Towards Data-Driven Nonlocal Density Functionals: Deep Learning DFT with Attention to approach Chemical Accuracy

Kirill Kulaev [1,2] **Bogdan Protsenko** [1] **Alexander Ryabov** [2] **Alexander Guda** [1] **Evgeny Burnaev** [2,3] **Vladimir Vanovskiy** [2]

[1]*Southern Federal University, The Smart Materials Research Institute, Rostov-on-Don, 344090, Russian Federation*
[2]*Skolkovo Institute of Science and Technology, Artificial Intelligence Center, Moscow, 121205, Russian Federation*
[3]*Autonomous Non-Profit Organization Artificial Intelligence Research Institute (AIRI), Learnable Intelligence Group, Moscow, 121170, Russian Federation*. Correspondence to: Alexander Ryabov ununbium17@gmail.com.

1. Introduction

Density functional theory (DFT) remains the workhorse for electronic-structure calculations, and its practical accuracy is largely set by the exchange–correlation (XC) functional [1]. Many widely used approximations - especially local and semilocal ones - use only the density and a small set of derivative-based ingredients [2], and therefore miss important nonlocal electronic effects. That is exactly where many hard cases live: stretched bonds, barrier heights, noncovalent binding, and related situations in which the exact XC contribution depends on density structure beyond the immediate local neighborhood [3].

In our previous work we showed that attention on atom-centered quadrature grids is a convenient way to inject such nonlocal dependence while staying inside stable convergence of a self-consistent field (SCF) [4]. The problem is cost. If every grid point attends to every other grid point, the operator scales as $O(N^2)$ in the number of grid points, which quickly dominates an SCF step. This abstract reports the best model from our subsequent trials: a hybrid neural XC functional that keeps the familiar B3LYP structure but replaces the expensive global attention with an atom-centered linear-scaling attention block.

2. Method

2.1 Hybrid XC functional

We work with a hybrid XC form

$$E_{\mathrm{xc}}[\rho] = \int \rho(\mathbf{r})\, \varepsilon_{\mathrm{xc}}(\mathbf{r})\, d\mathbf{r} \;+\; \alpha\, E_x^{\mathrm{HF}}, \tag{1}$$

and evaluate it self-consistently. The learned part produces a smooth XC energy density per particle $\varepsilon_{\mathrm{xc}}(\mathbf{r})$ on the molecular quadrature grid. Rather than learning XC "from scratch", we treat the network as an enhancement of the stable and widely used B3LYP [5] functional.

We implement separate exchange and correlation subnetworks. The exchange subnetwork outputs spin-channel enhancement factors that scale the Slater and B88 contributions, while the correlation subnetwork outputs an enhancement factor that scales the VWN+LYP combination (see Appendix. B).

2.2 Attention in DFT

A convenient way to express nonlocality on a molecular quadrature grid is to let each grid point $\mathbf{r}_i$ form a weighted average of information from other points $\mathbf{r}_j$, $i \neq j$ where the weights reflect both (i) similarity of the local electronic environments and (ii) their spatial separation. Starting from learned projections $q(x[\mathbf{r}]), k(x[\mathbf{r}]), v(x[\mathbf{r}])$ (where x is feature vector corresponding to point $\mathbf{r}$ (see Appendix. A), we denote $q[\mathbf{r}_i]$ for brevity) of local descriptors, a natural nonlocal operator is

$$\mathbf{y}(\mathbf{r}_i) = \sum_{j=1}^{N} \left(q([\mathbf{r}_i]^\top k[\mathbf{r}_j]\right)\; f(\mathbf{r}_i - \mathbf{r}_j)\, v[\mathbf{r}]_j\, w_j, \tag{2}$$

where $\mathbf{r}_j, w_j$ are grid points and quadrature weights and $f(\mathbf{r}_i - \mathbf{r}_j)$ is a smooth distance kernel. The complexity is $O(N^2)$.

