models failed to converge in general, which shows the crucial role of our AF.

In figure 1 the energy as a function of the particle number is displayed for two models: (A) a model trained with integer densities only and (B) a model using fractional densities, the shift in the loss function and the AF. The situation is clearly much better for model B, as the energy is very close to linear between integers and shows a cusp at $N = 2$. As shown in figure 6, this model is also able to reproduce the correct jump in the xc potential: here we display the xc potentials and densities for a randomly selected external potential at 2 and 2.01 electrons. On the left panel of figure 6 we can see that the xc potential of the basic

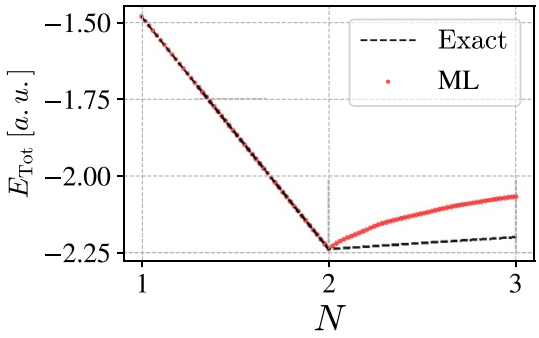

Figure 7. Comparison of the exact total energy and the total energy predicted by our best model (FDs, AF and xc jump incorporated) for the Helium atom as a function of charge.

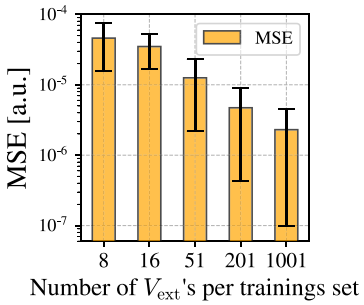

Figure 8. Comparison of the MSEs and variances for models (spin included) based on different numbers of external potentials contained in the training set.

model A is correct at 2 electrons but barely changes when going from 2 to 2.01. Model B, on the other hand, shows the correct uniform shift in its xc potential.

Finally, in figure 7 we illustrate the extrapolation potential of our model. To that effect we plot the exact total energy together with the values stemming from our best model (FDs, AF and xc jump incorporated) for a 1D Helium atom, with total charge going from $+1$ (cation), passing by 0 (neutral) to -1 (anion). It is well known that anions pose a considerable challenge for DFT due to the incorrect asymptotic behavior of most semi-local (and even hybrid) xc functionals [55]. Furthermore, we did not include any anion on the training set, so this is a proper extrapolation example. On a positive side, we can see from the figure that our model is correctly capable of binding the extra electron. However, and due to the well-known poor extrapolation capabilities of neural networks, the error in the total energy in the interval $[2, 3]$ is visibly larger than in the interval $[1, 2]$.

2.4. Dependency on the number of samples per training set

For the sake of completeness, we briefly demonstrate that the testing errors (MSE) decrease when increasing the number of physical systems in the training set. The results are illustrated in figure 8. Here, we trained models (spin included) with 8, 16, 51, 201, and 1001 external potentials per training set. Since all models have been trained with $\epsilon = (0.05, 0.2, 0.5, 0.8, 0.95)$ fractional densities, one v_{ext} per training sets equals 13 physical systems with different number of electrons. As expected, the errors decrease when adding more samples into the training data.

3. Conclusions

We trained a neural network as an exchange correlation functional that (a) depends explicitly on the number of particles; (b) yields total energies that are piece-wise linear between the integers and (c) reproduces the infamous derivative discontinuity of the exchange-correlation energy with remarkable accuracy. To do so we extended the sliding window convolution algorithm to systems with fractional number of particles, and developed a non-differentiable auxiliary function that allows the network to learn correctly the derivative discontinuity. The most efficient way to train a model yielding highly accurate predictions for the energy in the entire range of particle numbers is by (a) adding multiple fractional densities into the training data, (b)

training for the correct shift in the xc potential at integer particle numbers, and (c) incorporating an auxiliary function with a discontinuous derivative with respect to the particle number at integers.

