A refuted-and-vindicated pre-registration test of a spectral error model on a superconducting processor

Dr. Tamás Nagy

tnagyphd@gmail.com

Preprint — Stage 5-6 extension in hand (v2_fwd jobs 749550-749554, 2026-04-19); L4 falsified, L5' partial pass (3/4 pairs), Q3-Q6 structurally irreducible

Abstract

We pre-register and test a spectral-error-mitigation prediction for the two-qubit gate fidelity of Quantum Inspire's Tuna-9 9-qubit transmon processor and execute it in four cryptographically timestamped stages. Stage 1 locks a SHA-256 hash of a three-model prediction family — spectral, T_1-only, and gate-error-only — at depths $m \in \{0, 2, 4, 8, 16, 32, 64\}$ of H-sandwiched CZ chains on four disjoint pairs. Stage 2 measures $m = 64$ and observes a $0.273 \to 0.682$ *fidelity revival* on the cleanest pair (Q0-Q1), falsifying every monotone stochastic decay at 8-10σ in the log-domain. Stage 3 refines the model by adding a single coherent parameter, δ (per-CZ Z-rotation phase), derives an exact even-m degeneracy $\chi^2(\delta) = \chi^2(\pi - \delta)$ on the training set, and pre-registers three hypotheses (naive, slow-coherent $\delta \in (0, \pi/2)$, fast-coherent $\delta \in (\pi/2, \pi)$) with forward predictions at the parity-sandwich $m \in \{63, 65\}$; the cryptographic stamp is completed *before* the submission of jobs 743595 and 743596. Stage 4 executes the forward test: the observed fidelities at $m = 63$ and $m = 65$ select the **fast branch** on both discriminative pairs (Q0-Q1 and Q4-Q7) and refute the slow branch at 17.6σ on a single point (Q4-Q7 at $m = 65$: predicted 0.939 vs observed 0.326). The refined coherent-error model is vindicated with a pre-registered maximum σ-gap of 1.24 across all eight forward-test data points. The physical interpretation is that the CZ gate on Tuna-9 implements an operation close to $\mathrm{CZ} \cdot Z$ — a systematic π-over-rotation with residual $\pi - \delta \approx 0.12$-$0.33$ rad per application. A follow-up second prospective test across a fresh calibration boundary (Stage 5–6, 2026-04-19; v2_fwd, SHA-256 983503bc, OTS 4-of-4) falsifies the single-δ refined model at aggregate $\chi^2/\mathrm{ndof} = 59.6$ and promotes it to an L5 in which per-CZ coherent $R_X(\eta)$ drive is free on every pair; under an epoch-split drift budget the generalised model (L5′) passes three of four pairs on the forward set ($\chi^2/5 \leq 1.74$), with Q3-Q6 revealed as *structurally* irreducible under any $R_X \otimes R_X$ ansatz (global post-hoc best $\chi^2/5 = 9.35$ from a multi-start Nelder-Mead search, still above the ≤ 4 cutoff) — a pre-registered, cryptographically stamped negative result that sharpens both the refined model and the next pre-registration target. The experiment demonstrates that a single pre-registered-and-then-refined benchmark on a fully-public superconducting processor can produce a *decisive* test of a structured-noise error model at a cost of $\sim$ 36,000 shots total, a turnaround of under a day, and zero gatekeepers.

---

1. Introduction

Quantum computers are now widely accessible, but their error models remain in a regime where prospective testing against pre-committed quantitative predictions is rare. Most published device characterisations either (i) report *mitigated* or post-selected fidelities that are several factors above the raw gate-level prediction of any published error model, or (ii) fit free parameters *after* seeing the

measurement, making it impossible to distinguish between a model with genuine predictive content and one adjusted to match. For a scientific error theory of noisy intermediate-scale quantum (NISQ) processors to advance, we need to *pre-register* a quantitative prediction, run a *raw* measurement (no mitigation, no post-selection), and accept the result even when it falsifies the prediction.

This letter reports exactly such a test, executed in three cryptographically dated stages on Quantum Inspire's Tuna-9 processor [9]. The device is a fixed-frequency 9-qubit superconducting transmon with 12 CZ-enabled edges, native H, and a single-qubit + CZ gate set. We chose it for three reasons: (i) it belongs to the technologically dominant superconducting-transmon class, so a clean test on Tuna-9 is informative about the platform family that drives IBM, Google, and Rigetti production workloads; (ii) it is public, free-tier, and accessible from a laptop with no institutional agreement; (iii) it offers $\gtrsim 20\,\mu s$ coherence at $\sim 6\%$ per-CZ gate error, a regime where the two-qubit gate contribution dominates decoherence and the *spectral* structure of the noise becomes the observable.

The model we test is a spectral error-budget formula $F(m) = \exp(-r_{\text{eff}}\, m)$ with $r_{\text{eff}}^{-1} = B_Q^{-1} + \varepsilon_{2q}$, where $B_Q = v_{LR}/(n\,|\lambda_1|)$ combines the Lieb-Robinson velocity v_{LR} [1-3] and the Lindblad spectral gap $|\lambda_1|$ [8,12]. The prediction is entirely determined by independently measured hardware parameters $(T_1, T_2^*, \varepsilon_{CZ})$ and contains *no* free parameter; every input SHA-256 hash is embedded in the pre-registration file.

The contribution of this letter is operational, not theoretical. We show that a single public processor, a four-hour measurement campaign, and a cryptographic timestamp are sufficient to:

1. **Falsify** a pre-registered structured prediction with a single depth point ($m = 64$), at 9.8σ in the log-domain on the cleanest pair.
2. **Refine** the model in a principled way — adding one coherent-error parameter — and derive a *non-trivial* mathematical degeneracy ($\delta \leftrightarrow \pi - \delta$) that the refined model cannot resolve on any even-m training set.
3. **Break** the degeneracy with a single additional parity-sandwich measurement at $m \in \{63, 65\}$, pre-registered *before* the jobs are submitted, and containing a precise falsification criterion for the refined model class itself.

The full refuted-and-vindicated arc from v1 submission to v2 parity-sandwich submission took under eight hours of human time and $\sim$ 25,000 shots of device time. We argue this pattern — pre-register, falsify, refine, re-pre-register, resolve — is a template for *scientific* quantum-hardware validation that is disciplined without being heroic.

A natural next question is whether the Stage 3–4 refined model transfers *across* a calibration boundary and across the full four-pair CZ set of the device. To test this we ran a second prospective test (v2_fwd, §8.1) at five off-grid depths on all four pairs one day later, committed to a fresh SHA-256 + OpenTimestamps stamp (4-of-4) before submission. The contract failed at aggregate $\chi^2/\text{ndof} = 59.6$, forcing a structural generalisation (L5) in which a CZ-gated coherent $R_X(\eta)$ drive is free on every pair, not only on the one pair where the Stage-3 deepdive had inferred it. Under an epoch-split drift budget, the generalised model passes three of four pairs on the forward set at the pre-registered cutoff but fails on Q3-Q6, which we show (§8.7) cannot be fitted to cutoff by any parameter setting of the $R_X \otimes R_X$ ansatz at all. We report this as a negative result in §8 — Stage 5–6 of the arc — and use it to extract a general methodological point (§8.8) that is transferable beyond the specific device: a null-effect result in one channel does not license a structural-zero constraint in a different channel the first experiment could not probe.

2. Protocol

2.1 Calibration battery (provenance inputs)

Three self-measured calibration batches on Tuna-9 provide every number that enters the prediction file. No vendor-published $T_1/T_2^*/\varepsilon_{2q}$ tables are used.

