CAREB: CHEMICAL ACCURACY REACTION ENERGY DATASET AND BENCHMARK FOR NEURAL DENSITY FUNCTIONALS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce the largest openly available dataset for neural density functional theory, coupling 130 053 highly accurate reaction-energy targets to a unified reaction representation with DFT-level features for every molecular state. The dataset compiles and standardizes decades of DFA benchmarking across thermochemistry, kinetics, noncovalent interactions, barrier heights, atomization reactions, and DES15K dimer profiles, and is designed to support development of accurate DFT methods including neural functionals and ML Δ-learning on top of DFA baselines. In this work, we instantiate this setting by finetuning the Skala neural XC functional and training Δ-learning models in our benchmark, establishing strong baselines and highlighting chemical regimes where neural DFT still falls short of high-level accuracy. Additionally, a small representative subset of Gold Standard Chemistry Database was constructed by an evolutionary algorithm allowing model selection to be accelerated by orders of magnitude. By releasing all reference data, DFT-derived inputs, training code, and pretrained models, our work opens the way to fully reproducible development and systematic stress-testing of neural DFT architectures.

1 INTRODUCTION

Density functional theory (DFT) is the dominant tool for first-principles electronic-structure calculations in chemistry and materials science. Its appeal lies in a relatively favorable balance between computational cost and accuracy compared to high-level wavefunction methods, but it is still far from inexpensive, and its performance is determined largely by the choice of the exchange–correlation (XC) functional, whose exact form remains unknown. As a result, there is growing interest in using machine learning both to refine DFT accuracy beyond standard density functional approximations and to accelerate DFT-based workflows by learning surrogates for parts of the electronic-structure calculation.

In recent years, neural-network density functionals have emerged as a promising way to push DFT accuracy beyond what hand-designed DFAs can offer. DM21 and Skala are representative of the current state of the art: DM21 demonstrated that a deep neural network trained on carefully curated main-group data can substantially reduce delocalization and fractional-charge errors compared to conventional DFAs Kirkpatrick et al. (2021), while Skala showed that training a neural XC model on a very large collection of CCSD(T)-quality atomization energies leads to systematic accuracy gains across broad main-group chemical space Luise et al. (2025); Ehlert et al. (2026). At the same time, subsequent studies reveal that these neural XC functionals do not yet provide a robust, high-accuracy replacement for traditional DFAs across chemistry: DM21 and Skala fail to extrapolate reliably to transition-metal chemistry and in the case of DM21 are significantly slower than standard functionals in SCF and geometry optimizations Zhao et al. (2024); Kulaev et al. (2025). Moreover, full training datasets and training pipelines used for these neural XC functionals are not openly available, which makes their development effectively non-reproducible. For DM21, only the final functional form and a high-level description of the training objectives are public, while the underlying collection of molecular systems, reaction energies, and fractional-charge/spin constraints remains closed Kirkpatrick et al. (2021). Similarly, Skala relies on large-scale quantum-chemistry calculations that have not been released in a form suitable for reuse by the community Luise et al.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(2025); Ehlert et al. (2026). As a consequence, it is currently impossible to retrain or systematically extend these state-of-the-art neural XC functionals on the same data, and their impressive accuracies cannot be independently reproduced or stress-tested on alternative architectures.

Beyond XC functionals themselves, a broader class of neural DFT methods aims to accelerate or refine different components of electronic-structure workflows. These include models that learn corrections on top of baseline DFT energies and forces via Δ-learning Ramakrishnan et al. (2015); Smith et al. (2019), neural surrogates for Kohn–Sham Hamiltonians and densities that speed up or try to bypass the SCF procedure Li et al. (2022); Koker et al. (2024), electronic-structure-informed neural models that approximate DFT-level energies and forces at reduced cost Qiao et al. (2022), and equivariant networks that learn non-local density functionals Gao et al. (2024). While such methods can dramatically reduce computational cost or improve effective accuracy relative to a given DFA, they are often trained and evaluated within method-specific reference levels, input representations, and chemical domains, leaving open the question of how systematically they can be compared in a unified, high-level reaction-energy benchmark relevant to practical quantum chemistry.

In parallel, the data landscape for quantum-chemistry machine learning has evolved along two largely separate directions. On one side, large DFT-level datasets such as QM9 Ramakrishnan et al. (2014), nablaDFT/∇^2DFT Khrabrov et al. (2022), and related collections of drug-like molecules and materials structures (including QM7b, ANI-1, and large-scale inorganic databases such as the Materials Project and OQMD Montavon et al. (2013); Smith et al. (2017); Jain et al. (2013); Saal et al. (2013)) provide tens of thousands to millions of entries for training geometric and electronic-structure machine-learning models. These resources primarily serve the atomistic potential and neural-force-field community: they target fast surrogates for DFT-level energies and forces, but typically remain tied to a single DFA level of theory and are organized around molecular properties, conformations, Hamiltonians, densities, or structures rather than high-level reaction-energy targets.

On the other side, high-accuracy benchmarks like GMTKN55 Goerigk et al. (2017), GSCDB137 Liang & Head-Gordon (2025), and RDB7 Liu et al. (2025) were designed to assess human-crafted DFAs and provide carefully curated CCSD(T)-quality reference energies for main-group thermochemistry, kinetics, noncovalent interactions and isomerization reacctions. The MSR-ACC/TAE25 collection Ehlert et al. (2026) contributes 76 879 coupled-cluster total atomization energies at the CCSD(T)/CBS level, systematically sampling the space of small closed-shell molecules up to five heavy atoms, and has already been used internally to train neural XC functionals such as Skala Luise et al. (2025). These datasets underpin the DFT and functional-development community, but crucially expose only scalar energies and were never intended to provide the richer electronic-structure information that neural DFT models require.

For training many classes of neural DFT models, high-level reference energies alone are not sufficient. Depending on the model class, learning also requires standardized electronic-structure inputs or labels for every molecular entry, such as spin-resolved densities on a numerical grid, density matrices in a chosen basis, DFT Hamiltonians, and Coulomb and exact-exchange terms. Existing high-level benchmarks like GSCDB137 Liang & Head-Gordon (2025), RDB7 Liu et al. (2025), and MSR-ACC/TAE25 Ehlert et al. (2026) were designed to assess human-crafted DFAs and therefore ship only reference energies, without standardized DFT-level inputs. As a result, prior neural DFT works either train on private collections of densities and Hamiltonians, or reconstruct their own ad hoc training sets from scratch, and there is still no openly available resource that combines CCSD(T)-quality reaction energetics with the full set of DFT-derived quantities needed as inputs for modern neural models.

