Neural network distillation of orbital dependent density functional theory

Matija Medvidović [1,*] Jaylyn C. Umana [2,3,4] Iman Ahmadabadi [5,2,6] Domenico Di Sante [7] Johannes Flick [3,4,2] and Angel Rubio [8,2,9]

[1]Institute for Theoretical Physics, ETH Zürich, CH-8093 Zürich, Switzerland
[2]Center for Computational Quantum Physics, Flatiron Institute, 162 5th Avenue, New York, New York 10010, USA
[3]Department of Physics, City College of New York, New York, New York 10031, USA
[4]Department of Physics, The Graduate Center, City University of New York, New York, New York 10016, USA
[5]Joint Quantum Institute, NIST and University of Maryland, College Park, Maryland 20742, USA
[6]Department of Chemistry, Princeton University, Princeton, New Jersey 08544, USA
[7]Department of Physics and Astronomy, University of Bologna, I-40127 Bologna, Italy
[8]Max Planck Institute for the Structure and Dynamics of Matter, Luruper Chaussee 149, D-22761 Hamburg, Germany
[9]Initiative for Computational Catalysis, Flatiron Institute, 162 5th Avenue, New York, New York 10010, USA

(Received 29 October 2024; accepted 8 April 2025; published 2 May 2025)

Density functional theory (DFT) offers a desirable balance between quantitative accuracy and computational efficiency in practical many-electron calculations. Its central component, the exchange-correlation energy functional, has been approximated with increasing levels of complexity ranging from strictly local approximations to nonlocal and orbital dependent expressions with many tuned parameters. In this paper, we formulate a general way of rewriting complex density functionals using deep neural networks in a way that allows for simplified computation of Kohn-Sham potentials as well as higher functional derivatives through automatic differentiation, enabling access to highly nonlinear response functions and forces. These goals are achieved by using a recently developed class of robust neural network models capable of modeling functionals, as opposed to functions, with explicitly enforced spatial symmetries. Functionals treated in this way are then called *global density approximations* and can be seamlessly integrated with existing DFT workflows. Tests are performed for a dataset featuring a large variety of molecular structures and popular meta–generalized gradient approximation density functionals, where we successfully eliminate orbital dependencies coming from the kinetic energy density, and discover a high degree of transferability to a variety of physical systems. The presented framework is general and could be extended to more complex orbital and energy dependent functionals as well as refined with specialized datasets.

DOI: 10.1103/PhysRevResearch.7.023113

I. INTRODUCTION

The many-electron problem has been central to quantum physics for decades. With exact solutions out of reach in most cases, different approximate methods have been used in their place, offering controlled trade-offs between efficiency, accuracy, and applicability. Density functional theory (DFT) [1–5] should be contrasted with the accuracy of wave-function methods such as full configuration interaction (FCI) [6], coupled cluster [7,8], quantum Monte Carlo (QMC) [9–12], and Green's function methods such as dynamical mean-field theory (DMFT) [13–18] due to its considerably lower computational cost while still capturing the essential physics in most cases. This feature often makes DFT the only method that can access many-electron physics at large scales becoming the method of choice in solid state physics, quantum chemistry, and material science [19].

While the existence of the exact energy functional mapping from the electron densities to energies has been proven [1], its explicit form remains unknown. Approximate forms of the exchange-correlation (XC) energy functional have been constructed at varying levels of complexity [2,20–28]. More recently, expressive machine learning methods have been used to build representations of XC functionals from data [29–41], showing that achieving chemically accurate, efficient, and transferable functionals is feasible, at least within a targeted domain.

In this paper, we propose a method of constructing fully machine-learned nonlocal XC functionals based on neural networks, accurately reproducing known functionals. There are two benefits of constructing a neural network copy of known functionals. First, for meta–generalized gradient approximation (meta-GGA) functionals, we eliminate the orbital dependence of the kinetic contribution to the XC energy by directly controlling data generation and neural network inputs during training. Second, higher-order functional derivatives

*Contact author: mmedvidovic@ethz.ch

of the resulting XC functionals can easily be evaluated using automatic differentiation (AD) tools [42,43]. We call this method the global density approximation (GDA).

The neural network model used in this work is built around recent progress in transformer models that have recently revolutionized natural language and image processing [44,45]. We employ the linear version of the underlying attention mechanism [46–48] to customize the network architecture for functional learning while respecting underlying spatial symmetries.

The GDA scheme results in an approximate but pure density functional. The main contribution of this work is the regularized training scheme allowing us to construct the GDA functional using only density-energy pairs as a part of the training dataset. As a direct result, GDA approximations are generalizable between molecules and can be used in independent self-consistent field (SCF) calculations without retraining.

II. METHODS

A. The global functional distillation

Consider an isolated molecule with N electrons in an external potential $v_{\text{ext}}(\mathbf{r})$. We limit the discussion to isolated molecules in this paper but the discussion is equally applicable to ensembles of molecules and other quantum systems. A standard DFT calculation [3,4,49] outputs an approximation to the ground state density n_0 minimizing the total energy: $n_0(\mathbf{r}) = \text{argmin}_n E[n]$. The total energy functional $E[n]$ is commonly written as

$$E[n] = T[n] + E_{\text{ext}}[n] + E_H[n] + E_{\text{xc}}[n], \tag{1}$$

where the external contribution $E_{\text{ext}}[n]$ captures the effects of the external nuclear potential v_{ext} and the direct Hartree component E_H is directly computable given the density $n(\mathbf{r})$. We use atomic units throughout.

