Science
AAAS

Supplementary Materials for

Pushing the frontiers of density functionals by solving the fractional electron problem

James Kirkpatrick *et al*.

Corresponding authors: James Kirkpatrick, kirkpatrick@google.com; Aron J. Cohen, aroncohen@google.com

Science **374**, 1385 (2021)
DOI: 10.1126/science.abj6511

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

Materials and Methods

S1 Functional architecture

The exchange-correlation functional consists of a MLP $\mathbf{f}_\theta$ with a single set of weights, θ, that accepts a vector of features, $\mathbf{x}(\mathbf{r})$, at each grid point $\mathbf{r}$ and returns a vector of 3 enhancement factors. The energy is then computed as the integral

$$E_{\mathrm{xc}}^{\mathrm{MLP}}[\rho] = \int \mathbf{f}_\theta(\mathbf{x}(\mathbf{r})) \cdot \begin{bmatrix} e_{\mathrm{x}}^{\mathrm{LDA}}(\mathbf{r}) \\ e^{\mathrm{HF}}(\mathbf{r}) \\ e^{\omega\mathrm{HF}}(\mathbf{r}) \end{bmatrix} \mathrm{d}^3\mathbf{r}. \tag{S1}$$

Note that none of the input features are sensitive to long-range dispersion forces, and thus we simply augment the MLP energy with an empirical dispersion correction (see also Table S7).

We chose to use a D3 correction with Becke-Johnson damping ($\mathrm{D3_{BJ}}$) (*48*) with coefficients chosen to match those of B3LYP, and simply add the correction during and after training.

$$E_{\mathrm{xc}}^{\mathrm{DM21}} = E_{\mathrm{xc}}^{\mathrm{MLP}} + E_{\mathrm{D3(BJ)}}. \tag{S2}$$

In the following sections we provide additional details for the computation of the input features $\mathbf{x}(\mathbf{r})$, the MLP architecture and the objective functions used for training.

S1.1 Feature generation

The 11 features, $\mathbf{x}(\mathbf{r})$, supplied at each grid point are computed from a (spin indexed $\sigma \in \{\uparrow, \downarrow\}$) density matrix Γ_{ab}^{σ} and basis set ψ_a as follows (using Einstein summation throughout):

- The density $\rho^{\sigma}(\mathbf{r}) = \Gamma_{ab}^{\sigma}\psi_a(\mathbf{r})\psi_b(\mathbf{r})$ in each spin channel.
- The square norm of the gradient of the density in each channel and of the total density ($|\nabla\rho^{\uparrow}|^2, |\nabla\rho^{\downarrow}|^2, |\nabla(\rho^{\uparrow} + \rho^{\downarrow})|^2$).
- The kinetic energy density $\tau^{\sigma}(\mathbf{r}) = \frac{1}{2}\Gamma_{ab}^{\sigma}[\nabla\psi_a(\mathbf{r}) \cdot \nabla\psi_b(\mathbf{r})]$ in each channel.
- The local HF features

$$e_{\sigma}^{\omega\mathrm{HF}}(\mathbf{r}) = -\frac{\Gamma_{ac}^{\sigma}\Gamma_{bd}^{\sigma}}{2}\int \psi_a(\mathbf{r})\psi_b(\mathbf{r})\frac{\mathrm{erf}(\omega|\mathbf{r}-\mathbf{r}'|)}{|\mathbf{r}-\mathbf{r}'|}\psi_c(\mathbf{r}')\psi_d(\mathbf{r}')\,\mathrm{d}^3\mathbf{r}' \tag{S3}$$

  for both range-separated $\omega = 0.4$ and non-range-separated $\omega \to \infty$ in each spin channel. The single range-separated feature at $\omega = 0.4$ in atomic units was chosen empirically based on validation set performance.

We perform these feature computations with slightly different settings during training and testing to ensure that the large molecules in the benchmark sets (e.g. C_{60}) converge quickly. Additionally the 3 exchange energies that are enhanced by our network are computed as $e^{\mathrm{LDA}}(\mathbf{r}) = -2\pi[(3/4\pi)(\rho^{\uparrow} + \rho^{\downarrow})]^{4/3}$, $e^{\mathrm{HF}}(\mathbf{r}) = e_{\uparrow}^{\mathrm{HF}}(\mathbf{r}) + e_{\downarrow}^{\mathrm{HF}}(\mathbf{r})$, and $e^{\omega\mathrm{HF}}(\mathbf{r}) = e_{\uparrow}^{\omega\mathrm{HF}}(\mathbf{r}) + e_{\downarrow}^{\omega\mathrm{HF}}(\mathbf{r})$.

Overall, the functional can be described as a local range-separated hybrid using only the occupied orbitals. There have been several interesting attempts to construct local-hybrids in the literature (*49–51*) but none has yet clearly surpassed the performance of the simpler global hybrids.

For training we use PySCF (*19*) with the aug-pc-3 basis set, and the integrals in equation (S3) are performed without any approximation. We compute features on 8 different grids for each reaction, given by Treutler, Mura-Knowles, Delley, and Gauss-Chebyshev schemes at PySCF grid levels 2 and 3. The 8 grids of features are all labelled with the same reaction energy and each grid is treated as an independent training example so that the model does not specialise to any particular choice of grid.

For testing, we run all benchmarks using the (aug′-)def2-QZVP basis set (*43*) on Treutler grids at PySCF level 3. For efficiency, the local HF integrals are computed using the resolution of the identity method described in supplementary section S3.1.

Besides computing the raw input features $\mathbf{x}$, we also supply the gradients of these features $\partial\mathbf{x}/\partial\Gamma^{\sigma}_{ab}$ to compute the Fock matrices $F^{\sigma}_{ab} = (\mathrm{d}E/\mathrm{d}\mathbf{x}) \cdot (\partial\mathbf{x}/\partial\Gamma^{\sigma}_{ab})$ required for gradient regularization during training and for self-consistent calculations after training. The gradients of ρ^{σ}, $\nabla\rho^{\sigma}$ and τ^{σ} can all be easily computed on-the-fly from the atomic orbitals and their derivatives on the grid. However, naïvely precomputing $\partial e^{\omega\mathrm{HF}}_{\sigma}(\mathbf{r})/\partial\Gamma^{\sigma'}_{ab}$ produces objects that scale with size $\mathcal{O}(GB^2)$ where G is the number of grid points and B is the size of the basis set. This amount of data is prohibitively large to store and move during training, so instead we precompute and store a $\mathcal{O}(GB)$ sized object

$$\chi^{\sigma,\omega}_{a}(\mathbf{r}) = \Gamma^{\sigma}_{bd}\psi_b(\mathbf{r}) \int \frac{\mathrm{erf}(\omega|\mathbf{r}-\mathbf{r}'|)}{|\mathbf{r}-\mathbf{r}'|}\psi_a(\mathbf{r}')\psi_d(\mathbf{r}')\ \mathrm{d}^3\mathbf{r}', \tag{S4}$$

such that the full gradient $\partial e^{\omega\mathrm{HF}}_{\sigma}/\partial\Gamma^{\sigma'}_{ab} = -\psi_a\chi^{\omega,\sigma}_{b}\delta_{\sigma\sigma'}$ can be computed on-the-fly. By combining $\partial\mathbf{x}/\partial\Gamma^{\sigma}_{ab}$ with the factor $(\mathrm{d}E/\mathrm{d}\mathbf{x})$ computed by automatic differentiation, we obtain the

exchange-correlation contribution to the Fock matrix.

S1.2 Network architecture

The 11-dimensional feature vector $\mathbf{x}(\mathbf{r})$ at each grid point is independently passed through a shared neural network with the following architecture (here we drop the $\mathbf{r}$ argument for clarity). First $\mathbf{x}$ is passed through an element-wise squashing function $\log(|x_i| + \eta)$ where η is a small constant 10^{-4} to ensure numerical stability. The squashed result is then passed through a linear layer and a tanh non-linearity to produce a 256 element activation vector which is fed into a 6 layer MLP of width 256 at each layer (roughly 4×10^5 parameters in total). The MLP has elu nonlinearities (to encourage smooth derivatives) and layer normalisation between layers. We also find a small benefit in initialising the weight matrices close to the identity. The final layer of the MLP is projected to a vector of 3 elements by a linear layer and then passed through a scaled sigmoid to produce the enhancement factors $\mathbf{f}_\theta$ in equation (S1) that are bounded by $0 < f_i < 2$: a range inspired by the Lieb-Oxford bound and then empirically tuned. The final integral over all grid points is performed by quadrature.

The architecture is made spin symmetric by running the same network twice – once for each spin ordering of the input features – and then averaging the two sets of enhancement factors. In practice we found that the network can also learn to become approximately spin symmetric in a single pass by simply random flipping of spin channels per molecule during training, and both approaches to spin symmetrisation achieve similar benchmark scores.

S1.3 Training objective

The overall objective function is the sum of a regression and gradient regularization term, both with units of [energy]^2.

$$\mathcal{L} = \mathbb{E}_r[(\Delta E_{\mathrm{xc},r}^{\mathrm{DM21}} - \Delta E_{\mathrm{xc},r}^{*})^2] + \lambda \mathbb{E}_s[\delta E_{\mathrm{SCF},s}^2]. \tag{S5}$$

Here $\mathbb{E}$ denotes the expectation over the dataset of reactions (r) or systems (s), and the hyperparameter λ controls the weighting of the gradient term (we use a value $\lambda = 1$ throughout). $\Delta E_{\mathrm{xc},r}^{\mathrm{DM21}}$ denotes our model's prediction for the total reaction exchange-correlation energy (product minus reactant exchange-correlation energy) and $\Delta E_{\mathrm{xc},r}^{*}$ denotes an accurate exchange-correlation energy for the reaction computed from literature or CCSD(T) total energies.

The network is trained to convergence at 3×10^6 steps of an Adam optimiser with a learning rate that decays from 10^{-4} to 10^{-6} during training. At each step the regression and gradient components of the loss are computed on minibatches of size ~ 60 and 8, respectively. We train the model using TPU accelerators at FP32 precision with careful implementation of pairwise summation to retain accuracy in the final quadrature integral over the grid.

S1.3.1 SCF loss

A common concern with deep learned functionals is that the functional derivatives may be noisy, prohibiting use in self-consistent optimization. We solve this by regularizing the functional derivative using a cheap second-order perturbation theory result, and we find that this enables SCF calculations on all the benchmarks that we have run in this work. The spirit of our approach is to construct the Roothan-Hall equations using a Fock matrix $\mathbf{F}$ constructed from the neural network's derivatives, and to demand that the solution is close to reasonable Kohn-Sham orbitals from a traditional functional. More specifically, we use the Roothan-Hall equations to derive an approximate expression for the change in energy after a single SCF iteration starting from orbitals from a traditional functional (B3LYP). In equation (S5) we have a term that keeps this change, δE_{SCF}, small and hence regularizes the network gradients. The expression we obtain for δE_{SCF} is:

$$\delta E_{\mathrm{SCF}} = \frac{1}{2} \sum_{i \neq j} \frac{n_i - n_j}{\epsilon_i - \epsilon_j} \left[\mathbf{C}^\top \mathbf{F} \mathbf{C} \right]_{ij}^2 , \tag{S6}$$

where n, ϵ are the orbital occupations and energies, and $\mathbf{C}$ is a reasonable guess for the orbital coefficients taken from a traditional functional (and we drop the spin index for clarity). The remainder of this section is a derivation of equation (S6).