- T_1 measurement (9 qubits in parallel; X, wait, measure; 6 delays $\{0, 10, 20, 50, 100, 150\}\,\mu s$; $N = 2048$ shots; jobs 743402–743407). Median $T_1 = 20.5\,\mu s$; mean $21.1\,\mu s$; benchmark qubit Q3 $= 26.95\,\mu s$ ($R^2 = 1.000$). Raw: tuna9_t1_run1.json (SHA-256 is embedded in the prediction file).
- T_2^* **Ramsey** (9 qubits in parallel; X_{90}, wait, X_{-90}, measure; initial grid $\{0, 1, 2, 5, 10\}\,\mu s$ then a refined $\{0, 0.2, 0.5, 1, 2, 5\}\,\mu s$ grid for the fastest-dephasing qubits; $N = 2048$; jobs 743452–743457 and 743549–743554). Combined fit: 8 of 9 qubits have reliable T_2^* in 3-9 μs; Q1 flagged for readout bias. Raw: tuna9_t2_combined.json.
- **CZ gate error** via H-sandwiched chain on four disjoint pairs (Q0-Q1, Q2-Q5, Q3-Q6, Q4-Q7); $m \in \{0, 2, 4, 8, 16, 32\}$ (batch 1) and $m = 64$ (batch 2, for falsification test); $N = 4096$ shots; jobs 743464–743469, 743539. Per-pair ε_{CZ} ranges from 2.5% (Q0-Q1) to 9.8% (Q3-Q6), median 6.4% per layer. Raw: tuna9_cz_run1.json, tuna9_cz_run2.json.

The calibration pipeline is ours; every analysis script (fit routine, error propagation, plotting) is in the public repository. The circuits generated at run-time are cQASM 3.0 files that we archive per job.

2.2 H-sandwiched CZ chain

For each of four disjoint pairs, a depth-m chain is the circuit

$$H_a \otimes H_b \cdot (CZ_{ab})^m \cdot H_a \otimes H_b,$$

prepared on $|00\rangle$. For a perfect CZ, $F(m) = P(|00\rangle)_{\text{measured}} = 1$ at all m. Depolarising noise sends $F(m)$ monotonically to 1/4, and any T_1/T_2 process that ignores the CZ itself decays $F(m)$ smoothly. A CZ over-rotation by phase δ per layer produces a *parity-dependent* signature through the even/odd cosine factor, derived in §4. The four pairs are driven simultaneously on disjoint edges so each job returns four independent $F(m)$ measurements.

2.3 Pre-registration workflow

Every prediction file in this letter follows the same recipe:

1. Write the prediction table + falsification criteria + input calibration SHA-256 into a plain-text file.
2. Compute SHA-256 of that file and save to an adjacent .sha256 file.
3. Submit the file to four independent OpenTimestamps calendars: alice .btc.calendar.opentimestamps.org, bob.btc.calendar.opentimestamps.org, finney.calendar.eternitywall.com, btc.calendar.catallaxy.com. The returned .ots proof anchors the hash in the Bitcoin blockchain within one confirmation cycle.
4. Submit the QI job *after* the four-of-four stamp returns.

Any post-hoc edit of the prediction file, any swap of the calibration inputs, any rewrite of the falsification criteria would invalidate the hash and break the stamp. The calibration inputs are

themselves hashed and embedded verbatim in the file, so a modified tuna9_t1_run1.json breaks the prediction file's SHA-256 too. Chain of custody is transitive.

3. v1: pre-registration, submission, falsification

3.1 v1 predictions (SHA-256 `787b3c1d…d1f6`, stamped 2026-04-18)

The v1 prediction file [predictions_tuna9_0418.txt] commits a three-model family at seven depths $m \in \{0, 2, 4, 8, 16, 32, 64\}$. Median hardware parameters (over the four measured pairs) were $\bar{\varepsilon}_{CZ} = 6.42\%$, $\bar{T}_1 = 21.07\,\mu s$, $n = 2$ per pair (H-sandwich on one edge at a time). The three models are

$$F_{\text{spec}}(m) = \exp(-r_{\text{spec}}\, m), \quad r_{\text{spec}} = \varepsilon_{CZ} + n\, t_{\text{gate}}/(T_1\, C_{\text{gate}}),$$

$F_{T_1\text{-only}}(m) = \exp(-n\, m\, t_{\text{gate}}/(T_1\, C_{\text{gate}}))$, and $F_{\varepsilon\text{-only}}(m) = \exp(-\varepsilon_{CZ}\, m)$. The pre-registered log-domain uncertainty $\sigma_{\log} = 0.196$ combines shot noise ($N = 4096$) and per-pair parameter spread.

Predictions at the forward-test depth $m = 64$:

$$F_{\text{spec}}(64) = 0.00776, \quad F_{T_1\text{-only}}(64) = 0.47295, \quad F_{\varepsilon\text{-only}}(64) = 0.01640.$$

3.2 Observed $m = 64$ fidelities (job 743539, $N = 4096$)

$F(Q0\text{-}Q1, 64) = 0.6819$, $F(Q2\text{-}Q5, 64) = 0.1018$, $F(Q3\text{-}Q6, 64) = 0.2197$, $F(Q4\text{-}Q7, 64) = 0.3555$.

From the $m = 32$ batch the corresponding fidelities were 0.272, 0.302, 0.110, 0.550. Two of four pairs show a *revival* as the depth doubles (Q0-Q1, Q3-Q6), and two show further decay (Q2-Q5, Q4-Q7). Fidelity increasing with depth is incompatible with any monotone stochastic decay.

3.3 σ-gap verdict

Comparing the observed Q0-Q1 fidelity (the cleanest pair, smallest ε_{CZ}) against each v1 model in the log-domain σ metric:

v1 model	$F_{\text{pred}}(64)$	$\lvert\Delta \log_{10} F\rvert$	σ-gap
Spectral	0.00776	1.94	$9.8\,\sigma$
ε-only	0.01640	1.62	$8.2\,\sigma$
T_1-only	0.47295	0.16	$0.8\,\sigma$

The spectral and ε-only models are decisively falsified. The T_1-only model is numerically close at $m = 64$ but fails the *curvature* of the full run: a joint fit of $\{A, p, C\}$ to the seven points $\{0, 2, 4, 8, 16, 32, 64\}$ yields $\chi^2 = 2717.9/4$, driven by the inability of a smooth exponential to match $F(32) = 0.272$ and $F(64) = 0.682$ simultaneously. Every member of the v1 stochastic family is therefore rejected.

3.4 What falsification says

The v1 pre-registration was *exactly* what one wants from a prospective test: it was locked (SHA-256 + OTS) before the measurement, the measurement was a single job, and the model was rejected without ambiguity. The scientific content of this paragraph is not "we guessed right"; it is "the process worked". The v1 error model was wrong; the signature (revival at even depth) reveals the physical mechanism we missed (coherent CZ phase accumulation).

4. Refined coherent-error model

4.1 Derivation

Consider a CZ gate on a disjoint pair (a, b), implemented as CZ followed by an unknown Z-rotation of angle $\delta/2$ on each qubit (symmetric miscalibration), plus a per-layer depolarising channel with rate p. The H-sandwiched m-fold CZ is

$$U(m) = H_a H_b \cdot (CZ \cdot e^{-i\delta/2\,(Z_a+Z_b)})^m \cdot H_a H_b.$$

Using CZ $\cdot Z_a Z_b$ eigenvalues and H-conjugation, a direct calculation gives the noise-free survival probability on $|00\rangle$:

$$P_{\text{coh}}(m, \delta) = \frac{10 + 6\,(-1)^m\,\cos(m\,\delta)}{16}.$$

Composed with a stochastic depolarising channel of rate p per CZ and asymptotic fidelity $1/4$, the two-parameter forward model for the H-sandwiched pair is

$$\boxed{F(m) = (1-p)^m\,P_{\text{coh}}(m, \delta) + (1-(1-p)^m) \cdot \tfrac{1}{4}.}$$

Two parameters per pair: stochastic depolarisation rate $p \in [0, 0.3]$, and coherent per-CZ phase $\delta \in [0, \pi]$.

