Neural network distillation of orbital dependent density functional theory

Matija Medvidović
ETH Zürich
CH-8093 Zürich, Switzerland
mmedvidovic@ethz.ch

Jaylyn C. Umana
Center for Computational Quantum Physics
Flatiron Institute, New York 10010, USA
The Graduate Center, City University of New York
New York 10016, USA

Iman Ahmadabadi
Joint Quantum Institute, NIST and University of Maryland
College Park, Maryland 20742, USA
Center for Computational Quantum Physics
Flatiron Institute, New York 10010, USA
Princeton University
Princeton, New Jersey 08544, USA

Domenico Di Sante
University of Bologna
40127 Bologna, Italy

Johannes Flick
City College of New York
New York 10031, USA
The Graduate Center,
City University of New York
New York 10016, USA
Center for Computational Quantum Physics
Flatiron Institute, New York 10010, USA

Angel Rubio
Max Planck Institute for the
Structure and Dynamics of Matter
22761 Hamburg, Germany
Center for Computational Quantum Physics
Flatiron Institute, New York 10010, USA
Initiative for Computational Catalysis
Flatiron Institute, New York 10010, USA

Abstract

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.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science.

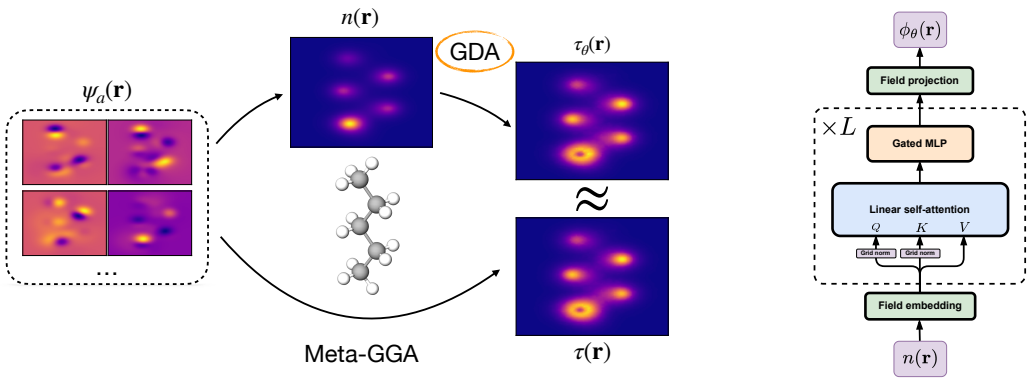

Figure 1: A schematic representation of the global density approximation scheme. **Left**: The approximation scheme in which the kinetic energy density τ is directly inferred from the density, eliminating the orbital dependence in all resulting functionals. **Right**: A diagrammatic representation of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

1 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 tradeoffs between efficiency, accuracy, and applicability. Density functional theory (DFT) [1–5] should be contrasted with the accuracy of wavefunction 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 [20, 2, 21–24, 20, 25–28]. 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 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.

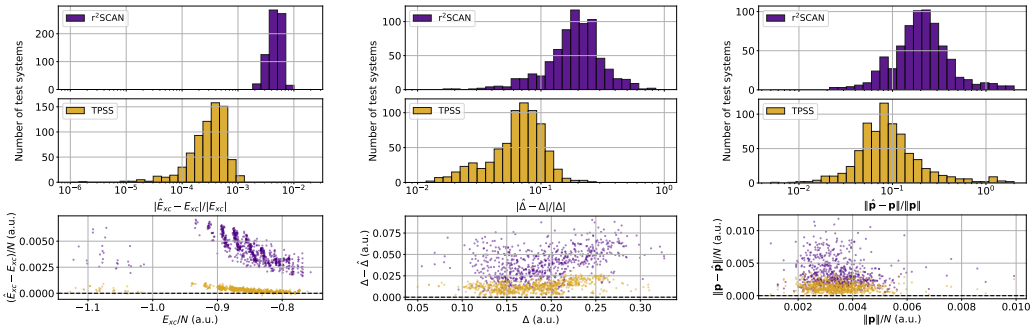

Figure 2: Error distributions of observables obtained from first-principles calculations using the GDA approximation. The surrogate functional was tested on 717 test molecular systems withheld during training, a random selection of 10% of the QM7 [54, 55] dataset. **Left**: XC energy values. **Middle**: HOMO-LUMO gap values. **Right**: Total dipole moments.

