PHYSDFTBENCH: WHICH EXACT CONDITIONS SURVIVE FROM ARCHITECTURE TO RELEASED NEURAL EXCHANGE–CORRELATION FUNCTIONALS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Whether released neural functionals inherit the exact conditions hand-built functionals secure by construction is an empirical question. We find no third-party audit of these releases: commonly cited audits evaluate no neural network. **PhysDFTBench** audits seven exact-condition families under declared protocols and a sealed coverage rule. It reads PBE, DM21's four released weights, Skala-1.1, three controls, our own correction and pcNN in part. **Deployed identities fail in released weights.** DM21 breaks spin permutation and the uniform-gas limit, and the fractional-charge tail and midpoint readings order the columns differently. For the conditions no deployed construction guarantees, which ones survive is a property of the trained weights, not of the architecture. DM21's supplementary material describes strict and approximate spin-symmetrisation routes. We test the released implementation: a reference-free test puts all four released DM21 weights outside their own round-off *in the deployed energy*, by 0.02–$2.2\,\mathrm{mHa}$. A 5,123-parameter neural correction is our controlled testbed. A loss term written directly on a criterion quantity need not reach that criterion even on its own probe. We also give a three-class rule for omitting the SCF response. In the unprotected class the omitted term need not vanish, and in our own loss it did not.

1 INTRODUCTION

Exact conditions are the load-bearing structure of density-functional approximations. They include the uniform-gas limit, the Lieb–Oxford bound (Lieb & Oxford, 1981; Lewin et al., 2022), piecewise linearity in electron number (Perdew et al., 1982), and the $-1/r$ decay of the exchange–correlation potential. A hand-built functional secures some of them *by construction*. For example, SCAN is presented as satisfying all seventeen conditions a meta-GGA can (Sun et al., 2015), a list excluding the last two of these four. A neural functional is different. Either its deployed construction guarantees compliance, or compliance is left to 10^5–10^6 weights fitted to energies. Absent such a guarantee, whether a condition survives the fit is an empirical question about the released weights and their deployment implementation, not a property of the architecture. Neither a shared architecture nor a list of training constraints establishes compliance. Existing audits of analytic functionals do not substitute for that measurement. Kanungo et al. (2025) write that "many empirical functionals…still satisfy them, at least in the regime of chemically relevant densities". The two audits cited (Pederson & Burke, 2023; Khanna et al., 2025) measure eight and nine *analytic* functionals in their main tables, hundreds more in the first's supplement, none a neural network. For released neural weights the closest reading is DM21 on hydrogenic ions (Schwalbe et al., 2022), and it grades total-energy error, not the conditions. Our own search found no third-party measurement of these conditions on these released weights (Appendix G).

Seven exact-condition families (M1–M7, Table A2, Appendix A) become criterion quantities on one benchmark. Four functionals are measured on it: PBE (Perdew et al., 1996), DM21 (Kirkpatrick et al., 2021), Skala-1.1 (Luise et al., 2025) and a 5,123-parameter neural correction of our own. It is a controlled testbed rather than a proposed state of the art: we use it to separate what architecture can guarantee from what only training can reach. Three hand-built controls (PBE0, LC-ωPBE, SCAN) join them later. **An exact condition can be violated in three ways.** It may be represented but not enforced, differentiated with an incomplete gradient, or reached on its training probe and not

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.

on the released weight. The audit measures the last. Our own correction intervenes on the first two. Each stage has its own literature, joined here on one set of released artefacts. The failures have a pattern, and it is not simply that construction wins. Two of our own rows are construction identities (M6 before regularisation), and they hold as deployed. pcNN's implied spin symmetry (Nagai et al., 2022) holds too. Its fully spin-polarised uniform-gas limit does not. What a construction buys is therefore only as good as the identity it rests on, and what no construction carries must be measured. M1 to M3 are not a neural weakness: only LC-ωPBE is inside the released gate on M1 and on M3 at $r{=}5$. The answer favours neither side. The conditions are not inherited. And yet a neural correction *can* hold several of them, two by construction; of the three in its loss, one misses its band, two are unproved (our 5,123-parameter correction, `W-13-final` in the appendix tables). It proves neither that it holds M7 nor that it fails it, and misses its own held-out gate in every round.

Context of exact conditions. *On the hand-built side the conditions are the design.* They name a hierarchy (Perdew & Schmidt, 2001) built under "more constraint satisfaction with fewer fits" (Perdew et al., 2005), and Perdew (2026) asks that a fit to large data be constrained by them. The literature argues both directions (Kaplan et al., 2023; Mardirossian & Head-Gordon, 2017; Medvedev et al., 2017), so for a neural functional the question is empirical, not a corollary. *On the neural side they enter as training data or loss terms.* DM21 trains on fractionally charged and fractional-spin systems, Skala's paper asks whether they can emerge from data (Kirkpatrick et al., 2021; Luise et al., 2025), and newer entries add detail from inside (Medvidović et al., 2025; Li et al., 2024). *Neither route is a third-party audit of the released weights* (Figure 1). *What exists already is adjacent to that question, not it.* A Comment and a Response in *Science* dispute DM21's generalisation beyond its training set (Gerasimov et al., 2022; Kirkpatrick et al., 2022). Many-condition audits on one benchmark (Pederson & Burke, 2023; Khanna et al., 2025) overlap our set only at M5 in its local form. An agentic search calls "explicitly enforced constraints" essential (Duston et al., 2026); conditions as losses ship as a library (Casares et al., 2024); and learned functionals are a line of their own (Snyder et al., 2012; Dick & Fernandez-Serra, 2020; Li et al., 2021; Kasim & Vinko, 2021; Kanungo et al., 2025; Navarro-Rodríguez et al., 2026; Eberhard et al., 2026).

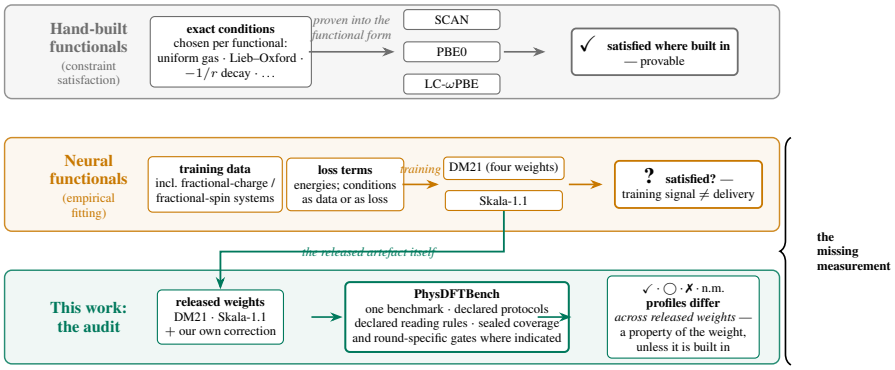

Figure 1: **Exact conditions reach a hand-built functional by construction, a neural one through its training, except where its deployed construction guarantees them.** *Schematic: not a measurement.*

Contributions.

- *A benchmark, and the third-party audit run with it.* PhysDFTBench grades *released neural weights* against seven exact-condition families at declared protocols, for the conditions a deployed construction does not guarantee. Its code, sealed point set and pre-registrations ship with the paper. Four weights of one architecture do not share one profile (Section 3).
- *A reference-free deployment test.* Because M4 is an identity, two energy evaluations through a released functional's own deployment interface put all four DM21 weights outside their own round-off, with no benchmark, no reference energy and no retraining.
- *A training selection rule.* The standard envelope theorem decides which exact-condition losses may omit the SCF response, and we audit its six premises one by one. A violation of that rule sat undetected, and symptomless, in an earlier version of our own loss.
- *A construction-based case study.* Our 5,123-parameter Δ-on-PBE correction enforces spin symmetry and the uniform-gas limit analytically, and its controls separate structural guarantees from readings under this protocol (Sections 4 and 6).

Scope of the benchmark. A missing condition is a generalisation gap wherever it would have bound, not a worse benchmark number. The audit needs no retraining. pcNN (Nagai et al., 2022) and CIDER24Xe (Bystrom et al., 2024) attach through its plugin route, and the coverage gate refuses CIDER24Xe's thermochemistry. *Which* of their rows come back live is a property of the pair (benchmark, functional).

2 BENCHMARK SUITE AND EVALUATION PROTOCOL

Each exact condition is evaluated with a *criterion* quantity, an evaluation protocol and a decision rule. The protocol fixes the basis, grid, point set and any reference density, and on the rows it reads at aug-cc-pVDZ, the movement to aug-cc-pVTZ is the yardstick an ordering has to beat. We evaluate individual models, our own checkpoints and library functionals, under these fixed rules, one per table column.

From a checklist to numbers with a bound. Table A2 has all seven, with thresholds, derivations and code pins. One benchmark means one basis and grid *per row*, and one common thermochemistry point set whose deviations are recomputed from raw energy columns and not copied from derived files. It also means one shared PBE reference density, for M5's cell *alone*. Three protocol differences survive (Table G1, notes).

Certificates and one-sided verdicts. A self-consistent solution bounds the energy from above and does not certify that the lowest solution was found, so criteria that need an optimiser yield one-sided verdicts. On the fragment-interaction error at $R = 6\,a_0$, conditional on Appendix A's premise, a value below -1 mHa *proves* failure while one inside it proves nothing (Figure 5). PBE's $+0.55$, read at a converged aufbau-constrained point, is not a pass, and neither is our own -0.0007. For M7 we therefore attach a convergence-independent *certificate*: a trial density evaluated with no SCF, whose energy bounds the functional's own minimum from above. A reading above that bound fails the criterion's *validity gate*, which caught our own largest error on an earlier model (Appendices A and C).

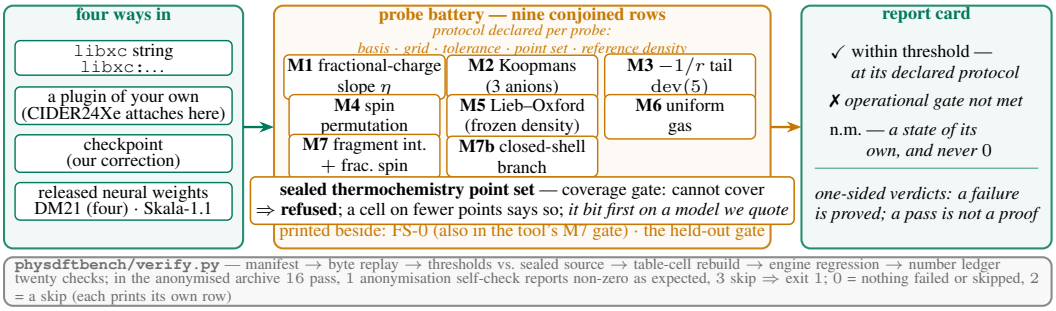

Figure 2: **The released benchmark, end to end: four ways to attach a functional, nine conjoined rows with a declared protocol each, and a sealed thermochemistry set behind a coverage gate.** Support: Table H2; rows and gates: Table H1; digests, protocol and the `verify.py` chain: Appendix H.

PhysDFTBench: the benchmark. It grades a functional against its sealed versions of these criteria with one command. The routes in, and the nine rows it grades over, are in Figure 2. Those rows include M7's closed-shell branch and thermochemistry, and FS-0 and our held-out gate are printed separately. *The criterion this paper reads on M7 is the one-sided d reading.* FS-0 is the lowest energy any small spin fraction reaches, measured from the integer state. It sits beside that reading as a lesion diagnostic, and the released tool conjoins the two and $\lambda_{\min}$ in its own M7 gate. (i) "Not measured" (*n.m.*) is a state of its own and never 0, never "not applicable". Every non-live cell names, in code, the measurement or refusal behind it. (ii) The thermochemistry point set is *sealed*, and a *coverage gate* holds it: a model that cannot cover that set is refused, and a cell on fewer points says so. One declared exception is applied to *every* model at once (Appendix I, graded that way by a different benchmark). Rule (ii) bound first on *a model we ourselves quote*. (iii) Whether a criterion can be read is a property of the pair (benchmark, functional). Skala-1.1's and DM21's M7 cells read *n.m.* for two different reasons, and neither reason means "no symmetry-broken solution exists". M3, M5 and M7 carry no ✓, for a different reason each, and the reading rule is why. M3 reaches its exact value only as $r \to \infty$. M5 scores distance from the $N{=}1$ equality, not violation

Table 1: **Seven exact-condition families, seven functionals, one benchmark**: the compact form of Table G1 (Appendix G), carrying the reading rules and protocol notes; row M4 excepted. ✓/✗ declared threshold met/not met, ◯ inside tolerance, not proved; n.m. not measured — never "not applicable"; no cell holds an estimate. The audit reports every cell, including the ones we lose. M5's exact value is 0. *M1 and M3 carry no glyph: they are readings, not verdicts* (Table G1: the glyph rule, M1's fit-window systematic, up to $+11.8$ pp, against a 2.5 pp gate, and LC-ωPBE inside M3's r=5 gate). *Four n.m. cells, three kinds*: † the run aborted in a library, not the functional; § the symmetry-broken descent *measured* to stall above that model's floor; ¶ the same stall *predicted* there, not reached. The last two carry a certificate, in brackets. ‡ *Read by a protocol other than the released tool's*, which returns *n.m.* on this row: the $57\times$ is a declared limit (Appendix H).

		audited				*hand-built*		
	condition **criterion quantity**	**PBE**	**DM21** (1 of 4)	**Skala**	**Ours**	**PBE0**	**LC-ωPBE**	**SCAN**
M4	spin permutation $\max \lvert\Delta E_{\mathrm{swap}}\rvert$ [mHa], exact 0	✓0	✗2.17	✓1.4e-11	✓0	✓0	✓0	✓1.4e-11
M6	uniform-gas limit $\max \lvert\Delta\varepsilon_{\mathrm{xc}}\rvert$ / our 0.2 mHa/e	✓1e-06×	✗124×	✗57×‡	✓8e-09×	✓3e-10×	✗1.12×	✓1e-06×
M1	fractional-charge tail η [%], exact 100, here 100.03	0.005	81.9	0.005	96.5	25.0	100.0	n.m.†
	$\delta(\frac{1}{2})$ [mHa], exact 0	−53.22	**−1.04**	−47.63	−13.30	−39.28	−12.11	n.m.†
M5	Lieb–Oxford at N=1, H∧He$^+$ $\lambda(\mathrm{He}^+)/\lambda_1 - 99.23$ [pp]	◯−0.87	◯+0.17	◯−0.00	◯+0.33	◯−0.40	✗+0.06	◯+0.13
M2	Anion-HOMO sign $\#\{\varepsilon_{\mathrm{HOMO}}(\mathrm{X}^-) < 0\}$, of 3	✗0/3	✗1/3	✗0/3	✓3/3	✗1/3	✓3/3	✗0/3
M3	$-1/r$ at finite r, tool gate r=5 $\lvert r\, w_{\mathrm{xc}}(3) + 1\rvert$, 0 as $r \to \infty$	0.553	0.369	0.462	0.229	0.426	0.141	0.517
M7	fragment interaction d [mHa], window $\lvert d\rvert \le 1$	◯+0.55	n.m.¶ ($\le$+2.76)	n.m.§ ($\le$+3.63)	◯−0.0007	◯+0.60	◯+0.61	◯−0.45

of the inequality: only the LC-ωPBE column exceeds the bound, and not exceeding it decides little. M7 is read one-sidedly, as above. Appendix H adds its caveats.

3 An architecture is not a warranty

All four released DM21 weights break the spin-permutation identity *in the deployed energy*. Both reads go through one released interface on one fixed density pair. No SCF runs between them. The energy moves by 0.02–2.2 mHa under a spin swap, and by 3.8 mHa on an over-occupied stress pair. PBE, Skala-1.1 and ours instead read only round-off. That floor is measured, not assumed. A second implementation, sharing the released graph, weights and densities, reproduces every reading (Appendix A). The test is one-sided. No finite set of agreements can establish the identity. At the SCF level that interface, run unmodified, gives $\max|D_\alpha - D_\beta|$ from 1.51×10^{-2} to 7.26×10^{-2} on a closed-shell anion. As a bare map, the released DM21 weight breaks the identity at nearly every grid point of H^- and F^- (Appendix A, with the downstream cost of a swap-symmetric wrapper). *The three readings are the deployed energy, the self-consistent density and the bare map, and none of them allows us to infer how DM21 was constructed.* An asymmetric map is compatible with an exactly symmetric deployed functional, and our own is one. Skala-1.1's fractional-charge linearity is at PBE's level, and both leave essentially the whole spurious tail. Its authors report a model that works "without relying on Hartree–Fock exchange" (Luise et al., 2025). With no long-range Fock term in the class, nothing cancels the $1/(4R)$ tail. However, the non-local layer reads nuclear coordinates, so it could reintroduce an R dependence through them, which we have not ruled out. Skala-1.1's own reported exchange fraction is another quantity again, not the criterion measured here (Appendix A). *So no warranty comes with the design.* What a function class can *represent* says nothing about what one set of weights *does*. A released checkpoint therefore has to be audited as an artefact.

To test this externally we evaluated one further released functional, pcNN (Nagai et al., 2022), after the benchmark had been frozen. Another group built it to impose asymptotic constraints *analytically*, and we registered its predictions before the run. Two of those constraints reach rows we can read, and we predicted both would hold. *We observe mixed results.* The spin-permutation invariance it implies holds, to round-off. The uniform gas does not. Evaluated on the same implementation used for Skala-1.1, pcNN reads $1.70\times$ our budget, rising with spin polarisation rather than switching on at it, while our sealed stage returns *no reading* on that row. Those inputs leave the constraint's feature surface. Of three untargeted readings two miss their gates, Koopmans and the $-1/r$ residual,

while FS-0 clears a lesion probe that is not a compliance verdict. *A feature-space anchor need not enforce the intended physical limit outside the manifold it anchors, and the untargeted criteria do not follow automatically* (Appendix K). **Condition profiles differ across released checkpoints.**

The four released DM21 weights (Kirkpatrick et al., 2021) share one architecture, and their training constraints differ in documented ways. They ship with an ablation table of what each was trained on. Table G2 reads the condition profile of each under the same audit protocols. *No released `DM21` weight takes more than one* of fractional-charge linearity, the anion-HOMO sign count for Koopmans and the uniform gas. The DM21 weight meets *none* of them. The two weights supervised on fractional charge do not deliver the criterion ($\eta = 81.9$ and 110.6 against 100.03), and the other rows are read out one by one there (Table G2). The contradicting case is in the same table. On M4 the four residuals bear no relation to the documented constraints, and we infer nothing causal from either half. *The ablation table shipped beside a weight states which constraints entered its training. It is not a certificate of which conditions that released weight satisfies.* The differences in Table G2 therefore motivate evaluating released checkpoints directly, and do not license attributing those differences causally to any single training choice.

4 LONG-RANGE EXCHANGE IS NOT ONE SCALAR

No single scalar summarises a learned functional's long-range exchange. Two diagnostics order the columns differently, and a functional design must explicitly specify the intended metric. PBE0, the long-range-corrected LC-ωPBE, and SCAN are read on the same benchmark under a rule fixed beforehand (Appendix G). LC-ωPBE (Vydrov & Scuseria, 2006; Iikura et al., 2001) tends to 100% exact exchange in the long-range limit by construction, at scale $1/\omega = 2.5\,a_0$. On the two cells the pre-registration put at issue, M2 and M3, it matches or beats us. It ties us on Koopmans and holds the smallest $-1/r$ residual, ahead of us by 0.0872 against a basis-set shift of 0.082 on the row. *That shift is measured, not a bound, and the differential basis convergence of this pair is not established*, so the ordering is a reading under the declared aug-cc-pVDZ protocol and not settled beyond it (Appendix G.3.2). *A 2006 functional with one range-separation parameter, fixed by its authors, holds by construction the long-range asymptotics a neural functional has to learn.* Its Koopmans $3/3$ and its $-1/r$ residual here are readings under this protocol, not further guarantees.

For a constant-coefficient hybrid the fractional-charge condition returns that same fraction: the benchmark recovers the column's built-in long-range exact-exchange coefficient, $\eta = a$ (Figure 3; Appendix A, note f). PBE0 (Adamo & Barone, 1999; Ernzerhof & Scuseria, 1999) measures its own mixing fraction to three digits, and LC-ωPBE measures 100.0%. *Our own reading here is not directly comparable*, because η is a term in our loss (14), Appendix D, on the same three separations. The $-1/r$ row carries a different two-point index, whose PBE0 residual is 0.06 at one column, not a transferable uncertainty (Appendix G). The two readings differ by that residual on the global hybrid and by 0.18–0.37 on *every* learned column. *Evaluated on the grid, the long-range coefficient is a field and not a number*: the DM21 column rises from nucleus to tail, ours dips and returns, and the two constant controls hold their constants at every shell (Appendix G.3.3).

Self-interaction error is similarly multifaceted. *The M1 condition splits, and we lose the second half.* The second half is $\delta(\frac{1}{2})$, the deviation from the straight line at half occupation, exact value 0. It reads linearity *at* that point rather than through the $1/R$ tail. On it the two fractional-charge DM21 weights sit ahead of every other measured column, ours included. No term of ours enters this reading of M1 (Tables 1 and G2, Figure G1b). Evaluating a different criterion yields yet another ranking. The closed-shell branch of the fragment-interaction probe orders the seven by long-range exchange fraction (Figure 5), and the tie at $a \approx 0$ already spans $12\,\mathrm{mHa}$. On this benchmark *ours is the worst of the three neural readings* (Figure G1a). All seven are readings, not proved minima. LC-ωPBE exceeds the one-electron Lieb–Oxford bound on H, though its printed He^+ value is inside, and it reads $1.12\times$ the uniform-gas budget (Appendices G.3.1 and G). SCAN (Sun et al., 2015) takes Koopmans $0/3$, and its constraint list contains neither Koopmans nor the potential's asymptote. On thermochemistry the controls beat us where long-range exchange helps, while `DM21`, partly in-sample on the two atomic subsets, beats all three on all three subsets and Skala-1.1, archived there, on two — one of them by 0.01 kcal/mol (Table G5).

Do these readings predict benchmark accuracy? Association varies in direction and size with the probe and the subset. Each comparison has only five to seven columns, and the suite attaches no

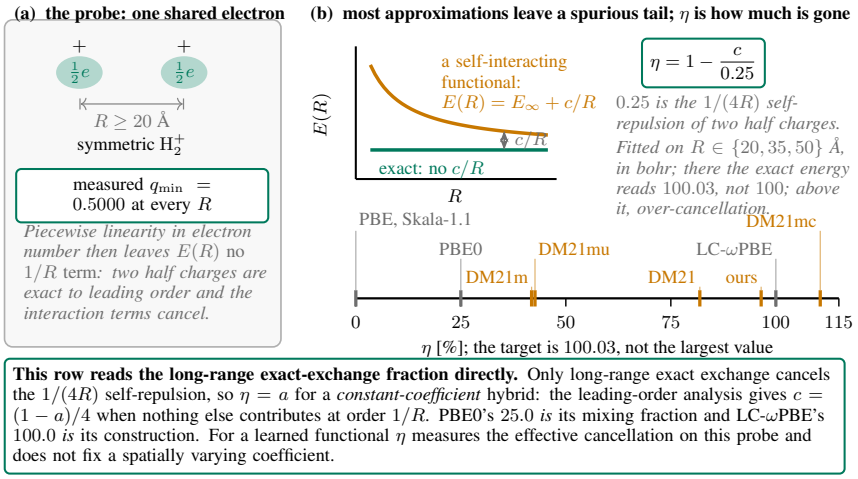

Figure 3: **What fractional-charge linearity measures, and why this row reads the long-range exact-exchange fraction** (Appendix A, note f). *Every number in (b) is a measurement (Tables G2 and G3).*

ordering verdict to any row. These data do not establish a robust relationship between condition readings and benchmark accuracy (Appendix G.6). That is a finding about the data, not the design, and not evidence the conditions are irrelevant.

5 How constraints should enter self-consistent learning

Optimisation in self-consistent training. A missing condition can be targeted through self-consistent training rather than guaranteed by construction, and that raises an optimisation question. Training through an SCF loop means differentiating through a fixed-point solver, where the exact gradient in general carries an SCF response term. A common shortcut invokes the envelope theorem to omit that response instead. While this shortcut is standard and not ours (Appendix B), what we add is an operational classification, six explicit proof obligations, and the numerical diagnostics behind them. Those diagnostics caught an incomplete gradient in our own earlier training implementation (the correction of Section 6). With the density matrix written as an orbital rotation κ of the converged solution D^*,

$$\frac{\mathrm{d}E}{\mathrm{d}\theta} = \underbrace{\frac{\partial E}{\partial \theta}}_{\text{explicit}} + \underbrace{\sum_{pq} \frac{\partial E}{\partial \kappa_{pq}} \frac{\mathrm{d}\kappa_{pq}}{\mathrm{d}\theta}}_{\text{response}}, \qquad \frac{\partial E}{\partial \kappa_{pq}} = (f_p - f_q) F_{pq}, \tag{1}$$

self-consistency *is* the statement that the occupied–virtual block of F vanishes, so every term of the response sum carries a zero factor. The shortcut rests on six premises, P1–P6 (Appendix B). The one easy to lose is P6, which requires that *the quantity being differentiated* be itself stationary at D^*.

Violation of premise P6. That loss contained an exact one-electron target, $E_{\mathrm{xc}}[n] = -E_{\mathrm{H}}[n]$ at $N = 1$ (Perdew & Zunger, 1981), imposed on the current parameters' self-consistent density. The differentiated $Q = E_{\mathrm{xc}}^{\theta} + E_{\mathrm{H}}$ is only part of the total energy, and $\partial Q/\partial \kappa \neq 0$. The response therefore survives, and finite differences put it at 0.8–1.1% of the explicit term on the two systems the term is imposed on, and 22% on a two-electron control where the identity does not hold (Appendix B). This violation occurred silently: the loss decreased, every SCF converged, the logs were clean.

Along a direction u orthogonal to both explicit gradients the implemented gradient is *identically zero* while the true derivative is not (Figure 4). The finding is not the subspace itself, but that the *true* gradient has a component outside it, and no step size attenuates or recovers it. Along the same u the total energy is flat to 10^{-12}, so the envelope theorem is intact and only the E_{xc} term is unprotected. Two limits bound it (Appendix B): u is a null direction of the one-electron term only, and that term is 4.3% of the *loss*, a share of the objective and not of the gradient. We do not convert it into a bound on the update. Where the premises hold, (1) verifies numerically, each premise audited to its own kind, measured or held by construction or unmonitored (Appendix B). $E_{\mathrm{xc}}[n] = -E_{\mathrm{H}}[n]$ holds for *any* one-electron density, so a fixed reference density kills the response term. Its price is that the condition is enforced at one density, not the one the functional generates. That moves the one-electron

target to class B from the tenth training round on. The other loss terms keep their own classifications. However, between refreshes the differentiated object is a frozen-density surrogate (Appendix D). The *repair itself* is still not isolated, because the first round to deploy it also changed the label set.

A selection rule. For a smooth objective $Q(\theta, \kappa^*(\theta))$ the term the shortcut omits in parameter direction i is $R_i = \langle \partial Q/\partial \kappa,\ \partial \kappa^*/\partial \theta_i \rangle$. The class of a term says which of two established guarantees, if either, makes R_i vanish. **(A) Protected:** the term is a smooth combination of *stationary SCF total energies*, so, subject to the premises, stationarity removes R_i term by term. **(B) Needs no protection:** the term is evaluated on a density independent of the differentiated θ, so the density response is identically zero. **(C) Unprotected:** neither guarantee has been established, so R_i need not vanish. Exits: argue it, measure it, move the term to (A) or (B), or solve the coupled-perturbed equations. A numerical check validates the inputs and tolerances it tests, not a global identity. Classes attach to *individual objective terms*, so one training loss may carry several. M6 carries no term in (14), Appendix D. It is held by construction. Of the terms that are there, the energy residuals, the fragment-interaction residual, the fractional-spin hinge and the $1/R$ coefficient are class A at a refresh, class B between refreshes; the chord residual is class B throughout (Appendix D). A loss on fixed molecular inputs is class B, one on stationary total energies class A, *whether ours or anyone's*. Skala-1.1 is the standing precedent: its supplement (Luise et al., 2025) states this shortcut, exact there as here wherever the differentiated solution is stationary.

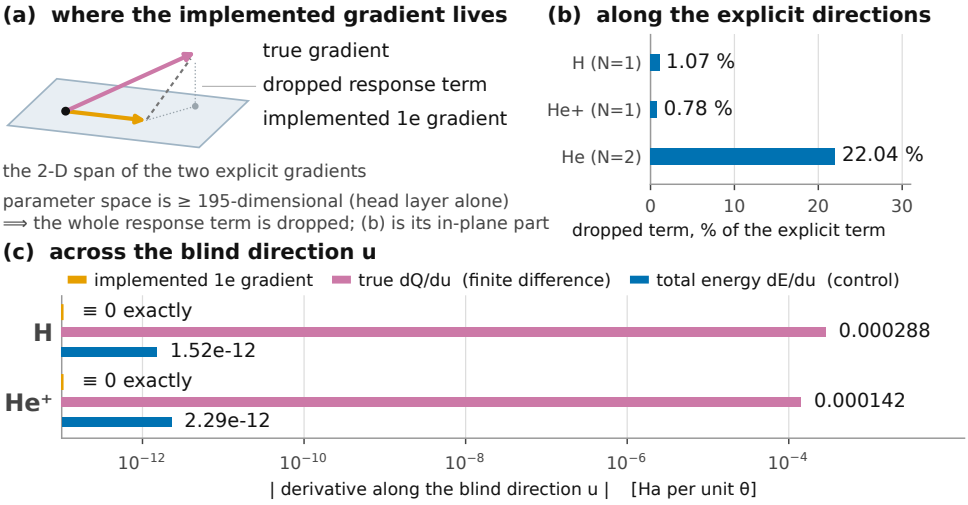

Figure 4: The one-electron target confines the implemented gradient to the plane of the two explicit gradients. The response part reaches outside it. **Panel by panel, with every number**: Appendix B.

6 Training on a condition does not secure it at deployment

This is an intervention, not a competing functional. *The shared ingredient* (Appendix G.3.3) is DM21's third energy density, $e_{\omega\mathrm{HF}}$ at $\omega = 0.4$, the long-range exact-exchange density that LC-ωPBE turns on at 100%. Our correction carries it too. However, the controls show that carrying it does not secure the measured condition profile (Table 1).

Our correction: the construction. It is additive on PBE, with no parameter outside the network:

$$E^{\theta}_{\mathrm{xc}}[\rho] = E^{\mathrm{PBE}}_{\mathrm{xc}}[\rho] + \int \sum_{i\le 3} f^{\theta}_i\big(x(\mathbf{r})\big)\, e_i(\mathbf{r})\, \mathrm{d}\mathbf{r}, \qquad f^{\theta}_i = S(x)\, \bar{f}^{\theta}_i(x), \tag{2}$$

where e_1 is LDA exchange and e_2, e_3 are *local* exact-exchange energy densities at $\omega = 0$ and $\omega = 0.4$. At each grid point the network reads eleven local quantities, and it outputs three dimensionless coefficients — the shape DM21's own head outputs — rather than $\varepsilon_{\mathrm{xc}}$ itself.

How a non-local quantity enters a pointwise network. The four exact-exchange channels are Hartree–Fock exchange *spread over grid points*, not integrated. Here $E^{\sigma}_b(\mathbf{r}) = \sum_a D^{\sigma}_{ab}\chi_a(\mathbf{r})$, and ν_{bd} is the basis-pair kernel of v_ω,

$$e^{\sigma,\omega}_{\mathrm{HF}}(\mathbf{r}) = -\tfrac{1}{2}\sum_{bd} E^{\sigma}_b(\mathbf{r})\, E^{\sigma}_d(\mathbf{r})\, \nu_{bd}(\mathbf{r}), \qquad \int e^{\sigma,\omega}_{\mathrm{HF}}\, \mathrm{d}\mathbf{r} = E^{\mathrm{HF},\sigma}_{\mathrm{x}}(\omega). \tag{3}$$

Two conditions are bought by construction, not fitted. The switch

$$S = 1-e^{-u},\ u = a_s p + a_\alpha(\alpha-1)^2,\ p = \frac{\hat{\sigma}}{4k_{\mathrm{F}}^2\rho^2},\ \alpha = \frac{\sum_\sigma \tau_\sigma - \tau_W}{\tau_{\mathrm{unif}}},\ |\nabla\rho|^2 \le \hat{\sigma} = 2(\sigma_{\alpha\alpha}+\sigma_{\beta\beta}) \tag{4}$$

is built from two dimensionless variables that a uniform gas sets to $p = 0$ and $\alpha = 1$. In the unregularised form $u = 0$, $S = 0$, the correction vanishes identically, and **M6 is an identity, not a fitted property**. An explicit $\alpha \leftrightarrow \beta$ average does the same for M4. *The switch enforces M6 at every θ, not a network that learned it*. The model trained without the switch reads $144\times$ the uniform-gas budget. As deployed, the correction reads 8×10^{-9} of it (Table 1). The network is $11 \to 64 \to 64 \to 3$, with **5,123** trainable parameters (Appendices E, G).

Training objective. Seven terms enter the loss (Appendix D), all but the η term squared residuals in hartree. Four are shared by every model here: two against our own VMC totals, evaluated on the model's own SCF density, plus the PPLB chord residual for fractional charge and the one-electron condition $E_{\mathrm{xc}}[\rho] = -E_{\mathrm{H}}[\rho]$. The chord is the straight line joining two integer electron counts. The last three are targeted directly: the fractional-spin hinge at $N = 1$, the fragment-interaction residual $d = E(\mathrm{H}_2) - 2E(\mathrm{H}) - \Lambda$, and the $1/R$ coefficient through η_θ (Appendix D). **Those three are the audit's own criterion quantities**, so what this model reaches on them is what *training* buys, not what generalises. The VMC labels supervise the energy-residual terms only. The M1, M5 and M7 targets are theory-fixed instead: an analytic η target, the $N = 1$ identity $E_{\mathrm{xc}}[\rho] = -E_{\mathrm{H}}[\rho]$, and a published Born–Oppenheimer reference. They inherit no Monte-Carlo label noise.

Exact exchange in the architecture secures no condition by itself. Its coefficient must be pinned, by construction or by a loss carrying weight where the tail lives. However, an energy loss integrated against a density decaying as $e^{-2\sqrt{2I}\,r}$ carries almost none in that tail. The two diagnostics are defined differently, an ambiguity the local-hybrid literature names (Jaramillo et al., 2003) and addresses by calibration (Arbuznikov & Kaupp, 2014).

The evaluated model is the final checkpoint of our run, W-13-final in the appendix tables. It *targets* three of the seven conditions in its loss: M1, the one-electron condition (M5) and fragment interaction (M7). Of those three it still misses M1's band at audit. The fractional-spin hinge sits beside that last term. The model is *final*, not *best*; *the round was chosen post hoc*. *This model still fails its own held-out accuracy gate in every round we have run*: on systems it never trained on, $68.6\,\mathrm{mHa}$ against a $5\,\mathrm{mHa}$ threshold fixed before training, $13.7\times$ over; 11.19 against a $2\,\mathrm{mHa}$ training gate.

Training labels and error calibration. Our labels are finite-sample neural-wavefunction variational Monte Carlo estimates (Pfau et al., 2020; von Glehn et al., 2023), not coupled-cluster values. They carry two error bars, one an empirical label-error proxy and one variational. Those **two errors must not be conflated**: the variational one is calibrated against a published compilation that tabulates no anions, and *the molecular labels carry neither calibration*. We built none (Appendix F).

Thermochemistry. On three GMTKN55 subsets (Goerigk et al., 2017) the correction halves the self-interaction subset, worsens ionisation potentials and leaves electron affinities unmoved ($+0.0008$ kcal/mol). Even so, *we are not the best model*. We are worst of the three neural functionals on all three subsets, six comparisons of six; three earlier models of ours trail in all 18 of theirs. At the system level neither direction is resolved. The point estimates describe the measured set and are not established population effects (Table G5; Appendix G).

7 Discussion

Two readings pull on the same region. Our M7 reading improves when the correction dies away in the far tail. The $-1/r$ condition instead wants it to *shape* that tail. *In our interventions* the two readings move in opposite directions, echoing the familiar delocalisation-versus-static-correlation tension (Cohen et al., 2008a;b; 2012). *We do not interpret this as a representational incompatibility between the conditions*, and did not run the ablation that would isolate its cause.

7.1 Implications for design

On the fragment-interaction reading we now meet the criterion as written, and that is not the same as proving it. Our model turns every reading at once, all three conjuncts of the sealed rule:

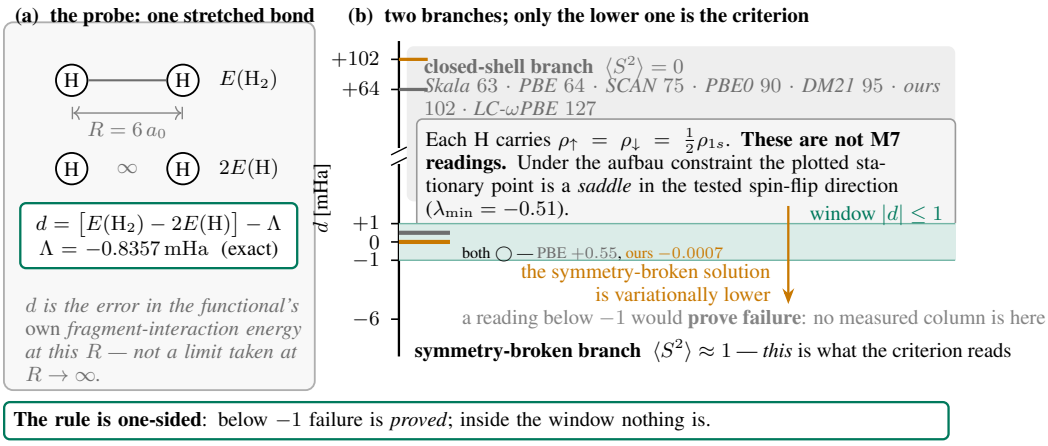

Figure 5: **What the M7 fragment-interaction criterion measures, and why a reading inside its window still proves nothing.** Λ is the high-precision variational Born–Oppenheimer value (Sims & Hagstrom, 2006). The closed-shell readings are Section 4's and the one-sided rule Section 2's, conditional as stated there. *Every number in (b) is a measurement, from Table G3 and Appendix G.*

the residual, its curvature and the fractional-spin margin. On the closed-shell branch (Table G3) **we are still behind PBE there**.

A longer-budget run (Appendix J) makes the trade-off explicit. With the M1 loss in both runs, a longer budget and the anneal horizon it stretches move η inside its gate and two untargeted metrics partway; they worsen M3 and cost the operational M7 gate its FS-0 margin on one run. A wider fit window is sensitivity, not a verdict. The two conditions the construction holds still hold.

These interventions carry one design implication: a condition has to be enforced by construction, or targeted and then measured. Analytical construction can still outrank empirical learning: *the hand-built controls answer the question they were run to answer, against us on one cell and level with us on the other* (Section 4).

7.2 Limitations, and what comes next

Four results above concern learned models and solvers, not chemistry. Five limitations qualify them. **Computational cost.** The two hybrid controls' price is not in Table G1. It is formally $O(N^4)$ in the exact-exchange integrals. **Scope of the conditions.** *"Seven conditions" is narrower than it sounds.* M1–M3 and M7 all probe self-interaction and delocalisation, and M1 and M5 probe one-electron systems. Fractional-spin constancy is measured at $N = 1$ only, so it is *unmeasured* at $N > 1$. So are Lieb–Oxford there, coordinate scaling and the derivative discontinuity. *At $N = 1$ every column is measured, ours by an earlier model* (Appendix A). **Basis set and reconstruction.** M3 shifts are measured on earlier and reported models, with no bound and no basis-set ladder, and its finite-basis reconstruction is uncalibrated against a directly computed potential (Appendix A). M2 has an aug-cc-pVTZ reading for our column and for PBE, and neither count moves. **Survey size.** *Three released families is still not a survey*, the third (pcNN) on four rows only. **Non-determinism.** Three sources are known and none is a functional property: branch selection in the stretched-bond diagnostic, a 10^{-6} Ha run-to-run scatter in the SCF solver that plain PBE shows too, and two unexplained discrepancies, of 0.045 and 0.024 mHa (Appendix G.12).

7.3 Conclusion

This work measures the distance between what a machine-learned functional's design guarantees and what its released weights deliver. Architectural intent and training penalties do not translate automatically into compliance. For the conditions no deployed construction guarantees, which ones survive is a property of the trained weights, not of the architecture. As the field moves toward foundation models for quantum chemistry, the assumption that a network implicitly inherits exact physics is the one most worth testing rather than presuming. This audit needs no retraining and reads only what a model's attachment supports. What it reports is not whether a functional is good but which of its presumed conditions survive the audit.