Our work pushes forward the on-going research on machine-learning functionals by incorporating the correct physics into the training process, thereby improving the path towards an exact functional [56]. We expect the most important finding of our work (namely, the crucial role played by the auxiliary function) to be easily transferable to 3D machine-learning functionals and therefore to play an important role in future developments in machine-learning DFT. As an outlook, we think it will be important to incorporate and test our results for realistic 3D systems or include in the training data diffusive anion states in the sense of [41, 42]. We also expect that our results can stimulate research on similar problems for ensemble DFT (understood as mixtures of ground and excites states), orbital free DFT or functional theories of reduced density matrices [57–62].

4. Methods

In this section we present all methods and computational details necessary to arrive at our results. First we discuss the generation of the training data and second the details of the network implementation, and the definition of the mean errors discussed and presented above.

4.1. Exact calculations

To train and test our models we created a set of 1-dimensional exact calculations and Kohn–Sham inversions. We sampled 1500 external (Coulomb-) potentials and computed the exact electronic ground state densities as well as the corresponding energies by solving the eigenvalue problem for the electronic Hamiltonian

$$H = -\sum_{i=1}^{N}\frac{\nabla_i^2}{2} - \sum_{j=1}^{K}\sum_{i=1}^{N}\frac{Z_j}{\sqrt{1+|R_j-r_i|^2}} + \sum_{i<j}\frac{1}{\sqrt{1+|r_i-r_j|^2}}, \tag{7}$$

where K is the total number of nuclei, the variables R_j and Z_j denote the position and charge of the jth nuclei respectively, $N \in \{1,2,3\}$ is the total number of electrons, and r_i is the position of the ith electron. We solve the exact ground-state problem with Octopus [63], using a grid spacing of 0.1 a.u. and a box size of 23 a.u. (leading to a grid with 231 points). To circumvent the integrability problem of the Coulomb interaction in 1D we used a softened interaction. The total number of nuclei K were set to be 1, 2 or 3, such that their individual charges satisfy $\sum_k Z_k = 3$. Their positions were randomly distributed with $|R_k| \leqslant 4$ a.u.

4.2. Spin densities

As Octopus [63] only provides directly spin-densities and wave-functions for two-particle systems we now discuss the problem of obtaining the spin densities for the three particle systems.

We start by noticing that solving the eigenvalue problem $\hat{H}\Phi = E\Phi$ for the Hamiltonian (7) yields all many-particle solutions, including both fermionic and bosonic states. A spin-adapted fermionic solution of the form $\Psi(r_1\,\sigma_1,\ldots,r_N\sigma_N)$ can be obtained by projecting any spatial solution Φ on the Young diagrams belonging to certain spin quantum numbers (S,M), with S being the total spin and $M = \sum_i \sigma_i$ [64, 65]. For a detailed description we refer to [64, 66, 67]. Let us denote a set of f primitive, degenerated, but orthogonal spin functions as $\{X(N,S,M;i)\}_{i=1\ldots f}$. A permutation $\mathbf{P}$ acting on a spin function can be expressed as a linear combination of all primitive spin functions:

$$\mathbf{P}X(N,S,M;i) = \sum_{j=1}^{f} X(N,S,M;j)U(P)^S_{ji}. \tag{8}$$

The expansion coefficients $U(P)^S_{ji}$ can be calculated using the orthogonality of $X(N,S,M;j)$ [68]. By taking into account the antisymmetrization $\mathcal{A} = \frac{1}{\sqrt{N!}}\sum_P(-1)^P\mathbf{P}$ of the product of the spin and spatial parts $\Phi X(N,S,M;i)$ one obtains a sum of products of spatial and spin functions, as follows:

$$\begin{aligned}\Psi_i &= \mathcal{A}\Phi X(N,S,M;i)\\ &= \frac{1}{\sqrt{N!}}\sum_P(-1)^p\mathbf{P}^r\Phi\mathbf{P}^\sigma X(N,S,M;i)\\ &= \frac{1}{\sqrt{f}}\sum_{j=1}^{f}X(N,S,M;j)\Phi^S_{ji},\end{aligned} \tag{9}$$