Our broader goal is to help bring DFT-like workflows closer to a high-level-accuracy regime across chemistry. We aim to make it possible to use familiar, DFT-like workflows while reaching coupled-cluster quality for reaction energies, barrier heights, and atomization energies through learned models. To this end, we integrate large, standardized DFT datasets with accurate reaction benchmarks by constructing a unified, high-level dataset that explicitly combines both sides of the learning problem: high-level *outputs* and DFT-level *inputs*. Concretely, we aggregate 130 053 reaction-energy targets from GMTKN55, GSCDB137, RDB7, MSR-ACC/TAE25, and DES15K.

For every molecular state (reactant, product, or transition state) we compute a unified set of DFT-derived quantities using a single DFT protocol, including electron densities on numerical grids,

one-electron Hamiltonians, Coulomb and exact-exchange contributions, DFA XC energies and dispersion corrections, together with basic metadata.

Our main contributions are:

- A standardized, openly released benchmark of 130 053 high-level **energy-difference** targets spanning thermochemistry, kinetics, noncovalent interactions, barrier heights, atomization-derived reactions, and non-covalent interactions. The benchmark is built from established CCSD(T)-quality sources, organized in a unified reaction representation, and paired with corresponding DFT-derived inputs, including densities, density matrices, Hamiltonians *et cetera*, for training and evaluating neural DFT models.
- A systematic empirical study of several neural DFT model variants, including non-local neural XC functionals, MACE-based adaptive hybrids, Δ-learning models, and XC enhancement-factor models and a DeePHF baseline trained anew (Appendix E). We establish strong baselines and identify the reaction classes and practically relevant chemical regimes where neural DFT still falls short of the desired high-level accuracy.

2 Dataset

2.1 Source Data and Coverage

Our dataset is built from several established high-level benchmarks with diverse reaction types and then extended with a standardized set of DFT features for each molecular state. GMTKN55 Goerigk et al. (2017) contributes 1 505 relative energies covering main-group thermochemistry, kinetics, and noncovalent interactions. GSCDB137 Liang & Head-Gordon (2025) adds 8 377 accurate energy differences for main-group and transition-metal systems, emphasizing isomerization, and other chemically challenging cases.

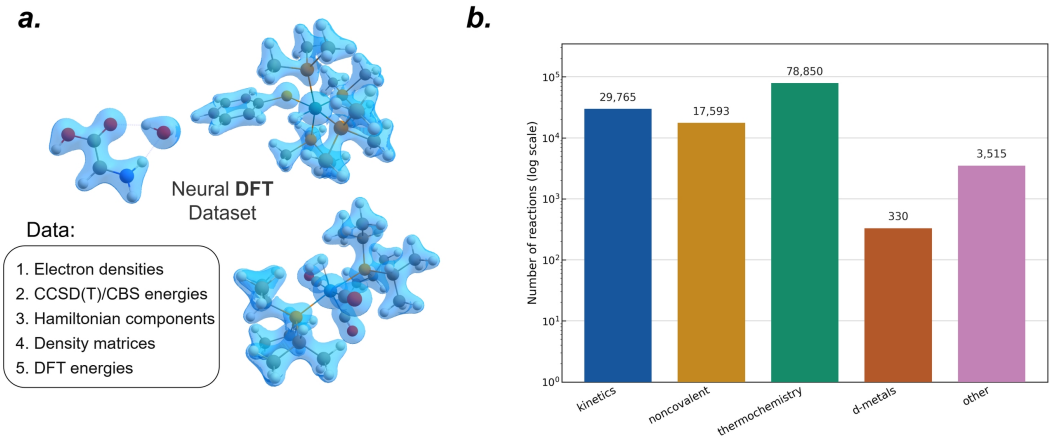

Figure 1: The list of main data types in the dataset and examples of systems from the study (a). Reaction category composition of the dataset (b).

For barrier heights and reaction profiles, we include energetic data from the RDB7 Liu et al. (2025) benchmark. RDB7 is a diverse data set of 11 926 reactions with high-accuracy reaction energies and barrier heights. or noncovalent interactions, we include DES15K Donchev et al. (2021), a chemically diverse subset of the DES370K dimer database with CCSD(T)/CBS interaction energies across compressed, near-equilibrium, and expanded geometries.

We include total atomization energies from the Microsoft Research Accurate Chemistry Collection MSR-ACC/TAE25 Ehlert et al. (2026). MSR-ACC/TAE25 provides 76 879 coupled-cluster total atomization energies at the CCSD(T)/CBS level (W1-F12-type protocol) for small molecules with up to five heavy atoms up to argon.

Table 1: Source benchmarks and number of energetic targets used in this work.

Dataset	Type	# energetic targets
GMTKN55	main-group thermochemistry/kinetics/NCI	1 505
GSCDB137	main-group and TM reactions	8 377
RDB7	barriers and reaction energies	28 641
MSR-ACC/TAE25	atomization-based reactions	76 879
DES15K	noncovalent dimer interactions	14 651
Total		130 053

2.2 Standardized Reaction Representation

For a reaction with stoichiometric coefficients c_i and state energies E_i, we define the reaction energy as

$$\Delta E = \sum_{i \in \text{products}} c_i E_i - \sum_{i \in \text{reactants}} c_i E_i. \tag{1}$$

For every molecular state appearing in these energetic data, we compute a fixed set of DFT-derived quantities and metadata using a PySCF Sun et al. (2020) unrestricted Kohn–Sham (UKS) protocol with the PBE0 functional and the def2-QZVP basis set. For each state, we store the nuclear charges and coordinates $\{Z_A, \mathbf{R}_A\}$, the total charge and spin multiplicity, the spin-resolved density matrices $P^\sigma_{\mu\nu}$, and the nuclear repulsion energy E_{nuc}. From the density matrices and the one-electron Hamiltonian $h_{\mu\nu}$ we evaluate and save the one-electron contribution $E_{1\text{e}} = \sum_{\mu\nu} P_{\mu\nu}\, h_{\nu\mu}$, the Hartree (Coulomb) energy $E_{\text{H}} = \frac{1}{2}\sum_{\mu\nu} P_{\mu\nu}\, J_{\nu\mu}[P]$, and the exact Hartree–Fock exchange energy $E_{\text{x}}^{\text{HF}} = -\frac{1}{2}\sum_{\sigma=\alpha,\beta}\sum_{\mu\nu} P^\sigma_{\mu\nu}\, K^\sigma_{\nu\mu}[P^\sigma]$, together with the corresponding sum $E_{\text{unxc}} = E_{\text{nuc}} + E_{1\text{e}} + E_{\text{H}}$. We also store the spin-resolved electron densities on a numerical integration grid, $\rho^\sigma(\mathbf{r}_k),\ \mathbf{r}_k,\ w_k, \quad k = 1, \ldots, N_{\text{grid}}$.