AI USE STATEMENT

We used generative AI tools for code implementation, drafting and copy-editing, literature retrieval and screening, parts of experimental design and claim drafting. They did not supply or fabricate any reported number, table entry or figure datum. The four schematic figures are hand-drawn vector graphics; the two with measured values print only numbers this paper's tables or appendices also print with provenance, none produced for the figure. We have reviewed all AI-assisted work: AI-drafted code was checked against its physical protocol before adoption, and every citation against its primary source. **We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.**

ETHICS STATEMENT

This work involves no human subjects, no personal data, and no data collection from people. It does report third-party measurements of published model weights and published functionals. We have confined those reports to residuals we measured ourselves on a declared benchmark, together with the protocol differences that affect them; we quote the original authors' own statements where we rely on them, and we make no claim about why any design decision was taken. In several places our own models are the ones that fail, and those cells are reported in the same tables.

Compute. This work ran on a single laptop: an Intel Core Ultra 9 285H with an NVIDIA GeForce RTX 5080 Laptop GPU, a 15 GiB memory budget, and CPU stages run at four threads. Its GPU generated the labels and ran two of the rounds in Table G7, including the reported one; its CPU ran the others and every grading and criterion run, pcNN's included, which runs in a software environment of its own (Appendix K). No large-scale compute was used. Training step budgets are listed in Appendix D; the benchmark card (Appendix H) states the protocol's wall-clock ceilings per system and per subset, and a grading run's per-stage timings are in the released judging products rather than here.

REPRODUCIBILITY STATEMENT

Software. PySCF (Sun et al., 2018; 2020), libxc (Lehtola et al., 2018), Dunning bases (Dunning, 1989; Kendall et al., 1992), GMTKN55 (Goerigk et al., 2017). *Benchmark.* Basis, grid, tolerances, reference density, thresholds: Section 2, Table A2; criterion code paths: Appendix A; SCF machinery and the aufbau caveat: Appendix C. *Training.* Objective (14); schedules, rules, initialisation: Appendix D. *Numbers.* Complete tables, unfilled cells included: Appendix G; derived quantities recomputed from raw columns. *Navigation for two things reviewers ask for.* The sealed thermochemistry point set ships with the archive, and its `lost_points` and `excluded_points` keys itemise every exclusion together with the model responsible. The four shared terms of (14) are classified one by one in Appendix D, under the rule of Section 5, and the finite-difference checks behind that assignment are in Appendix B. *Archive.* The anonymous supplement carries the sealed pre-registrations with their SHA-256 digests, which pin bytes only: that each preceded its runs rests on our records, not third-party time certification. It also ships the provenance ledger, benchmark code, reported checkpoint `2fb1dc5a...d6b41` and the raw files behind every table except those Appendix H lists as unsourced. One appendix number, the reported round's refresh-failure count, comes from a training log we do not ship, and it is the only one we have found. *Verifiers and limits.* Of two `verify.py` programs, the root one passes and exits zero; `physdftbench/verify.py` reads 16 pass, 1 fail, 3 skip in the anonymised archive, so exits 1. The failure is its `readme.9.out-anonymisation` self-count of redacted paths, which anonymising must change. The three skips all need a source root that reviewers lack. The audit half recomputes from archived outputs and M4's spin swap reruns end to end from a fresh unpack (23 of its thirty differences bitwise, the rest within one unit in the last place); other criteria are re-read, not re-derived, and the model half needs our labels. *A passing verifier is therefore not a statement that this paper has been reproduced.*

REFERENCES

Carlo Adamo and Vincenzo Barone. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.*, 110(13):6158–6170, 1999. doi: 10.1063/1.478522.

C.-O. Almbladh and U. von Barth. Exact results for the charge and spin densities, exchange-correlation potentials, and density-functional eigenvalues. *Phys. Rev. B*, 31(6):3231–3244, 1985. doi: 10.1103/PhysRevB.31.3231.

Brandon Amos and J. Zico Kolter. OptNet: Differentiable Optimization as a Layer in Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *PMLR*, pp. 136–145, 2017.

Alexei V. Arbuznikov and Martin Kaupp. Towards improved local hybrid functionals by calibration of exchange-energy densities. *J. Chem. Phys.*, 141(20):204101, 2014. doi: 10.1063/1.4901238.

D. T. Aznabaev, A. K. Bekbaev, and Vladimir I. Korobov. Nonrelativistic energy levels of helium atoms. *Phys. Rev. A*, 98(1):012510, 2018. doi: 10.1103/PhysRevA.98.012510.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep Equilibrium Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 688–699, 2019.

A. D. Becke. A multicenter numerical integration scheme for polyatomic molecules. *J. Chem. Phys.*, 88(4):2547–2553, 1988. doi: 10.1063/1.454033.

Kyle Bystrom and Boris Kozinsky. CIDER: An Expressive, Nonlocal Feature Set for Machine Learning Density Functionals with Exact Constraints. *J. Chem. Theory Comput.*, 18(4):2180–2192, 2022. doi: 10.1021/acs.jctc.1c00904.

Kyle Bystrom and Boris Kozinsky. Nonlocal machine-learned exchange functional for molecules and solids. *Phys. Rev. B*, 110(7):075130, 2024. doi: 10.1103/PhysRevB.110.075130.

Kyle Bystrom, Stefano Falletta, and Boris Kozinsky. Training Machine-Learned Density Functionals on Band Gaps. *J. Chem. Theory Comput.*, 20(17):7516–7532, 2024. doi: 10.1021/acs.jctc.4c00999.

Pablo A. M. Casares, Jack S. Baker, Matija Medvidović, Roberto dos Reis, and Juan Miguel Arrazola. GradDFT: a software library for machine learning enhanced density functional theory. *J. Chem. Phys.*, 160(6):062501, 2024. doi: 10.1063/5.0181037.

Subhas J. Chakravorty, Steven R. Gwaltney, Ernest R. Davidson, Farid A. Parpia, and Charlotte Froese Fischer. Ground-state correlation energies for atomic ions with 3 to 18 electrons. *Phys. Rev. A*, 47(5):3649–3670, 1993. doi: 10.1103/PhysRevA.47.3649.

Garnet Kin-Lic Chan and Nicholas C. Handy. Optimized Lieb-Oxford bound for the exchange-correlation energy. *Phys. Rev. A*, 59(4):3075–3077, 1999. doi: 10.1103/PhysRevA.59.3075.

Yixiao Chen, Linfeng Zhang, Han Wang, and Weinan E. DeePKS: A Comprehensive Data-Driven Approach toward Chemically Accurate Density Functional Theory. *J. Chem. Theory Comput.*, 17 (1):170–181, 2021. doi: 10.1021/acs.jctc.0c00872.

Aron J. Cohen, Paula Mori-Sánchez, and Weitao Yang. Fractional spins and static correlation error in density functional theory. *J. Chem. Phys.*, 129(12):121104, 2008a. doi: 10.1063/1.2987202.

Aron J. Cohen, Paula Mori-Sánchez, and Weitao Yang. Insights into Current Limitations of Density Functional Theory. *Science*, 321(5890):792–794, 2008b. doi: 10.1126/science.1158722.

Aron J. Cohen, Paula Mori-Sánchez, and Weitao Yang. Challenges for Density Functional Theory. *Chem. Rev.*, 112(1):289–320, 2012. doi: 10.1021/cr200107z.

Pablo del Mazo-Sevillano and Jan Hermann. Variational principle to regularize machine-learned density functionals: The non-interacting kinetic-energy functional. *J. Chem. Phys.*, 159(19): 194107, 2023. doi: 10.1063/5.0166432.

Sebastian Dick and Marivi Fernandez-Serra. Machine learning accurate exchange and correlation functionals of the electronic density. *Nat. Commun.*, 11(1):3509, 2020. doi: 10.1038/s41467-020-17265-7.

Thom H. Dunning. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.*, 90(2):1007–1023, 1989. doi: 10.1063/1.456153.

Titouan Duston, Jiashu Liang, Yuanheng Wang, Weihao Gao, Xuelan Wen, Nan Sheng, Weiluo Ren, Yang Sun, and Yixiao Chen. Agentic Discovery of Exchange-Correlation Density Functionals. *arXiv preprint arXiv:2605.05460*, 2026.

Eike S. Eberhard, Luca A. Thiede, Abdul Aldossary, Andreas Burger, Nicholas Gao, Vignesh Bhethanabotla, Alán Aspuru-Guzik, and Stephan Günnemann. Derivative Informed Learning of Exchange-Correlation Functionals. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2606.04279.

Matthias Ernzerhof and Gustavo E. Scuseria. Assessment of the Perdew–Burke–Ernzerhof exchange-correlation functional. *J. Chem. Phys.*, 110(11):5029–5036, 1999. doi: 10.1063/1.478401.

Shridhar R. Gadre, Lee J. Bartolotti, and Nicholas C. Handy. Bounds for Coulomb energies. *J. Chem. Phys.*, 72(2):1034–1038, 1980. doi: 10.1063/1.439270.

Igor S. Gerasimov, Timofey V. Losev, Evgeny Yu. Epifanov, Irina Rudenko, Ivan S. Bushmarinov, Alexander A. Ryabov, Petr A. Zhilyaev, and Michael G. Medvedev. Comment on "Pushing the frontiers of density functionals by solving the fractional electron problem". *Science*, 377(6606): eabq3385, 2022. doi: 10.1126/science.abq3385.

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, 2017. doi: 10.1039/C7CP04913G.

Nicholas C. Handy and Henry F. Schaefer III. On the evaluation of analytic energy derivatives for correlated wave functions. *J. Chem. Phys.*, 81(11):5031–5033, 1984. doi: 10.1063/1.447489.

Jan Hermann, Zeno Schätzle, and Frank Noé. Deep-neural-network solution of the electronic Schrödinger equation. *Nat. Chem.*, 12(10):891–897, 2020. doi: 10.1038/s41557-020-0544-y.

P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Phys. Rev.*, 136(3B):B864–B871, 1964. doi: 10.1103/PhysRev.136.B864.

Hisayoshi Iikura, Takao Tsuneda, Takeshi Yanai, and Kimihiko Hirao. A long-range correction scheme for generalized-gradient-approximation exchange functionals. *J. Chem. Phys.*, 115(8): 3540–3544, 2001. doi: 10.1063/1.1383587.

Juanita Jaramillo, Gustavo E. Scuseria, and Matthias Ernzerhof. Local hybrid functionals. *J. Chem. Phys.*, 118(3):1068–1073, 2003. doi: 10.1063/1.1528936.

Evgeny M. Kadilenko and Roland Grinis. The Power of Hellmann–Feynman Theorem: Kohn–Sham DFT Energy Derivatives with Respect to the Parameters of the Exchange-Correlation Functional at Linear Cost. *J. Phys. Chem. A*, 129(29):6694–6702, 2025. doi: 10.1021/acs.jpca.5c01771.

Bikash Kanungo, Jeffrey Hatch, Paul M. Zimmerman, and Vikram Gavini. Learning local and semi-local density functionals from exact exchange-correlation potentials and energies. *Sci. Adv.*, 11 (38):eady8962, 2025. doi: 10.1126/sciadv.ady8962.

Aaron D. Kaplan, Mel Levy, and John P. Perdew. The Predictive Power of Exact Constraints and Appropriate Norms in Density Functional Theory. *Annu. Rev. Phys. Chem.*, 74(1):193–218, 2023. doi: 10.1146/annurev-physchem-062422-013259.

M. F. Kasim and S. M. Vinko. Learning the Exchange-Correlation Functional from Nature with Fully Differentiable Density Functional Theory. *Phys. Rev. Lett.*, 127(12):126403, 2021. doi: 10.1103/PhysRevLett.127.126403.

Rick A. Kendall, Thom H. Dunning, and Robert J. Harrison. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.*, 96(9):6796–6806, 1992. doi: 10.1063/1.462569.

Vaibhav Khanna, Bikash Kanungo, Vikram Gavini, Ambuj Tewari, and Paul M. Zimmerman. Examining the Impact of Local Constraint Violations on Energy Computations in DFT. *J. Comput. Chem.*, 46(1):e70005, 2025. doi: 10.1002/jcc.70005.

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, 2021. doi: 10.1126/science.abj6511.

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. Response to Comment on "Pushing the frontiers of density functionals by solving the fractional electron problem". *Science*, 377(6606):eabq4282, 2022. doi: 10.1126/science.abq4282.

Wim Klopper, Rafał A. Bachorz, David P. Tew, and Christof Hättig. Sub-meV accuracy in first-principles computations of the ionization potentials and electron affinities of the atoms H to Ne. *Phys. Rev. A*, 81(2):022503, 2010. doi: 10.1103/PhysRevA.81.022503.

W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.*, 140(4A):A1133–A1138, 1965. doi: 10.1103/PhysRev.140.A1133.

Kirill Kulaev, Alexander Ryabov, Michael G. Medvedev, Evgeny Burnaev, and Vladimir Vanovskiy. On the practical applicability of DM21 neural-network DFT functional for chemical calculations: Focus on geometry optimization. *J. Chem. Phys.*, 163(7):074105, 2025. doi: 10.1063/5.0266500.

Susi Lehtola, Conrad Steigemann, Micael J.T. Oliveira, and Miguel A.L. Marques. Recent developments in libxc – A comprehensive library of functionals for density functional theory. *SoftwareX*, 7:1–5, 2018. doi: 10.1016/j.softx.2017.11.002.

Mel Levy, John P. Perdew, and Viraht Sahni. Exact differential equation for the density and ionization energy of a many-particle system. *Phys. Rev. A*, 30(5):2745–2748, 1984. doi: 10.1103/PhysRevA.30.2745.

Mathieu Lewin, Elliott H. Lieb, and Robert Seiringer. Improved Lieb-Oxford bound on the indirect and exchange energies. *Lett. Math. Phys.*, 112(5):92, 2022. doi: 10.1007/s11005-022-01584-5.

Li Li, Stephan Hoyer, Ryan Pederson, Ruoxi Sun, Ekin D. Cubuk, Patrick Riley, and Kieron Burke. Kohn-Sham Equations as Regularizer: Building Prior Knowledge into Machine-Learned Physics. *Phys. Rev. Lett.*, 126(3):036401, 2021. doi: 10.1103/PhysRevLett.126.036401.

Yang Li, Zechen Tang, Zezhou Chen, Minghui Sun, Boheng Zhao, He Li, Honggeng Tao, Zilong Yuan, Wenhui Duan, and Yong Xu. Neural-network Density Functional Theory Based on Variational Energy Minimization. *Phys. Rev. Lett.*, 133:076401, 2024. doi: 10.1103/PhysRevLett.133.076401.

Elliott H. Lieb and Stephen Oxford. Improved lower bound on the indirect Coulomb energy. *Int. J. Quantum Chem.*, 19(3):427–439, 1981. doi: 10.1002/qua.560190306.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing Millions of Hyperparameters by Implicit Differentiation. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *PMLR*, pp. 1540–1552, 2020.

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. *arXiv preprint arXiv:2506.14665*, 2025. arXiv version v6; the model release measured in this work is Skala-1.1, distributed by the authors at `https://github.com/microsoft/skala`.

Narbe Mardirossian and Martin Head-Gordon. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.*, 115(19):2315–2372, 2017. doi: 10.1080/00268976.2017.1333644.

Michael G. Medvedev, Ivan S. Bushmarinov, Jianwei Sun, John P. Perdew, and Konstantin A. Lyssenko. Density functional theory is straying from the path toward the exact functional. *Science*, 355(6320):49–52, 2017. doi: 10.1126/science.aah5975.

Matija Medvidović, Jaylyn C. Umana, Iman Ahmadabadi, Domenico Di Sante, Johannes Flick, and Angel Rubio. Neural network distillation of orbital dependent density functional theory. *Phys. Rev. Research*, 7:023113, 2025. doi: 10.1103/PhysRevResearch.7.023113.

Paula Mori-Sánchez, Aron J. Cohen, and Weitao Yang. Discontinuous Nature of the Exchange-Correlation Functional in Strongly Correlated Systems. *Phys. Rev. Lett.*, 102(6):066403, 2009. doi: 10.1103/PhysRevLett.102.066403.

Ryo Nagai, Ryosuke Akashi, and Osamu Sugino. Machine-learning-based exchange correlation functional with physical asymptotic constraints. *Phys. Rev. Research*, 4(1):013106, 2022. doi: 10.1103/PhysRevResearch.4.013106.

Sara Navarro-Rodríguez, Alec Wills, Kimberly J. Daas, María Camarasa-Gómez, and Marivi Fernández-Serra. Constraint-aware functional cloning for stable and transferable machine-learned density functional theory. *arXiv preprint arXiv:2605.10331*, 2026.

Mariana M. Odashima and K. Capelle. How tight is the Lieb-Oxford bound? *J. Chem. Phys.*, 127 (5):054106, 2007. doi: 10.1063/1.2759202.

Ryan Pederson and Kieron Burke. The difference between molecules and materials: Reassessing the role of exact conditions in density functional theory. *J. Chem. Phys.*, 159(21):214113, 2023. doi: 10.1063/5.0172058.

J. P. Perdew and Alex Zunger. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B*, 23(10):5048–5079, 1981. doi: 10.1103/PhysRevB.23.5048.

John P. Perdew. Imagining the future of density functional theory. *ZENtropy*, 1:3, 2026. doi: 10.20517/zentropy.2026.01. URL `https://www.oaepublish.com/articles/zentropy.2026.01`.

John P. Perdew and Mel Levy. Physical Content of the Exact Kohn-Sham Orbital Energies: Band Gaps and Derivative Discontinuities. *Phys. Rev. Lett.*, 51(20):1884–1887, 1983. doi: 10.1103/PhysRevLett.51.1884.

John P. Perdew and Karla Schmidt. Jacob's ladder of density functional approximations for the exchange-correlation energy. In *Density Functional Theory and Its Application to Materials*, volume 577 of *AIP Conf. Proc.*, pp. 1–20. AIP, 2001. doi: 10.1063/1.1390175.

John P. Perdew and Yue Wang. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B*, 45(23):13244–13249, 1992. doi: 10.1103/PhysRevB.45.13244.

John P. Perdew, Robert G. Parr, Mel Levy, and Jose L. Balduz. Density-Functional Theory for Fractional Particle Number: Derivative Discontinuities of the Energy. *Phys. Rev. Lett.*, 49(23): 1691–1694, 1982. doi: 10.1103/PhysRevLett.49.1691.

John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.*, 77(18):3865–3868, 1996. doi: 10.1103/PhysRevLett.77.3865.

John P. Perdew, Adrienn Ruzsinszky, Jianmin Tao, Viktor N. Staroverov, Gustavo E. Scuseria, and Gábor I. Csonka. Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits. *J. Chem. Phys.*, 123(6):062201, 2005. doi: 10.1063/1.1904565.

David Pfau, James S. Spencer, Alexander G. D. G. Matthews, and W. M. C. Foulkes. Ab initio solution of the many-electron Schrödinger equation with deep neural networks. *Phys. Rev. Research*, 2(3):033429, 2020. doi: 10.1103/PhysRevResearch.2.033429.

Kanun Pokharel, James W. Furness, Yi Yao, Volker Blum, Tom J. P. Irons, Andrew M. Teale, and Jianwei Sun. Exact constraints and appropriate norms in machine-learned exchange-correlation functionals. *J. Chem. Phys.*, 157(17):174106, 2022. doi: 10.1063/5.0111183.

Adrienn Ruzsinszky, John P. Perdew, Gábor I. Csonka, Oleg A. Vydrov, and Gustavo E. Scuseria. Spurious fractional charge on dissociated atoms: Pervasive and resilient self-interaction error of common density functionals. *J. Chem. Phys.*, 125(19):194112, 2006. doi: 10.1063/1.2387954.

Sebastian Schwalbe, Kai Trepte, and Susi Lehtola. How good are recent density functionals for ground and excited states of one-electron systems? *J. Chem. Phys.*, 157:174113, 2022. doi: 10.1063/5.0120515.

Rolf Seeger and John A. Pople. Self-consistent molecular orbital methods. XVIII. Constraints and stability in Hartree-Fock theory. *J. Chem. Phys.*, 66(7):3045–3050, 1977. doi: 10.1063/1.434318.

James S. Sims and Stanley A. Hagstrom. High precision variational calculations for the Born-Oppenheimer energies of the ground state of the hydrogen molecule. *J. Chem. Phys.*, 124(9):094101, 2006. doi: 10.1063/1.2173250.

John C. Snyder, Matthias Rupp, Katja Hansen, Klaus-Robert Müller, and Kieron Burke. Finding Density Functionals with Machine Learning. *Phys. Rev. Lett.*, 108(25):253002, 2012. doi: 10.1103/PhysRevLett.108.253002.

Jianwei Sun, Adrienn Ruzsinszky, and John P. Perdew. Strongly Constrained and Appropriately Normed Semilocal Density Functional. *Phys. Rev. Lett.*, 115(3):036402, 2015. doi: 10.1103/PhysRevLett.115.036402.

Qiming Sun, Timothy C. Berkelbach, Nick S. Blunt, George H. Booth, Sheng Guo, Zhendong Li, Junzi Liu, James D. McClain, Elvira R. Sayfutyarova, Sandeep Sharma, Sebastian Wouters, and Garnet Kin-Lic Chan. PySCF: the Python-based simulations of chemistry framework. *WIREs Comput. Mol. Sci.*, 8(1):e1340, 2018. doi: 10.1002/wcms.1340.

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, 2020. doi: 10.1063/5.0006074.

Alex J. W. Thom and Martin Head-Gordon. Locating Multiple Self-Consistent Field Solutions: An Approach Inspired by Metadynamics. *Phys. Rev. Lett.*, 101(19):193001, 2008. doi: 10.1103/PhysRevLett.101.193001.

Ingrid von Glehn, James S. Spencer, and David Pfau. A Self-Attention Ansatz for Ab-initio Quantum Chemistry. In *International Conference on Learning Representations (ICLR)*, 2023. ICLR 2023 poster; OpenReview forum xveTeHVlF7j.

Oleg A. Vydrov and Gustavo E. Scuseria. Assessment of a long-range corrected hybrid functional. *J. Chem. Phys.*, 125(23):234109, 2006. doi: 10.1063/1.2409292.

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, 2024. doi: 10.1039/D4CP00878B.

APPENDIX GUIDE

Each headline claim carries its evidence pointers in the main-text sentence that makes it, and the pre-registration's scope is in the caption of Table G7. This table is the map from a claim to the place that carries its evidence.

to check this	**read**
what each condition measures, and what it does *not* establish	A
every audited cell at full precision, with its protocol notes	G
the spin-permutation break, and what restoring it costs	A, M4
associations between condition readings and thermochemistry	G
the envelope argument, and the premise that broke	B
finite-basis Kohn–Sham, and the aufbau trap that decides M7	C
the loss, and when a fixed reference density is licensed	D, B
the model, and the two conditions that hold for any θ	E
the labels, and the two error bars each one carries	F
the released benchmark, its gates, and what it misses	H
two exploratory rounds, neither of them the reported model	I, J
two external construction tests, the second unreadable here	K
detail kept in the archive rather than in this PDF, and where: the archive's `MOVED_FROM_PDF.md` gives, for each block, what it is, the path that holds it, and what this PDF says about it.	

Round labels. Training rounds are named `W-`k in launch order, and that label is the hook into the released judging products, whose file names carry it. In prose a round is named by what distinguishes it; this table ties the two together.

in prose	label	what distinguishes it
the reported model	`W-13-final`	an explicit η term on the repair recipe
the earlier reported model	`W-9-final`	loosened the target-refresh tolerance
its best checkpoint	`W-9-best`	the reading that proves M7 failure, $d = -6.4062\,\mathrm{mHa}$
the larger-label round	`W-10`	more labels at fixed capacity — and a second change
its extra anneal	`W-10R`	$+10{,}000$ steps, warm-started, fresh schedule
the longer-budget run	`W-14`	$20{,}000 \to 60{,}000$ steps, anneal horizon with it
the discriminator round	`W-11`	arms `C0`, `R`, `W224`

Calibration notes for the compact matrix. M6 is reported in units of our 0.2 mHa/electron operational budget, not a community threshold, and its symbol does not grade magnitude: $1.12\times$ and $124\times$ both read ✗, so read the number. M5's symbol is the H $\wedge$ He^+ conjunction; the number printed beside it is He^+'s alone. M3 at $r = 3\,a_0$ is a finite-radius readout, not an asymptotic-limit proof. M7 is one-sided, with $d < -1\,\mathrm{mHa}$ proving failure while a reading inside $|d| \leq 1$ proves nothing.

What M4 reads in the compact matrix. *M4 there is the deployed energy on one fixed density pair, spins swapped*, taken as the largest over H^- and F^- on that column's *own* converged state;

the cross-fed and over-occupied pairs are controls, reported apart (Table A4). $\max|D_\alpha - D_\beta|$, spin breaking in the self-consistent *solution*, is a different quantity, and it is read instead by Table G1.

A THE SEVEN CONDITIONS: DEFINITION, WHY THE EXACT FUNCTIONAL SATISFIES THEM, HOW WE MEASURE THEM

Table A1: **What each row measures, and what it does not settle.** The seven are diagnostics associated with exact-condition families, not seven equivalent compliance tests, and the reading rules differ by row. Protocols are in Table A2; the values are in Table 1 and Table G1.

	What is measured	What it supports	What it does not establish
M1	η, the $1/R$ coefficient of symmetric H_2^+ over a finite fit window, and $\delta(\frac{1}{2})$, the half-occupation chord deviation	cancellation of that long-range tail term, and the curvature at one fractional point	piecewise linearity across the range of N
M2	the count of $\varepsilon_{\mathrm{HOMO}} < 0$ over H^-, O^-, F^-	that three systems pass a necessary sign check at the declared protocol	that the Koopmans equality error is small — it is 0.15–2.43 eV for our column (Table A3)
M3	$\lvert r\, w_{\mathrm{xc}}(3) + 1\rvert$, from a potential reconstructed on the HOMO at finite r	a tail diagnostic at that system, basis, radius and reconstruction	the exact $r \to \infty$ asymptote
M4	$\max\lvert\Delta E_{\mathrm{swap}}\rvert$ on fixed density-matrix pairs, in the deployed energy	a value beyond that column's numerical resolution is an identity counterexample	that zero on a finite set proves symmetry for all inputs; nor does SCF spin-breaking alone prove asymmetry
M5	λ/λ_1 at $N = 1$ on H and He^+, and the distance to the one-electron exact value	two separate readings: whether the bound is violated, and how large the one-electron self-interaction error is	that staying inside the bound implies the one-electron equality holds
M6	$\max\lvert\Delta\varepsilon_{\mathrm{xc}}\rvert$ against PW92 under a declared input construction, in units of our own budget; the controls and ours are read against PBE's uniform-gas value, itself PW92 to 2.4×10^{-7} mHa/electron	agreement with that reference, density window and polarisation range	that the uniform-gas limit holds or fails under any input representation
M7	d at $R = 6\,a_0$ for H_2 fragments, with a stability check and FS-0	a conditional finite-separation diagnostic and one-sided exclusions	size consistency, or fractional-spin constancy across the curve

Four labels for how a reading was obtained. A cell of *our* column is **targeted** when its condition enters this model's loss (M1, M5, M7; (14) and Appendix D), **untargeted** when no term of ours enters it (M2, M3), and held **by construction** when the architecture makes it an identity no fit can move (M4, M6). They say where a number came from, not how good it is: **targeted is not a mark of quality**, because a targeted cell reports what training bought on the probe that term was written against and nothing beyond. For M7 the one-sided rule below states that most sharply. **Held-out** is a different axis and **no cell of Table G1 is offered as one**, ours included: a condition no term of ours enters is not a probe withheld from training. In this paper held-out belongs to the accuracy gate of Table G6, which our model misses in every round, and nowhere else. The augmented η windows of Appendix J are *not* a held-out test: each still contains the three separations the loss was written against.

A.1 NOTES TO TABLE A2

[a] An anion-HOMO sign diagnostic: what a failing count indicts is the Koopmans relation, not binding, and it carries neither magnitude nor direction: see M2 below.

Table A2: The seven exact-condition families turned into measured quantities, with the protocol that fixes each. *The seven are not alike*: M4 is a direct identity test, M2 a necessary-condition sign test, M1, M3 and M6 are finite-protocol diagnostics, and M5 and M7 support one-sided exclusions at their declared bound and window. *Within each row*, the same basis, grid, point set and reference density for every functional measured on it; *across* rows the protocol changes, as the third column shows. "Exact" is what the exact functional must return, not a tolerance. **The lettered notes, derivations and code pins are in Appendix A.1 and A**.

	Criterion quantity	Protocol	Exact value / reading
M1	Piecewise linearity in N: $\eta = 1 - c/0.25$ from the fit $E(R) = E_\infty + c/R$ on symmetric H_2^+	aug-cc-pVQZ, grid 3, $R \in \{20, 35, 50\}$ Å, fitted in bohr	$\eta = 100\%$ for exact piecewise linearity; **the exact energy itself reads** 100.03 **on this window**[g]. Quoted to 0.1 pp (fit-window systematic $+1.0$ to $+2.8$ pp on ours, PBE and Skala-1.1, not common-mode; the `DM21` weight is larger, Table G1 note (1)). *Reads the long-range exact-exchange fraction of a column built to carry one*[f]
M2	Koopmans, $\varepsilon_{\mathrm{HOMO}} = -\mathrm{IP}$: count of $\varepsilon_{\mathrm{HOMO}}(\mathrm{X}^-) < 0$ over H^-, O^-, F^-	aug-cc-pVDZ, grid 3, UKS, fixed occupations	3 of 3: an anion-HOMO sign diagnostic, *not* the Koopmans violation itself and *not* whether the anion binds[a]
M3	$v_{\mathrm{xc}}(r) \to -1/r$ on H^-, shell-averaged, w_{xc} from the zero-projection identity[b]. **The gate is** $\mathrm{dev}(5) \le 0.05$; $\lvert\, r\, w_{\mathrm{xc}}(r{=}3) + 1 \,\rvert$ is a historical read-out with no threshold (both radii: Table G1)	aug-cc-pVDZ, grid 3; $r{=}3$ inside a validated window $[1.0, 6.5]\ a_0$[b]	0 as $r \to \infty$; at $r{=}3$ a finite-radius residual, smaller is better. Matched-control residual 0.025–0.030 where run; measured DZ→TZ shifts $[0.004, 0.082]$ — spreads, not bounds[b]
M4	$E_{\mathrm{xc}}[\rho_\alpha, \rho_\beta] = E_{\mathrm{xc}}[\rho_\beta, \rho_\alpha]$: (a) $\max \lvert D_\alpha - D_\beta \rvert$ on H^-, F^-; (b) $\max \lvert \Delta\langle S^2\rangle \rvert$ over 36 closed-shell systems; (c) $\max \lvert \Delta_{\mathrm{swap}} \rvert$, Table 1's ΔE_{swap}: the deployed energy read twice on one fixed density-matrix pair (L_3)	aug-cc-pVDZ, grid 3; (a) for all seven columns of Table G1, (b) for our models and the three controls; (c) every column, no SCF between the reads	0. Gate on (b): $\le 10^{-9}$; 0 of 36 above it, 0 of 35 for SCAN. (c) needs no tolerance
M5	Lieb–Oxford at $N = 1$, where also $E_{\mathrm{xc}} = -E_{\mathrm{H}}$ exactly: $\lambda = E_{\mathrm{xc}}/E_{\mathrm{x}}^{\mathrm{LDA}}$ against λ_1. **The gate is the conjunction**: λ on H *and* He^+, both $\le \lambda_1$. The printed reading is He^+'s signed distance from λ_{exact}	one shared PBE reference density; $E_{\mathrm{x}}^{\mathrm{LDA}}$ *unpolarised*, on the total density	Bound not exceeded, *and* $\lambda/\lambda_1 = 99.23\%$ — not 100%[c]
M6	Reduces to LDA (PW92) on the uniform gas: $\Delta\varepsilon_{\mathrm{xc}} = \varepsilon_{\mathrm{xc}}^{\mathrm{model}} - \varepsilon_{\mathrm{xc}}^{\mathrm{PW92}}$, window $r_s \in [2, 10]\ a_0 \times \zeta \in \{0, 1\}$	model grid against the PW92 reference	0. Budget 0.2 mHa/electron, **ours**, not a community standard[d]
M7	Fragment interaction at $R = 6\ a_0$, not size consistency at $R \to \infty$: $d = [E(H_2\ @\ R{=}6\ a_0) - 2E(\mathrm{H})] - \Lambda$, $\Lambda = -0.8357$ mHa[e]	aug-cc-pVDZ, grid 3; aufbau-constrained stationary point, spin-flip curvature test, and a variational bound. UKS, $\langle S^2\rangle \approx 1$	$\lvert d \rvert \le 1$ mHa, judged one-sidedly and conditionally, on the premise in Appendix A: a converged value is then an upper bound, so < -1 *proves* failure and inside the window proves nothing[e]

[b] w_{xc} is defined at (6) below and read back out of the converged HOMO, so no closed form is needed; the archived value and its owner are in M3 below.

[c] λ_1, the ambiguity window an older constant puts around it, and λ_{exact} are in M5 below; the exact functional does not sit at the bound, so the signed distance is the informative reading.

[d] Multiples of this budget are a scale, not a verdict against a published threshold. [e] $\Lambda = -1.0008357077 + 1$ Ha: the high-precision variational Born–Oppenheimer energy of Sims & Hagstrom (2006) Tab. VI at a tabulated grid point, printed there to 16 digits, minus $2E(\mathrm{H}) = -1$ Ha analytically — not a Monte-Carlo number. At finite R it also absorbs dispersion (≈ 0.14 mHa) and is not counterpoise-corrected. A validity gate ($d_{\mathrm{pool}} \le d_{\mathrm{ub}} + 0.10$ mHa) rejected four readings, each traced to a negative-curvature saddle.

[f] At $R \ge 20$ Å the spurious tail of the symmetric H_2^+ cation is the Hartree self-repulsion of two half-charges, $1/(4R)$, which only long-range exact exchange cancels. For a hybrid carrying a *constant* long-range fraction a, and provided nothing else contributes at order $1/R$, the leading-order analysis gives $c = 0.25(1-a)$ and hence $\eta = a$. Note g quantifies that condition for the exact functional, whose residual is an R^{-4} polarisation rather than a $1/R$ term. For a learned functional, whose long-range fraction is not a constant of its construction, η is the effective cancellation measured on this probe and does not by itself identify a spatially varying exchange coefficient (Section 4).

[g] $\eta = 100$ is the reading for a functional with *no* spurious tail. The exact Born–Oppenheimer energy has a real one: the residual polarisation of the H atom by the bare proton, $-\alpha/(2R^4)$ with $\alpha_{\mathrm{H}} = 4.5\, a_0^3$, which the same three-point c/R fit on $R \in \{20, 35, 50\}$ Å, converted to bohr, turns into $c = -7.14 \times 10^{-5}$, hence $\eta_{\mathrm{exact}} = 100.03$ on this window. LC-ωPBE's 100.024 is therefore 85% of that residual, not an overshoot past exactness; at the 0.1 pp precision this row is quoted to, the distinction changes no verdict.

Every note here fixes a protocol, not a verdict.

A.2 M1 — PIECEWISE LINEARITY IN ELECTRON NUMBER

Significance of M1: how much of the spurious $1/R$ tail of stretched H_2^+ each functional removes.

Statement. For the exact functional the ground-state energy is piecewise linear between integer electron numbers (Perdew et al., 1982): $E(M+w) = (1-w)E(M) + wE(M+1)$ for $0 \le w \le 1$. Approximate functionals curve — the origin of delocalisation error and of spurious fractional charges on dissociated fragments (Ruzsinszky et al., 2006; Cohen et al., 2008b; Mori-Sánchez et al., 2009).

Theoretical justification. In the ensemble formulation the ground state at non-integer N is a statistical mixture of the two adjacent integer states, whose energy is the corresponding convex combination.

Criterion quantity. Symmetric H_2^+ at large separation: there a delocalised solution puts exactly half an electron on each proton (measured $q_{\min} = 0.5000$), so under piecewise linearity the total energy must be *independent* of R to leading order in $1/R$. An approximate functional leaves a spurious tail, and fitting

$$E(R) = E_\infty + \frac{c}{R} \qquad \Longrightarrow \qquad \eta = 1 - \frac{c}{0.25} \tag{5}$$

gives $\eta = 100\%$ for a functional with no spurious tail and $\eta \approx 0$ for one with the full $1/(4R)$ tail. The constant 0.25 is the tail coefficient a self-interacting functional generates from a half-and-half charge distribution.

Protocol and measurement. aug-cc-pVQZ, grid level 3, fit range $R \in \{20, 35, 50\}$ Å. **Convert R to bohr before fitting**: the 0.25 in Eq. 5 is an atomic-unit constant. The basis is fixed in the criterion code rather than passed in at run time — so re-running with the basis given explicitly reproduces every model's archived value to $|\Delta\eta| \le 3.1 \times 10^{-11}$. Two protocol rules travel with this number. (i) η is quoted to 0.1 pp. Widening the fit range from $R \ge 20$ Å to $R \ge 12$, 8 and 5 Å gives 96.549, 96.731 and 99.312%. That protocol systematic of $+2.79$ pp dominates that fit's least-squares standard error (± 0.0012 pp). It is larger than the 2.5 pp tolerance the released gate reads this criterion at (Appendix H), so the window alone can move the reading by more than the gate's width. **That systematic is not common-mode**, and the two fits' standard errors differ by $\sim 300\times$ for a declared `conv_tol` reason and not an estimator one; both are per column in Table G1, note (1). Figure A1 shows the fitted quantity for each functional. (ii) A checkpoint whose fit rests on two points (one R having failed to converge) has zero degrees of freedom, so no error bar at all, and is never quoted as if it had one.

SCF branch selection. At these separations σ_g and σ_u are numerically degenerate, so the search decides the branch: **no irreducible-representation constraint is imposed anywhere in this row**. Every model instead walks the R chain, warm-started at each R from the previous converged density under ordinary aufbau occupations. The measured $q_{\min} = 0.5000$, the smaller of the two proton populations, at every R records a delocalised solution, which is what η needs; it does not by itself tell the two branches apart. The chain is $R = \{0.8, 1.06, 1.5, 2, 3, 5, 8, 12, 20, 35, 50\}$ Å, in that order. Re-running it reproduces the archived energies along that chain to 1.7×10^{-14} Ha. The θ_0 model needs no chain: with the last layer zeroed both the correction and its potential vanish identically (measured $E_\Delta = 0$ and $\max|v_\Delta| = 0$, digit for digit). Plain PBE's converged density is therefore a fixed point of the bridge, reproducing the archived PBE value on it to 2.2×10^{-16} Ha. *A cold start is not a substitute*: from scratch at $R = 20$ Å the same solver does not converge in 300 DIIS cycles, a property of the degeneracy and the reason the chain exists.

Skala-1.1's reported exchange fraction is not this criterion. Skala-1.1's own paper reports an "effective non-local exchange of 53.7%" on Grimme's molecular thermometer (Luise et al., 2025). On the criterion quantity of this subsection the same model reads $\eta = 0.005\%$. These are two different quantities.

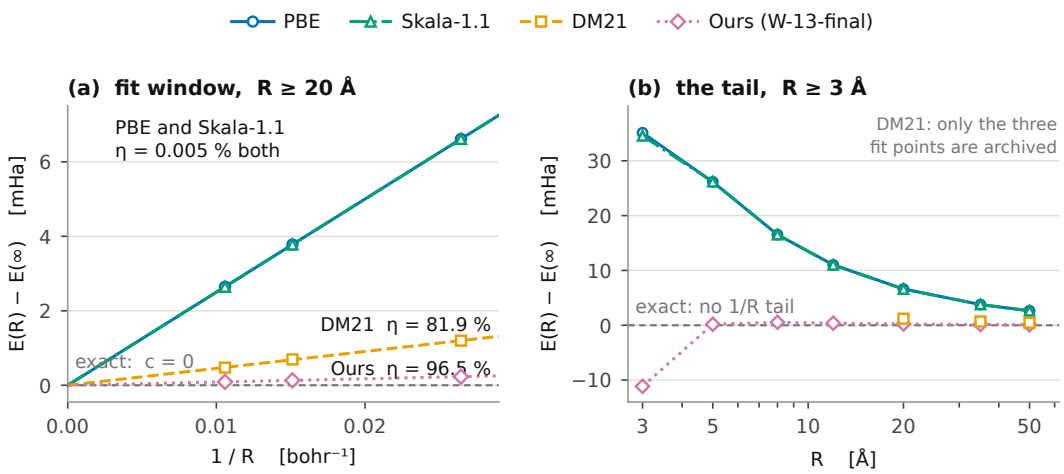

Figure A1: What η measures. At large R the symmetric H_2^+ ion carries exactly half an electron on each proton (measured $q_{\min} = 0.5000$), so under piecewise linearity $E(R)$ would carry no $1/R$ term; most approximations leave a spurious tail c/R, and $\eta = 1 - c/0.25$. (a) Inside the fit range the slope *is* c: PBE and Skala-1.1 both leave essentially the entire spurious tail ($\eta = 0.005\%$ for both — they agree closely, which is a statement about how little either removes and not a meaningful coincidence), while DM21 leaves about one sixth of it. In both panels the dashed line at zero is the *leading-order* statement that the exact functional has no $1/R$ tail, not a finite-R exact energy. (b) The same residual over the tail, $R \geq 3$ Å: DM21 and our correction do not scale the tail down, they change its shape. DM21 is drawn as markers only, because the archived baseline stored the three fit points rather than the full scan.