A successful evaluation of Eq. (1) depends on approximating the unknown kinetic and exchange-correlation (XC) functionals, $T[n]$ and $E_{\text{xc}}[n]$.. The Kohn-Sham (KS) DFT [2] framework approximates the ground state density as induced by an effective system of noninteracting electrons, $n(\mathbf{r}) = \sum_a n_a |\psi_a(\mathbf{r})|^2$, where n_a is the occupation of the single-particle orbital $\psi_a(\mathbf{r})$ with energies ϵ_a. Crucially, orbitals also allow for an approximate treatment of the kinetic energy contribution as $T = \int \mathrm{d}^3\mathbf{r}\,\tau(\mathbf{r})$ with the kinetic energy density τ given by

$$\tau(\mathbf{r}) = \frac{1}{2}\sum_a n_a |\nabla \psi_a(\mathbf{r})|^2. \tag{2}$$

For an overview of KS-DFT, we refer the readers to the Supplemental Material [50] and Refs. [3,4,49,51,52].

After the desired XC functional has been specified, the constrained minimization of the total energy functional given in Eq. (1) can proceed, yielding Kohn-Sham equations [2] outlined in the Supplemental Material [50]. These equations have to be solved in a self-consistent (SCF) manner, ensuring that $n(\mathbf{r}) = \sum_a n_a |\psi_a(\mathbf{r})|^2$ holds at all times.

The exact energy functional is unknown. However, approximations with increasing levels of complexity have been ordered into the so-called *Jacob's ladder* [20]. Higher rungs capture more details of local density neighborhoods at increased computational cost, starting with the local density approximation (LDA) which approximates electron XC effects as a uniform gas of interacting electrons of density $n(\mathbf{r})$ [53]. While providing a good approximation for systems with slowly varying densities, it often lacks accuracy for molecular systems. This motivated the development of the second rung, the generalized gradient approximation (GGA) [24,54], which includes local gradients $\nabla n(\mathbf{r})$ and $\nabla^2 n(\mathbf{r})$ as a local variable, offering better accuracy for a wider range of chemical systems.

Even higher rungs host the meta-GGA functionals [28,55–64], which capture more complex electronic interactions and offer improved accuracy for diverse systems, especially in terms of chemical reactivity and band gaps. These benefits come from including new local variables such as $\tau(\mathbf{r})$ and its orbital dependence via Eq. (2).

Mathematically, the capacity to numerically solve the KS equations and access observables relies on our ability to approximate functional derivatives of the total energy in Eq. (1) during the SCF loop. In this paper, we propose an approximation to efficiently estimate functional derivatives of common meta-GGA density functionals by systematically removing orbital dependence. We do this by fitting them to expressive parametrized neural network models with restricted input variables. After capturing the target functional to a satisfactory degree, derivatives of the model can be efficiently calculated using automatic differentiation tools [42,43,65].

We rewrite the kinetic energy density as $\tau \approx \tau_\theta[n]$, a nonlocal functional captured by an expressive deep neural network [66–68] with θ indicating a set of all free parameters in neural network subcomponents (see Supplemental Material [50] for details). This approximation allows us to trivially rewrite any meta-GGA functional as

$$E_{\text{xc}}^{\text{GDA}}[n] = \int \mathrm{d}^3\mathbf{r}\, n(\mathbf{r}) \varepsilon_{\text{xc}}(n(\mathbf{r}), \nabla n(\mathbf{r}), \tau_\theta[n](\mathbf{r})), \tag{3}$$

turning it into a nonlocal density functional.

B. The neural network model

In Eq. (3), we parametrize the kinetic energy density $\tau_\theta[n]$ as a deep neural network, capable of approximating functionals in a controlled way while respecting the underlying spatial symmetries. Additionally, to satisfy known exact scaling laws, we model the local dimensionless enhancement factor with a parametrized form $\phi_\theta[n](\mathbf{r})$, such that

$$\tau_\theta(\mathbf{r}) = \tau_W(\mathbf{r}) + e^{\phi_\theta(\mathbf{r})}(\tau_U(\mathbf{r}) + \eta\,\tau_W(\mathbf{r})) \tag{4}$$

is represented by a neural network with parameters $\theta \in \mathbb{R}^P$, based on a dimensionless kinetic energy indicator variable used in Ref. [62]. In Eq. (4), $\tau_U = \frac{3}{10}(3\pi^2)^{2/3} n^{5/3}$ is the uniform electron gas kinetic energy, and $\tau_W = |\nabla n|^2/8n$ is the von Weizsäcker [69] kinetic functional. The regularization parameter was set to $\eta = 10^{-3}$, as defined in Ref. [62] and informed by normalization heuristics [70,71] based on exact values of $\tau(\mathbf{r})$ in the dataset.

Parameterizing only dimensionless enhancement factors in Eq. (4) allows us to recover the correct scaling laws for the

