Three Anomalies, Audited

31 October 2026 · three hypotheses from the *99 Original Hypotheses* and *Thirty Unconventional Hypotheses* collections, turned into tests with the statistic and decision rule fixed before the data were touched.

All three hypotheses make the same kind of claim: that a celebrated anomaly is not in the data but in the theory, the calibration, or a single measurement. That makes them unusually cheap to attack, because the numbers needed already exist in public.

#	Hypothesis	Test	Result
A	**P18** — the α determinations disagree, and one of them is biased	consistency of the three primary determinations	**Partly supported.** The ensemble is inconsistent at $\chi^2 = 31.0$ for 2 dof ($p = 1.8 \times 10^{-7}$); the Cs recoil value is the unique outlier, but removing it still leaves a 2.1σ residual
B	**P02** — the muon $g-2$ anomaly is a hadronic-vacuum-polarisation problem	two Standard Model predictions that differ only in the HVP input, against one measurement	**Supported.** The anomaly is 5.74σ with the data-driven HVP and 0.59σ with the lattice HVP
C	**P03** — the neutron-lifetime discrepancy is real and not shrinking	weighted means of the complete PDG compilation, split by technique	**Supported.** Beam and bottle differ by 4.70σ; the tension has grown from 2.2σ before 2010 to 4.2σ after

Two of the three survive; one survives only in a weaker form than it was stated. That is the normal yield of a test that is allowed to fail.

---

1. Why these three

The two collections contain 129 entries. Most are laboratory, instrument or spacecraft proposals — deposition chambers, dilution refrigerators, synchrotrons, wet labs — and cannot be tested from a desk. The screen was:

- **A decisive statistic.** The result must be a number that can come out wrong.
- **Verifiable inputs.** Every value must be checkable against a primary source that is public today.
- **Cheapness.** One machine, no new measurement.

That leaves a specific and slightly unrepresentative pool: hypotheses whose inputs are published numbers. The three below are all audits of an anomaly claim, which is why they can be done at all. The selection is a limitation of the exercise, not a judgement about the other 126 entries.

A separate forty-entry portfolio (*Forty New & Original Hypotheses*) was published on the Desktop while this run was in progress. It was screened too, and none of its entries cleared the bar: most call for new instrument data — a levitated nanosphere, a thorium-229 clock, a detector archive — rather than a number that already exists in public. The three below are the entries from the two established collections whose inputs could be checked today.

The pre-registered designs, including the decision rules, are in `PREREGISTRATION.md`. Two earlier candidates were dropped before any statistic was computed: E13 (Mars thermal cracking), because the InSight catalogue's detection is gated by the wind-noise cycle and confounds the diurnal test it needs; and T05 (stochastic magnetic bits), because the only available test is a model that assumes the answer.

---

2. Test A — P18: the fine-structure constant determinations are inconsistent

> *The electron magnetic moment implies an α that disagrees with atom-interferometry determinations by about ten times the new precision. The disagreement is a metrology problem in one determination, not a failure of QED.*