To preserve orthonormality an SCF iteration must only result in transformation of $\mathbf{C}$ by an orthogonal matrix. Representing this orthogonal matrix as the exponential of an antisymmetric matrix $\mathbf{A}$ and expanding, we obtain the update to the orbitals as

$$\delta\mathbf{C} = \mathbf{C}e^{\mathbf{A}} - \mathbf{C} = \mathbf{C}\mathbf{A} + \tfrac{1}{2}\mathbf{C}\mathbf{A}^2 + \mathcal{O}(\mathbf{A}^3). \tag{S7}$$

This gives a change in the density matrix $\mathbf{\Gamma} = \mathbf{C}\mathbf{n}\mathbf{C}^\top$ of

$$\delta\mathbf{\Gamma} = \mathbf{C}\mathbf{n}\delta\mathbf{C}^\top + \delta\mathbf{C}\mathbf{n}\mathbf{C}^\top + \delta\mathbf{C}\mathbf{n}\delta\mathbf{C}^\top, \tag{S8}$$

where $[\mathbf{n}]_{ij} = n_i\delta_{ij}$ is a diagonal matrix of occupations. Substituting equation (S7) and using the antisymmetry of $\mathbf{A}$ yields

$$\delta\mathbf{\Gamma} = \mathbf{C}\left[\mathbf{A}\mathbf{n} - \mathbf{n}\mathbf{A} + \frac{1}{2}(\mathbf{n}\mathbf{A}^2 + \mathbf{A}^2\mathbf{n} - 2\mathbf{A}\mathbf{n}\mathbf{A}) + \mathcal{O}(\mathbf{A}^3)\right]\mathbf{C}^\top. \tag{S9}$$

Using the Fock matrix $F_{ab} = \partial E/\partial\Gamma_{ab}$ we can convert this into a change in energy:

$$\delta E = \sum_{ij}\left[\mathbf{C}^\top\mathbf{F}\mathbf{C}\right]_{ij}\left[(n_i - n_j)A_{ji} + \frac{1}{2}\sum_k (n_i + n_j - 2n_k)A_{jk}A_{ki}\right] + \mathcal{O}(\mathbf{A}^3). \tag{S10}$$

To find $\mathbf{A}$, consider that the SCF iteration guarantees that the coefficients $\mathbf{C} + \delta\mathbf{C}$ diagonalise the Fock matrix:

$$(\mathbf{C} + \delta\mathbf{C})^\top\,\mathbf{F}\,(\mathbf{C} + \delta\mathbf{C}) = \boldsymbol{\epsilon}. \tag{S11}$$

Inserting equation (S7) in equation (S11) and solving for $\mathbf{C}^\top\mathbf{F}\mathbf{C}$ gives

$$\left[\mathbf{C}^\top\mathbf{F}\mathbf{C}\right]_{ij} = (\epsilon_j - \epsilon_i)A_{ij} + \epsilon_i\delta_{ij} + \mathcal{O}(\mathbf{A}^2). \tag{S12}$$

The diagonal part of this equation gives a means of estimating the orbital energies to leading order: $[\mathbf{C}^\top\mathbf{F}\mathbf{C}]_{ii} = \epsilon_i$, and the off diagonal part gives a leading order expression for $\mathbf{A}$

$$A_{ji} = \frac{[\mathbf{C}^\top\mathbf{F}\mathbf{C}]_{ij}}{\epsilon_i - \epsilon_j} \qquad i \neq j. \tag{S13}$$

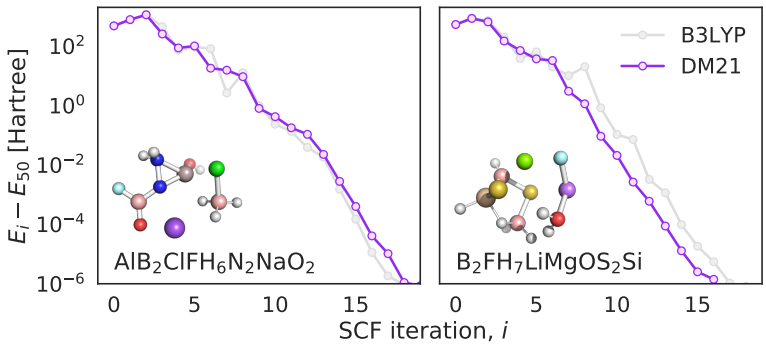


Figure S1: **Example SCF convergence trajectories**. We show the energy at each iteration relative to the energy at iteration 50, and plot the approach on a log-axis truncated at 10^{-6} Hartree, which is roughly the accuracy limit of our RI approximation (see section S3.1).

Additionally, equation (S12) simplifies equation (S10) to

$$\delta E = \frac{1}{2}\sum_{ij}(\epsilon_j - \epsilon_i)(n_j - n_i)A_{ji}^2 + \mathcal{O}(\mathbf{A}^3), \tag{S14}$$

Note that the first order in $\mathbf{A}$ cancels in the overall result and the energy change is second order in $\mathbf{A}$. Finally using equation (S13) in equation (S14) gives equation (S6).

In Fig. S1 we show example SCF convergence traces from the first two complexes in the MB16-43 dataset. We run this optimization in the def2-QZVP basis starting from the eigenstates of the 1-electron Hamiltonian to demonstrate that the stability of DM21 is comparable to B3LYP during these trajectories. We further analyse the quality of densities from DM21 in section S5.

S2 Training data

We use three types of data to train our functional: molecular properties labelled with energies from high accuracy wavefunction-based methods, FC and FS exact constraints expressed as data and data for evaluating the SCF loss in sectionS1.3.1. Each of these datasets is described below.

S2.1 Molecular properties dataset

The molecular dataset consists of 1161 energy differences, with most target values and geometries obtained from literature and the remainder computed by us. The reactions taken from literature include the atomisation energy of small molecules from W4-17 (*21*), W4-11 (*22*) and a set which we collectively refer to as W4-C consisting of 10 carbon-only compounds and 17 alkanes (that do not overlap with W4-17) compiled from Refs. (*23, 52*). We also include binding energies of small molecule complexes based on S66x8 (*24*) and X40x10 (*25*). For S66x8 we use the target values from (*53*) and remove the shortest dimers as recommended by those authors. For X40x10, we exclude systems containing atoms past Kr. We also generated our own energies for the following 222 systems:

- 20 total energies for atoms H-Ar, and ions K^+ and Ca^{2+}.
- 48 single, double and triple ionisation potentials for atoms He-Ar.
- 11 electron affinities for atoms H-Ar.
- 31 bounded cation dimers for H, He, Li, Be, Na, Mg and Al at 90%, 95%, 100%, 105% and 110% of the equilibrium bond lengths.
- 112 atomisation energies from the G3 set (*54*) filtered for systems where the contribution of perturbative triples in CCSD(T) is less than 5% as recommended in Ref. (*21*).

For these systems, we carried out all electron CCSD(T) calculations with extrapolation to the complete basis set limit. The extrapolation is based on (aug-)cc-pCV(Q+d)Z and (aug-)cc-pCV5Z, where aug- is used for systems containing an anion. The Hartree Fock, CCSD correlation energy and perturbative triples contribution are separately computed. Correlation and triples correction are extrapolated using a 2 point method with a Z^{-3} power law, whereas Hartree-Fock energies are extrapolated based on an exponential Karton-Martin scheme (*55*).

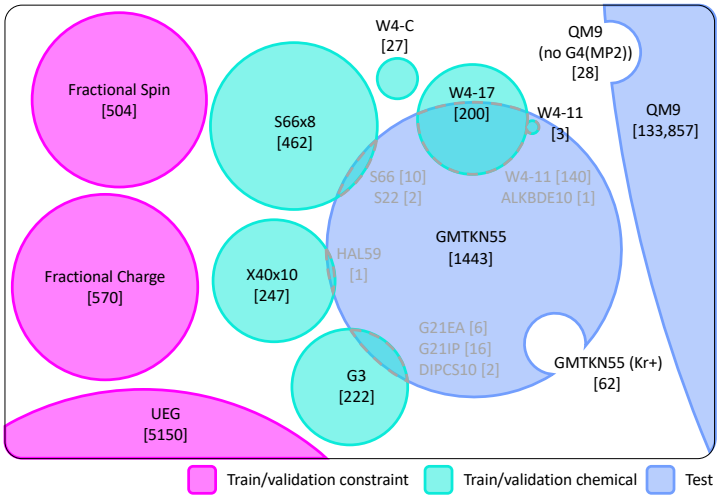


Figure S2: **Data sets used in this study**. The area of each set is proportional to the number of chemically distinct reactions in the set, and counts are shown in square brackets. Overlaps between GMTKN55 and the training set are computed by geometry matching and enumerated in grey. The UEG dataset is used in Supplementary section S4.2.

A summary diagram of the final dataset and any overlaps between the training and test sets is provided in Fig. S2. We measure overlaps between the training data and the GMTKN55 benchmark using a graph isomorphism test of all reactants and products. This test is run by generating a fully connected graph for each molecule with atoms at the nodes and all edges labelled with their length. We then test for isomorphism between graphs with a resolution of 0.01Å. An overlap is detected when two reactions are found where all reactants and products have isomorphic partners and the reaction stoichiometries agree (or the reaction is reversed). We find 178 reactions duplicated in both the training set and GMTKN55. Additionally we find 13 reactions that appear in both W4-17 and QM9, but for clarity we do not show this overlap in Fig. S2 since this represents less than 0.01% of the QM9 set.

S2.1.1 Accuracy vs. dataset size

We performed a preliminary experiment to investigate how the final model accuracy varies with training set size as follows: We took the W4-C, W4-17 and G3 sets and created a validation set

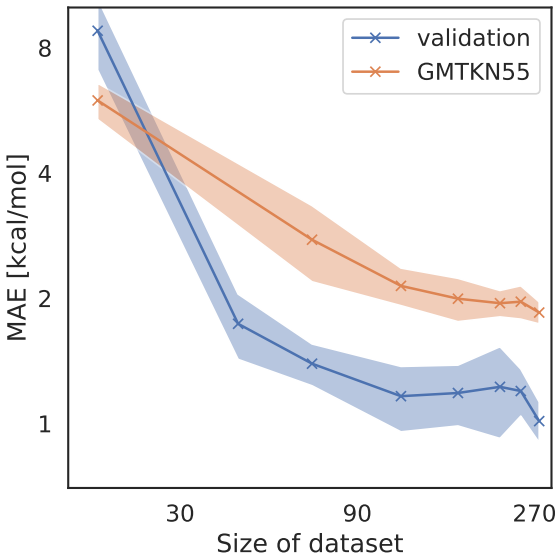


Figure S3: **Mean absolute error dependence on train set size**. Different models are trained with varying numbers of examples from the W4-C, W4-17 and G3 sets. Displayed is the MAE for held-out validation examples and the MAE for GMTKN55.

by selecting 20% of the data at random. The total atomic energies for H-Ar were always included in the training data and the remaining data was used to generate training sets of different sizes. We sampled 10 training sets for each dataset size and carried out a training run. Gradient training was always supervised with the same fixed SCF dataset from section S2.3. Figure S3 shows the dependence of functional error on train set size for the held out validation set and GMTKN55 (evaluated post-B3LYP). As the training dataset size increases, the validation error decreases and approaches chemical accuracy (1 kcal/mol) with around 300 examples. The error on GMTKN55 does not fall as low indicating that there is a distribution shift between the training data used in this investigation and GMTKN55. This provides evidence that our benchmark results are a test of generalization out of distribution.

S2.2 FC and FS datasets

Here we describe the FC and FS datasets and the Wu-Yang algorithm used to generate both.

FC dataset: From the linearity of energy with fractional electron number, the energy change in the reaction $\mathrm{X}^{f+} \rightarrow (1-f)\mathrm{X} + f\mathrm{X}^{+}$ is exactly zero for $0 \leq f \leq 1$ at a fixed geometry X. To construct a dataset of such reactions we use either a single neutral atom or monatomic anion (where bound) from H-Ar in the place of X and enumerate f in steps of 0.01. From the linearity of the density, we construct densities for X^{f+} from a linear combination of densities for X and X^{+}, and then use the Wu-Yang (*56*) method (see below) to provide Kohn-Sham orbitals for the fractional system with one fractional occupation for the frontier orbital.

FS dataset: Here we consider a linear combination of the two degenerate spin states with total spin, S, and $m_S = \pm S$ as in (*18*)

$$\rho[S,\gamma] = \left(\frac{1}{2} + \frac{\gamma}{2S}\right)\rho[S,S] + \left(\frac{1}{2} - \frac{\gamma}{2S}\right)\rho[S,-S], \tag{S15}$$

with $\rho[S, m_S]$ denoting the density of each spin state and all states $-S \leq \gamma \leq S$ being degenerate in energy. We use symmetry labels to ensure that the total occupation of each symmetry channel is the same for the two interpolation limits $\rho[S, \pm S]$. For example, in the case of the carbon triplet we have occupations $\rho[1,1] : [\mathrm{Be}]2p^1_{x\uparrow}2p^1_{y\uparrow}$ and $\rho[1,-1] : [\mathrm{Be}]2p^1_{x\downarrow}2p^1_{y\downarrow}$, which leads to the $\gamma = 0$ occupation $\rho[1,0] : [\mathrm{Be}]2p^{1/2}_{x\uparrow}2p^{1/2}_{y\uparrow}2p^{1/2}_{x\downarrow}2p^{1/2}_{y\downarrow}$. Note that this $\gamma = 0$ state is seen in the limit of closed shell stretching of C_2, and here we correctly label it with the same energy as the triplet.

To construct a dataset of zero-energy reactions $\rho[S,S] \rightarrow \rho[S,\gamma]$ we used all neutral atoms, cations and bounded anions for H-Ar with unpaired electrons. For each nucleus we enumerate γ in steps of 0.01. We again use the Wu-Yang method to recover Kohn-Sham orbitals with fractional occupation from the interpolated density $\rho[S,\gamma]$.