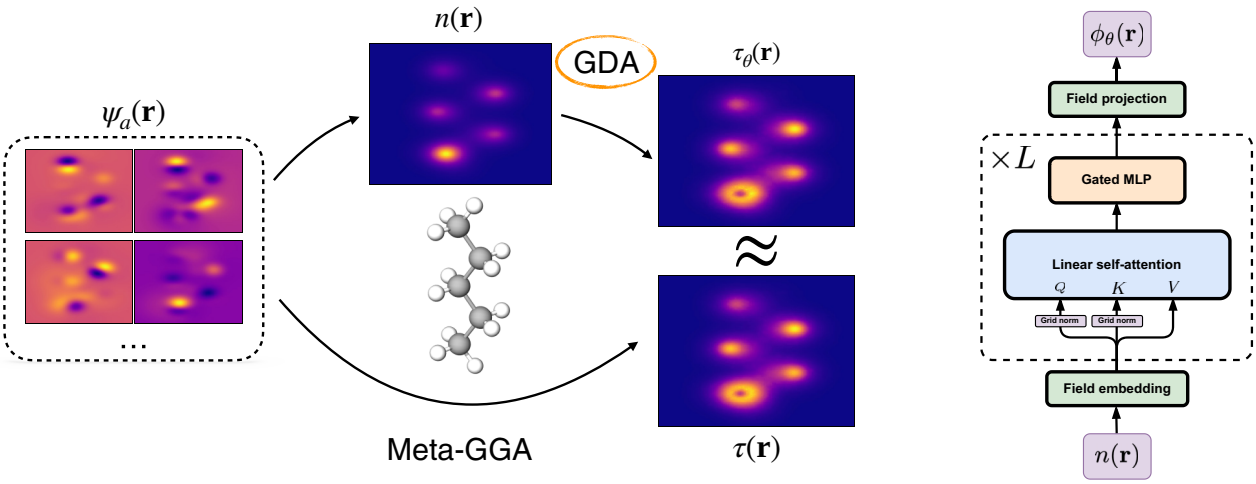

FIG. 1. A schematic representation of the global density approximation scheme. Left: The approximation scheme in which the kinetic energy density τ is directly inferred from the density, eliminating the orbital dependence in all resulting functionals. Right: A diagrammatic representation of the internal connectivity of the GDA model. We use $L = 3$ blocks and $d = 128$ as the dimension of the internal field representation.

output $\tau_\theta(\mathbf{r})$. Therefore, the GDA approximation inherits all scaling properties from parent meta-GGA functionals, such as the uniform scaling of the exchange functional.

Our architecture is similar to an encoder-only transformer [44] with linear attention modules [46,47]. Attention layers were chosen because of the permutation equivariance property—permuting input values on the grid also permutes outputs in the same way, making this architecture well suited for functional learning on grids.

We highlight two main differences in internal component design: a flexible density embedding layer, lifting the density into a high-dimensional representation while respecting underlying symmetries, and specialized linear attention layers to directly operate on DFT grids equipped with custom normalization subcomponents.

Local molecular densities can vary several orders of magnitude, depending on proximity to nuclei. To normalize density variations, we construct the following dimensionless d-component input field,

$$\phi(\mathbf{r}) = \overline{\ln n(\mathbf{r})}\, \boldsymbol{e}_n + \overline{\ln |\nabla n(\mathbf{r})|^2}\, \boldsymbol{e}_\gamma, \tag{5}$$

where $\boldsymbol{e}_n, \boldsymbol{e}_\gamma \in \mathbb{R}^d$ are trainable vectors and $\overline{f(\mathbf{r})}$ indicates subtracting the mean and dividing by the standard deviation of $f(\mathbf{r})$ with respect to the distribution $n(\mathbf{r})/N$. Input ϕ values are updated by subsequent network layers.

To make ϕ_θ invariant with respect to translations and rotations of input coordinates we use a symmetry-aware *positional encoding* sublayer. All computations are performed in a coordinate system where the mean value (dipole moment) of the density distribution $n(\mathbf{r})/N$ vanishes and the covariance matrix is diagonal. This choice eliminates the special Euclidean symmetries in SE(3), determining the resulting computational coordinate values up to molecular point-group symmetries.

The coordinates are then lifted to a d-dimensional representation using random Fourier features (RFFs) [72–74] combined with the density embedding of Eq. (5) using a gating mechanism described in Ref. [75]. The lifted density representation is then processed by a sequence of L blocks. One GDA block is defined as a stack of self-attention (SA) and gated multilayer perceptron (MLP) layers [75] with `SiLU` activations [76]. After L blocks, one final projection layer consisting of an element-wise gated MLP is applied, projecting the pointwise embeddings into a single real number per coordinate $\mathbf{r}$, which we then interpret as the final value of the field $\phi = \phi_\theta[n](\mathbf{r})$ in Eq. (4). We refer the reader to Fig. 1 for an overview of the internal connectivity and the Supplemental Material [50] for numerical and technical details.

The nonlocality in the GDA model is captured by the linear attention layer [46,47]. The input field ϕ is transformed as

$$\phi'(\mathbf{r}) = \int \mathrm{d}^3\mathbf{r}' n(\mathbf{r}')(Q(\mathbf{r}) \cdot K(\mathbf{r}'))V(\mathbf{r}'), \tag{6}$$

where $Q_i(\mathbf{r}), K_i(\mathbf{r}), V_i(\mathbf{r})$ are the *query*, *key*, and *value* parametrized local projections of the input field: $\sum_k W_{ik}^{Q,K,V}\phi_k(\mathbf{r})$. Furthermore, we employ rotary positional encoding (RoPE) [47,48] independently for each block to ensure that the product $Q(\mathbf{r}) \cdot K(\mathbf{r}') = \mathcal{K}(\mathbf{r} - \mathbf{r}')$ depends only on the relative coordinate, parametrizing a flexible continuous convolutional kernel.

Apart from facilitating translational invariance, we emphasize that the linear attention mechanism used in this work sidesteps the unfavorable quadratic scaling of simple functional evaluations. As a consequence, using the transformer architecture does not spoil asymptotic scaling properties of the SCF loop. Further details about the model used in this work are given in the Supplemental Material [50].