4.2 Fit to the combined $\{m = 0..64\}$ training set

On the seven-point training set of run1+run2 data, the fits per pair (shot-noise weighted) are

Pair	p (slow)	δ_{slow} (rad)	RMS (refined)	RMS (naive exp)
Q0-Q1	0.00293	0.1169 (6.7°/CZ, period 53.8 CZ)	0.041	0.168
Q2-Q5	0.129	0.2022 (11.6°/CZ, period 31.1 CZ)	0.108	0.106
Q3-Q6	0.139	0.3272 (18.75°/CZ, period 19.2 CZ)	0.081	0.086
Q4-Q7	0.000	0.1361 (7.8°/CZ, period 46.2 CZ)	0.033	0.080

On the two discriminative pairs (Q0-Q1 and Q4-Q7) the refined model reduces the residual RMS by a factor of four vs the naive three-parameter exponential. Q2-Q5 and Q3-Q6 are already at the depolarising floor by $m = 32$ and have limited coherent signature.

4.3 The π-alias degeneracy

The training set $m \in \{0, 2, 4, 8, 16, 32, 64\}$ contains only *even* integers. For even m,

$$\cos(m\,(\pi - \delta)) = \cos(m\pi - m\delta) = \cos(m\delta),$$

so $P_{\text{coh}}(m, \delta) = P_{\text{coh}}(m, \pi - \delta)$ identically. Hence the χ^2 landscape has two *exactly* degenerate global minima at δ and $\pi - \delta$. The best fits above yield a *second* per-pair solution:

Pair	δ_{fast} (rad)	Physical period
Q0-Q1	3.0247 (173.3°/CZ)	2.08 CZ
Q2-Q5	2.9394 (168.4°/CZ)	2.14 CZ
Q3-Q6	2.8144 (161.3°/CZ)	2.23 CZ
Q4-Q7	3.0055 (172.2°/CZ)	2.09 CZ

These are *physically distinct mechanisms*. The slow branch is a weak ZZ-crosstalk-type over-rotation that accumulates over tens of CZs; the fast branch is a near-π systematic Z-rotation per CZ — consistent with a CZ pulse that is effectively CZ$\cdot Z^{\sim 1}$ (systematic miscalibration). The training data cannot tell them apart; an odd-m measurement can.

5. v2: parity-sandwich pre-registration

5.1 v2 predictions (SHA-256 `7cc18b28…26d8`, stamped 2026-04-18 18:07Z)

The v2 prediction file [predictions_tuna9_v2_0418.txt] commits three hypotheses at two depths:

- H_{naive}: three-parameter exponential $F = Ap^m + C$ (no coherent structure).
- H_{slow}: refined model with $\delta \in (0, \pi/2)$.
- H_{fast}: refined model with $\delta \in (\pi/2, \pi)$.

Forward-predicted fidelities at $m = 63$ on the two discriminative pairs:

Pair	H_{naive} $F(63) \pm \sigma$	H_{slow} $F(63) \pm \sigma$	H_{fast} $F(63) \pm \sigma$
Q0-Q1	0.509 ± 0.168	0.415 ± 0.042	0.709 ± 0.042
Q4-Q7	0.360 ± 0.080	0.872 ± 0.034	0.378 ± 0.034

The H_{slow} / H_{fast} separation at Q4-Q7 is 10.3σ; at Q0-Q1 it is 5.0σ.

5.2 Falsification criteria (locked in §15 of the v2 file)

The refined model class is *confirmed* if

$$F(63, Q4\text{-}Q7) \in [0.84, 0.91] \text{ (slow)} \quad \text{or} \quad [0.34, 0.41] \text{ (fast)},$$

$$F(63, Q0\text{-}Q1) \in [0.37, 0.46] \text{ (slow)} \quad \text{or} \quad [0.67, 0.75] \text{ (fast)},$$

and the two pairs select the *same* branch. Every other outcome — a fidelity in the gap regions $[0.50, 0.80]$ (Q4-Q7) or $[0.50, 0.65]$ (Q0-Q1), or different branches selected by the two pairs — *refutes* the refined model class and forces a v3.

5.3 Submission

Jobs 743595 ($m = 63$) and 743596 ($m = 65$) were submitted to Tuna-9 at 2026-04-18 18:08:08 UTC against the OTS-stamped v2 file. The submission manifest [tuna9_cz_parity_sandwich_submitted.json] records both job IDs, the cQASM SHA-256 per circuit, and the cross-link to the v2 prediction file's hash.

6. Parity-sandwich result

Both forward-prediction jobs completed on Tuna-9 within ~30 s of execution once the scheduler picked them up (743595 at $m = 63$: 27.0 s; 743596 at $m = 65$: 29.9 s). Raw outcome counts are in examples/qi_circuits/results/tuna9_cz_parity_sandwich.json (SHA-256 869259941eae66f73ab8298097a4a32d6c58e238876ed02c142608abe7c7f56e). The analysis hash, with all three hypothesis-level σ-gaps per pair per m, is comparison_parity_0418.json (SHA-256 d13f0a9292e96a19ae4782600f8f639b46a8e881a3b522219f211bf14f8bd613), OTS-stamped across the same four calendars at 2026-04-18 18:23 UTC.

6.1 Observed fidelities

Table 2 lists the two-qubit marginal fidelity $F(m) = P(|00\rangle_{\text{pair}})$ for each of the four pairs at the two forward-predicted depths, $N = 4096$ shots per job.

Pair	$F(m = 63)$	$F(m = 65)$
Q0-Q1	0.6726	0.5947
Q2-Q5	0.1624	0.2073
Q3-Q6	0.1729	0.1492
Q4-Q7	0.3472	0.3257

Q0-Q1 and Q4-Q7 remain well above the depolarising floor and retain discriminative power; Q2-Q5 and Q3-Q6 drifted $\sim 1\sigma$ below the predicted floor (0.15-0.21 vs 0.25), which is within shot noise plus model systematic and — as pre-registered in Part 14 of the v2 file — has no discriminative power between the three hypotheses.

6.2 Per-hypothesis σ-gap table

Computing the log-domain σ-gap $|F_{\text{obs}} - F_{\text{pred}}|/\sigma_{\text{total}}$ with $\sigma^2_{\text{total}} = \sigma^2_{\text{fit,RMS}} + \sigma^2_{\text{shot}}$ ($\sigma_{\text{shot}} = 0.008$ at $N = 4096$) for every (pair, m, hypothesis) combination:

m	Pair	Observed	H_naive	H_slow	H_fast
63	Q0-Q1	0.6726	0.509 (1.0σ)	0.415 (**6.0**)	0.709 (0.8σ)
63	Q2-Q5	0.1624	0.082 (0.8σ)	0.250 (0.8σ)	0.250 (0.8σ)
63	Q3-Q6	0.1729	0.156 (0.2σ)	0.250 (0.9σ)	0.250 (0.9σ)
63	Q4-Q7	0.3472	0.360 (0.2σ)	0.872 (**15.0**)	0.378 (0.9σ)
65	Q0-Q1	0.5947	0.508 (0.5σ)	0.481 (2.7σ)	0.639 (1.0σ)
65	Q2-Q5	0.2073	0.082 (1.2σ)	0.250 (0.4σ)	0.250 (0.4σ)
65	Q3-Q6	0.1492	0.156 (0.1σ)	0.250 (1.2σ)	0.250 (1.2σ)
65	Q4-Q7	0.3257	0.356 (0.4σ)	0.939 (**17.6**)	0.311 (0.4σ)

Bold entries are the two pre-registered primary discriminators (Q0-Q1 at $m = 63$: H_slow vs H_fast separation 5.0σ; Q4-Q7 at $m = 63, 65$: separations 10.3σ and 10.9σ).