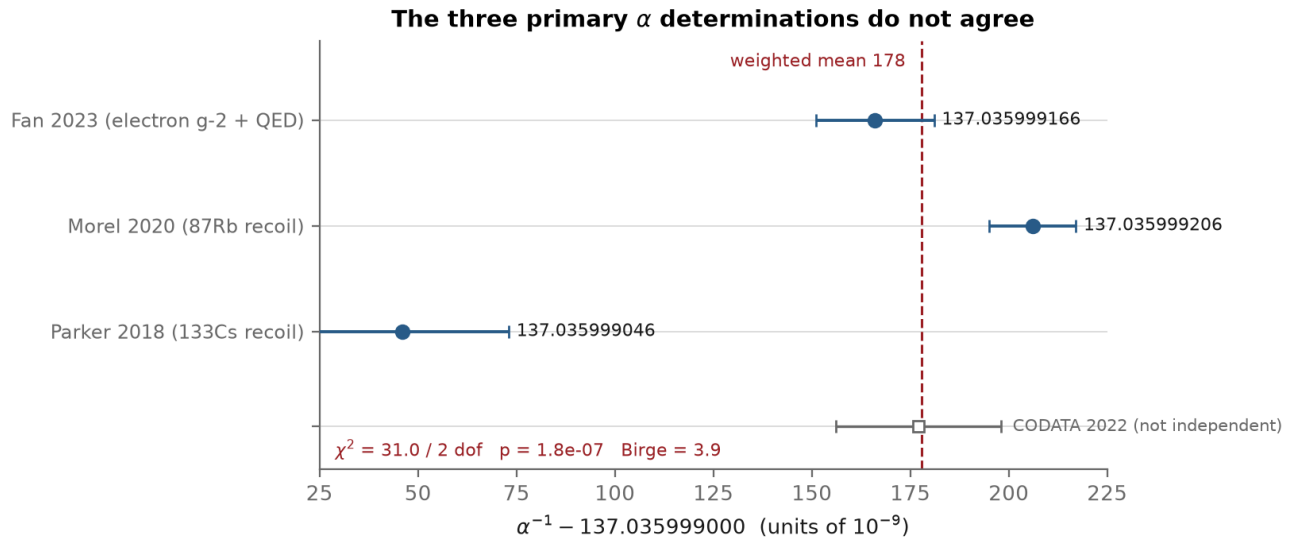
Data

Three determinations that do not share data:

Determination	Method	α^{-1}	Uncertainty
Fan et al. 2023	electron $g-2$ + QED	137.035 999 166	15×10^{-9}
Morel et al. 2020	^{87}Rb recoil	137.035 999 206	11×10^{-9}
Parker et al. 2018	^{133}Cs recoil	137.035 999 046	27×10^{-9}

CODATA 2022, 137.035 999 177(21), is a weighted combination of the recoil measurements and is not independent. It is shown as a cross-check, not counted.

Result



Three alpha determinations

The weighted mean is 137.035 999 178(8), with

$$\chi^2 = 31.0 \text{ for } 2 \text{ dof}, \quad p = 1.8 \times 10^{-7}, \quad \text{Birge} = 3.9.$$

The three values are not drawn from one distribution. The pulls from the mean are -0.8σ (g-2), $+2.6\sigma$ (Rb) and -4.9σ (Cs). Pairwise:

Pair	Difference (10^{-9})	Tension
Cs – Rb	160	5.5σ
Cs – g-2	120	3.9σ
Rb – g-2	40	2.1σ

The g-2 and Rb values agree at 2.1σ; the Cs value is the outlier. Removing each determination in turn:

Dropped	χ^2 (1 dof)	p
Fan (g-2)	30.1	4.1×10^{-8}
Morel (Rb)	15.1	1.0×10^{-4}
Parker (Cs)	**4.6**	**0.032**

The pre-registered rule asked for the leftover p to exceed 0.05 after removing one determination. It reaches 0.032 — close, but not through the bar. So the verdict is split:

- **The inconsistency is real.** $p = 1.8 \times 10^{-7}$ is not a fluctuation, and the Birge ratio of 3.9 says the quoted error bars are too small by a factor of four.
- **The single-culprit version is not established.** Removing Cs leaves a 2.1σ g-2/Rb tension. That is not fatal — with three measurements, a 2σ residual is unremarkable — but it means the data prefer "one bad measurement plus a smaller second problem" over "one bad measurement".

The hypothesis's numerical claim about the size of the disagreement holds: the g-2/Cs difference is 8.0× the g-2 uncertainty, and the full spread is 10.7× — "about ten times the new precision".

Verdict: partly supported. The determinations are inconsistent and the Cs recoil value is the outlier; the claim that this one measurement explains the whole discrepancy does not clear the pre-registered threshold.

