Supplementary Information

Scalable and Verifiable Quantum Benchmarking with Spectrally Flat Hidden Shifts

Lior Horesh, Vasileios Kalantzis, Shashanka Ubaru

Contents

Status of the claims made in this paper **3**

1 S1: Proof of the Spanning Lemma **5**

2 S2: Classical Query Complexity **5**
2.1 Public f: $\mathcal{O}(n/\alpha^2)$ classical queries suffice 6
2.2 Hidden injective oracles: the $\Omega(2^{n/2})$ barrier 6
2.3 Consequence for this paper 7

3 S3: Proof of the Noise Amplification Theorem **7**

4 S4: Shift Identity **8**

5 S5: Classical Information-Rate Bound in the Hidden-Injective-Oracle Model **8**

6 S6: Quadratic Oracle Decomposition **9**

7 S7: Degree-3 Oracle and T-Count **10**

8 S8: Noise Model Details **11**

9 S9: Retired Single-Oracle Degree-2 Campaign (Output-Support Control) **13**

10 S10: Scale-up Degree-2 Experiments ($n = 68$–154) **15**

11 S11: Higher-Degree Experiments **16**

12 S12: Lean 4 Formalization **20**

13 S13: Taxonomy of Problems Reducible to Hidden Shift **21**
13.1 Equivalence with the dihedral hidden subgroup problem 23
13.2 Lattice problems 23
13.3 Isogeny-based cryptography 24
13.4 Boolean hidden shift 24
13.5 Shifted Legendre symbol 24
13.6 Symmetric-key cryptanalysis 24

13.7 Spectral flatness transfer . . . 25
13.8 Summary of the reduction landscape . . . 25

14 S14: Unbounded-Degree Validation (M-M with $C^{(m)}Z$, m growing with n) 26

15 S15: Multi-Output Noise Amplification (Abelian-Group Hidden Shift) 34

16 S16: Cross-Backend F-vs-N_{2Q} Test of the 2Q-Only Noise Model 36

17 S17: Noise Theorem under General Pauli Noise 39

18 S18: Worked Cryptographic Example – Even-Mansour Key Recovery 40

19 S19: Sharper Pauli-Noise Bound for Diagonal Phase Oracles 41

20 S20: Structured Noise and the Per-Bit Recovery Statistic 43
20.1 Bit-flip readout model . . . 43
20.2 The full-bitstring mode is not structurally protected . . . 43
20.3 Per-bit majority vote is width-independent . . . 44
20.4 Graceful degradation and the single-wrong-bit signature . . . 44
20.5 General Hamming-ball noise . . . 44
20.6 Positively correlated noise: the joint mode outlives the per-bit vote . . . 45
20.7 The measured margin at the frontier . . . 47

21 S21: Cross-Architecture Demonstration 49

22 S22: Phase-State Parity Runs with a Fixed Matching (Hidden Matching Cells) 53
22.1 Problem setting, record and scope . . . 53
22.2 Registered design and scoring rule . . . 55
22.3 Exploratory extensions . . . 55
22.4 Execution . . . 55
22.5 Interpretation and reclassification . . . 55
22.6 Protocol and decoder . . . 56
22.7 Registered design versus extensions . . . 56
22.8 Per-instance results . . . 56
22.9 Pre-registered rerun on the second architecture . . . 56

23 S23: Powered Mean-Flatness Sweep, Per-Cell Record 58
23.1 Original registered hypothesis . . . 59
23.2 Initial min-flatness sweep . . . 59
23.3 Powered mean-flatness experiment . . . 59
23.4 Registered decision rule . . . 60
23.5 Results on the primary processor . . . 60
23.6 Replication on a second processor . . . 61
23.7 Replication on a second architecture . . . 61
23.8 Gate-content controls . . . 61
23.9 Scope and limitations . . . 61
23.10 Supporting data . . . 61

24 S24: Registration Chronology of Every Hardware Claim **62**
24.1 Registration policy . . . 62
24.2 Registered hardware studies . . . 63
24.3 Degree-2 ladder chronology . . . 63
24.4 Submission manifests . . . 64
24.5 Routing audit . . . 64
24.6 Updated submission workflow . . . 64
24.7 Analysis workflow . . . 65
24.8 Recovery statistics . . . 65
24.9 Statistical intervals . . . 65
24.10 Automated verification . . . 65
24.11 Machine-readable record . . . 66
24.12 Interpretation . . . 66

25 S25: Frontier Records and Diagnostics **68**
25.1 Routing audit of the degree-2 record . . . 68
25.2 Margin score of every frontier run . . . 68
25.3 Per-bit and best-rival diagnostics . . . 68
25.4 Two-component mixture model of the frontier runs . . . 69

26 S26: Protocol Card **69**

27 S27: Additional tables and figures **69**

Status of the claims made in this paper

Table 1 tracks every claim of the main text by its status (proved, demonstrated, conditional, disclaimed or open); the main-text Discussion points here. Section numbers are those of this Supplement.

Table 1: Status of the claims made in this paper. "Proved" means proved here or in the Supplement; where a row says Lean, the stated part is machine-checked unconditionally (the named external axioms of Section S12 are standalone records of unformalized standard steps, and no Lean theorem depends on them). "Demonstrated" means observed on IBM hardware. "Conditional" means contingent on a named external assumption. "Disclaimed" means the claim is explicitly not made at the demonstrated parameters.