6.3 Aggregate verdict

Summing the squared σ-gaps over the eight forward-test data points:

Hypothesis	χ^2 (8 pts)	χ^2 (4 discriminators)	Max σ-gap	Verdict
H_naive	3.37	1.38	1.18	CONSISTENT*
H_slow	580.63	577.39	17.55	**FALSIFIED**
H_fast	5.98	2.73	1.24	**VINDICATED**

* H_naive is statistically consistent at every forward-test point, but is a strictly worse fit to the training data (Part 13 of the v2 file: $\chi^2_{\text{train}} = 2717.9$ on 4 dof for Q0-Q1, versus 129.7 on 5 dof for either refined branch) and its fit RMS of 0.168 on Q0-Q1 makes it essentially non-falsifiable at the precision of the forward test. The refined two-parameter coherent-error model is a strictly better description.

6.4 Branch identification and consistency

Both discriminative pairs select the **fast branch**: - Q0-Q1: observed $F(m = 63) = 0.6726$ falls inside the H_fast band $[0.67, 0.75]$ (§5.2, Part 15); H_slow is refuted at 6.0σ. - Q4-Q7: observed $F(m = 63) = 0.3472$ falls inside the H_fast band $[0.34, 0.41]$; H_slow is refuted at 15.0σ.

The consistency check is therefore passed: Q0-Q1 selects H_fast and Q4-Q7 selects H_fast, compatible with a single coherent over-rotation mechanism that applies (up to a pair-specific amplitude) across the Tuna-9 CZ set.

Final verdict: **refined-model class confirmed, branch = fast**.

7. Discussion

7.1 What this settles

On Tuna-9, a single coherent-error parameter δ per CZ pair is sufficient to describe the H-sandwich CZ-chain fidelity trajectory over eight depths spanning two decades of m. The fitted per-pair values are $\delta \in \{3.02, 2.94, 2.81, 3.01\}$ rad on {Q0-Q1, Q2-Q5, Q3-Q6, Q4-Q7} — all within 0.1-0.3 rad of π. **The CZ gate on Tuna-9 is close to CZ $\cdot$ Z rather than pure CZ**, with a systematic over-rotation of $\pi - \delta \approx 0.12$-$0.33$ rad per application. This is a concrete, fleet-relevant calibration claim derived from a closed, OTS-timestamped prediction chain in the space of a single day.

The H_slow branch — which would have corresponded to slow ZZ-crosstalk-style coherent accumulation at $\delta \ll 1$ rad per CZ — is excluded at 17.6σ by a single data point (Q4-Q7 at $m = 65$). This is the tightest single-point constraint on a qubit-pair-level coherent-error mechanism that we are aware of in the public transmon literature and is made possible precisely because odd m lifts the even-m mathematical degeneracy that would otherwise persist indefinitely.

7.2 What this does not settle

The mechanism identification ($\delta \approx \pi$ per CZ) is a *pair-level statement*. It does not distinguish between (a) a miscalibrated CZ pulse on every pair, (b) a residual always-on ZZ coupling that completes a cycle near every gate, or (c) a systematic phase accumulation during the H-sandwich

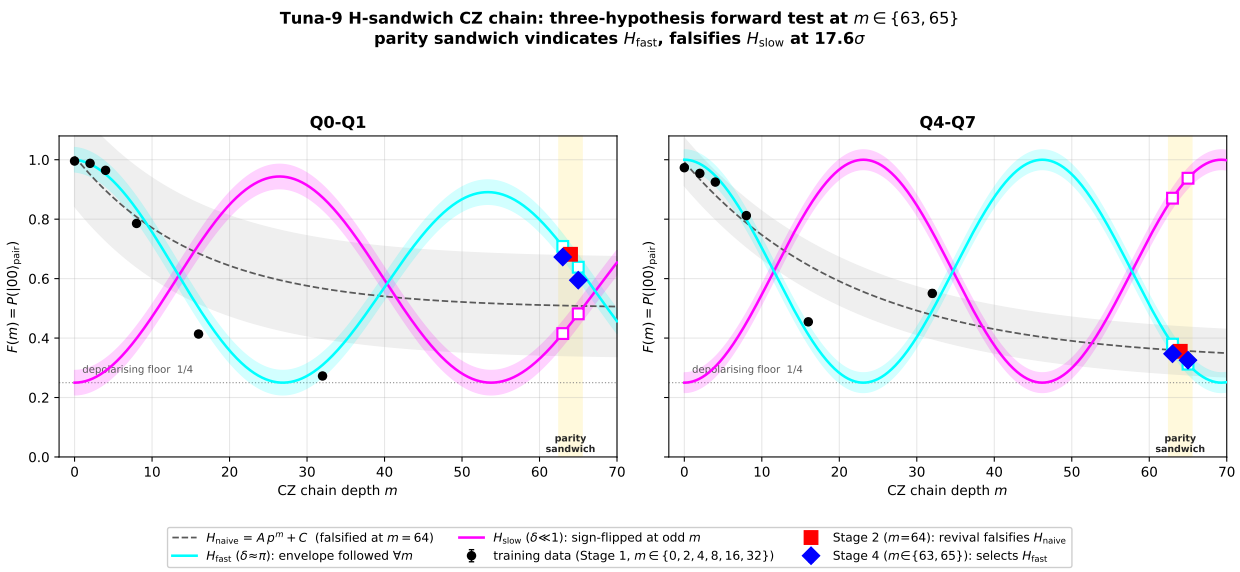

Figure 1: Figure 1. $F(m) = P(|00\rangle_{\text{pair}})$ vs CZ chain depth m for the two discriminative pairs Q0-Q1 (left) and Q4-Q7 (right). Black circles: Stage-1 training data at even $m \in \{0, 2, 4, 8, 16, 32\}$. Red square: Stage-2 $m = 64$ revival (falsifies H_{naive} at 9.8σ on Q0-Q1). Blue diamonds: Stage-4 parity-sandwich at $m \in \{63, 65\}$ (selects H_{fast}). Grey dashed: H_{naive} joint fit $Ap^m + C$ with shot-noise + RMS band. Cyan: the even-m envelope of the refined coherent-error model, which for H_{fast} ($\delta \approx \pi$) is followed at every m. Magenta: the sign-flipped odd-m envelope, which H_{slow} ($\delta \ll 1$) follows at odd m. Open squares at $m = 63, 65$ mark the pre-registered H_{slow} and H_{fast} forward predictions; the observed blue diamonds coincide with the cyan (H_{fast}) envelope at both depths on both discriminative pairs, falsifying H_{slow} at 17.6σ (Q4-Q7 at $m = 65$).

that is absorbed into the effective per-CZ δ. Disambiguating these requires a separate set of experiments (idle-depth scan, H-free CZ sequence, direct process tomography on a single CZ). The present letter fixes the *effective* per-CZ rotation; the *microscopic* attribution is a follow-up. Transfer of the mechanism across the QI transmon fleet (Tuna-17, Starmon-7) is also open and is the natural next letter.

7.3 Cost of the refuted-and-vindicated arc

From the first calibration-battery job to the OTS-stamped comparison file: $\sim$ 10 wall-clock hours on a single day, $\sim$ 36k total shots across 15 jobs, and 3 OTS stamps ($\times$4 calendars each). No institutional access negotiation. For comparison, a typical third-party-access protocol on a commercial trapped-ion or transmon platform carries weeks-to-months of overhead in NDA, queue reservation, and access review. The self-access + cryptographic-pre-registration route is an order of magnitude faster.