A.3 M2 — Koopmans

Significance of M2: whether the HOMO eigenvalue of a bound anion is negative, as Koopmans requires.

Statement. For the exact functional the highest occupied Kohn–Sham eigenvalue equals minus the ionisation potential (Perdew et al., 1982; Perdew & Levy, 1983; Levy et al., 1984). For an anion that is bound, this forces $\varepsilon_{\mathrm{HOMO}} < 0$.

Criterion quantity, and the easiest misreading. The quantity is the count of $\varepsilon_{\mathrm{HOMO}}(\mathrm{X}^-) < 0$ over H^-, O^- and F^-, and the exact functional reads 3 of 3. **This is an anion-HOMO sign diagnostic: not a measurement of the Koopmans violation itself, and not of whether the anion binds**. This row's easiest misreading: in the ΔSCF sense PBE and Skala-1.1 bind all three anions (Skala-1.1 gives EA $= +0.0322/ + 0.0535/ + 0.1279$ Ha) while their HOMO eigenvalues are positive.

Two limits of a sign count. First, it measures neither magnitude nor direction: a functional at $\varepsilon_{\mathrm{HOMO}} = -0.001$ Ha scores 3 of 3 while violating Koopmans by whatever its ΔSCF ionisation potential happens to be. The more informative statistic, $|\varepsilon_{\mathrm{HOMO}} + \mathrm{IP}^{\Delta\mathrm{SCF}}|$ per system, we report in Table A3. The count is what the sealed instrument scores; the residual is a reading printed beside it, not a second verdict. Second, a finite Gaussian basis artificially confines an unbound anion, and confinement *raises* a one-electron energy: at a fixed operator a larger variational basis lowers every eigenvalue, so a small basis biases $\varepsilon_{\mathrm{HOMO}}$ *upward*, not downward. *The premise of that step is not met here*: a self-consistent Kohn–Sham operator is built from the density it converges to, so the nested-subspace monotonicity of a *fixed* operator does not transfer to a self-consistent eigenvalue. aug-cc-pVDZ is a small basis for H^-, O^- and F^-, so that is a direction and not a measurement. At aug-cc-pVTZ our three eigenvalues all move further negative and the count stays $3/3$; PBE stays $0/3$; the other five columns were not re-run. As a heuristic it points at the cells this row *fails*, not the ones it passes, and it does not shield those passing cells. The cells it points at are the $0/3$ readings of PBE, Skala-1.1 and SCAN and the $1/3$ of PBE0 and DM21, on which the anion-HOMO sign-count clause of Section 3's headline rests. Every M2 verdict in this paper is therefore a verdict *at aug-cc-pVDZ*.

Table A3: **What the M2 sign count does not measure.** The Koopmans residual $\Delta_{\mathrm{IP}} = \varepsilon_{\mathrm{HOMO}}(\mathrm{X}^-) + \mathrm{IP}^{\Delta\mathrm{SCF}}$ in eV on the three anions M2 counts, with $\mathrm{IP}^{\Delta\mathrm{SCF}} = E(\mathrm{X}) - E(\mathrm{X}^-)$; the exact value is 0. **A column can read $3/3$ on the sign test and still miss the equality by electronvolts.** Ours and LC-ωPBE read $3/3$ in Table 1, and so does the released `DM21mc` weight, which is closer to the equality than we are on all three systems. The six non-DM21 columns share aug-cc-pVDZ, grid level 3 and $\mathrm{conv_tol} = 10^{-10}$; the four DM21 weights run at their released $10^{-6}/10^{-3}$, the same split Table G1 declares. O^- and F^- carry a repeat-evaluation floor near 10^{-6} Ha on this protocol, so no entry is quoted beyond 0.01 eV.

	$\mathbf{H}^-$	$\mathbf{O}^-$	$\mathbf{F}^-$	$\max_X \lvert\Delta_{\mathrm{IP}}\rvert$
PBE	$+2.50$	$+4.39$	$+5.31$	5.31
PBE0	$+1.56$	$+2.69$	$+3.19$	3.19
LC-ωPBE	-0.78	-0.04	$+0.50$	**0.78**
SCAN	$+2.35$	$+3.98$	$+4.68$	4.68
Skala-1.1	$+2.58$	$+3.64$	$+4.49$	4.49
`DM21`	$+1.35$	$+2.04$	$+1.86$	2.04
`DM21m`	$+1.03$	$+1.96$	$+2.18$	2.18
`DM21mc`	-0.03	$+0.35$	$+0.72$	0.72
`DM21mu`	$+1.60$	$+2.05$	$+2.17$	2.17
Ours	$+0.15$	$+1.26$	$+2.43$	2.43

A.4 M3 — Asymptotic $-1/r$ decay of the potential

Significance of M3: how far $r\, w_{\mathrm{xc}}(r) + 1$ sits from zero at $r = 3\, a_0$ on H^-.

Statement. For a finite system the exact exchange–correlation potential decays as $-1/r$ (Almbladh & von Barth, 1985; Levy et al., 1984): the exchange hole integrates to -1, and an electron far away sees the unscreened charge of the hole it left behind. The leading term is set by the hole, not by the net charge, so the statement holds at any total charge and covers the anion we measure it on. What differs between neutral systems, cations and anions is the sub-leading behaviour and the radius at which the asymptote sets in, which is why we compare only like with like.

Criterion quantity, and what w_{xc} is. Acting with the converged Kohn–Sham operator on the highest occupied orbital ϕ_{H} and solving (11) of Appendix C for the exchange–correlation part gives the *zero-projection identity*

$$w_{\mathrm{xc}}(\mathbf{r}) = \varepsilon_{\mathrm{H}} + \frac{1}{2}\frac{\nabla^2\phi_{\mathrm{H}}(\mathbf{r})}{\phi_{\mathrm{H}}(\mathbf{r})} - v_{\mathrm{ext}}(\mathbf{r}) - v_{\mathrm{H}}(\mathbf{r}), \tag{6}$$

which reads the potential *back out of the converged orbitals*, needs no closed form for the functional, and so is measurable for all four audited columns. It is a potential, not an energy density, so the exact asymptote is $-1/r$, not $-1/2r$, and the criterion quantity $\lvert\, r\, w_{\mathrm{xc}}(r) + 1\,\rvert$ has exact value 0 only as $r \to \infty$. Every reading below is a finite-radius residual under this protocol, not that limit. The readout is a shell average, not one grid point: half-width $0.5\, a_0$, trapezoidal radial weight, angular `sphere_grid(12,24)`, $\mathrm{d}r = 0.05$, $r_{\max} = 12.6$.

Four boundaries travel with (6). (i) It is a *single-orbital* read of a local potential reproducing that functional's own converged density, not a full Kohn–Sham inversion from a high-accuracy many-body density, which we did not do. (ii) Every column whose potential is an operator rather than a multiplicative function — through exact exchange (ours, DM21, PBE0, LC-ωPBE) or through τ (SCAN, Skala-1.1) — lives in the generalised Kohn–Sham framework (Appendix C), so for those (6) reads "the effective local potential seen by the HOMO" and we do not claim it is a multiplicative v_{xc} of those functionals; only PBE's potential is multiplicative. (iii) *In a finite Gaussian basis the reconstruction is not exact pointwise.* φ_{H} satisfies the projected eigenvalue equation, not the differential one at every point. Writing $R = (\hat{h} - \varepsilon_{\mathrm{H}})\varphi_{\mathrm{H}}$, the effective potential obeys $v_{\mathrm{eff}} = \varepsilon_{\mathrm{H}} + \frac{1}{2}\nabla^2\varphi_{\mathrm{H}}/\varphi_{\mathrm{H}} + R/\varphi_{\mathrm{H}}$. Equation (6) drops the last term, and that term is divided by an orbital that is small in the tail — where this row is read. We did not calibrate it: the finite-basis projection residual and the averaging convention both enter (6) and neither is measured here. The comparison against a directly computed multiplicative potential is well defined — both quantities can be taken

pointwise on one grid and then given the same shell average — and we list it as future work. (iv) On H^- the HOMO of PBE, SCAN and Skala-1.1 is unbound in this basis (M2 reads $0/3$ for each), so for those columns (6) reads the tail of an orbital the basis holds in place; the positive control below, a bound UHF HOMO, does not cover that case.

Protocol, the validated radius window, and the basis-set sensitivity. aug-cc-pVDZ, grid 3, on H^- for every functional. The radius $r = 3\,a_0$ is not a guess. Running the pipeline on the analytic positive control (same basis, UHF, where $r\,w_x \to -1$ has a pointwise closed form) reproduces the analytic curve to $|\Delta(r\,w_x)| \leq 0.10$ over $r \in [1.0, 6.5]\,a_0$ on H^- at this basis, so $r = 3$ sits well inside a validated window. The window is set by the HOMO decay rate: it reaches 6.5–$7.0\,a_0$ for H^- ($\kappa \approx 0.30$) but only $1.5\,a_0$ for He^+, which is why one diffuse anion, not a survey, carries it.

The basis-set sensitivity. The archived 0.2282 elsewhere in our records is *our own W-4 model at aug-cc-pVTZ on H^-*, against 0.1465 for the same model at aug-cc-pVDZ — a shift of $+0.082$. Across every model with both bases, the aug-cc-pVDZ → aug-cc-pVTZ shift of this quantity at $r = 3$ on H^- lies in $[0.004, 0.082]$. The matched-control residual — a one-electron problem with no electron–electron term, solved in the same basis with its decay rate fitted to that cell's density, whose exact $r\,v/Z_{\text{eff}}$ is -1 at every radius — is 0.025 for PBE and 0.030 for `DM21` (the other three released weights read 0.026–0.027); the other five columns, the reported model among them, carry none. The gap between the value we report (0.229) and the runner-up among the four audited columns (`DM21`, 0.369) is 0.140, i.e. $1.7\times$ the largest measured basis shift and $4.6\times$ `DM21`'s own residual. *Both are measured spreads, not bounds*: neither brackets the basis error of a column not run at the larger basis. **Three different quantities circulate on this row and none may be substituted for another**: that shift, the matched-control residual and the HF-control discrepancy (0.0643), which is one control's residual and not an established common offset. Applied symmetrically they leave the ordering against the released `DM21mc` weight unresolved by this benchmark, and leave the one against the LC-ωPBE control a reading *at the declared aug-cc-pVDZ protocol*, not an ordering established beyond it. That control has no aug-cc-pVTZ reading, so the pair's differential basis convergence is unverified. Both distances are in Appendix G.3.2.

Two further limits on that defence. The aug-cc-pVTZ readings exist for our own models and PBE only — DM21 and Skala-1.1 were not re-run at the larger basis — so the ordering is shown stable against the basis for the columns we could move, not for all four. Only the H^- row is resolvable at all, since on F^- the matched-control residual (0.142) exceeds the Skala-1.1–PBE difference (0.039).

Cross-environment check, and what it does and does not show. PBE reads 0.553360 in the DM21 environment and 0.553360 in ours. Both call the same PySCF and libxc implementation, so this is a *harness*-consistency check across two installations. It is **not** two independent implementations, and it says nothing about whether the criterion, the grid or the choice $r = 3$ is right.

A.5 M4 — Spin permutation

Significance of M4: whether swapping the two spin channels leaves the energy unchanged.

Statement. $E_{\text{xc}}[\rho_\alpha, \rho_\beta] = E_{\text{xc}}[\rho_\beta, \rho_\alpha]$. This is an identity, not an approximation target: relabelling the two spin channels cannot change the energy, and it needs no derivation.

Six levels, kept separate, and which of them the main table carries. (L0) the potential level, which on a closed-shell system can return zero for a functional that is not symmetric, a provably blind read; (L1) the network map itself; (L2) the density level, $\max|D_\alpha - D_\beta|$ on a closed-shell anion; (L3) the deployed energy, read twice on one fixed density-matrix pair; (L4) two independent self-consistent runs with the spin labels swapped; and (V-A) an aggregate over 36 closed-shell systems, $\max|\Delta\langle S^2\rangle|$, with gate $\leq 10^{-9}$ and a count of systems above it. The table below adds two auxiliary readings, (L4′) and (V-E), the latter the energy companion of V-A on the same 36 systems. L2 is measured for all seven columns on the same basis, grid and systems (the tolerances differ, as declared under Protocol below) and is the historical density-layer number kept in Table G1. L3 is measured for every column of this paper, is the number in the main table (Table 1) and is reported below. V-A is measured for our models and the three hand-built controls (Table G3), and is *not measured* at that protocol for the released neural columns, which do not carry a V-A cell in any table

of this paper. That fixed-density/self-consistent split is not peculiar to this benchmark: Zhao et al. (2024) report that DM21 stays accurate when evaluated on fixed B3LYP densities while 34 of their 117 reactions do not reach self-consistency under it.

The auxiliary levels, in one table. These read the same swap at levels other than the deployed energy. *They are reported so that the deployed-energy result is not the only measurement on this row, and none of them carries it.* Per-system data are in the archive.

	What it reads	Reading	What it cannot establish
(L4)	the F-atom doublet converged twice, occupations $(5\alpha, 4\beta)$ and the swap, from a guess spin-symmetric at $\max \lvert G_\alpha - G_\beta \rvert = 0$; each model read on the spread of its own four same-branch readings	the four released DM21 weights give 0.126 to $0.820\ \mathrm{mHa}$, 7.7 to 109 times those floors; PBE on two routes and B3LYP give $0.000355\ \mathrm{mHa}$ or less; ours gives $0.010692\ \mathrm{mHa}$ at 1.0 times its own floor, an upper bound of $0.011\ \mathrm{mHa}$ and not a claim of zero	the two routes do not share an energy tolerance (10^{-6} Ha against 10^{-10} Ha), so no cross-route margin is defined
(L4′)	the highest occupied eigenvalue, invariant under the label swap, whose own spread already absorbs the $2p$ orientation splitting	the four released weights separate by 17 to 417 times that spread — for the `DM21` weight, $24\ \mathrm{mHa}$ against a spread of $0.10\ \mathrm{mHa}$ — while the two controls and ours stay within a factor of two of theirs	Skala-1.1 is not readable on L4 at all: its six legs land on two orientations of the $2p$ hole, so the group difference is half that splitting and not a spin effect (matched on orientation, 1.4×10^{-6} to $1.6 \times 10^{-5}\ \mathrm{mHa}$)
(V-E)	what sets the resolution on a nominally closed-shell row: exchanging the two spin channels and averaging over 36 systems, at the released package's own recommended $\texttt{conv_tol} = 10^{-6}$ Ha	the converged total energy moves by a median of $0.0169\ \mathrm{mHa}$ and at most $0.2494\ \mathrm{mHa}$, the largest being 249 times that tolerance; none of the 36 converges spin-pure ($\langle S^2 \rangle$ from 4.6×10^{-8} to 2.1×10^{-3}) and the shift tracks that residual, $r = 0.977$ on a log scale	it isolates no cause. Run-to-run variation is smaller in every case (by $15\times$ at worst, a median of 5×10^3, exactly zero on five of the 36), so the resolution here is set by spin purity, not by repeat noise

An independent study calibrates this functional's own numerical noise two ways and works three to five orders tighter than the released tool recommends (Kulaev et al., 2025); our positive controls at that recommended tolerance return $0.000355\,\mathrm{mHa}$ or less, so the tolerance alone does not produce the shifts here. We isolate no cause and claim nothing about how the functional was built.

Protocol. aug-cc-pVDZ and grid level 3, with one integrator. **The levels do not share a tolerance.** L2 reads the released DM21 weights through their own published interface, at its `conv_tol`$= 10^{-6}$. PBE and Skala-1.1 are read at 10^{-10}, and that difference is declared rather than controlled. L3 needs no tolerance at all, because it is two single-point evaluations on one fixed density-matrix pair with no SCF between them.

Which of the six levels we report, and two readings of the released artefact. *(L2) SCF level.* Driven through their own published PySCF interface, unmodified and at that interface's official `conv_tol`$= 10^{-6}$, the released DM21 weights give $\max \lvert D_\alpha - D_\beta \rvert = 4.8 \times 10^{-2}$ on the converged spin density matrices of a closed-shell anion — the number in Table G1. Its protocol limit is the tolerance: PBE and Skala-1.1 are read at 10^{-10}, and for a quantity whose exact value is zero and whose comparison spans thirteen orders of magnitude, declaring that difference is not the same as controlling it. A DM21 run at 10^{-10} would settle it and we did not obtain one.

We ran the confound backwards instead. On this benchmark, at `conv_tol`$= 10^{-6}$ and otherwise identical protocol, PBE reads 9.2×10^{-16} on H^- and 2.5×10^{-15} on F^-, against 1.1×10^{-16} and 6.2×10^{-15} at 10^{-10} measured in the same process. LC-ωPBE reads 1.4×10^{-11} and 4.3×10^{-14} against 7.5×10^{-12} and 2.4×10^{-14}. Loosening the tolerance by four orders of magnitude moves neither model out of round-off — at this level the digits are not reproducible run to run even at fixed tolerance, only the magnitude is — and thirteen orders separate that magnitude from 4.8×10^{-2}. The declared tolerance difference therefore does not account for DM21's cell.

(L1) Network-map level. Feeding a grid of feature vectors x and their spin-swapped images Px directly to the `DM21` weight measures the violation of $N(x) = N(Px)$ with no SCF in between: $\max \lvert N(x) - N(Px) \rvert = 9.6578 \times 10^{-2}$ on H^-, at 4000 of 4000 grid points, and at 3809 of 4000 on F^-. A symmetric input ($x = Px$) reproduces itself at all 4000 points, and repeating the same input twice gives identical output, ruling out non-determinism. The other three weights read the

same way. Across the four, $\max|N(x) - N(Px)|$ runs from 8.0×10^{-3} to 3.9×10^{-1}, which is 0.8 to 15% of the largest output the same probe returns, and every one of the eight readings is at least 10^5 float32 units in the last place. The symmetric-input and repeat checks return zero and bitwise agreement for all four.

What may and may not be concluded from L1. A network map that is not symmetric is compatible with a *deployed functional* that is exactly symmetric: our own released functional is precisely that configuration, zero at the deployed level by an evaluate-twice-average wrapper while the bare network inside it is not. Evaluate-twice-average is one of two routes DM21's own supplementary material describes (§S1.2 of Kirkpatrick et al., 2021); the other, a single pass trained with random spin-channel flipping, it describes as making the network *approximately* spin symmetric. **L1 therefore establishes nothing about how DM21 was built, and we draw no inference from it about whether any symmetrisation was applied.** L3 decides the row of Table 1 and L2 the row of Table G1; L1 locates the residual below the SCF.

(L3) Deployed-energy level. Neither L2 nor L1 shows that the *deployed energy functional* violates $E_{\mathrm{xc}}[\rho_\alpha, \rho_\beta] = E_{\mathrm{xc}}[\rho_\beta, \rho_\alpha]$, because a symmetric functional can have symmetry-broken stationary points. L3 closes that gap by evaluating each functional's own deployed total energy twice on a *fixed* density-matrix pair — once as (D_α, D_β) and once as (D_β, D_α) — with no SCF in between, on the same molecule, basis (aug-cc-pVDZ), grid (level 3) and integrator. **The grid is fixed before either evaluation**, so the two orderings integrate on the same points *by construction*. That is not a formality: left to itself `pyscf` builds the grid on the first evaluation and may prune it *against whichever density matrix arrived first* — under the version the `DM21` readings were produced with, a lazily built H^- grid loses 96 of its 10 024 points that way; under the newer one the threshold defaults to zero and none are. **This measurement rests on neither default**: the grid is built and measured bitwise unchanged across both evaluations, and on the `DM21` columns it is injected from the archived state with its point count and atom positions asserted to match. Reading $\Delta_{\mathrm{swap}} = E_{\mathrm{tot}}(D_\alpha, D_\beta) - E_{\mathrm{tot}}(D_\beta, D_\alpha)$. Because T, V_{ne} and J depend only on $D_\alpha + D_\beta$, and exact exchange $K[D_\alpha] + K[D_\beta]$ is manifestly swap-symmetric, Δ_{swap} is the spin-permutation defect of the deployed exchange–correlation energy. We check that term by term on the DM21 cells, where the one-electron piece is bitwise identical between the two orderings in 24 of 24 and the Coulomb piece in 22 of 24. The remaining two differ by one unit in the last place of the Coulomb energy itself (7.1×10^{-15} Ha).

Three density pairs are used per column and system (H^- and F^-, both closed-shell anions at fixed occupations): (a) the column's own converged M4 state; (b) the converged (D_α, D_β) of the `DM21` weight, fed to every column as a common asymmetric input; and (c) a hand-made pair obtained from the symmetric PBE solution by moving 0.05 Tr $D_\beta S$ of occupation from the β to the α HOMO, which involves DM21 nowhere. *Round-off is measured, not assumed*: evaluating the same functional twice on the same fixed pair returns bitwise identical energies in 59 of 66 cells and differs by one unit in the last place (1.4×10^{-14} Ha) in the other seven.

On all three pairs and both systems, PBE, PBE0, LC-ωPBE, SCAN, Skala-1.1, our checkpoint and pcNN return a Δ_{swap} of at most a few units in that last place. It is bitwise zero in 37 of 42 cells and 1.4×10^{-14} Ha in the other five, and an independent re-run of the released package reaches 2.8×10^{-14} Ha in one cell.

What a tick on this row rests on. Two things, not one. The first is the measurement above: Δ_{swap} at that column's own arithmetic floor. The second is a construction-level guarantee that the identity holds for *any* input, and it is of a different kind in each case. For PBE, PBE0, LC-ωPBE and SCAN it is algebraic: the functional form is invariant under the swap. For Skala-1.1 it is architectural and stated by its own authors, who apply the semi-local network to each spin-ordering and average (Appendix G); we did not inspect their code. For our column it is the deployment wrapper's explicit $\alpha \leftrightarrow \beta$ average, and *the bare network is not symmetric* — that contrast is the point of this row for us. **A column that measured at its floor with no such guarantee would be a ◯ here and not a ✓**; no column in this table is in that position, so the two readings agree on every cell printed, but they are not the same rule. *Which* cells carry the residual is not reproducible from run to run, because the order of a grid sum sets it. The bound $|\Delta_{\mathrm{swap}}| \leq 3 \times 10^{-14}$ Ha is reproducible. All four released DM21 weights return a non-zero Δ_{swap} in all 24 of theirs, from 1.8×10^{-2} to 3.83 mHa, nine to eleven orders of magnitude above that bound. **The decisive control is pair (b) read on PBE**: on the very density pair that gives `DM21` -2.17 mHa ($\max|D_\alpha - D_\beta| = 4.8 \times 10^{-2}$ on H^-), PBE

returns Δ_{swap} bitwise zero, so the reading is a property of the functional evaluated and not of the asymmetry of the input. *The 3×10^{-14} Ha bound certifies determinism, not the single-precision resolution of the released graph*, whose inputs are `float32`. Read instead against the exchange–correlation energy it acts on, the defect runs from 1.7×10^{-6} to 4.9×10^{-3}. That is 14 to 4×10^4 times the `float32` epsilon.

On pair (c), which no DM21 quantity enters, the four weights read 0.083 to 3.83 mHa where the other seven read zero or one unit in the last place; *that pair carries $\mathrm{Tr}\, D_\alpha = 1.05$ (H^-) and 5.25 (F^-)*, above single-orbital occupation and outside the fixed-occupation sector of the SCF, so it is reported as a stress test, apart from (a), the cell's own reading, and (b), the decisive control.

Pair (c) reproduces in magnitude but not in digits. It is built from the F^- HOMO, and that HOMO is one member of a threefold-degenerate $2p$ shell. A fresh run of the released package therefore picks a different member. Then $\max |D_\alpha - D_\beta|$ reads 1.29×10^{-1} rather than the archived 1.83×10^{-1}, and `DM21` reads -3.826 rather than $-3.832\,\mathrm{mHa}$. Pairs (a) and (b) are archived density matrices and reproduce bitwise, which is the second reason the cell and its decisive control rest on (a) and (b) rather than on (c).

What rules out a silent re-convergence is, first, the call path. Every energy in this measurement is a single-point evaluation on a density handed in. Ten of the eleven columns read it as `mf.energy_elec(dm) + mol.energy_nuc()`, and ours reads the same single-point assembly through its own bridge. *No SCF kernel runs between the two evaluations of a pair.* The four DM21 columns run no SCF at all, because they reuse archived converged states. The driver's only SCF calls build density pair (a) for the other columns, before any swap is taken. Each cell ships the fingerprint of the density it was handed. The variational check is second, and by itself it could not exclude a re-convergence to the *same* minimum: of the 88 fixed-density energies from (a) and (b), none lies below its own column's SCF minimum in the same sector (criterion -10^{-12} Ha). The energy the `DM21` weight returns at its own archived state reproduces that state's archived total to the last digit (-0.5310136342773943 Ha).

L3 therefore establishes what L2 alone could not: for the released DM21 weights the spin-permutation identity fails in the deployed energy itself, not only in the density it converges to. As at L1, this is a reading of the released artefact and not a statement about how DM21 was built; DM21's own supplementary material (Kirkpatrick et al., 2021) describes both a strict and an approximate symmetrisation, and we draw no inference about which, if either, is present in these weights. Every cell, its round-off floor, its density fingerprint and its provenance pin are in `results/runs/2026-09-05_M4SWAP/`.

Two readings of that table are worth separating. **The cross rests on the grid axis and on nothing narrower**: `DM21`'s $2.1733\,\mathrm{mHa}$ is 4.72 times the widest movement any cell of the four released weights shows across five integration grids, and 116 times that of its own pair, so it is not a repeat-call artefact. **The ticks rest on two things at once**, a construction-level guarantee and a measurement at that column's arithmetic floor; the six columns that carry one read 0 or one ULP on *all six* of their measurements, the two controls included. *A column at the floor with no construction guarantee would be left unresolved here and not ticked*, exactly as the paragraph above states it, and no printed column is in that state. A floor-level reading on its own is a numerical check, not a tick. The shipped checker applies the two-part rule and prints, every time it runs, what a measurement-only reading would have marked instead.

What restoring the identity costs downstream. The audit says the released weights break M4 in the deployed energy; it does not say whether that matters. We wrapped those weights so that the deployed energy and its derivative channels are exactly symmetric under $\alpha \leftrightarrow \beta$, changed nothing else, and re-ran three GMTKN55 subsets. A third leg re-measures the bare functional today through a byte-identical driver, so the comparison is not confounded with machine drift: that drift is -0.00%, $+0.03\%$ and -0.28%, and the third is one system, `IP_76`, where today's run reaches a solution $1.72\,\mathrm{mHa}$ lower than the archive. With that reaction removed the drift is $+0.0079\,\mathrm{kcal/mol}$, and on all three legs 142 of the 144 systems converge on the same solver setting. Against today's bare leg the mean absolute deviation moves by -14.01% on SIE4x4, -19.15% on G21EA and $+10.04\%$ on G21IP; with each subset's largest single reaction dropped, -12.72%, -3.81% and $+1.08\%$. Read against each subset's own closed-shell readings, whose root mean square is 0.0049, 0.0650 and $0.0367\,\mathrm{kcal/mol}$ (below), those residuals stand at 93, 1.0 and 2.4 times that width. *SIE4x4 is*

Table A4: **M4 (L3): what each mark rests on.** $\max|\Delta E_{\mathrm{swap}}|$ over the *two* measurements that belong to the column — both probe systems on density pair (a), its own converged M4 state. Pairs (b) and (c) are controls, reported apart and not folded in; `DM21` reads $3.8320\,\mathrm{mHa}$ on (c), the over-occupied pair, which is the stress test and not this cell. *Widest scale exercised* is read off the archive, not declared: a column counts as read on the grid axis only where this archive holds it on more than one integration grid. **The grid scale is the most conservative this archive holds**: the widest movement of any cell of the four released `DM21` weights across five grids, which is `DM21mc`'s over-occupied pair (c) — so, unlike the measurement, the scale does draw on the controls. PBE reads bitwise 0 in all 120 of those evaluations. The lower block reads each released weight on its own pair (a): against that conservative scale only `DM21` and `DM21mu` clear it, against the movement of their own pair all four do, and each keeps one sign on all ten of its readings (`results/runs/2026-09-14_RECHECK/`). ULP is one unit in the last place of that column's own total energy, the smallest difference two spin assignments could show once summed in a different order. Every number here is recomputed from the archived raw results by `verify_m4_swap.py`, which also re-derives the marks and checks them against the printed table.

	$\max\|\Delta E_{\mathrm{swap}}\|$ [mHa]	widest scale exercised	construction guarantee	
PBE	0	grid, 0	algebraic	✓
`DM21`	2.1733	grid, $0.4605\,\mathrm{mHa}$	**none**	✗
4.72× the conservative scale, 116× its own pair (a): the one column failed on the grid axis				
Skala-1.1	1.4e-11	this swap only	architectural, stated	✓
Ours	0	this swap only	architectural, written out	✓
PBE0	0	this swap only	algebraic	✓
LC-ωPBE	0	this swap only	algebraic	✓
SCAN	1.4e-11	this swap only	algebraic	✓

weight	$\max\|\Delta E_{\mathrm{swap}}\|$ [mHa]	pair-(a) spread [mHa]	ratio of unrounded values
`DM21`	2.1733	0.0187	116×
`DM21mu`	1.5215	0.0187	82×
`DM21mc`	0.2074	0.0130	16×
`DM21m`	0.1204	0.0071	17×

resolved and G21IP is too, at 93 *and* 2.4 *times it. G21EA's residual sits on the width, and we quote no effect for it.* Self-interaction improves, ionisation energies degrade, and the wrapper costs 1.3 to 1.8 times the wall clock. On `IP_74` the bare functional reaches a solution the symmetric one does not, $41.2\,\mathrm{mHa}$ lower; *that is a difference between solutions, not one state moving*.

A second implementation. The readings above were repeated with an evaluator written from scratch. Everything but the released graph and weights, which cannot be rewritten, and the input density matrices is rebuilt: the basis-function values and gradients, the three densities, the local Hartree–Fock energy densities at both ω, and the energy assembly. It returns every archived total energy to 4.3×10^{-14} Ha and every archived Δ_{swap} to a relative 7.9×10^{-10}, 16 of the 24 bitwise and all 24 with the same sign. PBE reads bitwise zero on that pipeline in 24 of 24, and so do the one-electron and Coulomb terms under the swap (`results/runs/2026-09-14_XEVAL/`).

Two wider floors, and a defect that scales. That second evaluator also found a larger source of movement than repeat evaluation. Reading the same frozen graph through a different TensorFlow fetch list moves the total energy by up to 3.0×10^{-7} Ha; noise in this functional's own outputs is reported independently (Kulaev et al., 2025). That spread is wider than the repeat bound, and the grid is wider still. Re-running all 24 cells on four further grids moves each cell by 1.1×10^{-6} to 4.6×10^{-4} Ha. Twenty-two of the 24 keep their sign, and the two that do not are among the smallest. The smallest reading anywhere in that 24×5 matrix is $1.8 \times 10^{-3}\,\mathrm{mHa}$, which is 6 times the fetch spread. **PBE returns bitwise zero in all** 120 **evaluations, and that count, not any single margin, is what this row rests on** (`results/runs/2026-09-14_RECHECK/`).

The basis is not a second axis. On a fixed grid the deployed energy reads only the density, its gradient, the kinetic energy density and the two local Hartree–Fock densities at the grid points. cc-pVDZ is an exact subspace of aug-cc-pVDZ, returning bitwise identical basis-function values on

the canonical grid, so one density pair embedded in both gives Δ_{swap} bitwise equal for all four weights and for PBE. The basis changes which pairs exist, not how a pair is read. Six bases give 240 readings from 40 cells, every one non-zero, PBE bitwise zero in all 96, and 23 of the 40 changing sign against 2 of 24 on the grid axis. The sign belongs to the state. The defect also scales with the input asymmetry. Write $(D_\alpha, D_\beta) = (\bar{D} + t\delta, \bar{D} - t\delta)$, which holds $D_\alpha + D_\beta$ fixed and turns the swap into $t \to -t$. On H^- the slope is $-2.830\,\mathrm{mHa}$ per unit t and holds to 0.97% over three decades, the reading is exactly zero at $t = 0$, and PBE is bitwise zero at every t. Rounding does not scale that way. Two limits go with it. The odd-parity residual is a bookkeeping identity rather than an independent measurement, because $\bar{D} - t\delta$ and $\bar{D} + (-t)\delta$ are the same bits. And on F^-'s own-SCF pair the small-t readings fall into the single-precision floor.

A.6 M5 — Lieb–Oxford at $N = 1$

Significance of M5: the signed distance of $\lambda = E_{\mathrm{xc}}/E_{\mathrm{x}}^{\mathrm{LDA}}$ on He^+ from where the exact functional sits; its symbol is the bound on H and He^+ together.

Statement. The indirect Coulomb energy is bounded below by $-C \int \rho^{4/3}$. The constant has been tightened repeatedly (Lieb & Oxford, 1981; Chan & Handy, 1999; Odashima & Capelle, 2007; Lewin et al., 2022). At $N = 1$ the bound specialises to a tight constant against the LDA exchange energy of the same density (Gadre et al., 1980).

Criterion quantity. $\lambda = E_{\mathrm{xc}}/E_{\mathrm{x}}^{\mathrm{LDA}}$ evaluated on He^+. The sharp bound constant used here is $\lambda_1 = 1.4783078009278349363$, while the older constant gives 1.4785553293; values inside $[1.4783078009, 1.4785553293]$ therefore fall in an ambiguity window and we report both. *Where that constant comes from*: $\lambda_1 = (4/3)(\pi/3)^{1/3} C_1$ with the $N = 1$ constant $C_1 = 1.0918171857861821283$, first printed to four digits by Gadre et al. (1980) and stated by Lewin et al. (2022) as an attained supremum following from the Lane–Emden equation. The digits above are our own evaluation of that closed form, $C_1 = 3\pi/(4\pi\mu_1)^{2/3}$ with μ_1 the Lane–Emden mass at the $n = 3$ boundary. *The profile that attains the bound is that polytrope, not a hydrogenic density*, so no atom in this paper saturates it. The older rounded value $C_1 = 1.0920$ reproduces 1.4785553293. *That constant is the bound, not the exact ratio on this density*: on the frozen He^+ reference the exact functional sits at $\lambda_{\mathrm{exact}} = 1.4669423$, derived below. **The denominator convention**: $E_{\mathrm{x}}^{\mathrm{LDA}} = A_{\mathrm{LDA}} \int \rho^{4/3}$ with $A_{\mathrm{LDA}} = -\frac{3}{4}(3/\pi)^{1/3}$, i.e. the *spin-unpolarised* form evaluated on the *total* density, which is the convention consistent with $\lambda_1 = (4/3)(\pi/3)^{1/3} C_1$. Using the spin-scaled LSDA form on this fully polarised system instead would move λ by $2^{1/3}$.

Protocol, and the test system. One PBE reference density (aug-cc-pVDZ, grid 3), shared by every column of that row, the three hand-built controls of Appendix G.3 included. The comparison is therefore of functionals on one density, not of their self-consistent solutions. Two λ conventions circulate in our own records and must never be mixed; the constant above is the one used throughout this paper. **He^+ is one of the two systems carrying our one-electron loss term** (the other is H; Appendix D), and after the repair that term is imposed on *this same cached reference density*. For our column this row is therefore an evaluation on a training point: a check that the term does what it says, not evidence of generalisation.

The exact functional does not sit at 100%; the signed distance is the informative reading. At $N = 1$ the exact functional returns $E_{\mathrm{xc}} = -E_{\mathrm{H}}$ (Perdew & Zunger, 1981), so λ_{exact} is a definite number on the shared reference density, not a free parameter. On our He^+ reference density it is $\lambda_{\mathrm{exact}} = 1.4669423$, i.e. 99.23% of λ_1. An independent closed form on an idealised hydrogenic $1s$ density ($E_{\mathrm{H}} = 5Z/16$, $\int \rho^{4/3}$ analytic) gives 99.36%; the 0.13 pp gap is that idealised density against the PBE/aug-cc-pVDZ density actually shared — a cross-check on the numerical value, not a competing number. Reading the row as a signed one-electron self-interaction error on that density:

	PBE	DM21	Skala-1.1	Ours (W-13-final)
$\lambda(\mathrm{He}^+)$	1.4541355	1.4694254	1.4669050	1.4718586
λ/λ_1 [%]	98.36	99.40	99.23	99.56
signed deviation from exact [pp of λ_1]	-0.87	$+0.17$	-0.00	$+0.33$
E_{xc} error on the shared density [mHa]	$+5.38$	-1.04	$+0.02$	-2.07
rank by $\lvert\cdot\rvert$	4th	2nd	**1st**	**3rd**

So on the physics content of M5 we are third of the four audited columns — and fifth of seven once the three controls are read on the same row (Table G1) — on the side making E_{xc} too negative on a one-electron density. Exceeding λ_1 remains a definite failure — our own W-6 model did, at 100.25%, and W-7 onwards pulled it back — but staying inside it is not an achievement, since all of them do. *And the ranking does not separate every adjacent pair.* Read against the exact value of the row — 99.23% of the bound, not 100, at $N = 1$ where $E_{\mathrm{xc}} = -E_{\mathrm{H}}$ — our 0.07 pp margin to PBE0 falls within the anchor's own 0.13 pp reference-density shift, so that pair is unresolved at this protocol (Table G1).

Does this row depend on its reference density? It is read on one shared PBE density, so we re-read it on the exact one. He^+ carries one electron, so Hartree–Fock is exact there and the density is the analytic hydrogenic $1s$. On it λ_{exact} has a Z-independent closed form, $(80/81)(\pi^2/3)^{1/3} = 1.4689188$, which is 99.3649% of λ_1 and which the grid returns to 4.5×10^{-14}. **No verdict moves.** Every column stays inside the bound, the largest reading is 99.47%, and the column that exceeds it on H still does. **The order does.** Each signed deviation shifts down by 0.13 to 0.27 pp, the spread across columns is 0.14 pp, and six of the twenty-one pairs change place. SCAN reads -0.00 pp on the exact density against $+0.13$ on the shared one. We therefore read the order of the middle columns on this row as unresolved (`results/runs/2026-09-14_M5ABL/`).

A.7 M6 — Uniform electron gas limit

Significance of M6: how far each functional drifts from PW92 on a uniform gas.

Statement. On a uniform gas the exchange–correlation energy per electron must reduce to the uniform-gas value — one of the known exact constraints reviewed by Kaplan et al. (2023) — whose correlation part we take in the parametrisation of Perdew & Wang (1992). Perdew (2026) names exactness for uniform densities as one of the two constraints that matter most.

Criterion quantity. $\Delta\varepsilon_{\mathrm{xc}} = \varepsilon_{\mathrm{xc}}^{\mathrm{model}} - \varepsilon_{\mathrm{xc}}^{\mathrm{PW92}}$, read over the window $r_s \in [2, 10]\, a_0$ crossed with $\zeta \in \{0, 1\}$, where the PW92 reference reads 0. For the three controls and ours the reference is PBE's own uniform-gas value, itself PW92 to 2.4×10^{-7} mHa/electron (Table G1, note (7)), so our cell reads the correction alone. *Three point counts circulate for this row, and they count different sets*: the scored window is those 10 points, five radii by two polarisations; the main grid adds the two radii below $r_s = 2$, giving 14; and the 21-point table adds the half-polarised $\zeta = \frac{1}{2}$ to all seven radii. Only the first is scored.

Where the budget comes from. **That budget is ours**, taken from our own training gate; it is not a community standard, and the multiples should be read as a scale, not as a verdict against a published threshold.

Two protocol warnings. (i) The number depends on the window: for DM21 the main-window value is $124.20\times$ budget on the 21-point table, 154.61 on a 33-point denser grid, and 51–297 across windows. (ii) For a functional whose non-local features are defined relative to atomic centres, the quantity is not well defined without a convention for where the nuclei sit. Across the eight conventions we tried the reading spans $[44.2\times, 1168.5\times]$ budget, and at one (r_s, ζ) the spread *across conventions* is $1282\times$ it — larger than the effect (Table G2, note †), so the released tool records "not measured" for that cell and **no one of those eight is quotable on its own**; the number the tables print is the declared extrapolation limit of Appendix H, read by a protocol other than the tool's. Cross-checks over centres and shapes show that sequence is stable; they do not show that every input representation gives it.

A.8 M7 — FRAGMENT INTERACTION AT FINITE SEPARATION

Significance of M7: whether $E(\mathrm{H_2}@R{=}6\,a_0) - 2E(\mathrm{H})$ matches the exact value of the same difference.

Statement. The energy of two infinitely separated fragments equals the sum of the fragment energies. Failure of this is the same physics as M1: a functional that curves in N can lower its energy by putting fractional charge or fractional spin on each fragment (Ruzsinszky et al., 2006). *What M7 measures is not that limit*: it is the fragment-interaction error of $\mathrm{H_2}$ at $R = 6\,a_0$, defined next.

Criterion quantity, and the anchor. $d = [E(\mathrm{H_2}@R{=}6\,a_0) - 2E(\mathrm{H})] - \Lambda$; the window is $|d| \leq 1$ mHa. The anchor is *not* a Monte-Carlo number of ours:

$$\Lambda = E^{\mathrm{ref}}(\mathrm{H_2}@R{=}6\,a_0) - 2E^{\mathrm{exact}}(\mathrm{H}) = -1.0008357077 - 2\times(-0.5)\ \mathrm{Ha} = -0.8357\ \mathrm{mHa}, \quad (7)$$

with E^{ref} the high-precision variational Born–Oppenheimer total energy of Sims & Hagstrom (2006), Table VI, at a tabulated grid point — $R = 6\,a_0$ lies exactly on their grid, so no interpolation is involved — and $E^{\mathrm{exact}}(\mathrm{H}) = -\frac{1}{2}$ Ha analytically. Rounded from -0.8357077 mHa.

