Distance weighted self-attention for nonlocal density functional approximation by artificial neural network

Kirill Kulaev [1] **Bogdan Protsenko** [2] **Alexander Ryabov** [1] **Alexander Guda** [2] **Evgeny Burnaev** [1] **Vladimir Vanovskiy** [1]

[1] *Skolkovo Institute of Science and Technology, Artificial Intelligence Center, Moscow, 121205, Russian Federation*
[2] *Southern Federal University, The Smart Materials Research Institute, Rostov-on-Don, 344090, Russian Federation.*
Correspondence to: Alexander Ryabov A.Ryabov@skoltech.ru.

1. Introduction

One of the most popular and efficient methods of quantum chemistry is density functional theory (DFT), the key parameter of which is the exchange-correlation (XC) functional, which determines the electron-electron interaction. Almost all density functionals approximations (DFA's) are local, which leads to a lack of proper accounting for electron correlation, while accounting for nonlocal correlation requires significantly more resources. Different designs of neural network (NN) approximation of XC energy have been proposed [1, 2, 3, 4]. Up to this point, most work in the field of NN DFA has focused on feature engineering in framework of local functional [5, 4] when global architectures [6, 7] suffer from non self-consistent behavior that limits it's practical application. In this paper, we investigate a new architecture for this problem and show that it is possible to efficiently account for nonlocal interactions by applying the modified transformer architecture directly to an electron density grid.

2. Methods

Most of DFT simulations perform self-consistent calculation of the total potential energy of system:

$$E_{\text{tot}} = E_{\text{ext}} + E_{\text{J}} + E_{\text{xc}}, \tag{1}$$

Where E_{tot} is calculated from E_{ext} - the energy due to the external potential, Coulumb energy E_{J} and exchange-correlation energy E_{xc}

The challenge in DFT lies in approximating $E_{\text{xc}}[\rho]$ since its exact form is unknown. The continuos need in more precise DFA's led to their evolution from local approximation to generalized gradient and to meta-generalized global Hartree-Fock exchange correction. It has been shown that non-local functionals provide remarkably more accurate results especially for long-range interactions, such as Van-der-Waals complexes [8].

The functional we present in this study is a hybrid [9] functional, which means that it contains a part of "exact" Hartree-Fock (HF) exchange energy. Total exchange-correlation energy can be written as:

$$E_{\text{xc}} = \int e_{\text{xc}}[\rho(\mathbf{r}), \nabla\rho(\mathbf{r}), ...]\rho(\mathbf{r})d\mathbf{r} + \alpha E_{\text{x}}^{HF}, \tag{2}$$

where $\rho(\mathbf{r})$ is electron density, $e_{xc}(\mathbf{r})$ is XC energy per particle in point $\mathbf{r}$, E_{x}^{HF} is Hartree-Fock exchange energy, coefficient of HF exchange energy α.

2.1 Deep-learning DFA

The entire architecture of our DFA by NN consists of three components: encoder, distance-weighted self-attention (the idea of which is illustrated in the Fig.1) and decoder. Encoder and decoder are composed from three fully connected layers with residual connections and SoftPlus activation. The encoder's role consist in mapping the local features of electron density $\mathbf{x}(\mathbf{r})$ (that will be described in next section), at each point into a vector of higher dimensionality, which is fed to the attention mechanism as input. The distance-weighted self-attention block employs the following formula:

$$\mathbf{a}_i(\mathbf{q}_i, \mathbf{K}, \mathbf{V}, \mathbf{D}_i) = [\mathbf{D}_i \odot \text{SoftMax}(\frac{\mathbf{q}_i\mathbf{K}^{\text{T}}}{t})]\mathbf{V} \tag{3}$$

where $\mathbf{a}_i$ is attention vector for i point of electron density depending on query vector $\mathbf{q}_i$ of point, keys and values matrices $\mathbf{K}$ and $\mathbf{V}$ containing information of all points, SoftMax temperature t. It inputs the sequence of grid points with embeddings constructed by the encoder, the only difference from the original attention mechanism [10] is the consideration of spatial distances between the points on the grid. To do this, attention weights are scaled by $\mathbf{D}_i$ is a vector of distance-dependent weights, $\mathbf{D}_{ij} = f(d_{ij})$, where d_{ij} is distance between points i, j. Decoder projects the nonlocal embedding of electron density into the exchange-correlation energy per particle (e_{xc} from eq.2).

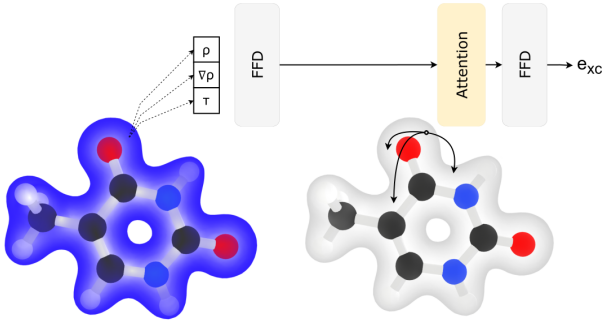

Fig. 1: Scheme of XC functional based on attention mechanism.