2 Methods

2.1 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 [49, 4, 3] 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 non-interacting 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

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

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

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. 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 $n(\mathbf{r})$ [52]. 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) [53, 24], 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 [56–60, 28, 61–65], which capture more complex electronic interactions and offer improved accuracy for diverse systems, especially in terms of

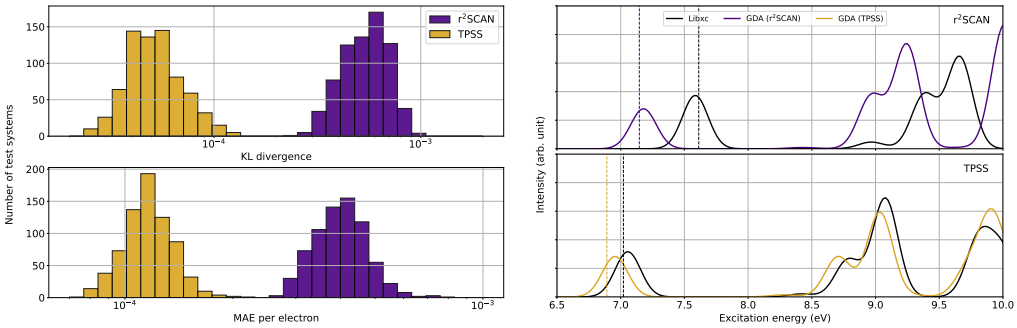

Figure 3: Comparison of first-principles densities calculated using independent KS-DFT calculations and an example linear-response calculation of dimethyl ether (CH_3OCH_3) spectra using TD-DFT where the predicted KS gaps are indicated by vertical lines in the same style as the corresponding spectra. 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.

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 parameterized 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, 66].

We rewrite the kinetic energy density as $\tau \approx \tau_\theta[n]$, a nonlocal functional captured by an expressive deep neural network [67–69] with θ indicating a set of all free parameters in neural network sub-components (see Supplemental Material 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.

2.2 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

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

is represented by a neural network with parameters $\theta \in \mathbb{R}^P$, based on a dimensionless kinetic energy indicator variable used in Ref. [63]. 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 [70] kinetic functional. The regularization parameter was set to $\eta = 10^{-3}$, as defined in Ref. [63] and informed by normalization heuristics [71, 72] 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 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 [47, 46]. 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 sub-components.

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:

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

where $\mathbf{e}_n\,, \mathbf{e}_\gamma \in \mathbb{R}^d$ 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* sub-layer. 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) [73–75] combined with the density embedding of Eq. 5 using a gating mechanism described in Ref. [76]. 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 multi-layer perceptron (MLP) layers [76] with `SiLU` activations [77]. After L blocks, one final projection layer consisting of an element-wise gated MLP is applied, projecting the point-wise 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 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: $\sum_k W^{Q,K,V}_{ik}\, \phi_k(\mathbf{r})$. Furthermore, we employ rotary positional encoding (RoPE) [48, 47] 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

The GDA model ϕ_θ is trained using gradient-based optimization of parameters θ using the RAdam optimizer [78–80] and the gradients [81] 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 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 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 [82, 83].

Including the second term in the cost function in Eq. 7 ensures that the kinetic contribution to the overall KS effective Hamiltonian $H_{\text{eff}} = {}^{\partial E}/_{\partial \Gamma}$ 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. [84, 38].

3 Results

We examine GDA approximations to two prominent meta-GGA functionals: r^{2}SCAN [63, 28, 61] and TPSS [56]. 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 [54, 55] 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 orbital (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 over estimate 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 de-orbitalizing 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 [85] $D_{\text{KL}}(n \parallel \hat{n}) = \frac{1}{N}\int \mathrm{d}^3\mathbf{r}\; n(\mathbf{r})\, \ln \frac{n(\mathbf{r})}{\hat{n}(\mathbf{r})}$ 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 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 [86, 87] library. Calculations are carried out in the Casida TD-DFT [88] 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 [89, 90] 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 ~ 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.

4 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 0 (PBE0) [91, 25, 92], 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.

Software and simulations

All DFT simulations were performed using a custom interface between the PySCF [82, 83] library and PyTorch [66, 81], 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. Code needed to reproduce results in this work or experiment with new results has been open-sourced and can be found at GitHub: `https://github.com/Matematija/global-density-approximation`.

Acknowledgments and Disclosure of Funding

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.

References

[1] P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Physical Review*, 136(3B):B864–B871, November 1964. ISSN 0031-899X. doi: 10.1103/PhysRev.136.B864. URL https://link.aps.org/doi/10.1103/PhysRev.136.B864.

[2] W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review*, 140(4A):A1133–A1138, November 1965. ISSN 0031-899X. doi: 10.1103/PhysRev.140.A1133. URL https://link.aps.org/doi/10.1103/PhysRev.140.A1133.