That reference is a 7034-term Hylleraas variational upper bound, tabulated to 16 decimal places at each R. That gives a round-off floor near 10^{-13} mHa, far below the pass window's width. The floor is a statement about the printed digits and *not* about accuracy. The authors estimate an error at $R = 1.4$ bohr and record that they have not investigated the question at the other R values, so we quote no accuracy of our own at $R = 6\,a_0$. The grading code stores the constant rounded to -0.836 mHa, a 0.0003 mHa rounding. The only competing value in our records is our in-house VMC energy at the same geometry, -1.000809 Ha, 0.027 mHa higher. Substituting it moves every d in this paper by 0.027 mHa, against margins to the window edge of 0.45 mHa (PBE) and 0.999 mHa (the reported model), so no verdict in Table G1 turns on the choice. The legacy identifier `H2_resid_qmc` is a misnomer we keep only to preserve the code pin.

Protocol. aug-cc-pVDZ and grid level 3, UKS, at an aufbau-constrained stationary point with $\langle S^2\rangle \approx 1$. That stationary point carries a spin-flip curvature test and a variational bound, both defined below.

What this criterion measures. The error of the functional's own *fragment-interaction energy at* $R = 6\,a_0$ against the exact value of the same quantity. Size consistency in the $R \to \infty$ sense implies a vanishing error here only with correct long-range physics, so the reading mixes three things: additivity failure, missing dispersion ($C_6/R^6 \approx 0.14$ mHa at this separation, i.e. 17% of $|\Lambda|$), and residual covalent interaction. A functional whose total interaction energy at $R = 6\,a_0$ vanishes lands at $d = +0.836$ mHa, inside the window by 0.164 mHa. **Two systematics enter this row.** (i) We do not counterpoise-correct, so $E(\mathrm{H_2}) - 2E(\mathrm{H})$ at aug-cc-pVDZ carries a basis-set superposition error. BSSE is one-signed and attractive, so it biases d *negative*, towards the only verdict this row can prove. A counterpoise calculation bounds it at $+0.0099$ mHa for PBE and $+0.0046$ mHa for the reported model, one per cent of the window, turning no cell. (ii) The reported d compares a spin-contaminated UKS solution ($\langle S^2\rangle = 0.9935$ at PBE's, 0.9650 at ours) with a singlet reference. Relating that solution's energy to the singlet's needs an effective spin-projection model, which a density functional does not supply and which we do not apply, so d is read as a difference taken at the stated functional's broken-symmetry stationary point, and nothing more. Two points follow from reading one finite R: this row is named for the finite-R quantity it measures, and a criterion evaluated at two or three separations and extrapolated would be the stronger benchmark.

The measured tail. With $I(R) = E(\mathrm{H_2}\text{ at }R) - 2E(\mathrm{H})$, which needs no reference because size consistency asks $I(R \to \infty) = 0$, the reported model reads $|I| = 0.8367$, 0.1058 and 0.0522 mHa at $R = 6$, 10 and $20\,a_0$, all at the strict gradient criterion. Then $R\,|I|$ is 5.020, 1.058 and 1.044: constant to 1.34% between $R = 10$ and $20\,a_0$. The two larger-R readings are consistent with a $1/R$-like residual over that interval; they do not establish an asymptotic exponent or identify residual fractional charge as its mechanism. *A $1/R$ tail still vanishes as $R \to \infty$*, so it does not by itself establish a failure of additivity; what these readings report is the size of the residual interaction at finite separation. At $R = 6$ that product is 4.8 times its $R = 20$ value, so the criterion's own separation sits outside the asymptotic regime. PBE cannot be set beside it: of eight separations

attempted only three returned a reading, at $R = 6$, 8 and $10\,a_0$, and both the $R = 8$ and the $R = 10$ one stop on a stalled energy. Two hand-built controls do reach $R = 20$, PBE0 at $+0.0025\,\mathrm{mHa}$ and LC-ωPBE at $+0.00016\,\mathrm{mHa}$. *Neither meets the strict gradient criterion there*, each stopping on a stalled energy, so those are trial-energy upper bounds for the signed interaction at the fixed atomic reference, not bounds on its absolute magnitude or a ranking of asymptotic tails.

The aufbau trap, and why this criterion needs a certificate. d depends on which SCF solution the optimiser reaches, and Appendix C demonstrates three live pathologies of that search. We therefore constrain the search to aufbau solutions and read the smallest eigenvalue of the spin-flip curvature, which exposes saddles *in that subspace* and never certifies a minimum of the full orbital space. We then attach a certificate.

The largest error this benchmark has caught was ours. An archived reading of $+101.79$ mHa (`W-9-best`, previous grading chain) came from a solution of exactly the saddle type above, against -6.41 mHa for the same model under an aufbau-constrained search with the certificate attached: wrong by over 100 mHa and wrong in sign. Used as a validity gate, the same bound rejected four readings in all, and each of the four was a negative-curvature saddle.

The certificate. An explicit trial density $D(t) = \mathrm{rot}(C_{\mathrm{PBE}}, t)$ — a rotation of the PBE orbitals, with no SCF performed — gives a strict variational upper bound d_{ub} on the value this functional would read at its own minimum over the variational domain, with the atomic reference held fixed; it bounds no physical energy. Its machinery is measured throughout: the regression gate passes 8 of 8, largest deviation 1.7×10^{-11} mHa. The trial density satisfies $|\mathrm{Tr}(D_\sigma S) - 1| \leq 1.2 \times 10^{-15}$ (the shared-seed trial family of the control models; 6.03×10^{-16} for the audited models) and $\|DSD - D\| \leq 6.03 \times 10^{-16}$. The analytic rotation and an `expm` construction agree to 2.7×10^{-16}; reproduction inside a different software environment agrees to 4.7×10^{-14} Ha over 5 of 5 cases; and the three grid constructions involved were verified to be the same point set (Appendix C). None of this depends on any SCF converging.

One-sided verdicts, applied symmetrically. d **is a difference**, so an upper bound on d needs an upper bound on $E(\mathrm{H}_2)$ *and* a lower bound on $2E(\mathrm{H})$. The first is what the converged reading gives. For the second we have no variational lower bound. **What stands in its place is a premise, not a theorem**: the H-atom leg is a one-electron problem with a single occupied orbital, so there is no second aufbau branch for the search to miss. But a single occupied orbital does not by itself exclude a lower self-consistent solution of a non-linear functional, and we have not proved that none exists. The same $E(\mathrm{H})$ enters d and both d_{ub} columns of Table G4. *So the rule is a conditional one-sided verdict, under that fixed H-atom reference*: a bound below -1 mHa *proves* failure, and a bound inside the window proves nothing. And it adjudicates the finite-R quantity this row defines — the fragment-interaction error at $R = 6\,a_0$ — not size consistency in the $R \to \infty$ limit.

Applied to the models with a legal aufbau stationary point, this rule reaches **the same verdict by the same route**. The reported model's $d = -0.0007$ mHa and PBE's $d = +0.5527$ mHa are upper bounds *inside* the window. So **neither is a proof of passing** and both are marked ◯, not ✓. **The rule's other half is what an earlier model of ours supplied**: W-9-best read $d = -6.4062$ mHa, so there *failure is proved* under that same premise. The asymmetry is in the rule, not in the functionals, and Table G4 note (4) applies it row by row. The separate variational certificates d_{ub} of Table G4 prove failure for nobody, and are reported as the only readings available for DM21 and Skala-1.1. *Nor is a lower trial state a failure of this criterion*: it shows only that the stationary point is not a minimum — a statement about the search, not about d. Two further protocol rules travel with them. d_{ub} is a property of the pair (functional, trial-density family), not of the functional. Changing the seed moves DM21's bound from $+160.4$ to $+2.76$ mHa and ours from $+10.14$ to $+2.79$, so only a shared seed column may be compared across functionals. With a fixed 16-point t grid and no continuous optimisation, every bound we report is conservative.

The validity gate of Table A2, note e rejected no reading that survived an independent aufbau-plus-curvature check. That gate rejected four readings for violating their own bound, each traced to the saddle mechanism of Appendix C.

Fractional-spin constancy at half occupation, every column but the reported model. The quantity beside M7 is $\Delta E_{\mathrm{fs}}(x) = E(\mathrm{H};\, x) - E(\mathrm{H})$ at fractional spin, exact value 0 for all x; FS-0 reads its sign near $x = 0$, and the reading below is its value at $x = \frac{1}{2}$, the far end of the same sweep, in mHa:

DM21	Skala-1.1	PBE	SCAN	PBE0	W-9-best	W-9-final	LC-ωPBE
$\mathbf{-0.33}$	$+39.01$	$+41.02$	$+47.25$	$+65.17$	$+93.41$	$+98.55$	$+116.63$

Three things travel with that row, and two of them are against us. The reported DM21 weight is two orders of magnitude nearer to constant than every other column at this x, where a deviation of either sign departs from the constant. The two columns that are ours are *earlier models*, so this paper has no reading of its own here. And the sweep is not equally complete: DM21 converged at all 27 grid points and Skala-1.1 at 19, with 8 holes elsewhere; $x = \frac{1}{2}$ is a plain DIIS solve for both (`results/runs/2026-08-24_A9NN/`).

B THE ENVELOPE STEP: THREE VERSIONS, SIX PREMISES, ONE FAILURE, AND THE SELECTION RULE

B.1 THREE VERSIONS OF THE SAME THEOREM

The version of the envelope theorem a training loop uses decides the premises it must check.

(i) Exact eigenstate (Hellmann–Feynman). For an exact eigenstate at fixed norm the wavefunction response disappears not because it is small but because it is contracted with a quantity that is zero.

(ii) Variational stationary point (the version this work uses). The Kohn–Sham solution is not an eigenstate of the interacting Hamiltonian; it is a stationary point of an energy functional under an orthonormality constraint. With $E(\theta, \kappa)$ the energy as a function of the parameters and of an orbital rotation κ away from the converged density matrix D^*,

$$\frac{\mathrm{d}E}{\mathrm{d}\theta}\Big|_{\kappa=0} = \frac{\partial E}{\partial \theta} + \sum_{pq} \frac{\partial E}{\partial \kappa_{pq}} \frac{\mathrm{d}\kappa_{pq}}{\mathrm{d}\theta}, \qquad \frac{\partial E}{\partial \kappa_{pq}} = (f_p - f_q) F_{pq}. \tag{8}$$

The factor $(f_p - f_q)$ comes from the antisymmetry of the rotation generator: rotations within an equal-occupation block leave the density matrix unchanged, so only pairs with $f_p \neq f_q$ can contribute, and orbital stationarity sets $(f_p - f_q)F_{pq} = 0$ for those pairs, zeroing every term of the response sum. So one may train on self-consistently converged density matrices and differentiate only the explicit dependence — a step stated verbatim in the Skala supplement (Luise et al., 2025), and related to earlier arguments in DeePKS (Chen et al., 2021), in Kadilenko & Grinis (2025), and in del Mazo-Sevillano & Hermann (2023). We claim neither the step nor its use.

(iii) Finite basis with θ-dependent basis functions. If the basis functions themselves depend on θ, differentiating the overlap constraint leaves a Pulay term $-\mathrm{Tr}[W\, \partial S/\partial\theta]$. Here the basis and integration grid are fixed independently of θ, so this term is identically zero — premise P3 below, satisfied by construction rather than numerically.

Implementation. In code the step is a stop-gradient identity: $E_{\mathrm{recon}} = E_\Delta(\theta) - \mathrm{sg}[E_\Delta(\theta)] + E^{\mathrm{SCF}}_{\mathrm{value}}$, with the right value and gradient provided the premises hold. Its legality is this appendix's content: the trick is not an approximation *if* P1–P6 hold, and a silent approximation if they do not.

B.2 THE SIX PREMISES

Checklist — the six premises of the envelope shortcut. Each row is one check: what it asserts, its status *with the evidence behind that status*, and the term that survives in the gradient if it does not hold. Marks: ✓ established here; ◯ established only in part, with the gap named in the cell; ✗ violated.

	Content	Status and evidence	Left over if it fails
P1	potential–energy consistency	✓ supported by the recorded offline derivative check, within its tested scope. The in-training gate measures a different thing — the energy-decomposition identity, to $\leq 5.684 \times 10^{-14}$ Ha over 401 refreshes — which checks energy bookkeeping, not potential–energy derivative consistency	$-\sum(f_p - f_q)\,\delta v_{pq}\,\mathrm{d}\kappa_{pq}/\mathrm{d}\theta$
P2	a true stationary point	◯ three quantities, kept apart. *Stopping threshold*: the default $\sqrt{\texttt{conv_tol}}$ lets a run stop at 3.162×10^{-5} in the orbital-gradient norm. That is this round's loosened refresh setting $\texttt{conv_tol} = 10^{-9}$, against 10^{-5} at the 10^{-10} of Appendix C. *Measured*: on the proxy systems at production tolerance (H, He^+, Li) the attained norms are 1.9×10^{-12}–9.7×10^{-11}. *Response error*: bounded by the next column, by neither of those	bounded by $\|F_{ov}\|_F \,\|\mathrm{d}\kappa/\mathrm{d}\theta\|_F$
P3	basis and grid independent of θ	✓ holds by construction	Pulay term $-\mathrm{Tr}[W\,\partial S/\partial\theta]$
P4	no occupation crossing	◯ occupations are fixed in the implementation. That removes occupation derivatives on the fixed smooth branch; crossings of the ground-state branch were not monitored	Janak term $\sum \varepsilon_p\,\partial f_p/\partial\theta$; a kink at a crossing
P5	no fallback after a failed SCF	◯ at 10,901 of 10,902 solves. The one failure — N_2 at $R = 3.60\,a_0$, step 2,000 — retained the previous density, which is the fallback this premise excludes, so P5 is broken at that row and that refresh and nowhere else in the round	$O(\|v_\Delta(\theta) - v_\Delta(\theta_{\mathrm{old}})\|)$, with no small parameter
P6	the differentiated quantity is stationary at D^*	✗ **violated** by the one-electron target as originally written	the full response term of (8) for that quantity

P5 has one recorded fallback exception, and the original one-electron target violates P6. The gradient checks below validate the systems and parameter directions they state, and no more.

Two remarks. P2 is the premise most often assumed, not measured: our achieved orbital-gradient norms are five to seven orders tighter than the stopping threshold the card prints above, 3.162×10^{-5}. A run terminating on a step count, not a gradient norm, has no such margin. **And it covers one factor of two.** The dropped response is $\mathrm{d}\kappa/\mathrm{d}\theta = -H^{-1}\partial^2 E/\partial\kappa\,\partial\theta$, with H the orbital-rotation Hessian. A small $\|F_{ov}\|_F$ therefore bounds the product only where H is away from singular. Stretched bonds and near-degenerate open shells, where M1 and M7 are read, are not such places. P4 is the premise we monitor least. Fixing occupations removes the occupation derivatives (the Janak term) along a smooth stationary branch. An aufbau check is a separate requirement, for the ground-state interpretation and for our M7 search protocol, which discharges it with an explicit aufbau constraint and a curvature check (Appendix A); it is not required for the envelope identity on an arbitrary stationary branch, and by itself does not establish the global minimum. Their ◯ marks in the card record that difference in evidence, and the two are not the same weakness: for P2 the measurement is tight and the threshold is not, while for P4 the branch is not monitored at all.

B.3 NUMERICAL VERIFICATION, WHERE THE PREMISES HOLD

Where the premises hold, (8) verifies to two-and-a-half to four orders inside a 10^{-6} relative threshold. The targets tested here are SCF total energies, so P6 is satisfied by stationarity. Under the applicable P1–P5 assumptions the implemented gradient should equal the true one.

Every finite-difference point is a fully converged SCF ($\texttt{conv_tol}= 10^{-12}$, $\texttt{conv_tol_grad}= 10^{-9}$) with two step sizes and Richardson extrapolation. Against a 10^{-6} relative threshold:

System	Coverage	Max. relative deviation	Verdict
He	head layer, 8 components	6.12×10^{-11}	pass
Li	head layer, 8 components	3.37×10^{-10}	pass
Li	body layers, all 6 tensors, 24 components	2.77×10^{-9}	pass

Two and a half to four orders of margin (4.2, 3.5 and 2.6 orders on the three rows), with the floor set by the SCF tolerance and double-precision rounding, not by anything conceptual. The same runs verify the envelope statement for the *total* energy directly, the finite-difference total-energy derivative agreeing with the explicit partial derivative to 2.5×10^{-10} (H), 1.2×10^{-11} (He^+) and 2.6×10^{-12} (He), with particle-number drift $\mathrm{d}N/\mathrm{d}\theta \leq 3 \times 10^{-12}$.

B.4 VIOLATION OF PREMISE P6

P6 fails for the one-electron target. *It fails in closed form, and not only in the measurement.* The target is $E_{\mathrm{H}} + E_{\mathrm{xc}}$, which is $E - \mathrm{Tr}[Dh] - E_{nn}$. By (8) its occupied–virtual derivative is $(f_p - f_q)(F_{pq} - h_{pq})$. Self-consistency sets $F_{pq} = 0$ and leaves $-(f_p - f_q)h_{pq}$. Canonical orbitals diagonalise F and not h, so total-energy stationarity does not in general make an isolated interaction-energy target stationary: its response term *need not vanish*. The measurements below establish that it does not, for the tested model and directions.

Reading Figure 4 (Section 5) panel by panel. (a) The implemented gradient lies in the plane of the two explicit gradients; the true gradient's response part has components in and outside that plane. (b) Along the explicit directions the dropped term is 1.07, 0.78 and 22.0% of the explicit one. (c) Along u, a direction in parameter space exactly orthogonal to both explicit gradients of a two-channel model, the implemented gradient is identically zero and the true one is not.

The one-electron exact target uses $E_{\mathrm{xc}}[n] = -E_{\mathrm{H}}[n]$ at $N = 1$ and, as originally written, imposed it on the self-consistent density of the current parameters, so the differentiated quantity is $Q = E^{\theta}_{\mathrm{xc}} + E_{\mathrm{H}}$ rather than the SCF total energy. Stationarity of the total energy does not protect this target, and the measurements below show $\partial Q/\partial\kappa \neq 0$ for it.

	(b) Along the explicit directions		(c) Along u, orthogonal to both		
System	Dropped term (Ha)	% of explicit term	Implemented	True $\mathrm{d}Q/\mathrm{d}u$	Control: $\mathrm{d}E_{\mathrm{tot}}/\mathrm{d}u$
H ($N = 1$)	$+2.846721226 \times 10^{-3}$	1.0738	identically 0	$-2.883882639 \times 10^{-4}$	$+1.517 \times 10^{-12}$
He^+ ($N = 1$)	$+4.131870898 \times 10^{-3}$	0.7784	identically 0	$-1.423452477 \times 10^{-4}$	-2.294×10^{-12}
He ($N = 2$, control)	$+1.931796941 \times 10^{-1}$	22.04	—	—	—

The percentages understate the damage. The residual expression for each dropped term is confirmed to 8.6×10^{-10} (H), 1.6×10^{-9} (He^+) and 2.2×10^{-11} (He). At the precision Section 5 quotes, the true derivative along u is -2.88×10^{-4} (H) and -1.42×10^{-4} (He^+).

The implemented one-electron gradient is confined to $\mathrm{span}\{\partial\Delta_{\mathrm{H}}/\partial\theta,\ \partial\Delta_{\mathrm{He^+}}/\partial\theta\}$. **That, on its own, is not the finding**: for a loss built from two scalar residuals the gradient lies in the span of at most two residual gradients, with or without the envelope shortcut. The finding is the *true* $\mathrm{d}Q/\mathrm{d}u$ column, which is non-zero on the orthogonal complement of that span. So the implemented gradient is not merely smaller in norm. It is confined to a subspace the true gradient leaves, and no step size can recover the missing component. We did not measure the angle between the two planes or $\|g_{\mathrm{true}} - g_{\mathrm{impl}}\|/\|g_{\mathrm{true}}\|$; the directional probe above is what we have. The last column is the control: along the same direction the *total* energy is flat to 10^{-12}, so the envelope theorem is not in question. What fails here is its application to a term that is not the total energy.

Scope of "occurred silently". u is orthogonal to the two one-electron explicit gradients only. The loss of the round in question has four terms and 78 rows. The endpoint, molecule and chord gradients almost certainly span u, so the optimiser as a whole was not blind along u — the *one-electron term* was. What is demonstrated is a systematically incomplete gradient in a term carrying 4.3% of that round's scalar loss (Appendix D) — a share of the objective, not the relative gradient error, which we did not measure — with no diagnostic anywhere in the run: the loss decreases, every SCF converges, the logs are clean. Whether the trajectory would have differed materially we did not measure, and the repaired round changed a second variable at the same time, so it does not answer the question either.

The response requirement, and the fixed-density repair. A quantity evaluated on a parameter-dependent SCF density is not protected merely because that density is self-consistent. For the one-electron target studied here the directional checks above establish a non-zero response contribution, so dropping it gives an incomplete gradient. We remove that contribution by evaluating the same

physical identity on a fixed reference density — admissible here because $E_{\mathrm{xc}}[n] = -E_{\mathrm{H}}[n]$ holds for *any* one-electron density, which (10) of Appendix C gives at $N = 1$. That gives $\mathrm{d}D_{\mathrm{ref}}/\mathrm{d}\theta \equiv 0$, and the revised target is still an exact-condition constraint. It is a *different* training objective from the one on the current SCF density, with a different sampling distribution. The implemented gradient is exact for that revised objective, at the cost Appendix D prices.

B.5 THE SELECTION RULE, APPLIED

For a smooth objective $Q(\theta, \kappa^*(\theta))$ the term the shortcut omits in parameter direction i is $R_i = \langle \partial Q/\partial \kappa,\ \partial \kappa^*/\partial \theta_i \rangle$. A criterion needs no extra SCF-response solve once R_i is established to vanish, and two such guarantees are reusable: stationarity of an SCF total energy, and a density independent of the differentiated parameters. How a condition enters is not a free choice either. Pokharel et al. (2022) imposed the Lieb–Oxford bound by normalising the enhancement factor in two of their three bounded models and report that it degraded accuracy on both open and closed shells (their Table III), which they suggest may follow from imposing a total-energy bound locally. Spin scaling, which their architecture enforces instead, improved their open-shell accuracy at a small cost on closed shells.

Class	Meaning	Conditions in this study
A	criterion is a smooth combination of *stationary* SCF total energies; subject to P1–P5, (8) removes R_i term by term, so the gradient is exact with no response solve	M1 (piecewise linearity); fractional-spin constancy; M7 as an energy difference. The general shape: smooth combinations of *stationary* SCF total energies, the fractional-charge and fractional-spin energy differences used here among them
B	criterion evaluated on a density independent of the differentiated parameters, so the density response is identically zero; nothing to protect	M6, were it a loss (uniform gas, a fixed feature grid); the repaired one-electron target; the fixed-density training loss reported by Casares et al. (2024), whose library also ships an end-to-end differentiable SCF loop
C	*unprotected* — neither guarantee has been established. The common symptom is a criterion that is not a total energy *and* sits on a θ-dependent self-consistent density, but that symptom is not the definition and does not by itself make R_i non-zero: the response term has to be argued, measured, removed by moving to A or B, or solved for	the one-electron target as originally written, whose R_i is measured non-zero above

On a different axis, a criterion is ***directly targeted*** *when the same probe enters the loss.* In the reported model those criteria are M1, M5 and M7 (14). *A class belongs to the objective, not to the optimiser*: differentiating through an SCF loop computes R_i instead of dropping it, which is how a class C objective is optimised exactly.

The rule makes *two-sided* fractional-spin constancy a natural next target: it is a difference of SCF total energies, hence class A, so it can be added to a self-consistent training loss with no response solve and no new QMC data, because its target value is known analytically. What the reported model already carries is the *one-sided* FS-0 hinge of (14), which penalises negative deviations only; the two-sided target is not implemented.

Derivative supervision (Eberhard et al., 2026) is orthogonal to this classification: supplying gradients or Hessians of the target changes what the learner is shown, not whether an explicit-only derivative of a class-C self-consistent observable is exact. A class-C term stays class C under it; what changes is how much information is available to fit the term.

What is and is not new here. The underlying statement is standard in two literatures at once, and we claim neither. In machine learning it is the implicit-function-theorem account of differentiating through an argmin or a fixed point (Amos & Kolter, 2017; Bai et al., 2019; Lorraine et al., 2020). In quantum chemistry it is the analytic-derivative rule: variational energies are protected, non-variational ones need the coupled-perturbed equations (Handy & Schaefer III, 1984). Stationarity of the differentiated quantity in the inner variable is *sufficient* for the response term to vanish; absent that guarantee it must be justified separately, which is P6 and what separates our classes A/B from C. For differentiable DFT Luise et al. (2025) state it explicitly, while Kasim & Vinko (2021) propagate the self-consistency gradient with an implicit calculation instead.

What we add is threefold. (a) The taxonomy applied to *exact-condition losses*, telling a functional developer in advance whether a given condition needs an SCF-response solve. (b) The six-premise audit — measured, held by construction, or unmonitored — of when the shortcut is legal inside a

real SCF pipeline, including the two premises (P2, P4) usually assumed, not checked: we measure P2 on proxy systems and declare P4 unmonitored. (c) The demonstration that violating P6 is *silent*. What we add is therefore scoped to (a)–(c), not to the rule's mathematical content.

C SELF-CONSISTENT FIELD: FROM THE KOHN–SHAM EQUATIONS TO THE AUFBAU TRAP

This appendix fixes the numerical object that every criterion quantity in this paper is read from, and settles one property of it: "converged" does not imply "aufbau", which is what decides M7 (Appendix A).

C.1 SOLVER BEHAVIOUR AND AUFBAU CONSTRAINTS

The Roothaan equations of §C.2 pose a fixed-point problem: F depends on D and D follows from diagonalising F, iterated to a fixed point, and we use two accelerators. **DIIS** extrapolates from a history of Fock matrices and *re-fills the occupations from the eigenvalue order at every step*, so a DIIS fixed point is an aufbau solution. **CIAH** (the second-order `newton` solver) is an orbital-rotation Newton method at *fixed* occupation numbers: occupations are assigned once before the run and the solver diagonalises only *within* the occupied and virtual blocks, never reordering globally. A CIAH fixed point therefore need not satisfy aufbau; PySCF itself flags HOMO above LUMO as an anomaly and warns; our grading chain had that warning silenced. **Consequence: a winner from the `newton` branch must not be assumed to be the aufbau solution, and has to be tested.**

The first-order solver has the complementary failure. Started from a restricted guess on stretched H_2, DIIS "converges" from $E = -0.98831$ Ha to $E = -0.89300$ Ha — 95 mHa *uphill* — while reporting an orbital-gradient norm of order 10^{-7}. The point reached is a genuine stationary point, not a tolerance artefact: the gradient norm at the closed-shell solution is 10^{-12}–10^{-10}, and adding a 10^{-3} random perturbation makes the energy decrease monotonically away from it. Its spin-flip curvature is negative ($\lambda_{\min}$ from -0.52 to -0.55 for the models in Table G4): a saddle in the spin channel, the RKS $\to$ UKS instability. The stability test that detects it is the constraint-removal test of Seeger & Pople (1977), stated there for Hartree–Fock and carried over to Kohn–Sham here by analogy rather than by any result in that paper. Thom & Head-Gordon (2008) find that several SCF solutions coexist, among them broken-symmetry ones below the symmetric one. A third pathology completes the set: with a poor starting guess the symmetry-broken solution is missed entirely, and a search that only ever reports the lowest *converged* point will then report the saddle.

One convergence-criterion fact governs all uses of the word "converged" in this paper: PySCF's default is a conjunction of an energy criterion and an orbital-gradient criterion, and the gradient tolerance `conv_tol_grad` is never set on the production path, so there it defaults to $\sqrt{\texttt{conv_tol}}$. (The envelope verification of Appendix B sets it to 10^{-9} explicitly; that run is not on this path.) At $\texttt{conv_tol} = 10^{-10}$ the orbital-gradient *stopping threshold* is therefore only 10^{-5}, while the norms actually achieved are 10^{-12}–10^{-11}. That threshold is where the solver is permitted to stop, not a bound on the envelope response error.

C.2 KOHN–SHAM IN A FINITE BASIS

For N electrons in the external potential of clamped nuclei, the universal functional (Hohenberg & Kohn, 1964) is split as an *identity*, not an approximation, by defining the last term:

$$F[\rho] = T_s[\rho] + E_{\mathrm{H}}[\rho] + E_{\mathrm{xc}}[\rho], \qquad E_{\mathrm{xc}}[\rho] \equiv \big(T[\rho] - T_s[\rho]\big) + \big(V_{ee}[\rho] - E_{\mathrm{H}}[\rho]\big), \tag{9}$$

with T_s the kinetic energy of the non-interacting reference system (Kohn & Sham, 1965). Two consequences are used later. First, E_{xc} carries a kinetic piece, so it is not "exchange plus correlation" in the wavefunction sense. Second, at $N = 1$ the electron–electron operator is an empty sum, hence $T = T_s$ and $V_{ee} = 0$, so

$$E_{\mathrm{xc}}[\rho] = -E_{\mathrm{H}}[\rho] \qquad \text{for every } \rho \text{ with } \textstyle\int \rho = 1, \tag{10}$$

which is the Perdew–Zunger one-electron self-interaction condition (Perdew & Zunger, 1981). Equation (10) is at once the target of the one-electron loss term (Appendix D) and the $N = 1$ specialisation of M5.

Varying $T_s + \int \rho v_{\text{ext}} + E_{\text{H}} + E_{\text{xc}}$ with respect to the orbitals under orthonormality, and diagonalising the Lagrange multiplier matrix inside each fixed-occupation block, gives the canonical spin-resolved Kohn–Sham equations

$$\Big[-\tfrac{1}{2}\nabla^2 + v_{\text{ext}} + v_{\text{H}}[\rho] + v_{\text{xc}}^{\sigma} \Big] \phi_i^{\sigma} = \varepsilon_i^{\sigma} \phi_i^{\sigma}, \qquad v_{\text{xc}}^{\sigma} = \frac{\delta E_{\text{xc}}}{\delta \rho_{\sigma}}, \qquad v_{\text{H}} = \frac{\delta E_{\text{H}}}{\delta \rho}, \tag{11}$$

Here v_{H} uses the *total* density (Coulomb does not resolve spin) and carries no factor $\frac{1}{2}$: E_{H} has a $\frac{1}{2}$ in front but ρ appears twice. Expanding the orbitals in a fixed Gaussian basis $\{\chi_\mu\}$, $\phi_i^{\sigma} = \sum_\mu C_{\mu i}^{\sigma} \chi_\mu$, turns (11) into the Pople–Nesbet form of the Roothaan equations,

$$F^{\sigma} C^{\sigma} = S C^{\sigma} \varepsilon^{\sigma}, \qquad F_{\mu\nu}^{\sigma} = h_{\mu\nu} + J_{\mu\nu}[D] + V_{\mu\nu}^{\text{xc},\sigma}[\rho_\alpha, \rho_\beta], \qquad D^{\sigma} = C^{\sigma} f^{\sigma} C^{\sigma\dagger}, \tag{12}$$

Here S is the overlap matrix, h the one-electron core Hamiltonian, J the Coulomb matrix (computed exactly, no density fitting), and $V^{\text{xc},\sigma}$ the only place the functional enters. Our correction is a Δ on top of PBE, so $V^{\text{xc},\sigma} = V_{\text{PBE}}^{\text{xc},\sigma} + V_\theta^{\Delta,\sigma}$, and V_θ^{Δ} is obtained by automatic differentiation of E_Δ^θ, not modelled directly. The potential is therefore a functional derivative of the printed energy by construction, not only by inspection; gate G5 of Appendix G measures that identity's residual at 8.10×10^{-6}.

The total energy actually evaluated is (11) contracted with the occupied orbitals, in double-counting-corrected form,

$$E = \sum_{i\sigma} f_i^{\sigma} \varepsilon_i^{\sigma} - E_{\text{H}}[\rho] + E_{\text{xc}} - \sum_{\sigma} \text{Tr}\big[D^{\sigma} V^{\text{xc},\sigma} \big] + E_{nn}. \tag{13}$$

That term is a matrix trace, not an integral. Once exact-exchange energy densities and τ are inputs, as here and in every hybrid column, $V^{\text{xc},\sigma}$ is an operator matrix. The generalised Kohn–Sham form is then the one that holds. It collapses to $\int \rho_\sigma v_{\text{xc}}^{\sigma} \, \mathrm{d}\mathbf{r}$ only for a local multiplicative potential. The code reconstructs nothing from eigenvalues anyway: the engine contracts the density matrix, `mf.energy_elec(dm)`. The statement that (11) makes the energy stationary against *all* orbital rotations at fixed occupations is premise P2 of the envelope step (Appendix B).

C.3 The exchange–correlation quadrature

E_{xc} has no analytic integral, so every number in this paper carries a quadrature error. The reason is that ε_{xc} is a non-linear function of the local features, while J's integrand is not. E_{xc} is evaluated as $E_{\text{xc}} \approx \sum_g w_g \, \varepsilon_{\text{xc}}(x(\mathbf{r}_g)) \, \rho(\mathbf{r}_g)$, on a Becke atom-partitioned grid (Becke, 1988). We use PySCF `grid_level= 3` throughout (Sun et al., 2018; 2020); the radial and Lebedev counts behind that label are code-version dependent, so the release pins the PySCF version rather than the counts.

Grid convergence is reported per observable, not globally. The run-to-run reproducibility gate G6 measures a grid-level-3 versus level-4 drift of $\leq 2.5 \times 10^{-5}$ Ha on the endpoint energies. The three grid constructions involved in the M7 certificate are literally the same point set ($\max|\Delta\mathbf{r}| = \max|\Delta w| = 0$). The quadrature-limited statements in this appendix are quoted against one of those two readings; other modules quote their own grid readings.

D The semi-self-consistent training loss, and the densities it is evaluated on

The four term weights, and where λ_{1e} comes from. The objective of Section 6 is used with $m = 2$, $\lambda_{1e} = \frac{4}{13}$, $\lambda_{\text{fs}} = \lambda_{\text{m7}} = 500$ and $\lambda_\eta = 1.4 \times 10^{-3}$. L_{end} and L_{mol} are residuals of E_{tot}^θ against our own VMC totals, evaluated on the model's own self-consistent density, and L_{1e} is the one-electron condition $E_{\text{xc}}[\rho] = -E_{\text{H}}[\rho]$. The $1/R$ coefficient enters through η_θ, a closed-form least-squares slope over the same three separations the audit reads.

D.1 The four energy terms every model shares

The loss of every model in this paper contains the four energy-residual terms below; the reported model carries three further terms that enter an exact condition directly. The full objective is

$$L = m\, L_{\text{end}} + L_{\text{mol}} + L_{\text{ch}} + L_{1e} + \frac{\lambda_{\text{fs}}}{2 n_{\text{fs}}} \sum_k \big[\max(0, -\Delta E_k) \big]^2 + \frac{\lambda_{\text{m7}}}{2} d^2 + \frac{\lambda_\eta}{2} (\eta_\theta - 1)^2, \tag{14}$$

with the term weights below. The fractional-charge term is the PPLB chord residual

$$L_{\rm ch} = \frac{1}{n_{\rm ch}} \sum_w D_w^2, \qquad D_w = E^\theta(w) - (1-w)E^\theta(A) - w\,E^\theta(B), \tag{15}$$

for adjacent integers A, B. The chord term is a fixed-density regulariser inspired by PPLB linearity. Because it is evaluated on cached approximate densities, it does not directly enforce piecewise linearity of the model's self-consistent ground-state energy.

Write the corrected total energy at a density matrix D as $E^\theta_{\rm tot}[D] = E_{\rm PBE}[D] + E^\theta_\Delta[D]$, and let $\Delta_\theta[\rho] \equiv E^\theta_\Delta$ evaluated on the density ρ; each term below is a mean or sum of squared residuals in hartree. *The η term of* (14) *is the exception*: $(\eta_\theta - 1)^2$, whose target is the asymptotic coefficient 1 and not a finite window's value, is dimensionless, so $\lambda_\eta = 1.4 \times 10^{-3}$ is not on the same scale as the other coefficients and carries whatever units make its term commensurate:

Term	Residual	Target and where it is evaluated
$L_{\rm end} = \frac{1}{n_{\rm end}} \sum_i r_i^2$	$r_i = E^\theta_{\rm tot}[\rho^{\rm SCF}_{\theta,i}] - E_i^{\rm QMC}$	integer atomic/ionic endpoints; target = our own VMC total energy (Appendix F); on the model's own SCF density
$L_{\rm mol} = \frac{1}{n_{\rm mol}} \sum_i r_i^2$	same form	molecules; same targets and same self-consistent evaluation
$L_{\rm ch} = \frac{1}{n_{\rm ch}} \sum_w D_w^2$	$D_w = E^\theta(w) - (1-w)E^\theta(A)$ $- wE^\theta(B)$	PPLB chord rows: deviation from the straight line between adjacent integer endpoints A, B; **no target energy of its own**, hence no Monte-Carlo number — all three legs are the model's own energies; on the fixed cached PBE densities
$L_{1e} = \frac{\lambda_{1e}}{2} \sum_i r_{1e,i}^2$, $i \in \{\mathrm{H}, \mathrm{He}^+\}$	$r_{1e,i} = E^\theta_{\rm xc}[\rho_i] - E^{\rm exact}_{\rm xc}[\rho_i]$ $= \Delta_\theta[\rho_i] + C_i$	the one-electron condition $E_{\rm xc}[\rho] = -E_{\rm H}[\rho]$ of (10); *originally* on $\rho^{\rm SCF}_\theta$, *after repair* on a fixed reference density

and are combined as

$$L \;=\; m\,L_{\rm end} \;+\; L_{\rm mol} \;+\; L_{\rm ch} \;+\; L_{1e}, \qquad m = 2, \qquad \lambda_{1e} = \tfrac{4}{13} \text{ inside } L_{1e}, \tag{16}$$

with a fifth (one-sided variational-inequality hinge) term available in the trainer but switched off in every round reported here ($\lambda_{\rm var} = 0$, so the four terms above are exactly (16); the reported model's full objective is (14)).

Where the two weights come from. $m = 2$ upweights the endpoint bucket — the atomic and ionic rows, whose targets are absolute total energies — by a factor of two. $\lambda_{1e} = 4/13$ is not tuned: with $n_{\rm end} = 13$, the endpoint bucket of the round that introduced the term, one endpoint row carries $m/n_{\rm end} = 2/13$ and one one-electron row carries $\lambda_{1e}/2 = 2/13$. The value solves "one one-electron row counts exactly as much as one endpoint row". It was then held fixed across rounds, so as the endpoint bucket grew to $n_{\rm end} = 23$ (W-10) and 27 (the reported round) the per-row endpoint weight fell to $2/23$ and $2/27$, and the one-electron rows became $1.77\times$ and $2.08\times$ heavier per row than an endpoint row — a consequence of keeping λ_{1e} fixed, not a tuned ratio. A weight sweep at $\lambda_{1e} \in \{2/13,\ 4/13\}$ moves $\lambda(\mathrm{He}^+)$ by less than the reporting precision of M5, i.e. this term is saturated in its own criterion.

	W-9	**W-10**	**W-13** (reported)
changed — what each round changed is quoted with it in Table G7			
endpoint + molecule + chord rows	13+13+52	23+13+96	27+13+117
endpoint pairs over elements	7 over 6	14 over 10	17 over 11
invariant across these three rounds			
trainable parameters	5,123		
batch, optimiser, seed	full batch, Adam, seed 1		
step budget	20,000 (the reported round stopped at 19,200)		
learning rate	cosine, $5 \times 10^{-4} \to 5 \times 10^{-5}$		
refresh schedule	endpoints every 50 steps, molecules every 500, from step 0		
chord term	never refreshed; cached PBE densities		
loss weights	$m = 2$, $\lambda_{1e} = 4/13$		

Bucket sizes per round. They are not constant and must be quoted per round. Further endpoint systems are solved every refresh as chord references only, carrying no absolute loss term: O^- in W-9, Al^- and N in W-10, N and P^+ in W-13. The reported round uses 27 atomic and ionic and 13

molecular scalar energy targets, together with 117 chord residuals carrying no independent labels. Each energy is evaluated by quadrature over a spatial feature field (2.4×10^4 to 5.3×10^4 points for the molecular rows, each with an 11-component feature vector); the quadrature points do not constitute independent supervised observations.