In addition, we store the semilocal DFA part of the B3LYP Becke (1993) exchange-correlation energy $E_{\text{xc}}^{\text{B3LYP}}[\rho] = \int \varepsilon_{\text{xc}}^{\text{B3LYP}}[\rho(\mathbf{r}), \nabla\rho(\mathbf{r})]\, \mathrm{d}\mathbf{r}$ and the D3 dispersion corrections evaluated for B3LYP functional. Finally, we record basic metadata such as the number of atoms, the number of grid points. This feature set is intended as a generic input or output representation for a broad class of neural DFT methods, including learned exchange–correlation functionals $E_{\text{xc}}[\rho]$, delta-learning corrections on top of baseline DFAs with subsequent reconstruction of relative reaction energies.

2.3 Evaluation Metrics

The main evaluation metric is the mean absolute error (MAE), defined for a set of reactions $\mathcal{R}$ as

$$\text{MAE} = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \left| \Delta E_i - \Delta E_i^{\text{ref}} \right|,$$

where ΔE_i and ΔE_i^{ref} are the predicted and reference reaction energies, respectively.

To compare performance across datasets with different intrinsic scales, we also report the normalized error ratio (NER). For a given DFA, it is defined as

$$\text{NER}_{\text{DFA}} = \frac{1}{N_d} \sum_d \frac{\text{MAE}_{d,\text{DFA}}}{\sigma_d},$$

where $\mathcal{D}_{\text{active}}$ is the set of active datasets, $\text{MAE}_{d,\text{DFA}}$ is the mean absolute error for dataset d and the given DFA, and σ_d is the corresponding standard error.

2.4 Validation Subset Construction

We construct the validation subset so that it is maximally small while still representative of the full benchmark. The goal is similar to that of GMTKN-style slim benchmarks: the subset should preserve the overall error pattern and relative difficulty of the original dataset, rather than being merely a random reduced sample.

For GSCDB137, we first remove reactions with large density grids ($N_{\text{points}} > 8 \cdot 10^3$) in order to keep the benchmark computationally manageable. The remaining reactions form the candidate pool $\mathcal{C}$. For each DFA, we evaluate the MAE over the full pool and over a candidate subset $\mathcal{S} \subset \mathcal{C}$, denoted by $\mathrm{MAE}^{\text{full}}_{\text{DFA}}$ and $\mathrm{MAE}^{\mathcal{S}}_{\text{DFA}}$, respectively.

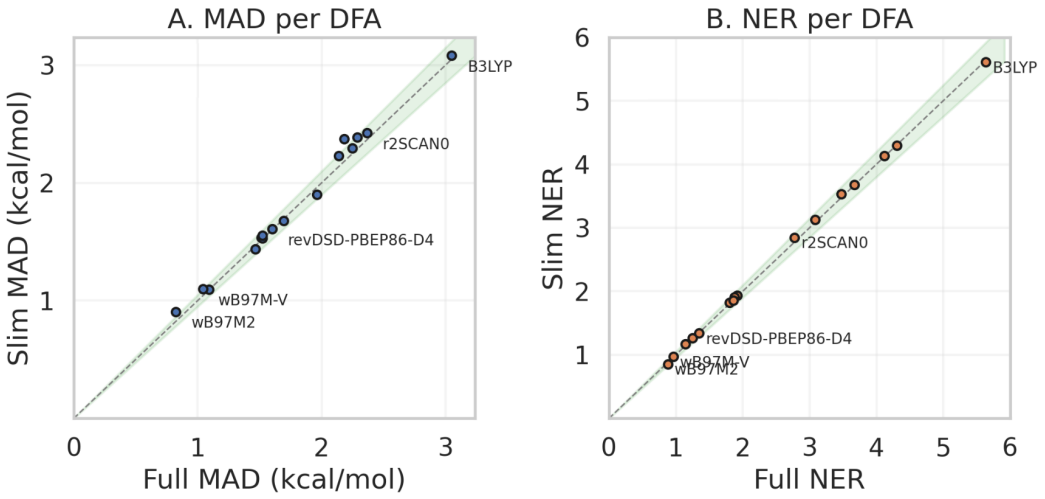

Figure 2: Parity plots comparing the full and slim benchmark values for MAE and NER across DFAs. The diagonal line indicates perfect agreement between the full benchmark and the validation subset.

To quantify how well the subset reproduces the full benchmark, we define

$$\Delta_{\mathrm{MAD}}(\mathcal{S},\mathcal{C}) = \frac{1}{|\mathcal{D}|} \sum_{\mathrm{DFA}\in\mathcal{D}} \frac{\left|\mathrm{MAE}^{\mathcal{S}}_{\mathrm{DFA}} - \mathrm{MAE}^{\text{full}}_{\mathrm{DFA}}\right|}{\mathrm{MAE}^{\text{full}}_{\mathrm{DFA}}},$$

$$\Delta_{\mathrm{NER}}(\mathcal{S},\mathcal{C}) = \frac{1}{|\mathcal{D}|} \sum_{\mathrm{DFA}\in\mathcal{D}} \frac{\left|\mathrm{NER}^{\mathcal{S}}_{\mathrm{DFA}} - \mathrm{NER}^{\text{full}}_{\mathrm{DFA}}\right|}{\mathrm{NER}^{\text{full}}_{\mathrm{DFA}}},$$

$$\Delta_{\max}(\mathcal{S},\mathcal{C}) = \max_{\mathrm{DFA}\in\mathcal{D}} \frac{\left|\mathrm{NER}^{\mathcal{S}}_{\mathrm{DFA}} - \mathrm{NER}^{\text{full}}_{\mathrm{DFA}}\right|}{\mathrm{NER}^{\text{full}}_{\mathrm{DFA}}};$$

The subset is then selected so that these error profiles remain as close as possible. For this purpose, we use the normalized error ratio to ensure that the slim subset matches the difficulty of the full benchmark, not only its average error magnitude. The selection objective combines these two criteria together with a penalty for the worst-case deviation:

$$F(\mathcal{S}) = \Delta_{\mathrm{MAD}}(\mathcal{S},\mathcal{C}) + \Delta_{\mathrm{NER}}(\mathcal{S},\mathcal{C}) + 5\,\Delta_{\max}(\mathcal{S},\mathcal{C}).$$

The final validation subset is obtained by minimizing this objective with the evolutionary algorithm described in Appendix B.