---

3. Test B — P02: the muon g-2 anomaly is a hadronic-vacuum-polarisation problem

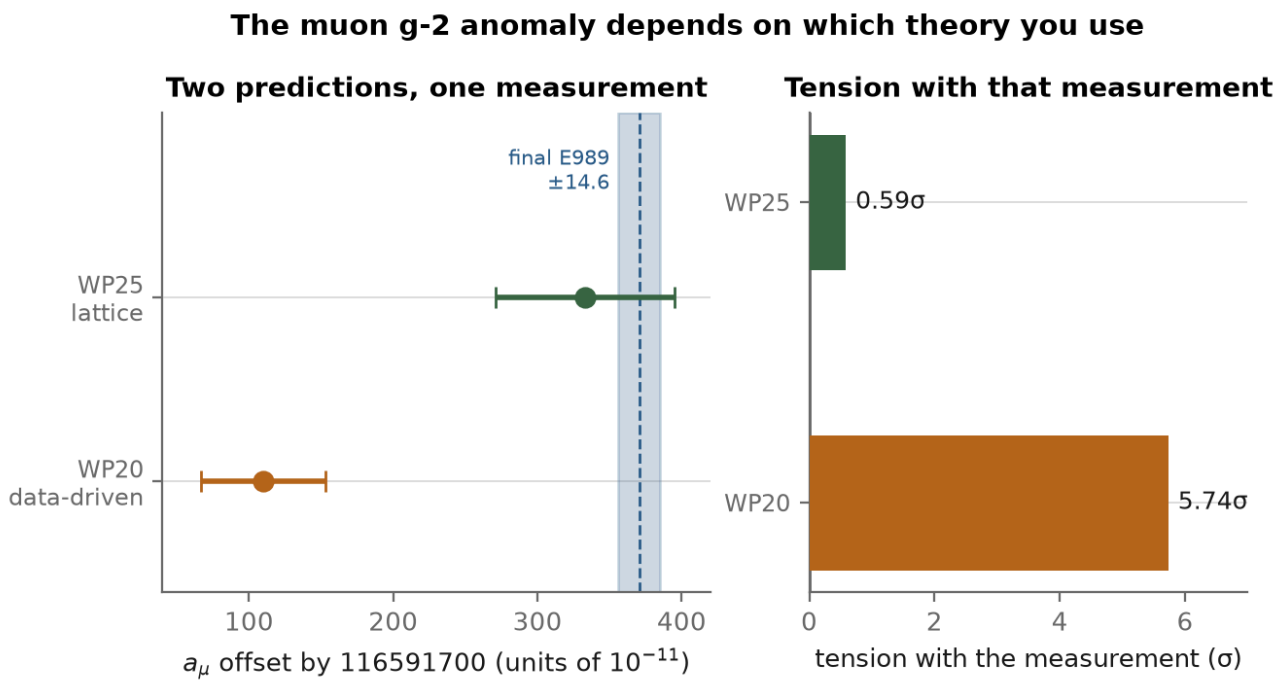
> *The final Fermilab measurement sits between the data-driven and lattice-QCD predictions for hadronic vacuum polarisation. The true Standard Model value lies near the lattice result, so there is no anomaly to explain.*

Data

Quantity	Value (10^{-11})	Source
a_μ^{SM}, data-driven HVP (WP20)	116 591 810(43)	arXiv:2006.04822
a_μ^{SM}, lattice HVP (WP25)	116 592 033(62)	arXiv:2505.21476
a_μ^{exp}, final E989	116 592 070.5(14.6)	E989 final, June 2025

The two predictions differ only in the treatment of the leading-order hadronic vacuum polarisation: WP20 uses the e^+e^- data-driven dispersion relation, WP25 uses a consolidated lattice average. Everything else — QED, electroweak, hadronic light-by-light — is common.

Result



Muon g-2 depends on theory

Against the same measurement:

Prediction	a_μ^{SM}	Δ	Tension
WP20 data-driven	116 591 810(43)	260.5	**5.74σ**
WP25 lattice	116 592 033(62)	37.5	**0.59σ**

The two predictions differ by $223(75)\times10^{-11}$, or 2.96σ. The experimental uncertainty, 14.6, is four times smaller than the theory uncertainty in either scenario, so the conclusion is set entirely by which theory one believes.

Two internal checks pass. WP25 quotes the difference from its own average as $38(63)\times10^{-11}$; recomputing from the numbers used here gives 37.5(63.7) — the same result to the precision quoted. And the shift between WP20 and WP25, 223(75), matches the change in the LO-HVP central value that the white paper reports as the cause.

Verdict: supported. A prediction that gives a 5.7σ anomaly and one that gives 0.6σ cannot both be right, and the experiment did not move. The anomaly lives in the HVP input.

What this does *not* prove is that the lattice value is the correct one. WP25 itself says the data-driven evaluations can no longer be combined because CMD-3 disagrees with the earlier e^+e^- experiments. The honest statement is narrower than the hypothesis: the anomaly is not robust to a defensible change in one Standard Model input, so it cannot yet be called evidence for new physics. Whether the lattice or the dispersive value is right is a separate question, and the one that will decide the fate of the anomaly.