Claim	**Status**
Quantum query upper bound $\mathcal{O}(n/\alpha^2)$	Proved analytically (Section S1); not Lean-formalized
Classical recovery for public f in $\mathcal{O}(n)$ queries	Proved (Section S2); no exponential query separation on the demonstrated bent instances
$\Omega(2^{n/2})$ classical query barrier for injective abelian-group oracles	Proved (Section S2). The matching quantum algorithm is Simon's for $\mathbb{Z}_2^n$ (polynomial queries) and Kuperberg's sieve for cyclic groups (subexponential). Neither regime is run here
Majority-vote suppression by $\Theta(\alpha^2 \cdot 2^n)$	Dominance-ratio algebra of main-text Corollary 3 proved unconditionally in Lean, the uniform case only; the concentration step is the named Hoeffding axiom (Section S12)
Per-bit recovery under i.i.d. flips; multi-output and structured-Pauli extensions	Proved (Sections S15, S17, S19; an earlier $2^n/n$-sharper claim is withdrawn in S19)
Joint-mode advantage under positively correlated noise	Proved (association bound plus burst-noise regime, Section S20); measured on both hardware records (overdispersion 1.4–2.1×)
Degree-2 recovery on Heron r2 (dual-oracle)	Demonstrated on two processors (registration chronology in S24): 40/40 through $n = 20$ by both statistics; 10/10 at $n = 24$ by full-string vote at measured $F = 0.9$–3.0×10^{-3} (processor B 5/5 stable under resampling; processor A 5/5, three of five unstable); one vote-carried recovery at $n = 32$ at the shot floor (the mode survives resampling with probability 0.02 naive, 0.79 tail-smoothed); none beyond (main-text Table 3)
Earlier 427-run single-oracle campaign	**Reclassified**: output distribution provably independent of s for bent f, a design error found in a post-hoc internal audit; retained only as an output-support control (Sections S9–S10)
Non-Clifford (CCZ) recovery at $n = 8$–140	Demonstrated, 156/259 across five backends (complete record, Section S11); interleaved layouts at large n (entanglement confined to the monomial support)
Benchmark metric ranks backends by measured 2Q error	Demonstrated, pre-registered; Spearman $\rho = -0.90$ across five processors, second registered test after a disclosed failed first (Section S16)
One-way phase-state messages (Hidden Matching cells)	**Reclassified**: the registered protocol fixed Bob's matching to the shift matching on the top qubit ($k = q - 1$), a property of the registered design found in a post-hoc internal audit (2026-09-21). For that problem a classical one-way message of one bit (matching fixed and known) or at most q bits (any of the q shift matchings, Bob's choice private) succeeds with certainty, so the $\Omega(\sqrt{N})$ bound [1] does not apply to the runs and no communication advantage is demonstrated. Retained as a prepare-and-measure record of the phase states' parity success at $q = 5$ to 8 (registered $q = 6$ at 0.752 unmitigated on processor B, $q = 8$ at 0.5735 on processor C), a fidelity readout of the phase state through one Hadamard (Section 22)
Cryptographic transfer (lattice, isogeny, symmetric-key)	Conditional on hardness of the underlying problem and a bounded-query adversary model
Classical intractability of unbounded-degree instances	**Open**; our search closed both natural routes without finding a classically hard noise-friendly instance (Section S14)
Computational supremacy or query separation at demonstrated parameters	**Disclaimed**; the demonstrations show majority-vote recovery persisting under a quantified departure from the premise of the noise-amplification theorem, not that premise or its count ratio
Cryptographic-parameter cryptanalysis	**Disclaimed**; fault-tolerant-era device-side problem (Table 8)

1 S1: Proof of the Spanning Lemma

Lemma 1.1 (Spanning from non-uniform sampling)**.** *Let D be a probability distribution over $\mathbb{F}_2^n$ with $D(\xi) \geq \mu > 0$ for all $\xi \in \mathbb{F}_2^n$. Then Q independent samples $\xi_1, \ldots, \xi_Q \sim D$ span $\mathbb{F}_2^n$ with probability at least $1 - \delta$, provided*

$$Q \;\geq\; \frac{n \ln 2 + \ln(1/\delta)}{\mu \cdot 2^{n-1}}.$$

Proof. The event "samples do not span $\mathbb{F}_2^n$" is equivalent to all samples lying in some proper subspace of $\mathbb{F}_2^n$. Every proper subspace is contained in a maximal subspace (of codimension 1), and $\mathbb{F}_2^n$ has exactly $2^n - 1$ maximal subspaces (one per nonzero linear functional).

Fix a maximal subspace V with $|V| = 2^{n-1}$. The complement $\mathbb{F}_2^n \setminus V$ has 2^{n-1} elements, each with probability $\geq \mu$ under D. So

$$D(V) \;=\; 1 - D(\mathbb{F}_2^n \setminus V) \;\leq\; 1 - 2^{n-1}\mu.$$

The probability that all Q samples lie in V is at most $(1 - 2^{n-1}\mu)^Q$. By a union bound over the $2^n - 1$ maximal subspaces:

$$\begin{aligned} \Pr[\text{not spanning}] &\leq (2^n - 1)(1 - 2^{n-1}\mu)^Q \\ &< 2^n \exp(-2^{n-1}\mu Q). \end{aligned} \tag{1}$$

Setting (1) $\leq \delta$ and solving:

$$Q \;\geq\; \frac{n \ln 2 + \ln(1/\delta)}{2^{n-1}\mu}.$$

For the hidden shift algorithm, $\mu = \alpha^2/2^n$, so $2^{n-1}\mu = \alpha^2/2$ and $Q \geq 2(n \ln 2 + \ln(1/\delta))/\alpha^2 = \mathcal{O}(n/\alpha^2)$. □

Remark 1.2 (Tighter bound via coupon collecting)**.** *The union bound above is loose. A tighter analysis conditions on the rank achieved after each sample. Let $r_k = \dim \operatorname{span}(\xi_1, \ldots, \xi_k)$. When the current rank is $r \leq n - 1$, exactly $2^n - 2^r \geq 2^{n-1}$ vectors lie outside the span, each drawn with probability at least μ, so the probability that the next sample increases the rank is $p_r \geq \mu\,(2^n - 2^r)$. The expected number of samples to go from rank r to $r + 1$ is at most $1/p_r \leq 1/\big(\mu\,(2^n - 2^r)\big)$. Summing, with $k = n - r$:*

$$\mathbb{E}[\textit{samples to span}] \;\leq\; \frac{1}{\mu\, 2^n} \sum_{k=1}^{n} \frac{1}{1 - 2^{-k}} \;\leq\; \frac{n+2}{\mu\, 2^n},$$

since $\sum_{k \geq 1} \big(1/(1 - 2^{-k}) - 1\big) \leq 2$. With $\mu \geq \alpha^2/2^n$ this is at most $(n + 2)/\alpha^2$. A Markov or Chernoff bound then gives $Q = \mathcal{O}(n/\alpha^2)$ with high probability, matching the union bound up to constants (compare the full-rank asymptotics of random matrices over $\mathbb{F}_2$ [2]).

2 S2: Classical Query Complexity

This section corrects a claim made in earlier versions of this work. We had asserted an unconditional $\Omega(2^{n/2})$ classical query lower bound for the α-flat hidden shift problem. That statement is false when the function f is public, which is the setting of every hardware demonstration in this paper: the collision argument used to prove it requires the function to be a hidden injective black box, a hypothesis that the public ± 1-valued bent functions violate. For public f the problem is easy in the query model, and a genuine collision barrier survives only for injective oracles on a group.

2.1 Public f: $\mathcal{O}(n/\alpha^2)$ classical queries suffice

Proposition 2.1 (Level-set recovery)**.** *Let $f \in \mathcal{F}_\alpha$ be public and $g(x) = f(x \oplus s)$. A classical algorithm that queries g at $q = \lceil (n + \log_2(1/\delta) + 1)/\log_2(1/(1-\alpha^2/2)) \rceil$ uniformly random points recovers s with probability at least $1-\delta$. For bent f ($\alpha = 1$), $q = n + \log_2(1/\delta) + 1$ suffices.*