For MSR, DES15K, and RDB7, the procedure is much simpler. These benchmarks are more homogeneous, so a random structural sample already provides a good approximation of the error distribution. In other words, a random hold-out is usually sufficient, and there is no strong need for a separately optimized slim subset.

3 Neural DFT Training

3.1 Training Protocol

We evaluate two independent training settings. The first setting covers broad main-group chemistry using a mixed reaction pool from GSCDB, MSR, and RDB7. The second setting is a dedicated

transition-metal (TM) experiment and is evaluated separately; it is not a continuation of the main-group run. All models were trained on a single workstation equipped with an NVIDIA RTX 5090 GPU, an Intel Xeon Silver 4510 CPU, and 4×64GB DDR5 ECC memory.

In both settings, we use the pretrained Microsoft Skala functional as the trainable exchange–correlation model. Reaction energies are represented as

$$\Delta E = \sum_{i \in \text{products}} c_i \, E_i - \sum_{i \in \text{reactants}} c_i \, E_i,$$

where E_i denotes the total energy of molecular state i and c_i is the corresponding stoichiometric coefficient. For each molecular state, the total energy is written as

$$E_i = E^{(i)}_{\text{nuc+1e+H}} + E^{(i)}_{\text{xc,Skala}} + E^{(i)}_{\text{disp}},$$

with $E^{(i)}_{\text{nuc+1e+H}}$ being the sum of nuclear repulsion, one-electron, and Hartree (Coulomb) terms from the underlying Kohn–Sham energy decomposition, $E^{(i)}_{\text{xc,Skala}}$ the neural XC contribution, and $E^{(i)}_{\text{disp}}$ the D_3(BJ) dispersion correction (B3LYP parameterization). This same energy expression is used in training and evaluation.

The optimization objective is a baseline-normalized weighted mean-squared error in kcal/mol:

$$\mathcal{L} = \frac{1}{N} \sum_{r=1}^{N} \left(\frac{\Delta E^{(r)}_{\text{pred}} - \Delta E^{(r)}_{\text{ref}}}{b_r + \varepsilon} \right)^2, \qquad \varepsilon = 10^{-8},$$

where N is the number of reactions in a batch and b_r is the baseline MAE assigned to reaction r. For GSCDB, b_r is subset-dependent; for MSR and RDB7, benchmark-level global baseline constants are used. Optimizer settings for all trainable models are summarized in Appendix F.

Main-group training. We first form GSCDB train part by applying the molecular-size filter ($N_{atoms} < 60$). The compact GSCDB slim validation subset is then removed from this pool, and the remaining reactions define a fixed GSCDB train dataset. Then we construct a balanced training pool from the remaining reactions. The goal is to preserve coverage of the main reaction types while reducing the dominance of very large categories through adaptive subsampling.

More specifically, if n_c is the number of available reactions in category c, we define a retained fraction f_c by interpolating between $f_{\min}$=0.1 and $f_{\max}$=0.8 on a logarithmic scale,

$$f_c = f_{\max} - t_c \, (f_{\max} - f_{\min}), \qquad t_c = \frac{\log n_c - \log n_{\min}}{\log n_{\max} - \log n_{\min}},$$

and then keep

$$k_c = \max(k_{\min}, \text{round}(f_c n_c))$$

reactions from each category. This compresses large categories more aggressively while retaining a higher fraction of small and rare categories.

MSR and RDB7 are included to increase chemical coverage, with their training subsets chosen to have the same size as the GSCDB training subset. Validation slices for MSR and RDB7 are fixed to 100 reactions each, yielding a balanced three-source training mixture in which no single benchmark dominates optimization.

Table 2: Main-group split sizes used in training.

Source	Train reactions	Validation reactions
GSCDB	2 428	300
MSR	2 428	100
RDB7	2 428	100
Total	7 284	500

Targeted transition-metal training. In the transition-metal (TM) experiment, we perform a small-scale targeted finetuning run. Reactions are constructed from the dedicated TM reaction set and split by stratified partitioning within each GSCDB subset containing d-metals. This results in 221 reactions in the training split, 28 reactions in the validation subset (used for early stopping) and 28 reactions in the test split. The same weighted MSE objective is used as in the main-group setup.

4 RESULTS

Figure 3 summarizes the outcome of the main-group training setup on the GSCDB holdout set using a mirrored horizontal bar plot, where the left bars show the normalized error ratio (NER; mean over subsets) and the right bars show the mean absolute error (MAE) in kcal/mol. Both metrics are evaluated on the same pool of reactions, so shorter paired bars indicate simultaneous improvements in relative and absolute accuracy. In this main-group regime, the two MACE-based models Batatia et al. (2022) exhibit markedly different behavior. The Δ-MACE model, which learns an additive correction on top of the B3LYP-based energy, improves over the uncorrected B3LYP-D3(BJ) baseline, although it remains less accurate than the strongest conventional DFAs and the fine-tuned Skala functional. In contrast, the α-MACE adaptive-hybrid model, which predicts a system-dependent exact-exchange mixing coefficient, does not improve upon the B3LYP baseline in this experiment. Together, these results illustrate that CAREB enables direct comparison of distinct neural-DFT strategies under a common reaction-energy benchmark.

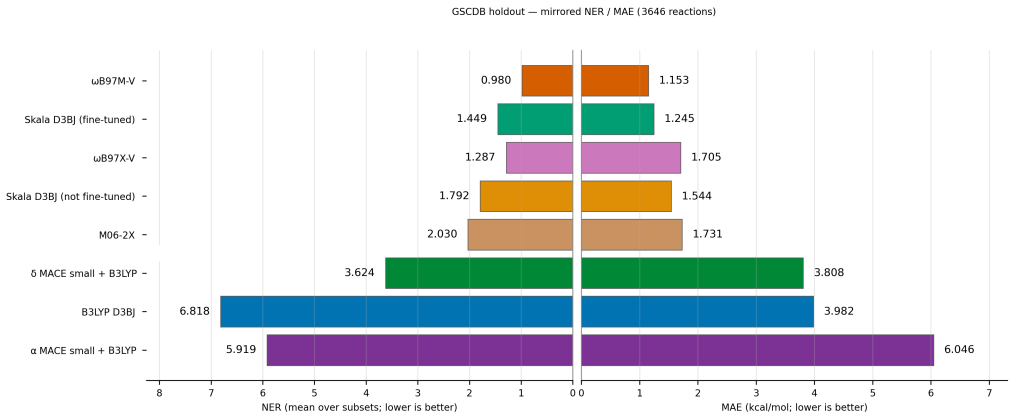

Figure 3: Mirrored horizontal bar plot of NER (left) and MAE (right) on the GSCDB holdout set ($n \approx 3{,}600$ reactions) obtained after main-group finetuning of Skala and training of the MACE-based delta learning C and adaptive hybrid D baselines corrections to B3LYP.