[3] Miguel A.L. Marques, Carsten A. Ullrich, Fernando Nogueira, Angel Rubio, Kieron Burke, and Eberhard K. U. Gross, editors. *Time-Dependent Density Functional Theory*, volume 706. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-35422-2. doi: 10.1007/b11767107. URL http://link.springer.com/10.1007/b11767107. Series Title: Lecture Notes in Physics.

[4] Carlos Fiolhais, Fernando Nogueira, Miguel 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, editors. *A Primer in Density Functional Theory*, volume 620 of *Lecture Notes in Physics*. Springer, Berlin, Heidelberg, 2003. ISBN 978-3-540-03083-6 978-3-540-37072-7. doi: 10.1007/3-540-37072-2. URL http://link.springer.com/10.1007/3-540-37072-2.

[5] Giovanni Onida, Lucia Reining, and Angel Rubio. Electronic excitations: density-functional versus many-body Green's-function approaches. *Reviews of Modern Physics*, 74(2):601–659, June 2002. doi: 10.1103/RevModPhys.74.601. URL https://link.aps.org/doi/10.1103/RevModPhys.74.601.

[6] John A. Pople, Martin Head-Gordon, and Krishnan Raghavachari. Quadratic configuration interaction. A general technique for determining electron correlation energies. *The Journal of Chemical Physics*, 87(10):5968–5975, November 1987. ISSN 0021-9606. doi: 10.1063/1.453520. URL https://doi.org/10.1063/1.453520.

[7] Rodney J. Bartlett and Monika Musiał. Coupled-cluster theory in quantum chemistry. *Reviews of Modern Physics*, 79(1):291–352, February 2007. ISSN 0034-6861, 1539-0756. doi: 10.1103/RevModPhys.79.291. URL https://link.aps.org/doi/10.1103/RevModPhys.79.291.

[8] Bogumil Jeziorski and Hendrik J. Monkhorst. Coupled-cluster method for multideterminantal reference states. *Physical Review A*, 24(4):1668–1681, October 1981. URL https://link.aps.org/doi/10.1103/PhysRevA.24.1668.

[9] Federico Becca and Sandro Sorella. *Quantum Monte Carlo Approaches for Correlated Systems*. Cambridge University Press, Cambridge, 2017. ISBN 978-1-107-12993-1. doi: 10.1017/9781316417041. URL https://www.cambridge.org/core/books/quantum-monte-carlo-approaches-for-correlated-systems/EB88C86BD9553A0738BDAE400D0B2900.

[10] Dian Wu, Riccardo Rossi, Filippo Vicentini, Nikita Astrakhantsev, Federico Becca, Xiaodong Cao, Juan Carrasquilla, Francesco Ferrari, Antoine Georges, Mohamed Hibat-Allah, Masatoshi Imada, Andreas M. Läuchli, Guglielmo Mazzola, Antonio Mezzacapo, Andrew Millis, Javier Robledo Moreno, Titus Neupert, Yusuke Nomura, Jannes Nys, Olivier Parcollet, Rico Pohle, Imelda Romero, Michael Schmid, J. Maxwell Silvester, Sandro Sorella, Luca F. Tocchio, Lei Wang, Steven R. White, Alexander Wietek, Qi Yang, Yiqi Yang, Shiwei Zhang, and Giuseppe Carleo. Variational benchmarks for quantum many-body problems. *Science*, 386 (6719):296–301, October 2024. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adg9774. URL https://www.science.org/doi/10.1126/science.adg9774.

[11] Matija Medvidović and Javier Robledo Moreno. Neural-network quantum states for many-body physics. *The European Physical Journal Plus*, 139(7):631, July 2024. ISSN 2190-5444. doi: 10.1140/epjp/s13360-024-05311-y. URL http://arxiv.org/abs/2402.11014.

[12] Joonho Lee, Hung Q. Pham, and David R. Reichman. Twenty Years of Auxiliary-Field Quantum Monte Carlo in Quantum Chemistry: An Overview and Assessment on Main Group Chemistry and Bond-Breaking. *Journal of Chemical Theory and Computation*, 18(12):7024–7042, December 2022. ISSN 1549-9618, 1549-9626. doi: 10.1021/acs.jctc.2c00802. URL https://pubs.acs.org/doi/10.1021/acs.jctc.2c00802.

[13] Antoine Georges and Gabriel Kotliar. Hubbard model in infinite dimensions. *Physical Review B*, 45(12):6479–6483, March 1992. doi: 10.1103/PhysRevB.45.6479. URL https://link.aps.org/doi/10.1103/PhysRevB.45.6479.