Wu-Yang algorithm and extensions The fractional electron conditions concern densities

on the linear interpolation between the nearby integer ground states, but to compute the kinetic energy density τ and the local HF features $e^{\omega\mathrm{HF}}$ we must convert these interpolated densities to Kohn-Sham orbitals. To achieve this we implement the Wu-Yang inversion algorithm (*57*) that given a density $\rho^*(\mathbf{r})$ finds the one-electron potential v_s and the corresponding orbitals ϕ_i – that diagonalise $(\hat{T} + \hat{v}_\mathrm{s})$ – by solving the maximisation

$$\begin{aligned} v_\mathrm{s} &= \underset{v}{\mathrm{argmax}} \left(W_\mathrm{s}[v(\mathbf{r})] \right) \\ \text{where} \quad W_\mathrm{s}[v(\mathbf{r})] &= \sum_i n_i \langle \phi_i | \hat{T} | \phi_i \rangle + \int v(\mathbf{r}) \left(\sum_i n_i |\phi_i(\mathbf{r})|^2 - \rho^*(\mathbf{r}) \right) \mathrm{d}\mathbf{r}. \end{aligned} \tag{S16}$$

We extend this to fractional electron calculations by constraining the (fractional) orbital occupations n_i and the orbital symmetries to those dictated by the interpolations above. In practice this optimisation is performed by finding optimal expansion coefficients b_t for expanding $v(\mathbf{r})$ with a basis set $\{g_t(\mathbf{r})\}$ as a correction to the external and Fermi-Amaldi potentials:

$$v(\mathbf{r}) = v_\mathrm{ext}(\mathbf{r}) + \frac{N-1}{N} \int \frac{\rho^*(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \mathrm{d}\mathbf{r}' + \sum_t b_t g_t(\mathbf{r}), \tag{S17}$$

where $N = \sum_i n_i$. For this method to converge reliably for all the monatomic systems in our dataset we find that we must use the large aug-pc-3 basis set for expressing ϕ_i and the even larger basis set generated by PySCF's `aug_etb` function applied to aug-pc-3 for expanding v. With these basis sets and the symmetry constraints no additional regularisation is needed to converge the calculations.

S2.3 SCF dataset

The SCF dataset consists of a set of small individual molecules for which we can efficiently compute the regularization term in equation S6. To evaluate the Fock matrix $F^\sigma_{ab} = (\mathrm{d}E/\mathrm{d}\mathbf{x}) \cdot (\partial\mathbf{x}/\partial\Gamma^\sigma_{ab})$ in this expression during training we can precompute and store the input feature derivatives $(\partial\mathbf{x}/\partial\Gamma^\sigma_{ab})$. The derivatives for the local HF $e^{\omega\mathrm{HF}}$ features are obtained through

$\chi_a^{\sigma,\omega}(\mathbf{r})$ in equation S4, which have space complexity GB and set the limit for the largest system that can be handled during training. We include as many of the systems from the molecular training set as possible, and simply limit the size of the basis set and the grid level for larger molecules where GB can become impractically large.

For all the atoms and diatomic molecules in the regression set we generate the features for the SCF loss at PySCF grid level 2 and use the largest basis set in the aug-pc-X family such that the number of basis functions is less than 128. For larger neutral molecules we use grid level 1 and the largest basis set with less than 128 basis functions from cc-pCV(Q+d)Z, cc-pCV(T+d)Z, cc-pV(T+d)Z or cc-pV(D+d)Z. Any system with more than 32,000 grid points is excluded. With these filters we produce an SCF dataset of 931 systems, which we also ran using the same four radial schemes as above (Treutler, Mura-Knowles, Delley, and Gauss-Chebyshev). When training on FC or FS data, the fractional orbitals obtained from the Wu-Yang procedure described above are used to generate inputs for the SCF loss.

S3 Benchmark evaluation

Here we describe our procedure for running the large scale GMTKN55 and QM9 benchmark calculations. We additionally provide details of the resolution of the identity (RI) approach we use for accelerating computation of the local HF features and an alternating optimization scheme used to enable fractional occupation of the frontier orbitals in Fig. 2 in the main text. Finally we also give details of the calculations used to run compressed H_{24} in Fig. 3B.

All GMTKN55 and QM9 benchmark calculations are run using the aug′-def2-QZVP basis set and started from B3LYP initial guess orbitals with a restricted ansatz for closed shell systems and unrestricted otherwise. The functional to benchmark is run from this guess on a Treutler grid at PySCF level 3 for up to 25 SCF cycles, or until a convergence tolerance of 10^{-7} Hartree is reached. If the energy at the end of the SCF cycles is higher than at the start we revert back

to the initial guess and run a new SCF calculation with a level shift of 0.5. If we again see an increase in energy then we run with the first order gradient descent algorithm described in section S3.2. The majority of the calculations reach the convergence tolerance in the first SCF run. We find that even if the strict tolerance is never met for a system, all calculations are at least converged at the sub kcal level, so we take the energy at the final cycle and do not exclude any calculations. All calculations use resolution of the identity (see supplementary section S3.1) to accelerate exchange and coulomb integrals.

S3.1 Resolution of the identity

Computation of the local HF features dominates the execution time of our functional. Resolution of the identity (RI) approaches (*58*), significantly accelerate this step. After feature computation, running the MLP on all grid points only scales linearly with grid size, so in practice the runtime of our method in PySCF (*19*) is comparable to existing hybrid functionals.

Here we provide the details of the resolution of the identity method for efficiently evaluating the local HF features $e_\sigma^{\omega\mathrm{HF}}$ and the component of their gradient $\chi_a^{\sigma,\omega}$ in equations (S3) and (S4). For this discussion we consider the $\omega \to \infty$ limit for clarity, but the finite ω case is simple to recover by inserting $\mathrm{erf}(\omega|\mathbf{r}|)/|\mathbf{r}|$ as the kernel in the integrals in place of $1/|\mathbf{r}|$ as appropriate. Additionally we drop the σ label to compact the notation and use Einstein summation throughout.

First we analyse an inefficient but exact approach for calculating these quantities (which is used to generate the features for the molecules in the training set). We can write χ_a as

$$\chi_a(\mathbf{r}) = \Gamma_{bc}\psi_c(\mathbf{r})\nu_{ab}(\mathbf{r}) \qquad \text{where} \qquad \nu_{ab}(\mathbf{r}) = \int \frac{\psi_a(\mathbf{r}')\psi_b(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}^3\mathbf{r}'. \tag{S18}$$

Each of the the elements of $\nu_{ab}(\mathbf{r})$ requires expensive computation of an integral, and there are GB^2 elements in total, were G is the number of grid points and B the size of the basis set. These integrals dominate the exact computation of both χ_a and e^{HF}.

The approximate resolution of the identity (RI) approach with careful precontraction improves the run time as follows: We compute the following $B'B^2$ integrals using an auxiliary basis Ψ_m of size $B' \ll G$:

$$I_{nab} = \iint \frac{\psi_a(\mathbf{r})\psi_b(\mathbf{r})\Psi_n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}^3\mathbf{r}\mathrm{d}^3\mathbf{r}'. \tag{S19}$$

Now we can approximate an atomic orbital-atomic orbital pair as $(\psi_a\psi_b) \approx A_{mab}\Psi_m$ using the fitting coefficients A_{mai} given by

$$A_{mab} = [\mathbf{S}^{-1}]_{mn} I_{nab}, \qquad \text{where} \qquad S_{mn} = \iint \frac{\Psi_m(\mathbf{r})\Psi_n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}^3\mathbf{r}\mathrm{d}^3\mathbf{r}'. \tag{S20}$$

The matrix elements of the potential from the HF exchange term is then

$$\begin{aligned} k_{ab} &= \Gamma_{cd} \iint \frac{\psi_a(\mathbf{r})\psi_c(\mathbf{r})\psi_b(\mathbf{r}')\psi_d(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}^3\mathbf{r}\mathrm{d}^3\mathbf{r}' \\ &\approx \Gamma_{cd} A_{mac} I_{mbd} = (C_{di} I_{mbd})\,[\mathbf{S}^{-1}]_{mn}\,(C_{ci} I_{nac})\,. \end{aligned} \tag{S21}$$

The leading complexity in the optimal contraction order of the final expression is $\mathcal{O}(B'^2BN)$, where N is the number of occupied orbitals. Now expanding $\chi_a(\mathbf{r})$ using the AO basis $\psi_a(\mathbf{r})$ with least-squares coefficients gives:

$$\chi_a(\mathbf{r}) \approx \left(k_{ab}\left[\mathbf{s}^{-1}\right]_{bc}\right)\psi_c(\mathbf{r}) \qquad \text{where} \qquad s_{ab} = \int \psi_a(\mathbf{r})\psi_b(\mathbf{r})\,\mathrm{d}^3\mathbf{r}. \tag{S22}$$

The dominant complexity remains in the computation of k_{ab}. Overall the quartic scaling of this approach is worse than the exact route to $\chi_a(\mathbf{r})$ above. However the number of integrals to compute is reduced by a significant factor B'/G and the prefactor for the runtime of the remaining linear algebra steps is small.

We could then go on to compute $e^{\mathrm{HF}} = -(1/2)\Gamma_{ab}\psi_a\chi_b$ without altering the dominant complexity. However, we were cautious of the approximation in equation (S22) because it is not clear that the AO basis is sufficiently expressive to accurately expand χ_a. We considered

that this approximation (in a large basis set) suffices for χ_a itself (which is only used to compute the gradient that directs steps in an SCF optimisation loop) but $e_{\mathrm{x}}^{\mathrm{HF}}$ is fed as a feature directly into the neural network, so we decided to avoid the AO expansion approximation for computing e^{HF}. Instead, we compute e^{HF} as follows:

$$e^{\mathrm{HF}}(\mathbf{r}) = -\frac{1}{2}\Gamma_{ab}\Gamma_{cd}\int \frac{\psi_a(\mathbf{r})\psi_c(\mathbf{r})\psi_b(\mathbf{r}')\psi_d(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}^3\mathbf{r}' \tag{S23}$$

$$\approx -\frac{1}{2}\left[C_{ai}\psi_a(\mathbf{r})\right]\left[C_{cj}\psi_c(\mathbf{r})\right]\left[C_{bi}C_{dj}I_{mbd}\right]\int \frac{\Psi_m(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}^3\mathbf{r}', \tag{S24}$$

This requires only GB integrals and a linear algebra leading order complexity of $GB'N^2$. As future work we could investigate whether this additional care is necessary for e^{HF}, or whether the 'free' computation of e^{HF} from the approximate χ_a would suffice.

Additionally we use similar methods to evaluate the matrix elements of the Coulomb term required to evaluate the total energy. Concretely, we perform the contraction

$$j_{ab} = \Gamma_{cd}\iint \frac{\psi_a(\mathbf{r})\psi_b(\mathbf{r})\psi_c(\mathbf{r}')\psi_d(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\mathrm{d}^3\mathbf{r}\mathrm{d}^3\mathbf{r}' = \left(C_{ci}C_{di}I_{mcd}\right)\left[\mathbf{S}^{-1}\right]_{mn}I_{nab}. \tag{S25}$$

The optimum ordering for this contraction has complexity of $\mathcal{O}(B'B^2N)$, and this can be reduced by recycling the intermediate $(C_{ci}I_{mcd})$ computed in equation (S21).

For all benchmarking, we use the large aug′-def2-QZVP basis set (*43*) for the AOs and we use PySCF's `aug_etb` function to generate a corresponding even-tempered auxilary basis set. We compared the exact approach with the RI approach on the diet-GMTKN55-100 (*59*) (excluding the 5 longest-running reactions) and found that the mean error only changed at the 0.01 kcal/mol level.

S3.2 Alternating optimisation for fractional occupation

The usual SCF optimization procedure assumes that Kohn-Sham orbitals have integer occupation. Here we describe a simple first-order method to extend this to fractional occupation that

we found can eliminate the 'hump' in restricted binding curves shown in Fig. 2D in the main text.

Given an energy function E that can be evaluated on an density matrix, the minimisation problem to solve is

$$\min_{\mathbf{C},\mathbf{n}} E\left(\mathbf{\Gamma}\right) \quad \text{s.t.} \quad \mathbf{C}^\top \mathbf{S}\mathbf{C} = \mathbf{1}, \quad 0 \leq n_i \leq 2, \quad \sum_i n_i = N, \tag{S26}$$

where we have used notation from S1.3.1 and N is the total electron count. We perform this optimisation by alternating between optimisation of $\mathbf{C}$ at fixed $\mathbf{n}$, and $\mathbf{n}$ at fixed $\mathbf{C}$ as described below.

First order orbital optimisation: In an orthogonal basis $\bar{\mathbf{C}} = \mathbf{S}^{1/2}\mathbf{C}$, the first order gradient descent step to optimize the energy is given by $\Delta\bar{\mathbf{C}} = -\eta\bar{\mathbf{J}}$, where $\bar{J}_{ij} = \partial E/\partial \bar{C}_{ij}$. In general, this step does not preserve the orthonormality of $\bar{\mathbf{C}}$, so we instead make a projected step as follows. We first write the step as a multiplicative transformation

$$\bar{\mathbf{C}} + \Delta\bar{\mathbf{C}} = \bar{\mathbf{C}} \exp(\mathbf{A} + \mathbf{M}) \tag{S27}$$

where $\mathbf{A}$ and $\mathbf{M}$ are antisymmetric and symmetric matrices to be found. Inserting the definition of $\Delta\bar{\mathbf{C}}$ and matching first order terms in the series expansion for small $\mathbf{A}$ and $\mathbf{M}$ we obtain

$$\mathbf{A} = -\frac{\eta}{2}\left(\bar{\mathbf{C}}^\top\bar{\mathbf{J}} - \bar{\mathbf{J}}^\top\bar{\mathbf{C}}\right) \tag{S28}$$

$$\mathbf{M} = -\frac{\eta}{2}\left(\bar{\mathbf{C}}^\top\bar{\mathbf{J}} + \bar{\mathbf{J}}^\top\bar{\mathbf{C}}\right). \tag{S29}$$

Then to enforce that the step preserves orthonormality we project the multiplicative transformation to an orthogonal matrix by dropping the symmetric $\mathbf{M}$ term.

$$\bar{\mathbf{C}}_{t+1} = \bar{\mathbf{C}}_t \exp(\mathbf{A}), \tag{S30}$$

where t indexes the optimization steps. This procedure matches (*60*) to first order.