In order to speed up the computations the attention is computed for electron density near each atom separately with some hyperparameter R_{attn} defining the radius around the atom within which attention is computed, after that the result is normalized by the

number of atoms that influence current point and finally mapped onto a single grid.

In the proposed version of our functional we use pre-processing and post-processing of NN's inputs and outputs to account for physical constraints, HF exchange coefficient α of 0.25 (non-empirical estimation used in hybrid PBE0 functional [11]) and the distance dependence in Attention block expressed in exponential form: $f(d) = e^{-0.5d^3}$, providing a smooth lowering of the impact from more distant points, R_{attn} of 2.3 Bohr chosen empirically close to the length of a typical covalent bond in our dataset.

2.2 Features of electron density

We use the next electron density descriptors: reduced density gradient $s(\mathbf{r}) = \frac{|\nabla\rho(\mathbf{r})|}{\rho(\mathbf{r})^{4/3}}$, spin-polarization $\zeta(\mathbf{r}) = \frac{\rho^{\downarrow}(\mathbf{r})-\rho^{\uparrow}(\mathbf{r})}{\rho(\mathbf{r})}$ and kinetic energy-density $\tau^{\sigma}(\mathbf{r}) = \frac{1}{2}\sum_{i}^{\text{occup}} |\nabla\Psi_i^{\sigma}(\mathbf{r})|^2$ depending on molecular orbitals $\Psi_i^{\sigma}(\mathbf{r})$ in spin channels $\sigma \in (\uparrow, \downarrow)$.

Then we apply preprocessing and logarithmic scaling of electron density features to construct input vector $\mathbf{x}(\mathbf{r})$ which is described in Appendix "NN's inputs". This helps to balance the input values and ensures the model independence of the swapping of spin channels σ.

2.3 Post-processing of NN's outputs

For a more convenient implementation of constraints we divide output of neural network into two parts: $h_1(\mathbf{r})$ and $h_2(\mathbf{r})$. Exchange-correlation energy per particle from the eq. 2 sums up both factors $e_{xc}(\mathbf{r}) = f_{\sigma}(\mathbf{r}) + f_{\beta}(\mathbf{r})$.

First output h_1 parametrizes formula ensuring it's correct asymptotic behavior in the case of the uniformly spin-polarized electron gas [12] by construction.

$$f_{\sigma}(\mathbf{r}) = h_1(\mathbf{r}) e_{\text{X}}^{\text{LDA}}(\mathbf{r}) \frac{1}{2}[(1+\zeta(\mathbf{r})^{4/3}) + (1-\zeta(\mathbf{r})^{4/3})]$$

$$e_{\text{X}}^{\text{LDA}}(\mathbf{r}) = -\frac{3}{4}(\frac{3}{\pi})^{1/3}\rho(\mathbf{r})^{1/3}$$

The second NN's output $h_2(\mathbf{r})$ strongly depends on attention weights and accounts for non-local interactions, β is the constant inspired by the VV10 [13] functional construction and has a value of 0.03 E_h (hartree energy units) that was selected empirically.

$$f_{\beta}(\mathbf{r}) = \beta - h_2(\mathbf{r})\rho(\mathbf{r})^{1/6} \tag{4}$$

2.4 Data and training details

Training data represented in energies of chemical reactions, therefore we used least-squares training objective $\mathcal{L}$ between reference reaction energy ΔE_{Ref} and reaction energy calculated by the neural network where ν_i is a stoichiometric coefficient (with negative values for reagents) of substance (S) in reaction (r) from the training dataset.

$$\mathcal{L} = \frac{1}{N}\sum_{r}^{N}(\sum_{i}^{S} E_i^{\text{tot}}\nu_{i,r} - \Delta E_{r,\text{Ref}})^2 \tag{5}$$

We constructed training dataset from the accurate reaction energies and corresponding molecular electron densities. For this purpose, we combined next datasets: G3, W4-17, S66x8 and G21EA, G21IP, MB-165, BH76 from the GMTKN database [14]. G3, G21EA and G21IP include back-corrected experimental energy differences, when W4-17 [15], BH76, S66x8, MB-165 contain reaction energies calculated using high-level wave function theory methods. Molecular electron densities of all systems were calculated using PySCF [16] unrestricted Kohn-Sham DFT with the PBE0 functional and def2-QZVPD basis set.

We trained the model for 1000 epochs, for first 150 epochs we used batch size of 16 and learning rate of $3 \cdot 10^{-4}$, then we reduced them to 8 and $3 \cdot 10^{-5}$, respectively.

2.5 Results

We report the functional employing distance weighted self-attention mechanism allows self-consistent calculations and achieves high precision on a various benchmarks not represented in the training sample. Table 1 shows the mean absolute deviation in reaction energy in kcal/mol on test benchmarks for our "Attentive Density" (AD) functional and several classic empirical (B3LYP, M06-2X) and non-empirical (SCAN) XC functionals. Calculations were carried out with def2-TZVP basis set.