The GDA model ϕ_θ is trained using gradient-based optimization of parameters θ using the RAdam optimizer [77–79] and the gradients [80] of the scalar cost function

$$\mathcal{C}(\theta) = \frac{\|\phi_\theta - \phi_0\|_G^2}{\|\phi_0\|_G^2} + \lambda \frac{\|\mathcal{T} - \mathcal{T}_0\|_F^2}{\|\mathcal{T}_0\|_F^2} \tag{7}$$

consisting of two terms. The first term is the scaled mean-squared error (MSE) for ϕ itself where the unweighted grid

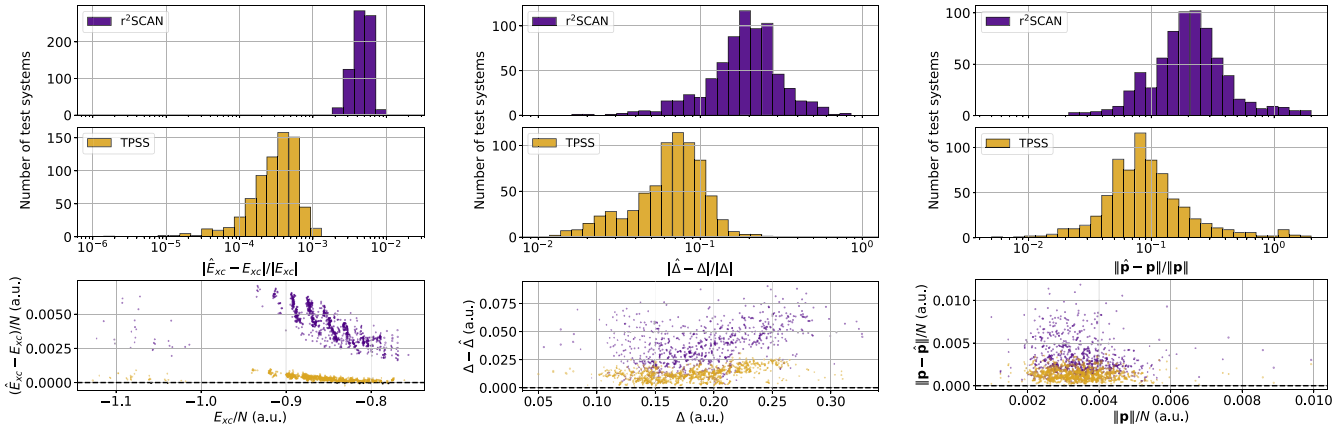

FIG. 2. Error distributions of observables obtained from first-principles calculations using the GDA approximation. The surrogate functional was tested on 717 test molecular systems withheld during training, a random selection of 10% of the QM7 [84,85] dataset. Left: XC energy values. Middle: HOMO-LUMO gap values. Right: Total dipole moments.

norm $\|f\|_G^2 = \sum_i f(\mathbf{r}_i)^2$ is defined over all DFT grid points $\{\mathbf{r}_1, \ldots, \mathbf{r}_{N_g}\}$, where $\phi_0(\mathbf{r})$ is evaluated from precalculated τ values in the dataset, by inverting Eq. (4). The second term in Eq. (7) enforces the correct predicted kinetic energy matrix $\mathcal{T}$ in the KS orbital basis, which can be easily predicted from the one-body electronic density matrix Γ as

$$\mathcal{T}_{ab} = \sum_{\mu\nu} C_{\mu a} C_{\nu b} \frac{\partial T[n]}{\partial \Gamma_{\mu\nu}} \tag{8}$$

by using standard AD routines and the linear combination of atomic orbitals (LCAO) expansion coefficients $C_{\mu a}$ (see Supplemental Material [50]). Reference values $\mathcal{T}_0$ can be precalculated as $\mathcal{T}_{0,ab} = \frac{1}{2}\int \mathrm{d}^3\mathbf{r} \nabla\psi_a \cdot \nabla\psi_b$ from easily accessible basis set integrals [81,82].

Including the second term in the cost function in Eq. (7) ensures that the kinetic contribution to the overall KS effective Hamiltonian $H_{\text{eff}} = \frac{\partial E}{\partial \Gamma}$ is well approximated by the GDA surrogate functional when solving the KS equation $C^\top H_{\text{eff}} C = \epsilon$. We find that including such a gradient cost term is key for making the resulting neural network functional converge in a practical DFT SCF calculation, allowing us to use it to obtain energies, orbitals, densities, or observables from first principles, without ever referring to the parent functional. Similar regularization methods have been proposed in Refs. [38,83].

III. RESULTS

We examine GDA approximations to two prominent meta-GGA functionals: r^{2}SCAN [28,60,62] and TPSS [55]. We use a single neural network GDA model for evaluation of all molecules for all three functionals. Each functional is tested in both ground state KS-DFT and linear response time-dependent density functional theory (TD-DFT) calculations. The architecture outlined in Fig. 1 with $L = 3$ blocks and the embedding dimension of $d = 128$ is used and optimized using $\lambda = 1$ in Eq. (7). Molecular geometries from the QM7 [84,85] dataset are used, consisting of 7165 organic molecules of up to 23 atoms, and featuring a large variety of molecular structures such as double and triple bonds, cycles, carboxy, cyanide, amide, alcohol, and epoxy. Corresponding KS-DFT calculations were done using the r^{2}SCAN functional to obtain input densities and target τ values.

We independently calculate and compare several physical observables: the total energy, the molecular dipole moment, the KS highest occupied molecular orbital–lowest unoccupied molecular crbital (HOMO-LUMO) gap, and the self-consistent density itself. Error distributions are shown in Fig. 2, demonstrating that the resulting surrogate GGA functionals converge to physical results independently from the source meta-GGA functional, using only density inputs as shown in Eq. (3).