For robust convergence, we adaptively change η at each step by recursively halving it from an initial value 0.1 until $E(\mathbf{\Gamma}_{t+1}) < E(\mathbf{\Gamma}_t)$ and $E(\mathbf{\Gamma}_{t+1}) - E(\mathbf{\Gamma}_t) \leq k_{\mathbf{C}} \sum_{ij} \bar{J}_{ij}[\bar{\mathbf{C}}_{t+1} - \bar{\mathbf{C}}_t]_{ij}$. The second condition states that the magnitude of the true energy change is not less than a factor $k_{\mathbf{C}} = 0.5$ of the first order approximation, and thus that the optimisation step stays within a trust region where the first order approximation is qualitatively valid.

Line search for occupations: The updated orbitals have energies $\epsilon_i = \partial E / \partial n_i$ and we construct a candidate new occupation $\tilde{\mathbf{n}}$ by Aufbau filling of these orbitals. To improve on this candidate, consider that all points along the convex sum $\mathbf{n}_{t+1} = \lambda \tilde{\mathbf{n}} + (1 - \lambda)\mathbf{n}_t$ obey the constraints on $\mathbf{n}$ required by equation (S26), and therefore the task is reduced to finding an optimal λ. We start this search at $\lambda = 1$ and recursively halve it until the true energy is reduced and the magnitude of the reduction is not less than a factor $k_{\mathbf{n}} = 0.1$ of the first order prediction. Orbital and occupation optimisation are alternated until the total energy is converged to within a given tolerance (10^{-5} Hartree).

S3.3 Evaluation of compressed H_{24}

To run H_{24} with a compressed H-H bond length of 0.48Å in Fig. 3B we found that it was necessary to optimize a custom even-tempered basis (expressed as $\alpha\beta^i$. Specifically the basis we used is a $10s3p$ basis set with exponents $\alpha\beta^i, i = 0, ..., n$, $n_s = 10, \alpha_s = 35.0, \beta_s = 0.54$ and $n_p = 3, \alpha_p = 0.58, \beta_p = 0.42$. This is generated in PySCF using

```
pyscf.gto.expand_etbs([(0, 10, 35.0, 0.54),
                       (1, 3,  0.58, 0.42)])
```

We used a quasi-Newton method and projected out linear dependence in the basis set with a threshold of 10^{-6} to optimize the orbitals in this basis and plot the spin density $\rho_\uparrow(\mathbf{r}) - \rho_\downarrow(\mathbf{r})$. Additionally we show the line density $n_\sigma(z)$ in each spin channel σ along the z-direction aligned

with the chain, computed as

$$n_\sigma(z) = \int \rho_\sigma(\mathbf{r})\, \mathrm{d}x\mathrm{d}y. \tag{S31}$$

Supplementary Text

S4 Further data and constraint studies

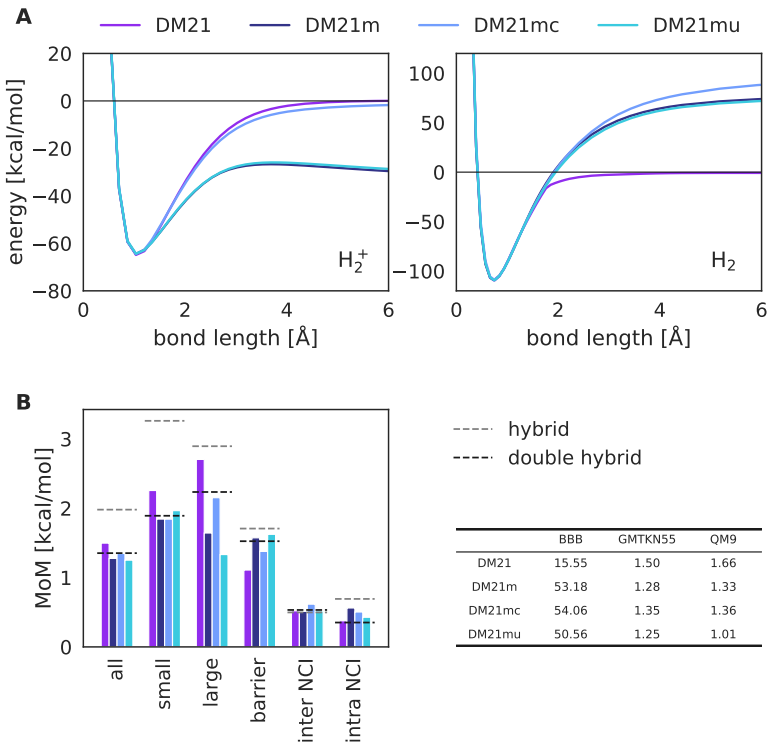
S4.1 Constraint ablation

To assess the effect of training using fractional electron constraints, and to highlight the flexibility of our framework for enforcing different exact constraints, we have also developed three other variants of DM21. All of these variants, named DM21m, DM21mc and DM21mu, were trained on the molecular dataset described in section S2.1 with additional constraint data as follows:

	FC	FS	UEG
DM21	✓	✓	
DM21m			
DM21mc	✓		
DM21mu			✓

The UEG constraint set imposes the uniform electron gas exact limit (see section S4.2). We find that all of these variants show state-of-the-art performance on the large benchmark sets GMTKN55 and QM9, and we provide further analysis of the particularly strong performance of DM21mu on QM9 by studying behaviour on a recurring carbon cage motif in section S4.2.

Fig. S4 gives a summary of the performance of these variants on the archetypal stretched H_2^+ and H_2 systems together with benchmark statistics. Here we see that both FC and FS are required to solve the BBB benchmark set, and in section S7 we further see that only DM21 gives a correct description of the diradical transition states considered in the main text.



	BBB	GMTKN55	QM9
DM21	15.55	1.50	1.66
DM21m	53.18	1.28	1.33
DM21mc	54.06	1.35	1.36
DM21mu	50.56	1.25	1.01

Figure S4: **Ablation of constraints**. **A** Stretching of the archetypal H_2^+ and H_2 systems to illustrate FC and FS behaviour. DM21 and DM21mc correctly describe H_2^+, but only DM21 correctly describes H_2. **B** GMTKN55 MoM performance and benchmark summaries for the DM21 variants. We include reference performance of the best hybrid (PW6B95:$D3_0$) and double hybrid (DSD-PBEP86:$D3_{BJ}$) from (*43*).

S4.2 UEG Constraint

DM21mu serves as an example to demonstrate how our approach can be applied to incorporate different constraints, and specifically DM21mu obeys the UEG limit. To enforce this this we supply a dataset of infinite-extent uniform systems represented to the network as individual grid points with density ($\rho^\uparrow$, $\rho^\downarrow$) in each spin channel. The dataset contains elements of the ($\rho^\uparrow$, $\rho^\downarrow$) plane from (10^{-4}, 10^{-4}) to (10^{-1}, 10^{-1}), and we supply 4950 unique examples evenly spaced in a log space grid inside the lower half plane. These are augmented to cover the full plane by spin

flipping. Additionally we supply 100 examples of fully polarised gas $10^{-4} \leq \rho^{\uparrow} \leq 10^{-1}; \rho^{\downarrow} = 0$ (which are again augmented by spin flipping) and 100 examples lying exactly on the unpolarised diagonal. The features for these systems are computed as:

$$|\nabla \rho^{\sigma}| = 0 \tag{S32}$$

$$\tau^{\sigma}(\mathbf{r}) = \frac{3}{10}(3\pi^2)^{2/3}\rho^{\sigma}(\mathbf{r})^{5/3} \tag{S33}$$

$$e_{\sigma}^{\omega \mathrm{HF}} = -a_{\sigma}\left(\frac{6}{\pi}\right)^{1/3}\left[\sqrt{\pi}\,\mathrm{erf}\left(a_{\sigma}^{-1}\right) + a_{\sigma}(b_{\sigma} - c_{\sigma})\right]\rho^{\sigma}(\mathbf{r})^{4/3} \tag{S34}$$

where $a_{\sigma} = \omega/(6\pi^2\rho^{\sigma})^{1/3}$, $b_{\sigma} = \exp(-a_{\sigma}^{-2}) - 1$ and $c_{\sigma} = (a_{\sigma}^2 b_{\sigma} + 1)/2$. Equation (S34) is taken from (*61*) and gives the correct $\omega \to \infty$ limit to recover the expression for $e_{\mathrm{x}}^{\mathrm{LDA}}$.

The performance of DM21m and DM21mu on the plane for $2 \times 10^{-4} \leq \rho^{\sigma} \leq 50$ is shown in Fig. S5A. Without the UEG data, DM21m shows a few 10s of percent deviation from the LDA functional over this region. This can be reconciled with the good performance of DM21m on most molecular systems because such systems contain very few UEG-like regions at high ρ, and the absolute scale of the energy at small ρ is negligible. DM21m has therefore learned that only loose agreement with the UEG is sufficient to explain the small molecules that it saw during training.

However, strained rings and cage systems are exceptions where training on the UEG constraint can lead to improvements at the kcal/mol level. We investigated why DM21mu significantly outperforms DM21m on the QM9 benchmark (error of 1.0 kcal/mol vs. 1.3) and found that the largest energetic differences arise from a recurring strained hydrocarbon cage motif that has an unusual C-C non-bonded distance of around 1.9 Å. An example is shown in C_8H_9N in Fig. S5B. Plotting ρ vs. the reduced gradient $s = (24\pi^2)^{-1/3}|\nabla\rho|/\rho^{4/3}$, we find that this strained system has regions with low gradient (and low hessian) at reasonably high density ($\rho \sim 0.1$ bohr^{-3}) in the cage centre. At these densities DM21m under-predicts the magnitude of E_{xc} for the unpolarized UEG by up to 10%, whereas DM21mu has $< 0.01\%$ error. This means

that DM21mu moves density into the cage centre and the energy predictions are improved at the kcal/mol level. Similar effects are seen on the strained propane ring, but the effect is reduced in unstrained rings such as cyclopentane where the UEG-like low gradient region appears at lower density and therefore lower absolute energy scale.

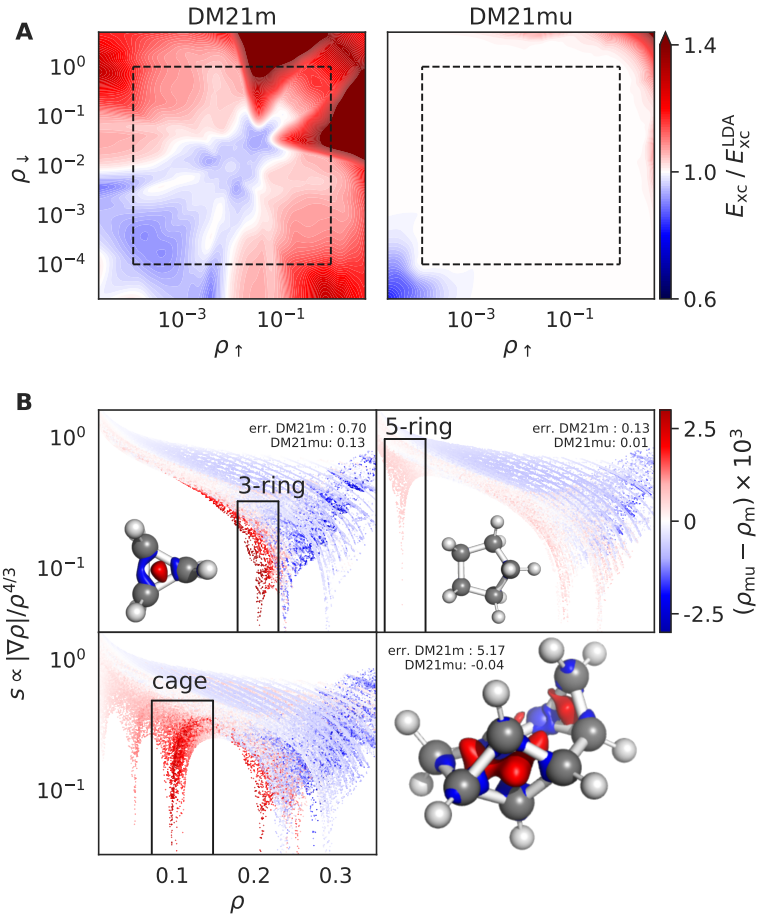


Figure S5: **The UEG constraint**. **A** Performance of DM21m and DM21mu on the UEG relative to the LDA functional. The dashed black line indicates the training region for DM21mu beyond which the functional is being asked to extrapolate. **B** The reduced gradient s vs. the density ρ coloured by the difference between the DM21mu and DM21m densities. Isosurfaces at a density difference of 0.001 (atomic units) show that DM21mu puts more density in low gradient (low hessian) regions that occur at ring and cage centres. The effect is more pronounced in strained structures and correlates with the atomisation energy error (err.) shown vs. G4(MP2) labels.