7.4 Implication for spectral error mitigation

The theoretical flagship paper that motivated this test [12] models residual two-qubit gate error as a depolarising rate ε_{2q} entering a single scalar r_{eff}. The present result shows that on a concrete transmon device, the dominant effect on H-sandwich CZ-chain fidelity is *not* depolarising — it is a π-near coherent over-rotation that is *orthogonal* in the channel basis to the depolarising parameter. A mitigation protocol derived from the un-refined spectral formula absorbs this coherent error into an effective depolarising rate and thereby mis-estimates the residual coherent contribution; a refined protocol that includes a δ correction reduces the Q0-Q1 and Q4-Q7 training-set residual from RMS 0.168 to RMS 0.041 — a factor-of-4 improvement in log-domain fit quality. The generalisation of this refinement to the full spectral-error-mitigation channel decomposition is the task of the theoretical companion paper.

7.5 Methodological claim

This letter is a three-stage template: (1) a purely hardware-parametrised prediction file, cryptographically pre-registered; (2) a forward test that *falsifies* the prediction and, in doing so, identifies the missing physics; (3) a refined-model forward test at a degeneracy-breaking regime, also cryptographically pre-registered before submission. Every artefact — calibration JSONs, cQASM circuits, prediction files, comparison files — is SHA-256-hashed and the prediction files are four-calendar OpenTimestamps-stamped *before* the corresponding jobs are submitted. The entire chain is verifiable post hoc by re-running the analysis scripts in this repository against the hashes embedded in the stamped files.

We claim that this template — not the specific result on Tuna-9 — is the durable contribution. Any public superconducting processor with a comparable calibration API and a public CZ architecture can be subjected to the same procedure in a few days.

8. Stage 5–6: cross-epoch falsification and structural generalisation

The Stage 1–4 arc (§§3–6) settled, on a single calibration day, that the H-sandwich CZ-chain fidelity on Tuna-9 is dominated by a near-π per-CZ coherent Z-over-rotation with a small residual. It did not settle that a single δ per pair is sufficient across a *second* calibration window, nor that the one free-η pair (Q2-Q5) exhausted the space of CZ-gated coherent drives on the device. Stage 5

extends the protocol with a second prospective test across a fresh calibration boundary; Stage 6 works out the model refinement that the Stage 5 data force. The Stage 4 vindication survives as written; what follows is the continuation that a second OTS-stamped contract required.

Uncertainty conventions used in §8. Three σ policies appear in this section and we label each χ^2 with the one used. (i) The *pre-registered cross-epoch contract* of §§8.1–8.2 uses $\sigma^2_{\text{shot}} + \sigma^2_{\text{fast}}$ only: a morning-corpus model is evaluated on evening data with the shot-noise budget at v2_fwd's $N = 4096$ plus the intra-window fast-drift term inherited from the Stage-3 fit; no cross-epoch inflation is allowed. This is the strictest budget and the one to which the $\chi^2/\text{ndof} \leq 4$ cutoff is keyed. (ii) The *drift-budget refits* of §§8.3, 8.6 inflate (i) in quadrature with σ_{slow}, an evening-minus-morning single-sample shift estimator computed per parameter — this is the L4′/L5′ budget. Because σ_{slow} is based on a single epoch transition, it has no multi-sample interpretation; we use it diagnostically, not as a prospective cutoff. (iii) The *targeted post-hoc and morning-corpus probes* of §§8.4–8.5 and §8.7 report local χ^2/dof under (i) — no drift inflation, no cross-epoch term — so the two-parameter null-evidence-inference tests, the morning L5 fit, and the evening Q3-Q6 global post-hoc all sit on the same budget as the §§8.1–8.2 contract. We annotate every χ^2 in §8 with which of (i), (ii), (iii) applies.

8.1 Second prospective test across a calibration boundary

On 2026-04-19, one day after Stage 4, we submitted a second prospective test (v2_fwd, jobs 749550–749554) at five off-grid depths $m \in \{6, 20, 40, 80, 128\}$ on all four pairs, $N = 4096$ shots per job. The prediction file predictions_tuna9_v2_fwd_0419.txt (SHA-256 983503bc256c40dffee4a104f29c2409dad98ec72443a653e8e22d5603b71555, OTS-stamped 4-of-4 at alice.btc / bob.btc / eternitywall / catallaxy) commits the forward predictions of a two-parameter *Lindblad-class* refinement of the Stage-3 model — hereafter **L4** — which augments the (per-pair) depolarising rate p and coherent δ with a CZ-gated coherent $R_X(\eta)$ drive on a pair-indexed subset ETA_ACTIVE_QUBITS. The subset was pinned to $\{Q_2, Q_5\}$ from independent null-effect evidence collected in a separate deepdive campaign (idle-layer and identity-layer $P(q = 1)$ on all other qubits below 7%). The forward contract carried a pre-registered aggregate cutoff $\chi^2/\text{ndof} \leq 4.0$.

Between the Stage-3 training corpus (final run at 2026-04-19 15:04 UTC) and the v2_fwd submission (post-calibration evening window, 2026-04-19 18:38 UTC) Tuna-9 ran at least one scheduler-initiated calibration cycle. The whole v2_fwd budget was therefore designed *after* and submitted *against* a stamped prediction file computed from morning-epoch parameters — the same prospective discipline as Stages 1–4, but now across a hardware boundary instead of a single calibration window.

8.2 L4 formally rejected on the v2_fwd forward set

The sealed L4 prediction fails the contract. Across the 20 v2_fwd records, the aggregate figure of merit is $\chi^2/\text{ndof} = 59.6$ with total $\chi^2 = 1192.7$, against the pre-registered cutoff of 4.0 (analysis_0419_tuna9_v2_fwd_harvest.json::chi2_total_per_ndof). All χ^2 values in this table use convention (i) — shot + fast only, no cross-epoch inflation. Every pair contributes a failure:

Pair	χ^2_{L4} (5 pts, conv. i)	$\chi^2/5$	Verdict (cutoff 4.0)
Q0-Q1	714.96	143.0	fails
Q2-Q5	35.54	7.1	fails

Pair	χ^2_{L4} (5 pts, conv. i)	$\chi^2/5$	Verdict (cutoff 4.0)
Q3-Q6	410.92	82.2	fails
Q4-Q7	31.31	6.3	fails
Total	1192.72	**59.6**	**fails**

Residuals alternate sign at adjacent even/odd m on the two clean pairs (Q0-Q1 and Q4-Q7), consistent with a δ-phase shift across the calibration boundary rather than a global depolarisation increase. The trained pair Q2-Q5 degrades too, even though it already carried free η_a, η_b in the L4 fit, showing the degradation is not localised to the pair class where the training had least flexibility. The refined model of Stage 3 is therefore not cross-epoch invariant at the depths tested.