Linear-scaling attention methods, including Euclidean Fast Attention (EFA) [6], obtain $O(N)$ cost by rewriting the kernel as an inner product in a finite feature space, which in our setting corresponds to approximating the spatial kernel by a low-rank factorization (for example via random Fourier features, RFF) [7]. With a feature map $\phi(\mathbf{r}) \in \mathbb{R}^{F_x}$ such that $k(\mathbf{r}_1 - \mathbf{r}_2) \approx \phi(\mathbf{r}_1)^\top \phi(\mathbf{r}_2)$, Eq. (2) becomes

$$\mathbf{y}(\mathbf{r}_i) = \sum_{j=1}^{N} \left(q[\mathbf{r}_i]^\top k[\mathbf{r}_j]\right)\; \left(\phi[\mathbf{r}_i]^\top \phi[\mathbf{r}_j]\right)\; v[\mathbf{r}]_j\, w_j, \tag{3}$$

, and can be rearranged into two contractions,

$$C_{fgd} = \sum_{j=1}^{N} k[\mathbf{r}_j]_f\, \phi_g[\mathbf{r}_j]\, v[\mathbf{r}]_{j,d}\, w_j, \tag{4}$$

$$y_{i,d} = \sum_{f,g} C_{fgd}\, q[\mathbf{r}_i]_f\, \phi_g[\mathbf{r}_i]. \tag{5}$$

where g frequency index and d is dimension of NN representation. This yields $O(N)$ complexity and is mathematically appealing.

In SCF, however, the relevant quantity is not only the XC energy but its functional derivative. In our experiments, global Fourier Features factorizations were numerically unstable, leading to unstable SCF convergence in our tests: small oscillations in the nonlocal term lead to noisy XC potential v_{xc} unless the approximation is heavily smoothed or uses very high ranks.

We therefore keep the real-space, distance-gated form of Eq. (2) but make it linear-scaling by routing interactions through atom-centered coarse points. This is closely aligned with the coarse-point positional

encoding used in Skala [8]: rather than approximating *all* pairwise grid–grid distances, we summarize the environment around each atom and broadcast it back to nearby grid points with smooth envelopes and SO(3)-equivariant angular channels. The result is an attention-like operator that remains smooth and differentiable in real space and is practical inside SCF.

3. Methods

3.1 E3-equivariant Factorized Attention Block (AttF)

Let $\{\mathbf{R}_a\}_{a=1}^{A}$ be the nuclear coordinates and define the edge set under cutoff r_c:

$$E = \{(i,a) \;:\; d_{ia} = \|\mathbf{r}_i - \mathbf{R}_a\| \le r_c\}.$$

The block maps an input embedding $\mathbf{h}_i \in \mathbb{R}^{d_{\text{emb}}}$ to a correction $\Delta\mathbf{h}_i$. First we project to a lower-dimensional latent space using linear layer with learned down-projection matrix $W_{\downarrow}$,

$$\mathbf{u}_i = \mathrm{SiLU}(W_{\downarrow}\mathbf{h}_i) \in \mathbb{R}^d. \tag{6}$$

For each edge (i,a) we compute a radial weight vector $\mathbf{g}(d_{ia}) \in \mathbb{R}^d$ and a smooth cutoff envelope $p(d_{ia}; r_c)$ (polynomial with $p(r_c) = 0$ and vanishing derivative), and set $\mathbf{r}_{ia} = \mathbf{g}(d_{ia}) \odot p(d_{ia}; r_c)$. We also compute spherical harmonics $Y(\widehat{\mathbf{d}}_{ia})$ up to $\ell_{\max} = 2$, where $\widehat{\mathbf{d}}_{ia} = (\mathbf{r}_i - \mathbf{R}_a)/d_{ia}$. An equivariant tensor product $\mathrm{TP}_{\downarrow}$ produces an edge representation,

$$\mathbf{q}_{ia} = \mathrm{TP}_{\downarrow}\big(\mathbf{u}_i,\; Y(\widehat{\mathbf{d}}_{ia})\big), \qquad \mathbf{q}_{ia} \leftarrow \mathbf{r}_{ia} \odot \mathbf{q}_{ia}, \tag{7}$$

where the last multiplication is applied blockwise over irreps (channel-wise scaling of each irrep slice) [9].