S4.3 Density vs. Energy Functional

We train our functionals to regress high accuracy energies, but we only supply B3LYP orbital features as the inputs during training. Table S2 gives justification for our choice to invest more resource in the energy labels than the densities. Here we computed orbitals for the W4-11 dataset by self consistent calculation with a number of different functionals and also obtained the more expensive relaxed CCSD densities coupled with a Wu-Yang inversion to generate KS orbitals. We then ran different energy functionals to compute the energy of each of these densities and report the error on the entire W4-11 set for each density functional/energy functional combination. For a fixed choice of energy functional, there is not significant variation in the error between densities, with the M06-2X energy functional showing the greatest standard deviation of 0.65 kcal/mol across the combinations we tried. In contrast, the choice of energy functional for a fixed density has a large effect, e.g. the B3LYP density reveals a large standard deviation of 4.2 kcal/mol across the energy functionals tried. We include DM21m in our investigation and confirm that it is largely insensitive to the underlying density functional when used as an energy functional and its SCF densities fit the trends seen in other functionals when used as a density functional. This lack of sensitivity on the input density is not specific to W4-11, but applies in general. For example S1 shows results for the barrier height set BH76 and the 'mindless benchmarking' set MB08 (*62*). These observations show that what drives the error in density functionals is usually the failure of functionals to represent the energy of a given density, rather than the fact they are optimized for different densities. This statement is true most of the time, but it should be noted that errors in the functional, especially related to fractional charge error can lead to qualitative errors in density as observed in section S5.2 and in figure 2. Nevertheless, the fact that the functional used to optimize the density has a weak influence on the predicted energy of a reaction leaves us free to choose the B3LYP functional for generating input data, but do not expect that this would have significant effect on our results.

Density	basis-set	PBE	BLYP	M06-L	B3LYP	PBE0	M06-2X	CAMB3LYP	DM21m	st.dev.
PBE	def2-QZVP	15.50	7.75	4.25	4.26	3.62	4.06	3.75	0.87	4.16
B3LYP	cc-pCVTZ	15.73	7.88	4.11	4.03	3.72	3.94	3.69	0.94	4.24
B3LYP	def2-QZVP	15.23	7.53	4.15	4.00	3.47	3.42	3.47	0.45	4.19
BLYP	def2-QZVP	15.44	7.81	4.10	4.29	3.71	5.12	3.74	1.03	4.09
M06-2X	def2-QZVP	14.36	6.53	4.25	4.32	3.39	3.01	3.42	0.84	3.83
DM21m	def2-QZVP	15.03	7.43	3.99	4.13	3.45	3.20	3.60	0.48	4.12
CCSD	cc-pCVTZ	15.64	7.84	4.28	4.42	4.25	3.67	4.37	1.46	4.08
st.dev.		0.43	0.44	0.10	0.14	0.27	0.65	0.29	0.32	

Table S1: Investigation of input density sensitivity. The errors on W4-11 for each density functional/energy functional combination are shown and we summarise the density sensitivity by standard deviation down the columns and energy sensitivity at fixed density by standard deviation across the columns. WY(CCSD) are densities obtained by Wu-Yang inversion of relaxed CCSD densities to KS orbitals. All values in kcal/mol.

Density	PBE	BLYP	M06-L	B3LYP	PBE0	M06-2X	CAMB3LYP	DM21m
BH76								
PBE	9.30	8.43	3.75	4.39	3.72	1.28	2.65	1.77
BLYP	9.27	8.47	3.87	4.41	3.66	1.29	2.63	1.87
M06 L	9.18	8.50	3.86	4.71	3.98	1.24	3.00	2.24
B3LYP	9.00	8.19	3.91	4.70	4.11	1.09	3.07	2.17
PBE0	8.85	7.97	3.74	4.65	4.16	1.15	3.14	2.15
M06 2X	8.27	7.41	3.33	4.35	3.92	1.21	3.03	2.14
CAMB3LYP	8.56	7.71	3.72	4.57	4.12	1.16	3.18	2.22
DM21m	8.55	7.73	3.68	4.43	3.91	1.13	3.04	2.38
st. dev.	0.38	0.41	0.18	0.15	0.18	0.07	0.21	0.20
MB08								
PBE	8.92	10.99	13.42	8.29	8.61	4.35	9.39	3.58
BLYP	8.85	10.91	13.37	8.25	8.40	4.24	9.30	3.65
M06 L	9.12	10.72	13.22	8.28	9.03	4.78	9.39	3.66
B3LYP	8.78	10.94	13.34	8.11	8.49	4.56	8.94	2.92
PBE0	8.72	10.93	13.36	8.08	8.62	4.78	8.92	2.75
M06 2X	8.60	11.28	12.96	8.31	8.13	4.77	8.75	2.58
CAMB3LYP	8.52	11.11	13.61	8.23	8.56	4.69	8.84	2.91
DM21m	8.60	11.29	14.06	8.51	8.83	4.81	9.46	2.73
st. dev.	0.20	0.19	0.32	0.13	0.27	0.22	0.29	0.45

Table S2: Investigation of input density sensitivity. The mean absolute deviation on BH76 and MB08 (*62*) for each density functional/energy functional combination are shown and we summarise the density sensitivity by standard deviation down the columns. All values in kcal/mol. Note that for the MB08 dataset, two reactions were excluded for which M06-2X or DM21m failed to converge

S5 Density assessment

S5.1 Self-consistent density

To verify the quality of self-consistent densities from DM21, we follow the procedure from Fig. 1 of Ref. (*63*). We compute the RMSD error versus an accurate CCSD reference density for the systems considered in (*63*) and compare the maximum RMSD on this set for historic functionals through time. In Fig. S6 we show that DM21 does not show large density errors of any of the systems considered in this procedure.

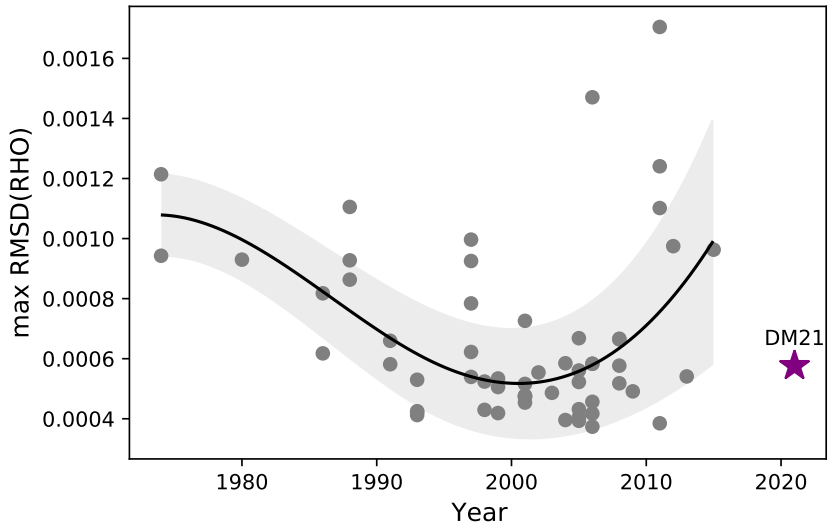


Figure S6: **Self-consistent density error** The maximum RMSD error of self-consistent DFT densities vs. accurate relaxed CCSD reference densities, all calculations use the aug-cc-pwCV5Z basis set. The systems are a set of $2e^-$, $4e^-$ and $10e^-$ atomic species: B^+, B^{3+}, C^{2+}, C^{4+}, N^{3+}, N^{5+}, O^{4+}, O^{6+}, F^{5+}, F^{7+}, Ne^{6+}, Ne^{8+} and Ne. Grey dots are from the conventional functionals listed in (*63*).

S5.2 Dipole moment prediction

As a further test for the quality of the charge densities produced, we ran our functionals on a benchmark set of 152 dipole moments computed at a high level of theory (CCSD(T)/CBS)

from (*64*). Following (*64*) we define the error in terms of relative regularised error η as:

$$\eta = \frac{\mu - \mu_{\text{ref}}}{\max(\mu_{\text{ref}}, 1 \text{ debye})} \times 100\% \tag{S35}$$

The results are given in Table S3, and show that our functionals are competitive with the best hybrid functionals, significantly surpassing performance of the B3LYP functional which supplied the training densities.

	RMSD error (%)	MAE error (%)	MAX error (%)
DM21m	5.42	3.33	30.58
DM21mu	5.49	**3.07**	44.24
DM21mc	6.44	4.18	34.65
DM21	**4.62**	3.20	**18.30**
PW6B95	5.05	3.26	24.15
wB97X-V	5.52	3.71	22.76
M06-2X	6.58	3.96	35.36
B3LYP	7.13	4.11	45.66
SCAN	8.60	5.67	31.89
PBE	11.70	8.30	45.14

Table S3: Percentage relative regularised error for dipole prediction of several local and hybrid functionals. Regularised error is defined as in (*64*).

S6 HOMO eigenvalues

In this section we show the behaviour of DM21 on fractional charge for all atoms and compare the deviation from linearity with the predictions from the highest occupied molecular orbital. As explained in the main text, the exact functional predicts that the energy change between a system with n and $n-1$ electrons is linear. In contrast, most existing functional approximations are concave with respect to the exact condition. This concavity means that the energy can be artificially lowered by delocalising fractions of charges over sub-parts of the molecule. DM21 was explicitly trained such that this condition is enforced. As shown in figure S7, this training

results in deviation from linearity that rarely exceed tens of meV, whereas a popular hybrid density functional such as B3LYP has much larger deviations. Recent work has trained an ML model to tune the range-separation parameter to achieve linearity on a system-by-system basis (*65*).

Another well known result in density functional theory is that the energy of the highest molecular orbital (HOMO) corresponds to the derivative of the energy functional with respect to number of electrons (Janak's theorem). As such, if a functional perfectly obeyed the linearity condition, the HOMO energy would equal the vertical ionization potential measured as the energy difference between a system with n and $n-1$ electrons. For this reason we have added to figure S7 guides to the eye that show where the HOMO energy would predict the energy of the ionized molecule to lie. Note that even for DM21 there are discrepancies between the prediction from the HOMO energy and the actual energy of a ionized atom. This is because, despite the functional closely obeying linearity, it's infinitesimal derivative is still rather noisy. DM21 is therefore still effectively free of charge delocalisation error in the sense that a calculation on a charged dimer would not spuriously delocalise charge, but the molecular orbital energies need to be treated with care. We additionally checked the same conclusion is reached for molecular systems in the G21IP set (Fig. S8).

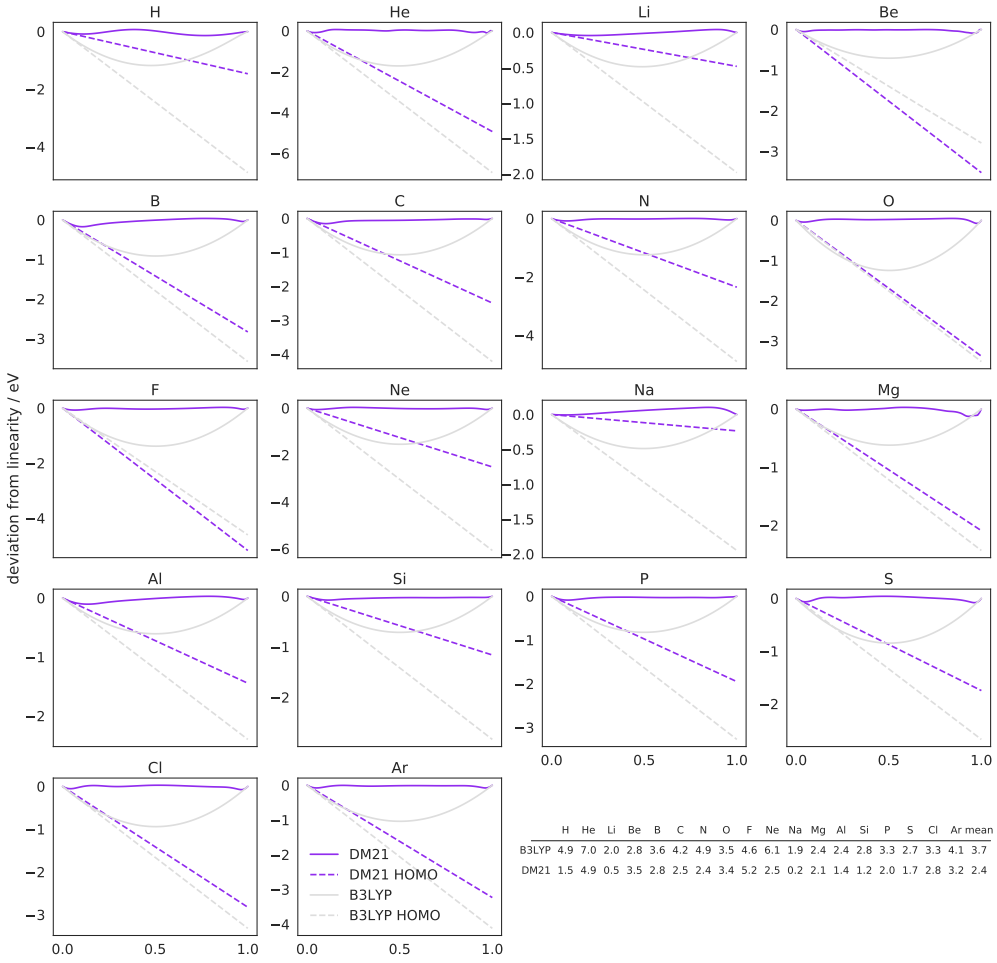
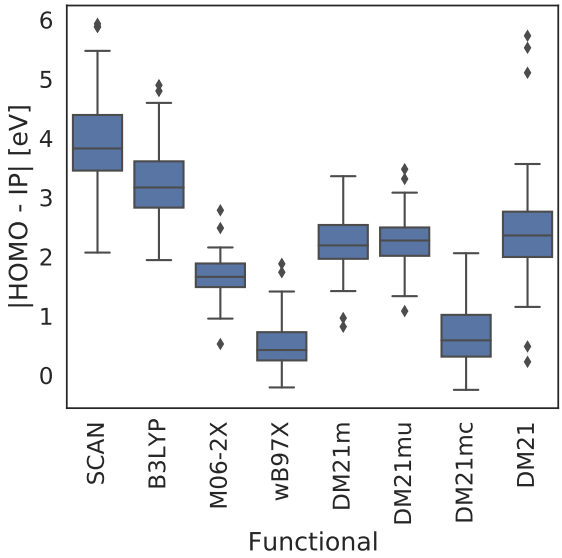