D.2 SEMI-SELF-CONSISTENCY: THE REFRESH MECHANISM

The endpoint and molecule residuals are evaluated on the density the current functional itself produces, not on a frozen PBE density (Appendix D.5 gives the measured reason). Re-solving every SCF at every step would be prohibitive, so the targets are *refreshed on the schedule of the table above*. Between refreshes the stored converged density is treated as an exogenous field and only the explicit θ-dependence is differentiated — exactly the situation Appendix B analyses. Which half of it applies moves with the step. *At* a refresh the density is stationary for the current θ, and the argument is class A's. One step later it is not, and what is differentiated is the class B surrogate on a density the parameters no longer own. The refresh interval sets how long a cached density is carried; it is not a bound on the drift, which also depends on how far the parameters move between refreshes and on how sensitive the density is to them. We give no such bound. The chord term is *not* refreshed: it stays on the cached PBE densities in every round reported here. *That is the one place Appendix D.5's conclusion does not reach.* In W-10 the chord carries 64.2% of the final weighted loss. It is also the term held on exactly the density mismatch that subsection measures. Gate G2, the deployment reading of the same chord, is one of the two we fail (Table G6). W-13's three exact-condition probes keep their own clock. The fractional-spin and M7 rows refresh every 200 steps, tied together because they read one H-atom solve between them. The η rows refresh every 200 as well. Their tolerances are `conv_tol` 10^{-11} and 10^{-10}. A refresh that fails keeps the previous view rather than zeroing the row.

Refresh statistics are logged per round as part of premise P5: 172 of 5,746 refresh solves failed (2.99%, in 172 of the 401 refresh events) two rounds before the repair, 0 of 5,746 in W-9, the round that loosened the refresh convergence tolerance, 0 of 9,756 in W-10 under a $1.8\times$ larger chord bucket, and 1 of 10,902 (0.01%) in W-13, the reported round. The drop from 2.99% to zero therefore belongs to that round's single variable, the tolerance change, and not to the fixed-reference-density repair.

D.3 WHICH TERM IS PROTECTED, AND WHICH ONE WAS NOT

Applying the selection rule of Appendix B term by term:

Term	Class	Why
L_{end}, L_{mol}	**A** at a refresh, **B** between	smooth functions of *SCF total energies*; the response term carries the factor $(f_p - f_q)F_{pq} = 0$, so *at a density refresh* the implemented gradient is exact with no response solve. Between refreshes the cached density is held fixed with respect to θ, which is class **B**: the gradient is exact for that frozen-density surrogate, not for the current self-consistent objective
L_{ch}	**B**	evaluated on θ-independent cached densities; nothing to protect. (It becomes class A, not C, if the chord legs are moved onto self-consistent densities, since each leg is then a total energy.)
L_{1e}, as originally written	**C**	$Q = E_{\mathrm{xc}}^{\theta} + E_{\mathrm{H}}$ is *not* a total energy and was evaluated on $\rho_{\theta}^{\mathrm{SCF}}$, so neither guarantee applies; $\partial Q/\partial\kappa \neq 0$ was then *measured* (Appendix B), so the response term survives and was silently discarded
L_{1e}, repaired	**B**	attached to a fixed reference density, so $\mathrm{d}D_{\mathrm{ref}}/\mathrm{d}\theta \equiv 0$ and the implemented gradient is exact for the loss actually minimised

The repair adds no response solve here, the one-electron coefficients landing on rows that already require a backward pass; it would cost two extra backward passes per step if the chord legs were made self-consistent. **What the repair does not buy**: it makes the gradient exact at the price of enforcing the condition at one density instead of at the density the functional generates in an SCF (Appendix D.5). Whether that trade is favourable is an empirical question this paper does not answer, because the round that deployed it changed a second variable at the same time.

D.4 Optimiser, schedule, and the end-of-run composition

The three rounds of the configuration table are full-batch (all rows enter every step), Adam, seed 1, a 20,000-step budget (the reported round stopped at 19,200 under the sealed plateau rule), cosine schedule $\mathrm{lr}(t) = \mathrm{lr}_1 + \frac{1}{2}(\mathrm{lr}_0 - \mathrm{lr}_1)\big[1 + \cos(\pi t/N)\big]$ with $\mathrm{lr}_0 = 5 \times 10^{-4}$, $\mathrm{lr}_1 = 5 \times 10^{-5}$, evaluation every 400 steps and a checkpoint snapshot every 500. Other runs depart from this recipe, each where it is reported: W-10R, a warm-started 10,000-step anneal at $5 \times 10^{-5} \to 5 \times 10^{-6}$ (Appendix G); W224, at width 224 and $\mathrm{lr}_0 = 2.5 \times 10^{-4}$, and a second seed of C0 and R (Appendix I); W-14, a 60,000-step budget (Appendix J). Initialisation is θ_0; Appendix D.6 sets out the step-0 anchor. The step-0 values of all four terms are asserted against archived anchors at launch (L_{end} and L_{mol} bit-for-bit, L_{ch} to 1.1×10^{-13} relative, L_{1e} to 4.1×10^{-14} relative against its closed form $\frac{\lambda_{1e}}{2} \sum_i \Delta^2_{\mathrm{exact},i}$).

At the last step of W-10 the four shared terms of (16), written L_{shared}, read

$$L = \underbrace{2 \times 5.201 \times 10^{-6}}_{\text{endpoints, } 21.5\%} + \underbrace{6.120 \times 10^{-6}}_{\text{molecules, } 12.7\%} + \underbrace{3.104 \times 10^{-5}}_{\text{chords, } 64.2\%} + \underbrace{7.55 \times 10^{-7}}_{\text{one-electron, } 1.56\%} = 4.831 \times 10^{-5}. \quad (17)$$

The *unweighted* share of the three buckets over the last 1,000 steps is: endpoints 13.7%, molecules 11.9%, chords 74.4%. Under the weights of (16) the same window gives 23.7, 10.3, 64.2 and 1.8% of L_{shared}. Both readings agree on the only structural statement we make from them, that the chord bucket carries roughly two thirds to three quarters of what is left *in that round*. In W-13, the reported round, the same window gives 21.3, 15.2 and 63.6% unweighted and 34.6, 12.3, 51.7 and 1.4% weighted, so there the chord bucket carries between a half and two thirds; the one-electron term is 1.49% of L_{shared} at its last step. Every share in this section is of L_{shared}; the three further terms of the reported model's objective (14) are outside this denominator.

The one-electron term is therefore small but not absent. It contributed to the scalar loss at the 1.5–4.3% level across these rounds, computed here from (17) and the last-step share of each round, which runs from 4.31% in the first to 1.49% here. Only in the first, before the repair, was its response component discarded — *in full* rather than attenuated, as Appendix B measures.

D.5 Why we train on self-consistent densities: the two ways to train a Δ-correction

Writing the total energy of the corrected functional at a density matrix D as $E^{\theta}_{\mathrm{tot}}[D] = E_{\mathrm{base}}[D] + E^{\theta}_{\Delta}[D]$, there are two training regimes. *Non-self-consistent*: evaluate the loss on a fixed baseline density ρ_{PBE}, never solving an SCF during training. *Self-consistent*: re-solve the SCF with the current correction in place and evaluate the loss on the solution the model actually produces.

The first is cheaper and a legitimate way to fit energies. Its failure mode is a mismatch between training and deployment: at deployment nobody hands the functional ρ_{PBE}, so a correction that reproduces target energies *on* ρ_{PBE} can move the density elsewhere once it enters an SCF, and the energies then follow the moved density.

Measured, on our own model family. Train a GGA-level feature model without exact exchange non-self-consistently, then evaluate it both ways. The fractional-charge suppression factor on Li reads 45.3× evaluated non-self-consistently (a pass by our gate) and 0.9× when the same checkpoint is evaluated self-consistently (a failure) — a collapse by a factor of 50. That feature model carries no exact-exchange channels and the run is an overall failure, so this must not be summarised as "non-self-consistent training passed".

D.6 The step-0 anchor

The correction is initialised so θ_0 gives $f \equiv 0$ and hence $\Delta_{\theta_0} \equiv 0$: at step 0 the model *is* the baseline functional, every gate reading there is an independently known number, and the first training step is a regression test. The fixed reference density used by the one-electron target is the same cached ρ_{PBE} that defines that anchor, so the fixed-density target does not move it.

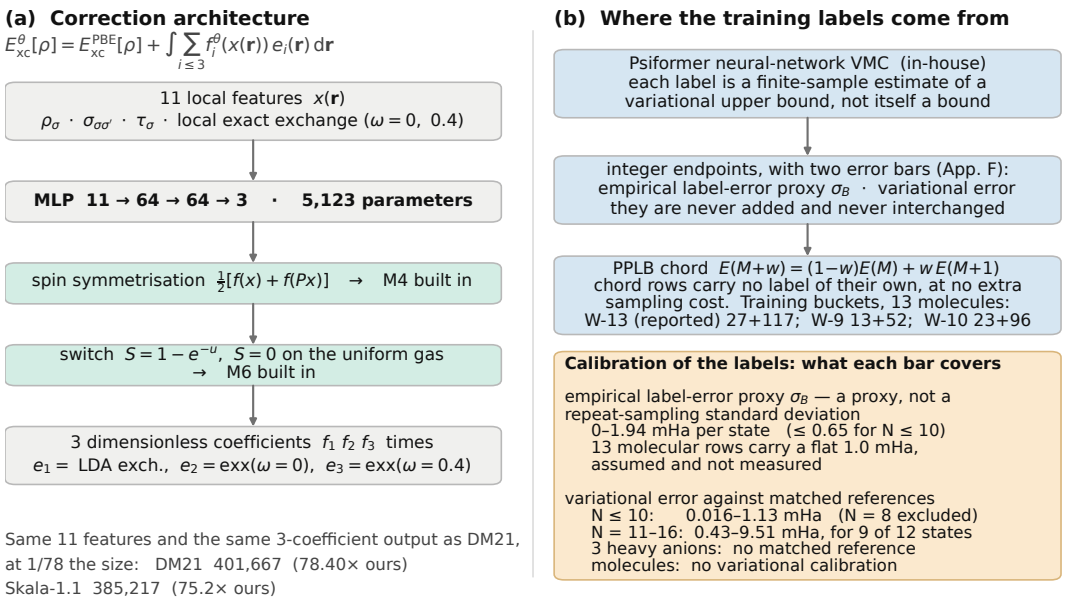

Figure E1: (a) 11 local features feed a 5,123-parameter MLP whose output is spin symmetrised and gated by a switch, so M4 holds by construction and M6 does for the unregularised switch (its deployed agreement is measured), and three dimensionless coefficients multiply three energy densities of known scaling — the shape of DM21 at $1/78$ the size. (b) The label data flow. In-house Psiformer VMC supplies the integer endpoints as *finite-sample* estimates; the PPLB-inspired chord regulariser expands them into chord rows at no extra sampling cost, and those rows carry no label of their own (Appendices F and D). *The bucket sizes are per round*: W-13, the round whose numbers we report, trained on 27 endpoints, 13 molecules and 117 chord rows; W-9 on $13 + 13 + 52$ and W-10 on $23+13+96$. Notation: ω is the range-separation parameter of the exact-exchange channels and w is the fractional electron number of the PPLB chord — two different quantities, which must never share a symbol.

E THE ELEVEN FEATURES, THE SWITCH, AND THE SYMMETRISATION

Eleven local quantities go in. At each grid point $\mathbf{r}$ the network reads

$$x = \big(\underbrace{\rho_\alpha,\ \rho_\beta}_{\text{densities}},\ \underbrace{\sigma_{\alpha\alpha},\ \sigma_{\alpha\beta},\ \sigma_{\beta\beta}}_{\text{gradient invariants}},\ \underbrace{\tau_\alpha,\ \tau_\beta}_{\text{kinetic-energy densities}},\ \underbrace{e_{\mathrm{HF}}^{\alpha,0},\ e_{\mathrm{HF}}^{\beta,0},\ e_{\mathrm{HF}}^{\alpha,\omega},\ e_{\mathrm{HF}}^{\beta,\omega}}_{\text{local exact exchange}}\big),$$

and Figure E1(a) is the map of what happens to them. In the compact notation of Section 6 these eleven are ρ_σ, $\sigma_{\sigma\sigma'}$, τ_σ and the four $e_{\mathrm{HF}}^{\sigma,\omega}$.

E.1 GUARANTEES: THE TWO CONDITIONS THAT HOLD FOR ANY θ

The switch: why M6 holds for any θ. The switch multiplies the symmetrised network output and turns it off exactly where the gas is uniform.

$$f^\theta(x) = S(x)\,\bar{f}_\theta(x), \qquad S = 1 - e^{-u}, \qquad u = a_s\,p + a_\alpha(\alpha - 1)^2, \tag{18}$$

with $a_s = a_\alpha = 4.0$, one pair of variables rather than a spin sum, and every denominator on the total density $\rho = \rho_\alpha + \rho_\beta$:

$$p = \frac{\hat{\sigma}}{4k_{\mathrm{F}}^2\rho^2}, \qquad \alpha = \frac{\sum_\sigma \tau_\sigma - \tau_W}{\tau_{\mathrm{unif}}}, \qquad \hat{\sigma} = 2(\sigma_{\alpha\alpha} + \sigma_{\beta\beta}) \ \ge\ |\nabla\rho|^2,$$

$$k_{\mathrm{F}} = (3\pi^2\rho)^{1/3}, \quad \tau_W = \frac{\hat{\sigma}}{8\rho}, \quad \tau_{\mathrm{unif}} = \sum_\sigma \frac{k_{\mathrm{F}\sigma}^5}{20\pi^2}, \quad k_{\mathrm{F}\sigma} = (6\pi^2\rho_\sigma)^{1/3}$$

(α the iso-orbital indicator, never a spin label), with ρ floored at 10^{-30}, an additive 10^{-60} in both denominators, and the squash offset 10^{-4} the switch inverts. These regularisers make the imple-

mented α agree with the unregularised expression to implementation precision inside the density window we read, not identically on every density.

Why $\hat{\sigma}$, and not $|\nabla\rho|^2$. The features reach the switch squashed as $\log(|x| + 10^{-4})$, which loses the sign of the one signed input $\sigma_{\alpha\beta}$, so $|\nabla\rho|^2 = \sigma_{\alpha\alpha} + 2\sigma_{\alpha\beta} + \sigma_{\beta\beta}$ is not recoverable and the switch reads same-spin invariants only. The bound above is Cauchy–Schwarz, then arithmetic–geometric means; it is an equality exactly where $\nabla\rho_\alpha = \nabla\rho_\beta$, the unpolarised case as well as $\sigma_{\alpha\alpha} = \sigma_{\beta\beta} = 0$, so it costs nothing where M6 is claimed, and being a polynomial it keeps $\partial\hat{\sigma}/\partial\sigma_{\sigma\sigma} = 2$ as a channel vanishes. So α here is not SCAN's: on a fully polarised one-electron density $\hat{\sigma} = 2|\nabla\rho|^2$, so $\alpha = -\frac{5}{6}2^{-2/3}p < 0$ by construction where SCAN's is zero, while $(\alpha - 1)^2 \geq 1$ there keeps $S \geq 1 - e^{-4}$ and leaves the uniform gas untouched.

What it buys. For the unregularised quantities above, a uniform gas has every σ vanishing, so $\hat{\sigma} = 0$ and $p = \tau_W = 0$, while $\tau_\sigma = \tau_{\text{unif}}(\rho_\sigma)$ makes the numerator of α its denominator term by term, so $\alpha = 1$ at any polarisation. Then $u = 0$, $S = 0$, $\Delta \equiv 0$ for every θ: the unregularised construction reduces to its baseline, which reduces to LDA, so there M6 holds as an identity and not as a fitted property. The deployed implementation carries the regularisers stated above; its agreement with this identity is a numerical result over the declared density and polarisation window. **The switch enforces it at every θ, not a network that learned it**, and the control prices that guarantee: the one model we trained without the switch reads $144\times$ the uniform-gas budget. Deployed, the correction reads 8×10^{-9} of the budget relative to the PBE baseline (Table 1).

Why both variables are needed. With p alone, every point where both spin gradients vanish — bond midpoints, density maxima — would switch the correction off, though the kinetic-energy density there is not the uniform-gas one; α keeps it alive, $S = 0$ needing $p = 0$ *and* $\alpha = 1$. We claim only the direction M6 asks for, uniform gas $\Rightarrow S = 0$, and not its converse: $\alpha = 1$ constrains the kinetic-energy density alone, and both are pointwise, so a point at which they take their uniform values need not lie in a uniform neighbourhood.

Why $1 - e^{-u}$ rather than $u/(1 + u)$. It saturates quickly ($u = 4.6$ gives $S = 0.99$), so on real molecular grids $S \approx 1$ and the switch does not interfere with what the network has learned. The scalar map is also smooth, with $\partial S/\partial u = e^{-u} \in (0, 1]$ bounded and no kink. That bounds $u \mapsto S$ and nothing else: a functional derivative also carries $\partial u/\partial\rho$, whose negative powers of the density are held in the vacuum by the two floors above, not by the choice of map. The no-kink requirement is a lesson paid for: an earlier $\log(|\zeta| + 10^{-4})$ feature had a kink at $\zeta = 0$ that amplified $\partial/\partial\rho$ by 10^4 into hartree-scale jumps in the Kohn–Sham potential.

The same transform still carries a cusp. $\sigma_{\alpha\beta}$ is the one signed input and $\log(|x| + 10^{-4})$ is even in it. On the reported arm the coefficients stay continuous across $\sigma_{\alpha\beta} = 0$ (largest jump 4.6×10^{-8}) while $\partial f/\partial\sigma_{\alpha\beta}$ reverses sign there, autograd and finite differences agreeing to 10^{-4}: a cusp, not a jump. Fixed-density, fixed-grid energies do not cross it — on a uniform gas $\sigma_{\alpha\beta} \equiv 0$, so M6 is read *at* the cusp — while the potential that M1, M3 and M7 reach through an SCF has a corner on that surface. We attribute no reported failure to it, and the transform is the one the released DM21 code uses.

The symmetrisation: why M4 holds for any θ. $\bar{f}_\theta = \frac{1}{2}[g_\theta(x) + g_\theta(Px)]$, where g_θ is the raw network output and P is the involution on the eleven features that swaps the two spin channels, $P = (1, 0, 4, 3, 2, 6, 5, 8, 7, 10, 9)$ in zero-based index order, mapping $\sigma_{\alpha\alpha} \leftrightarrow \sigma_{\beta\beta}$ while fixing $\sigma_{\alpha\beta}$. Since $P^2 = \text{id}$, the symmetrised map satisfies $\bar{f}_\theta(Px) = \bar{f}_\theta(x)$ identically, so M4 holds by construction for any θ, at the cost of one extra forward evaluation. The same construction is described in DM21's supplementary material, and we do not claim it; what we report in Section 3 is a measurement of released weights.

E.2 Implementation: what each input is, and where the parameters sit

ρ_σ. Spin-resolved densities, with $\rho = \rho_\alpha + \rho_\beta$. The polarisation ζ is *not* an independent input, and neither ζ nor $\rho_\alpha\rho_\beta$ adds information the network cannot already form. *Lemma.* If $|\Psi\rangle$ is an

eigenstate of $\hat{S}^2$ with $M_S = 0$, its spin densities are equal at every point, $\rho_\alpha \equiv \rho_\beta = \frac{1}{2}\rho$. *Proof.* $\hat{F} = \exp(-i\pi \hat{S}_y)$ is unitary, commutes with $\hat{S}^2$ and sends $M_S \to -M_S$, so it fixes such a state up to a phase while exchanging the two spin-density operators; the two densities therefore coincide. *What the lemma does not settle*: it speaks only at a fixed total density, so a more accurate *total* density is not ruled out by it.

$\sigma_{\sigma\sigma'}$. Pairwise dot products of the spin-density gradients, $\sigma_{\alpha\beta} = \nabla\rho_\alpha \cdot \nabla\rho_\beta$ included. The cross term is a dot product of two different channels, not a squared magnitude, so *it can be negative*; by Cauchy–Schwarz $\sigma_{\alpha\beta}^2 \le \sigma_{\alpha\alpha}\sigma_{\beta\beta}$, hence $\sigma_{\alpha\alpha} = \sigma_{\beta\beta} = 0 \Rightarrow \sigma_{\alpha\beta} = 0$, which is what makes the uniform-gas switch above exact rather than approximate.

τ_σ. Kohn–Sham kinetic-energy density in the $\frac{1}{2}\sum_i^{\mathrm{occ}} |\nabla\psi_{i\sigma}|^2$ convention. The factor $\frac{1}{2}$ is not a notational preference. It must match the convention used in τ_{unif} and τ_W inside the switch, so we pin it with an identity, not by inspection: $\int \tau_\sigma \,\mathrm{d}\mathbf{r} = \mathrm{Tr}(D_\sigma T_{\mathrm{kin}})$, measured ratio $1 + 2 \times 10^{-13}$.

$e_{\mathrm{HF}}^{\sigma,\omega}$ — local exact exchange. With $\nu_{bd}(\mathbf{r}) = \int \chi_b(\mathbf{r}')v_\omega(\mathbf{r},\mathbf{r}')\chi_d(\mathbf{r}')\,\mathrm{d}\mathbf{r}'$ and $E_b^\sigma(\mathbf{r}) = \sum_a D_{ab}^\sigma \chi_a(\mathbf{r})$,

$$e_{\mathrm{HF}}^{\sigma,\omega}(\mathbf{r}) = -\tfrac{1}{2}\sum_{bd} E_b^\sigma(\mathbf{r})\, E_d^\sigma(\mathbf{r})\, \nu_{bd}(\mathbf{r}), \qquad \int e_{\mathrm{HF}}^{\sigma,\omega}(\mathbf{r})\,\mathrm{d}\mathbf{r} = E_{\mathrm{x}}^{\mathrm{HF},\sigma}(\omega). \tag{19}$$

The defining property is the integral: this is Hartree–Fock exchange *spread over grid points*, so genuinely non-local information is presented in a form a pointwise network can read.

A counter-intuitive convention in the two ω channels. The kernel in (19) is the full Coulomb kernel at $\omega = 0$ and the *long-range* erf kernel at $\omega > 0$ — not the short-range erfc one. Consequently the $\omega > 0$ channel tends to zero as $\omega \to 0^+$, and the $\omega = 0$ channel is a separate case, not that limit: the two are discontinuous by construction. The correct description of the pair is "full exchange" and "long-range exchange at $\omega = 0.4$", never "two screening lengths".

A normalisation warning. $\varepsilon_{\mathrm{xc}}$ is a per-particle quantity, $E_{\mathrm{xc}} = \int \varepsilon_{\mathrm{xc}}\rho$, while e_{HF} and the e_i of the main text are per-volume, $E = \int e$. The two differ by a factor of ρ, and any expression moved between them must be checked.

The parameter count. The network is $11 \to 64 \to 64 \to 3$ with biases: $11{\cdot}64 + 64 = 768$, $64{\cdot}64 + 64 = 4160$, $64{\cdot}3 + 3 = 195$, total $\mathbf{5{,}123}$ trainable parameters. Both hidden layers are $\tanh$, and the head's three outputs pass a symmetric soft clamp $\mathrm{sc}(x) = \mathrm{sp}(x{+}M) - \mathrm{sp}(x{-}M) - [\mathrm{sp}(M) - \mathrm{sp}(-M)]$, sp the softplus and $M = 4$, before the switch multiplies them. There is no sigmoid anywhere. Both are C^∞ with $\mathrm{sc}(0) = 0$ exactly, which is what V_θ^Δ needs, being an automatic derivative through them. That is $1/78.4$ of DM21's count and $1/75.2$ of Skala-1.1's, counted over the enhancement-factor network weights (Appendix G).

F THE QUANTUM-MONTE-CARLO LABELS

Top line: the two error bars, and what each one covers.

	What it is	$N \le 10$	$N = 11$–16	molecular rows
label-error proxy σ_B	empirical, $\sqrt{\sigma_A^2 + \sigma_{\mathrm{sys}}^2}$; the column the training pack ships. *Not* a repeat-sampling standard deviation	≤ 0.65 mHa; the analytic hydrogen endpoint exactly 0	0.78–1.94 mHa	flat 1.0 mHa assumed, not measured
variational	one-sided and positive, not reducible by sampling; calibrated rather than estimated	0.016–1.13 mHa measured ($N = 8$ excluded)	0.43–9.51 mHa measured for nine of twelve; the three anions have no reference of this protocol	no calibration

They are never added and never interchanged.

F.1 What variational Monte Carlo delivers, and what its two error bars mean

For a trial state Ψ_θ the energy expectation is an average over configurations drawn from $|\Psi_\theta|^2$ of the *local energy* $E_L(\mathcal{R}) = \hat{H}\Psi_\theta(\mathcal{R})/\Psi_\theta(\mathcal{R})$,

$$E[\Psi_\theta] = \frac{\langle\Psi_\theta|\hat{H}|\Psi_\theta\rangle}{\langle\Psi_\theta|\Psi_\theta\rangle} = \int |\Psi_\theta|^2 E_L \,\mathrm{d}\mathcal{R} \;\approx\; \frac{1}{M}\sum_{m=1}^{M} E_L(\mathcal{R}_m), \tag{20}$$

sampled by Metropolis; an exact eigenstate would make E_L constant, and that zero-variance property is where the method's accuracy comes from. The left-hand side of (20) is an *upper bound* on the ground-state energy for any *fixed* Ψ_θ; the right-hand side is a finite-sample estimate of that bound and not itself a bound, so a label can fall below the true energy by chance. We run *pure* VMC — no fixed-node projection at any stage — so the residual against an exact reference is a *variational error* (unrecovered correlation energy), never a fixed-node bias.

The estimator, and the two error bars. Each state's energy is the mean of the last fifth of its optimisation trajectory (the "tail window"); the standard deviation over that window uses $\mathrm{ddof} = 1$ and $\mathrm{sem} = \mathrm{sd}/\sqrt{n_{\mathrm{tail}}}$. Except for P, whose re-run introduced a different label protocol and is treated separately below, it is a plain mean, not a clipped or median-of-means estimator. Two error bars then travel with every row of the label pack. The first is **sampling**, $\sigma_A \equiv \mathrm{sem}$ over the tail window — a stretch of an optimisation trajectory, not a sample from one fixed $|\Psi_\theta|^2$, with its stationarity, autocorrelation and effective sample size uncalibrated. Then

$$\sigma_B = \sqrt{\sigma_A^2 + \sigma_{\mathrm{sys}}^2}, \qquad \sigma_{\mathrm{sys}}(N) = 2.8\,N^{2.34}\;\mu\mathrm{Ha}, \tag{21}$$

where σ_{sys} is *half* of the absolute-error power law $5.6\,N^{2.34}\,\mu$Ha fitted across the 21 states of the light-atom label run ($R^2 = 0.906$), a floor, so no state is quoted more precisely than the family trend supports. σ_B **is therefore an empirical label-error proxy, not a statistical error bar**: σ_{sys} is a fitted trend in N, never a repeat-sampling spread, so neither σ_A nor σ_B carries a confidence statement. On a fractional-occupation row the sampling part propagates in quadrature, $\sigma_A(w) = \sqrt{((1-w)\,\mathrm{sem}_A)^2 + (w\,\mathrm{sem}_B)^2}$, and σ_{sys} is taken from the larger endpoint. The analytic hydrogen endpoint ($E = -\frac{1}{2}$ Ha exactly) carries $\sigma_A = \sigma_{\mathrm{sys}} = \sigma_B = 0$. The second is **variational** (§F.3). σ_B is the quantity Table G7 reads endpoint residuals against; its largest value in the training bucket is 1.94 mHa (S, $N = 16$). *And keeping the lower of two runs* (P, below) is a *selection* among finite-sample estimates, not the variational principle applied to exact expectation values; its bias is not excluded.

The ansatz. It is the Psiformer (von Glehn et al., 2023), a neural-network ansatz of the kind introduced by Hermann et al. (2020) and Pfau et al. (2020), in which each orbital depends on all electron coordinates. We run it with sixteen dense generalised determinants, the default of the reference implementation, and do not claim that this suffices for every state in the pack.

F.2 Sampling error, per state

Twelve heavy states were added in the round that produced the larger bucket. Their measured σ_A and σ_B, in mHa, read from the shipped label pack:

state	N	σ_A	σ_B	state	N	σ_A	σ_B
Mg^+	11	0.159	0.782	Si^-	15	0.329	1.616
Mg	12	0.243	0.970	P^+	14	0.482	1.430
Al^+	12	0.233	0.967	$P^\dagger$	15	0.59	1.69
Al	13	0.283	1.167	P^-	16	0.555	1.922
Al^-	14	0.378	1.398	S	16	**0.619**	**1.941**
Si^+	13	0.324	1.177	Si	14	0.410	1.407

† **P is the one state that has been re-measured** (Appendix F.3, note †). *The stationarity of a tail window is not calibrated anywhere in this pack.* One further state, Mg^{2+}, diverged in mid-optimisation (26,556 of 40,001 logged rows non-finite) and is not in the bucket at all.

The molecular labels carry no calibration. The thirteen molecular rows enter the pack with a flat $\sigma_B = 1.0\,\mathrm{mHa}$ on every row: CH_4, H_2O, Li_2, NH_3, LiH at six separations and N_2 at three. That is an assumed value, not a measurement: there is no per-molecule tail-window statistic in the pack and no non-relativistic reference of the grade used below, so neither error bar is calibrated for them. Molecular nodal surfaces are harder than atomic ones and we have not tested whether the atomic calibration transfers. Over the last $1{,}000$ steps they carry 10.3% of W-10's weighted L_{shared} and 12.3% of the reported round's (Appendix D).

Where the sampling error does *not* enter. Not in the chord bucket. A chord residual is $D_w = E^\theta(w) - (1-w)E^\theta(A) - wE^\theta(B)$, three energies of the model itself and no label (Appendix D; the code path carries no target column for these rows). So that bucket's 64% of the same W-10 total is invariant to any offset or noise in the endpoint labels. What it does enter is the endpoint bucket, 23.7% of it, and through it the residual decomposition of Table G7.

F.3 Variational error: the calibration, and what it does not cover

What the residual is measured against. The reference values are non-relativistic, clamped-nucleus, all-electron ground-state total energies, and they come from a named chain, not from experiment: for the neutral atoms Li–Ne, the "Exact" column of Pfau et al. (2020), Table I, which compiles the correlation energies of Chakravorty et al. (1993) with Hartree–Fock limits; for He, the independent high-precision value of Aznabaev et al. (2018), which is *not* from that compilation; and for ions, the neutral value plus the corresponding ionisation potential or electron affinity with estimated relativistic effects removed (Klopper et al., 2010).

Two provenance facts travel with it. (i) For ions there are two equally legitimate routes: the derived one just described, and the same compilation's directly tabulated ionic values. They differ by up to $0.18\,\mathrm{mHa}$, so **every ionic reference carries a $\approx 0.2\,\mathrm{mHa}$ systematic** that the neutral chain does not. (ii) The heavy-bucket calibration below is read from the full text of Chakravorty et al. (1993); the light-atom chain above was not re-read from that text.

The calibration. Using one Psiformer architecture and one protocol throughout, the variational error against those references is

N	2 (He)	3	4	5	6	7	9 (F)	10 (Ne)
variational error [mHa]	0.016	0.074	0.154	0.315	0.402	0.502	1.11	1.13
in units of that state's σ_A	1.8	3.0	4.0	7.1	6.3	6.8	11.6	9.1

All eight are positive, as the variational principle requires of the quantity being estimated — but positivity of a *finite-sample* estimate is informative only in proportion to its own sampling error, the second row. At $N = 2$ the deviation is $1.8\,\sigma_A$ and the sign carries little information; from $N = 3$ upward it is 3 to $12\,\sigma_A$. *σ_A is an uncalibrated empirical scale, not a calibrated standard error*, so these multiples say how far each state sits above its own noise and are not z-scores.

The excluded point, and what we did not do about it. $N = 8$ (O) is absent because its tail window contains $n_{\mathrm{tail}} = 164$ steps, a truncated trajectory, against $\sim 10^3$–10^4 for the others. The criterion is therefore a property of the *trajectory*, not of the residual — the form of exclusion least able to bias the fit. But it was applied after the runs, not fixed in advance, and **we have not refitted the power law with $N = 8$ restored**, so we cannot report how much the exponent would move.

The measurement that replaced the extrapolation. For nine of the twelve heavy states the earlier power law is **superseded by a measurement** against the same-protocol compilation of Chakravorty et al. (1993). Their Table XI, "estimated nonrelativistic stationary-point-nucleus atomic energies", is non-relativistic, all-electron and clamped point nucleus, like our labels, so no conversion is needed:

state	N	σ_A [mHa]	σ_B [mHa]	Δ_{A} [mHa]	Δ_{B} [mHa]	correlation energy recovered [%]
Mg^+	11	0.16	0.78	$+0.42$	$+0.43$	99.89
Mg	12	0.24	0.97	$+2.44$	$+2.36$	99.46
Al^+	12	0.23	0.97	$+1.42$	$+1.72$	99.62
Al	13	0.28	1.17	$+3.67$	$+3.98$	99.15
Si^+	13	0.32	1.18	$+4.72$	$+4.97$	98.98
Si	14	0.41	1.41	$+5.72$	$+6.12$	98.79
P^+	14	0.48	1.43	$+7.03$	$+6.77$	98.70
$P^\dagger$	15	0.59	1.69	$+7.30$	$+7.34$	98.64
S	16	0.62	1.94	$+9.86$	$+9.51$	98.43
Al^-	14	0.38	1.40	*no reference*		—
Si^-	15	0.33	1.62	*no reference*		—
P^-	16	0.56	1.92	*no reference*		—

† **P is the one row that has been re-measured**, at the same configuration; we keep the lower of the two readings — a selection (§F.1), not the variational principle applied to two exact expectation values. It is also the one row at the label protocol that re-run introduced (a robust truncation of the tail window, not a plain mean). **The other eight rows are at the old protocol**; recomputing them at the new one moves them by at most $0.214\,\mathrm{mHa}$ (Mg), below every σ_B in the column.

The two extraction routes, and five things that travel with the table. Δ_{A} is Table XI read directly; Δ_{B} reconstructs the same reference as $E = E_{\mathrm{HF}} + E_{\mathrm{c}}$ from Tables III and X of the same compilation. The two differ by up to $0.40\,\mathrm{mHa}$ (worst case Si: $+5.72$ against $+6.12$), entirely because Table XI prints only three decimals for $N \geq 12$ and four below. That is why **every statement in this paper at the $0.5\,\mathrm{mHa}$ level, and the %CE column above, use route B**. They are printed together because neither is a derived form of the other.

(i) *It is positive in all nine on both routes*, as the variational principle requires, and the same extraction reproduces the light-atom row above to the last printed digit on the seven states the two chains share.

(ii) *The earlier N-only extrapolation does not describe these states*: it was wrong at both ends and cations do not follow the neutral curve, so the row above is a direct calibration, limited by its coverage. **We have refitted nothing on these points.**

(iii) *The reference is itself semi-empirical*. Its authors bound their per-increment uncertainty by "less than ± 1 mhartree" only for $N = 3$–10, writing that for 12–18 electrons "the present predictions of total energies require further careful explorations". Eight of our nine states sit in that second regime and carry an unquantified reference uncertainty; the ninth, Mg^+ at $N = 11$, is in neither regime the compilation bounds, so nothing published bounds its reference error either. For two of the smaller heavy residuals (Al^+, Mg) the "measured" variational error is therefore not resolved against the reference's own uncertainty.

(iv) *The anomalies did not close into a trend*: the re-run removed P's break of monotonicity in N, but three further states of the same calibration run, not in this table (S^+, Cl^+, Cl), break it again in the same sense, a lower-N state reading the larger error (S^+ above S, Cl^+ above Cl), so **neither N nor the valence-electron count orders that run's errors**.

(v) *The three anions remain extrapolated*: that compilation covers $N \leq Z$ and tabulates no electron affinities, so Al^-, Si^- and P^- have no reference of this protocol and no measured variational error.

The two kinds of target pull against each other. The absolute energies above carry a one-signed variational bias that cancels only partly between adjacent charge states (below), while the one-electron, fractional-spin and M7 terms are exact relations carrying none. Fitting both at once can put the two in conflict; we do not decompose that trade-off here. The held-out endpoint gate G1 and the ionisation column of Table G5 are where it would show.

Which error dominates, and on which quantity. On the heavy bucket neither bar may be quoted alone, and above $N = 13$ the variational one is the larger.

On *differences* between adjacent charge states of the same element — what a chord segment spans — the variational error partly cancels, but by less than we assumed. On the four cation chords the difference of the two endpoint errors is now measured at 0.57 to $2.26\,\mathrm{mHa}$ (Mg 1.93, Al 2.26, Si

1.15, P 0.57; P read 3.72 before its re-run), against the 0.60–0.75 mHa previously assumed. Three of the four exceed that band by 1.5 to 3.0× while P now falls just below it. Two of the four (Mg, Al) also exceed by 1.4× the $\sqrt{2}\,\sigma_B$ heuristic scale, while two do not. Those two assumed values belong to the Al and P *anion* chords, precisely the ones that still have no reference. So nothing measured here bounds them, and σ_B alone is not the error of such a difference.

What is not compared with what. The power law is deliberately not refitted on the nine new points, which are its test. Fixed-node diffusion Monte Carlo upper bounds belong to a different method class and are never placed in the same column as the variational errors above.

F.4 The chord construction, and why those rows are free

For the exact functional the ground-state energy is piecewise linear in electron number between adjacent integers (Perdew et al., 1982), for the ensemble reason Appendix A gives:

$$E(M+w) = (1-w)E(M) + wE(M+1), \qquad 0 \le w \le 1. \tag{22}$$

This is an identity, not a fit, so once the two integer endpoints are computed the whole segment is known: nine fractional rows per pair at $w \in \{0.1, \ldots, 0.9\}$ cost no sampling at all. That is the pack. A round's chord bucket (Appendix D) is smaller: twelve rows, all at $w \ge 0.5$, have no cached density, the validation element's two pairs are held out, and a fixed-seed split withholds 15% of the atomic rows, so 2 to 9 rows per pair remain — 117 over 17 pairs in the reported round.

(22) *motivates* the chord regulariser, so these rows contain no new physical information. They carry no independent QMC label, and that regulariser — defined once in Appendix D and evaluated on cached approximate densities — does not enforce the model's self-consistent $E(N)$ linearity. The H_2^+ tail coefficient η is targeted separately in the reported model. A segment whose upper endpoint is an anion unbound in the complete-basis limit would enforce (22) where the exact $E(N)$ is flat instead; our pairs are the ones listed in the pack, and we have not audited them against that condition.

G Complete tables, verdict summaries, provenance and search scope

G.1 The seven-column summary, with its reading rules

Table G1 is the full form of Table 1, its $\delta(\frac{1}{2})$ row aside and its M4 row read on the density layer rather than on the deployed energy (Appendix A), with the per-cell reading rule and the protocol note that the compact table cannot carry.

Two rows of this table carry two different readings associated with long-range exact exchange, and they need not agree. M1 reads it directly as η. The M3 row supports a second, two-point reading, calibrated to 0 at PBE and 1 at LC-ωPBE. At PBE0 that second scale reads 0.31 against a declared 0.25, a residual of 0.06 at that one column (Appendix G.3.3 builds it).

How each row reads. M1: 96.5 against 81.9 for the `DM21` weight at the same basis and grid; the 14.7 pp gap is comparable to that weight's own fit-window systematic[(1)], and η is a term in our loss and not in that weight's, so it reads what training buys and *is not claimed as an advantage*. η also reads back a long-range fraction a column was *built* with: PBE0's 25.0 is its mixing fraction, LC-ωPBE's 100.0 its construction[(1)]. **M2**: tied best with the released `DM21mc` weight, *not ours alone* — and matched by LC-ωPBE. **M3**: smallest of the four audited columns, but behind the LC-ωPBE control — its 0.141 is 0.087 below us, past the largest shift measured here (0.082), so **that ordering is a reading at the declared protocol, and a loss for us**, and the pair's differential basis convergence is unverified; against `DM21mc` (0.2589, a gap of 0.030) it is not resolved by this benchmark. At r=5, where the benchmark's gate is, **of the columns read there, only LC-ωPBE is inside it**[(9)]. **M4**: tied clean with PBE, Skala-1.1 and all three controls; all four released DM21 weights break it in the deployed energy (Table A4); their density-layer readings here are 1.5e-02 to 7.3e-02[(2)]. **M5**: no column exceeds the bound here, which decides little. As a signed distance from the exact value we are **third of the four audited columns** (+0.33 pp; Skala −0.00, DM21 +0.17, PBE −0.87) and **fifth of seven** once the controls are read on the same row (PBE0 −0.40, LC-ωPBE +0.06, SCAN +0.13), and He^+ is a training system for us[(4)]. **LC-ωPBE does exceed the bound on H** (101.3%). **M6**: our cell reads the correction alone, against PBE's uniform-gas value — zero *by construction* for the unregularised switch (not learned), 8×10^{-9} of the budget as deployed; DM21 was not trained on the uniform gas[(3)], and LC-ωPBE's 1.12× is entirely its short-range exchange model (Appendix G.3.1). **M7**: **no converged reading here is below** −1, so this row proves nothing against any column, ours included. Earlier models of ours read below −1 and did. For the other two only variational bounds exist ($\le +2.76$, $\le +3.63$)[(5)].

Notes to Table G1. Thermochemistry (Table G5): of the three neural functionals we are the worst on all three standard subsets; on the same point sets PBE0 is the best of the three controls on the two atomic subsets and LC-ωPBE on the self-interaction subset, where it also beats us.