The GDA functional is able to accurately predict XC energies and potentials over a large range of diverse test molecules, eliminating orbital dependence from input functionals. Gap values are predicted within 10% for TPSS with higher errors in r^{2}SCAN calculations where GDA approximations tend to overestimate the gap. A similar trend persists with dipole moments, demonstrating qualitative accuracy. Since all of the results have been obtained with a single transferable neural network model, the GDA approximation scheme eliminates the need to train multiple models for deorbitalizing different XC functionals and can serve as a starting point for more fine-tuned approximations. Furthermore, we speculate that such broad generalization for larger models trained on more diverse datasets can be used to train foundation functionals to then fine tune on downstream problems with limited data.

As a direct global divergence measure between two densities, we consider the Kullback-Liebler (KL) divergence [86] $D_{\text{KL}}(n \parallel \hat{n}) = \frac{1}{N}\int \mathrm{d}^3\mathbf{r}\, n(\mathbf{r}) \ln \frac{n(\mathbf{r})}{\hat{n}(\mathbf{r})}$ commonly used in statistical literature, where $\hat{n}$ is the density obtained by independent SCF convergence using the GDA approximation of the original functional. In addition, we also compare direct mean absolute errors (MAEs) of density values defined as $D_{\text{MAE}}(n, \hat{n}) = \frac{1}{N}\int \mathrm{d}^3\mathbf{r}\, |n(\mathbf{r}) - \hat{n}(\mathbf{r})|$. In all cases the trained model produces a good approximation to the ground state density indicating good transferability, as can be seen in the left panel of Fig. 3. We see that TPSS densities are reproduced more accurately indicating that the generated dataset based on QM7 is better suited to some typical densities encountered with that functional. A more expansive dataset with more

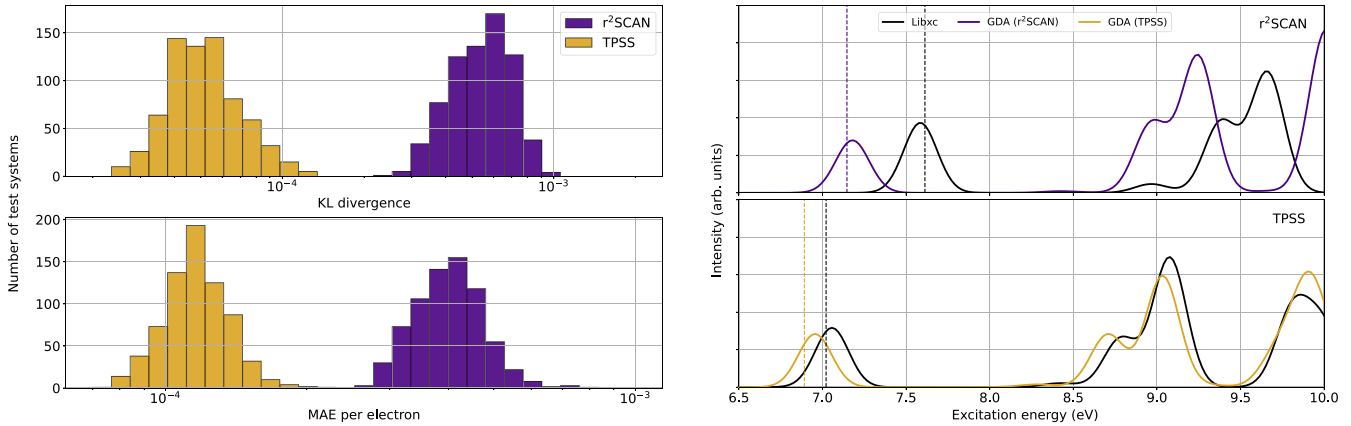

FIG. 3. Comparison of first-principles densities calculated using independent KS-DFT calculations and an example linear-response calculation of dimethyl ether (CH_3OCH_3) spectra using TD-DFT where the predicted KS gaps are indicated by vertical lines in the same style as the corresponding spectra. We note that the spectral data are shown as a proof of principle, as an example of a calculation that can be performed with more specialized GDA functionals using automatic differentiation.

diverse densities and geometries is likely to close the gap in Fig. 3.

To showcase the performance of GDA approximations in the excited-state regime, we present a qualitative demonstration in the right panel of Fig. 3. Here, we simulate absorption spectra via linear response for a single test molecule. We compare spectra generated using the GDA model with the parent functionals accessed through the LIBXC [87,88] library. Calculations are carried out in the Casida TD-DFT [89] formalism for 100 excited states, and we focus on a low-energy range to highlight the ability of our model to replicate physically meaningful features before the ionization threshold [90,91] in a fashion similar to the original functional. Unsurprisingly, the GDA scheme captures the behavior of TPSS. However, for most test molecules in our dataset, the excited-state energy values are qualitative. For example, in the case of r^{2}SCAN, we observe spectral shifts (on the order of $\sim$1 eV in Fig. 3) but still capture the overall shape and widths of the affected features. The GDA model has a higher error in r^{2}SCAN with the expectation that the predictions can be improved by using larger GDA neural networks trained on more specialized datasets. Therefore, the spectral data is shown as a proof of principle, as an example of a linear response calculation that can be performed with fine-tuned GDA functionals using automatic differentiation.

IV. CONCLUSION AND OUTLOOK

We introduce another class of approximations, enabling first-principles replacement of orbital dependent meta-GGA functionals. Using these approximations, arbitrary derivatives of source functionals can be constructed, enabling access to different results and phenomena (e.g., highly nonlinear responses).

As a proof of principle, this approximation is accurate and resource efficient with a high degree of transferability between different functionals and molecular systems. Tests on periodic systems are left for future research. In addition, our GDA approximation scheme formally enables the use of orbital-free DFT (OF-DFT) calculations at the meta-GGA level of theory. However, robust OF-DFT solvers for Gaussian-type basis sets are still an active area of research with experimental support for nonlocal functionals at best. Therefore, extending the GDA approximation to OF-DFT is left for future research.