Proof. Query x_i returns $b_i = f(x_i \oplus s)$, consistent with a candidate s' iff $f(x_i \oplus s') = b_i$. Fix $s' \neq s$ and write $d = s \oplus s' \neq 0$. Over uniform x,

$$\Pr_x[f(x \oplus s') \neq f(x \oplus s)] = \frac{|S_{s,s'}|}{2^n}, \qquad 2^n - 2\,|S_{s,s'}| = \sum_x f(x \oplus s)\, f(x \oplus s') = C_f(d),$$

the autocorrelation of f at offset d. By Parseval, $C_f(d) = 2^{-n} \sum_\xi |\hat{f}(\xi)|^2 (-1)^{d \cdot \xi}$. With $|\hat{f}(\xi)|^2 \geq \alpha^2 2^n$ and $\sum_\xi |\hat{f}(\xi)|^2 = 2^{2n}$, one gets $|C_f(d)| \leq (1-\alpha^2)2^n$, hence $\Pr_x[\text{eliminate } s'] \geq \alpha^2/2$. For bent f, $C_f(d) = 0$ exactly, so this probability is exactly $1/2$. The q queries are independent, so $\Pr[s' \text{ survives all}] \leq (1-\alpha^2/2)^q$. A union bound over the $2^n - 1$ wrong shifts is below δ at the stated q. The unique survivor is s. □

The classical query count is thus $\mathcal{O}(n/\alpha^2)$, matching the quantum count of Theorem 5 up to constants. Solving the surviving constraints is Gaussian elimination for degree 2 and, for bounded ANF degree, the polynomial-time confluent path-sum rewriting of Amy and Stinchcombe [3]. There is no exponential quantum query advantage on these instances.

2.2 Hidden injective oracles: the $\Omega(2^{n/2})$ barrier

The collision argument is correct in a different model. Let $h\colon G \to R$ be an *injective* oracle on a finite abelian group G of order N, given as a black box with its description withheld, and let $g(x) = h(x + s)$. Recovering s requires a claw $h(x) = g(y)$, i.e. $x = y + s$.

Theorem 2.2 (Collision lower bound, injective oracles)**.** *Any classical randomized algorithm recovering s from q queries to a black-box injective h and $g = h(\cdot + s)$ succeeds with probability at most $\mathcal{O}(q^2/N)$, so $q = \Omega(\sqrt{N})$ queries are required.*

Proof sketch. This is the classical birthday argument: with the description of h withheld and s uniform, the transcript of q queries to h and g reveals nothing about s until a claw $h(x) = g(y)$ appears, and for uniformly random s each of the $\leq q^2/4$ query pairs is a claw with probability $\mathcal{O}(1/N)$. A union bound gives the stated success probability. (This is a statement about classical queries, and the well-known quantum collision lower bound of Aaronson–Shi [4] concerns quantum query complexity and is not what is used here.) □

This regime covers the cryptographic reductions of Section S13 (injective class-group actions, LWE samplers, block ciphers as permutations). First, the bound needs injectivity and a hidden description, neither of which holds for the public bent functions run on hardware. Second, the matching quantum algorithm depends on the group and is never the $\mathcal{O}(n/\alpha^2)$-query bent recovery, whose sign-extraction step needs the public dual that an injective group oracle does not provide. For $G = \mathbb{Z}_2^n$ the problem is Simon's problem [5]: the function $F(x, b)$ equal to $h(x)$ for $b = 0$ and to $g(x)$ for $b = 1$ is two-to-one on $\mathbb{Z}_2^{n+1}$ with hidden period $(s, 1)$, so $\mathcal{O}(n)$ quantum queries recover s against the $\Omega(2^{n/2})$ classical bound, the exponential black-box gap of Simon's own problem. For cyclic $G = \mathbb{Z}_N$ the problem is the dihedral hidden subgroup problem, the matching quantum algorithm is Kuperberg's subexponential sieve [6], and the gap is subexponential. Neither regime is run in this paper.

2.3 Consequence for this paper

No claim of exponential quantum query advantage is made for the demonstrated instances. The remaining contributions are the structural noise-resilience mechanism (Theorem 3.1, and its dominance-ratio algebra is formalized in Lean, with the concentration step carried as an explicit axiom) and the zero T-count of the degree-2 oracles. The query and cryptographic-transfer discussions in the main text are scoped accordingly.

3 S3: Proof of the Noise Amplification Theorem

Theorem 3.1 (Majority-vote recovery under a probability margin, full statement)**.** *Consider one round of the hidden shift algorithm executed with S independent shots. Let $p(\eta)$ be the probability that a single shot produces the outcome η, let $F = p(\xi^*)$ be the probability of the correct outcome ξ^*, and let $\Delta := F - \max_{\eta \neq \xi^*} p(\eta) > 0$ be the margin of the correct outcome over its best rival. Then the probability that ξ^* is the most frequent outcome satisfies*

$$P(\xi^* \text{ is mode}) \;\geq\; 1 - 2^n \exp\left(-\frac{S\Delta^2}{4F + 2\Delta}\right),$$

and $S = (4F + 2\Delta)\,\Delta^{-2}\left(n \ln 2 + \ln(1/\delta)\right)$ shots suffice for failure probability at most δ. This is the main text's Theorem 2. Lemma 20.6 (Section S20) is the same statement in that section's notation.

Proof. Fix a wrong outcome η and set $X_t = \mathbf{1}[x_t = \xi^*] - \mathbf{1}[x_t = \eta] \in \{-1, 0, 1\}$, with mean $\Delta_\eta = F - p(\eta) \geq \Delta$ and variance at most $F + p(\eta) \leq 2F$. The mode can fail to η only if $\sum_t X_t \leq 0$, a deviation of $S\Delta_\eta$ below the mean. Bernstein's inequality bounds this by $\exp\big(-S\Delta_\eta^2/(2(F + p(\eta)) + \frac{2}{3} \cdot 2\Delta_\eta)\big) \leq \exp\big(-S\Delta^2/(4F + 2\Delta)\big)$. A union bound over the $2^n - 1$ wrong outcomes gives the bound, and solving $2^n \exp(-S\Delta^2/(4F + 2\Delta)) \leq \delta$ for S gives the budget. □

Remark 3.2 (Hoeffding form)**.** *Hoeffding's inequality in place of Bernstein's bounds the same deviation by $\exp(-S\Delta_\eta^2/2)$, so $P(\xi^* \text{ is mode}) \geq 1 - (2^n - 1)\exp(-S\Delta^2/2)$ and $2\Delta^{-2}(n \ln 2 + \ln(1/\delta))$ shots suffice, weaker than the theorem by the factor $2/(4F + 2\Delta)$ in the exponent whenever $2F + \Delta < 1$. Every shot budget quoted in this paper uses the theorem's bound, and Table 16 (Section S20) lists both.*

Corollary 3.3 (Uniform redistribution, full statement)**.** *If $F > 2^{-n}$ and each of the $2^n - 1$ incorrect outcomes has probability at most $(1 - F)/(2^n - 1)$, then $\Delta \geq F - (1 - F)/(2^n - 1) > 0$ and Theorem 3.1 applies with that margin. In the Hoeffding form of the remark this reads*

$$P(\xi^* \text{ is mode}) \;\geq\; 1 - (2^n - 1)\exp\left(-\frac{S\big(F - (1 - F)/(2^n - 1)\big)^2}{2}\right),$$

the statement of this theorem in earlier versions of this paper and the form that the multi-output (Theorem 15.1) and Pauli (Theorems 17.1, 19.2) statements specialize to. This is the main text's Corollary 3. The premise is a sufficient condition, not a model of the device (Section S8 measures the departure).