Although the main training and holdout evaluations above are performed post-SCF on frozen PBE0/def2-QZVP densities, we additionally ran fine-tuned Skala self-consistently on the GSCDB slim validation subset (300 reactions). All required molecular states converged in the SCF loop. Relative to the post-SCF evaluation on the same PBE0 densities (MAE 1.84 kcal/mol, NER 1.29), the self-consistent Skala/def2-QZVP energies give MAE 1.00 kcal/mol and NER 1.41.

Figure 4 summarizes the behavior of the Skala model in the targeted TM finetuning setup described in Section 3.1. The left panel shows the evolution of the weighted mean-squared error on the validation split during finetuning, indicating stable optimization and a consistent reduction in loss over epochs. The right panel reports the distribution of signed reaction-energy errors on the held-out TM test set for baseline DFAs, the untuned Skala functional, and the finetuned Skala model.

Overall, finetuning on the compact TM-focused subset narrows the error distribution of Skala relative to its untuned checkpoint giving a roughly unbiased error profile around zero. Strong hybrid DFAs (e.g., ωB97X-V, ωB97M-V, and B3LYP-D3(BJ)) nevertheless remain more accurate on this small held-out TM test set (28 reactions), so the experiment is best read as a proof-of-concept for targeted domain adaptation rather than as evidence that fine-tuned Skala surpasses standard hybrids.

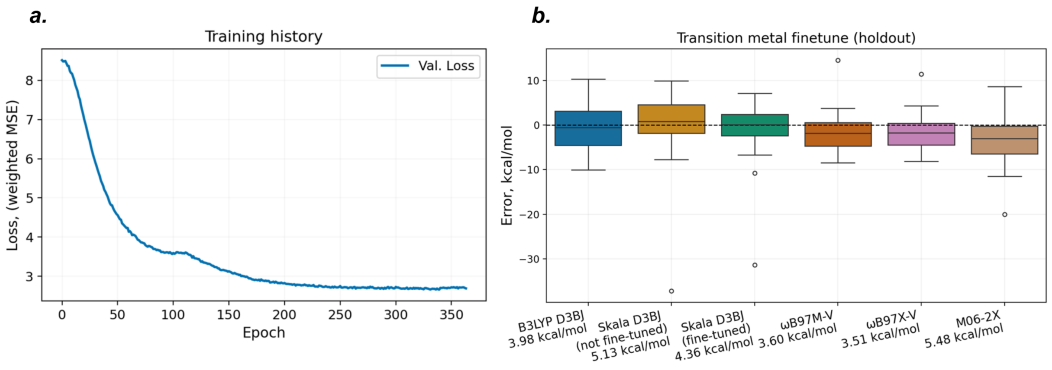

Figure 4: Performance of the targeted transition-metal finetuning experiment. (a) Validation loss over training epochs for the Skala-based neural XC model, showing stable convergence on the TM-focused reaction pool. (b) Distribution of signed reaction-energy errors on the hold-out TM set for conventional DFAs, the untuned Skala baseline, and the finetuned Skala model; labels report per-method MAE in kcal/mol.

5 Discussion

In contrast to prior neural density functionals such as DM21 and Skala, which are built on proprietary training collections, our benchmark is explicitly designed as an open, end-to-end resource for developing and stress-testing neural DFT models. While DM21 publicly exposes only the final functional form and high-level training objectives, the reaction energies with corresponding electron densities and fractional-charge/spin data remains closed, making it impossible to retrain or systematically extend the model on the original data Kirkpatrick et al. (2021). Similarly, Skala relies on large-scale CCSD(T)-quality atomization energies that have not been released in a form suitable for reuse, so alternative architectures cannot be evaluated under identical data and protocol conditions Luise et al. (2025); Ehlert et al. (2026). By aggregating 130 053 CCSD(T)-quality reaction-energy targets from GMTKN55, GSCDB137, RDB7, MSR-ACC/TAE25, and DES15K and pairing them with standardized PBE0/def2-QZVP DFT-level inputs for every molecular state, we for the first time provide an openly available training and evaluation layer that spans thermochemistry, kinetics, non-covalent interactions, barrier heights and atomization reactions within a unified representation. This makes it possible, for the first time, to compare competing neural DFT architectures head-to-head on the same high-accuracy reaction data and low-level electronic-structure inputs.

Beyond the dataset itself, our experiments show that the benchmark can support and directly compare several distinct families of neural DFT models under a common training and evaluation protocol. In addition to Skala-style non-local neural XC, we train a Δ-learning MACE model that predicts high-level reaction energies on top of a B3LYP-consistent baseline, as well as an adaptive-hybrid variant where MACE outputs a system-dependent mixing coefficient between Hartree–Fock exchange and DFA exchange–correlation. This juxtaposition shows that qualitatively different modeling strategies—reshaping the XC functional, learning an explicit energy correction, or adjusting the hybrid mixing—can all reach competitive accuracy when given the same CCSD(T)-quality reaction targets and standardized DFT-level inputs Brémond et al. (2026); Khan et al. (2025). At the same time, fine-tuning Skala on a compact transition-metal split improves over the untuned checkpoint in a chemically distinct regime, but does not outperform strong hybrid DFAs on the held-out TM test reactions; given the small TM sample, we treat this as a targeted stress test rather than broad coverage of transition-metal chemistry. Together, these results indicate that our benchmark enables a fair and systematic comparison of neural XC, Δ-learning, and adaptive-hybrid approaches, including on a transition-metal holdout. We additionally construct a GSCDB-based slim validation subset using an evolutionary optimization procedure that minimizes distortions in DFA-wise MAE and NER profiles relative to the full pool, improving over simpler random or stratified sampling strategies. In combination, the open release of rich electronic-structure inputs, the inclusion of transition-metal reactions, and the carefully engineered validation protocol establish our benchmark not only as a

data source, but as the first comprehensive and reproducible testbed for next-generation neural DFT frameworks.