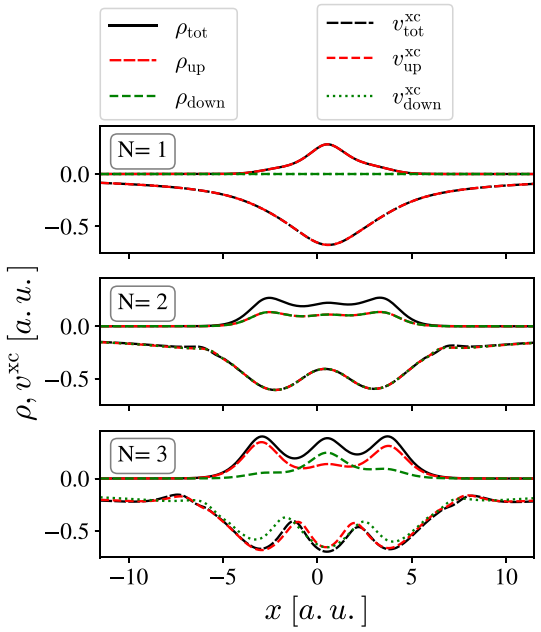

Figure 9. Densities and corresponding xc potentials for a certain external potential for different integer particle numbers.

where $\mathbf{P}^r$ and $\mathbf{P}^\sigma$ denote that the permutations operates on spatial and spin coordinates respectively, and

$$\Phi^S_{ji} = \sqrt{\frac{f}{N!}} \sum_P U(P)^S_{ji} (-1)^p \mathbf{P}^r \Phi\left(\mathbf{r}_1, \ldots, \mathbf{r}_N\right). \tag{10}$$

For the ground state of $N = 3$ two linear independent spin-eigenfunctions can be chosen:

$$X(3, 1/2, 1/2; 1) = \frac{1}{\sqrt{6}} \left[2\uparrow\uparrow\downarrow - (\uparrow\downarrow\uparrow + \downarrow\uparrow\uparrow)\right], \tag{11a}$$

$$X(3, 1/2, 1/2; 2) = \frac{1}{\sqrt{2}} \left(\uparrow\downarrow\uparrow - \downarrow\uparrow\uparrow\right). \tag{11b}$$

The (normalized) spatial parts in equation (9) and the corresponding spin densities $\rho_\uparrow(x)$ and $\rho_\downarrow(x)$ can then be found. For instance,

$$\begin{aligned}\rho_\uparrow(x) = &\iint \left[5|\tilde{\Phi}_1|^2 + 9|\tilde{\Phi}_2|^2\right] (\mathrm{d}x_2 \mathrm{d}x_3 + \mathrm{d}x_1 \mathrm{d}x_3) \\ &+ \iint \left[2|\tilde{\Phi}_1|^2 + 18|\tilde{\Phi}_2|^2\right] \mathrm{d}x_1 \mathrm{d}x_2,\end{aligned} \tag{12}$$

with $\tilde{\Phi}_1 = 2\,\Phi^{S=1/2}_{11}$ and $\tilde{\Phi}_2 = 2\,\Phi^{S=1/2}_{21}/\sqrt{3}$.

4.3. Kohn–Sham inversion

For the inversion of the densities we used the optimization algorithm proposed in [51], that casts the inverse DFT problem of finding the $v_{\mathrm{xc}}(\boldsymbol{r})$ that yields a given density $\rho(\boldsymbol{r})$ as a constrained optimization problem. The conjugate gradient method was used to update the xc potentials $v^\sigma_{\mathrm{xc}}(\boldsymbol{r})$. A constant weight function $w \equiv 1$ was also used.

The densities of $N = 1, 2, 3$ particles were mixed, allowing us to generate a fractional densities for each external potential. For instance, the set $\{1, 1.5, 2, 2.5, 3\}$ contains additional $\epsilon = 0.5$ fractional densities beside the integer ones. If the inversion algorithm did not converge to a MSE below 1.5×10^{-7} for a given density, we removed all samples corresponding to that external potential. The computed xc potentials were shifted by a constant to be in agreement with Koopman's theorem [69].