Corollary 3.4 (Concrete bound at illustrative parameters)**.** *At $n = 24$ with the 2Q-model parameters $F = 0.254$, $S = 256$ (model illustration; the measured frontier values are $F \sim 10^{-3}$ at $S = 16{,}384$, main-text Table 29):*

$$F - \frac{1 - F}{2^n - 1} = 0.254 - 4.45 \times 10^{-8} \approx 0.254.$$

$$P(\textit{error per circuit}) \le 2^{24} \exp\left(-\frac{256 \times 0.254^2}{4 \times 0.254 + 2 \times 0.254}\right) = 1.68 \times 10^7 \cdot e^{-10.8} \approx 3.3 \times 10^2,$$

which exceeds 1 *and is therefore vacuous from the Bernstein-plus-union bound alone (the Hoeffding form of the remark gives* 4.4×10^3*). However, the model's expected count gap (*$\mu_c \approx 65$ *vs.* $\mu_w \approx 10^{-5}$*) is so extreme that the exact binomial calculation gives* $P(\textit{error}) < 10^{-100}$ *(see the proof of Proposition 3.5 below).*

Proposition 3.5 (Tight bound via expected counts)**.** *When* $\mu_c = S \cdot F \gg 1$ *and* $\mu_w = S \cdot (1 - F)/(2^n - 1) \ll 1$*, the probability that any wrong outcome receives even one count is*

$$P(\exists\, \eta : W_\eta \ge 1) \le (2^n - 1)(1 - (1 - p_w)^S) \le (2^n - 1) \cdot S \cdot p_w = S(1 - F),$$

and the probability that any wrong outcome with ≥ 1 *count exceeds* μ_c *counts is superexponentially small. For our parameters, the expected number of wrong outcomes receiving* ≥ 1 *count ranges from 104 (*$n = 8$*) to 191 (*$n = 24$*), but each receives at most 1–2 counts versus* $\mu_c = 65$–891 *for the correct outcome.*

4 S4: Shift Identity

Proposition 4.1 (Shift identity)**.** *For* $f\colon \mathbb{Z}_2^n \to \{\pm 1\}$ *and* $g(x) = f(x \oplus s)$*:*

$$\hat{g}(\xi) = (-1)^{s \cdot \xi}\, \hat{f}(\xi) \quad \textit{for all } \xi \in \mathbb{Z}_2^n.$$

Proof.

$$\begin{aligned}
\hat{g}(\xi) &= \sum_{x \in \mathbb{Z}_2^n} f(x \oplus s)(-1)^{x \cdot \xi} \\
&= \sum_{y \in \mathbb{Z}_2^n} f(y)(-1)^{(y \oplus s) \cdot \xi} \quad (y = x \oplus s) \\
&= (-1)^{s \cdot \xi} \sum_{y \in \mathbb{Z}_2^n} f(y)(-1)^{y \cdot \xi} \\
&= (-1)^{s \cdot \xi}\, \hat{f}(\xi).
\end{aligned}$$

□

5 S5: Classical Information-Rate Bound in the Hidden-Injective-Oracle Model

The separation below holds only in the *hidden-injective-oracle* model of Section S2 (Theorem 2.2). For public f the classical level-set recovery (Proposition 2.1) extracts $\Theta(1)$ bits per query, so the quantum and classical per-query information rates coincide up to constants. The exponential information-rate gap is a property of the hidden-oracle regime, where it is the information-theoretic counterpart of the collision lower bound.

Two statements hold, one in each access model. In the *public* model, each quantum round of Algorithm 1 produces (ξ_i, b_i) with $b_i = s \cdot \xi_i \pmod 2$. Since ξ_i is drawn from a full-support distribution, b_i is a near-uniform linear functional of s and carries $\Theta(1)$ bits, and with noise $I_Q \ge 1 - h(p_{\text{eff}}/2)$ for binary entropy h and effective per-round error p_{eff}. The classical level-set recovery extracts $\Theta(1)$ bits per query in the same model, so there is no rate separation for public f. In the *hidden-injective-oracle* model the sign-extraction step is unavailable (it evaluates the public dual

$\hat{f}$), so the quantum rate of *this* algorithm is not defined there, and we claim no quantum rate in that model. The best known quantum solver is Kuperberg's subexponential sieve. What does hold in the hidden model is a classical rate collapse:

Theorem 5.1 (Classical information-rate collapse, hidden-injective-oracle model)**.** *For q adaptive classical queries to a hidden injective oracle on an abelian group of order 2^n, the total mutual information with the hidden shift obeys*

$$I_C(q) \;=\; \mathcal{O}\Big(\frac{q^2}{2^n}\Big) \;\; \textit{bits}, \qquad I_C^{\textit{per query}} \;=\; \mathcal{O}\Big(\frac{q}{2^n}\Big),$$

so any classical solver needs $q = \Omega(2^{n/2})$ queries to collect $\Omega(1)$ bits, matching the collision bound of Theorem 2.2.

Proof sketch. Each classical query x_i returns $g(x_i) = f(x_i \oplus s)$. The mutual information $I(s; g(x_i) \mid x_i, g(x_1), \ldots, g(x_{i-1}))$ is bounded by the probability that $g(x_i)$ distinguishes s from a uniformly random alternative s'. By the collision analysis (Theorem 2.2), this probability is at most $\mathcal{O}(i/2^n)$ after i queries, giving total information $\sum_{i=1}^{q} \mathcal{O}(i/2^n) = \mathcal{O}(q^2/2^n)$ bits. □

We do not state a quantum-to-classical rate *ratio* in the hidden model: the numerator would require a quantum algorithm with constant information rate against hidden injective oracles, and no such algorithm is known (Kuperberg's sieve extracts subexponentially little per query at comparable total cost). The information-theoretic content of this section is the collapse of the classical rate. It establishes no separation.

6 S6: Quadratic Oracle Decomposition

For $f(x) = \bigoplus_{i<j} A_{ij} x_i x_j \oplus \bigoplus_i c_i x_i$ over $\mathbb{F}_2$, the phase oracle $O_f \colon |x\rangle \mapsto (-1)^{f(x)}|x\rangle$ decomposes as:

$$O_f = \prod_{\substack{i<j \\ A_{ij}=1}} \mathrm{CZ}_{i,j} \;\cdot\; \prod_{\substack{i \\ c_i=1}} Z_i.$$

Both CZ and Z are Clifford gates on superconducting hardware.

For the shifted oracle $g(x) = f(x \oplus s)$:

$$\begin{aligned} g(x) &= \bigoplus_{i<j} A_{ij}(x_i \oplus s_i)(x_j \oplus s_j) \oplus \bigoplus_i c_i(x_i \oplus s_i) \\ &= \bigoplus_{i<j} A_{ij} x_i x_j \;\oplus\; \bigoplus_i \Big(c_i \oplus \bigoplus_j A_{ij} s_j\Big) x_i \;\oplus\; \text{const}. \end{aligned}$$

The quadratic part is unchanged. The linear part becomes $c'_i = c_i \oplus \bigoplus_j A_{ij} s_j$. The constant term is a global phase and does not affect measurement. Thus O_g has the same CZ structure as O_f with modified Z gates.