Figure S7: **Deviation from linearity for DM21 and B3LYP for selected atoms**. The full lines show the results of post-B3LYP fractional charge calculations whereas the dashed line shows the projection based on the highest occupied molecular orbital (HOMO) for the neutral system. Values on the x-axis denote the fraction of an electron removed from a neutral system. All energies are relative to the expected linear change in energy between the charged and neutral atoms.



!h

Figure S8: **Deviation of the HOMO from the ionization potential**. Results are presented for the G21IP set. DM21mc shows improved alignment between the HOMO and the ionization potential, but as with traditional functionals, the orbital energy eigenvalues of the neural functionals should be treated with care.

S7 Diradical transition states

In these sections we provide additional details and examples of reaction transition states with diradical character and provide evidence that incorporation of the FS constraint leads to better description.

S7.1 Bicyclobutane reaction barrier heights

The isomerisation of bicyclobutane into butadiene can procede via two mechanisms (*66*): a conrotatory movement of the methylene group and a disrotatory one. The disrotatory pathway in particular is challenging for ab-initio theory as it has strong biradical character. In order to characterise how accurately different functionals capture this transition state, we consider both the forward and backwards cyclobutane to gauche-butadiene reaction barrier heights and show

the error with respect to high accuracy diffusion Monte-Carlo results (*42*). Note that of all the functionals presented, only DM21 has no errors greater than 4 kcal/mol. These energies were computed using an unrestricted ansatz and a def2-QZVP basis set.

	conrotatory		disrotatory	
	fwd	bwd	fwd	bwd
DM21m	9.93	7.05	2.53	-0.35
DM21mu	8.64	5.64	2.72	-0.29
DM21mc	10.62	8.73	1.70	-0.19
DM21	2.36	-0.40	0.09	-2.67
M06-L	-7.28	1.58	-29.45	-20.58
MN15-L	-5.02	10.37	-26.56	-11.17
BLYP	-2.32	2.99	-4.43	0.88
PBE	2.55	-3.60	0.40	-5.75
PBE0	8.00	-0.73	-1.13	-9.86
PW6B95	8.60	1.90	3.27	-3.43
M06	8.97	-1.79	2.75	-8.01
ωB97M-V	10.57	5.56	3.08	-1.93
ωB97X-V	11.89	3.57	2.27	-6.05
M06-2X	12.81	4.13	3.74	-4.94

Table S4: Errors (in kcal/mol) for several functionals for the forward (fwd) and backward (bwd) reactions for the bicyclobutane to gauche-butadiene transition. Errors are based on diffusion Monte-Carlo calculations from (*42*).

S7.2 Dehydro-Diels-Alder barrier heights

In the main text we discussed the case of isomerisation of bicylobutane to show cases where even unrestricted Kohn-Sham can lead to broken bonds which are not fully polarised and thus can lead to overestimation of barrier heights using hybrid density functionals. In this section we further investigate this effect by looking at the case of 2+4 dehydro-Diels-Alder reaction barrier heights which have been studied by Aida et al in (*67*). Dehydro-Diels-Alder reactions can proceed via either a concerted pathway, where both bonds form at once and a stepwise pathway where each bond forms sequentially. We computed barrier heights with different functionals,

our findings are shown in table S5. Note that as is the case for bicylobutane, we find significantly lower barrier heights for the stepwise pathway when using the functional trained on both fractional charge and spin constraints. In this case, reliable reference numbers are not available, but based on our experience with bicyclobutane, we tend to trust DM21.

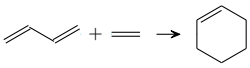
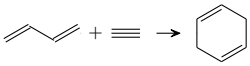
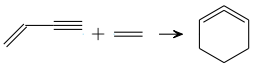
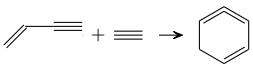
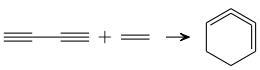
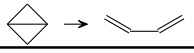
		$\langle S^2 \rangle$ B3LYP	PBE	B3LYP	M06-2X	DM21m	DM21mu	DM21mc	DM21
	TSc	0.00	16.10	26.31	20.66	17.62	18.84	18.65	20.65
	TSstep1	0.54	26.43	35.94	36.95	35.45	36.03	37.22	33.29
	TSint	0.99	28.60	33.73	30.36	31.53	31.63	32.73	31.68
	TSstep2	0.89	31.72	38.60	33.94	34.27	34.87	35.59	35.23
	Prod	0.00	-43.14	-34.38	-46.20	-45.63	-44.99	-44.43	-45.63
	TSc	0.00	15.93	26.10	21.79	19.70	21.10	19.48	22.44
	TSstep1	0.45	22.00	32.87	36.39	34.89	34.94	36.39	29.53
	TSint	0.96	22.06	28.47	28.43	28.91	29.69	30.01	25.37
	TSstep2	0.50	17.83	30.27	30.47	29.09	29.42	30.81	24.19
	Prod	0.00	-62.81	-54.00	-61.49	-60.30	-59.92	-59.46	-61.41
	TSc	0.00	20.44	32.23	28.23	25.10	26.29	25.03	27.47
	TSstep1	0.44	21.46	32.40	35.69	34.36	34.71	35.77	30.45
	TSint	1.03	23.01	28.92	29.13	30.81	30.89	31.60	29.94
	TSstep2	0.52	21.86	34.02	34.38	31.98	32.47	34.25	27.12
	Prod	0.00	-21.00	-8.80	-17.29	-19.40	-18.36	-17.72	-19.28
	TSc	0.00	20.76	32.31	30.00	27.64	29.00	26.35	29.79
	TSstep1	0.39	18.08	29.99	35.75	34.36	33.55	35.91	27.86
	TSint	1.03	15.47	22.40	25.96	27.42	28.08	28.28	23.06
	TSstep2	0.67	14.68	26.24	29.28	27.82	28.34	29.51	19.02
	Prod	0.00	-38.10	-25.23	-29.34	-31.80	-30.97	-29.94	-33.12
	TSc	0.00	25.00	37.27	34.68	31.44	32.66	30.01	33.34
	TSstep1	0.42	20.81	31.99	35.37	35.08	35.21	35.60	29.68
	TSint	1.01	23.74	30.04	30.95	32.83	32.92	33.50	31.05
	TSstep2	0.40	24.83	38.12	39.19	36.93	36.92	38.71	30.54
	Prod	0.00	-11.61	2.15	-3.23	-6.94	-5.36	-5.16	-7.75
	TSc	0.00	25.53	37.12	36.19	33.91	35.29	31.17	35.59
	TSstep1	0.37	17.68	29.91	35.88	35.77	35.41	36.41	27.33
	TSint	1.04	15.59	22.96	26.75	27.98	28.87	29.67	21.94
	TSstep2	0.67	15.85	28.07	31.88	30.48	30.92	31.73	20.86
	Prod	0.00	-66.86	-54.02	-56.89	-60.73	-59.66	-58.07	-62.46
Bicyclobutane 	ΔTS conrotatory	0.29	47.84	47.19	57.06	52.02	52.48	54.61	46.65
	ΔTS disrotatory	1.00	60.75	55.83	64.05	61.80	62.16	61.41	59.70

Table S5: $\langle S^2 \rangle$ values and energies (in kcal/mol) for the reactions from (*67*). All systems are nominal singlets. Reaction labels correspond to the naming in the original paper, for each 2+4 addition reaction we show a concerted pathway with a single transition state TSc and a step-wise pathway characterised by two transition states TSstep1 and TSstep2 together with an intermediate state TSint. Note that large deviations between DM21mc and DM21 correlate with cases with intermediate spin contamination, i.e. in the transition states of the stepwise pathway. All calculations carried out with a def2-TZVP basis set. For comparison we show the same analysis carried out on the forward reaction barrier height for isomerisation of bicyclobutane at the bottom of the table.

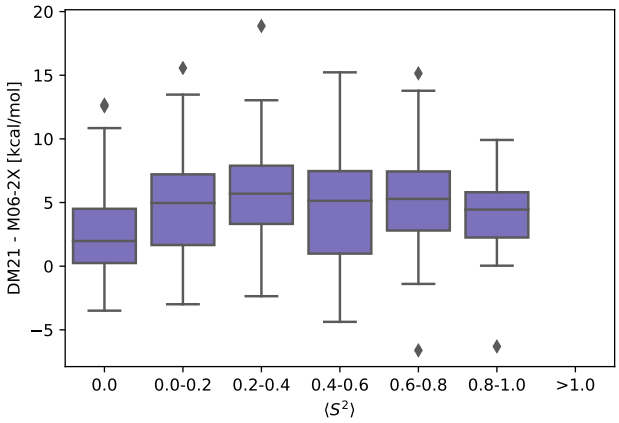


Figure S9: **Barrier heights for the automatically generated transitions states**. Box plot showing the difference in energy (in kcal/mol) between M06-2X and DM21 for the reactions from (*68*) which displayed diradicaloid character.

S7.3 Automatically generated transition states.

In the main text we argued that DM21 tends to predict lower energy barriers for transition states with partial diradical character compared to hybrid density functionals such as M06-2X and ωB97X. In order to investigate if this is the case, we looked at a set of 4,000 transition states from organic reactions automatically generated by Grambow et. al. and based on the GDB-7 dataset (*68*). We carried out B3LYP calculations with a split valence basis set and identified all transition state with an $\langle S^2 \rangle$ greater than 0 (we found 537), we then computed barrier heights for those 537 barrier heights, together with 100 randomly selected barrier heights with $\langle S^2 \rangle = 0$ with DM21 and M06-2X and a triple zeta basis set (def2-TZVP). The results are shown in figure S9 and show that indeed the hybrid density functional overpredicts barrier heights compared to DM21 in cases where S^2 is non zero, and that the effect is strongest when spin contamination has intermediate values.

S8 Detailed benchmark results

Here we provide additional information for the BBB, GMTKN55 and QM9 benchmark evaluations.

S8.1 Bond breaking benchmark (BBB)

The ground state energy of neutral dimers H_2, Li_2, C_2, N_2, F_2 were obtained by optimising the FermiNet wavefunction ansatz with the variational Monte Carlo method (*69, 70*). FermiNet is a flexible neural network which can achieve highly accurate energies for small molecules (exceeding 99% of the correlation energy in the complete basis set limit). We used the architecture and training schedule as given in Ref. (*69*): three layers each with 256 hidden units for the one-electron stream, 32 hidden units for the two-electron stream, and a wavefunction formed from a linear combination of 16 determinants. We optimised FermiNet using the K-FAC optimisation algorithm (*71*), with a learning rate of $1/(10^4+t)$, where t is the iteration index, using a batch of 4096 MCMC configurations, and 10 MCMC all-electron steps between parameter updates, each with a proposal drawn from a normal distribution with standard deviation of 0.02. Optimisation was performed for 150,000-200,000 iterations and the final energy estimated from 50,000 local energy evaluations, each separated by 10 MCMC all-electron steps. The auto-correlation was accounted for using a blocking analysis (*72*). For the charged dimers H_2^+, He_2^+, Li_2^+, B_2^+, Ne_2^+ and Al_2^+ we obtained UCCSD(T) energies from extrapolation from cc-pVXZ basis sets using up to quintuple zeta.

The errors were computed by considering stretching from 0.5 to 10 Å in 50 equally separated steps. Errors are only considered for parts of the binding curve that are bound according to the oracle (QMC or UCCSD(T)).