In figure 9 we show a few examples of such spin-densities and inverted xc potentials. For the fractional densities plotted in figure 10 the shift of the xc potential caused by the Δ_{xc} jump is clearly visible.

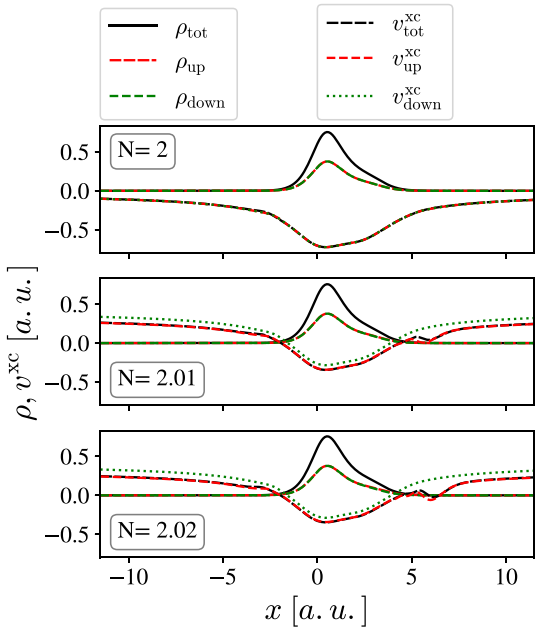

Figure 10. Densities and corresponding xc potentials for a certain external potential for different fractional particle numbers.

4.4. Network implementation

The neural networks were implemented in pytorch [70] using pytorch-lightning [71] to simplify the training process. As discussed before we used the sliding window convolution as proposed in [14]. We also attempted to use fully convolutional networks but were not able to achieve the same performance with them. For each network the number of zero-paddings at the system's boundaries was set to $(\kappa-1)/2$ (κ is an odd number in our calculations). For all models presented in this paper, we chose a kernel size of $\kappa = 201$, corresponding to highly nonlocal funtionals (see the exact calculations in the section 4). Each local density was fed into a fully connected network with SILU [72] activation functions, and the output layer returned the *local* functions described in the previous sections. The use of SILU activation functions guaranteed the smoothness of the exchange correlation potential and its higher order derivatives. All hidden layer sizes were set to 32. Including the window convolution we used four hidden layers. The additional SWC unit for the AF shared the same hyperparameters. The network weights were optimized with ADAM while using a cyclic learning rate scheduler. We used a learning rate of 7×10^{-4} with a batch size of 30. The only exceptions were models trained with the xc-jump in the loss function. In this case each batch contained one density, corresponding to a random external potential. If the Δ_{xc}-shift was incorporated in the loss function, the densities per batch consisted of *all* fractional densities of a certain external potential. For these networks we kept the learning rate used before, but changed the batch size to one. The models were trained for 30 000 epochs and the model with the best validation loss was selected for testing.

4.5. Mean errors

In this section we discuss all the errors discussed in the main body of the paper. For a set of M external potentials $\{v^{(1)}_{\mathrm{ext}},\dots,v^{(M)}_{\mathrm{ext}}\}$ the 'averaged mean squared error' (MSE) is defined as:

$$\mathrm{MSE} = \frac{1}{KM}\sum_{m=1}^{M}\sum_{k=1}^{K}\left[\Delta E_{\mathrm{tot}}(\tilde{n}_k; v^{(m)}_{\mathrm{ext}})\right]^2, \tag{13}$$

and the 'averaged variance' (Var) is defined as:

$$\mathrm{Var}^2 = \frac{1}{KM}\sum_{m=1}^{M}\sum_{k=1}^{K}\left[\Delta E_{\mathrm{tot}}(\tilde{n}_k; v^{(m)}_{\mathrm{ext}}) - \mu(v^{(m)}_{\mathrm{ext}})\right]^2, \tag{14}$$

where

$$\Delta E_{\mathrm{tot}}(\tilde{n}_k; v^{(m)}_{\mathrm{ext}}) = E^{\mathrm{ML}}_{\mathrm{tot}}(\tilde{n}_k; v^{(m)}_{\mathrm{ext}}) - E^{\mathrm{Exact}}_{\mathrm{tot}}(\tilde{n}_k; v^{(m)}_{\mathrm{ext}})$$