More general orbital dependent functionals have shown great potential in overcoming the limitations of their pure density functionals. Hybrid functionals, such as Becke three-parameter Lee-Yang-Parr (B3LYP), Heyd-Scuseria-Ernzerhof (HSE), and Perdew-Burke-Ernzerhof (PBE0) [25,92,93], which combine GGA functionals with a fraction of orbital dependent exact exchange, have achieved superior accuracy in predicting molecular geometries, reaction barriers, and electronic properties. The GDA treatment of the Fock operator is a topic of ongoing research. The GDA approach allows reformulating orbital-dependent functionals as pure but nonlocal density functionals, offering a promising direction to bridge the gap between *pure* DFT and the quantum chemical accuracy.

ACKNOWLEDGMENTS

M.M. acknowledges support from the CCQ Graduate Fellowship in computational quantum physics under the Grant No. 653217. J.C.U. acknowledges support from the CCQ Graduate Fellowship in computational quantum physics under the Grant No. 1165064.

DATA AVAILABILITY

Code needed to reproduce results in this work or experiment with new results has been open-sourced and can be found at GitHub [94].

All DFT simulations were performed using a custom interface between the PYSCF [81,82] library and PYTORCH [65,80], used for automatic generation of XC potentials for KS-DFT calculations and kernels for linear-response TD-DFT calculations. All calculations were performed using the correlation-consistent polarized valence double zeta (cc-pVDZ) basis set at grid level 1 in PySCF. Convergence tolerance was set to 10^{-6} Ha.

[1] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, Phys. Rev. **136**, B864 (1964).

[2] W. Kohn and L. J. Sham, Self-consistent equations including exchange and correlation effects, Phys. Rev. **140**, A1133 (1965).

[3] *Time-Dependent Density Functional Theory*, edited by M. A. L. Marques, C. A. Ullrich, F. Nogueira, A. Rubio, K. Burke, and E. K. U. Gross, Lecture Notes in Physics Vol. 706 (Springer, Berlin, 2006).

[4] *A Primer in Density Functional Theory*, edited by C. Fiolhais, F. Nogueira, M. A. L. Marques, R. Beig, B. G. Englert, U. Frisch, P. Hänggi, K. Hepp, W. Hillebrandt, D. Imboden, R. L. Jaffe, R. Lipowsky, H. V. Löhneysen, I. Ojima, D. Sornette, S. Theisen, W. Weise, J. Wess, and J. Zittartz, Lecture Notes in Physics Vol. 620 (Springer, Berlin, 2003).

[5] G. Onida, L. Reining, and A. Rubio, Electronic excitations: density-functional versus many-body Green's-function approaches, Rev. Mod. Phys. **74**, 601 (2002).

[6] J. A. Pople, M. Head-Gordon, and K. Raghavachari, Quadratic configuration interaction. A general technique for determining electron correlation energies, J. Chem. Phys. **87**, 5968 (1987).

[7] R. J. Bartlett and M. Musiał, Coupled-cluster theory in quantum chemistry, Rev. Mod. Phys. **79**, 291 (2007).

[8] B. Jeziorski and H. J. Monkhorst, Coupled-cluster method for multideterminantal reference states, Phys. Rev. A **24**, 1668 (1981).

[9] F. Becca and S. Sorella, *Quantum Monte Carlo Approaches for Correlated Systems* (Cambridge University Press, Cambridge, UK, 2017).

[10] D. Wu, R. Rossi, F. Vicentini, N. Astrakhantsev, F. Becca, X. Cao, J. Carrasquilla, F. Ferrari, A. Georges, M. Hibat-Allah, M. Imada, A. M. Läuchli, G. Mazzola, A. Mezzacapo, A. Millis, J. Robledo Moreno, T. Neupert, Y. Nomura, J. Nys, O. Parcollet *et al.*, Variational benchmarks for quantum many-body problems, Science **386**, 296 (2024).

[11] M. Medvidović and J. Robledo Moreno, Neural-network quantum states for many-body physics, Eur. Phys. J. Plus **139**, 631 (2024).

[12] J. Lee, H. Q. Pham, and D. R. Reichman, Twenty years of auxiliary-field quantum Monte Carlo in quantum chemistry: An overview and assessment on main group chemistry and bond-breaking, J. Chem. Theory Comput. **18**, 7024 (2022).

[13] A. Georges and G. Kotliar, Hubbard model in infinite dimensions, Phys. Rev. B **45**, 6479 (1992).

[14] A. Georges, G. Kotliar, W. Krauth, and M. J. Rozenberg, Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions, Rev. Mod. Phys. **68**, 13 (1996).

[15] L. Hedin, New method for calculating the one-particle Green's function with application to the electron-gas problem, Phys. Rev. **139**, A796 (1965).

[16] F. Aryasetiawan and O. Gunnarsson, The *GW* method, Rep. Prog. Phys. **61**, 237 (1998).

[17] L. Reining, The *GW* approximation: Content, successes and limitations, WIREs Comput. Mol. Sci. **8**, e1344 (2018).

[18] D. Golze, M. Dvorak, and P. Rinke, The *GW* compendium: A practical guide to theoretical photoemission spectroscopy, Front. Chem. **7**, 377 (2019).

[19] K. Burke, Perspective on density functional theory, J. Chem. Phys. **136**, 150901 (2012).

[20] J. P. Perdew and K. Schmidt, Jacob's ladder of density functional approximations for the exchange-correlation energy, AIP Conf. Proc. **577**, 1 (2001).