S8.2 GMTKN55

The GMTKN55 collection of benchmark sets is often analysed using three different aggregation methods that vary in their weighting of the accuracy on NCI systems (*43*). The simplest of these is the MoM metric presented in the main text which does not do any weighting and the other schemes WTMAD-1,2 are presented below:

$$\begin{aligned}
\text{MoM} &= \frac{1}{55}\sum_i \bar{s}_i\,; \\
\text{WTMAD-1} &= \frac{1}{55}\sum_i w_i \bar{s}_i\,; \\
\text{WTMAD-2} &= \frac{1}{1505}\sum_i n_i \frac{56.84\ \text{kcal/mol}}{\bar{t}_i}\bar{s}_i,
\end{aligned} \tag{S36}$$

where $\bar{s}_i$ is the mean absolute error on the i^{th} subset, $\bar{t}_i$ is the mean absolute reaction energy in subset i, n_i is the number of reactions in subset i and w_i is a weighting of 10 if $\bar{t}_i < 7.5\ \text{kcal/mol}$ and 0.1 if $\bar{t}_i > 75\ \text{kcal/mol}$. Note that the values in the denominators reflect the fact that there are 1505 reactions across 55 sub-benchmarks. A table of all the values of $\bar{s}_i$ for DM21 functionals and selected traditional functionals together with the MoM and WTMAD-1,2 scores is presented in Table S6. We note that all of the DM21 variants outperform or are on-par with the best traditional hybrid functional on all metrics.

62 reactions in GMTKN55 involve atoms heavier than Kr for which we used the def2-ECP potential as explained in (*36*). We did not use resolution of the identity for these reactions and we used a quasi-Newton optimization method as we found that convergence was not stable for large atoms using PySCF's `aug_etb` auxiliary basis. In general we would not recommend running the DM21 family on atoms beyond Kr since the training data does not contain any heavy atoms or pseudopotentials. Nevertheless, we were surprised to discover that DM21 has good performance on these 62 systems in GMTKN55 that use unseen heavy atoms, and scores on the full GMTKN55 sets are shown in the main text and table S6.

	DM21	DM21m	DM21mc	DM21mu	revPBE:D3$_{BJ}$	MN15L:D3$_0$	SCAN:D3$_0$	PW6B95:D3$_0$	B3LYP:D3$_{BJ}$	M06-2X:D3$_0$	ωB97X-V	DSD-PBEP86:D3$_{BJ}$
AL2X6	0.82	0.76	1.64	0.89	2.07	1.35	1.95	0.76	2.71	0.90	1.21	0.31
ALK8	3.42	1.82	4.08	1.90	3.60	3.23	3.12	3.04	2.48	2.31	0.95	2.28
ALKBDE10	4.22	2.60	2.87	2.66	5.16	4.48	19.21	4.05	4.40	4.79	4.07	3.01
BH76RC	1.22	0.64	0.85	0.49	2.76	2.43	3.38	1.48	2.25	1.18	1.81	0.86
DC13	4.23	2.50	4.39	3.05	8.87	8.25	7.30	6.75	10.14	7.08	6.29	2.97
DIPCS10	2.07	1.55	1.27	2.23	4.81	10.46	4.92	2.70	4.73	3.16	4.10	3.90
FH51	1.04	0.81	0.73	0.80	3.34	2.55	2.69	1.57	2.61	1.20	2.30	0.87
G21EA	1.34	1.24	1.28	1.31	2.75	2.40	3.64	1.28	1.91	1.76	1.84	1.50
G21IP	1.81	0.81	1.14	1.02	4.20	3.46	4.69	2.77	3.55	2.64	2.96	2.10
G2RC	1.38	0.84	0.98	0.88	6.16	6.73	6.29	3.01	2.73	1.92	3.91	1.83
HEAVYSB11	3.42	3.15	1.44	3.40	2.72	6.47	6.79	2.09	3.30	8.16	1.39	0.92
NBPRC	1.64	1.97	0.96	2.25	1.98	1.93	2.28	1.76	2.00	0.95	1.43	0.88
PA26	1.74	1.25	1.26	1.47	4.73	2.25	3.14	2.53	2.87	1.23	2.65	1.10
RC21	1.02	1.16	1.53	1.43	4.85	2.00	6.53	2.91	2.44	1.63	3.53	1.77
SIE4x4	4.92	8.43	5.20	7.64	23.43	10.99	17.99	15.43	18.06	8.67	11.49	5.04
TAUT15	0.45	0.41	0.68	0.43	1.55	0.70	1.72	0.87	1.16	0.77	0.72	0.46
W4-11	3.04	0.52	0.80	0.63	7.57	3.41	4.02	2.42	3.40	3.16	2.78	2.72
YBDE18	2.84	2.72	2.07	2.89	4.41	4.20	3.36	3.29	4.72	2.40	2.03	1.53
BSR36	0.51	0.51	1.10	0.53	1.80	3.55	1.48	3.51	3.35	2.48	2.11	1.36
C60ISO	11.35	2.49	6.59	1.42	9.82	5.94	6.05	1.59	2.22	6.88	13.74	7.56
CDIE20	0.35	0.34	0.51	0.33	1.50	1.78	1.47	1.07	1.00	0.54	0.63	0.47
DARC	1.04	1.81	2.08	1.54	3.71	2.78	2.13	3.75	8.03	2.16	4.31	1.32
ISO34	0.58	0.60	0.49	0.55	1.50	1.88	1.30	1.27	1.78	1.23	1.17	0.41
ISOL24	1.56	2.06	1.44	1.76	4.56	3.55	3.34	3.80	5.80	2.74	2.98	1.12
MB16-43	6.65	5.35	5.79	4.46	27.11	20.42	16.56	8.04	24.84	15.68	32.51	6.46
PArel	0.84	0.77	0.74	0.89	1.53	2.19	1.48	1.00	1.18	0.97	0.63	0.50
RSE43	1.45	0.86	0.62	0.51	2.31	1.23	1.29	2.04	1.72	0.63	0.98	0.90
BH76	2.08	2.09	1.90	2.32	8.32	1.81	7.71	4.06	5.70	2.34	1.83	1.28
BHDIV10	1.34	1.72	1.99	1.99	7.83	2.08	6.51	2.43	3.22	1.04	0.85	1.71
BHPERI	0.77	2.12	1.78	1.74	6.29	1.78	5.17	0.98	1.18	1.35	2.07	2.45
BHROT27	0.19	0.38	0.20	0.32	0.37	0.87	0.82	0.54	0.41	0.36	0.31	0.21
INV24	0.90	0.84	0.73	1.34	2.18	2.02	1.16	1.17	1.05	1.28	1.22	0.75
PX13	0.90	2.57	1.07	2.41	8.75	6.38	8.23	1.33	4.33	5.32	2.56	2.51
WCPT18	1.57	1.30	1.98	1.25	7.22	1.81	6.12	1.44	2.28	1.88	1.71	1.77
ADIM6	0.11	0.13	0.13	0.07	0.25	3.88	0.04	0.55	0.11	0.27	0.16	0.06
AHB21	0.24	0.25	0.32	0.21	1.04	2.29	1.61	0.34	0.33	0.95	0.34	0.43
CARBHB12	0.51	0.61	0.53	0.48	1.10	1.21	1.35	0.43	0.88	0.25	0.33	0.61
CHB6	1.06	1.24	1.16	1.40	0.90	0.64	0.44	1.26	1.41	1.42	0.87	1.12
HAL59	0.58	0.40	0.45	0.43	0.72	0.59	0.99	0.32	0.57	0.35	0.30	0.44
HEAVY28	0.17	0.21	0.29	0.24	0.29	0.58	0.32	0.10	0.34	0.33	0.18	0.18
IL16	0.49	0.78	0.84	0.70	0.77	2.40	0.91	0.90	0.76	0.47	1.02	0.23
PNICO23	0.17	0.27	0.27	0.33	0.88	0.40	0.98	0.29	0.48	0.29	0.19	0.40
RG18	0.57	0.33	0.33	0.62	0.09	0.17	0.24	0.24	0.13	0.23	0.10	0.15
S22	0.20	0.30	0.28	0.21	0.43	1.82	0.44	0.35	0.30	0.34	0.22	0.31
S66	0.18	0.19	0.22	0.17	0.28	1.66	0.45	0.22	0.26	0.22	0.12	0.21
WATER27	1.96	1.45	2.60	1.37	3.51	12.00	10.65	0.97	4.07	3.70	1.30	2.26
ACONF	0.13	0.31	0.22	0.17	0.09	0.69	0.15	0.15	0.05	0.27	0.03	0.25
Amino20x4	0.19	0.26	0.22	0.24	0.37	0.92	0.22	0.29	0.21	0.30	0.19	0.13
BUT14DIOL	0.17	0.17	0.35	0.13	0.31	1.10	0.40	0.35	0.31	0.13	0.04	0.05
ICONF	0.18	0.22	0.19	0.19	0.32	0.53	0.31	0.28	0.29	0.32	0.26	0.14
IDISP	1.37	2.34	1.99	1.97	3.14	7.54	2.45	3.46	3.57	2.07	2.59	1.43
MCONF	0.22	0.59	0.28	0.28	0.44	1.60	0.47	0.39	0.22	0.55	0.24	0.38
PCONF21	0.27	0.26	0.44	0.34	0.87	4.10	0.49	0.55	0.53	1.09	0.30	0.27
SCONF	0.36	0.43	0.41	0.16	0.51	0.92	0.60	0.23	0.30	0.26	0.15	0.13
UPU23	0.50	0.48	0.44	0.34	0.47	1.67	0.44	0.54	0.61	0.50	0.59	0.38
MOM	1.50	1.28	1.35	1.25	3.76	3.35	3.6	1.98	2.9	2.09	2.45	1.35
WTMAD-1	1.99	2.15	2.14	2.04	4.67	6.77	4.64	2.99	3.59	2.77	2.31	1.81
WTMAD-2	3.97	3.88	3.95	3.95	8.29	11.49	8.03	5.49	6.36	4.91	3.92	3.16

Table S6: Mean absolute errors for all DM21 functionals and selected traditional functionals on each of the 55 GMTKN55 subsets considered in the main text.

Additionally, to justify our use of D3 corrections, we present results for functionals DM21:no-D3 in the m, mc and mcs cases, which were trained and evaluated without D3 correction on the energy labels. Excluding the D3 correction reduces performance on WTMAD-1,2 as expected, and can also reduce performance on the MoM as the functional is disrupted attempting to explain the training data using features that do not capture the required long range behaviour. Note that these evaluations to justify D3 were performed in preliminary experiments which did exclude the 62 GMTKN55 reactions containing atoms larger than Kr. We include ωB97X-V in Table S7 as a reference for the strongest traditional hybrid functional on this reduced subset of GMTKN55.

	GMTKN55($\leq$Kr)		
	MoM	WTMAD-1	WTMAD-2
ωB97X-V	2.29	2.34	3.89
DM21	1.50	1.96	3.81
DM21m	1.27	2.11	3.76
DM21mc	1.36	2.09	3.70
DM21:no-D3	1.42	2.70	5.33
DM21m:no-D3	1.48	2.83	5.48
DM21mc:no-D3	1.50	2.88	4.89

Table S7: GMTKN55 Performance of DM21 functionals trained with and without D3 corrections evaluated on the subset of 1443 GMTKN55 reactions containing atoms up to Kr.

S8.3 QM9

In table S8 we provide mean absolute errors of isomerisation energies versus the G4(MP2) results from Ref. (*38*) across the full QM9 benchmark for all the functionals that we ran. This list includes the best performing functionals on GMTKN55 from Fig. 4(b), and covers ωB97X(-V), M06-2X, TPSS (*73*) and TPSSh (*74*) which were highlighted in Ref. (*13*).

Functional	no-D3	D3	Functional	no-D3	D3
DM21m		1.33	DM21mu		**1.01**
DM21mc		1.36	DM21		1.66
ωB97X	2.12	2.24*	PBE	3.22	
M08-HX	2.17		M05-2X	3.29	
M06-2X	2.20	2.21	SCAN	3.57	3.62
PW6B95	2.52	2.24	TPSS	3.66	
MPW1B95	2.93		N12	3.98	
wB97	3.08		B3LYP	4.43	3.94
MN15-L	3.13		revTPSS	4.59	
TPSSh	3.15		BLYP	5.57	
CAMB3LYP	3.20				

Table S8: MAE on the QM9 benchmark for all functionals tested. We ran D3(0) corrections for selected traditional functionals with no consistent improvement seen on this dataset. * indicates the ωB97X-V functional with VV10 long range correction rather than D3 corrected ωB97X.

References and Notes

1. P. Hohenberg, W. Kohn, Inhomogeneous electron gas. *Phys. Rev.* **136** (3B), B864–B871 (1964).

2. W. Kohn, L. J. Sham, Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).

3. R. Van Noorden, B. Maher, R. Nuzzo, The top 100 papers. *Nature* **514**, 550–553 (2014).

4. A. J. Cohen, P. Mori-Sánchez, W. Yang, Insights into current limitations of density functional theory. *Science* **321**, 792–794 (2008).