Combined-oracle implementation (degree 2). The degree-2 hardware experiments use the combined-oracle circuit $H^{\otimes n} - O_f O_g - H^{\otimes n}$. All CZ gates of O_f and O_g cancel pairwise (same A), leaving only Z gates with mask $c_i \oplus c'_i = \bigoplus_j A_{ij} s_j = (As)_i$. The Hadamard sandwich maps the residual Z-mask to a deterministic basis-state outcome $\xi = As \bmod 2$, so a single shot yields a linear constraint on s (Algorithm 1). For higher ANF degree, the combined-oracle residual additionally

contains CCZ/CZ gates from the shift expansion of the high-degree terms, and the recovery via $s = A^{-1}\xi$ becomes statistically fragile when only a single high-degree monomial is present (the residual's Walsh–Hadamard structure is symmetric enough that the mode of the noisy output drifts off As by a non-negligible margin, and we documented this empirically on processors D, A and E during method development). The existing degree-3 hardware results in S11 use *multiple* cubic monomials, which break this symmetry. The unbounded-degree single-monomial extension in S14 therefore uses the full Rötteler structure $H^{\otimes n} - O_g - H^{\otimes n} - O_{\overline{F}} - H^{\otimes n}$ instead, in which the noiseless output is $|s\rangle$ directly and mode-shift errors do not propagate through an inversion.

Algorithm 1 Combined-oracle hidden shift recovery (hardware implementation)

Require: quadratic part $A \in \mathbb{F}_2^{n\times n}$ of f; oracle O_g with shift s; round count Q; shots S
Ensure: hidden shift $s \in \mathbb{F}_2^n$

build the combined-oracle circuit $\mathcal{C}\colon |0^n\rangle \xrightarrow{H^{\otimes n}} \xrightarrow{O_f O_g} \xrightarrow{H^{\otimes n}}$ ▷ CZ gates cancel pairwise; only Z residual remains for degree-2
$M \leftarrow \varnothing$
for $i = 1, \ldots, Q$ **do**
    execute $\mathcal{C}$ for S shots; collect outcomes $\{\xi_{i,t}\}_{t=1}^{S} \subset \mathbb{F}_2^n$
    $\xi_i \leftarrow \mathrm{mode}_t(\xi_{i,t})$ ▷ Theorem S3.1 (S3)
    $M \leftarrow M \cup \{\xi_i\}$
end for
stack $\{\xi_i = As \bmod 2\}$ and solve by Gaussian elimination over $\mathbb{F}_2$ (rank-deficient batches: stack more constraints until rank $= n$)
return s

7 S7: Degree-3 Oracle and T-Count

For a degree-3 term $x_i x_j x_k$, the corresponding phase gate is $\mathrm{CCZ}_{i,j,k}$, the doubly-controlled Z gate. In the standard Clifford+T decomposition:

$$\mathrm{CCZ} = (6\ \mathrm{CNOT}) + (4\ T) + (3\ T^\dagger),$$

contributing 7 non-Clifford gates per cubic term. A degree-3 function with m cubic terms has T-count $7m$ per oracle call and $7mQ$ total.

This circuit lies outside the stabilizer (Gottesman–Knill) class: the CCZ gates are non-Clifford, so it is not simulable by stabilizer-only methods. This does not by itself imply classical intractability. These instances remain classically easy (polynomial-time shift recovery), and their matrix-product-state bond dimensions are measured directly later in this supplement (S11). The claim here is only that the degree-≥ 3 oracle exercises non-stabilizer resources, unlike the Clifford degree-2 oracle.

Hardware validation. We executed the degree-3 hidden shift algorithm at $n = 8$ and $n = 12$ with $m = 2$ cubic terms (2 CCZ = 14 T-gates per circuit) on processor D and processor F, achieving 20/20 success (Table 2). The average transpiled depth of ∼113–156 reflects the additional CCZ decomposition overhead compared to the Clifford circuits (depth ∼39 at $n = 8$).

Table 2: Early degree-3 (2-CCZ) validation runs at $n = 8, 12$.

n	Cubic terms	CCZ	Runs	Success	Depth	Backend
8	(1,6,7),(4,6,7)	2	10	10/10	119 ± 10	D
8	(1,6,7),(4,6,7)	2	5	5/5	113 ± 8	F
12	(1,6,7),(4,6,7)	2	5	5/5	156 ± 7	F

8 S8: Noise Model Details

The single-shot fidelity F is estimated from the hardware gate error rates as follows. Each qubit participates in $(n-1)$ CZ gates (in the worst case for a complete graph oracle). The per-qubit error probability after all CZ gates is

$$q_{\text{qubit}} = 1 - (1 - p_{2Q})^{n-1},$$

where p_{2Q} is the median two-qubit gate error rate. The probability of a correct n-qubit measurement is approximately

$$F \approx (1 - q_{\text{qubit}})^n = (1 - p_{2Q})^{n(n-1)}.$$

This per-qubit convention counts each CZ error at both endpoints, so its exponent $n(n-1)$ is twice the gate count $n(n-1)/2$: it is the square of, and deliberately more pessimistic than, the per-gate form $F_{\text{pred}} = (1 - p_{2Q})^{N_{2Q}}$ that Section S16 tests against measured data (and finds optimistic by about a decade in absolute level). Both are heuristics. Where the two sections quote model fidelities, S16's per-gate convention is the one under test.

n	p_{2Q}	q_{qubit}	F	Backend
8	2.59×10^{-3}	0.018	0.870	G
12	2.48×10^{-3}	0.027	0.721	D
16	2.59×10^{-3}	0.038	0.551	G
16	2.48×10^{-3}	0.037	0.551	D
20	2.59×10^{-3}	0.048	0.389	G
20	2.48×10^{-3}	0.046	0.397	D
24	2.48×10^{-3}	0.056	0.254	D

The fact that $F = 0.254$ at $n = 24$ (only 1 in 4 shots correct) yet the algorithm succeeds 15/15 is entirely explained by the majority-vote theorem: with 256 shots, the correct outcome receives ~ 65 counts while the strongest competitor receives $\sim 10^{-5}$ counts.