Downsampling forms atom summaries by quadrature-weighted accumulation,

$$\mathbf{H}_a = \sum_{(i,a)\in E} \mathbf{q}_{ia}\, w_i. \tag{8}$$

To obtain an attention-like message, we compute an edge gate by taking a normalized inner product between $\mathbf{q}_{ia}$ with $\mathbf{H}_a$ in each ℓ-block via a normalized inner product and applying SiLU:

$$\mathbf{s}_{ia} = \mathrm{SiLU}\left(p(d_{ia}; r_c) \sum_{\ell=0}^{\ell_{\max}} \frac{\langle \mathbf{q}_{ia}^{(\ell)}, \mathbf{H}_a^{(\ell)} \rangle}{\sqrt{2\ell+1}}\right) \in \mathbb{R}^d. \tag{9}$$

Upsampling broadcasts $\mathbf{H}_a$ back to edges through a second tensor product,

$$\mathbf{z}_{ia} = \mathrm{TP}_{\uparrow}\big(\mathbf{H}_a,\; Y(\widehat{\mathbf{d}}_{ia})\big), \tag{10}$$

and applies a distance normalization envelope $\nu(d_{ia})$ [8] that is normalized per grid point to reduce neighbor-count bias [8],

$$\tilde{w}_{ia} = \frac{\nu(d_{ia})}{\sum_{a':(i,a')\in E} \nu(d_{ia'}) + 0.1}. \tag{11}$$

The final edge message is then formed by combining radial weights, normalized envelope, and the gate,

$$\mathbf{m}_{ia} = (\mathbf{r}_{ia} \odot \tilde{w}_{ia} \odot \mathbf{s}_{ia}) \odot \mathbf{z}_{ia}, \tag{12}$$

and aggregated to grid points, post-processed, and damped near the core,

$$\Delta\mathbf{h}_i = W_{\uparrow}\,\mathrm{SiLU}\left(W_{\text{post}} \sum_{(i,a)\in E} \mathbf{m}_{ia}\right) \odot \exp\big(-\rho(\mathbf{r}_i)\big). \tag{13}$$

In this case $W_{\uparrow}$ and W_{post} are trainable matrices. Because all steps are sums over edges (i,a) within a cutoff, the computational cost is $O(|E|)$ and scales linearly with system size on standard atom-centered grids. [6, 7].

4. Data, training, and validation

Training uses reaction energies paired with molecular electron densities evaluated on atom-centered quadrature grids. The final training set contains 1464 reactions and is assembled by combining a selected subset of MSR_TAE25, S66x8 noncovalent interaction curves and RDB7 [10].

AttF is trained on organic systems and, in our best configuration, reaches a mean absolute error (MAE) of 1.38 kcal/mol on GMTKN_slim [11], close to the ∼ 1 kcal/mol "chemical accuracy" target and in the range of high-level methods, and 1.74 kcal/mol on an RDB7 validation set, placing it between the range-separated hybrid ωB97X:D_3 (2.15 kcal/mol) and the double-hybrid ωB97M(2) (1.36 kcal/mol). On GMTKN subsets dominated by covalent thermochemistry and barrier heights, AttF often lowers the MAE relative to ωB97M-V by about 40–80% (e.g., YBDE18, BSR36, MB16-43, DIPCS10), whereas several noncovalent and conformational subsets governed by dispersion and hydrogen-bonding effects (BUT14DIOL, TAUT15, SCONF, UPU23, PNICO23, HAL59) show larger errors than ωB97M-V, in some cases by factors of 2–5; this pattern suggests that the dispersion model inherited from B3LYP-D3(BJ) is still suboptimal for certain noncovalent regimes and that a more refined, potentially learned, dispersion correction is a natural route to further reduce the remaining discrepancies.