[14] Antoine Georges, Gabriel Kotliar, Werner Krauth, and Marcelo J. Rozenberg. Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions. *Reviews of Modern Physics*, 68(1):13–125, January 1996. ISSN 0034-6861, 1539-0756. doi: 10.1103/RevModPhys.68.13. URL https://link.aps.org/doi/10.1103/RevModPhys.68.13.

[15] Lars Hedin. New Method for Calculating the One-Particle Green's Function with Application to the Electron-Gas Problem. *Physical Review*, 139(3A):A796–A823, August 1965. doi: 10.1103/PhysRev.139.A796. URL https://link.aps.org/doi/10.1103/PhysRev.139.A796.

[16] F. Aryasetiawan and O. Gunnarsson. The GW method. *Reports on Progress in Physics*, 61 (3):237, March 1998. ISSN 0034-4885. doi: 10.1088/0034-4885/61/3/002. URL https://dx.doi.org/10.1088/0034-4885/61/3/002.

[17] Lucia Reining. The GW approximation: content, successes and limitations. *WIREs Computational Molecular Science*, 8(3):e1344, 2018. ISSN 1759-0884. doi: 10.1002/wcms.1344. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1344.

[18] Dorothea Golze, Marc Dvorak, and Patrick Rinke. The GW Compendium: A Practical Guide to Theoretical Photoemission Spectroscopy. *Frontiers in Chemistry*, 7:377, 2019. ISSN 2296-2646. doi: 10.3389/fchem.2019.00377.

[19] Kieron Burke. Perspective on density functional theory. *The Journal of Chemical Physics*, 136(15), April 2012. ISSN 0021-9606, 1089-7690. doi: 10.1063/1.4704546. URL https://pubs.aip.org/jcp/article/136/15/150901/941589/Perspective-on-density-functional-theory.

[20] John P. Perdew and Karla Schmidt. Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conference Proceedings*, 577(1):1–20, July 2001. ISSN 0094-243X. doi: 10.1063/1.1390175. URL https://doi.org/10.1063/1.1390175.

[21] A. Seidl, A. Görling, P. Vogl, J. A. Majewski, and M. Levy. Generalized Kohn-Sham schemes and the band-gap problem. *Physical Review B*, 53(7):3764–3774, February 1996. doi: 10.1103/PhysRevB.53.3764. URL https://link.aps.org/doi/10.1103/PhysRevB.53.3764.

[22] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2): 785–789, January 1988. doi: 10.1103/PhysRevB.37.785. URL https://link.aps.org/doi/10.1103/PhysRevB.37.785.

[23] Axel D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, April 1993. ISSN 0021-9606. doi: 10.1063/1.464913. URL https://doi.org/10.1063/1.464913.

[24] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple. *Physical Review Letters*, 77(18):3865–3868, October 1996. ISSN 0031-9007. doi: 10.1103/PhysRevLett.77.3865. URL https://link.aps.org/doi/10.1103/PhysRevLett.77.3865.

[25] Jochen Heyd, Gustavo E. Scuseria, and Matthias Ernzerhof. Hybrid functionals based on a screened Coulomb potential. *The Journal of Chemical Physics*, 118(18):8207–8215, May 2003. ISSN 0021-9606. doi: 10.1063/1.1564060. URL https://doi.org/10.1063/1.1564060.

[26] Stefan Grimme. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of Computational Chemistry*, 27(15):1787–1799, 2006. doi: https://doi.org/10.1002/jcc.20495. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20495.

[27] Alexandre Tkatchenko and Matthias Scheffler. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Physical Review Letters*, 102(7):073005, February 2009. doi: 10.1103/PhysRevLett.102.073005. URL https://link.aps.org/doi/10.1103/PhysRevLett.102.073005.

[28] Jianwei Sun, Adrienn Ruzsinszky, and John P. Perdew. Strongly Constrained and Appropriately Normed Semilocal Density Functional. *Physical Review Letters*, 115(3):036402, July 2015. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.115.036402. URL https://link.aps.org/doi/10.1103/PhysRevLett.115.036402.

[29] John C. Snyder, Matthias Rupp, Katja Hansen, Klaus-Robert Müller, and Kieron Burke. Finding Density Functionals with Machine Learning. *Physical Review Letters*, 108(25):253002, June 2012. doi: 10.1103/PhysRevLett.108.253002. URL https://link.aps.org/doi/10.1103/PhysRevLett.108.253002.

[30] Mihail Bogojeski, Leslie Vogt-Maranto, Mark E. Tuckerman, Klaus-Robert Müller, and Kieron Burke. Quantum chemical accuracy from density functional approximations via machine learning. *Nature Communications*, 11(1):5223, October 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19093-1. URL https://www.nature.com/articles/s41467-020-19093-1.