Scale-up results ($n = 28$–64). Using $p_{2Q} = 2.5 \times 10^{-3}$ (Heron r2 median), the model predicts the fidelities below. The actual circuit depths are significantly lower than worst-case predictions because the Maiorana–McFarland construction yields oracles with $\sim n^2/4$ CZ gates rather than the maximum $n(n-1)/2$.

n	Depth (actual)	F_{pred}	F_{meas}	Dom. ratio	Group size	Success
28	292	15.1%	0.2%	4.8×10^7	2.7×10^8	9/9
32	398	8.4%	0.2%	3.9×10^8	4.3×10^9	8/8
36	474	4.3%	0.2%	3.1×10^9	6.9×10^{10}	9/9
40	596	2.0%	0.1%	2.2×10^{10}	1.1×10^{12}	9/9
44	718	0.9%	0.05%	1.6×10^{11}	1.8×10^{13}	9/9
48	790	0.4%	0.024%	1.0×10^{12}	2.8×10^{14}	9/9
52	915	0.1%	0.024%	6.2×10^{12}	4.5×10^{15}	9/9
56	1033	0.04%	0.024%	3.2×10^{13}	7.2×10^{16}	9/9
60	1102	0.01%	0.024%	1.6×10^{14}	1.2×10^{18}	9/9
64	1292	0.004%	0.024%	7.6×10^{14}	1.8×10^{19}	9/9

Even at $n = 64$ with fidelity below 0.005%, the dominance ratio exceeds 10^{14}, predicting reliable recovery with ∼8,192 shots *under uniform redistribution of the lost mass*. The measured dual-oracle record (main-text Table 3 and Extended Data Table 29) falsifies that premise quantitatively. At $n = 24$ ($S = 16{,}384$ shots; uniform per-string floor $u = (1 - F)/(2^n - 1)$) the best wrong outcome alone carries 9.2×10^3–1.9×10^4 times u on processor B, where that rival sits at Hamming distance 5–7, and 3.1×10^3–1.5×10^4 times u on processor A (distance 1 in two runs, 7 in three). The strings within Hamming distance 2 of the answer are denser than uniform by far less: over the ten runs the 24 distance-1 strings average 2.1×10^2–5.2×10^3 times u each, the 276 distance-2 strings 2.6×10^1–3.6×10^2, and the 300-string neighborhood together carries 4.1×10^1–7.4×10^2 times its uniform mass, 1.8–2.5 times the answer's own count on processor B and 1.5–9.0 times on processor A. The measured count ratio at the $n = 24$ frontier is a factor ∼3 rather than $\sim 10^6$, and recovery stops at $n \geq 32$. The table above gives the model's prediction and makes no recovery claim. The F_{meas} column and the "Success" column refer to the retired single-oracle control (Section 9), whose F_{meas} is the mode fraction of a uniform distribution (the $1/S$ shot floor), not a fidelity to a correct outcome.

Routing audit (2026-09-10). Pinning an `initial_layout` fixes only the starting positions of the logical qubits. The preset router may still place SWAP-induced two-qubit gates on any coupler of the map it is given, including couplers outside the selected chain. After this was identified program-wide, every hardware job on the account was audited retrospectively: the submitted ISA circuits were retrieved from the runtime, two-qubit gates were counted per physical coupler, and each coupler was scored against the calibration in force at submission (DEAD: any gate on a coupler with reported error ≥ 0.5; FLAGGED: any gate on a coupler with error $> 10^{-2}$; CLEAN otherwise). Every non-clean job of this paper was then classified by the share of its two-qubit error budget sitting on flagged or dead couplers (material above 20%). *Void (dead couplers):* the ten processor H $(14, 6)$ circuits of 2026-05-01 and the fifteen processor F $(16, 7)$ circuits of 2026-05-04 (46–460 CZ per circuit on couplers at error 1.0; excluded from the pooled counts of Section S14); the three exploratory processor A Protocol C cells $n = 100$–154 (main-text Table 3); 25 of the 36 Nighthawk R1 non-Clifford frontier circuits of Section S21 (all nine processor E $n = 100$ circuits and sixteen of eighteen processor I circuits); and the sixteen Nighthawk R1 degree-2 circuits at $n \geq 64$, whose recoveries stand. *Flagged with a material share:* the fifteen processor B $(16, 7)$ and fifteen processor D $(14, 6)$ circuits of 2026-05-04 (couplers up to 0.118, 44–72% of the budget), eleven of the fifteen processor D $(18, 8)$ circuits, the twelve chain-pre-selected configurations of the 2026-06-02 mitigation ablation (couplers up to 6.4×10^{-2}, 17–29%), the Protocol C cells $n \geq 32$ on processor B (20–69%) and $n = 56$, 72, 80 on processor A (23–50%), seven processor E $n = 80$

frontier circuits (5–45%), and two processor I frontier circuits (52–68%). *Flagged with a minor share* (below 20%, and no action beyond this note): ten processor D $(14, 6)$ trials of 2026-05-01 (coupler at 1.1×10^{-2}), five processor H non-Clifford replicates (two gates each), six 2026-05-01 ladder trials in the S16 dataset, eight of the 108 discrimination-test circuits of Section S16 (seven on processor E), the four processor C stage-1 timing circuits (three of 499 gates), and processor A Protocol C $n = 40$ and 64. *Unused:* forty processor B jobs of 2026-07-01 (a superseded confidence-interval run, all canceled unexecuted), two processor E jobs that exceeded the execution-time limit, and about 2,100 processor B and processor A jobs of 2026-05-05 to 05-11 attributed by circuit fingerprint to an $n = 100$ degree-3 family run from a second workstation, none of which feeds any statement here. Every processor C Hidden Matching and discrimination job, the eighteen circuits of the 2026-09-11 vetted-map rerun (Section S21, audited before and after submission), every Heron Hidden Matching job, the α-sweep, the Protocol C cells $n \leq 24$ on both devices and $n = 32$, 48 on processor A, and the $(8, 3)$, $(10, 4)$, $(12, 5)$, $(20, 9)$ ladder cells are clean. Rule applied throughout: a positive result on a flagged circuit (a recovery) is kept, since extra noise can only lower a recovery rate. A negative whose verdict rests on dead or materially flagged circuits is voided or restated on clean data in place. The one negative not reproduced on clean data at the time of the audit (the Nighthawk R1 non-Clifford frontier) was rerun on processor C on a vetted coupling map under a pre-registration filed before submission (2026-09-11), and the collapse reproduced on clean-routed circuits (Section S21). One observation bears on the DEAD class on Nighthawk R1: degree-2 recovery succeeded with 90–96% of the nominal two-qubit error budget on couplers reported at error 1.0, which is consistent with those entries marking uncalibrated rather than physically dead couplers. The negatives on such couplers remain uninterpretable either way. Every driver in this repository now transpiles on a vetted coupling map (couplers above 10^{-2}, dead couplers, and bad qubits removed). It counts two-qubit gates per coupler on the ISA circuits and refuses to submit if any gate sits off the vetted set. It writes a tracked submission manifest (job id, backend, time, circuit fingerprint) before any retrieval.