5. Results

The results show that the main contribution of this work is the design of the AttF nonlocal block, which enables a functional with high accuracy at linear-scaling cost. Overall, our AttF model achieves near-chemical accuracy on GMTKN_slim (1.38 kcal/mol) and competitive performance on RDB7 (1.74 kcal/mol), being better or on par with state-of-the-art range-separated hybrid functionals (see Appendix C for the GMTKN subset-wise results).

Acknowledgments

The study was supported by the Russian Science Foundation grant No. 24-41-02035, `https://rscf.ru/en/project/24-41-02035/`.

References

[1] Giovanni Vignale et al. Dft exchange: sharing perspectives on the workhorse of quantum

chemistry and materials science. *Physical Chemistry Chemical Physics*, 24(48):29579–29643, 2022.

[2] John P. Perdew and coauthors. Semi-local exchange-correlation approximations in density functional theory. *arXiv preprint*, 2026.

[3] Kristian Berland and Per Hyldgaard. The vdW-DF family of nonlocal exchange-correlation functionals. *Reports on Progress in Physics*, 78(6):066501, 2015.

[4] Kirill Kulaev, Bogdan Protsenko, Alexander Ryabov, Evgeny Burnaev, and Vladimir Vanovskiy. Distance weighted self-attention for nonlocal density functional approximation by artificial neural network. AI4X – Accelerate (oral presentation), 2025.

[5] A. D. Becke. Density-functional thermochemistry. III. the role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, 1993.

[6] J. Thorben Frank, Stefan Chmiela, Klaus-Robert Müller, and Oliver T. Unke. Euclidean fast attention – machine learning global atomic representations at linear cost, 2025.

[7] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, 2007.

[8] Giulia Luise, Chin-Wei Huang, Thijs Vogels, Derk P. Kooi, Sebastian Ehlert, Stephanie Lanius, Klaas J. H. Giesbertz, Amir Karton, Deniz Gunceler, Megan Stanley, Wessel P. Bruinsma, Lin Huang, Xinran Wei, José Garrido Torres, Abylay Katbashev, Rodrigo Chavez Zavaleta, Bálint Máté, Sékou-Oumar Kaba, Roberto Sordillo, Yingrong Chen, David B. Williams-Young, Christopher M. Bishop, Jan Hermann, Rianne van den Berg, and Paola Gori-Giorgi. Accurate and scalable exchange-correlation with deep learning, 2025.

[9] Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks, 2022.

[10] Xiao Liu, Kevin A. Spiekermann, Angiras Menon, William H. Green, and Martin Head-Gordon. Revisiting a large and diverse data set for barrier heights and reaction energies: best practices in density functional theory calculations for chemical kinetics. *Physical Chemistry Chemical Physics*, 27(25):13326–13339, 2025.

[11] Tim Gould and Stefan Vuckovic. "slim" benchmark sets for faster method development. *Journal of Chemical Theory and Computation*, 21(13):6517–6527, 2025.

Appendix A. Input feature preprocessing

At each quadrature point $\mathbf{r}_i$ we start from spin densities and semilocal quantities provided by the DFT code. For numerical stability we add a small constant ε (10^{-7} depending on the quantity) before divisions and logarithms, and we clamp denominators away from zero.

For spin-polarized systems we use spin densities $\rho^\uparrow, \rho^\downarrow$ and define

$$\rho = \rho^\uparrow + \rho^\downarrow, \qquad \zeta = \frac{\rho^\downarrow - \rho^\uparrow}{\rho}. \tag{A1}$$

We form gradient invariants

$$\gamma_{\uparrow\uparrow} = |\nabla\rho^\uparrow|^2, \quad \gamma_{\downarrow\downarrow} = |\nabla\rho^\downarrow|^2, \quad \gamma_{\uparrow\downarrow} = \nabla\rho^\uparrow \cdot \nabla\rho^\downarrow, \tag{A2}$$

and use kinetic energy densities $\tau^\uparrow, \tau^\downarrow$. For spin-unpolarized inputs we set $\rho^\uparrow = \rho^\downarrow = \rho/2$ and split τ analogously; the gradient invariants are constructed consistently from the shared gradient.