Glyphs: ✓/✗ declared threshold met/not met, ◯ inside tolerance, not proved; "n.m." is a state, not a glyph: not measured — not the same as not applicable, and no cell is filled with an estimate; a "×" inside a number is a multiplier, never a verdict; LC-ωP heads the LC-ωPBE column. The M4 and M6 entries for our column hold *by construction*, not by training.

Table G1: Exact-condition diagnostics for seven functionals measured on one benchmark: *within each row*, the same basis, grid, point set and reference density for every column (the protocol differs *between* rows; see Table A2). **The glyph legend, the three control columns and the protocol declarations are in the notes below the table.**

		audited				*hand-built (by construction)*		
	Reported quantity	**PBE**	**DM21**	**Skala**	**Ours**	**PBE0**	**LC-ωP**	**SCAN**
M1	η [%]; exact 100, this window 100.03	0.005	81.9	0.005	**96.5**	25.0	100.0	n.m.[6]
M2	Count of $\varepsilon_{\mathrm{HOMO}} < 0$ over three anions	0/3	1/3	0/3	**3/3**	1/3	**3/3**	0/3
M3	$\lvert\, r\, w_{\mathrm{xc}}(3) + 1 \,\rvert$; exact 0 only as $r \to \infty$	0.553	0.369	0.462	**0.229**	0.426	**0.141**	0.517
	same row at r=5: $\mathrm{dev}(5)$, gate ≤ 0.05	0.478	0.367	0.544	**0.099**	0.358	**0.040**	0.469
M4	$\max \lvert D_\alpha - D_\beta \rvert$, exact 0	1.8e-15	4.8e-02	1.7e-16	3.2e-15	2.0e-15	1.6e-11	9.4e-15
M5	$\lambda(\mathrm{He}^+)/\lambda_1$ [%]; bound ≤ 100, **exact** 99.23	98.4	99.4	99.2	99.6	98.8	99.3	99.4
M6	$\max \lvert \Delta\varepsilon_{\mathrm{xc}} \rvert$, in units of the budget	1e-06×[7]	124×	57×[10]	8e-09×	3e-10×	1.12×	1e-06×[7]
M7	d [mHa], window $\lvert d \rvert \leq 1$[8]	◯ +0.55	n.m.	n.m.	◯ −0.0007	◯ +0.60	◯ +0.61	◯ −0.45

The three right-hand columns are hand-built controls, with no parameter fitted in this work — PBE0 (Adamo & Barone, 1999), LC-ωPBE (Vydrov & Scuseria, 2006) and SCAN (Sun et al., 2015) — run after the audit under a reading rule fixed before the runs. *Hand-built is not parameter-free*: LC-ωPBE carries a range-separation constant $\omega = 0.4\, a_0^{-1}$ fixed by its own authors, and SCAN's coefficients are fixed against appropriate norms; only PBE0's $\alpha = \frac{1}{4}$ is derived. A cell they hold by construction is not a defeat of the audit but its calibration: read them against the four audited columns row by row.

(1) η is the $1/R$ coefficient *only*; the deviation from linearity at half occupation, $\delta(\frac{1}{2})$, is in Tables 1 and G2, and it orders the columns differently. η is quoted to 0.1 pp; DM21's is 81.9 ± 0.4, the ordinary-least-squares standard error of its own three-point fit, against ± 0.0012 pp for ours on the same three points, a ratio following from the `conv_tol` difference below. `DM21mc` reaches $\eta = 110.6\%$, i.e. $\lvert 100 - \eta \rvert = 10.6$ against our 3.5, and it overshoots rather than undershoots, so the sign of the pathology differs ($c < 0$ is a spurious attractive tail). *The two are not comparable as evidence*: η is a term in our loss and not in that weight's. Widening the fit range moves η by $+2.79$ pp for us, $+1.13$ for PBE, $+1.02$ for Skala-1.1 and $+11.76$ for `DM21`, the largest shift on this row. Our gap to that weight therefore moves with the window too, from 14.7 pp at $R \geq 20$ Å down to 5.7 pp at $R \geq 5$ Å. That $+11.76$ is refitted from a later run of the same probe at the same basis and grid, because the baseline archive for that weight stores only the three fit points. The two runs agree on those three points to 1.1×10^{-15} Ha. *And that reading is calibrated, not definitional*: note f derives $\eta = a$ for a functional *built* to carry a long-range fraction a; a learned column need carry none, so its η is a reading on that scale and not a mixing coefficient.

(2) DM21's supplementary material (Kirkpatrick et al., 2021) describes both a strict symmetrisation (evaluate twice, average) and an approximate one (random spin flips during training) and reports similar benchmark scores for the two. What we measure is the released weights, not the architecture. The three control columns were produced on a replica of this benchmark — same systems, basis, grid, occupations and definitions, with the functional as the only substitution — whose PBE model reproduces the archived M2 and M3 cells of this table (Appendix G.3.2).

(3) For Skala-1.1 the released tool returns n.m., because the quantity is not well defined for an atom-partitioned non-local term without a convention for where the nuclei sit. Across the eight conventions we tried the main-window deviation was always $\geq 44\times$ budget, but the spread across conventions (256 mHa/electron) exceeds the effect, so no single one of those eight should be quoted on its own; the $57\times$ printed is the declared extrapolation limit of Appendix H (note (10)), read by a protocol other than the tool's. For DM21, its own ablation table records that it was trained on fractional charge and spin but not on the uniform gas; a separately released weight that was trained on the uniform gas holds it to $0.92\times$ on this same benchmark.

(4) The exact value 99.23% follows from $E_{\mathrm{xc}} = -E_{\mathrm{H}}$ at $N = 1$ on the same shared reference density (Appendix A); an independent closed form on an idealised hydrogenic density gives 99.36%.

(5) Every cell in the *Ours* column is the same checkpoint, M7 included: the model reported here reaches an aufbau stationary point of its own with no negative spin-flip curvature ($\lambda_{\min} = +0.36$, both spin channels aufbau). "Not measured" for DM21 and Skala-1.1 reflects search effort, not a property of those functionals. What the criterion reads is the fragment-interaction error at a finite $R = 6\, a_0$ against an exact Born–Oppenheimer reference (Sims & Hagstrom, 2006), which mixes additivity failure with missing dispersion; see Appendix A.

(6) SCAN's M1 cell is empty for a tooling reason, not a property of that functional: at large R its H_2^+ SCF produces a singular DIIS extrapolation matrix, and the `except` clause that `pyscf` 2.13.1's first-order `lib/diis.py` uses to re-raise it names `numpy.linalg.linalg.LinAlgError`, an attribute `numpy` 2.4 no longer has, so the run aborts inside the library's own error handler. The second-order route completes but costs 614 s for the single point $R = 2.0$ Å, against 28–52 s for an entire 11-point M1 scan on this benchmark, so eleven such points were outside the compute budget of this work. A guard-mode variant did return a number ($\eta = -2.3 \times 10^{-5}\%$) and was **rejected**, because on the same route PBE's positive control reads $7.2 \times 10^{-4}\%$ against its archived $5.1 \times 10^{-3}\%$; the cell stays empty rather than estimated. *The released bench has since repaired that path, and the cell still reads "not measured".* With the third rung of its SCF ladder alive, one of two identical runs converged all eleven R points and returned $\eta = 0.00515\%$ — structurally the right answer, since H_2^+ is one-electron and the reading coincides with `r2SCAN`'s to eight significant figures in the fitted coefficient. The second run did not converge, and a reading that does not reproduce on demand is not a shipped reading, so the bench prints "not measured" here too and keeps both runs (Appendix H). Reachable is not the same claim as measured. `R2SCAN`, which regularises the same form, covers all eleven points here on one run, and returns $0/3$ on Koopmans and $\mathrm{dev}(5) = 0.4694$, within 10^{-5} of SCAN's own value. The fragility is removable. These two readings are not.

(7) PBE's cell is its uniform-gas residual against `libxc`'s `PW_MOD` parameterisation of PW92 (gate H2, Appendix G.3.1), 2.4082×10^{-7} mHa/electron, the benchmark's floor and 1.2×10^{-6} of the budget. The controls and our column are read against PBE's own uniform-gas value, so SCAN, which reduces to the uniform gas *by construction*, returns that same residual (2.4082×10^{-7}, to seven digits), while PBE0 and ours sit below the floor because they are differences from PBE, not from PW92. Neither PBE nor SCAN is a miss.

(8) This row has two readings, and the released bench prints both columns rather than collapsing them. The verdict used here is one-sided: a converged d is itself an upper bound, so a value below -1 mHa *proves* failure while one inside the window proves nothing —

which is why every cell on this row that was measured at all is ◯. The pre-registered conjunction is stricter, $|d| \leq 1\,\mathrm{mHa}$ *and* $\lambda_{\min} > 0$ *and* FS-0 ≥ 0, and it moves two cells of this table: PBE passes all three ($+0.55$, $+0.45$, and the benchmark's PBE control reads FS-0 $= +0.57\,\mathrm{mHa}$) and so does the reported model (-0.0007, $+0.3606$, FS-0 $= +0.4377\,\mathrm{mHa}$); both would read ✓ under it, whereas PBE0, LC-ωPBE and SCAN have no archived FS-0, so for them the conjunction is not evaluable and ◯ stands under either reading. **The conjunction is a gate outcome, not a proof.** $\lambda_{\min} > 0$ rules out negative curvature in the spin-flip subspace only — a *negative* $\lambda_{\min}$ would prove instability, while a positive one proves nothing about a minimum of the full orbital space — and FS-0 ≥ 0 is a one-sided reading on five occupations; neither supplies the lower bound that $|d| \leq 1$ would need. A column that passes all three has had no *proof* of failure produced against it, which is why this table marks it ◯ and not ✓. The rule still bites: a *converged* reading below -1 would be a proved failure, and on this row no column is one — earlier models of ours were, which is how the rule earned its place here.

(9) $r = 3$ **is this paper's read-out and** $r = 5$ **is the released benchmark's gate**; the two are the same criterion at two radii, a threshold at one does not transfer to the other, and *both are finite-radius residuals*, not the $r \to \infty$ limit the condition states. Read across the two radii, our gap to `DM21` widens from 0.140 to 0.268, and **Skala-1.1's residual *grows* with** r ($0.462 \to 0.544$) while every other column's shrinks. *Read at H^- only.* On O^- and F^- the aug-cc-pVDZ tail is exhausted before $r = 5$: on *all seven* models the profile crosses zero between $r \approx 6.3$ and $7.0\,a_0$ and reaches $+10$ to $+18$ by $r = 12$, and LC-ωPBE's F^- cell moves from 0.193 at $r = 3$ to 0.471 at $r = 5$ for that reason. Extending $r = 5$ to those two systems would need a basis check we have not run.

(10) Read by a protocol other than the released tool's, which returns *n.m.* here. The limit is $11.34\,\mathrm{mHa/e}$ (Appendix H).

Protocol difference to declare: the DM21 family runs at its official `conv_tol`$= 10^{-6}$, our models and the archived PBE and Skala-1.1 models at 10^{-10}; the three controls at 10^{-10} on M1–M6 and 10^{-11} on M7.

G.2 All released weights against all seven conditions

Table G2: Seven conditions against the four released DM21 weights, PBE, Skala-1.1 and our checkpoint, all on one benchmark. Training constraints per DM21 weight are from its own ablation table: `DM21` fractional charge and spin, `DM21m` none, `DM21mc` fractional charge, `DM21mu` uniform gas. The DM21 family runs at its official `conv_tol`$= 10^{-6}$; the other columns at 10^{-10}, all seven on the same canonical grid.

	DM21	DM21m	DM21mc	DM21mu	PBE	Skala-1.1	Ours
M1 η [%], exact 100, this window 100.03	81.86	41.85	110.57	42.66	0.0051	0.0051	96.52
least-squares SE of the fit [pp]	0.360	0.027	0.161	0.002	0.001	0.001	0.001
$\lvert 100 - \eta \rvert$ (to the asymptotic 100), lower better	18.14	58.15	10.57	57.34	99.99	99.99	**3.48**
$\delta(\frac{1}{2})$ [mHa]‡, exact 0	-1.04	-29.73	$\mathbf{+0.25}$	-28.86	-53.22	-47.63	-13.30
M2 count of $\varepsilon_{\mathrm{HOMO}} < 0$	1/3	1/3	**3/3**	1/3	0/3	0/3	**3/3**
ΔSCF electron affinity > 0	3/3	3/3	3/3	3/3	3/3	3/3	3/3
$\max_X \lvert \varepsilon_{\mathrm{HOMO}} + \mathrm{IP}^{\Delta\mathrm{SCF}} \rvert$ [eV], exact 0	2.04	2.18	**0.72**	2.17	5.31	4.49	2.43
M3 $\lvert r\, w_{\mathrm{xc}}(3) + 1 \rvert$, 0 only as $r \to \infty$	0.3685	0.3263	0.2589	0.3554	0.5534	0.4622	**0.2286**
M4 $\max \lvert D_\alpha - D_\beta \rvert$, exact 0	4.8e-02	1.5e-02	2.0e-02	7.3e-02	1.8e-15	1.7e-16	3.2e-15
M5 $\lambda(\mathrm{He}^+)/\lambda_1$ [%]	99.40	n.m.	n.m.	n.m.	98.36	99.23	99.56
M6 $\max \lvert \Delta\varepsilon_{\mathrm{xc}} \rvert$ / budget[§]	124.20	99.35	84.37	**0.92**	1.2e-06	56.71[†]	7.6e-09
M7 d [mHa], window $\lvert d \rvert \leq 1$	n.m.	n.m.	n.m.	n.m.	$+0.55$	n.m.	-0.0007

[§] In multiples of the 0.2 mHa/electron budget, which is ours and not a community standard; multiply by 0.2 for mHa/electron. The two units must not be read off the same row. [‡] $\delta(\frac{1}{2}) \equiv E_{\mathrm{H}}(N{=}\frac{1}{2}) - \frac{1}{2}E_{\mathrm{H}}(N{=}1)$: the deviation from the piecewise-linear line at half occupation of the hydrogen atom, exact value 0. Measured in the same run as η on the same benchmark, the two are tied by the closure identity $E_\infty = E_{\mathrm{H}} + 2\,\delta(\frac{1}{2})$, residual $\leq 7.6 \times 10^{-5}$ Ha for all seven columns. η **and** $\delta(\frac{1}{2})$ **are two readings of one condition and they order the columns differently**: η is the coefficient of the $1/R$ tail of stretched H_2^+, $\delta(\frac{1}{2})$ the curvature at the half-occupied point, and **on the second we are behind the two DM21 weights whose own ablation table records a fractional-charge training constraint**, and ahead of the other two. Ours reads compliance with a condition entered as a training target (Section 6), although this particular segment, $N{=}0 \to 1$ on H, was not itself one of the trained chord rows (Appendix D). Skala-1.1's value comes from its own baseline archive on this benchmark. The two controls carrying this row are in Table 1; SCAN's cell is not measured, for the reason given in Table G3, note a. [†] Not well defined without a convention for nuclear positions. Across the eight conventions tried the main-window deviation was always $\geq 44\times$ budget, with a spread across conventions of 256.47 mHa/electron, larger than the effect. Those eight have one limit, and it is what this cell prints (Appendix H). M5 and M7 were run for the `DM21` weight only. The M4 entries for PBE and Skala-1.1 are recomputed as the maximum over the two raw system columns rather than taken from a summary row.

Motivation for evaluating Skala-1.1, and what its stated constraints are. They "imposed only minimal constraints to maximize model flexibility, making it interesting to explore whether physical constraints can emerge from data", and test one such condition themselves with a positive result; Section 3 extends it to conditions they did not check. *Minimal is not none, and the list is theirs*: they "build in some of the most energetically relevant constraints (such as the high-density uniform coordinate scaling, size-consistency, and the Lieb-Oxford lower bound)" by construction, they apply the semi-local MLP "twice, once to each spin-ordering … followed by an averaging operation", and they pass the result through "a scaled sigmoid … yielding a bounded enhancement factor that enforces the Lieb-Oxford lower bound". Two of those meet a row here and read consistent with the statement: M4 reads 1.4×10^{-11} and M5 reads -0.00 pp. Size consistency motivates M7, a finite-R diagnostic that samples it only at $R = 6\,a_0$ and reads *n.m.* here. **These are readings beside a stated**

guarantee and not tests of it: the constants are not shared, we did not inspect their code, and an *n.m.* cell records what this benchmark did not reach, not a property of the functional.

Two readings that only this table supports, and one that costs us. First, *no released weight holds M2 and M6 together, and none reads close to exact on M1*: `DM21mc` takes Koopmans 3 of 3 at $84\times$ the uniform-gas budget, while `DM21mu` holds the uniform gas at $0.92\times$ and is 57 pp from exact linearity with Koopmans 1 of 3. Second, our advantage is narrower than the summary table suggests: on M2 we are tied with `DM21mc`, and on M1, where we now read closer to exact than that weight does ($|100 - \eta| = 3.48$ against its 10.57), η is a term in our loss and not in its. Third, M3 is the only condition where we are ahead of every column here on a reading that is not our own training target, and even that ordering is unresolved: our lead over `DM21mc` is inside the row's largest measured basis shift, and Appendix G.3.2 puts a hand-built functional below us there too (it prices all three rulers).

G.3 Hand-built controls: the full table, the gates, the three M3 uncertainties, and the long-range fraction

Their verdict rule was fixed beforehand. A pre-registration (SHA-256 `c812654f...89bd`) named the three models, the two cells at issue (M2, M3), and three verdict tiers: (A) all three controls miss both cells, and the claim under test stands; (B) any control takes M2 $3/3$ *and* M3 ≤ 0.1794, and that claim is withdrawn from the abstract and becomes a case study; (C) any control takes one of the two. *The* 0.1794 *is the sealed bar as written*, taken from the model reported when that document was signed; it is quoted unchanged and is not the reported model's present reading.

Tier B was triggered, by LC-ωPBE, and Section 4 and the abstract are written to it. The remaining criteria (M1, M4–M7, thermochemistry) were pre-declared reported-but-not-adjudicated by the tier rule.

Cryptographic verification of the pre-registrations. This paper prints two pre-registration digests. The one for the control pre-registration above is `c812654f...89bd`; the element-coverage one is `c3c5a0ad...7f6e`, and it names *the file as sealed* at 2026-08-12 19:50:57: a later appended section leaves its §§1–6 byte-identical but changes the file's hash, so the sealed text ships as a separate extracted file, whose digest is the one printed here. **Both are self-hashes**: with the commit time and file mtimes they indicate that a document existed unchanged before the runs it governs. They are not third-party timestamps, and we claim no priority beyond that ordering.

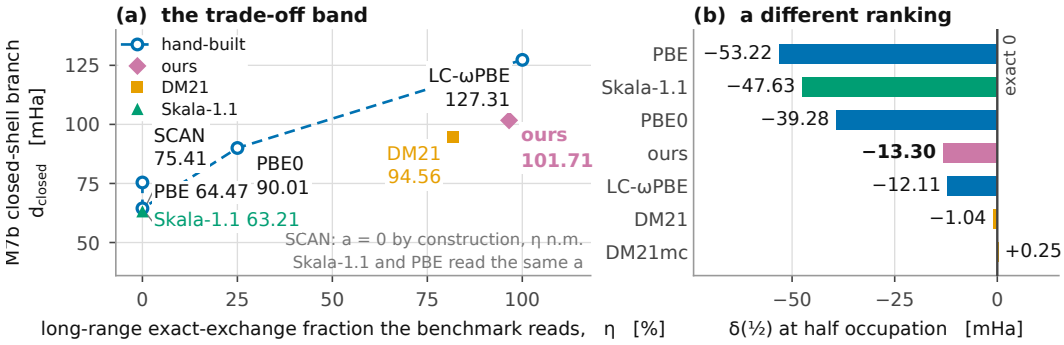

Figure G1: Hand-built functionals and ours *lie along a band* on the delocalisation–static-correlation trade-off (a), and it is the band a learned local mixing would have to leave. The tie at $a \approx 0$ alone spans 12 mHa, a fifth of the total range. All seven are measured now; ours reads worst of the three neural columns, DM21's under a damping ladder (this appendix). Half-occupation linearity (b) ranks them differently: three criteria, not one scalar.

The two released neural columns on the closed-shell branch. That branch is now measured for them too, on the same benchmark and definition (H_2 at $R = 6\,a_0$, aug-cc-pVDZ, grid 3, $d_{\text{closed}} = E_{\text{closed}} - 2E(\text{H}) - \Lambda$): **DM21** +94.56 mHa and **Skala-1.1** +63.21. On Table G3's M7b row DM21's sits between PBE0's 90.01 and our 101.71, while Skala-1.1's falls just below PBE's 64.47 (Figure G1a). The reproduction gate on that pipeline is PBE: 64.46878032827735 against the archived 64.46878032827713, i.e. 2.1×10^{-13} mHa. **DM21's entry carries a protocol that has to travel with it.** A literal transcription of the restricted solver every other column uses

Table G3: The three hand-built controls, with our reported model and PBE beside them. **This table carries only what is not already in Table G1 or Table G4**: the library settings, the M1 tail coefficient, the 36-system M4 aggregate, the M5 reading on H rather than He^+, and the closed-shell branch. Every criterion cell for these columns is in Table G1, and M7's gradient norm, $\langle S^2 \rangle$, curvature and certificate are in Table G4. Upper block: what the library was asked for, read back out of it rather than from memory — `pyscf` composes the range-separated exchange as $E_x^{\text{exact}} = \alpha K + \beta K_{\text{SR}}(\omega)$, so $(\alpha, \beta) = (1, -1)$ is exactly "no short range, all long range". **M4 (b) here and M4 (a) in Table G1 are two different measurements and must not be compared with each other.**

`pyscf`/`libxc` string	`PBE`	—	`PBE0`	`LC_wPBE`	`SCAN`
rung	GGA	Δ-on-GGA	hybrid	RS hybrid	meta-GGA
`libxc` functional id	101+130	—	406	478	263+267
(ω, α, β)	$(0, 0, 0)$	ω=0.4, coefficients learned	$(0, 0.25, 0)$	$(\mathbf{0.4}, \mathbf{1}, \mathbf{-1})$	$(0, 0, 0)$
tail coefficient c, 0 as $R \to \infty$	0.24999	0.00870	0.18748	−6.1e-05	n.m.[a]
M4 (b) $\max \lvert\Delta\langle S^2\rangle\rvert$, 36 systems	—	2.1e-11	1.6e-11	2.5e-10	2.9e-10[b]
count above the 10^{-9} gate	—	0/36	0/36	0/36	0/35[b]
M5 $\lambda(\mathrm{H})/\lambda_1$ [%]	99.31	99.44	99.76	**101.35**	99.36
M7b closed-shell branch [mHa]	64.47	101.71	90.01	**127.31**	75.41
$\lambda_{\min}$ there[c]	n.m.	−0.5096	−0.3611	−0.5278	−0.3410

[a] See Table G1, note (6): a library incompatibility in the solver's own error handler, not a property of SCAN. No estimate is entered.
[b] SCAN's `li+` cell did not converge, and its $\Delta\langle S^2\rangle$ there (0.89) is a non-convergence artefact that would otherwise dominate the aggregate; the number shown is over the 35 cells that did converge, and both readings are given, not the better one. [c] $\lambda_{\min}$ is the smallest eigenvalue of the finite-difference spin-flip Hessian at the stationary point in the row above it, on the same benchmark as the audited columns ($h \in \{2, 1\} \times 10^{-3}$ with h^2 Richardson, zero band 10^{-6}); for the three controls the stationary point was reproduced to the archived value before the curvature was taken. PBE's closed-shell $\lambda_{\min}$ was not computed. M7b is the reading on the closed-shell branch, a fractional-spin-like state mixed with finite-R static correlation, and not the symmetry-broken M7 readout; the broken-symmetry M7 readings, and the bounds d_{ub} with the seed rule that governs them, are in Table G4 and its note (3). The *Ours* column of this row is W-13-final, the reported model, whose broken-symmetry stationary point is an aufbau-constrained stationary point stable in the tested spin-flip subspace; among our earlier models only W-9-best reached one.

— bare DIIS, no damping, no level shift — *diverges* on DM21 from both starting points we tried (200 cycles each, final $\|\Delta D\|$ of 1.05 and 0.81). The value quoted is what that functional's own released protocol converges to with a level shift of 0.2 and damping 0.3: $\mathrm{tr}(DS)$ within 1.6×10^{-15} of 2, and the restricted and spin-compensated unrestricted routes agreeing on the same density to $|\Delta| = 0$, digit for digit. A scheduling accident ran that step twice, pricing its re-run scatter for free at 6.8×10^{-8} Ha (6.8×10^{-5} mHa), four orders below the smallest gap in the sequence above.

That pair is two runs, not a floor: repeating the reading on the released benchmark spreads it over 2.35×10^{-3} mHa, a re-run floor of about 2×10^{-6} Ha set by the protocol's *gradient* threshold (10^{-3}), still $\sim$1900× below this cell's 4.55 mHa distance to the M7b gate, so every reading rounds to 94.56 mHa (Appendix H). **The claim is "the solution that protocol converges to under that damping ladder", not "the lowest closed-shell solution".** Skala-1.1 needed no ladder (167 cycles from `minao` at `conv_tol` 10^{-11}). The restricted solver is not bitwise reproducible against itself (three runs spread by 2.2×10^{-16} Ha, its transcription by 4.4×10^{-16}), so the transcription check is a tolerance gate ($\leq 10^{-12}$ Ha), not an equality.

G.3.1 Three self-verification gates, an adapter check, and attribution

(H1) On a uniform gas, full-range exact exchange must equal Slater/LDA exchange: measured $\max|\Delta| = 0$, digit for digit. (H2) PBE must reduce to PW92 there: measured 2.4×10^{-7} mHa/electron against `LDA_C_PW_MOD`, and 3.7×10^{-4} mHa/electron against `LDA_C_PW` — the gap is `libxc`'s two parameterisations of PW92, not physics, and the smaller number is the like-for-like one.

SCAN's M6 cell is that same number: its measured main-window deviation is 2.4081776×10^{-7} mHa/electron against gate H2's 2.4081775×10^{-7} for PBE, agreeing to seven significant figures (Table G1, note (7)).

(H3, attribution) LC-ωPBE's *short-range semilocal* exchange model, compared against analytic short-range exact exchange on the same grid, deviates by 0.224442 mHa/electron — digit for digit its entire M6 miss. So the 1.12× in Table G1 is its short-range exchange approximation, and neither its exact-exchange component nor its correlation.

Adapter harmlessness was tested the same way: the same class with the string `PBE` reproduces the parent PBE model bitwise on 5 of 5 cells at one thread, while the parent model reproduces *itself*

only to $\sim 10^{-13}$ at four threads. The bitwise statement is therefore made at one thread, and the four-thread scatter is 10^4 below the project's 10^{-9} re-run gate.

Two things the M5 rows say that the bound does not. *First, a positive control.* SCAN reads $\lambda(\mathrm{H}) = 1.468824$ and $\lambda(\mathrm{He}^+) = 1.468835$ — one number to 1.1×10^{-5}, what freedom from one-electron self-correlation plus the uniform scaling covariance of exchange predicts. No other column does it: the two systems differ by 0.0140 for PBE, 0.0138 for PBE0, 0.0304 for LC-ωPBE and 0.0018 for our model — a stronger validation of this row than "no column exceeds the bound". The two released neural columns read $\lambda(\mathrm{H})/\lambda_1 = 99.66\%$ for `DM21` and 99.37% for Skala-1.1, so LC-ωPBE is the only column of the seven above the bound on either system. *Second, LC-ωPBE's excursion past the bound on H is not a cost of long-range exchange.* At $N = 1$ the exact statement is $E_{\mathrm{xc}}[n] = -E_{\mathrm{H}}[n]$; a one-electron density's long-range exact exchange is common to model and exact answer and cancels identically, leaving

$$E_{\mathrm{xc}}^{\mathrm{LC}}[n] - \big(-E_{\mathrm{H}}[n]\big) = \underbrace{\big(E_{\mathrm{x}}^{\mathrm{SR,model}} - E_{\mathrm{x}}^{\mathrm{SR,exact}}\big)}_{\text{short-range exchange model}} + \underbrace{E_{\mathrm{c}}^{\mathrm{PBE}}[n]}_{\text{self-correlation}} .$$

Measured on the same cached reference density that carries the whole row, the excursion is $-6.613\,\mathrm{mHa}$: -5.850 (88.5%) PBE correlation on a one-electron density, -0.763 (11.5%) the short-range exchange model. The split closes to 6×10^{-18} Ha and reproduces the archived semilocal components digit for digit. So the H row records that removing the exchange self-interaction error stops it *masking* a correlation self-interaction error PBE carries too, and He^+ stays inside the bound because that contribution falls with Z. The printed $\lambda(\mathrm{H})$ row has five columns; `DM21`'s and Skala-1.1's are quoted above instead.

G.3.2 Benchmark fidelity, and three M3 uncertainties not to be mixed

The replica reproduces the archived M3 at H^- for plain PBE (0.5534) and the archived reference model's $\varepsilon_{\mathrm{HOMO}}$ and $r\,w_{\mathrm{xc}}$ at F^- to the printed digits. 6 of the 8 regression cells agree to $\leq 10^{-9}$. The two that do not are both O^- (1.6×10^{-7} to 2.0×10^{-6}), which the archived grading of the same benchmark shows too — a known non-determinism of that open-shell SCF (Appendix G.12), not something the replica introduced. *Those two magnitudes are that comparison's, not a constant of the benchmark*: they were taken before the O^- occupancy patch, which moves that solution by about $1.9\,\mu$Ha, and no verdict on this page depends on either. The two cells that changed a claim were recomputed independently: the $\varepsilon_{\mathrm{HOMO}}$ signs with bare `pyscf UKS` under ordinary aufbau occupations (all nine signs identical), and $r\,w_{\mathrm{xc}}(3)$ with an independently written implementation of the zero-projection identity, agreeing to better than 10^{-6}.

Finally, **three different quantities circulate on the M3 row and none of them may be substituted for another**: (i) the *HF-control discrepancy* of the $r = 3$ reading against an analytic Hartree–Fock positive control on H^-, 0.0643. It is one control's residual; **that it is a common additive offset across functionals is not established here**, since projection residual, orbital decay and shell-averaging error need not be the same for each. (ii) The *matched-control residual*, where one was run: an analytic one-electron problem with no electron–electron term is solved in the *same element and the same basis*, with Z_{eff} chosen so its exact decay rate matches the measured decay rate of that cell's density, then pushed through the same shell-averaging pipeline. Its exact $r\,v/Z_{\mathrm{eff}}$ is -1 at every radius, so its deviation at $r = 3$ is basis and pipeline error. It reads 0.025 for PBE and 0.026–0.030 for the four `DM21` weights; the reported model, Skala-1.1 and the three controls carry none. (iii) The measured aug-cc-pVDZ→aug-cc-pVTZ *shifts* across the models holding both bases, $[0.004,\ 0.082]$.

Our distance to LC-ωPBE on that row, 0.0872, is larger than all three: than (i), than the top of (ii) and than the top of (iii). *All three are measured spreads, not bounds*, so the ordering is a reading under the declared aug-cc-pVDZ protocol, and a loss for us, rather than an ordering established beyond it. The reported model's own aug-cc-pVTZ shift on this row is 0.081. The LC-ωPBE control carries no aug-cc-pVTZ reading on this row, so the *differential* basis convergence of that pair is unverified. The same arithmetic puts the released `DM21mc` weight (0.2589) at 0.0303 from ours, below the top of (iii), so **the M3 ordering against that weight is unresolved as well**. That is the scope in which "smallest on M3" has to be read, and the scope we use everywhere in this paper.

The same row at the benchmark's own radius. The released bench gates M3 at $r = 5$ and declares $r = 3$ a historical read-out carrying no threshold (Table H3), so the two radii travel together in Table G1: $\mathrm{dev}(5)$ reads LC-ωPBE 0.0399, ours 0.0987, PBE0 0.3583, DM21 0.3670, SCAN 0.4694, PBE 0.4784, Skala-1.1 0.5436. Under the released gate LC-ωPBE alone reads inside it (0.0399 against 0.05), a margin inside the positive control's validated ± 0.10, so this paper prints the number and not that verdict. The limit that it is read at H^- only is in Table G1, note (9).

G.3.3 The long-range fraction: a second reading, and the coefficient itself

Beside M1's direct reading, the M3 row admits a second, two-point one,

$$a_{\mathrm{M3}} = \frac{\mathrm{dev}_{\mathrm{PBE}} - \mathrm{dev}}{\mathrm{dev}_{\mathrm{PBE}} - \mathrm{dev}_{\mathrm{LC}}}, \qquad \mathrm{dev} = |\, r\, w_{\mathrm{xc}}(3) + 1\,|,$$

calibrated to 0 at PBE (0.5534) and 1 at LC-ωPBE (0.1414).

One property makes it usable and one limits it. It is *invariant under any common additive offset*, since such an offset shifts numerator and denominator alike — a conditional property, not a claim that the HF-control discrepancy is one. But it is a *two-point interpolation of a quantity that is not linear in* a, so its accuracy has to be read off a column whose a is known — PBE0, which reads $a_{\mathrm{M3}} = 0.309$ against a declared 0.25, a residual of about 0.06 at that one column. **That is a calibration at exactly one point**: the scale is anchored at two columns by construction (PBE $= 0$, LC-ωPBE $= 1$) and validated at one. a_{M3} **is therefore an empirical comparison index, not an estimator with a model-independent error bound**, and the 0.06 is that column's residual and not a uniform scale error that may be carried to any other column or used to call a difference resolved.

On that ruler the columns read PBE0 0.31, ours 0.79, DM21 0.45 and Skala-1.1 0.22, against η of 0.250, 0.965, 0.819 and 0.00005. The two rulers differ by that 0.06 on the global hybrid and by 0.37 on DM21, 0.22 on Skala-1.1 and 0.18 on ours. That is why the DM21 reading is corroborated by the direct measurement of the coefficient below, not left resting on this scale alone. **We report that as a measurement, not as a diagnosis**: a disagreement of that size is what a coefficient varying with the local environment produces — what "local range-separated hybrid" names — but the two rows also differ in system, basis and observable. *The paragraph below separates them*: read off the grid on both systems on one pipeline, the coefficient is a constant in neither learned column.

The coefficient itself, read off the grid. Three facts fix what is being printed. (i) Both architectures are $E_{\mathrm{xc}} = E_{\mathrm{xc}}^{\mathrm{semilocal}} + \int \sum_i c_i(\mathbf{r})\, e_i(\mathbf{r})$ with e_2, e_3 the exact-exchange energy densities at $\omega = 0$ and $\omega = 0.4$; since $\mathrm{erf}(\omega r)/r \to 1/r$, what carries the $-1/r$ tail is the *sum* $c_2 + c_3$. **On this probe the energy-side ruler η reads the net cancellation of that tail, which for a constant-coefficient hybrid with no other $1/R$ term is that sum;** a_{M3} **does not read it**, because it is read off the potential and picks up a $\delta c/\delta\rho$ term (below).

(ii) DM21's released graph computes $c_i = 2\sigma(z_i/2)$ from a linear layer — one element-wise sigmoid per channel, **no coupling between channels and no normalisation**, so each c_i is free in $(0, 2)$ and the pair in $(0, 4)$. Both constants are read out of the released `SavedModel`, the supplement's Eq. (S1) stating the functional form, not the output map. Ours applies a symmetric soft clamp, $|c_i| < 4$, times the uniform-gas switch, with no sigmoid; the two parameterisations are not the same set.

(iii) The three channels were identified, not assumed. The graph's own output columns agree with independently computed e_{LDA}, e_{HF} and $e_{\omega\mathrm{HF}}$ to 2×10^{-7} relative, which is within the numerical precision of a `float32` graph. $\sum_g w_g \sum_i c_i e_i$ agrees with the graph's own energy sum and with the SCF's E_{xc} to 6×10^{-9} Ha. And the exchange energy densities agree between two independently written implementations running in two `pyscf` versions to 5×10^{-16} relative. The grid is the same point set for all five models ($\max|\Delta\mathbf{r}| = \max|\Delta w| = 0$).

	H^-, aug-cc-pVDZ (the M3 row)				H_2^+ at 20 Å, aug-cc-pVQZ (the M1 row)				$c_3/(c_2+c_3)$	
$\langle c_2+c_3\rangle$	r=0.5	**r=3**	r=8	all	r=1	r=3	**r=8**	all	H^-	H_2^+
DM21 (released)	0.387	**0.691**	0.884	0.616	0.547	0.750	**0.884**	0.578	0.357	0.320
Ours (W-13-final)	0.833	**0.766**	0.876	0.795	0.962	0.858	**0.882**	0.933	0.470	0.456
PBE0	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.000	0.000
LC-ωPBE	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
θ_0 (= PBE)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	—	—
the two rulers	a_{M3}: DM21 0.449, ours 0.788, PBE0 0.309				η: DM21 0.819, ours 0.965, PBE0 0.250					

Each entry is an $|e_{\omega\mathrm{HF}}|$-weighted mean over a shell of half-width $0.5\,a_0$ about that radius from the nearest nucleus, or over the whole grid; both profiles stop at $r = 12.17\,a_0$, the largest distance to the nearest nucleus the level-3 Becke grid samples, and $c_2 + c_3$ has reached its plateau well before that.

The two constant controls come through the same statistics digit for digit — 0.250 and 1.000 at every shell, every density bin and every global mean, 0 violations in 6 model–system pairs, and θ_0 reads exactly 0 — which licenses reading the two learned rows as measurement, not as an artefact of the estimator. They also caught a mismatched accumulation order in an earlier version of that weighted mean, and priced it at 110–138 last-place violations per pair, largest 1.3×10^{-15} — which is what a constant row is for.

What the table says. Neither ruler reproduces the coefficient where that ruler is read — compare the two bold columns above against the ruler row beneath them — and each learned column misses in one direction on both rulers: DM21 low on both (0.449 against 0.691, 0.819 against 0.884), ours high on both (0.788 against 0.766, 0.965 against 0.882). Neither coefficient is a constant: DM21 climbs from 0.387 near the nucleus to 0.884 in the tail, while ours starts high, dips and comes back (0.833, 0.766, 0.876), and the ranges of the individual coefficients are wider still (DM21 $c_2 \in [0.044, 0.700]$ on H^-; ours $c_3 \in [0.271, 0.537]$). **So "the effective long-range exchange fraction" is not a well-defined scalar for either learned column**: η and a_{M3} are *two differently defined diagnostics*, not two projections of one field. They are taken at different radii of different systems, and one of them reads the potential rather than the coefficient. They need not agree, and that is a property of how each is defined, not an error in either ruler.

Two mechanisms belong with that statement, and only the second is ours. First, the ambiguity has a name outside this paper: in the **local hybrid** construction (Jaramillo et al., 2003), exchange energy densities carry a gauge freedom whose artefacts Arbuznikov & Kaupp (2014) address by **calibration**, so a local mixing reproducing a given energy is fixed only up to that ambiguity. Our object is a multi-channel learned coefficient field, not a local hybrid, so this is a precedent and *not* a proof about our columns.

Second, and specific to which ruler is the outlier: a_{M3} is read off the *potential*, and for $E_{\mathrm{xc}} = \int c(\mathbf{r})\, e(\mathbf{r})\, \mathrm{d}\mathbf{r}$ with c a functional of the density, the functional derivative is not $c(\mathbf{r})$ times the exchange potential — it acquires a $\delta c/\delta\rho$ term. So the potential ruler is *not* expected to equal the $|e_{\omega\mathrm{HF}}|$-weighted mean of $c_2 + c_3$ even at the same radius on the same system. The table above agrees for DM21, where the potential ruler is the outlying one ($a_{\mathrm{M3}} = 0.449$ against a local 0.691 at $r = 3$, while $\eta = 0.819$ sits near its tail 0.884). On ours it is η that departs (0.965 against a tail of 0.882, with $a_{\mathrm{M3}} = 0.788$ near its local 0.766), so this mechanism does not account for our column. The two columns also divide the exact-exchange coefficient differently between the channels: weighted by $|e_{\omega\mathrm{HF}}|$, the $\omega = 0.4$ channel carries 47% of ours and 32–36% of DM21's. **Every sentence here is a measurement of released weights on our benchmark. We make no claim about how any of them was trained, and none about what any of them should have learned.**

G.4 Fragment interaction at $R = 6\,a_0$: stationary points and certificates

The verdict cells of Table G1 are read off the left half of Table G4.

The locus of what the term had to repair is fractional spin near an integer, not fractional charge — measured on an earlier model of ours, not re-measured for the one reported here.

G.5 Thermochemistry under both protocols

Table G5's cross-family readings are indicative only, and its own notes say why.