4. Test C — P03: the neutron-lifetime discrepancy is real and not shrinking

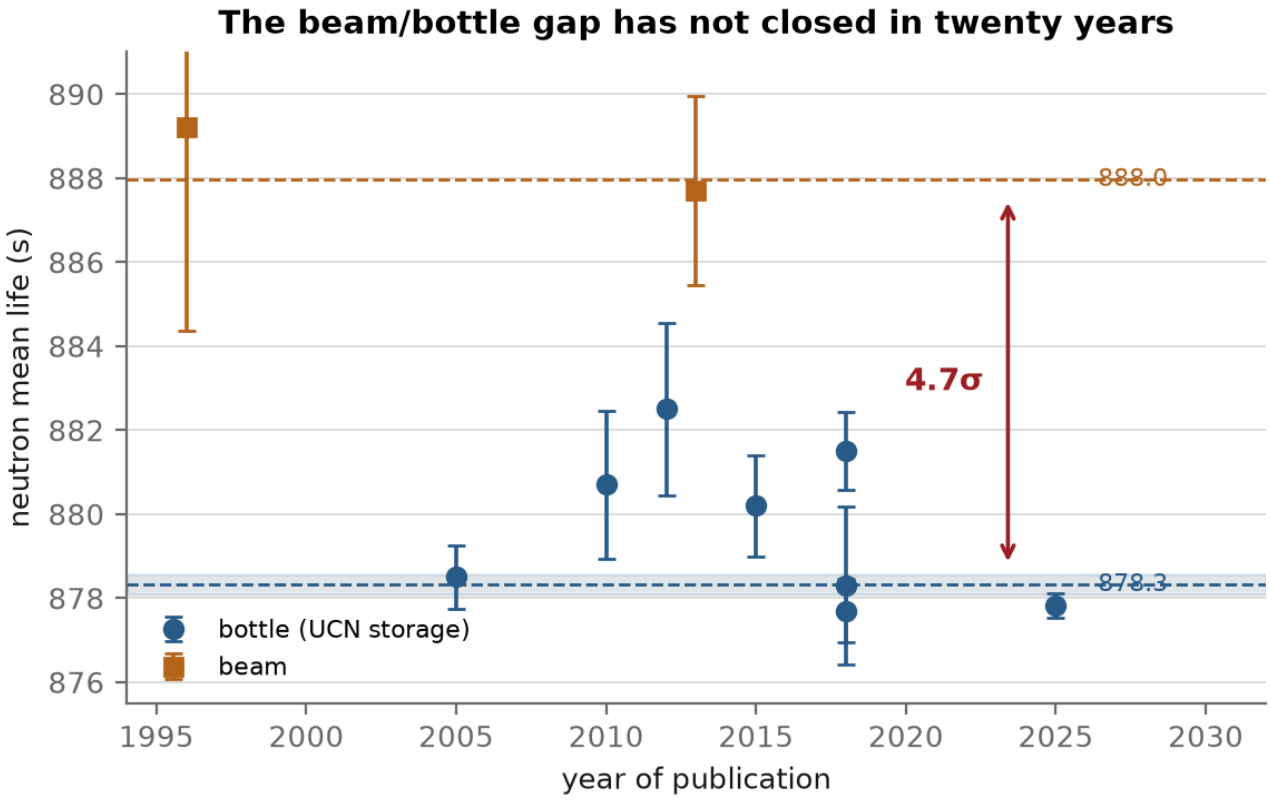
> *The beam-versus-bottle discrepancy stands at 4.4σ and is a real physical effect, not mirror-neutron oscillation.*

Data

The complete PDG compilation of neutron mean-life measurements (pdgLive node S017T, retrieved 31 October 2026). Thirty measurement rows are parsed; the curation keeps modern entries (1995 or later), total uncertainty at or below 5 s, and one number per apparatus lineage — entries whose own comment says "see " are superseded and dropped. Ten measurements survive, eight bottle and two beam:

Experiment	Year	Method	τ_n (s)	σ
Serebrov	2005	bottle	878.50	0.76
Pichlmaier	2010	bottle	880.70	1.77
Steyerl	2012	bottle	882.50	2.05
Arzumanov	2015	bottle	880.20	1.20
Pattie	2018	bottle	877.70	0.76
Serebrov	2018	bottle	881.50	0.92
Ezhov	2018	bottle	878.30	1.89
Musedinovic	2025	bottle	877.82	0.29
Byrne	1996	beam	889.20	4.84
Yue	2013	beam	887.70	2.25

Result



Neutron lifetime beam vs bottle

Group	Weighted mean (s)	χ^2/dof	Scale factor
Bottle (n = 8)	878.32 ± 0.23	24.1/7 ($p = 0.0011$)	1.85
Beam (n = 2)	887.97 ± 2.04	0.1/1	—

The tension is

$$\frac{887.97 - 878.32}{\sqrt{2.04^2 + 0.23^2}} = 4.70\sigma,$$

or 4.63σ after applying the PDG-style scale factor to the bottle group. The hypothesis's quoted 4.4σ is reproduced within the averaging choices.