6 LIMITATIONS

Our benchmark also reflects several inherent limitations of current quantum-chemistry datasets. First, as in most high-accuracy electronic-structure benchmarks, the data are dominated by gas-phase small and medium-sized molecules assembled from established reaction-energy and atomization-energy benchmarks, while condensed-phase environments, very large systems and other properties (geometries, forces, spectroscopic observables) remain underrepresented. Goerigk et al. (2017); Khrabrov et al. (2022)

Second, all DFT-level features are generated with a single, fixed UKS protocol (PBE0/def2-QZVP with standard numerical settings), which ensures internal consistency but does not capture the full diversity of functionals, basis sets, and convergence choices used in practical workflows. Third, although the reference values are based on CCSD(T)-quality protocols, they can fall short, for example, in strongly correlated or more multi-reference regimes Leszczyk et al. (2022).

7 CONCLUSION

We introduced the first open benchmark that couples 130 053 CCSD(T)-quality reaction energies to a unified reaction representation with standardized PBE0/def2-QZVP DFT-level inputs for every molecular state, spanning both main-group and transition-metal chemistry in a single, reusable resource. On top of this benchmark we, for the first time, train and evaluate side-by-side a non-local neural XC functional (Skala), a MACE-based Δ-learning model, and a MACE-based adaptive hybrid under exactly the same data and protocol, establishing directly comparable high-accuracy baselines for three qualitatively different neural DFT paradigms Luise et al. (2025); Ramakrishnan et al. (2015); Khan et al. (2025); Brémond et al. (2026); Smith et al. (2019). Targeted fine-tuning on a compact transition-metal split further shows that Skala can be improved in a chemically distinct regime via domain-focused retraining, while remaining behind strong hybrid baselines on the held-out TM test set. By releasing the full dataset, electronic-structure features, code, and pretrained models, we aim to establish this benchmark as a standard, fully reproducible testbed for the next generation of neural DFT methods and their stress-testing across realistic reaction chemistry.

REPRODUCIBILITY STATEMENT

The CAREB dataset is publicly available via an anonymous link at `https://dataset.chemdft.workers.dev/`. The release contains the reaction-level reference targets, molecular-state metadata, and standardized DFT-derived quantities described in Section 2.3. Details of the electronic-structure protocol and reaction representation are given in Section 2.3. Slim GSCDB validation subset is described in Section 2.4 and defined in the CAREB dataset release. Training protocols and model-specific details are provided in Section 3.1 and the appendices. Trained checkpoints are released with the dataset.

AI USE STATEMENT

We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, support qualitative and thematic data analysis, interpret results and generation of synthetic data sets, assist with translation, cleaning and reformation of dataset. Writing proofs is not applicable to this work. Additionally, we used generative AI tools to propose keywords for a research paper. We have reviewed all AI-assisted work: we checked LLM-generated language corrections. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gabor Csanyi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems 35*, NeurIPS 2022, pp. 11423–11436. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022. doi: 10.52202/068431-0830. URL http://dx.doi.org/10.52202/068431-0830.

Axel D Becke. Density-functional thermochemistry. III. the role of exact exchange. *J. Chem. Phys.*, 98(7):5648–5652, April 1993.

E. Brémond, A. J. Pérez-Jiménez, C. Adamo, and J. C. Sancho-García. Contemporary dft: learning from traditional and recent trends for the development and assessment of accurate exchange–correlation functionals. *Physical Chemistry Chemical Physics*, 28(6):3779–3796, 2026. ISSN 1463-9084. doi: 10.1039/d5cp03373j. URL http://dx.doi.org/10.1039/d5cp03373j.

Alexander G. Donchev, Andrew G. Taube, Elizabeth Decolvenaere, Cory Hargus, Robert T. McGibbon, Ka-Hei Law, Brent A. Gregersen, Je-Luen Li, Kim Palmo, Karthik Siva, Michael Bergdorf, John L. Klepeis, and David E. Shaw. Quantum chemical benchmark databases of gold-standard dimer interaction energies. *Scientific Data*, 8(1), February 2021. ISSN 2052-4463. doi: 10.1038/s41597-021-00833-x. URL http://dx.doi.org/10.1038/s41597-021-00833-x.

Sebastian Ehlert, Jan Hermann, Thijs Vogels, Victor Garcia Satorras, Stephanie Lanius, Marwin Segler, Klaas J. H. Giesbertz, Derk P. Kooi, Kenji Takeda, Chin-Wei Huang, Giulia Luise, Rianne van den Berg, Paola Gori-Giorgi, and Amir Karton. Accurate chemistry collection: Coupled cluster atomization energies for broad chemical space. *Scientific Data*, 13(1), April 2026. ISSN 2052-4463. doi: 10.1038/s41597-026-07200-8. URL http://dx.doi.org/10.1038/s41597-026-07200-8.

Nicholas Gao, Eike Eberhard, and Stephan Günnemann. Learning equivariant non-local electron density functionals, 2024. URL https://arxiv.org/abs/2410.07972.

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. *Phys. Chem. Chem. Phys.*, 19(48):32184–32215, December 2017.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.*, 1(1):011002, July 2013.

Danish Khan, Alastair J. A. Price, Bing Huang, Maximilian L. Ach, and O. Anatole von Lilienfeld. Adapting hybrid density functionals with machine learning. *Science Advances*, 11(5), January 2025. ISSN 2375-2548. doi: 10.1126/sciadv.adt7769. URL http://dx.doi.org/10.1126/sciadv.adt7769.

Kuzma Khrabrov, Ilya Shenbin, Alexander Ryabov, Artem Tsypin, Alexander Telepov, Anton Alekseev, Alexander Grishin, Pavel Strashnov, Petr Zhilyaev, Sergey Nikolenko, and Artur Kadurin. nablaDFT: Large-Scale conformational energy and hamiltonian prediction benchmark and dataset. *Phys. Chem. Chem. Phys.*, 24(42):25853–25863, November 2022.

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 1095-9203. doi: 10.1126/science.abj6511. URL http://dx.doi.org/10.1126/science.abj6511.

Teddy Koker, Keegan Quigley, Eric Taw, Kevin Tibbetts, and Lin Li. Higher-order equivariant neural networks for charge density prediction in materials. *Npj Comput. Mater.*, 10(1), July 2024.

Kirill Kulaev, Alexander Ryabov, Michael Medvedev, Evgeny Burnaev, and Vladimir Vanovskiy. On the practical applicability of modern dft functionals for chemical computations. case study of dm21 applicability for geometry optimization, 2025. URL https://arxiv.org/abs/2501.12149.

Aleksandra Leszczyk, Mihály Máté, Örs Legeza, and Katharina Boguslawski. Assessing the accuracy of tailored coupled cluster methods corrected by electronic wave functions of polynomial cost. *J. Chem. Theory Comput.*, 18(1):96–117, January 2022.