is the difference between the predicted energy and the exact energy and $\mu(v_{\text{ext}}^{(m)}) = \overline{\Delta E_{\text{tot}}}(v_{\text{ext}}^{(m)})$ is the mean value averaged over particle numbers corresponding to the densities belonging to a certain external potential $v_{\text{ext}}^{(m)}$. K denotes the number of equally distributed densities (with corresponding particle number $\tilde{n}_k$) (in the range $[1,3]$). The averaged MSEs and variances for a sufficiently large number of external potentials provide information about the accuracy of the prediction of the total energy for arbitrary fractional densities in general.

The computation of the derivative of the total energy with respect to the particle number offers the opportunity to estimate the success of including the AF in the network or the xc jump in the training. To that aim we have analyzed the gap

$$\Delta_\epsilon(N; v_{\text{ext}}^{(m)}) := \left.\frac{\partial E_{\text{Tot}}(v_{\text{ext}}^{(m)})}{\partial n}\right|_{n=N+\epsilon} - \left.\frac{\partial E_{\text{Tot}}(v_{\text{ext}}^{(m)})}{\partial n}\right|_{n=N-\epsilon}, \tag{15}$$

where $\epsilon > 0$, by means of the more appropriate numerical derivative

$$\left.\frac{\partial E_{\text{Tot}}}{\partial n}\right|_{n=\tilde{n}} \approx \left(\frac{E_{\text{Tot}}(\tilde{n}+\Delta; v_{\text{ext}}^{(m)}) - E_{\text{Tot}}(\tilde{n}-\Delta; v_{\text{ext}}^{(m)})}{2\,\Delta}\right), \tag{16}$$

where $\tilde{n}$ is a computed fractional particle number in the interval $[1,3]$ and Δ is the corresponding spacing. Since a non-differentiability at integer particle numbers occur in theory, the numerical derivatives (whose derivation is based on the assumption of having fully differentiable functions) have been computed separately on open intervals $(N, N+1)$. Using equations (15) and (16) we are able to evaluate the 'mean percentage error' between the estimated and the actual gap, namely:

$$\begin{aligned}&\text{gapE}(N,\epsilon)_{\text{mean}}\\ &= \frac{1}{M}\sum_{m=1}^{M}\left|\frac{\Delta_\epsilon^{\text{ML}}(N; v_{\text{ext}}^{(m)}) - \Delta_\epsilon^{\text{Exact}}(N; v_{\text{ext}}^{(m)})}{\Delta_\epsilon^{\text{Exact}}(N; v_{\text{ext}}^{(m)})}\right|,\end{aligned} \tag{17}$$

where $\Delta_\epsilon^{\text{Exact}}(N; v_{\text{ext}}^{(m)})$ is just the difference between ionization potential and electron affinity according to equation (2). We explicitly underscored the denotation of the error by "mean", since we will additionally evaluate $\text{gapE}(N,\epsilon)_{\text{median}}$, which is the midpoint of the sorted distribution consisting of all the summands in equation (17). For the evaluation of all the errors described before we chose $K = 100$ and $M = 100$. We estimate the gap errors at $N = 2$ with $\epsilon = 5 \times 10^{-9}$, i.e. $\text{gapE}(N = 2, \epsilon = 5 \times 10^{-9})$.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.5091466.

Acknowledgment

C L B-R acknowledges funding from "BiGmax", the Max Planck Society's research network on big-data-driven materials science.

Code availability

All codes to reproduce, examine, and improve our proposed analysis are freely available online [73].

Conflict of interest

The authors declare no competing interests.