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

6. C. Li, X. Zheng, N. Q. Su, W. Yang, Localized orbital scaling correction for systematic elimination of delocalization error in density functional approximations. *Natl. Sci. Rev.* **5**, 203–215 (2018).

7. N. Q. Su, C. Li, W. Yang, Describing strong correlation with fractional-spin correction in density functional theory. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9678–9683 (2018).

8. R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, Big data meets quantum chemistry approximations: The Δ-machine learning approach. *J. Chem. Theory Comput.* **11**, 2087–2096 (2015).

9. O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, K.-R. Müller, Machine learning force fields. *Chem. Rev.* **121**, 10142–10186 (2021).

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

11. L. Li, J. C. Snyder, I. M. Pelaschier, J. Huang, U.-N. Niranjan, P. Duncan, M. Rupp, K.-R. Müller, K. Burke, Understanding machine-learned density functionals. *Int. J. Quantum Chem.* **116**, 819–833 (2016).

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

13. A. V. Sinitskiy, V. S. Pande, Physical machine learning outperforms "human learning" in quantum chemistry. arXiv:1908.00971 [physics.chem-ph] (2019).

14. K. Ryczko, D. A. Strubbe, I. Tamblyn, Deep learning and density-functional theory. *Phys. Rev. A* **100**, 022512 (2019).

15. Y. Chen, L. Zhang, H. Wang, E. Weinan, DeePKS: A comprehensive data-driven approach towards chemically accurate density functional theory. *J. Chem. Theory Comput.* **17**, 170–181 (2021).

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

17. J. P. Perdew, R. G. Parr, M. Levy, J. L. Balduz Jr., Density-functional theory for fractional particle number: Derivative discontinuities of the energy. *Phys. Rev. Lett.* **49**, 1691–1694 (1982).

18. A. J. Cohen, P. Mori-Sánchez, W. Yang, Fractional spins and static correlation error in density functional theory. *J. Chem. Phys.* **129**, 121104 (2008).

19. Q. Sun, T. C. Berkelbach, N. S. Blunt, G. H. Booth, S. Guo, Z. Li, J. Liu, J. D. McClain, E. R. Sayfutyarova, S. Sharma, S. Wouters, G. K.-L. Chan, PySCF: The Python-based simulations of chemistry framework. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **8**, e1340 (2018).

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

21. A. Karton, N. Sylvetsky, J. M. L. Martin, W4-17: A diverse and high-confidence dataset of atomization energies for benchmarking high-level electronic structure methods. *J. Comput. Chem.* **38**, 2063–2075 (2017).

22. A. Karton, S. Daon, J. M. L. Martin, W4-11: A high-confidence benchmark dataset for computational thermochemistry derived from first-principles W4 data. *Chem. Phys. Lett.* **510**, 165–178 (2011).

23. A. Karton, A. Tarnopolsky, J. M.L. Martin, Atomization energies of the carbon clusters C_n ($n = 2-10$) revisited by means of W4 theory as well as density functional, G *n*, and CBS methods. *Mol. Phys.* **107**, 977–990 (2009).

24. B. Brauer, M. K. Kesharwani, S. Kozuch, J. M. L. Martin, The S66x8 benchmark for noncovalent interactions revisited: Explicitly correlated ab initio methods and density functional theory. *Phys. Chem. Chem. Phys.* **18**, 20905–20925 (2016).

25. M. K. Kesharwani, D. Manna, N. Sylvetsky, J. M. L. Martin, The X40×10 halogen bonding benchmark revisited: Surprising importance of (*n*-1)d subvalence correlation. *J. Phys. Chem. A* **122**, 2184–2197 (2018).

26. L. Li, S. Hoyer, R. Pederson, R. Sun, E. D. Cubuk, P. Riley, K. Burke, Kohn-Sham equations as regularizer: building prior knowledge into machine-learned physics. arXiv:2009.08551 [physics.comp-ph] (2020).

27. Y. Zhao, D. G. Truhlar, The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **120**, 215–241 (2008).

28. J.-D. Chai, M. Head-Gordon, Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* **128**, 084106 (2008).

29. A. J. Cohen, P. Mori-Sánchez, W. Yang, Challenges for density functional theory. *Chem. Rev.* **112**, 289–320 (2012).

30. M. Fuchs, Y.-M. Niquet, X. Gonze, K. Burke, Describing static correlation in bond dissociation by Kohn-Sham density functional theory. *J. Chem. Phys.* **122**, 094116 (2005).

31. A. D. Becke, *Density Functionals*, E. R. Johnson, Ed. (Springer, 2014), vol. 365, pp. 175–186.

32. J. C. Genereux, J. K. Barton, Mechanisms for DNA charge transport. *Chem. Rev.* **110**, 1642–1662 (2010).

33. M. Motta, D. M. Ceperley, G. K.-L. Chan, J. A. Gomez, E. Gull, S. Guo, C. A. Jiménez-Hoyos, T. N. Lan, J. Li, F. Ma, A. J. Millis, N. V. Prokof'ev, U. Ray, G. E. Scuseria, S. Sorella, E. M. Stoudenmire, Q. Sun, I. S. Tupitsyn, S. R. White, D. Zgid, S. Zhang, Towards the solution of the many-electron problem in real materials: Equation of state of the hydrogen chain with state-of-the-art many-body methods. *Phys. Rev. X* **7**, 031059 (2017).

34. N. P. Bauman, J. Shen, P. Piecuch, Combining active-space coupled-cluster approaches with moment energy corrections via the CC(P;Q) methodology: Connected quadruple excitations. *Mol. Phys.* **115**, 2860–2891 (2017).

35. H. V. Pham, K. N. Houk, Diels-Alder reactions of allene with benzene and butadiene: Concerted, stepwise, and ambimodal transition states. *J. Org. Chem.* **79**, 8968–8976 (2014).

36. L. Goerigk, S. Grimme, A general database for main group thermochemistry, kinetics, and noncovalent interactions: Assessment of common and reparameterized (meta-)GGA density functionals. *J. Chem. Theory Comput.* **6**, 107–126 (2010).

37. R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022 (2014).

38. H. Kim, J. Y. Park, S. Choi, Energy refinement and analysis of structures in the QM9 database via a highly accurate quantum chemical method. *Sci. Data* **6**, 109 (2019).

39. Y. Zhang, C. Lane, J. W. Furness, B. Barbiellini, J. P. Perdew, R. S. Markiewicz, A. Bansil, J. Sun, Competing stripe and magnetic phases in the cuprates from first principles. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 68–72 (2020).

40. J. Kirkpatrick, B. McMorrow, D. H. P. Turban, A. L. Gaunt, J. S. Spencer, A. G. D. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau, L. Román Castellanos, S. Petersen, A. W. R. Nelson, P. Kohli, P. Mori-Sánchez, D. Hassabis, A. J. Cohen, Zenodo (2021); doi:10.5281/zenodo.5482370.

41. J. P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).

42. R. Berner, A. Lüchow, Isomerization of bicyclo[1.1.0]butane by means of the diffusion quantum Monte Carlo method. *J. Phys. Chem. A* **114**, 13222–13227 (2010).

43. L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, S. 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**, 32184–32215 (2017).

44. Y. Zhang, W. Yang, Comment on "Co-generalized gradient approximation made simple". *Phys. Rev. Lett.* **80**, 890 (1998).

45. Y. Zhao, D. G. Truhlar, Design of density functionals that are broadly accurate for thermochemistry, thermochemical kinetics, and nonbonded interactions. *J. Phys. Chem. A* **109**, 5656–5667 (2005).

46. S. Kozuch, J. M. L. Martin, DSD-PBEP86: In search of the best double-hybrid DFT with spin-component scaled MP2 and dispersion corrections. *Phys. Chem. Chem. Phys.* **13**, 20104–20107 (2011).

47. N. Mardirossian, M. Head-Gordon, ωB97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. *Phys. Chem. Chem. Phys.* **16**, 9904–9924 (2014). ω

48. S. Grimme, S. Ehrlich, L. Goerigk, Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32**, 1456–1465 (2011).

49. A. D. Becke, Density functionals for static, dynamical, and strong correlation. *J. Chem. Phys.* **138**, 074109 (2013).

50. J. P. Perdew, V. N. Staroverov, J. Tao, G. E. Scuseria, Density functional with full exact exchange, balanced nonlocality of correlation, and constraint satisfaction. *Phys. Rev. A* **78**, 052513 (2008).

51. M. Haasler, T. M. Maier, R. Grotjahn, S. Gückel, A. V. Arbuznikov, M. Kaupp, A local hybrid functional with wide applicability and good balance between (de)localization and leftâ€“right correlation. *J. Chem. Theory Comput.* **16**, 5645–5657 (2020).

52. A. Karton, D. Gruzman, J. M. L. Martin, Benchmark thermochemistry of the $C_{(n)}H_{(2n+2)}$ alkane isomers (n = 2-8) and performance of DFT and composite ab initio methods for dispersion-driven isomeric equilibria. *J. Phys. Chem. A* **113**, 8434–8447 (2009).

53. M. K. Kesharwani, A. Karton, N. Sylvetsky, J. M. L. Martin, The S66 non-covalent interactions benchmark reconsidered using explicitly correlated methods near the basis set limit. *Aust. J. Chem.* **71**, 238 (2018).

54. L. A. Curtiss, K. Raghavachari, P. C. Redfern, J. A. Pople, Assessment of Gaussian-3 and density functional theories for a larger experimental test set. *J. Chem. Phys.* **112**, 7374–7383 (2000).

55. A. Karton, J. M. L. Martin, Comment on:“Estimating the Hartree-Fock limit from finite basis set calculations”. *Theor. Chem. Acc.* **115**, 330–333 (2006).

56. Q. Wu, W. T. Yang, A direct optimization method for calculating density functionals and exchange-correlation potentials from electron densities. *J. Chem. Phys.* **118**, 2498 (2003).

57. H. Bahmann, M. Kaupp, Efficient self-consistent implementation of local hybrid functionals. *J. Chem. Theory Comput.* **11**, 1540–1548 (2015).

58. T. Gould, ‘Diet GMTKN55’ offers accelerated benchmarking through a representative subset approach. *Phys. Chem. Chem. Phys.* **20**, 27735–27739 (2018).

59. Z. Wen, W. Yin, A feasible method for optimization with orthogonality constraints. *Math. Program.* **142**, 397–434 (2013).

60. Y. Tawada, T. Tsuneda, S. Yanagisawa, T. Yanai, K. Hirao, A long-range-corrected time-dependent density functional theory. *J. Chem. Phys.* **120**, 8425–8433 (2004).

61. M. Korth, S. Grimme, "Mindless" DFT benchmarking. *J. Chem. Theory Comput.* **5**, 993–1003 (2009).

62. M. G. Medvedev, I. S. Bushmarinov, J. Sun, J. P. Perdew, K. A. Lyssenko, Density functional theory is straying from the path toward the exact functional. *Science* **355**, 49–52 (2017).

63. D. Hait, M. Head-Gordon, How accurate is density functional theory at predicting dipole moments? An assessment using a new database of 200 benchmark values. *J. Chem. Theory Comput.* **14**, 1969–1981 (2018).

64. A. Fabrizio, B. Meyer, C. Corminboeuf, Machine learning models of the energy curvature vs particle number for optimal tuning of long-range corrected functionals. *J. Chem. Phys.* **152**, 154103 (2020).

65. J. Shen, P. Piecuch, Combining active-space coupled-cluster methods with moment energy corrections via the CC(P;Q) methodology, with benchmark calculations for biradical transition states. *J. Chem. Phys.* **136**, 144104 (2012).

66. A. Ajaz, A. Z. Bradley, R. C. Burrell, W. H. H. Li, K. J. Daoust, L. B. Bovee, K. J. DiRico, R. P. Johnson, Concerted vs stepwise mechanisms in dehydro-Diels-Alder reactions. *J. Org. Chem.* **76**, 9320–9328 (2011).

67. C. A. Grambow, L. Pattanaik, W. H. Green, Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Sci. Data* **7**, 137 (2020).

68. D. Pfau, J. S. Spencer, A. G. Matthews, W. M. C. Foulkes, Ab initio solution of the many-electron Schrödinger equation with deep neural networks. *Phys. Rev. Res.* **2**, 033429 (2020).

69. J. S. Spencer, D. Pfau, A. Botev, W. Foulkes, Better, faster fermionic neural networks. arXiv:2011.07125 [physics.comp-ph] (2020).

70. J. Martens, R. Grosse, Optimizing neural networks with kronecker-factored approximate curvature, in *ICML Proceedings* (ICML, 2015), pp. 2408–2417.

71. H. Flyvbjerg, H. G. Petersen, Error estimates on averages of correlated data. *J. Chem. Phys.* **91**, 461–466 (1989).

72. J. Tao, J. P. Perdew, V. N. Staroverov, G. E. Scuseria, Climbing the density functional ladder: Nonempirical meta-generalized gradient approximation designed for molecules and solids. *Phys. Rev. Lett.* **91**, 146401 (2003).

73. V. N. Staroverov, G. E. Scuseria, J. Tao, J. P. Perdew, Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes. *J. Chem. Phys.* **119**, 12129–12137 (2003).