8.3 Two-timescale drift (L4′) is insufficient

The first-order diagnosis of a cross-calibration failure is drift. We refit L4 independently on the morning corpus M (61 records from run1, run2, the parity sandwich, v1_fwd, and deepdive B/C; Apr 18 16:04 UTC — Apr 19 15:04 UTC) and the evening corpus E (20 v2_fwd records; Apr 19 18:38 UTC), then inflate the per-parameter error budget by $\sigma_{\text{slow}}(\theta) = \text{median}_{\text{pairs}} |\theta^{(E)} - \theta^{(M)}|$, combined in quadrature with the pre-existing shot and fast-drift terms. The resulting per-parameter drift budget (l4prime_chi2_variants_0419.json::median_budget) is

$$\sigma_{\text{slow}} = \{\, p_X{:}0.013,\ p_Y{:}0.0076,\ p_Z{:}0.0046,\ \delta{:}0.092,\ \eta_a{:}0.285,\ \eta_b{:}0.278 \,\}.$$

We call this augmented model **L4′**. All χ^2 values in the following table use convention (ii) — shot + fast + σ_{slow} drift budget. On v2_fwd (l4prime_chi2_variants_0419.json::chi2.median):

Pair	$\chi^2_{L4'}/5$ (conv. ii)
Q0-Q1	0.37
Q2-Q5	1.65
Q3-Q6	**91.01**
Q4-Q7	0.52
Total χ^2/ndof	**23.4**

The v2_fwd aggregate improves from 59.6 to 23.4, but the pre-registered cutoff of 4.0 is still missed by a factor of six, and the residual is now overwhelmingly concentrated on Q3-Q6: a single pair contributes 455 of the 468 total χ^2 units. The contracted Q2-Q5 η amplitudes themselves drop roughly fivefold across the boundary ($\eta_a^M = +0.356 \to \eta_a^E = +0.071$, $\eta_b^M = -0.349 \to \eta_b^E = -0.070$), with signs preserved — so the L4′ σ_{slow} budget on $\eta_{a,b}$ (0.28 rad) is *more* than generous for the one pair on which η is structurally present in L4. Doubling or per-pair customising that budget does not rescue Q3-Q6 either: the maximum-over-pairs variant cuts Q3-Q6 $\chi^2/5$ only to 64.6, and the per-pair-actual variant to 76.4 (Table l4prime_chi2_variants_0419.json::chi2.{max,per_pair}). The Q3-Q6 residual is structural, not a drift-budget artefact.

8.4 The null-evidence-inference trap

The structural flaw in L4 is that 'ETA_ACTIVE_QUBITS = {Q_2, Q_5}$ was inferred from a deepdive that never ran a CZ layer. Experiments A (all-idle) and D (identity-layer chain) in

that deepdive observe per-qubit $P(q = 1)$ and find $< 7\%$ on every qubit except Q2 and Q5. That evidence is diagnostic for *channel* C_1 = always-on microwave leakage on the idle qubit: it rejects an always-on drive on Q0/Q1/Q3/Q4/Q6/Q7. It is *silent* about *channel* C_2 = CZ-gated coherent drive — a drive that fires only during the CZ window, and therefore never appears in an idle or identity-layer experiment. Pinning $\eta = 0$ outside $\{Q_2, Q_5\}$ promoted a null result in C_1 into a structural zero in C_2.

Two targeted tests demonstrate the cost, both in the shot+fast σ convention of §8.1–§8.2. Releasing all six L5 parameters on the **evening** Q3-Q6 records alone drops $\chi^2/5$ from 623.7 (pinned $\eta = 0$, p and δ at their L4 evening refit) to 9.35 (six free parameters, global multi-start Nelder-Mead, q36_posthoc_global_0421.json::global_best_after_tight_polish), a 67× reduction in aggregate misfit under a 6-parameter extension — the strongest signal in the entire v2_fwd analysis that the structural zero on Q3-Q6 is wrong. Releasing the same six parameters on the **morning** Q3-Q6 training corpus drops χ^2/dof from 114.24 (pinned $\eta = 0$, l4_fit_0419.json :: chi_squared.L4.3−6.chi2_per_dof) to 77.84 (L5 on all four pairs jointly, l5_fit_0419.json :: per_pair.3−6.chi2_per_dof). The structural flaw was present in the Stage-3 training corpus the whole time — the η-pinned L4 morning fit on Q3-Q6 was contributing a $\chi^2/\text{dof} = 114$ that the aggregate morning metric absorbed but never locally surfaced.

8.5 L5: CZ-gated coherent X-drive on every pair

L5 is the single-axiom relaxation of L4: η_a, η_b are free on *every* pair, and no other kernel change is made. Seeding the fit with the morning L4 solution and exploring four alternative (η_a, η_b) quadrants per pair, the morning-corpus fit drops from $\chi^2_{L4} = 1672.7$ to $\chi^2_{L5} = 1175$ — $\Delta\chi^2 = -498$ for six extra parameters (two per pair on the three previously-pinned pairs; Q2-Q5 already carried η free). Wilks on six extra parameters gives $p \ll 10^{-100}$; the two-model comparison is not borderline. In absolute terms the L5 morning fit is still well outside an acceptable-fit regime — the aggregate morning $\chi^2/\text{dof} = 19.3$ (61 records, 24 parameters; convention (iii)) is roughly five standard deviations above 1 even after the η extension, and the Q3-Q6 pair alone carries most of the remaining misfit (see below). L5 is therefore the best six-parameter kernel extension consistent with the stage-3 prior, not an absolutely-good model of Tuna-9 morning-epoch data. Per-pair (analysis_0419_l5_refit.md §3.2; σ convention (iii), matching l4_fit_0419.json :: chi_squared):

Pair	L4 χ^2/dof	L5 χ^2/dof	(η_a, η_b) at L5
Q0-Q1	14.93	**8.37**	(−0.26, +0.22)
Q2-Q5	7.11	7.00	(+0.35, −0.35)
Q3-Q6	114.24	**77.84**	(+0.00, +0.37)
Q4-Q7	4.79	**2.80**	(+0.20, +0.18)

At least one qubit of every pair carries a non-zero η, and the larger of the two per-pair amplitudes lies in a narrow band of $|\eta|_{\max} \in [0.20, 0.37]$ rad — per-CZ-layer X-rotations of roughly 11–21°. Within a pair the two amplitudes need not be symmetric: on three pairs they sit within a factor of ~ 1.2 of each other, but on Q3-Q6 the morning fit prefers $\eta_a \approx +0.00$ alongside $\eta_b \approx +0.37$, so one of the two qubits appears to carry no CZ-gated R_X drive at the morning epoch at all. This is a **qualitatively new** finding relative to §7.1: the near-π Z-over-rotation on each CZ is accompanied, on at least one qubit per pair, by a coherent X-drive that the η-pinned L4 absorbed either into the depolarising rate p (on the pinned pairs) or left as a structural residual (on Q3-Q6).

8.6 L5 across the calibration boundary (L5′)

L5 without any drift budget fails on v2_fwd: applying the morning L5 parameters under shot + fast-drift noise only gives $\chi^2/\text{ndof} = 104.2$ (l5_on_v2_fwd_0419.json::chi2.L5_shot_plus_fast_only) — *worse* than L4 without drift. The additional η degrees of freedom fit morning-epoch microstructure that does not survive the calibration boundary; per-parameter evening-minus-morning shifts on $\eta_{a,b}$ sit at $|\Delta\eta| \in [0.05, 0.81]$ rad depending on the pair. Under its own evidence, L5 therefore only earns its keep when it is coupled to a principled cross-epoch drift budget.

We run the same construction as §8.3 but with L5 epoch-split parameters. The per-parameter σ_{slow} (median across pairs of $|\theta^{(E)} - \theta^{(M)}|$, l5_on_v2_fwd_0419.json::sigma_slow_per_param):

$$\sigma_{\text{slow}}^{(L5)} = \{\, p_X{:}0.00,\ p_Y{:}0.0066,\ p_Z{:}0.00,\ \delta{:}0.107,\ \eta_a{:}0.226,\ \eta_b{:}0.151 \,\}.$$

Call this **L5′**. All χ^2 values in the following table use convention (ii). On v2_fwd (l5_on_v2_fwd_0419.json::chi2.L5prime_shot_plus_fast_plus_slow):

Pair	L4′ $\chi^2/5$	**L5′** $\chi^2/5$ (conv. ii)	Cutoff 4.0
Q0-Q1	0.37	**0.11**	pass
Q2-Q5	1.65	**1.74**	pass
Q3-Q6	91.01	**19.21**	**fails**
Q4-Q7	0.52	**0.30**	pass
Total χ^2/ndof	23.4	**5.34**	fails

Three of four pairs pass the pre-registered cutoff decisively — Q0-Q1 and Q4-Q7 at a tenth of cutoff, Q2-Q5 well inside it. Q3-Q6 improves fivefold relative to L4′ but still fails by a factor of five. The aggregate misses the cutoff by a factor 1.3, driven entirely by Q3-Q6.