Table G4: Left: aufbau-constrained stationary points, with the spin-flip curvature $\lambda_{\min}$ separating a saddle from a point that is *stable in the tested spin-flip subspace*; a positive $\lambda_{\min}$ does not certify a minimum of the full orbital space. Right: variational upper bounds from an explicit trial density $D(t) = \mathrm{rot}(C, t)$, computed without any SCF — with $C = C_{\mathrm{PBE}}$ in the shared-seed column and that row's own closed-shell orbitals in the own-seed column. The three control rows were run on the same two benchmarks after the audit, their shared-seed bounds last of all, on a script that reproduces PBE's archived shared-seed value bit for bit.

	d [mHa]	$\|g\|_2$ [§]	$\langle S^2 \rangle$	$\lambda_{\min}$	spin-flip	d_{ub} shared seed[(3)]	own seed
PBE	$+0.5527$	3.25e-10	0.9935	$+0.4513$	stable	$+2.7625$	$+2.7625$
DM21	n.m.	—	—	—	—	$+2.7592$	—[‡]
Skala-1.1	n.m.	—	—	—	—	$+3.6291$	$+4.8778$
W-13-final	-0.0007[(4)]	3.96e-11	0.9650	$+0.3606$	stable	$+2.7905$	$+10.1364$
W-9-best	-6.4062[(4)]	2.62e-11	0.9695	$+0.6297$	stable	$+1.2626$	$+10.1493$
same model, archived	$+101.7859$[(5)]	—	—	—	saddle (old chain)	—	—
W-9-final	$+106.3294$[(1)]	4.92e-12	0.0000	-0.5475	saddle	-0.3591	$+10.8510$
W-8-final	$+106.3670$[(1)]	6.22e-11	0.0000	-0.5477	saddle	—	—
W-6-best	$+106.5984$[(1)]	1.29e-10	0.0000	-0.5159	saddle	—	—
hand-built controls (Appendix G.3)[(2)]							
PBE0	$+0.5994$	3.5e-10	0.9949	$+0.5175$	stable	$+3.2362$	$+6.6373$
LC-ωPBE	$+0.6057$	1.1e-09	0.9953	$+0.5762$	stable	$+3.3346$	$+13.3622$
SCAN	-0.4519	2.4e-11	0.9891	$+0.3497$	stable	$+2.5365$	$+3.0998$

[§] **One selection rule for this row, across all five columns.** Each M7 stationary point is reached from several starts; the members of one point — energies agreeing to $\leq 10^{-9}$ Ha, i.e. 10^{-6} mHa against a 1 mHa window — differ only in how hard the solver was driven. That row therefore reports the *best-converged member of the reported stationary point*, the rule stated in the code that produced the PBE column (`auf_table.py`, cluster tolerance 10^{-9} Ha). The reported model's stationary point comes from the judging chain instead, whose record stores every start, so the rule is applied to it directly: its cluster holds three members and the best-converged is 4.0e-11 against 6.0e-09 for the single run the chain reports. Under the alternative rule — the residual of the single run reported — the row is 9.0e-09, 6.0e-09, 1.1e-08, 1.9e-08, 3.4e-09; two columns then sit above our stricter 10^{-8} threshold instead of none, and *no verdict moves under either rule*, since every cell of this table is decided by d (Appendix A). [‡] Not printed. DM21's archived own-seed run starts from an early restricted solve that did not converge (the package lists it among its known caveats), distinct from the damped closed-shell solution of Appendix G.3. A trial energy from such a seed is still an upper bound, but far too loose to exclude anything at this threshold; DM21's entry is its shared-seed certificate.
(1) The three "$+106$" rows are not values of d in the M7 sense for those functionals: the same residual, read on a saddle branch, is not an admissible M7 verdict reading. They are the statement that under the aufbau constraint our search reached only a saddle for those models. *Reaching a saddle does not prove that no minimum exists*; it is why the validity gate marks them invalid, not failing.
(2) Every one of the three controls sits at a stationary point stable within the spin-flip subspace measured ($\lambda_{\min} > 0$ on the same finite-difference benchmark, with the stationary point reproduced to $\leq 9 \times 10^{-16}$ Ha of the archived value first), so their ◯ verdicts in Table G1 carry the same certificate PBE's does.
(3) d_{ub} is a property of the pair (functional, trial-density family), not of the functional, so only the shared-seed column may be compared across rows; with a fixed 16-point t grid every bound is conservative. That column spans -0.36 to $+3.63$ mHa across the nine rows carrying one, and the three controls move by -3.40, -10.03 and -0.56 mHa between the two seeds — the seed dependence itself, not a property of those functionals. No d_{ub} is below -1 mHa, so that column proves failure for nobody.
(4) Under the fixed-atomic-reference premise stated in Appendix A, what did prove failure at $R = 6\,a_0$ for an earlier model of ours is a converged aufbau value itself — W-9-best's $d = -6.4062$ mHa — an upper bound on the value at the global minimum for the same reason PBE's $+0.5527$ is: a lower solution may exist and would lie further outside the window on the same side. Readings *inside* the window, PBE's and the reported model's alike, prove nothing, so both are ◯ and not ✓ (Table G1, note (8)).
(5) The italic row is the archived reading for W-9-best under the previous grading chain, printed here so that the $+101.79$ mHa of Appendix A is traceable; it is a spin-saddle artefact, the same residual on a different branch and not an admissible M7 reading for that model, and is used nowhere else in this paper.
(6) The closed-shell branch discussed in Section 7: PBE's reading on this benchmark is $+64.47$ mHa, against the three $\langle S^2 \rangle = 0.0000$ rows above; it is not a row here because the stationary-point and certificate columns were not computed for it.

The riders that travel with this table. They are printed with the table itself, among them the three Coulomb treatments in play across the reference values, our runs and DM21's protocol — for which reason cross-family thermochemistry is indicative only.

Recomputation. Recomputed versus archived: $\max|\Delta| = 8.9 \times 10^{-12}$ kcal/mol, a summation-order effect of the interpreter that produced the archived derived columns, not a difference in the raw energies. The three hand-built columns were run afterwards and are placed on *exactly* the point sets of the panel they appear in, recomputed from their own raw energies; the statistics columns compare our model with PBE only and are unchanged. Percentages and ΔMAD are computed from the unrounded values, so recomputing them from the displayed columns can differ in the last digit.

Protocol, coverage and excluded points. aug-cc-pVDZ, exact Coulomb, grid 3, `conv_tol`$= 10^{-10}$, except DM21, which follows its official protocol (B3LYP warm start, `conv_tol`$= 10^{-6}$). Dispersion is omitted for every column alike. **Both neural columns are archived readings, not live ones**, and the released bench prints them as archived. `dm21` re-runs whole ($144/144$ systems). `skala` does not: its archived products were made at a 720 s per-system ceiling, a live re-run converges 139 of 144, two of the five failures carry a point of the sealed judging set, and the stage therefore *refuses* that model instead of scoring a shrunken set — so the Skala-1.1 column here is the archived one and **is not currently reproducible by the released benchmark** (Appendix H). Its SCF trajectory does not repeat run to run either. **Three different Coulomb treatments are involved** across the reference values, our runs and DM21's own evaluation, so cross-family comparisons here are indicative and are not one-benchmark comparisons in the sense of Table G1. The deviations also mix functional error with basis-set incompleteness, since aug-cc-pVDZ is compared against near-CBS reference values and we ran no basis ladder. **The three control columns were run at the protocol of our own models**, not the basis a hybrid is normally used at, so they bound what these functionals do *here* and not in general. **DM21's supplement reports 178 reactions shared between its training set and GMTKN55, including 6 G21EA and 16 G21IP points**; we could not map those onto the common point sets, so the two atomic rows are partly in-sample for that one column. For our model, most points involve elements represented by training endpoints (Figure G2 splits them by coverage). We have not established

Table G5: Thermochemistry on three GMTKN55 subsets, reported under both protocols because they disagree in sign. MAD in kcal/mol, recomputed from the raw energy column for every model. Panel (a) is the load-bearing protocol; panel (b) is a trend only; panel (c) is the reported model's point-by-point *disadvantage* against DM21. **The CI and Holm columns are point-level and descriptive**, and they are the reported model's: its SIE4x4 improvement over PBE passes point-level Holm correction (note e). Among our three earlier models, not shown here, six comparisons pass inside the two declared families (three favourable against PBE, three unfavourable against DM21, all on SIE4x4). None can reach significance at the system level (note a). LC-ωP heads the LC-ωPBE column. **Panels (a) and (b) are different point sets and must not be read across.**

		audited				*hand-built*			*ours versus the reference column*			
Subset	n	**PBE**	**Ours**	**DM21**	**Skala**	**PBE0**	**LC-ωP**	**SCAN**	**ΔMAD**	**rel.**	**95% CI (BCa)** *point-level*	**Holm p** *point-level*
(a) Common point set — load-bearing												
G21EA	23	3.29	3.29	1.91	1.92	2.85	3.14	3.69	$+0.00$	$+0.0\%$	$[-1.06, +1.58]$	1.000
G21IP†	33	3.93	4.29	2.26	2.18	3.72	4.97	5.01_{32}[d]	$+0.36$	$+9.1\%$	$[-0.82, +2.16]$	1.000
SIE4x4	11	19.54	9.32	3.98	6.92	11.23	5.24	14.08	-10.22	-52.3%	$[-13.47, -7.34]$	0.0264[e,a]
(b) Full point set — trend only, not load-bearing												
G21EA	25	3.44	3.39	—	—	2.86	3.35	3.80	-0.05	-1.5%[b]	not done	not done
G21IP	36	3.90	4.44	—	—	3.84	5.30	4.87_{35}[d]	$+0.54$	$+13.9\%$	not done	not done
G21IP†	35	3.98	4.29	—	—	3.75	5.13	4.99_{34}[d]	$+0.31$	$+7.7\%$	not done	not done
SIE4x4	16	23.13	10.51	—	—	13.91	9.11	17.72_{15}[d]	-12.62	-54.6%	not done	not done
(c) Ours versus DM21, on the same 11 points												
SIE4x4	11	—	9.32	3.98	—	—	—	—	$+5.34$	$+134\%$[c]	$[+3.26, +7.37]$	—[c]

† marks the rows with `IP_76` (P_2^+) removed, *symmetrically* from every model: a fresh run of the same released code reaches a solution $1.7165\ \mathrm{mHa}$ lower than the archived one along a different convergence path, so the cell is undecidable on this benchmark. The "Ours" column is the reported model in every panel and row.

[e] The reported model is not in the three-model family of this column's other Holm values, so its Holm is taken in the most conservative family we computed ($m = 27$, all our models against PBE across all three subsets); it stays below 0.05 under every family definition we examined, and the other two rows read 1.000 under all of them. [a] Point-level p / cluster-design floor, **not two p-values of one kind**. The 11 points are three dissociation curves (H_2^+, $(NH_3)_2^+$, $(H_2O)_2^+$) at $4 + 3 + 4$ separations (ICC $= 0.67$, $n_{\mathrm{eff}} = 3.9$), so the unit of inference is $n = 3$ curves, not 11 points, and $0.25 = 2/8$ is that design's floor, not a measured p, so *no significance is claimed* for this row — **nor for panel (c), which sits on the same 11 points**. What is robust is direction and rank: all 11 of 11 points move the same way, the largest single point accounts for only 18% of the effect, and the CI stays negative after dropping the most favourable point.

[b] The mean on this row is fragile in the opposite direction to the one the earlier model's was. The median paired difference is -0.09, the same sign as the ΔMAD, and we are better on 14 of the 25 points; but one adverse point (F^-, $+10.58\ \mathrm{kcal/mol}$) carries the difference between median and mean — without it alone the ΔMAD would read -0.49 rather than -0.05. *No claim is made on this row*; the panel is trend only.

[c] We are worse than DM21 on every one of the 11 points (0 better, 11 worse). *All six* of the reported model's comparisons against DM21 and Skala-1.1 put us behind on the point estimate, and so do all 18 of our three earlier models (W-9-best, W-9-final, W-8-best), the declared $m = 18$ family; three of those 18 reach point-level Holm significance, all on SIE4x4 against DM21 ($p = 0.0176, 0.0176, 0.0312$). The reported model is not in that family, so its Holm cell is empty; its own point-level sign-test p on this row is $2/2^{11} \approx 0.00098$, uncorrected. In the other declared family the same three models' SIE4x4 improvements over PBE reach point-level Holm $p = 0.0088$ each, an $m = 9$ value and *not* the panel (a) cell, which is note e's $m = 27$ value. At the system level the SIE4x4 comparisons cannot attain significance under the three-curve sign-flip design of note a; the atomic-subset comparisons show no Holm-significant difference in the reported point-level analysis. Neither is a reason to claim parity.

[d] SCAN's SCF did not converge on some systems, so its cells stand on fewer points than the row's n; *the subscript on a marked cell is that cell's own point count*: panel (a) G21IP† 32 of 33 (`li+`); panel (b) G21IP 35 of 36, G21IP† 34 of 35, SIE4x4 15 of 16. Every other cell in this table stands on the full n of its row. Missing points are never filled with an estimate; on the 32 panel-(a) points SCAN can do, the other columns read PBE 3.89, ours 4.36, DM21 2.27, Skala 2.21, PBE0 3.68, LC-ωPBE 4.97.

pointwise training-set overlap for Skala-1.1 and therefore do not label its atomic-subset results out-of-sample. *Excluded points, itemised.* Of the 16 SIE4x4 points, 5 are missing from the common set: He_2^+ at 1.0, 1.5 and 1.75 (our W-8 models), He_2^+ at 1.25 (Skala-1.1 and our W-8 models), and $(NH_3)_2^+$ at 1.5 (Skala-1.1 alone) — so 4 are attributable to us and 1 is not, and $16 - 4 = 12$ is the wrong arithmetic. Restoring the three that only our models failed gives $n = 14$ and raises PBE's MAD from 19.54 to 22.71, so the restriction is conservative against us — but that points we could not compute were removed from a benchmark has to be stated. G21EA loses 2 of 25, both to our own models; G21IP loses P_2^+ symmetrically and two further points to Skala-1.1. Four of the 11 SIE4x4 points are H_2^+, the system our M1 criterion and our η term both target, and are therefore in-domain for our objective.

Statistics. BCa intervals from 10^5 point-level (not clustered) bootstrap resamples, master seed 20260812; Holm correction inside two declared families, $m = 9$ for three earlier models of ours (W-9-best, W-9-final, W-8-best) against PBE and $m = 18$ for the same three against DM21 and Skala-1.1; note e's $m = 27$ covers the reported model. At $n = 11$ the bootstrap standard error of the MAD is itself about 2 kcal/mol. **No interval or p-value was recomputed for the control columns**, and none is claimed for them.

G.6 Do the condition readings predict the benchmark?

Each of the seven conditions supplies one or more read-out rows, thirteen in all, and each row is rank-correlated against the mean absolute deviation of each thermochemistry subset on the load-bearing point set, over the columns for which both exist: never more than seven, six for three rows and five

for three. **No cell of the** 39 **reaches** $p < 0.05$ **before correction** (the smallest raw p is 0.0583, M6 × SIE4x4) and every Holm-adjusted p is 1.000, so no robust association is established. It was the raw p that kept every cell out, not the correction: at $n = 5$ and $n = 6$ the smallest attainable two-sided permutation p ($2/5! = 1/60$ and $2/6! = 1/360$) could clear Holm only from rank 37 and 22, where $0.05/(40-k)$ first reaches them. Of the 30 cells with a declared better direction, the sign of ρ agrees with it in 15. No row carries an ordering verdict for a correlation to predict, and with $n \leq 7$ the comparison has almost no power, which is itself the result: *it is a finding about these columns and subsets, not evidence that the conditions are irrelevant*. The full grid (Spearman ρ, Kendall τ_b, exact p) ships as `results/tab/T_condbench_grid.json`, parsed from the table source; the recomputed subset deviations and the Holm table are in `results/runs/2026-09-04_T1W13/`. A post hoc check of the one non-uniform support set (SCAN converges 32 of the 33 G21IP† points) reproduces that row's ordering with no pair exchanged, so no cell moves.

G.7 The eight training gates, and how a checkpoint is chosen

M1–M7 ask whether a functional satisfies an exact condition. G1–G8 ask a different question: *is this round's Δ_θ worth deploying*. They are ours, built around our own label buckets and self-consistency bridge. Only two of them, G3 and G7, are defined for an arbitrary functional at all. The caption says which probes those two sit on and what the other six read for DM21 and Skala-1.1, which is not a statement about those functionals.

Table G6: The eight training gates. G3 and G7 are *defined on the same probes* as M1 and M7, at different thresholds. G3 gates $\eta \geq 50\%$ where M1 reads $|\eta - 100.03| \leq 2.5$ pp, and G7 gates $|d| \leq 1\,\mathrm{mHa}$ where M7's verdict is one-sided. So **a gate outcome here is not a criterion verdict there**: the same model reads *pass* in this table and ◯ in Table G1, whose M1 row prints a number and no glyph at all. A read-out with no threshold, a gate the released tool defines, and a gate we set for training are three ways of reporting, not three verdicts. G3 and G7 are the only two defined for any functional; the other six depend on our label buckets and self-consistency bridge and read "not measured on this protocol" for DM21 and Skala-1.1. A negative degradation is an improvement.

Gate	What it protects against	Threshold	W-13-final
G1	absolute endpoint energies not learned: residual against our QMC labels	train ≤ 2 / holdout $\leq 5\,\mathrm{mHa}$	**fail** $(11.19 / 68.55)^*$
G2	fractional-charge chord deviation not suppressed enough relative to PBE	train $\geq 50\times$ / holdout $\geq 10\times$	**fail**
G3 (M1)	H_2^+ identity closure and η	closure $\leq 10^{-4}$; $\eta \geq 50\%$	pass ($\eta = 96.5$)
G4	the H_2 curve: equilibrium contaminated, or stretched region not improved	equilibrium $\lvert\mathrm{err}\rvert \leq 2\,\mathrm{mHa}$ and no worse than PBE; stretched $\geq 3\times$ better	pass ($1.58\,\mathrm{mHa}$; $6.5\times$)
G5	potential and energy inconsistent: residual of the generalised Janak identity	$\leq 10^{-4}$	pass (8.10×10^{-6})
G6	numerical irreproducibility: same calculation twice, and grid drift	rerun $\leq 10^{-9}$	pass (1.4×10^{-14})
G7 (M7)	fragment interaction at $R = 6\,a_0$	$\lvert d\rvert \leq 1\,\mathrm{mHa}$	pass ($\lvert d\rvert = 0.0007\,\mathrm{mHa}$)
G8	H_2 atomisation energy degraded relative to PBE	degradation $\leq 20\%$	pass (-90.8%)
Total			**6/8**

In Table G6, the * on G1 marks two independent runs of the same calculation on the same checkpoint: that table's probe reads $68.55\,\mathrm{mHa}$ and the judging chain 68.58, a spread of $0.024\,\mathrm{mHa}$ that is not a difference of definition; it exceeds the open-shell band of Appendix G.12, is unexplained and is listed there under item (iii). Both round to the 68.6 quoted in Section 6, and no verdict on this page depends on which is taken.

Three protocol notes. *First*, G1 measures *absolute* energies, and improvement on its train side is close to a tautology, so we make no claim from it. *Second*, G2 is a suppression *factor* relative to PBE with a demanding threshold, so failing it does not mean losing to PBE. *Third*, G5 measures, not

assumes, the integrability that Appendix C shows this architecture holds by construction. Losing it is the theoretical risk a model that learns v_{xc} directly carries and this one does not.

Best versus final. Each round writes two checkpoints: the *final* one is the state at the last step, the *best* one is selected by $S = \mathrm{G1} + \mathrm{G2} + \mathrm{G4}$, an in-training score that **does not contain** η. The selection rule therefore does not optimise η directly. That is not the same as freedom from selection effects, since the terms of S can move η indirectly, and it is not a reason the numbers cannot be compared. Every η reported in Section 6 is a final model, so that comparison is like for like.

G.8 Round-by-round verdicts

What the chain does not settle. Capacity was not measured in this chain, so "too few parameters" stays a candidate explanation, not a finding, while the chord bucket carries most of the residual: it grew $1.85\times$ with no added parameters and kept 64% of the weighted residual.

Why not an earlier model. Earlier models look better on single conditions — 0.118 on M3 against our 0.229, and W-7-final's η of 96.53 against 96.52 — but both broke M4 and the M3 model broke M5.

Earlier rounds. Earlier rounds each changed more than one variable. A best model's η is read after a selection that the terms of S can move it through, and a final model's is not, so best models are not comparable on η with a final one; a round enters this appendix only through the change recorded for it in Table G7.

Table G7 records the reported round and three earlier ones at the granularity a reader needs to judge that chain; the round-by-round chart itself ships with the archive rather than being printed here.

One sealed side criterion of the reported round was crossed, and it is ours. That round's sealed document fixed a band on thermochemistry beside its tier rule: each of the three subsets to degrade by at most 10% against the denominator model, on the load-bearing reduced point set. The reported model degrades electron affinities by 19.4%, ionisation potentials by 8.7% and the self-interaction subset by 1.2%, so the first of the three is outside its band. The tier the same document assigns does not depend on the side criteria and is unchanged. We report the breach because the rule was fixed before the measurement, which is the only reason a fixed rule is worth anything.

The round was chosen after judging: its sealed document kept W-9-final as the reported model. *Held-out numbers from models trained on different label row sets are not to be subtracted*, the grading regime accounting for $\leq 0.04\,\mathrm{mHa}$ of any such difference (Appendix I).

What the decision record carries. Three disclosures from the round-by-round record are printed here. The reported model was **chosen post hoc**: eligibility was bounded in advance ($\lambda/\lambda_1 \leq 100\%$; met, 99.4/99.6), but the sealed document named a different one. That same round **crossed the sealed thermochemistry band on electron affinities**, $+19.4\%$ against $\leq 10\%$. And *"inactive" is a statement about the logged steps*: the fractional-spin hinge is written every 20th step and is zero at all 961 of them, its margin bottoming at $+0.015\,\mathrm{mHa}$ (step 400), so an excursion between two logged steps is not excluded by the log.

The sealed split has three disjoint groups: elements already covered before the data round, elements the round adds, and elements neither round covers, as a control. It carries four predictions: (P-1) the new group improves by ≥ 10 pp against PBE; (P-2) the new-versus-covered gap shrinks; (P-3) the control improves by less than half the new group's gain; (P-4) the covered group degrades by ≤ 5 pp. On electron affinities all four hold: the newly covered group catches up with the already-covered one (-21.8% against -21.7% versus PBE, from a 12.0 pp gap) while the control does not follow. On ionisation potentials P-1 and P-4 fail: the new group improves by 3.2 pp but stays 61.4% worse than PBE, and the covered group degrades by 6.9 pp. **No prediction here bounds coverage alone**, because the larger-label round also put the one-electron target on a fixed reference density (Table G7), and that change acts on every element.

Table G7: The reported round and three earlier ones, each with what it changed. W-13-final is the checkpoint whose numbers we report; its change is measured against the R model of Appendix I, not against the row above it, so the four rows are not a chain. *Scope of the pre-registration.* Each round's own gate thresholds were fixed in a sealed document before that round was launched, and those documents and their SHA-256 hashes are part of the release. *The reporting decision is not among them.* The sealed document for the reported round framed it as a single-model discriminating experiment and, in its opening line, *named a different model as the one that would be reported* — W-9-final, the earlier reported model. The decision to report this round instead was taken after it had been judged, is recorded with its date, and is therefore **post hoc**; the ground given for it is that round's own verdict tier, which found the targeted condition bought to within 90.0% of the gap while the cost fell on thermochemistry and the held-out residual. What that document did fix in advance is one bound on the choice — a report-model eligibility limit of $\lambda/\lambda_1 \leq 100\%$ on both one-electron systems — which the model met (99.4 and 99.6%); W-9, the round on which most of the earlier gate history was accumulated, carried no pre-registered reporting decision and no such bound. The pre-registration covers W-10 and the element-coverage split of Figure G2, and it does not cover the thresholds of the exact-condition audit itself (Table A2), which are ours and are declared rather than sealed. *Three different things carry that word in this paper and they are not interchangeable*: the control experiment's verdict tiers, sealed before those columns were run; the released tool's own thresholds, nine of twelve quoted from a sealed pre-registration line by line and three declared as engine constants (Appendix H); and the reading rules of Table A2, which are ours and are declared here and nowhere sealed.

Round[(1)]	Changed	η [%]	G1 train / holdout [mHa]	Refresh failures	Gates
W-13	explicit L_η, against the R model of Appendix I	96.52	11.19 / 68.55	1 of 10,902	**6/8**
W-9	refresh convergence tolerance	84.69	12.13 / 339.44	0 of 5,746	4/8
W-10	+12 heavy-element states; one-electron target on a fixed reference density[(2)]	71.44	13.55 / 57.26[(2)]	0 of 9,756	4/8
W-10R	+10,000-step anneal, warm-started from W-10's final weights with a fresh schedule ($5\times10^{-5} \rightarrow 5\times10^{-6}$), on GPU[(3)]	73.91[(3)]	13.50 / 63.43	—	4/8

(1) Where each column is read from. Each row's η and G1 pair come from that round's own judging product; for the reported round they are the eight-gate probe's readings, the same ones Table G6 prints, whose note $*$ separates the two independent runs of G1. The refresh counts come from each round's training log, and the gate count from the same eight-gate reading.
(2) W-10: checkpoints, residual decomposition, and two cautions on the label scale. W-10 finished its 20,000-step budget with no guard triggered, and its *best* and *final* checkpoints are the same parameters, all six tensors agreeing exactly; their largest held-out residuals nonetheless differ, by more than the declared run-to-run scatter and unexplained (Appendix G.12), *and a difference of that size is not, by itself, evidence of a benchmark fault. Residual decomposition*, regrouped by true membership because the grading chain's fixed lists predate this round's move of the carbon family to validation: 25 training-list endpoints (23 with an absolute-energy term; Al^- and N as chord references only) at $1.978\,\mathrm{mHa}$ mean absolute error, of which the 12 newly added heavy atomic and ionic states sit at 1.779; the carbon validation family at 10.393; the seven held-out systems at 27.739. σ_B is a composite empirical scale, not a stochastic error bar, so a ratio near one is not evidence that training is limited by label noise; the variational error of the same labels is a separate bias (Appendix F). Loss composition and endpoint conventions are in Appendix D.
(3) W-10R: not a doubled budget. A separate 10,000-step anneal of the row above (20,000 $\rightarrow$ 30,000 steps in all), declared with three deviations: a first launch truncated at step 1600 by a guard, that guard neutralised for the restart, and anchor-measurement mode with the arbiter off. Its G1 covers 35 of 36 systems (N^+ failed its SCF); its η cost $+5.2\%$ on electron affinities and $+2.7\%$ on ionisation potentials.

G.9 THE M4 RESIDUAL CLASS, AND WHAT IT COSTS IN OUR OWN FAMILY

Our own earlier family, not the intervention evidence. What restoring the spin-permutation identity costs downstream is measured on the released `DM21` weights (Appendix A); this panel is our *own* earlier family.

There M4-class residuals accompany benchmark outliers of 74.8–$98.0\,\mathrm{kcal/mol}$ and a fragment-interaction reading outside the legacy two-sided window — in **two different models** of that family, J-5 and W-4-best.

On a frozen density that is strictly additive by construction, the broken-symmetry model's own non-additivity matches its fragment-interaction reading in size and sign; restricting the functional to its spin-permutation-even part drops it by a factor of about 4000.

Two limits travel with the claim: J-5's d falls inside a $2\,\mathrm{mHa}$ window while the spin-flip violation ranges over $12.4\,\mathrm{mHa}$ along one training line, so that pass is a lottery and not evidence of size consistency; and causation is *not* established, since the violation grows by $+6.44\,\mathrm{mHa}$ across a step that moved no target, more than the $+5.91\,\mathrm{mHa}$ of one that did. The statement this appendix supports is that these residuals *accompany* those outliers and that gate miss, and that is all it says.

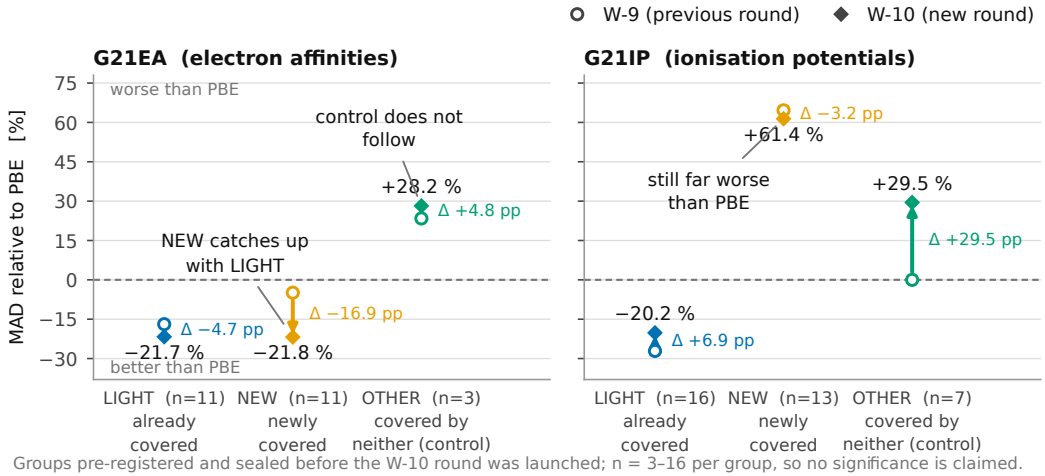

Figure G2: The pre-registered element-coverage test; groups were sealed before the run and are disjoint. Arrows run from W-9 (open marker) to W-10 (filled) on a shared axis where positive means worse than PBE, and the Δ beside each arrow is the change in percentage points, which is what the predictions are about. On ionisation potentials the covered group degrades by 6.9 pp and the control by 29.5 pp; two of the four pre-registered predictions fail there. With $n = 3$–16 per group no significance is claimed. Groups are full point sets, not the common sets of Table G5.

	J-5	W-4-best	(PBE)	
M4 (a) $\max \lvert D_\alpha - D_\beta \rvert$, H^-	2.26e-01	2.29e-01	8.3e-17	
the same at F^-	3.94e-01	2.67e-01	1.8e-15	
M4 (b) $\max \lvert \Delta\langle S^2 \rangle \rvert$, 36 closed shells	4.55e-01	2.13e-01	4.5e-12	
the three benchmark outliers, recomputed from the raw energy columns				
IP(F) [kcal/mol]	+97.97	—	+6.40	DM21 +3.56, Skala-1.1 +0.94
EA(F) [kcal/mol]	+76.65	—	+6.55	DM21 +3.68, Skala-1.1 +1.86
IP(HF) [kcal/mol]	+74.84	—	+0.84	DM21 +1.79, Skala-1.1 +0.52
fragment interaction, legacy two-sided gate (bold: not met)				
d [mHa], window $\lvert d \rvert \leq 1$	−0.08	**+11.89**	+0.55	
spin-flip violation $E(H, \alpha) - E(H, \beta)$ [mHa]	−0.72	+11.64	0 by constr.	
non-additivity on a strictly additive frozen density [mHa]	−0.72	+11.64	+0.0002	
the same after symmetrising the functional	+0.0022	+0.0028	—	

G.10 Cost, and what it is not a claim about

The correction carries 5,123 trainable parameters against DM21's 401,667 and Skala-1.1's 385,217, counting in each case the trainable weights of the enhancement-factor network, and its whole label set was produced on one laptop.

None of that is a claim of efficiency over either functional. DM21's network carries all of E_{xc}, not a correction to it, so the comparison is between two different objects; DM21 does not disclose its training cost, while Skala-1.1's model card reports about 48 hours on eight A100 GPUs for the functional training alone, which is not the cost of generating its labels; and our own deployment cost is dominated by the four local exact-exchange energy densities the features require, not by the weights, so a smaller network does not make an evaluation cheaper here. The parameter count supports a narrower statement: the two conditions this model holds by construction, M4 and M6, are held at this size, not by capacity.

G.11 Missing cells

Three cells of Table G1 are empty and stay empty, and the count is itself a limitation of the audit. Two of the three sit on the released neural weights, where a general statement would be most informative. (i) and (ii) The *symmetry-broken* M7 stationary point for DM21 and Skala-1.1 — the branch the criterion is read on. For DM21 the unrestricted H_2 SCF at $R = 6\,a_0$ did not converge from any of the initial guesses we tried; for Skala-1.1 the reading we obtained exceeded that functional's own strict variational upper bound by a factor of 33, proving it was not a global

minimum. *The closed-shell branch of the same system is measured for both* (Appendix G.3), for DM21 under a declared damping ladder; the two are different branches and neither substitutes for the other, and the shared-seed bounds in Table G4 close that asymmetry.

(iii) **M1 for SCAN**, for the tooling reason in Table G1, note (6), and still *not measured* after the released bench repaired that path, because one of two identical runs did not converge.

One further quantity is unmeasured without being a cell of that table, so the count is not the whole gap: the 36-system spin aggregate for DM21 and Skala-1.1, never run on that protocol. The ΔSCF electron-affinity count for our own column is now measured and reads $3/3$.

G.12 Known non-determinism

Three sources are declared, and none of them is a functional property. (i) *Branch selection in the stretched-bond diagnostic behaves as a lottery*: 7 of 51 evaluations land on the bad branch, so two rounds' values on that diagnostic are never differenced.

(ii) *A 10^{-6} Ha run-to-run scatter on open-shell anions*, which plain PBE shows too and is located in the SCF solver, forbids any claim of agreement at the 10^{-9} level; the criterion quantities affected are of order unity, five to six orders above the scatter.

(iii) *Two discrepancies exceed both bands.* First, the two gradings of the larger-label round's identical final and best weights give largest held-out residuals of 57.262 and $57.217\,\mathrm{mHa}$, a $0.045\,\mathrm{mHa}$ difference that is unexplained. It is 0.09% of the 52.2–$52.3\,\mathrm{mHa}$ by which the two readings miss this quantity's own $\leq 5\,\mathrm{mHa}$ held-out gate, which is why it changes no conclusion. *The distance that matters is that one*, on the same set; a cross-round improvement is not a denominator this paper's rounds support. Second, the $0.024\,\mathrm{mHa}$ between the two runs of G1 on the reported model (Table G6, note $*$), also unexplained, turns no verdict either.

G.13 Search scope for the claims of Section 1

The statement that none of the seven conditions carries a third-party reading of the criterion quantity for these released weights is bounded by the following search. Sources: CrossRef, Europe PMC, arXiv and OpenReview. Language: English. Cut-off: August 2026. The most likely place a competing audit would appear is the literature citing the two functional papers, which we traced: all 332 works citing DM21 and all 30 citing Skala-1.1 were screened at title level, 22 at abstract level (the search log ships with the archive).

It surfaced no competing audit, and several adjacent readings that the shipped log names. The one named here is the self-consistent evaluation of released DM21 on 36 hydrogenic one-electron ions (Schwalbe et al., 2022), the norm on which M5 lives. Its criterion is a relative total-energy error on self-consistent densities, not any criterion quantity of Table 1. The others grade accuracy or optimiser behaviour and read no criterion quantity either (Zhao et al., 2024; Kulaev et al., 2025). We treat them as we do the published M1 Comment: adjacent evidence, named here, filling no cell. Two author-side checks belong beside them. The `README` of the code released with Kirkpatrick et al. (2021) notes that a closed-shell system run unrestricted "can give a small difference between spin densities and eigenvalues with a negligible effect on the energy"; M4 here reads the deployed energy itself, on one fixed density. The Skala-1.1 authors test the positivity of the kinetic correlation energy on several atoms, with a data ablation (Luise et al., 2025, Sec. 3.3). Neither fills a cell of Table 1.

The same trace read two recent neural entries against this benchmark's purpose. The deorbitalised meta-GGA of Medvidović et al. (2025) hard-wires the von-Weizsäcker term and reports its scaling as inherited *by construction*, which is the class of claim an audit exists to read off the released artefact. Its code is released and its canonical weights are not, so it joins the stewardship candidates. The variationally trained network of Li et al. (2024) reports that energy minimisation pins the occupied manifold and not the unoccupied one, the mechanism of Section 5 seen independently.

Still not done: non-English literature, conference posters, patents, and full texts behind paywalls that we could not retrieve — those are recorded as bibliographic-level only in our reference ledger, and no statement in this paper depends on their contents. One of the eight functionals in the main table of Pederson & Burke (2023), CASE21, is machine-learned rather than hand-built, but it is a

fitted analytic form and not a network. What we did find on the two non-empty conditions: for M1, a published Comment on DM21 whose grounds are a training–evaluation overlap and re-examined examples, not a residual on any criterion quantity here (Gerasimov et al., 2022), with a Response (Kirkpatrick et al., 2022); and for M6, that same Comment and the DM21 authors' own ablation weights, which is the source of the `DM21mu` column above. Every citation in this paper is graded by how it was verified, and citations that could not be verified were removed, not downgraded.

H PHYSDFTBENCH: THE BENCHMARK CARD

This appendix is the card of the benchmark Section 2 describes. It collects the execution protocols and the implementation diagnostics behind them; the audited cells themselves are measured elsewhere and only re-printed here.

H.1 THE ROWS, THEIR PROBES AND THEIR GATES

Table H1 defines the benchmark row by row; the subsections after it say what is measured live, under what protocol, and what a reader can check without running anything.

H.2 SUPPORT MATRIX: WHAT IS MEASURED LIVE, AND WHAT IS READ OUT OF AN ARCHIVE

Table H2 says, for each way a functional can be attached, which rows come back live and which do not: six of the sixteen cells in the two released-neural rows are non-live, each with its own note.

H.3 PROTOCOL, AND THE NUMBERS THAT ARE DECLARED RATHER THAN SHARED

The benchmark stage adds no benchmark code: each kind runs through the driver that produced its own archived products, with the archived command line word for word — Table H1's thermochemistry protocol, plus a 600 s ceiling per system and 10 800 s per subset. A run that asks for another basis, grid or tolerance is *refused*, not warned about: the product's file name encodes basis and grid but not the tolerance, so it would carry an archived model's name while not being comparable with it.

Three declared departures travel with the rows they affect. **(i)** `skala`'s per-system ceiling is 720 s, the value its archived products were made with. It is a scheduling ceiling and not a protocol, but it decides *which* systems converge, so the stage records the value that really ran. Basis, grid, exact Coulomb and `conv_tol` are the shared ones, asserted for that model like every other. **(ii)** `dm21` runs at its own released `conv_tol` 10^{-6} / `conv_tol_grad` 10^{-3} with a B3LYP warm start, which is the protocol difference Table G1 declares. **(iii)** `skala`'s model card puts *single-pass evaluation at a fixed density* out of scope, and three of its rows are read that way. M6 admits no self-consistent reading at all; those three come out ✓, ◯ and ✗, not uniformly against it; and what is audited here is a released artefact under a declared protocol, not the deployment its authors recommend.

H.4 WHAT CAN BE CHECKED, AND THE RELEASE DIGEST

Two programs of that name ship, and a reader should run both. The twenty-check one, drawn in Figure 2, is `physdftbench/verify.py`; `verify.py` at the package root is a separate, larger ledger check over the shipped numbers. The package's own `README` itemises the twenty, and a skipping gate prints its own row, so the denominator is part of the state. **In the anonymised archive the twenty-check program exits non-zero, and that is expected** (Figure 2 prints the split). The root-level program exits zero with every check passing. Neither outcome is a statement about any functional. **Two verification strengths, and which cells have which.** That a file is present and the printed number was read from it is one statement; that a shipped program recomputes the number from the raw result is a stronger one. The first holds for every cell this appendix tabulates, with the one exception the reproducibility statement names — the refresh-failure count, whose training log is not part of the release. The second holds for what the root-level `verify.py` covers — the manifest, the two seal digests, the thermochemistry panel, the audit tables' criterion cells read back from the archived `json` by key path, the coefficient grid and the fragment-interaction and fractional-spin extras — plus the M4 (L3) row and the Koopmans-residual table of Appendix A.