He Li, Zun Wang, Nianlong Zou, Meng Ye, Runzhang Xu, Xiaoxun Gong, Wenhui Duan, and Yong Xu. Deep-learning density functional theory hamiltonian for efficient ab initio electronic-structure calculation. *Nat. Comput. Sci.*, 2(6):367–377, June 2022.

Jiashu Liang and Martin Head-Gordon. Gold-standard chemical database 137 (GSCDB137): A diverse set of accurate energy differences for assessing and developing density functionals. *J. Chem. Theory Comput.*, 21(24):12601–12621, December 2025.

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. ISSN 1463-9084. doi: 10.1039/d5cp01181g. URL http://dx.doi.org/10.1039/D5CP01181G.

Giulia Luise, Chin-Wei Huang, Thijs Vogels, Derk P. Kooi, Sebastian Ehlert, Stephanie Lanius, Klaas J. H. Giesbertz, Amir Karton, Deniz Gunceler, Stefano Battaglia, Gregor N. C. Simm, P. Bernát Szabó, 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. URL https://arxiv.org/abs/2506.14665.

Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.*, 15(9):095003, September 2013.

Zhuoran Qiao, Anders S Christensen, Matthew Welborn, Frederick R Manby, Anima Anandkumar, and Thomas F Miller, 3rd. Informing geometric deep learning with electronic interactions to accelerate quantum chemistry. *Proc. Natl. Acad. Sci. U. S. A.*, 119(31):e2205221119, August 2022.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data*, 1(1):140022, August 2014.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Big data meets quantum chemistry approximations: The Δ-machine learning approach. *J. Chem. Theory Comput.*, 11(5):2087–2096, May 2015.

James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and C Wolverton. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *JOM (1989)*, 65(11):1501–1509, November 2013.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost, 2018. URL https://arxiv.org/abs/1804.04235.

Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. ANI-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data*, 4(1):170193, December 2017.

Justin S Smith, Benjamin T Nebgen, Roman Zubatyuk, Nicholas Lubbers, Christian Devereux, Kipton Barros, Sergei Tretiak, Olexandr Isayev, and Adrian E Roitberg. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.*, 10(1):2903, July 2019.

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. *J. Chem. Phys.*, 153(2):024109, July 2020.

Jin Xiao, Yingfeng Zhang, Bowen Li, Shuwen Zhang, Ya Gao, Wei Chen, Han Wang, John Z. H. Zhang, and Tong Zhu. A deep learning-augmented density functional framework for reaction modeling with chemical accuracy. *JACS Au*, 5(8):3892–3903, 2025. ISSN 2691-3704. doi: 10.1021/jacsau.5c00541. URL http://dx.doi.org/10.1021/jacsau.5c00541.

Heng Zhao, Tim Gould, and Stefan Vuckovic. Deep mind 21 functional does not extrapolate to transition metal chemistry. *Phys. Chem. Chem. Phys.*, 26(16):12289–12298, April 2024.

A SAMPLING STRATEGY ABLATION

Table 3: Ablation of validation-subset sampling strategies on GSCDB. Lower values indicate better agreement with full-pool error profiles.

Strategy	Δ_{MAD}	Δ_{NER}	$\Delta_{\max}$	$F(S)$
Evolutionary optimized	0.4036	**0.0399**	**0.1225**	**1.0562**
Random uniform	0.3476	0.0907	0.2543	1.7098
Stratified by category	0.3714	0.1032	0.3018	1.9836
Stratified by subset	**0.2196**	0.1026	0.3512	2.0782
Stratified by PBE0 deviation	0.2893	0.1074	0.3888	2.3405

Table 3 compares simple subset-construction baselines against the optimized slim reference in terms of profile fidelity to the filtered full pool. Here, Δ_{MAD} quantifies average relative distortion of DFA-wise MAE profiles, Δ_{NER} quantifies average relative distortion of DFA-wise NER profiles, and $\Delta_{\max}$ captures the worst-case relative NER deviation across DFAs.

The evolutionary optimized reference subset provides the strongest overall robustness, achieving the lowest Δ_{NER}, $\Delta_{\max}$, and $F(S)$. Among simple baselines, random uniform sampling gives the best average composite performance, whereas stratified alternatives can improve specific average terms (e.g., Δ_{MAD} for dataset-stratified sampling) but tend to degrade worst-case behavior, resulting in higher $F(S)$.

B EVOLUTIONARY SELECTION OF THE SLIM VALIDATION SUBSET

The slim GSCDB validation subset is obtained by minimizing $F(\mathcal{S})$ from Section 2.4 over subsets of fixed size $|\mathcal{S}| = 300$. Quotas per reaction category are set by proportional allocation so that category composition mirrors the pool.

We optimize $F(\mathcal{S})$ with a multi-start genetic algorithm that preserves category quotas under mutation and crossover. Independently seeded populations of size 60 are evolved for up to 400 generations with elitism (10 individuals), mutation rate 0.12, and early stopping after 100 unimproved generations. The best individual across 12 runs is retained as the slim validation set.

C APPENDIX. Δ-LEARNING ON A B3LYP BASELINE (TRAINABLE MACE)

We train the Materials Project MACE small model end-to-end on the same reaction-level objective as in the section "Training Protocol". For each molecular state i, MACE produces a scalar $E^{(i)}_{\mathrm{mace}}$; the

total energy E_i is a fixed B3LYP-consistent baseline assembled from stored fields plus a correction $E^{(i)}_{\mathrm{mace}}$.

The baseline uses only quantities released with each state:

$$E^{(i)}_{\mathrm{base}} = E^{(i)}_{\mathrm{nuc+1e+H}} + E^{(i)}_{\mathrm{xc,\,B3LYP}} + D^{(i)}_{3,\,\mathrm{B3LYP}} + \alpha\, E^{(i)}_{xHF}, \tag{2}$$

(the factor $\alpha = 0.20$ matches B3LYP exact exchange mixture). The predicted molecular energy is

$$E_i = E^{(i)}_{\mathrm{base}} + E^{(i)}_{\mathrm{mace}} \tag{3}$$

Optimization uses the same main-group dataset (GSCDB, MSR, RDB7), splits, and weighted mean-squared loss in kcal/mol (Section 3.1).