ORCID iDs

Johannes Gedeon https://orcid.org/0000-0002-5060-7535
Jonathan Schmidt https://orcid.org/0000-0001-5685-6404
Matthew J P Hodgson https://orcid.org/0000-0002-2256-6860
Jack Wetherell https://orcid.org/0000-0003-1190-3830
Carlos L Benavides-Riveros https://orcid.org/0000-0001-6924-727X
Miguel A L Marques https://orcid.org/0000-0003-0170-8222

References

[1] Hohenberg P and Kohn W 1964 Inhomogeneous electron gas *Phys. Rev.* **136** B864
[2] Jones R O 2015 Density functional theory: its origins, rise to prominence and future *Rev. Mod. Phys.* **87** 897
[3] Kohn W and Sham L 1965 Self-consistent equations including exchange and correlation effects *Phys. Rev.* **140** A1133
[4] Lehtola S, Steigemann C, Oliveira M J T and Marques M A L 2018 Recent developments in LIBXC—a comprehensive library of functionals for density functional theory *SoftwareX* **7** 1
[5] Perdew J and Schmidt K 2001 Jacob's ladder of density functional approximations for the exchange-correlation energy *AIP Conf. Proc.* **577** 1
[6] Deng L and Li X 2013 Machine learning paradigms for speech recognition: an overview *IEEE Trans. Audio Speech Lang. Process.* **21** 1060
[7] Kalita B, Li L, McCarty R J and Burke K 2021 Learning to approximate density functionals *Acc. Chem. Res.* **54** 818
[8] Brockherde F, Vogt L, Li L, Tuckerman M, Burke K and Müller K-R 2017 Bypassing the Kohn–Sham equations with machine learning *Nat. Commun.* **8** 872
[9] Li L, Hoyer S, Pederson R, Sun R, Cubuk E, Riley P and Burke K 2021 Kohn–Sham equations as regularizer: building prior knowledge into machine-learned physics *Phys. Rev. Lett.* **126** 036401
[10] Margraf J and Reuter K 2021 Pure non-local machine-learned density functional theory for electron correlation *Nat. Commun.* **12** 344
[11] Moreno J R, Carleo G and Georges A 2020 Deep learning the Hohenberg–Kohn maps of density functional theory *Phys. Rev. Lett.* **125** 076402
[12] Ryczko K, Strubbe D A and Tamblyn I 2019 Deep learning and density-functional theory *Phys. Rev.* A **100** 022512
[13] Lubasch M, Fuks J I, Appel H, Rubio A, Cirac J I and Bañuls M-C 2016 Systematic construction of density functionals based on matrix product state computations *New J. Phys.* **18** 083039
[14] Schmidt J, Benavides-Riveros C L and Marques M A L 2019 Machine learning the physical nonlocal exchange-correlation functional of density-functional theory *J. Phys. Chem. Lett.* **10** 6425
[15] Nagai R, Akashi R and Sugino O 2020 Completing density functional theory by machine learning hidden messages from molecules *npj Comput. Mater.* **6** 43
[16] Denner M M, Fischer M H and Neupert T 2020 Efficient learning of a one-dimensional density functional theory *Phys. Rev. Res.* **2** 033388
[17] Lieb E 1983 Density functionals for Coulomb systems *Int. J. Quantum Chem.* **24** 243
[18] Ludeña E and Karasiev V 2002 *Kinetic Energy Functionals: History, Challenges and Prospects* (Singapore: World Scientific) pp 612–65
[19] Kraisler E and Kronik L 2013 Piecewise linearity of approximate density functionals revisited: implications for frontier orbital energies *Phys. Rev. Lett.* **110** 126403
[20] Sham L J and Schlüter M 1983 Density-functional theory of the energy gap *Phys. Rev. Lett.* **51** 1888
[21] Perdew J, Parr R, Levy M and Balduz J 1982 Density-functional theory for fractional particle number: derivative discontinuities of the energy *Phys. Rev. Lett.* **49** 1691
[22] Grüning M, Marini A and Rubio A 2006 Effect of spatial nonlocality on the density functional band gap *Phys. Rev.* B **74** 161103