Table 1: Benchmark results.

	AD	B3LYP	SCAN	M06-2X
BHPERI	**0.76**	4.15	4.82	1.44
PX13	**1.76**	10.45	6.58	5.12
BHDIV10	2.23	4.36	7.54	**0.89**
BSR36	3.31	11.24	11.72	**3.25**
WCPT18	**1.71**	2.11	8.05	2.48
Average	**1.91**	5.74	7.79	2.61

3. Conclusion

We have presented a novel approach to incorporate nonlocalities into density functional theory through the distance-weighted self-attention mechanism. By leveraging a modified transformer architecture, we have demonstrated that it is feasible to efficiently capture the complexities of electron interactions across a broader spatial range while maintaining computational efficiency. Our hybrid functional effectively combines the strengths of both local and nonlocal approximations, enabling more accurate modeling of chemical systems.

Acknowledgments

The authors are grateful to the support of the Russian Science Foundation (Grant No. 24-41-02035).

Appendix A. NN's inputs

The input vector $\mathbf{x}(\mathbf{r})$ is composed from electron density features according to the following formulas:

$$\mathbf{x}_1(\mathbf{r}) = \log \rho(\mathbf{r})^{1/3}$$
$$\mathbf{x}_2(\mathbf{r}) = \log(\frac{1}{2}[(1+\zeta(\mathbf{r})^{4/3}) + (1-\zeta(\mathbf{r})^{4/3})])$$
$$\mathbf{x}_3(\mathbf{r}) = \log s(\mathbf{r})$$
$$\mathbf{x}_4(\mathbf{r}) = \log \frac{\tau^{\downarrow}(\mathbf{r}) + \tau^{\uparrow}(\mathbf{r})}{\rho(\mathbf{r})^{5/3}[(1+\zeta(\mathbf{r}))^{5/3} - (1-\zeta(\mathbf{r}))^{5/3}]}$$

References

[1] Alexander Ryabov, Iskander Akhatov, and Petr Zhilyaev. Application of two-component neural network for exchange-correlation functional interpolation. *Scientific Reports*, 12(1), August 2022.

[2] Alexander Ryabov, Iskander Akhatov, and Petr Zhilyaev. Neural network interpolation of exchange-correlation functional. *Scientific Reports*, 10(1), May 2020.

[3] Ryo Nagai, Ryosuke Akashi, and Osamu Sugino. Machine-learning-based exchange correlation functional with physical asymptotic constraints. *Physical Review Research*, 4(1), February 2022.

[4] 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.

[5] Ryo Nagai, Ryosuke Akashi, and Osamu Sugino. Completing density functional theory by machine learning hidden messages from molecules. *npj Computational Materials*, 6(1), May 2020.

[6] Weiyi Gong, Tao Sun, Hexin Bai, Shah Tanvir ur Rahman Chowdhury, Peng Chu, Anoj Aryal, Jie Yu, Haibin Ling, John P. Perdew, and Qimin Yan. Incorporation of density scaling constraint in density functional design via contrastive representation learning. *Digital Discovery*, 2(5):1404–1413, 2023.

[7] Wenze Li, Donghan Wang, Zirui Yang, Huijie Zhang, LiHong Hu, and GuanHua Chen. Deepnci: Dft noncovalent interaction correction with transferable multimodal three-dimensional convolutional neural networks. *Journal of Chemical Information and Modeling*, 62(21):5090–5099, December 2021.

[8] Narbe Mardirossian and Martin Head-Gordon. wb97x-v: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. *Physical Chemistry Chemical Physics*, 16(21):9904, 2014.

[9] A Becke. Density-functional thermochemistry. iii. the role of exact exchange (1993) j. *Chem. Phys*, 98:5648, 1993.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[11] 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.

[12] G. L. Oliver and J. P. Perdew. Spin-density gradient expansion for the kinetic energy. *Physical Review A*, 20(2):397–403, August 1979.

[13] Oleg A Vydrov and Troy Van Voorhis. Nonlocal van der waals density functional: the simpler the better. *J. Chem. Phys.*, 133(24):244103, December 2010.

[14] Lars Goerigk, Andreas Hansen, Christoph Bauer, Stephan Ehrlich, Asim Najibi, and Stefan Grimme. A look at the density functional theory zoo with the advanced gmtkn55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Physical Chemistry Chemical Physics*, 19(48):32184–32215, 2017.

[15] Amir Karton, Nitai Sylvetsky, and Jan M. L. Martin. W4-17: A diverse and high-confidence dataset of atomization energies for benchmarking high-level electronic structure methods. *Journal of Computational Chemistry*, 38(24):2063–2075, July 2017.

[16] Qiming Sun, Timothy C. Berkelbach, Nick S. Blunt, George H. Booth, Sheng Guo, Zhendong Li, Junzi Liu, James McClain, Elvira R. Sayfutyarova, Sandeep Sharma, Sebastian Wouters, and Garnet Kin-Lic Chan. The python-based simulations of chemistry framework (pyscf), 2017.