[31] Kyle Bystrom and Boris Kozinsky. CIDER: An Expressive, Nonlocal Feature Set for Machine Learning Density Functionals with Exact Constraints. *Journal of Chemical Theory and Computation*, 18(4):2180–2192, 2022. doi: 10.1021/acs.jctc.1c00904. URL https://doi.org/10.1021/acs.jctc.1c00904.

[32] Sebastian Dick and Marivi Fernandez-Serra. Machine learning accurate exchange and correlation functionals of the electronic density. *Nature Communications*, 11(1):3509, July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17265-7. URL https://www.nature.com/articles/s41467-020-17265-7.

[33] Xiangyun Lei and Andrew J. Medford. Design and analysis of machine learning exchange-correlation functionals via rotationally invariant convolutional descriptors. *Physical Review Materials*, 3(6):063801, June 2019. doi: 10.1103/PhysRevMaterials.3.063801. URL https://link.aps.org/doi/10.1103/PhysRevMaterials.3.063801.

[34] Yixiao Chen, Linfeng Zhang, Han Wang, and Weinan E. DeePKS: A Comprehensive Data-Driven Approach toward Chemically Accurate Density Functional Theory. *Journal of Chemical Theory and Computation*, 17(1):170–181, January 2021. ISSN 1549-9618, 1549-9626. doi: 10.1021/acs.jctc.0c00872. URL https://pubs.acs.org/doi/10.1021/acs.jctc.0c00872.

[35] Ryo Nagai, Ryosuke Akashi, and Osamu Sugino. Completing density functional theory by machine learning hidden messages from molecules. *npj Computational Materials*, 6(1):43, May 2020. ISSN 2057-3960. doi: 10.1038/s41524-020-0310-0.

[36] Pablo A. M. Casares, Jack S. Baker, Matija Medvidović, Roberto dos Reis, and Juan Miguel Arrazola. GradDFT. A software library for machine learning enhanced density functional theory. *The Journal of Chemical Physics*, 160(6):062501, February 2024. ISSN 0021-9606. doi: 10.1063/5.0181037. URL https://doi.org/10.1063/5.0181037.

[37] Beatriz G. Del Rio, Brandon Phan, and Rampi Ramprasad. A deep learning framework to emulate density functional theory. *npj Computational Materials*, 9(1):158, August 2023. ISSN 2057-3960. doi: 10.1038/s41524-023-01115-3. URL https://www.nature.com/articles/s41524-023-01115-3.

[38] James Kirkpatrick, Brendan McMorrow, David H. P. Turban, Alexander L. Gaunt, James S. Spencer, Alexander G. D. G. Matthews, Annette Obika, Louis Thiry, Meire Fortunato, David Pfau, Lara Román Castellanos, Stig Petersen, Alexander W. R. Nelson, Pushmeet Kohli, Paula Mori-Sánchez, Demis Hassabis, and Aron J. Cohen. Pushing the frontiers of density functionals by solving the fractional electron problem. *Science*, 374(6573):1385–1389, December 2021. ISSN 0036-8075. doi: 10.1126/science.abj6511.

[39] Johannes T. Margraf and Karsten Reuter. Pure non-local machine-learned density functional theory for electron correlation. *Nature Communications*, 12(1):344, January 2021. ISSN 2041-1723. doi: 10.1038/s41467-020-20471-y. URL https://www.nature.com/articles/s41467-020-20471-y.

[40] Kyle Bystrom and Boris Kozinsky. Nonlocal Machine-Learned Exchange Functional for Molecules and Solids. *Physical Review B*, 110(7):075130, August 2024. ISSN 2469-9950, 2469-9969. doi: 10.1103/PhysRevB.110.075130. URL http://arxiv.org/abs/2303.00682.

[41] Kyle Bystrom, Stefano Falletta, and Boris Kozinsky. Training Machine-Learned Density Functionals on Band Gaps. *Journal of Chemical Theory and Computation*, 20(17):7516–7532, September 2024. ISSN 1549-9618. doi: 10.1021/acs.jctc.4c00999. URL https://doi.org/10.1021/acs.jctc.4c00999.

[42] Charles C. Margossian. A review of automatic differentiation and its efficient implementation. *WIREs Data Mining and Knowledge Discovery*, 9(4), July 2019. ISSN 1942-4787. doi: 10.1002/widm.1305. URL https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1305.

[43] Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic Differentiation in Machine Learning: a Survey. *Journal of Machine Learning Research*, 18(153):1–43, 2018. URL http://jmlr.org/papers/v18/17-468.html.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. June 2017. URL http://arxiv.org/abs/1706.03762. arXiv: 1706.03762.

[45] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. October 2020. URL http://arxiv.org/abs/2010.11929.