[23] Andrade X and Aspuru-Guzik A 2011 Prediction of the derivative discontinuity in density functional theory from an electrostatic description of the exchange and correlation potential *Phys. Rev. Lett.* **107** 183002
[24] Hodgson M J P, Kraisler E, Schild A and Gross E K U 2017 How interatomic steps in the exact Kohn–Sham potential relate to derivative discontinuities of the energy *J. Phys. Chem. Lett.* **8** 5974
[25] Mirtschink A, Seidl M and Gori-Giorgi P 2013 Derivative discontinuity in the strong-interaction limit of density-functional theory *Phys. Rev. Lett.* **111** 126402
[26] Mori-Sánchez P and Cohen A 2014 The derivative discontinuity of the exchange–correlation functional *Phys. Chem. Chem. Phys.* **16** 14378
[27] Mosquera M and Wasserman A 2014 Derivative discontinuities in density functional theory *Mol. Phys.* **112** 2997
[28] Baerends E J 2020 On derivatives of the energy with respect to total electron number and orbital occupation numbers. A critique of Janak's theorem *Mol. Phys.* **118** e1612955
[29] Perdew J 1985 What do the Kohn–Sham orbital energies mean? How do atoms dissociate? *Density Functional Methods In Physics* ed R Dreizler and J da Providência (New York: Plenum) pp 265–308
[30] Cohen A J, Mori-Sánchez P and Yang W 2012 Challenges for density functional theory *Chem. Rev.* **112** 289
[31] Yang W, Zhang Y and Ayers P 2000 Degenerate ground states and a fractional number of electrons in density and reduced density matrix functional theory *Phys. Rev. Lett.* **84** 5172
[32] Perdew J and Levy M 1983 Physical content of the exact Kohn–Sham orbital energies: band gaps and derivative discontinuities *Phys. Rev. Lett.* **51** 1884
[33] Eich F and Hellgren M 2014 Derivative discontinuity and exchange-correlation potential of meta-GGAs in density-functional theory *J. Chem. Phys.* **141** 224107
[34] Kümmel S and Kronik L 2008 Orbital-dependent density functionals: theory and applications *Rev. Mod. Phys.* **80** 3
[35] Ruzsinszky A, Perdew J, Csonka G, Vydrov O and Scuseria G 2006 Spurious fractional charge on dissociated atoms: pervasive and resilient self-interaction error of common density functionals *J. Chem. Phys.* **125** 194112
[36] Perfetto E and Stefanucci G 2012 Missing derivative discontinuity of the exchange-correlation energy for attractive interactions: the charge Kondo effect *Phys. Rev.* B **86** 081409
[37] Li C and Yang W 2017 On the piecewise convex or concave nature of ground state energy as a function of fractional number of electrons for approximate density functionals *J. Chem. Phys.* **146** 074107
[38] Slater J C and Johnson K H 1972 Self-consistent-field $X\alpha$ cluster method for polyatomic molecules and solids *Phys. Rev.* B **5** 844
[39] Ferreira L G, Marques M and Teles L K 2008 Approximation to density functional theory for the calculation of band gaps of semiconductors *Phys. Rev.* B **78** 125116
[40] Baerends E J 2018 Density functional approximations for orbital energies and total energies of molecules and solids *J. Chem. Phys.* **149** 054105
[41] Peach M J G, Teale A M, Helgaker T and Tozer D J 2015 Fractional electron loss in approximate DFT and Hartree–Fock theory *J. Chem. Theory Comput.* **11** 5262–8
[42] Anderson L N, Oviedo M B and Wong B M 2017 Accurate electron affinities and orbital energies of anions from a nonempirically tuned range-separated density functional theory approach *J. Chem. Theory Comput.* **13** 1656–66