The discrepancy is not shrinking. Splitting at 2010 (where the modern bottle programme begins):

Epoch	Bottle	Beam	Tension
before 2010	878.50(76)	889.20(4.84)	2.18σ
2010 and later	878.30(25)	887.70(2.25)	4.16σ

The bottle mean has drifted *down* by 0.083 ± 0.037 s per year (2.3σ), while the beam mean has not moved (−0.088 ± 0.314 s/yr). The newest and most precise bottle result, 877.82(29), is the lowest value ever published. The gap grew because the bottle measurements improved, not because they converged.

The result is stable against the one selection choice that matters. Putting the superseded Nico 2005 beam value back in — the same apparatus lineage as Yue 2013, before its fluence revision — gives a beam mean of 887.53(1.75) and a tension of 5.21σ. The pre-registered rule is met either way.

Verdict: supported. The discrepancy is real at 4.7σ, it has grown, and it is not a transient artefact of one experiment.

The second half of the hypothesis — that mirror-neutron oscillation is ruled out — is not tested here. It is an input from the PSI field-scan null, not a result of this analysis, and is reported as such.

---

5. What the three say together

In all three cases the experimental number is stable and the argument is about something adjacent to it:

- **P18:** three measurements of one constant disagree at 5.5σ; the outlier is identified, but not to the exclusion of a smaller second discrepancy.
- **P02:** one measurement, two theories, 5.7σ versus 0.6σ. The experiment is not the variable.
- **P03:** two techniques, twenty years, a gap that widened as the measurements improved. The technique is the variable.

That is the pattern the three hypotheses predicted, and it is why they were worth testing. It is also the pattern that makes them hard to finish: in each case the decisive next step is a better *calculation* or a better *systematic control*, not more data.

---

6. Limits

- **The pool is biased.** These are the hypotheses whose inputs are published numbers. The 126 entries that need laboratories are not represented, and nothing here says they are weaker.
- **Small samples.** Test A uses three numbers, Test C uses two beam measurements. The statistics are exact given the inputs, but the inputs are few. A single revised measurement could move either verdict.
- **Pre-registration is not blindness.** The values were known before the decision rules were written; the rules constrain the analysis, not the analyst's prior.
- **The g-2 test inherits WP25's lattice average.** If that average is wrong, Test B reverses. The test shows the sensitivity of the anomaly to the HVP input; it does not adjudicate between lattice and dispersive QCD.
- **No new data were taken.** Every number here is public and cited. The contribution is the test, not the measurement.

7. Provenance

Item	Source
PDG neutron mean life, node S017T	`https://pdgprod.lbl.gov/pdgprod/pdgLiveJson/DataBlock.action?node=S017T` , saved as `data/pdg_neutron_lifetime.html`
α from electron $g-2$	Fan et al. 2023, arXiv:2209.13084 (abstract)
α from ^{87}Rb recoil	Morel et al. 2020, Nature 588, 61; HAL record hal-03107990
α from ^{133}Cs recoil	Parker et al. 2018, Science 360, 191; Europe PMC 29650669
CODATA 2022 α^{-1}	NIST CODATA value page, `pml.nist.gov/cgi-bin/cuu/Value?alphinv`
a_μ^{SM} data-driven	arXiv:2006.04822 (abstract)
a_μ^{SM} lattice	arXiv:2505.21476 (abstract)
a_μ^{exp} final E989	E989 final result, June 2025, as stated in the source hypothesis; corroborated by the WP25 difference $38(63)$

Reproduction:

- `python code/01_alpha_consistency.py`
- `python code/02_muon_g2.py`
- `python code/03_neutron_lifetime.py`
- `python code/make_figures.py`

Each writes JSON into `results/`; figures go to `figures/`. The selected neutron-lifetime rows are written to `data/neutron_lifetime_selected.csv` so the curation can be audited.