[46] Shuhao Cao. Choose a Transformer: Fourier or Galerkin, November 2021. URL http://arxiv.org/abs/2105.14995. arXiv:2105.14995 [cs, math].

[47] Zijie Li, Kazem Meidani, and Amir Barati Farimani. Transformer for Partial Differential Equations' Operator Learning, April 2023. URL http://arxiv.org/abs/2205.13671.

[48] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding, November 2023. URL http://arxiv.org/abs/2104.09864.

[49] Richard M Martin. *Electronic structure: basic theory and practical methods*. Cambridge university press, 2020.

[50] R. O. Jones. Density functional theory: Its origins, rise to prominence, and future. *Rev. Mod. Phys.*, 87(3):897–923, August 2015. doi: 10.1103/RevModPhys.87.897. URL https://link.aps.org/doi/10.1103/RevModPhys.87.897.

[51] Stephan Kümmel and Leeor Kronik. Orbital-dependent density functionals: Theory and applications. *Rev. Mod. Phys.*, 80(1):3–60, January 2008. doi: 10.1103/RevModPhys.80.3. URL https://link.aps.org/doi/10.1103/RevModPhys.80.3.

[52] J. P. Perdew and Alex Zunger. Self-interaction correction to density-functional approximations for many-electron systems. *Physical Review B*, 23(10):5048–5079, May 1981. ISSN 0163-1829. doi: 10.1103/PhysRevB.23.5048. URL https://link.aps.org/doi/10.1103/PhysRevB.23.5048.

[53] David C Langreth and MJ Mehl. Beyond the local-density approximation in calculations of ground-state electronic properties. *Physical Review B*, 28(4):1809, 1983. URL https://journals.aps.org/prb/abstract/10.1103/PhysRevB.28.1809.

[54] Lorenz C. Blum and Jean-Louis Reymond. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *Journal of the American Chemical Society*, 131(25):8732–8733, July 2009. ISSN 0002-7863, 1520-5126. doi: 10.1021/ja902302h. URL https://pubs.acs.org/doi/10.1021/ja902302h.

[55] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole Von Lilienfeld. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters*, 108(5):058301, January 2012. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.108.058301. URL https://link.aps.org/doi/10.1103/PhysRevLett.108.058301.

[56] Jianmin Tao, John P. Perdew, Viktor N. Staroverov, and Gustavo E. Scuseria. Climbing the Density Functional Ladder: Nonempirical Meta–Generalized Gradient Approximation Designed for Molecules and Solids. *Physical Review Letters*, 91(14):146401, September 2003. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.91.146401. URL https://link.aps.org/doi/10.1103/PhysRevLett.91.146401.

[57] John P. Perdew, Adrienn Ruzsinszky, Gábor I. Csonka, Lucian A. Constantin, and Jianwei Sun. Workhorse Semilocal Density Functional for Condensed Matter Physics and Quantum Chemistry. *Physical Review Letters*, 103(2):026403, July 2009. doi: 10.1103/PhysRevLett.103.026403. URL https://link.aps.org/doi/10.1103/PhysRevLett.103.026403.

[58] Yan Zhao and Donald G. Truhlar. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *The Journal of Chemical Physics*, 125(19):194101, November 2006. ISSN 0021-9606. doi: 10.1063/1.2370993. URL https://doi.org/10.1063/1.2370993.

[59] Jianwei Sun, Bing Xiao, and Adrienn Ruzsinszky. Effect of the orbital-overlap dependence in the meta generalized gradient approximation. *The Journal of Chemical Physics*, 137(5): 051101, August 2012. ISSN 0021-9606. doi: 10.1063/1.4742312. URL https://doi.org/10.1063/1.4742312.

[60] Jianwei Sun, Robin Haunschild, Bing Xiao, Ireneusz W. Bulik, Gustavo E. Scuseria, and John P. Perdew. Semilocal and hybrid meta-generalized gradient approximations based on the understanding of the kinetic-energy-density dependence. *The Journal of Chemical Physics*, 138(4):044113, January 2013. ISSN 0021-9606. doi: 10.1063/1.4789414. URL https://doi.org/10.1063/1.4789414.

[61] Jianwei Sun, Richard C. Remsing, Yubo Zhang, Zhaoru Sun, Adrienn Ruzsinszky, Haowei Peng, Zenghui Yang, Arpita Paul, Umesh Waghmare, Xifan Wu, Michael L. Klein, and John P. Perdew. Accurate first-principles structures and energies of diversely bonded systems from an efficient density functional. *Nature Chemistry*, 8(9):831–836, September 2016. ISSN 1755-4349. doi: 10.1038/nchem.2535. URL https://www.nature.com/articles/nchem.2535.