The exchange subnetwork uses per-spin log-scaled inputs

$$x^\sigma_{x,1} = \log\Big((\rho^\sigma)^{1/3} + \varepsilon\Big), \tag{A3}$$

$$x^\sigma_{x,2} = \log\big(\sqrt{\gamma_{\sigma\sigma}} + \varepsilon\big), \tag{A4}$$

$$x^\sigma_{x,3} = \log(\tau^\sigma + \varepsilon), \tag{A5}$$

with $\sigma \in \{\uparrow, \downarrow\}$.

The correlation subnetwork uses spin-averaged inputs based on

$$s = \frac{\sqrt{\gamma_{\uparrow\uparrow} + \gamma_{\downarrow\downarrow} + 2\gamma_{\uparrow\downarrow}}}{(\rho^{1/3} + \varepsilon)^4}, \qquad \mathrm{SS} = \tfrac{1}{2}\Big((1+\zeta)^{4/3} + (1-\zeta)^{4/3}\Big), \tag{A6}$$

and the dimensionless kinetic-energy combination

$$\mathrm{DS} = (1+\zeta)^{5/3} + (1-\zeta)^{5/3}, \qquad t = \frac{\tau^\uparrow + \tau^\downarrow + \varepsilon}{(\rho^{1/3} + \varepsilon)^5\,(\mathrm{DS} + \varepsilon)}. \tag{A7}$$

The final correlation inputs are

$$x_{c,1} = \log\Big(\rho^{1/3} + \varepsilon\Big), \tag{A8}$$

$$x_{c,2} = \log(s + \varepsilon), \tag{A9}$$

$$x_{c,3} = \log(\mathrm{SS} + \varepsilon), \tag{A10}$$

$$x_{c,4} = \log(t + \varepsilon). \tag{A11}$$

Appendix B. Network integration with B3LYP exchange and correlation

Our architecture follows an encoder–decoder design, where the encoder transforms local electronic descriptors into latent embeddings, decoder reconstructs scalar enhancement factors used to modulate exchange and correlation energies. The nonlocal block is inserted before the decoder stage and

refines the exchange and correlation parts of a B3LYP-like functional with dispersion correction (Becke–Johnson, D3BJ). We keep the analytic structure of the underlying hybrid functional and learn smooth multiplicative corrections in the form of enhancement factors.

We use two separate subnetworks: one for exchange and one for correlation. The exchange subnetwork produces spin-resolved enhancement factors $f_\theta^\sigma(\mathbf{r})$ for $\sigma \in \{\uparrow, \downarrow\}$, while the correlation subnetwork produces a single enhancement factor $f_\theta^c(\mathbf{r})$ constructed from spin-averaged inputs. Both subnetworks follow the same encoder–nonlocal–decoder pattern on the quadrature grid.

At each grid point i we compute a local feature vector $\mathbf{x}_i$ (the exact preprocessing is described in Appendix A). The encoder maps $\mathbf{x}_i$ to a latent embedding $\mathbf{h}_i$ using a small residual MLP with normalization,

$$\mathbf{h}_i = \mathrm{LayerNorm}\big(\mathrm{SiLU}\big(W_1\,\mathrm{SiLU}(W_0\mathbf{x}_i)\big) + \mathbf{x}_i\big)\,. \quad \text{(A12)}$$