[21] A. Seidl, A. Görling, P. Vogl, J. A. Majewski, and M. Levy, Generalized Kohn-Sham schemes and the band-gap problem, Phys. Rev. B **53**, 3764 (1996).

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

[23] A. D. Becke, Density–functional thermochemistry. III. The role of exact exchange, J. Chem. Phys. **98**, 5648 (1993).

[24] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. **77**, 3865 (1996).

[25] J. Heyd, G. E. Scuseria, and M. Ernzerhof, Hybrid functionals based on a screened Coulomb potential, J. Chem. Phys. **118**, 8207 (2003).

[26] S. Grimme, Semiempirical GGA-type density functional constructed with a long-range dispersion correction, J. Comput. Chem. **27**, 1787 (2006).

[27] A. Tkatchenko and M. Scheffler, Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data, Phys. Rev. Lett. **102**, 073005 (2009).

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

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

[30] M. Bogojeski, L. Vogt-Maranto, M. E. Tuckerman, K.-R. Müller, and K. Burke, Quantum chemical accuracy from density functional approximations via machine learning, Nat. Commun. **11**, 5223 (2020).

[31] 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).

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

[33] 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).

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

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

[36] P. A. M. Casares, J. S. Baker, M. Medvidović, R. dos Reis, and J. M. Arrazola, GradDFT. A software library for machine learning enhanced density functional theory, J. Chem. Phys. **160**, 062501 (2024).

[37] B. G. Del Rio, B. Phan, and R. Ramprasad, A deep learning framework to emulate density functional theory, npj Comput. Mater. **9**, 158 (2023).

[38] J. Kirkpatrick, B. McMorrow, D. H. P. Turban, A. L. Gaunt, J. S. Spencer, A. G. D. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau, L. R. Castellanos, S. Petersen, A. W. R. Nelson, P. Kohli, P. Mori-Sánchez, D. Hassabis, and A. J. Cohen, Pushing the frontiers of density functionals by

solving the fractional electron problem, Science **374**, 1385 (2021).
[39] J. T. Margraf and K. Reuter, Pure non-local machine-learned density functional theory for electron correlation, Nat. Commun. **12**, 344 (2021).
[40] K. Bystrom and B. Kozinsky, Nonlocal machine-learned exchange functional for molecules and solids, Phys. Rev. B **110**, 075130 (2024).
[41] K. Bystrom, S. Falletta, and B. Kozinsky, Training machine-learned density functionals on band gaps, J. Chem. Theory Comput. **20**, 7516 (2024).
[42] C. C. Margossian, A review of automatic differentiation and its efficient implementation, WIREs Data Mining Knowl. Discov. **9**, e1305 (2019).
[43] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind, Automatic differentiation in machine learning: A survey, J. Mach. Learn. Res. **18**, 1 (2018).
[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, arXiv:1706.03762.
[45] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, An image is worth 16×16 words: Transformers for image recognition at scale, arXiv:2010.11929.
[46] S. Cao, Choose a transformer: Fourier or Galerkin, arXiv:2105.14995.
[47] Z. Li, K. Meidani, and A. B. Farimani, Transformer for partial differential equations' operator learning, arXiv:2205.13671.
[48] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, RoFormer: Enhanced transformer with rotary position embedding, arXiv:2104.09864.
[49] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods* (Cambridge University Press, Cambridge, UK, 2020).
[50] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.7.023113 for more details on the neural network architecture and training, extended performance benchmarks and DFT conventions.
[51] R. O. Jones, Density functional theory: Its origins, rise to prominence, and future, Rev. Mod. Phys. **87**, 897 (2015).
[52] S. Kümmel and L. Kronik, Orbital-dependent density functionals: Theory and applications, Rev. Mod. Phys. **80**, 3 (2008).
[53] J. P. Perdew and A. Zunger, Self-interaction correction to density-functional approximations for many-electron systems, Phys. Rev. B **23**, 5048 (1981).
[54] D. C. Langreth and M. J. Mehl, Beyond the local-density approximation in calculations of ground-state electronic properties, Phys. Rev. B **28**, 1809 (1983).
[55] J. Tao, J. P. Perdew, V. N. Staroverov, and G. E. Scuseria, Climbing the density functional ladder: Nonempirical meta–generalized gradient approximation designed for molecules and solids, Phys. Rev. Lett. **91**, 146401 (2003).
[56] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, L. A. Constantin, and J. Sun, Workhorse semilocal density functional for condensed matter physics and quantum chemistry, Phys. Rev. Lett. **103**, 026403 (2009).
[57] Y. Zhao and D. G. Truhlar, A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions, J. Chem. Phys. **125**, 194101 (2006).
[58] J. Sun, B. Xiao, and A. Ruzsinszky, Communication: Effect of the orbital-overlap dependence in the meta generalized gradient approximation, J. Chem. Phys. **137**, 051101 (2012).
[59] J. Sun, R. Haunschild, B. Xiao, I. W. Bulik, G. E. Scuseria, and J. P. Perdew, Semilocal and hybrid meta-generalized gradient approximations based on the understanding of the kinetic-energy-density dependence, J. Chem. Phys. **138**, 044113 (2013).
[60] J. Sun, R. C. Remsing, Y. Zhang, Z. Sun, A. Ruzsinszky, H. Peng, Z. Yang, A. Paul, U. Waghmare, X. Wu, and others, Accurate first-principles structures and energies of diversely bonded systems from an efficient density functional, Nat. Chem. **8**, 831 (2016).
[61] A. P. Bartók and J. R. Yates, Regularized SCAN functional, J. Chem. Phys. **150**, 161101 (2019).
[62] J. W. Furness, A. D. Kaplan, J. Ning, J. P. Perdew, and J. Sun, Accurate and numerically efficient r^2SCAN meta-generalized gradient approximation, J. Phys. Chem. Lett. **11**, 8208 (2020).
[63] A. D. Becke and E. R. Johnson, A simple effective potential for exchange, J. Chem. Phys. **124**, 221101 (2006).
[64] F. Tran and P. Blaha, Accurate band gaps of semiconductors and insulators with a semilocal exchange-correlation potential, Phys. Rev. Lett. **102**, 226401 (2009).
[65] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, PyTorch: An Imperative Style, High-Performance Deep Learning Library, Advances in Neural Information Processing Systems 32 (2019).
[66] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning. Nature (London) **521**, 436 (2015).
[67] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).
[68] A. Dawid, J. Arnold, B. Requena, A. Gresch, M. Płodzień, K. Donatella, K. A. Nicoli, P. Stornati, R. Koch, M. Büttner, R. Okuła, G. Muñoz-Gil, R. A. Vargas-Hernández, A. Cervera-Lierta, J. Carrasquilla, V. Dunjko, M. Gabrié, P. Huembeli, E. van Nieuwenburg, F. Vicentini *et al.*, Modern applications of machine learning in quantum sciences, arXiv:2204.04198.
[69] C. F. V. Weizsäcker, Zur Theorie der Kernmassen, Z. Phys. **96**, 431 (1935).
[70] S. Subramanian, P. Harrington, K. Keutzer, W. Bhimji, D. Morozov, M. Mahoney, and A. Gholami, Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior, arXiv:2306.00258.
[71] M. McCabe, B. Régaldo-Saint Blancard, L. H. Parker, R. Ohana, M. Cranmer, A. Bietti, M. Eickenberg, S. Golkar, G. Krawezik, F. Lanusse, M. Pettee, T. Tesileanu, K. Cho, and S. Ho, Multiple physics pretraining for physical surrogate models, arXiv:2310.02994.
[72] A. Rahimi and B. Recht, Random features for large-scale kernel machines, in *Advances in Neural Information Processing Systems*, edited by J. Platt, D. Koller, Y. Singer, and S. Roweis (Curran Associates, Red Hook, NY, 2007), Vol. 20.
[73] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng, Fourier features let networks learn high frequency functions in low dimensional domains, arXiv:2006.10739.