[62] Albert P. Bartók and Jonathan R. Yates. Regularized SCAN functional. *The Journal of Chemical Physics*, 150(16):161101, April 2019. ISSN 0021-9606. doi: 10.1063/1.5094646. URL https://doi.org/10.1063/1.5094646.

[63] James W. Furness, Aaron D. Kaplan, Jinliang Ning, John P. Perdew, and Jianwei Sun. Accurate and Numerically Efficient r^2SCAN Meta-Generalized Gradient Approximation. *The Journal of Physical Chemistry Letters*, 11(19):8208–8215, October 2020. ISSN 1948-7185, 1948-7185. doi: 10.1021/acs.jpclett.0c02405. URL https://pubs.acs.org/doi/10.1021/acs.jpclett.0c02405.

[64] Axel D. Becke and Erin R. Johnson. A simple effective potential for exchange. *The Journal of Chemical Physics*, 124(22):221101, June 2006. ISSN 0021-9606. doi: 10.1063/1.2213970. URL https://doi.org/10.1063/1.2213970.

[65] Fabien Tran and Peter Blaha. Accurate Band Gaps of Semiconductors and Insulators with a Semilocal Exchange-Correlation Potential. *Physical Review Letters*, 102(22):226401, June 2009. doi: 10.1103/PhysRevLett.102.226401. URL https://link.aps.org/doi/10.1103/PhysRevLett.102.226401.

[66] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic Differentiation in PyTorch. In *NIPS 2017 Workshop Autodiff Decision Program Chairs*, 2017. doi: 10.1145/24680.24681. ISSN: 01635778.

[67] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–44, May 2015. ISSN 1476-4687. doi: 10.1038/nature14539. URL http://www.ncbi.nlm.nih.gov/pubmed/26017442.

[68] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL http://www.deeplearningbook.org.

[69] Anna Dawid, Julian Arnold, Borja Requena, Alexander Gresch, Marcin Płodzień, Kaelan Donatella, Kim A. Nicoli, Paolo Stornati, Rouven Koch, Miriam Büttner, Robert Okuła, Gorka Muñoz-Gil, Rodrigo A. Vargas-Hernández, Alba Cervera-Lierta, Juan Carrasquilla, Vedran Dunjko, Marylou Gabrié, Patrick Huembeli, Evert van Nieuwenburg, Filippo Vicentini, Lei Wang, Sebastian J. Wetzel, Giuseppe Carleo, Eliška Greplová, Roman Krems, Florian Marquardt, Michał Tomza, Maciej Lewenstein, and Alexandre Dauphin. Modern applications of machine learning in quantum sciences. April 2022. URL http://arxiv.org/abs/2204.04198.

[70] C. F. V. Weizsäcker. Zur Theorie der Kernmassen. *Zeitschrift für Physik*, 96(7-8):431–458, July 1935. ISSN 1434-6001, 1434-601X. doi: 10.1007/BF01337700. URL http://link.springer.com/10.1007/BF01337700.

[71] Shashank Subramanian, Peter Harrington, Kurt Keutzer, Wahid Bhimji, Dmitriy Morozov, Michael Mahoney, and Amir Gholami. Towards Foundation Models for Scientific Machine Learning: Characterizing Scaling and Transfer Behavior. May 2023. URL http://arxiv.org/abs/2306.00258.

[72] Michael McCabe, Bruno Régaldo-Saint Blancard, Liam Holden Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Geraud Krawezik, Francois Lanusse, Mariel Pettee, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. Multiple Physics Pretraining for Physical Surrogate Models, 2023. URL https://arxiv.org/abs/2310.02994.

[73] Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf.

[74] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. June 2020. URL http://arxiv.org/abs/2006.10739.

[75] Yang Li, Si Si, Gang Li, Cho-Jui Hsieh, and Samy Bengio. Learnable Fourier Features for Multi-Dimensional Spatial Positional Encoding. June 2021. URL http://arxiv.org/abs/2106.02795.

[76] Noam Shazeer. GLU Variants Improve Transformer, February 2020. URL https://arxiv.org/abs/2002.05202v1.

[77] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning. February 2017. URL http://arxiv.org/abs/1702.03118.

[78] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the Variance of the Adaptive Learning Rate and Beyond. August 2019. URL http://arxiv.org/abs/1908.03265.

[79] Diederik P. Kingma and Jimmy 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, December 2015.

[80] Ilya Loshchilov and Frank Hutter. Fixing Weight Decay Regularization in Adam. 2017. URL http://arxiv.org/abs/1711.05101.