The nonlocal block then produces a context vector $\mathbf{y}_i$, the AttF block described in Sec. 2.2), which is added to the local embedding before decoding. The decoder outputs an enhancement factor through a bounded activation,

$$f_\theta(\mathbf{r}_i) = \sigma_{\mathrm{out}}\big(W_2\,\mathrm{LayerNorm}\big(\mathrm{SiLU}\big(W_1'(\mathbf{h}_i + \mathbf{y}_i)\big)\big)\big)\,, \quad \text{(A13)}$$

$$\sigma_{\mathrm{out}}(x) = \frac{2}{1+\exp(-x/2)}. \quad \text{(A14)}$$

The choice of σ_{out} keeps f_θ close to 1 when the decoder logit is near zero, which helps keep the learned correction controlled in regions that are weakly represented by the training data.

For exchange, the enhancement factor is spin-resolved and multiplies the B3LYP exchange energy density assembled from Slater and B88 components,

$$e_x^\sigma(\mathbf{r}) = f_\theta^\sigma(\mathbf{r})\,\Big(a_{\mathrm{Slater}}\,e_{x,\mathrm{Slater}}^\sigma(\mathbf{r}) + a_{\mathrm{B88}}\,e_{x,\mathrm{B88}}^\sigma(\mathbf{r})\Big). \quad \text{(A15)}$$

Here a_{Slater} and a_{B88} are the standard B3LYP mixing coefficients for the local (Slater) and gradient-corrected (B88) exchange contributions.

For correlation, the enhancement factor multiplies the VWN (local) and LYP (gradient-corrected) correlation combination,

$$e_c(\mathbf{r}) = f_\theta^c(\mathbf{r})\,\Big(a_{\mathrm{VWN}}\,e_{c,\mathrm{VWN}}(\mathbf{r}) + a_{\mathrm{LYP}}\,e_{c,\mathrm{LYP}}(\mathbf{r})\Big), \quad \text{(A16)}$$

with the usual B3LYP coefficients a_{VWN} and a_{LYP}. In both exchange and correlation, the enhancement-factor form preserves the baseline analytic structure and introduces a learnable correction that can use the nonlocal context $\mathbf{y}_i$ to capture nonlocal effects not represented by semilocal descriptors alone, while remaining well-behaved under self-consistent iteration.

Appendix C. Benchmark results

On GMTKN, AttF tends to have its advantages on covalent thermochemistry and reaction subsets, where it frequently reduces the MAE relative to ωB97M-V by 40–80% (e.g., YBDE18, BSR36, MB16_43, DIPCS10). In contrast, several noncovalent and conformational benchmarks dominated by delicate dispersion and hydrogen-bonding effects (such as BUT14DIOL, TAUT15, SCONF, UPU23, PNICO23, HAL59) exhibit increased errors compared to ωB97M-V, sometimes by factors of 2–5. This pattern suggests that the current dispersion treatment inherited from B3LYP-D3(BJ) is not yet fully optimal for all noncovalent regimes, and that a more carefully tuned or learned dispersion correction is a promising avenue to reduce these remaining discrepancies. While some noncovalent and conformational subsets show MAEs that are larger than those of ωB97M-V by factors of 2–5, the absolute differences in these cases remain below 1 kcal/mol, with the exception of the DARC set, where the gap is larger. In other words, even where AttF is relatively worse in a multiplicative sense, the degradation typically stays within a sub-chemical-accuracy window in absolute terms.

Table A1: Largest relative MAE changes [kcal/mol] of AttF vs. ωB97M-V on GMTKN.

Subset	MAE (AttF)	MAE (ωB97M-V)	ΔMAE [%]
YBDE18	0.54	2.90	−81.4
BSR36	0.14	0.46	−69.6
C60ISO	3.71	11.85	−68.7
MB16_43	6.31	14.52	−56.5
DIPCS10	2.32	5.20	−55.4
BUT14DIOL	0.29	0.05	+480.0
DARC	3.62	0.75	+382.7
PNICO23	1.14	0.26	+338.5
TAUT15	1.16	0.33	+251.5
SCONF	0.59	0.17	+247.1