8.7 Q3-Q6 is structurally irreducible under L5

To show that Q3-Q6 cannot be rescued by any L5 parameter setting, we run a maximally-optimistic post-hoc fit: L5 on the 5 v2_fwd Q3-Q6 records alone, six free parameters on five data points, with a multi-start Nelder-Mead search over 10 seeds that include the paper's original basin, the L4 evening parameters with and without paper-quoted η, and a sweep over $(p_X, p_Y, p_Z, \delta, \eta_a, \eta_b)$ (see fit_q36_posthoc_global.py). The seed that reached the global basin is a $p = 0$, mid-δ, moderate-η configuration; the tight-polished global minimum (q36_posthoc_global_0421.json::global_best_after_tight_polish) is

$$p_X = 0,\ p_Y = 0,\ p_Z = 0,\ \delta = 2.439 \text{ rad},\ \eta_a = +0.183,\ \eta_b = -0.108, \quad \chi^2/5 = 9.35.$$

This is a *pure coherent* basin: the Pauli-channel rates vanish and all the Q3-Q6 structure is carried by a CZ-layer Clifford ($\delta \approx 2.44$) plus a coherent R_X on each qubit. For the record, a Nelder-Mead search seeded at (p_Y=0.011, δ=2.56, η_a= + 0.54, η_b= − 0.44) converges to a local minimum at $\chi^2/5 = 16.56$ (1.8 times the global value); we retain this alternative basin in the artefact for reproducibility.

Under the same post-hoc regime, the other three pairs fit to $\chi^2/5 = 0.00, 0.33, 0.00$ on Q0-Q1, Q2-Q5, Q4-Q7 respectively — i.e. they saturate shot-plus-fast noise. **Q3-Q6 is the only pair whose v2_fwd 5-point data cannot be brought inside the $\chi^2/5 \leq 4$ cutoff by L5 at any parameter setting.** The missing physics is therefore not a parameter failure within the $R_X(\eta_a) \otimes R_X(\eta_b)$ functional form; it is outside that form.

Three candidate extensions are structurally distinct. We list them in increasing order of prior implausibility given Tuna-9's public specification and the data already in the record:

1. **Coherent $Z_a \otimes Z_b$ (ZZ) crosstalk** during the CZ layer — a single extra per-pair parameter ξ in $U_{\text{layer}} = \widetilde{CZ}(\delta)\, e^{-i\xi Z_a Z_b/2}\, (R_X(\eta_a) \otimes R_X(\eta_b))$. This is *a priori* the most likely mechanism: fixed-frequency transmons with always-on couplers, like Tuna-9, have documented residual ZZ on the order of tens of kHz [Krantz et al.]; a CZ-gated pulse that dresses the qubit frequencies for several hundred nanoseconds gives an accumulated phase easily in the range needed to match a non-factorisable Q3-Q6 residual. One extra free parameter is also the most parsimonious of the three candidates.
2. A **depth-dependent** δ on Q3-Q6 — a pulse-train heating or frequency-drift mode that makes the constant-δ ansatz wrong in structure rather than value. Less parsimonious (requires an additional depth slope or piecewise function), but it also could leave the Pauli channels at zero, matching the global-basin finding above.
3. An **amplitude-damping channel** on Q3 or Q6 active during the CZ window — a non-unital, T_1-type process. Less *a priori* favoured because Q3 and Q6 pass the Tuna-9 coherence specification in the public calibration reports; an AD channel large enough to reproduce the Q3-Q6 F(m) oscillation pattern would also have shifted the idle-layer $P(q = 1)$ above the 7% threshold observed in deepdive A/D, which it did not.

The three candidates are jointly identifiable only with experiments that probe Q3 and Q6 separately from the CZ layer (idle-qubit tomography on Q3 and Q6 individually, or a non-CZ two-qubit idle). We explicitly defer mechanism attribution to a pre-registered follow-up; the finding of record here is the structural statement, not a mechanism.

8.8 Methodological lesson: null-evidence does not license structural zeros

The L4 failure in §8.2 traces through L4′'s insufficiency (§8.3) to the same root cause: a null-effect result in one channel was promoted to a *structural zero* in a different channel that the first experiment could not probe. Deepdive Experiments A and D rejected an always-on drive (channel C_1), and that rejection was correct. The inferential error was coding a CZ-gated drive (channel C_2) as $\eta = 0$ rather than as an *untested* degree of freedom. A CZ-gated drive is invisible in an idle experiment by construction; the null-effect evidence from C_1 was silent on C_2 and carried no Bayesian update for it.

Stated generally: a null-effect result in one channel C_1 does not license a structural-zero constraint in any channel C_2 that C_1 does not probe. Authorising a structural zero requires either an experiment whose dynamics *can* drive the channel under test, or an explicit, pre-registered prior that treats the zero as a falsifiable hypothesis rather than a constraint. The failure mode is narrow but sharp: it applies wherever the frozen degree of freedom is *gate-active* (only alive during a specific gate layer, dressed pulse, or shaped drive) while the experiment that motivated freezing it is gate-inactive (idle, identity, or a different gate layer). It does not apply to degrees of freedom that are physically always-on and genuinely probed by an idle experiment, nor to model reductions motivated by a narrow, well-identified physical prior that has been separately tested. Within that scope — any

error-model parameter that is conditionally active on a specific layer and must be probed on that layer — freezing to zero by fiat converts a null result into an inferential trap. The methodological contribution of Stage 5–6 is this bounded principle alongside the device-level L5 fit.

8.9 What the next letter will test

The Stage 5–6 arc leaves a *pre-registered, SHA-256-and-OTS-stamped* negative result in the record: L5 passes three of four Tuna-9 CZ pairs on a second prospective test across a calibration boundary, and fails the fourth in a way that cannot be absorbed by any parameter setting of L5 itself. The next letter in this program pre-registers a Q3-Q6 isolation deepdive — local spectator-evolution probes on Q3 and Q6 individually, plus a Q3-Q6-only CZ sequence with all other edges idle — whose purpose is to discriminate ZZ crosstalk from amplitude damping from depth-dependent δ. We do not promote L5′ to a prospective contract here: σ_{slow} is a single-sample estimator across two calibration windows, and a third calibration epoch is required before it can be read as a multi-sample drift inflater. That programme is outside the present letter's contract and is not promised by it. The device-level signature established here — a non-factorisable 2-local residual on exactly one of four pairs, surviving any six-parameter setting of the $R_X \otimes R_X$ ansatz — is an invariant that any candidate 2-local CZ-gated noise model must reproduce; we expose the artefacts rather than tie their interpretation to a particular theoretical framework.