[81] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy,

Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[82] Qiming Sun, Timothy C. Berkelbach, Nick S. Blunt, George H. Booth, Sheng Guo, Zhendong Li, Junzi Liu, James D. McClain, Elvira R. Sayfutyarova, Sandeep Sharma, Sebastian Wouters, and Garnet Kin-Lic Chan. PySCF: the Python-based simulations of chemistry framework. *WIREs Computational Molecular Science*, 8(1):e1340, 2018. doi: https://doi.org/10.1002/wcms.1340. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1340.

[83] Qiming Sun, Xing Zhang, Samragni Banerjee, Peng Bao, Marc Barbry, Nick S. Blunt, Nikolay A. Bogdanov, George H. Booth, Jia Chen, Zhi-Hao Cui, Janus J. Eriksen, Yang Gao, Sheng Guo, Jan Hermann, Matthew R. Hermes, Kevin Koh, Peter Koval, Susi Lehtola, Zhendong Li, Junzi Liu, Narbe Mardirossian, James D. McClain, Mario Motta, Bastien Mussard, Hung Q. Pham, Artem Pulkin, Wirawan Purwanto, Paul J. Robinson, Enrico Ronca, Elvira R. Sayfutyarova, Maximilian Scheurer, Henry F. Schurkus, James E. T. Smith, Chong Sun, Shi-Ning Sun, Shiv Upadhyay, Lucas K. Wagner, Xiao Wang, Alec White, James Daniel Whitfield, Mark J. Williamson, Sebastian Wouters, Jun Yang, Jason M. Yu, Tianyu Zhu, Timothy C. Berkelbach, Sandeep Sharma, Alexander Yu. Sokolov, and Garnet Kin-Lic Chan. Recent developments in the PySCF program package. *The Journal of Chemical Physics*, 153(2):024109, July 2020. ISSN 0021-9606. doi: 10.1063/5.0006074. URL https://doi.org/10.1063/5.0006074.

[84] Pablo del Mazo-Sevillano and Jan Hermann. Variational principle to regularize machine-learned density functionals: The non-interacting kinetic-energy functional. *The Journal of Chemical Physics*, 159(19), November 2023. ISSN 0021-9606. URL https://doi.org/10.1063/5.0166432.

[85] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951. ISSN 0003-4851. doi: 10.1214/aoms/1177729694. URL http://projecteuclid.org/euclid.aoms/1177729694.

[86] Miguel A. L. Marques, Micael J. T. Oliveira, and Tobias Burnus. Libxc: A library of exchange and correlation functionals for density functional theory. *Computer Physics Communications*, 183(10):2272–2281, October 2012. ISSN 0010-4655. doi: 10.1016/j.cpc.2012.05.007. URL https://www.sciencedirect.com/science/article/pii/S0010465512001750.

[87] Susi Lehtola, Conrad Steigemann, Micael J. T. Oliveira, and Miguel A. L. Marques. Recent developments in libxc — A comprehensive library of functionals for density functional theory. *SoftwareX*, 7:1–5, January 2018. ISSN 2352-7110. doi: 10.1016/j.softx.2017.11.002. URL https://www.sciencedirect.com/science/article/pii/S2352711017300602.

[88] Mark E. Casida. Time-Dependent Density Functional Response Theory for Molecules. In *Recent Advances in Density Functional Methods*, volume Volume 1 of *Recent Advances in Computational Chemistry*, pages 155–192. World Scientific, November 1995. ISBN 978-981-02-2442-4. doi: 10.1142/9789812830586_0005. URL https://www.worldscientific.com/doi/abs/10.1142/9789812830586_0005.

[89] Wing-Cheung Tam and C. E. Brion. Electron impact spectra of some alkyl derivatives of water and related compounds. *Journal of Electron Spectroscopy and Related Phenomena*, 3 (4):263–279, January 1974. ISSN 0368-2048. doi: 10.1016/0368-2048(74)80079-4. URL https://www.sciencedirect.com/science/article/pii/0368204874800794.

[90] Peter J. Linstrom and William G. Mallard. The NIST Chemistry WebBook: A Chemical Data Resource on the Internet. *Journal of Chemical & Engineering Data*, 46(5):1059–1063, September 2001. ISSN 0021-9568. doi: 10.1021/je000236i. URL https://doi.org/10.1021/je000236i.

[91] 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. *The Journal of Physical Chemistry*, 98(45):11623–11627, November 1994. ISSN 0022-3654. doi: 10.1021/j100096a001. URL https://doi.org/10.1021/j100096a001.

[92] Carlo Adamo and Vincenzo Barone. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *The Journal of Chemical Physics*, 110(13):6158–6170, April 1999. ISSN 0021-9606. doi: 10.1063/1.478522. URL https://doi.org/10.1063/1.478522.