9 S9: Retired Single-Oracle Degree-2 Campaign (Output-Support Control)

Reclassification note (this section and S10). The 427 runs tabulated here and in Section 10 executed the single-oracle circuit $H^{\otimes n} O_g H^{\otimes n}$ and solved the linear system $\{s \cdot \xi_i = b_i\}$ with the sign bits b_i computed classically from the planted shift. For a bent function the single-oracle output distribution satisfies $|\hat{g}(\xi)|^2/4^n = 2^{-n}$ for every ξ: it is exactly uniform and carries no information about s, and the sign bits are not physically obtainable from this circuit (they require the dual construction, main-text Protocol C). This property is elementary from bentness, but we did not recognize it when the campaign was designed and run: it surfaced in an internal audit of the analysis chain after the campaign completed, at which point we reclassified these runs and designed, registered and executed the dual-oracle ladders of main-text Table 3 as the recovery record (the registration was committed after the processor A ladder ran and before the processor B ladder executed, see S24). "Success" in these tables is therefore precisely the event that the measured outcomes span $\mathbb{F}_2^n$, a property that holds with probability $\approx 1 - 2^{-(Q-n)}$ for uniformly random outcomes. The single failure below (rank 31 at $Q = n = 32$) matches the random-matrix expectation over $\mathbb{F}_2$ [2]. These runs are retained as an *output-support (rank) control*, a weak test that the device's output distribution has not collapsed onto a low-dimensional support, and as a transpiled depth/routing calibration of the oracle CZ network at scale. No recovery claim in the main text rests on them. The degree-2 recovery record is the dual-oracle ladder record of main-text

Table 3, with its registration chronology stated and tabulated in S24. The same reclassification applies to the Nighthawk degree-2 runs of Section S21.

Full data for this control campaign (427 runs), the non-Clifford degree-3 experiments (259 runs, Section 11), and the unbounded-degree runs (S14) is provided in the accompanying JSON files. This section lists the complete control record at $n \leq 64$: 287 runs, every executed run in the repository, deduplicated by job ID and machine-generated from the raw result files (with per-run source-file provenance). The 140 scale-up runs at $n = 68$–154 are tabulated separately in Section 10. The key statistics are:

Table 3: Complete degree-2 hardware record at $n = 8$–64: 287 runs, 286/287 success. The single failure is in the $(n = 32, Q = 32)$ cell, where all three runs reached rank 31. Two resolved the residual bit correctly and one did not.

n	Q	Runs	Success	Mean rank	Mean depth	Backend
8	16	6	6/6	8.0	39	G
8	32	6	6/6	8.0	39	G
8	64	6	6/6	8.0	40	G
12	32	10	10/10	12.0	72	D
12	48	10	10/10	12.0	71	D
12	64	10	10/10	12.0	72	D
16	32	10	10/10	16.0	131	G
16	48	10	10/10	16.0	132	G
16	64	10	10/10	16.0	131	G
16	32	10	10/10	16.0	139	D
16	48	10	10/10	16.0	142	D
16	64	10	10/10	16.0	140	D
20	32	10	10/10	20.0	189	G
20	48	10	10/10	20.0	190	G
20	64	10	10/10	20.0	190	G
20	32	10	10/10	20.0	193	D
20	48	10	10/10	20.0	191	D
20	64	10	10/10	20.0	188	D
24	32	5	5/5	24.0	319	D
24	48	5	5/5	24.0	322	D
24	64	5	5/5	24.0	322	D
28	32	5	5/5	28.0	293	D
28	48	3	3/3	28.0	289	D
28	64	3	3/3	28.0	292	D
32	32	3	2/3	31.0	399	D

continued on next page

Table 3 continued from previous page

n	Q	Runs	Success	Mean rank	Mean depth	Backend
32	48	3	3/3	32.0	399	D
32	64	3	3/3	32.0	396	D
36	48	3	3/3	36.0	474	D
36	64	3	3/3	36.0	472	D
36	96	3	3/3	36.0	477	D
40	48	3	3/3	40.0	599	D
40	64	3	3/3	40.0	595	D
40	96	3	3/3	40.0	593	D
44	48	3	3/3	44.0	719	D
44	64	3	3/3	44.0	716	D
44	96	6	6/6	44.0	715	D
48	64	6	6/6	48.0	792	D
48	96	6	6/6	48.0	790	D
48	128	6	6/6	48.0	794	D
52	64	3	3/3	52.0	914	D
52	96	3	3/3	52.0	911	D
52	128	3	3/3	52.0	920	D
56	64	3	3/3	56.0	1035	D
56	96	3	3/3	56.0	1032	D
56	128	3	3/3	56.0	1033	D
60	96	3	3/3	60.0	1106	D
60	128	3	3/3	60.0	1101	D
60	192	3	3/3	60.0	1099	D
64	96	3	3/3	64.0	1298	D
64	128	3	3/3	64.0	1288	D
64	192	3	3/3	64.0	1291	D
Total		287	286/287			

10 S10: Scale-up Degree-2 Experiments ($n = 68$–154)

An additional 140 degree-2 experiments were executed at $n = 68$–154 on five Heron backends, all achieving full rank and correct shift recovery. Table 4 lists the complete per-backend breakdown, deduplicated by job ID and machine-generated by the same script as Section 9. In total, the 427 degree-2 experiments reported in the main text comprise the 287 runs at $n \leq 64$ in Section 9 plus the 140 scale-up runs below.

Accounting note. One further submission at $(n = 72, Q = 128)$ errored client-side before reaching the device (an HTTPS connection failure during submission, and no job ID was issued) and was

Table 4: Scale-up degree-2 experiments ($n = 68$–154), complete record. All runs achieved full rank. The sign bits were computed from the planted shift, so these runs certify output support and depth tolerance, not recovery (Section S9).

n	Q	Runs	Success	Mean depth	Backend
68	128–256	18	18/18	1398	D
72	128–256	18	18/18	1527	D
76	192–320	18	18/18	1678	D
76	320	2	2/2	1668	B
80	192–320	18	18/18	1853	D
80	192–320	9	9/9	1850	A
80	192–256	6	6/6	1856	B
100	192–320	9	9/9	2607	D
100	256–320	6	6/6	2592	F
100	192	1	1/1	2620	B
120	192–320	10	10/10	3506	D
120	192–256	9	9/9	3497	B
130	192	3	3/3	3924	B
130	192	1	1/1	3929	F
140	192	3	3/3	4462	B
140	192	1	1/1	4500	H
150	192	4	4/4	5018	B
154	192	4	4/4	5307	B
Total		140	140/140		

resubmitted successfully. Unexecuted submissions are not runs and do not enter the record. An earlier version of this manuscript reported a curated subset of 345 runs. The tables in Sections 9 and 10 supersede that count with the complete repository record of 427 executed runs, which includes duplicate-configuration batches and single-run probes that the curated count omitted.

11 S11: Higher-Degree Experiments

To demonstrate that recovery extends beyond degree-2 oracles, we executed the hidden shift algorithm with degree-3 Boolean functions whose oracles require CCZ (Toffoli) gates. Each CCZ decomposes into T-gates, placing these circuits outside stabilizer (Gottesman–Knill) simulation. As the bond-dimension measurements below show, the statement concerns non-stabilizer content only. We claim no classical hardness.