[74] Y. Li, S. Si, G. Li, C.-J. Hsieh, and S. Bengio, Learnable Fourier features for multi-dimensional spatial positional encoding, arXiv:2106.02795.

[75] N. Shazeer, GLU variants improve transformer, arXiv:2002.05202.

[76] S. Elfwing, E. Uchibe, and K. Doya, Sigmoid-weighted linear units for neural network function approximation in reinforcement learning, arXiv:1702.03118.

[77] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, On the variance of the adaptive learning rate and beyond, arXiv:1908.03265.

[78] D. P. Kingma and J. Lei Ba, Adam: A method for stochastic optimization, in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (International Conference on Learning Representations, ICLR, 2015).

[79] I. Loshchilov and F. Hutter, Fixing weight decay regularization in Adam, arXiv:1711.05101.

[80] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai *et al.*, PyTorch: An imperative style, high-performance deep learning library, in *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (ACM, New York, 2019), p. 721.

[81] 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, S. Wouters, and G. K.-L. Chan, PySCF: the Python-based simulations of chemistry framework, WIREs Comput. Mol. Sci. **8**, e1340 (2018).

[82] Q. Sun, X. Zhang, S. Banerjee, P. Bao, M. Barbry, N. S. Blunt, N. A. Bogdanov, G. H. Booth, J. Chen, Z.-H. Cui, J. J. Eriksen, Y. Gao, S. Guo, J. Hermann, M. R. Hermes, K. Koh, P. Koval, S. Lehtola, Z. Li, J. Liu *et al.*, Recent developments in the PySCF program package, J. Chem. Phys. **153**, 024109 (2020).

[83] P. del Mazo-Sevillano and J. Hermann, Variational principle to regularize machine-learned density functionals: The non-interacting kinetic-energy functional, J. Chem. Phys. **159**, 194107 (2023).

[84] L. C. Blum and J.-L. Reymond, 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13, J. Am. Chem. Soc. **131**, 8732 (2009).

[85] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. Anatole Von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning, Phys. Rev. Lett. **108**, 058301 (2012).

[86] S. Kullback and R. A. Leibler, On information and sufficiency, Ann. Math. Stat. **22**, 79 (1951).

[87] M. A. L. Marques, M. J. T. Oliveira, and T. Burnus, Libxc: A library of exchange and correlation functionals for density functional theory, Comput. Phys. Commun. **183**, 2272 (2012).

[88] S. Lehtola, C. Steigemann, M. J. T. Oliveira, and M. A. L. Marques, Recent developments in Libxc—a comprehensive library of functionals for density functional theory, SoftwareX **7**, 1 (2018).

[89] M. E. Casida, Time-dependent density functional response theory for molecules, in *Recent Advances in Density Functional Methods: (Part I)* (World Scientific, Singapore, 1995), pp. 155–192.

[90] W.-C. Tam and C. E. Brion, Electron impact spectra of some alkyl derivatives of water and related compounds, J. Electron Spectrosc. Relat. Phenom. **3**, 263 (1974).

[91] P. J. Linstrom, W. G. Mallard *et al.*, NIST standard reference database number 69, NIST Chemistry WebBook 20899 (2018).

[92] 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).

[93] C. Adamo and V. Barone, Toward reliable density functional methods without adjustable parameters: The PBE0 model, J. Chem. Phys. **110**, 6158 (1999).

[94] https://github.com/Matematija/global-density-approximation