[43] Hait D and Head-Gordon M 2018 Delocalization errors in density functional theory are essentially quadratic in fractional occupation number *J. Phys. Chem. Lett.* **9** 6280–8
[44] Cohen A, Mori-Sánchez P and Yang W 2008 Insights into current limitations of density functional theory *Science* **321** 792
[45] Mori-Sánchez P, Cohen A J and Yang W 2008 Localization and delocalization errors in density functional theory and implications for band-gap prediction *Phys. Rev. Lett.* **100** 146401
[46] Benavides-Riveros C L, Lathiotakis N N and Marques M A L 2017 Towards a formal definition of static and dynamic electronic correlations *Phys. Chem. Chem. Phys.* **19** 12655
[47] Wagner L O, Stoudenmire E M, Burke K and White S R 2012 Reference electronic structure calculations in one dimension *Phys. Chem. Chem. Phys.* **14** 8581
[48] Loos P-F, Ball C and Gill P 2015 Chemistry in one dimension *Phys. Chem. Chem. Phys.* **17** 3196
[49] Kreibich T, Lein M, Engel V and Gross E K U 2001 Even-harmonic generation due to beyond-Born–Oppenheimer dynamics *Phys. Rev. Lett.* **87** 103901
[50] Baker T E, Stoudenmire E M, Wagner L O, Burke K and White S R 2015 One-dimensional mimicking of electronic structure: the case for exponentials *Phys. Rev.* B **91** 235141
[51] Kanungo B, Zimmerman P M and Gavini V 2019 Exact exchange-correlation potentials from ground-state electron densities *Nat. Commun.* **10** 4497
[52] von Barth U and Hedin L 1972 A local exchange-correlation potential for the spin polarized case. I *J. Phys.* C **5** 1629
[53] Hellgren M and Gross E K U 2012 Discontinuities of the exchange-correlation kernel and charge-transfer excitations in time-dependent density-functional theory *Phys. Rev.* A **85** 022514
[54] Nair V and Hinton G 2010 Rectified linear units improve restricted Boltzmann machines *Proc. 27th Int. Conf. on Machine Learning (ICML'10)* vol 27 pp 807–14
[55] Jensen F 2010 Describing anions by density functional theory: fractional electron affinity *J. Chem. Theory Comput.* **6** 2726–35
[56] Medvedev M, Bushmarinov I, Sun J, Perdew J and Lyssenko K 2017 Density functional theory is straying from the path toward the exact functional *Science* **355** 49
[57] Senjean B and Fromager E 2018 Unified formulation of fundamental and optical gap problems in density-functional theory for ensembles *Phys. Rev.* A **98** 022513
[58] Loos P-F and Fromager E 2020 A weight-dependent local correlation density-functional approximation for ensembles *J. Chem. Phys.* **152** 214101
[59] Benavides-Riveros C L, Wolff J, Marques M A L and Schilling C 2020 Reduced density matrix functional theory for bosons *Phys. Rev. Lett.* **124** 180603
[60] Kraisler E and Schild A 2020 Discontinuous behavior of the Pauli potential in density functional theory as a function of the electron number *Phys. Rev. Res.* **2** 013159
[61] Cioslowski J 2020 Off-diagonal derivative discontinuities in the reduced density matrices of electronic systems *J. Chem. Phys.* **153** 154108
[62] Schmidt J, Fadel M and Benavides-Riveros C L 2021 Machine learning universal bosonic functionals *Phys. Rev. Res.* **3** L032063
[63] Andrade X *et al* 2015 Real-space grids and the Octopus code as tools for the development of new simulation approaches for electronic systems *Phys. Chem. Chem. Phys.* **17** 31371
[64] Pauncz R 2000 *The Construction of Spin Eigenfunctions* (Boca Raton, FL: CRC Press)
[65] Whitfield J 2013 Communication: spin-free quantum computational simulations and symmetry adapted states *J. Chem. Phys.* **139** 021105
[66] Greiner W 1994 *Quantum Mechanics. Symmetries* (Berlin: Springer)
[67] McWeeny R 1992 *Methods of Molecular Quantum Mechanics* (New York: Academic)
[68] Porter F 2009 *Group Theory* (available at: www.hep.caltech.edu/~fcp/math/)
[69] Gritsenko O and Baerends E 2002 The analog of Koopmans' theorem in spin-density functional theory *J. Chem. Phys.* **117** 9154
[70] Paszke A *et al* 2019 Pytorch: an imperative style, high-performance deep learning library *Advances in Neural Information Processing Systems 32* (Curran Associates, Inc.) pp 8024–35
[71] Falcon W A *et al* 2019 Pytorch lightning *GitHub* (available at: https://github.com/PyTorchLightning/pytorch-lightning)
[72] Hendrycks D and Gimpel K 2020 Gaussian error linear units (GELUs) (arXiv:1606.08415)
[73] Gedeon J *et al* 2021 EnsembleDFT-ML Github (available at: https://github.com/JoGed/EnsembleDFT-ML)