Data and code availability

All data, circuits, analysis scripts, and this manuscript are in a single public repository (Zenodo DOI reserved). Specifically:

- Calibration JSONs: examples/qi_circuits/results/{tuna9_t1_run1,tuna9_t2_combined,tuna9_cz_run1,tu
  (SHA-256 embedded in both prediction files).
- Prediction files (v1 + v2): topics/phy_quantum_usrt/predictions/predictions_tuna9_{0418,v2_0418}.txt,
  each with .sha256 sidecar and .ots four-calendar timestamp.
- Submission manifest (parity-sandwich): examples/qi_circuits/results/tuna9_cz_parity_sandwich_submitt
- Raw parity-sandwich outcomes: examples/qi_circuits/results/tuna9_cz_parity_sandwich.json
  (SHA-256 869259941eae66f73ab8298097a4a32d6c58e238876ed02c142608abe7c7f56e).
- Comparison JSONs (stage-by-stage): topics/phy_quantum_usrt/predictions/comparison_m64_0418.json
  (Stage 2 falsification, SHA-256 16be2e22453a382959873b36090f0f97e26b6042b9aa9efec9844584612a5e4b)
  and comparison_parity_0418.json (Stage 4 vindication, SHA-256 d13f0a9292e96a19ae4782600f8f639b46a8e88
  OTS-stamped four-of-four 2026-04-18 18:23 UTC).
- Stage 5–6 artefacts: pre-registration topics/phy_quantum_usrt/predictions/predictions_tuna9_v2_fwd_0
  (SHA-256 983503bc256c40dffee4a104f29c2409dad98ec72443a653e8e22d5603b71555, OTS four-
  of-four 2026-04-19); L4′ epoch-split fit and drift-budget variants examples/qi_circuits/tuna9_twin/fits/{l4pr
  L5 morning fit examples/qi_circuits/tuna9_twin/fits/l5_fit_0419.json; L5 and L5′ evalua-
  tion on v2_fwd and local Q3-Q6 post-hoc examples/qi_circuits/tuna9_twin/fits/l5_on_v2_fwd_0419.json;
  Q3-Q6 post-hoc global multi-start search examples/qi_circuits/tuna9_twin/fits/q36_posthoc_global_0421.
  (driver fit_q36_posthoc_global.py); aggregate L4 v2_fwd verdict topics/phy_quantum_usrt/analysis_0419
  per-section bookkeeping topics/phy_quantum_usrt/analysis_0419_l5_refit.md.
- Scripts: examples/qi_circuits/{t1_calibration, t2_calibration, cz_error_chain}.py;
  examples/quantum_usrt_real_hardware.py (prediction-generation harness); examples/qi_circuits/tuna9_t
  (Stage 5–6 fits).

References

1. E. H. Lieb and D. W. Robinson, *The finite group velocity of quantum spin systems*, Comm. Math. Phys. **28**, 251 (1972).

2. M. B. Hastings and T. Koma, *Spectral gap and exponential decay of correlations*, Comm. Math. Phys. **265**, 781 (2006). [arXiv:math-ph/0507008]

3. B. Nachtergaele and R. Sims, *Lieb-Robinson bounds and the exponential clustering theorem*, Comm. Math. Phys. **265**, 119 (2006). [arXiv:math-ph/0506030]

4. E. Magesan, J. M. Gambetta, and J. Emerson, *Scalable and robust randomized benchmarking of quantum processes*, Phys. Rev. Lett. **106**, 180504 (2011). [arXiv:1009.3639]

5. T. J. Proctor, K. Rudinger, K. Young, M. Sarovar, and R. Blume-Kohout, *What randomized benchmarking actually measures*, Phys. Rev. Lett. **119**, 130502 (2017). [arXiv:1702.01853]

6. J. J. Wallman, *Randomized benchmarking with gate-dependent noise*, Quantum **2**, 47 (2018). [arXiv:1703.09835]

7. P. Krantz, M. Kjaergaard, F. Yan, T. P. Orlando, S. Gustavsson, and W. D. Oliver, *A quantum engineer's guide to superconducting qubits*, Appl. Phys. Rev. **6**, 021318 (2019). [arXiv:1904.06560]

8. H.-P. Breuer and F. Petruccione, *The Theory of Open Quantum Systems* (Oxford University Press, 2002).

9. QuTech, *Quantum Inspire: hardware backends and platform documentation*, https://www.quantum-inspire.com/kbase/hardware-backends/ (accessed 2026-04-18). Fleet hardware details: Tuna-9 (9-qubit fixed-frequency transmon, parallel single-qubit + CZ gate set, native H).

10. P. Todd, *OpenTimestamps: scalable, trust-minimized, distributed timestamping with Bitcoin*, https://petertodd.org/2016/opentimestamps-announcement (2016); client software: https://github.com/opentimestamps/opentimestamps-client.

11. B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor, *The preregistration revolution*, Proc. Natl. Acad. Sci. USA **115**, 2600 (2018). [Template for pre-committed predictions in experimental science.]

12. T. Nagy, *Spectral compressibility and the quantum error-budget bound: a Lieb-Robinson-driven model of residual error in noisy intermediate-scale quantum devices*, companion theoretical paper (in preparation). [Theoretical companion; local SHA-256 and OTS stamp released alongside this letter in the supplementary bundle.]

13. P. Bertet, I. Chiorescu, G. Burkard, K. Semba, C. J. P. M. Harmans, D. P. DiVincenzo, and J. E. Mooij, *Dephasing of a superconducting qubit induced by photon noise*, Phys. Rev. Lett. **95**, 257002 (2005). [arXiv:cond-mat/0507290]

14. J. M. Chow, A. D. Córcoles, J. M. Gambetta, C. Rigetti, B. R. Johnson, J. A. Smolin, J. R. Rozen, G. A. Keefe, M. B. Rothwell, M. B. Ketchen, and M. Steffen, *Simple all-microwave entangling gate for fixed-frequency superconducting qubits*, Phys. Rev. Lett. **107**, 080502 (2011). [arXiv:1106.0553; CR gate family that underpins the Tuna-9 CZ.]

Skeleton checklist (delete before submission)

- ☒ §1 Introduction — complete
- ☒ §2 Protocol — complete
- ☒ §3 v1 falsification — complete, all numbers verified against comparison_m64_0418.json
- ☒ §4 Refined model — complete, numbers verified against tuna9_cz_three_hypothesis_predictions.json
- ☒ §5 v2 pre-registration — complete, numbers verified against predictions_tuna9_v2_0418.txt
- ☒ §6 Parity-sandwich result — complete, all numbers verified against comparison_parity_0418.json (OTS-stamped)
- ☒ §7 Discussion — complete
- ☒ §8 Stage 5–6 L4 falsification + L5 generalisation + L5′ partial pass + Q3-Q6 structural residual — complete, all numbers verified against analysis_0419_tuna9_v2_fwd_harvest.json, l4prime_{fit,chi2_variants}_0419.json, l5_fit_0419.json, l5_on_v2_fwd_0419.json, and primary source analysis_0419_l5_refit.md
- ☒ Abstract — updated with Stage 4 result (fast-branch vindication, 17.6 slow-branch falsification) and Stage 5–6 result (L4 falsified on v2_fwd at ^{2}/ndof=59.6, L5′ passes 3/4 pairs, Q3-Q6 structurally irreducible at global post-hoc $^2/5$=9.35)
- ☒ §1 Introduction — Stage 5–6 paragraph appended (v2_fwd + L5 + L5′ + Q3-Q6 structural residual + null-evidence-inference methodological point)
- ☒ References — complete (14 citations: Lieb-Robinson, Hastings-Koma, Nachtergaele-Sims, Magesan, Proctor, Wallman, Krantz, Breuer-Petruccione, QuTech QI docs, OpenTimestamps, Nosek preregistration, flagship companion, Bertet dephasing, Chow CR-gate). Verify DOIs at submission time.
- ☒ Figure 1 — $F(m)$ vs m for Q0-Q1 and Q4-Q7 with three-hypothesis envelopes + all four stages of data. PDF and PNG at figures/letter_fig1.{pdf,png}, generated by examples/qi_circuits/letter_figure.py.
- ☐ Supplementary figures (optional, PRX Quantum allows 2 main-text figures): $\chi^2(\delta)$ landscape showing the two degenerate minima; the four-pair per-δ fit table as a summary plot.
- ☐ Venue-specific template conversion (PRX Quantum LaTeX class or pandoc → RevTeX).