Table H1: The nine conjoined rows and the two printed beside them. **These are the released tool's gates, not this paper's reading rules**: the paper reads M1 and M3 as numbers with no threshold and its M4(a) cells ungated (Table G1), while the tool must return a verdict, so it applies the thresholds below. Table H3 below sets the two regimes side by side, row by row, with where they part. "Exact" is what the exact functional reads on *this* probe, which for M1 is 100.03 rather than 100 because of the polarisation tail of the fit range. **One gate carries a declared caveat, and the released tool prints it on the row itself and not in this appendix alone.** The M7b gate is PBE0's own reading rounded to two decimals, which is why PBE0 fails the row named after it by $0.0038\,\mathrm{mHa}$. It is not quietly repaired, and the reason is not that a sealed number is inconvenient to change: M7b is one of the nine pre-registered conjuncts, so re-deriving it would alter the judging chain rather than a threshold, and that needs a new pre-registration; it is to be re-derived from an external anchor before it is unsealed. The M6 glyph records whether the declared budget is met; it does not encode the magnitude of an exceedance (LC-ωPBE's 0.2244 and DM21's 24.8 mHa per electron both read ✗), which is reported numerically.

row	quantity	probe	gate
M1	$\eta = 1 - c/0.25$, c the $1/R$ tail coefficient of symmetrically stretched H_2^+	H_2^+, 11 separations computed, 3 in the fit range; aug-cc-pVQZ, grid 3, `conv_tol` 10^{-10}	$\lvert\eta - 100.03\rvert \le 2.5$ pp
M2	Koopmans: $\#\{\varepsilon_{\mathrm{HOMO}} < 0\}$ over the three anions H^-, O^-, F^-	aug-cc-pVDZ, grid 3, fixed occupations	$= 3/3$
M3	$\mathrm{dev}(r) = \lvert r\, w_{\mathrm{xc}}(r) + 1\rvert$ at H^- (shell-weighted zero-projection profile)	aug-cc-pVDZ, grid 3; 288 angular pts, radial step 0.05, centres every $0.25\,a_0$	$\mathrm{dev}(5) \le 0.05$; $r = 3$ carries none
M4	$\max \lvert D_\alpha - D_\beta\rvert$ over H^-, F^- (density)	aug-cc-pVDZ, grid 3, fixed occupations	$\le 10^{-9}$
M5	$\lambda[n] = E_{\mathrm{xc}}[n]/(A_{\mathrm{LDA}}\int n^{4/3})$, H and He^+	*frozen* shared PBE reference density; no SCF, so the row is bitwise $E_{\mathrm{xc}} =$ PBE+Δ via the model's own coefficients	both $\le \lambda_1$
M6	$\max \lvert\Delta\varepsilon_{\mathrm{xc}}\rvert$ on the uniform gas	analytic features, $r_s \in [2, 10]\,a_0$, $\zeta \in \{0, 1\}$; no SCF and no grid	$\le 1\times$ budget ($0.2\,\mathrm{mHa}$/electron, ours)
M7	$d = [E(H_2@R{=}6a_0) - 2E(H)] - \Lambda$, symmetry-broken branch, $\Lambda = -0.8357\,\mathrm{mHa}$	aug-cc-pVDZ, grid 3, `conv_tol` 10^{-11}; 8 starts, lowest converged point with $\langle S^2\rangle \ge 0.5$	$\lvert d\rvert \le 1\,\mathrm{mHa}$ *and* $\lambda_{\min} > 0$ *and* FS-0 ≥ 0
M7b	d_{closed} on the closed-shell branch	the same probe, restricted solution	$\le 90.01\,\mathrm{mHa}$
thermo	MAD on three GMTKN55 subsets, recomputed from the raw per-system energies	aug-cc-pVDZ, exact Coulomb, grid 3, `conv_tol` 10^{-10}; the *sealed* common set, $n = 23/33/11$	all three better than PBE *and* ΔMAD $\le -1.0\,\mathrm{kcal/mol}$ on EA, IP
FS-0 held-out	$\min_x[E(H;1{-}x,x) - E(H;1,0)]$ $\max \lvert E_{\mathrm{SCF}} - E_{\mathrm{label}}\rvert$, held-out endpoints	$x \in \{0.002 \ldots 0.02\}$, M7's protocol our labels only; the *maximum* is the statistic, the mean printed beside it	≥ 0; a *term* of M7‡ reported, never conjoined

‡ **FS-$0 \ge 0$ is a one-sided lesion probe, not a compliance test.** The exact condition on that degree of freedom is fractional-spin constancy, $\Delta E_{\mathrm{fs}}(x) \equiv 0$, which *both* signs violate; this gate detects only the negative sign, the one that creates a spurious symmetry-broken minimum. *In the project-side sweep of Appendix A* — not in the released package, whose archive carries FS-0 for PBE and the reported model only — every hand-built column passes it while violating fractional-spin constancy by 0.4–1.0 mHa at $x = 0.002$ and by 41–117 mHa at $x = \frac{1}{2}$. Passing FS-0 therefore means "no anomalous sign", never "holds the condition".

Three objects share the name M7, and this is where they are separated. *The d component* is the numerical test $\lvert d\rvert \le 1$ mHa, which the implementation reports pass or fail. *The tool's aggregate verdict* for the row is the three-way conjunction Table H1 prints, and that is what the sealed engine decides the row on. *This paper's symbol* is neither: the ◯/✓ distinction is the one-sided rule of Section 2 and of Table G1's note (8), applied by us on top of the returned verdict. A reader who runs the release therefore sees a pass where this paper prints a circle; the circle is the weaker of the two readings, never the stronger.

Of three further sweeps, the M4 (L3) swap is rebuilt and two are not: the condition-to-benchmark grid of Appendix G.6 and the fractional-spin grid ship their raw files and this paper's numbers were read from them, not re-derived. Neither program recomputes the round of Appendix J. A green run is a statement about the cells the ledger covers, not about the paper. The swap carries this paper's strongest reading, which is why it was the first rebuilt: `verify_m4_swap.py` parses the shipped table source rather than carrying a copy of it, so an edit to the table moves the check with it. `verify_m2_koopmans.py` is the weaker kind and says so in its own header, because that table is written inline in the paper and this archive has no copy of it to parse. *Where a check is absent the archive names which of three reasons applies*: a dependency that cannot be shipped, an interface the plugin route does not carry, or a test that does not apply to that model. None of the three is a model failing a threshold, and the report keeps the two apart.

The current model's entry point. The root-level `verify.py` reads the reported model's cells from its `W13-final_*` products under `grader_w11/out/`; the sealed ledger stays as the record of what the tool measured when it was cut. The tool's M4 gate is not moved to match this paper's

Table H2: *live* = produced by a `grade` run; *archive* = read out of the registered archive; *n.m.* = the stage ran end to end and came back without a number *because of something it measured*; *no* = the stage refuses before it starts, for that kind. Each non-live cell is a statement about the pair (benchmark, functional).

how the functional is attached	M1	M2–M4	M5	M6	M7	M7b	FS-0	thermo
library string (`libxc:...`)	live	live	live	live	live	live	live	live
plugin module of the reader's own	live	live	live	live	live	live	live	live
checkpoint (ours)	live	live	live	live	live	live	live	live
`skala`	live	live	live	*n.m.*	*n.m.*[*]	live	live	archive[†]
`dm21`	live	live	live	live	*n.m.*[*]	live	*no*[**]	archive[**]
plugin (CIDER24Xe), partial[§]	*n.m.*[§]	live	*no*[¶]	*no*[¶]	live[‖]	live	live	*n.m.*[§]

[*] d and $\lambda_{\min}$ only, and measured on every run, not assumed: each model's own energy floor puts a descent's stall above the residual at which the chain declares it converged. For `skala` that stall is *measured* at 7.9×10^{-5} against a predicted 2.1×10^{-5}; for `dm21` it is *predicted* from the same floor and not reached. It is *not* a claim that either functional has no symmetry-broken solution. *Skala's* M6 is *n.m.* for a different reason: that reading is not quotable from any one of the eight declared nucleus/grid conventions. The spread, and the numbers that make it, are in Appendix A, M6.

Why those eight numbers have one limit. The released model's non-local layer couples each grid point only to the nucleus that owns it, through a Gaussian radial gate. That gate decays; it has no compact support. What the released widths give is a *numerical* cut-off: **in the released `float32` evaluation the coupling underflows to zero beyond a finite distance**, and on the blocks used here the reading is already bitwise constant well inside it. A uniform gas has no nucleus, so beyond that distance the reading is a constant, $11.34\,\mathrm{mHa}$ per electron or $56.7\times$ our budget — which is what the isolated convention already carries. Placing the nucleus far away returns it bitwise, and the robustness of that limit, the tensor-by-tensor check that the class we load the released state into is the released architecture, and the quadrature and cell-volume extrapolations behind it are all recorded in `results/runs/2026-09-14_M6UEG/` and `_RECHECK/`. **The tables print that limit and the released tool still returns *n.m.* here**, because its rows must be decidable from the declared protocol alone and this one is not. *One pipeline produced it*, on which four control columns and one Skala convention are bitwise at four threads. An anchor, not a second implementation.

[†] **The gate that refused a model we quote.** Both neural kinds were run whole for $0.1.6$; `skala` converged 139 of 144, and two of the five failures carry a point of the sealed judging set, so `thermo` *refused the model* rather than score a short set (the abort message names only the first: rule (i), the message is not the state). Skala-1.1's quotable thermochemistry is therefore the archived one (Table G5).

[**] The two `dm21` cells the notes did not reach. *FS-0 is* `no`*, not n.m.*: the criterion is defined for that weight and the released attachment cannot produce it, which is the distinction rule (i) of Section 2 draws. It is not that the quantity was never measured for `dm21` — a project-side sweep did measure fractional-spin constancy on that weight, and Appendix A reports what it found; that sweep is not this package and its reading does not fill this cell. *Thermo is* `archive` for the same reason the `skala` row is: the numbers this paper quotes are the archived ones and the bench prints them as archived.

[§] **CIDER24Xe (Bystrom et al., 2024) is a partial-coverage plugin demonstration, not an audited column** (its feature set originates with Bystrom & Kozinsky (2022)); four of the eight cells in this row come back non-live, so no table carries it as a column. It attaches through the reader-plugin route (Bystrom & Kozinsky, 2024) (weights: Zenodo `10.5281/zenodo.13336814`, `sha256 f2650a94...3aae1`; the vendor's own regression value is reproduced to 3.6×10^{-9} Ha). Its M1 is *n.m.* for the engine's registered defect, as `libxc:SCAN`'s is; its thermochemistry is *n.m.* by the coverage gate (9 of 144 systems non-converged). Measured: M4 3.3×10^{-15} (the density gate), M7b $83.8\,\mathrm{mHa}$, M2 $1/3$, M3 0.412; no causal reading is attached. [¶] not CIDER readings: the plugin hook evaluates those two stages from the *declared* `libxc` string, so they read R2SCAN — verified bitwise against the shipped `out/R2SCAN` ($6/6$ identical) and therefore marked *no*, not printed. [‖] *The released tool's* M7 conjunction is met ($d = +0.61\,\mathrm{mHa}$, $\lambda_{\min} = +0.52$ and FS-0 $= +0.67\,\mathrm{mHa}$, the last read from the FS-0 stage) — the tool's verdict, not this paper's validity requirement, which the row does not meet: the stationary point carries `strict_aufbau=false` with $\mathrm{gap}_{\min} = -0.163$ Ha.

reading (Table H3): its value is pinned to a line of the round's pre-registration, and moving a sealed threshold once the readings are in is the move this paper audits for.

The live judging route ships with its data pack, anonymised, as `physdftbench_datapack/` (11.4 MB): as unpacked, `verify_datapack.py` reports five passed, none failed, five skipped, and after one `localize.py` run nine passed, none failed, two skipped. No number in this paper comes from it.

Ledger verification. The tool's own ledger covers 473 cells at release $0.1.11$, unchanged at $0.2.0$: 352 rebuild from the archive, 121 are not recomputable from what the package carries, and 0 mismatch; the package's `README` itemises the 121, and **none of them is a load-bearing printed number**. The two mismatches the gate once reported were double-rounded literals, corrected and recorded in the package's `CHANGELOG`; neither verdict nor ordering moved.

The release tarball this paper was prepared against ships its own `MANIFEST.sha256`, which pins its bytes, and the offline verifier that reads it. That manifest is a file *inside* the archive, so the archive and the manifest have two different digests and both are named below rather than either being described as the other.

`physdftbench-0.1.11.tar.gz`	`a3ddd90b9b2bf778c35e02bbfba7722f0654aa14f60f322cbc38961b3d6235e0`
its `MANIFEST.sha256`	`27f6174f63169fa324ebcb0af0411a1537435bed523f06cbb4fcb33517fc5012`
the anonymised copy of it	`49b8ddcaa10c7907d8da4334d648ab254ae4fcc5d487f950c77f971dc0ba1ef9`

Anonymising rewrote the paths the manifest records, so the copy in this supplement has a digest of its own by design; the unrewritten original travels beside it as `MANIFEST.upstream.sha256`.

Table H3: **The paper's reading and the sealed tool's gate, row by row.** It **adds no measurement**: each row's evidence is its criterion's subsection of Appendix A and its rows of Tables G1 and G2. `python3 -m physdftbench paper-rules` prints the two side by side for the four rows its declared-rules file lists, for the reported model, W-13-final; the sealed ledger's *Ours* column is the legacy W-9-final (Appendix H).

	Paper reading	Sealed tool gate	Where they part
M1	η as a number, no threshold	$\lvert\eta - 100.03\rvert \le 2.5$ pp; the exact functional reads 100.03 on this window	the tool emits a verdict: the reported model's 96.52 is 3.51 pp from 100.03, outside its band
M2	anion-HOMO sign count, $3/3$	the same	nowhere
M3	$\mathrm{dev}(3)$ at H^- as a number, no threshold	$\mathrm{dev}(5) \le 0.05$ at H^-	the tool emits a verdict at $r = 5$: LC-ωPBE's 0.040 is the only pass among the columns read there; the reported model reads 0.099
M4	the deployed energy, spins swapped on one fixed density (L3)	$\max \lvert D_\alpha - D_\beta\rvert \le 10^{-9}$, the density layer (L2)	two different measurements; their verdicts agree on all 11 archived columns, which the paper does not rely on
M5	$\lambda \le \lambda_1$ on H and He^+, marked ◯ when met; Table 1 prints He^+'s distance to its exact value, Table G1 the ratio	the same bound on both	the symbol only: the tool passes what the paper marks ◯; the printed number has no gate
M6	a multiple of the 0.2 mHa/electron budget	the same budget	Skala-1.1's $57\times$ is read by a separate pipeline of ours; the tool returns n.m.
M7	one-sided: only $d < -1$ mHa proves failure; inside the window reads ◯	$\lvert d\rvert \le 1$ mHa, $\lambda_{\min} > 0$ and FS-0 ≥ 0, all three	the symbol is ours, applied on top of the tool's verdict

One object, one name. Four things here carry a version or an arm label and none of them is another:

the model this paper reports	W-13-final
the supplement's reporting interface	$0.2.0$
its sealed engine, pre-registration and digest	$0.1.11$, byte identical
the shipped cell ledger's *Ours* column	W-9-final, the $0.1.10$-era cut

The supplement is $0.2.0$ because its *reporting interface* changed: beside each sealed rule it prints this paper's reading rule for the four rows its declared-rules file lists (M1, M3, M4, M7), from a file whose header records them as declared rather than sealed and written after the results were in hand. Its sealed engine, pre-registration and the digest above are the byte-identical $0.1.11$ artefact, and none of $0.1.6$–$0.1.11$ changed any criterion, threshold, basis, grid, tolerance or assembly line (release $0.1.11$ records the ledger's W-9-final *Ours* column in `known_caveats.json`). **A green ledger run is therefore not a reproduction of this paper**; the supplementary archive's `verify.py` re-reads the reported model's cells from that model's own judging products.

No third-party data set, model or code is redistributed.

H.5 Caveats a reader should carry with the cells

(1) A reading is not its digits. Several rows sit at or below their own measured per-model re-run floor — the M4 density layer most of all, where seven independent readings of one model span a factor of 17.6. Those cells are to be read as "$\le$ this model's own floor". **(2) The per-model floor is per model.** For `skala` it is an order of magnitude wider (1.9×10^{-5} Ha over three repeats of one atom in a single process), and its SCF trajectory does not repeat either. **(3) The two verdict columns are two questions**, printed and not collapsed. **(4) One model reaches the benchmark stage through a graft** that does not carry that model's own response function, so it must be run through its own archived driver; routing it anywhere else reproduces a crash that driver was written to fix. That is a caveat of our graft, not a property of the functional.

I The discriminator round (exploratory; not a primary result)

Read this appendix under one rule. Its checkpoints and the model this paper reports were trained under the *same* label regime ("v2") and on the *same* rows (27 endpoint, 13 molecule, 117 chord, recorded identically in all four training logs), so nothing separates them on the data. **What separates the models here from it is the number of variables between them, not the data.** Only **R** is a single-variable contrast against the reported model — one added loss term, the rest of the recipe

word for word. **C0** and **W224** each differ from it in more than one place, so their cells are printed *beside* the reported model's and never subtracted from them. The label-regime dependence was measured, not assumed, and is small on both quantities: η reads the same under both regimes to 1.8×10^{-9} pp on the same checkpoints, and the held-out residual to $\leq 0.04\,\mathrm{mHa}$.

This appendix was not graded by the released PhysDFTBench. Its grader is the project-side chain of this round (`W-11`), whose `--drop-points` facility the package's own provenance file records as a "project-side judging-protocol feature, *deliberately NOT pulled into the package*".

Everything below is reported because it measures the mechanisms Sections 3 and 7 could only name; none of it changes a cell of the main tables.

I.1 Three models, and what each changes

The round asks which of three factors the gap between our column and the released neural columns is made of: the loss, the capacity, or the data. Three checkpoints were trained on the same v2 labels for the same 20,000 steps. **C0** is the denominator: the v2 label regime, $W = 64$, 5,123 parameters. **R** adds the two repair terms and nothing else (`--lam-fs 500 --lam-m7 500`). **W224** widens the network to $W = 224$ (53,763 parameters, $10.49\times$). It *also* halves the initial learning rate to 2.5×10^{-4}. That rate is the first rung of a pre-registered learning-rate ladder, and it passed and stopped the probe there.

That second change is declared, not discovered: the W model moves two things at once, so its verdict reads "capacity *and* the learning rate that goes with it", never "capacity alone". The pure-capacity contrast, a model at the old rate, was not run.

The round's own stop line was crossed rather than met, and two readings of it are in circulation. Against the sealed reference C0 sits 5.00 pp outside the band; the judging product stores the other reading, against a different arm and marked there as "not a criterion, reported alongside". Both readings and the sealed rule are in the round's record. **The round continued by an explicit recorded decision, not by passing.**

The mitigation, and its limit: every comparison below takes C0 as its denominator, not an earlier model, so no verdict here rests on a cross-round difference. One reading would have done so: a four-point η trend against earlier models. It is not reported.

I.2 Impact on fragment interaction and computational cost

M7 is a conjunction of three readings, and the R model turns the two C0 failed (Table I1); the third was already passing. d moves from -7.367 to $+0.000329\,\mathrm{mHa}$, inside a $1\,\mathrm{mHa}$ window; the spin-flip curvature stays positive, so the point is stable in the tested spin-flip subspace; and FS-0, the fractional-spin term that is the lesion the symmetry-broken branch finds, moves from -3.172 to $+0.496\,\mathrm{mHa}$. It is the only cell that changes verdict anywhere in this round. **What that change is, stated under our own rule.** C0's $-7.367\,\mathrm{mHa}$ is a converged reading below the window and therefore a *proved* failure; R's $+0.000329\,\mathrm{mHa}$ is one *inside* the window, which under that rule proves nothing (Table G1, note (8)). **The convergence-independent certificate says so**: R's shared-seed variational upper bound (own seed $+11.4364$) is *outside* the window, as Table I1 shows for every model but C0. So the honest statement is that *the proof of failure is removed* and not that size consistency is proved. That removal is the largest verdict change in this study, *on the seed where the failure was proved*. A second C0 seed does not fail M7 at all.

The side criteria, with the two cells that moved against the model. η *improves* by 1.15 pp on the first seed and 1.34 pp on the second, so the direction reproduces. The magnitude does not: within one model the two seeds differ by 3.22 pp on C0 and 3.41 pp on R, which is larger than either improvement. At two seeds per model that is a statement about resolution and not about size. M3 moves by $+0.019$, inside the sealed ≤ 0.029 tolerance for that row.[1] Two cells moved the other way: **M7b** $98.98 \to 101.21\,\mathrm{mHa}$ ($+2.23$, about 10^3 times the re-run floor Appendix G.3 derives for that cell) and **the held-out residual** $25.6 \to 29.3\,\mathrm{mHa}$ **on the mean and** $53.1 \to 59.4$ **on the maximum**. The two seeds differ by $1.22\,\mathrm{mHa}$ on C0 and $1.30\,\mathrm{mHa}$ on R, and they do not agree

[1] That ≤ 0.029 is a sealed number, taken from the matched-control residual for $\mathrm{dev}(3)$ as it stood when the round was sealed, and not derived for this row, which is $\mathrm{dev}(5)$.

on the sign of this cost: $+2.23$ on the first and -0.29 on the second. On the first seed it clears the weaker half of the thermo gate, like C0, and fails only the $\Delta\mathrm{MAD} \leq -1.0$ half. Its ionisation subset lands within $0.004\,\mathrm{kcal/mol}$ of PBE (3.9708 against 3.9748, its denominator at 3.80). On the second seed it misses that half too: the same subset lands $0.055\,\mathrm{kcal/mol}$ above PBE. *On the first seed the repair costs M7b and one atomic subset. It costs the held-out residual on both.* What it buys is narrower than one seed suggested. On the seed where C0's failure is *proved* the repair removes that proof. On the second seed C0 reads $+0.8577\,\mathrm{mHa}$, inside the window, so there is no proved failure there to remove. The asymmetry is in the evidence, not in the functionals. C0's first-seed failure carries a certificate inside the window and its second-seed pass does not.

Table I1: M7 is three readings; the repair model turns the two C0 failed, and the curvature row was already passing in both. PBE is the benchmark's control on the same probe, printed for scale and not as a column to subtract from. **The bottom two rows are a gate outcome and a certificate, not a proof**: a converged d inside the window proves nothing, and the variational bound d_{ub} — which needs no optimiser — is outside the window for every model but C0. **The reported model is not a column here**: its own M7 cells are in Table G4, from the same chain on the same probe, and of the three models above only R is a single-variable contrast against it.

component	gate	PBE	C0	R	W224
d, symmetry-broken branch [mHa]	$\lvert d\rvert \leq 1$	$+0.5527$	-7.3672	$\mathbf{+0.000329}$	-4.0541
$\lambda_{\min}$, spin-flip curvature	> 0	$+0.4513$	$+0.5671$	$\mathbf{+0.3345}$	$+0.5241$
FS-0 at $x = 0.002$ [mHa]	≥ 0	$+0.5652$	-3.1719	$\mathbf{+0.4963}$	-1.4957
d_{ub}, shared-seed certificate	—	$+2.7625$	$+0.3701$	$+3.0608$	$+2.9712$
M7 verdict, one-sided		◯	✗ *proved*	◯	✗ *proved*
the sealed three-part gate		met	not met	**met**	not met

One of the two repair terms was inactive at every logged step. Reading all 1001 logged training rows of that model, the fractional-spin hinge L_{fs} is zero on all 1001 rows. Its violation counter never leaves 0, and the quantity it guards stays positive throughout (smallest value over the whole run $+0.0846\,\mathrm{mHa}$). **Its measured gradient norm is identically zero on all $1{,}000$ rows that carry it as well** (the step-0 row logs it as undefined), which is the direct evidence rather than the inference: **at every step the log records, the repair was done by the M7 term alone**.

One limit on that log: it samples every 20th step of $20{,}000$. And that smallest sampled margin ($+0.0846\,\mathrm{mHa}$, at step 400) is smaller than the largest step-to-step excursion anywhere in the log ($0.0913\,\mathrm{mHa}$). So an excursion through zero between two logged rows near step 400 is not excluded by the log alone. The opposite claim — that both terms contributed — is the natural one from the command line, and the logged steps do not establish it. What the round did *not* test is whether that guard rail matters on data where it does fire.

The cost, stated with the result. In checkpoint R one system pays, and it is a pathology, not a slowdown. CO^+ (an open-shell cation of the ionisation subset) does not converge under the R functional: it hits the 600 s protocol ceiling at 620 s on a near-idle machine. A longer diagnostic run, outside the grading protocol and used for nothing else, still returned no converged energy; its per-stage timings are in the round's record. The whole robustness ladder was exhausted, not merely truncated. **The same system converges on plain DIIS in 25–40 s on every other model** (25.3 and 25.7 s on the two v1 models, 40.0 on C0, 36.8 on W224). The widest of them has $10.49\times$ the parameters, and it covers the full point set without a gap. So the pathology attaches to checkpoint R and not to capacity; the reported model carries the same two repair terms and converges on that point on plain DIIS. **R's repair, its measured costs above, and this SCF failure are one result and have to be quoted together, and the cost is not optional.**

What the released gate would have done with this model. The sealed ionisation set has 33 points. R is scored on 32, the missing one being the CO^+ point above. Under PhysDFTBench's own coverage gate a missing sealed point refuses the reading rather than shortening the set. It did that to Skala-1.1 over two sealed points and to CIDER24Xe over nine non-converged systems (Appendix H). **Had the released benchmark graded this model, its ionisation cell would have been refused and not scored.** The head of this appendix says that benchmark did not grade it, and no main-table cell depends on it. Two things keep the reduced reading worth printing. Every model quoted beside R is read on the same 32 points, PBE and the two neural references included. And the drop is itself

worth $0.041\,\mathrm{kcal/mol}$ on PBE, which reads 3.9341 on the full 33, the 3.93 of Table G5, and 3.9748 here. That is ten times the 0.004 above, so this set does not resolve the near-tie.

I.3 Width 224 at its halved learning rate: η moves, thermochemistry does not

The capacity model was graded against a pre-registered three-tier rule: tier A if both atomic subsets improve on C0 by at least $1.0\,\mathrm{kcal/mol}$, tier B if one does or both improve by half that, tier C otherwise. **Both cells came out with the wrong sign**: $\Delta\mathrm{MAD} = +0.402$ on electron affinities and $+0.508$ on ionisation potentials, *worse* than the ten-times-smaller network. The degradation is not resampling noise: the bootstrap intervals exclude zero ($[+0.116, +0.754]$ and $[+0.254, +0.923]$, BCa 95%, $B = 10^5$) at $p_{\mathrm{boot}} = 8.7 \times 10^{-3}$ and 1.4×10^{-4}. That is tier C: *widening to* $10.49\times$ *the parameters together with the halved learning rate did not improve atomic thermochemistry*; it raised η ($+5.88$ pp) and slightly improved the held-out residual. The M3 residual moves by $+0.0009$, the M7 lesion halves without changing the verdict, and it costs $2.59\times$ the wall clock. The self-interaction subset improves by -0.446 ($p_{\mathrm{boot}} = 0.023$) and is reported rather than adjudicated, because the sealed rule named the two atomic subsets only.

What p_{boot} is, what the correction does to it, and what none of it can reach. p_{boot} is a two-sided paired bootstrap tail probability on the point-level statistic $\theta = \overline{d}$, with points resampled with replacement (*not* clustered) $B = 10^5$ times; its floor on this machine is $2/(B+1) = 2.0 \times 10^{-5}$. The estimator, the random-stream construction and the seeds are in the method record. The bias-and-acceleration correction is applied to the intervals (BCa) and *not* to p_{boot}, and the two are computed from the same resamples, so they are not two independent pieces of evidence.

Holm's step-down correction over the three subsets leaves every verdict where it was: adjusted p_{boot} is 4.2×10^{-4} (ionisation potentials), 1.74×10^{-2} (electron affinities) and 2.27×10^{-2} (self-interaction); restricting the family to the two atomic subsets the sealed rule names gives 2.8×10^{-4} and 8.7×10^{-3}. The two atomic degradations also survive the correction on a Wilcoxon signed-rank test (8.8×10^{-3} and 2.71×10^{-2}); **the self-interaction improvement does not** (6.74×10^{-2}). Nor are its eleven points independent — three dissociation curves (ICC $= 0.67$, $n_{\mathrm{eff}} = 3.9$), so no system-level significance is claimed for it either. Nor is any of them a training target. Their neutral fragments are training rows, but no cation dimer is, and neither is H_2^+ at any separation.

The limit that no correction repairs. These intervals resample *evaluation points*. The conclusion is about a *training* design choice, and there is **one training run per configuration**: no seed replicate, and therefore no training variance inside any interval here. The SCF pathology above is also $n = 1$ per model. The M7 and η readings above carry two seeds per model.

One alternative explanation this round cannot exclude: the capacity model runs at half the production learning rate on an unchanged 20,000-step budget, so "under-trained at a halved rate" is not separated from "capacity does not help" by this round. The rate itself was not tuned to the outcome. A three-rung ladder was pre-registered to be run largest first and stopped at the first rung passing four conditions fixed before any of them was launched; the gates and their values are in the method record. The first rung passed, so the two smaller rungs were never run. Nothing in this round licenses a capacity claim at the production learning rate.

I.4 What this round settles, and what it does not

The loss can turn an exact condition, at the price the subsection above prices on the first seed: $+2.23\,\mathrm{mHa}$ on M7b, a tie with PBE on one atomic subset where its denominator passes, $+3.66/+6.29\,\mathrm{mHa}$ on the held-out mean and maximum, and, in checkpoint R, an SCF pathology on an open-shell cation that neither the wider network nor the reported model reproduces. *Width together with its learning rate* raises the fractional-charge reading and measurably *worsens* both atomic thermochemistry subsets. *The data regime* is the factor this round did **not** isolate: the model that would separate the label *values* from the row *set* was never run. So "the deficit lives in the data" remains a hypothesis that this round narrows by elimination and does not test. Three further limits: the capacity verdict rests on one training run per configuration; the repair terms were used at a fixed weight whose measured gradient share came out about four times the design target, so that weight is not calibrated; and L_{fs}'s value is untested, since it was inactive at every step the log records.

J THE LONGER-BUDGET RUN (NOT THE REPORTED MODEL): WHAT THE EXTRA STEPS AND THE SLOWER ANNEAL BUY, AND WHAT THEY SPEND

Read this appendix under one rule. The longer-budget run (`W-14`) is not the model this paper reports. It is one intervention on the reported model, and *one* setting separates them: the step budget, $20{,}000 \to 60{,}000$. That one setting changes two things at once, because the trainer takes the horizon of the cosine anneal from it. So the two runs sit at different points of their schedules at every step they share and the learning rate at a given step is not the same in both. **What the leg buys is therefore bought by budget and learning-rate schedule jointly, and is not attributable to step count alone.** Buckets $(27 + 13 + 117)$, label regime, width, parameter count, seed and the two learning-rate endpoints are identical, so the leg from the reported model to the longer-budget run isolates the budget and anneal schedule. The earlier rounds set beside it below change more at once.

J.1 PRE-REGISTERED EVALUATION METRICS

Five layers, in this order: the targeted quantity; a second ruler on the same condition; the same ruler on a window widened with separations the loss never fitted; the cost on every other criterion; and the construction invariants. A round in which the targeted quantity improves and other rows degrade is an intervention map, not a candidate recipe, and the reporting model does not move. That is the case here (Table J1).

Table J1: The five-layer readout of W-14, against the reported model. η is a term in both models' losses, so layer 1 says what the intervention buys on the probe it trains on and nothing about generalisation. **Layer 4 is not uniformly untargeted**: M7 enters both losses as well, through the fragment-interaction residual d and the fractional-spin hinge, so its conjunction row and the FS-0 row under it are targeted readings too. The quantities no term of ours is written against are $\delta(\frac{1}{2})$ — a second reading of the M1 condition, on a segment that was not one of the trained chord rows — M3's $\mathrm{dev}(5)$, and M7b's closed-shell branch. Layer 3 widens a window the loss already fits rather than withholding one. **Row 3 is window sensitivity, not a second verdict**: the 100.03 reference is what the exact energy reads *on the* $\{20, 35, 50\}$ *Å window* because of its polarisation tail (note g of Appendix A), and moves with the window.

	quantity	reported model	W-14	reading
1	M1 η [%]	96.52	**100.49**	targeted; $\lvert\eta - 100.03\rvert = 0.46$, inside 2.5
2	$\delta(\frac{1}{2})$ [mHa]	-13.30	-7.32	untargeted, same condition; moves, not to zero
3	η, $R \geq 8$ Å window	—	100.86	augmented window; $+0.37$ pp on row 1
	η, $R \geq 5$ Å window	—	103.83	augmented further; $+3.34$ pp on row 1
4	M3 $\mathrm{dev}(5)$	0.0987	0.1287	worse
	M7 conjunction	met	**not met**	fails on FS-0 alone
	FS-0 [mHa]	$+0.4377$	$\mathbf{-0.0089}$	the term that flips
	M7b [mHa]	101.71	97.24	better, still outside 90.01
5	M4 (a) [gate $\leq 10^{-9}$]	3.2×10^{-15}	3.5×10^{-15}	density gate met (identity: App. E)
	M6 [×budget, gate ≤ 1]	7.6×10^{-9}	1.0×10^{-8}	deployed reading, gate met

J.2 THE COST IS NOT A ONE-ROUND ACCIDENT: FOUR MODELS SHOW WHERE IT LANDS

The longer-budget run's chain has a common window of its own: G21EA 23, G21IP 32, SIE4x4 11 points. That is *not* the load-bearing set of Table G5, whose G21IP† row has 33. On it ΔMAD against PBE (negative is better than PBE) reads:

model	G21EA	G21IP	SIE4x4
C0	-0.34	-0.18	-10.16
R	-0.53	-0.00	-10.34
the reported model	$+0.00$	$+0.34$	-10.22
W-14	$+0.54$	$+0.56$	-10.34

The two windows agree on G21EA and SIE4x4 and differ only on G21IP, where the reported model reads $+0.36$ on the load-bearing set; the four models are compared here only against each other, inside the one window. G21IP degrades at every step; G21EA improves from C0 to R and degrades after it; SIE4x4 has no trend, moving inside a 0.18 band. The cost is therefore concentrated on electron affinities and ionisation potentials, not on self-interaction. *Only the last leg of that model*

order is the budget-and-schedule change by itself. C0→R changes more than one thing; R→W-13 adds the η term alone (Appendix I).

J.3 Every checkpoint on the ladder is reported: none was chosen

The round's sealed addendum fixed, before any of them was graded, that the post-plateau checkpoints would be graded and reported *in full* and that the verdict would read the final one. Three intermediate checkpoints and the S-selected *best* were graded on the same benchmark (Table J2).

Table J2: The longer-budget run's ladder, all reported. M7 is the sealed three-way conjunction $|d| \leq 1\,\mathrm{mHa} \wedge \lambda_{\min} > 0 \wedge \text{FS-0} \geq 0$. **The three intermediate checkpoints meet it and the final one does not**; the verdict reads the final one, as fixed in advance. The *best* checkpoint is selected by the in-training score S of Appendix G.7, which contains no η, and it has the largest M1 deviation of the five; the final one loses M7.

checkpoint	η [%]	$\lvert\eta - 100.03\rvert$	M1	FS-0 [mHa]	M7
step 30,000	100.46	0.43	pass	+0.2142	pass
step 40,000	100.81	0.78	pass	+0.0361	pass
step 50,000	100.76	0.73	pass	+0.0053	pass
final (60,000)	100.49	0.46	pass	**−0.0089**	**fail**
best (S-selected)	87.71	12.32	**fail**	+0.2326	pass

What the ladder shows, and what it must not be used for. On the checkpoints sampled, the sign change of FS-0 falls between step 50,000 and step 60,000: the conjunction is met at the 50,000 rung and lost at the final one. *Those checkpoints do not resolve where in that interval the crossing happened, nor that the conjunction held continuously before it* — the ladder is a sample, not a trace. It is *not* licence to report the third rung: the reporting rule was fixed before any rung was graded, and it reads the final checkpoint. The *best* row is why such a rule is needed — the in-training score picks the checkpoint with the largest M1 deviation. *One limitation travels with the ladder*: between steps 58,400 and 58,600 the reference SCF changed branch, $\langle S^2 \rangle$ going from 0.957 to 0.999, which rearranges the fractional-spin profile.

J.4 Benchmark continuity

The reported model was graded before a correction to the M2/M3 stage of the benchmark and the longer-budget run after it. *Re-running the reported model's checkpoint on the corrected benchmark leaves every criterion verdict unchanged*; the per-row drift is in the verification report that ships with the archive.

K Two external construction tests; the second could not be read

Why this functional and not another. pcNN (Nagai et al., 2022) is the released neural family we could find whose published construction makes the longest explicit list of constraints an *architectural* claim rather than a training one. Its Table I lists ten conditions imposed analytically on the form, and the paper places the model alongside PBE and SCAN as functionals that satisfy their physical constraints analytically. Among the ten are the uniform electron gas limit, as X3 and C1, and the exact spin scaling relation, as X2. It was built by a different group, before this benchmark existed, and with no view to it. Skala-1.1 states construction-level guarantees too (Appendix G); the contrast is the length of the list.

What was fixed before the run. The benchmark was already sealed. We recorded, before attaching pcNN, the mapping from its Table I to our criteria and two directional predictions that follow from that mapping alone: M4 and M6 would pass, because exact spin scaling and the uniform-gas limit are two of the ten. *The first was the weaker inference, and we record why*: exact spin scaling constrains the *exchange* part, so reading it as a prediction for the full E_{xc} needs an invariance argument for the correlation part under the swap, which that Table I does not list. For M1, M2, M3, M5 and M7 we recorded *no prediction*, because its Table I names none of them — a refusal to predict, not a

prediction of failure. The registration, the adapter and its digest are in the archive; the adapter adapts the interface only and touches no criterion, threshold or validity gate.

What the benchmark could reach. Five readings came back, four of them on the nine conjoined rows and one on FS-0 printed beside them. The boundary is set by our tooling, not by pcNN. M1, M7, M7b, M4(b) and thermochemistry are `no`: each of those stages resolves its model through a loader this attachment cannot reach. M5 is `no` for a different reason: our sealed instrument's criterion quantity on that row is read off a three-coefficient network of our own architecture, which pcNN does not have. The condition is defined for it; *our sealed measurement of it* is not. M6 begins as the same case, and the note below the table says what a separate pipeline of ours reads there. The M4(a) row is the tool's density gate (L2); the paper's M4, the deployed energy with spins swapped on a fixed density (L3), reads 1.4×10^{-14} Ha over six swaps.

criterion	registered prediction	reading	gate
M4(a) density gate (L2)	**pass**	✓ 1.045×10^{-13}	$\leq 10^{-9}$
M6 uniform gas	**pass**	✗ $1.70\times^{\ddagger}$	$\leq 1\times$
M2 Koopmans	none recorded	✗ $0/3$	$3/3$
M3 $\mathrm{dev}(5)$ @H^-	none recorded	✗ 0.4691	≤ 0.05
FS-0 [mHa]	none recorded	✓ $+0.3978$	≥ 0
M1 · M5 · M7 · M7b · M4(b) · thermo	none recorded	*no*	—

The prediction we registered for M6 was wrong, and we report it as wrong. The separate pipeline that reads Skala-1.1's M6 cell reads pcNN too, so we mark it ‡ as Table 1 marks that one. It gives 0.340 mHa per electron, $1.70\times$ our budget. The scored M6 window samples $\zeta \in \{0, 1\}$ only; with $\zeta = 0.5$ added as an *auxiliary* sweep it is $0.0017\times$ the budget at $\zeta = 0$, $0.78\times$ at the auxiliary $\zeta = 0.5$ and $1.70\times$ at $\zeta = 1$. The unpolarised reading is small but it is not at the benchmark floor (Table G1, note (7)). Its own back end returns infinities on three of the ten window points; a second reads all ten and agrees on the other seven to 1.5×10^{-5} mHa per electron.

Where the deviation sits. The uniform-gas constraint the released code puts on the correlation factor is anchored at $(\bar{s}, \bar{\tau}) = (0, 0)$, which is its Table I row C1. That code normalises the kinetic-energy density by the *unpolarised* Thomas–Fermi value of the total density, so a spin-polarised uniform gas does not sit at $\bar{\tau} = 0$: with $h(\zeta) = \frac{1}{2}[(1+\zeta)^{5/3} + (1-\zeta)^{5/3}]$ it sits at $\bar{\tau} = \tanh(h(\zeta) - 1)$, which is 0, 0.139 and 0.528 at $\zeta = 0$, 0.5 and 1. The released code reproduces those to five digits at $r_s = 2$ (the residual is its own 10^{-7} density floor). The exchange half is not exposed: that code evaluates it through the spin-scaling relation, one channel at a time at $2n_\sigma$, where $\bar{\tau}$ vanishes identically, and Table I carries a spin-scaling row for exchange and none for correlation. Consistent with the split, the correlation factor carries 99.99% of the $1.70\times$: at $r_s = 2$, $\zeta = 1$ the two contributions are $+2 \times 10^{-5}$ and -0.340 mHa per electron. The constraint fixes that factor's value at the anchor and not its slope, which we read as 6.4×10^{-3} there, and $\bar{\tau}$ grows as ζ^2 near $\zeta = 0$, so the deviation climbs rather than switching on. **This is a mechanism consistent with the readings; whether it is an implementation error we have not determined.** *Intent is not the criterion here*: the deviation is a property of the released code's output, whether or not it was meant. We infer nothing about why the code is built this way, and we do not attribute the deviation to training.

Protocol residual, declared. pcNN must run in its own software environment, with library versions the other columns do not share; the vendor-anchor check leaves 50–83 μHa unattributed between two of them. *That is an energy scale for FS-0, not a common error bar*: M2–M4 are adjudicated under their own gates and M6 is reported per electron, and the FS-0 margin, $+0.3978$ mHa or 397.8 μHa, exceeds the 83 μHa upper end — *an observed scale, not a bound*. A margin under 0.1 mHa would have been reported *inconclusive*.

What it does and does not add. It adds one independent instance of Section 3's sentence about our own architecture, and one qualification of it. The instance: spin symmetry, written into the form for exchange and not listed for correlation, holds in the deployed energy. The qualification: the uniform-gas limit is written into the form too, yet on the scored window it stays inside the budget only at its C1 anchor, on unpolarised input ($0.0017\times$, not at the floor), and reads $1.70\times$ the budget on a fully polarised gas. Beyond the form, the two untargeted *criteria* the benchmark could read are not inherited automatically. A third untargeted quantity, FS-0, comes back non-negative, which Table H1 reads as "no anomalous sign", never "holds the condition", so it joins neither count. It is not a third audited column — five of its nine rows carry no reading — and no cell of Tables G1 or G2 depends on it.

A second construction test, which could not be read. Pokharel et al. (2022) release six models, three with analytic constraints; **none of their nine rows can be read**, because the models need the density Laplacian, which our tooling does not carry — a limit of our tooling, not of those models, from which no verdict follows.