D Appendix. Adaptive hybrid

Analogously, we train MACE small end-to-end and obtain a per-molecule mixing weight $\alpha^{(i)} \in (0, 1)$ from MACE's scalar output $E^{(i)}_{\mathrm{mace}}$, using the same reaction objective and data protocol as Section 3.1. The MACE output is post-processed with a sigmoid function and interpreted as a system-dependent hybrid mixing coefficient, in close analogy to the adaptive hybrid functionals of Khan et al. Khan et al. (2025), which learn optimal exact-exchange admixtures with machine learning to reduce delocalization errors.

$$E^{(i)}_{\mathrm{xc,\,mix}} = \alpha^{(i)}\, E^{\mathrm{HF},(i)}_{x} + \left(1 - \alpha^{(i)}\right) E^{\mathrm{B3},(i)}_{x} + E^{(i)}_{c,\,\mathrm{B3LYP}} + D^{(i)}_{3,\,\mathrm{B3LYP}}, \tag{4}$$

$$E_i = E^{(i)}_{\mathrm{nuc+1e+H}} + E^{(i)}_{\mathrm{xc,\,mix}}. \tag{5}$$

E Learned exchange–correlation enhancement factors

Alongside fine-tuning a pretrained Skala XC checkpoint and training MACE-based Δ-learning / adaptive-hybrid corrections to B3LYP, we also train neural XC models *anew* (random initialization; not continued from a pretrained XC checkpoint). Rather than learning an additive energy correction, the primary construction learns a multiplicative exchange–correlation enhancement factor applied to a fixed r^2SCAN XC density,

$$\varepsilon_{\mathrm{xc}}(\mathbf{r}) = s_\theta(\mathbf{r})\, \left(\varepsilon_x + \varepsilon_c\right)_{r^2\mathrm{SCAN}}(\mathbf{r}), \tag{6}$$

so that the molecular energy becomes

$$E_i = E^{(i)}_{\mathrm{nuc+1e+H}} + \int s_\theta(\mathbf{r})\, \left(\varepsilon_x + \varepsilon_c\right)_{r^2\mathrm{SCAN}}(\mathbf{r})\, \mathrm{d}\mathbf{r} + D^{(i)}_{3,\,r^2\mathrm{SCAN}}. \tag{7}$$

We train three models from random initialization on the same mixed reaction pack (detailed below): (i) Skala neural XC producing s_θ ($\approx 2.75 \times 10^5$ parameters; nonlocal Skala block with embedding dimension 256 and nonlocal feature width 16, concat merge); (ii) a compact local MLP producing an analogous XC enhancement ($\approx 2.65 \times 10^5$ parameters; same embedding width 256, but with the nonlocal block disabled); and (iii) a DeePHF model Xiao et al. (2025) ($\approx 2.36 \times 10^5$ parameters; six hidden layers of width 206) adds an orbital-descriptor correction $\Delta E = \sum_I F_{\mathrm{NN}}(\mathrm{desc}_I)$ from eigenvalues of local projected density matrices, so that $E_i = E^{(i)}_{\mathrm{nuc+1e+H}} + E^{(i)}_{\mathrm{xc},\,r^2\mathrm{SCAN}} + D^{(i)}_{3,\,r^2\mathrm{SCAN}} + \Delta E^{(i)}$.

Training uses a mixed reaction dataset drawn from GSCDB, MSR-ACC/TAE25, RDB7, DES15K with a GSCDB-slim validation split. The GSCDB training count of 4 353 is not the full GSCDB137 release (8 377): it is the pool that remains after excluding GMTKN55-overlapping reactions, excluding the slim validation subset of Section 2.4, applying a molecular-size cutoff. In the released ablation configuration, the pack then adds 1 000 MSR, 1 000 RDB7, 1 000 DES15K for training (7 353 reactions total) and use Adam optimizer (learning rate 1.5×10^{-4}, no warmup, cosine schedule over 3.2×10^5 steps, RMS grad clip 10^{-4}, batch size 8) and loss function $\mathcal{L} = (E_{\mathrm{pred}} - E_{\mathrm{ref}})^2/(10^{-4} + |E_{\mathrm{ref}}|)$ (with energies in Hartree). On the GSCDB validation split, representative checkpoints reach MAE ≈ 0.49 kcal/mol (Skala), ≈ 0.73 kcal/mol (MLP), and ≈ 0.83 kcal/mol (DeePHF), showing that CAREB supports training XC enhancement and DeePHF-style models without a pretrained XC warm start. Figure 5 overlays the validation MAE trajectories for the three models on this pack.

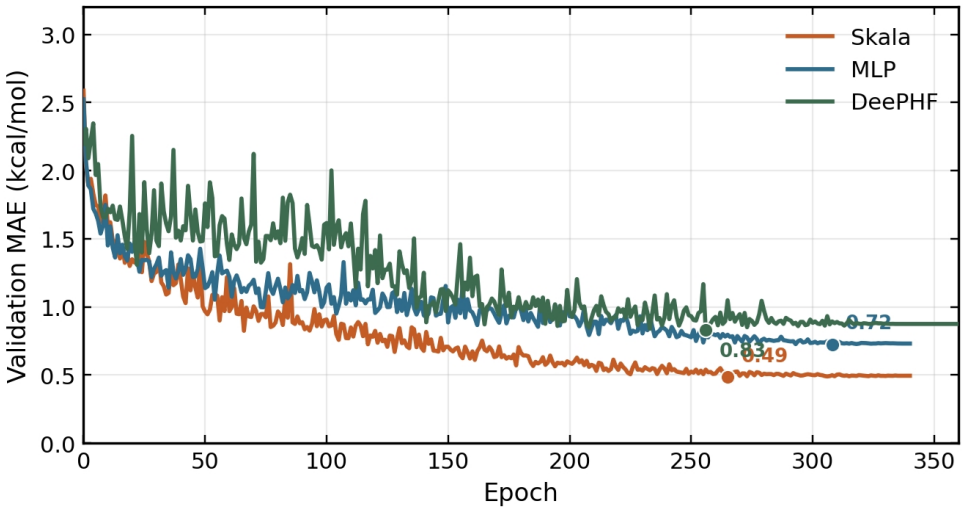

Figure 5: GSCDB validation MAE during training on the shared CAREB reaction pack (GSCDB / MSR / RDB7 / DES15K / water) for Skala-style and MLP XC enhancement factors and a DeePHF CorrNet baseline. Markers indicate the best validation checkpoint of each run.

F Optimization hyperparameters

Skala fine-tuning as well as the MACE-based Δ-learning and adaptive-hybrid baselines—use Hugging Face Adafactor Shazeer & Stern (2018) (`transformers.Adafactor`) with a cosine learning-rate schedule. The learning rate 5×10^{-5}, training proceeds for up to 5×10^{5} optimization steps, and early stopping is performed on the validation split.