The non-Clifford accounting below is *complete*: every degree-3 multi-CCZ hardware run in the repository is tabulated, including all failures. The table is machine-generated from the raw result files. Each run is assigned a protocol tier from fields recorded in its result file:

- **T0** shakedown / off-protocol: the first-day run (2026-03-23), $Q = 128$ runs at $n \geq 80$ (the protocol uses $Q = 128$ only up to $n = 64$, and at $n \geq 80$ it was superseded by $Q \geq 256$), and

optimization-level-3 runs (the protocol specifies level 2 with noise-aware routing). Excluded from the headline count and listed here for completeness.

- **T1** hardware-optimized baseline: optimization level 2, no clean-topology selection, no dynamical decoupling (includes the small-n runs).
- **T2** clean-topology baseline: clean-topology qubit selection, no dynamical decoupling.
- **T3** chunked XY4 + twirling: XY4 dynamical decoupling and Pauli twirling, the strongest recipe (cf. the pre-registered mitigation ablation).
- **T4** pre-registered replicates.

The headline count in the main text pools tiers T1–T4: **156/259**. Per-tier subtotals are T1 83/118, T2 37/100, T3 10/11, and T4 26/30 (T0, excluded: 2/19). The tiers are not comparable head-to-head: T2 contains most of the deep-CCZ frontier probes, and T3 is the strongest recipe applied at the deepest cells. The recipe dependence is part of the benchmark's readout. Dynamical decoupling is the load-bearing mitigation factor in the pre-registered ablation (+0.70 main effect on top-1 recovery).

Table 5: Complete degree-3 (multi-CCZ) hardware record: every run in the repository, by cell, backend, and protocol tier.

n	CCZ	Backend	Tier	Succ.	Trials
8	2	F	T1	5	5
8	2	F	T1	8	10
8	2	D	T0	2	5
8	2	D	T1	10	10
12	2	F	T1	5	5
12	2	F	T1	10	10
20	2	D	T1	5	5
32	2	D	T1	5	5
48	2	D	T1	5	5
48	2	D	T1	5	5
48	10	B	T2	1	3
48	12	B	T2	0	2
64	2	D	T1	3	5
64	2	D	T1	2	5
64	10	B	T2	0	3
80	2	A	T0	0	5
80	2	A	T1	4	5
80	4	A	T1	5	5
80	6	B	T2	1	3
80	6	A	T0	0	1
80	6	A	T1	4	5
80	8	B	T2	1	3
80	8	H	T4	10	10
80	10	B	T2	0	3

Table 5 (continued)

n	CCZ	Backend	Tier	Succ.	Trials
80	10	H	T2	0	3
80	12	B	T2	0	3
80	12	H	T2	0	3
90	8	B	T2	2	3
100	2	B	T1	0	2
100	2	B	T2	5	5
100	2	A	T0	0	3
100	2	A	T1	1	4
100	2	A	T2	3	3
100	2	D	T0	0	5
100	2	D	T1	2	5
100	2	H	T1	4	5
100	4	B	T2	3	3
100	4	B	T2	4	5
100	4	H	T2	0	3
100	6	B	T2	0	3
100	6	B	T2	1	3
100	6	F	T3	1	1
100	6	A	T3	3	3
100	6	H	T2	0	3
100	6	H	T4	10	10
100	8	B	T2	0	3
100	8	B	T2	0	3
100	8	B	T2	0	3
100	8	D	T3	0	1
100	8	H	T2	0	3
100	10	B	T3	1	1
108	6	A	T3	1	1
112	2	B	T2	4	5
116	2	B	T2	5	5
116	6	B	T2	1	3
120	2	H	T1	0	5
120	2	H	T2	0	3
120	6	B	T2	0	3
124	8	B	T2	0	3
126	2	A	T1	0	4
126	2	D	T1	0	3
126	2	H	T1	0	5
130	2	B	T2	3	3
136	8	B	T2	0	3
140	2	B	T2	3	3
140	6	B	T3	1	1
140	8	B	T3	1	1
140	8	B	T3	1	1
140	10	B	T3	1	1

Table 5 (continued)

n	CCZ	Backend	Tier	Succ.	Trials
140	10	H	T4	6	10
Tier T0 subtotal				2	19
Tier T1 subtotal				83	118
Tier T2 subtotal				37	100
Tier T3 subtotal				10	11
Tier T4 subtotal				26	30
Protocol total (T1–T4)				**156**	**259**
Grand total (incl. T0)				158	278

First, the $(100, 6)$ processor A entry uses nonlocal cubic-term placement, a deliberately harder routing variant. All other cells use local cubic terms. Second, the $n = 120$ and $n = 126$ failures (0/20) ran exclusively on processor H, processor A, and processor D, while the $n = 112$–140 2-CCZ successes ran on processor B with clean topology. This backend dependence at depth is the effect the discrimination test of S16 quantifies. Third, the repository also contains deep-genuine *degree-2* control runs from the same campaign (for example $n = 126$: 1/5 on processor A, 5/5 on processor D, 2026-04-13). These are controls for the degree-3 circuit family and belong to neither the retired 427-run single-oracle control campaign (Sections S9–S10) nor the 259-trial degree-3 record.

A worst-case entanglement-injection argument suggests $\chi \gtrsim 2^{2k}$ as a *lower bound* for faithful matrix-product-state (MPS) simulation of the k-CCZ circuits. Direct measurement (paragraph "Measured MPS bond" below) shows that the k-dependence of this bound is not realized by the hardware-optimized construction used here.

Measured MPS bond. We measured the bond dimension of these deep-genuine circuits directly (quimb `CircuitMPS`, qubits in the interleaved local ordering). The sharpest probe fixes a singular-value truncation tolerance (10^{-8}) and reads off the bond a faithful MPS then requires, with no bond cap, so the measurement is not cap-limited. Table 6 reports the result across register size n and CCZ count k. First, the bond is nearly *flat in* k: at $n = 24$ it is 2048 for $k = 2, 4, 6, 8$ and only 2512 at $k = 10$, and at $n = 26$ it ranges 2048–3160 across $k = 2$–8 (flat to within 1.6×, against the 4096× spread the 2^{2k} bound asserts over the same range, and the $k = 10$ cell at $n = 26$ exhausted memory and is not reported). The per-term multiplication asserted by the entanglement-injection argument is not realized by this construction. Second, the bond grows with *register size*: 24 at $n = 12$, 384 at $n = 20$, 2048 at $n = 24$, 2048–3160 at $n = 26$, consistent in order of magnitude with a generic central-cut scrambling cost of $\Theta(2^{n/2})$ (the measured values sit at 0.25–0.5 of $2^{n/2}$ at $n = 24$–26, and the $n = 24 \to 26$ step is 1.0×–1.5×, below the 2× that pure $2^{n/2}$ growth would give, so $2^{n/2}$ should be read as the trend's scale, not a fitted rate). Together, these two facts show that the $\chi \gtrsim 2^{2k}$ figure overstates the k-dependence (the bond does not multiply per CCZ term), but the circuits are *not* asymptotically MPS-tractable either, since the bond keeps growing with n. At the demonstrated sizes ($n \leq 26$) the required bond is in the low thousands and the circuits remain MPS-simulable. The 2^{2k} worst-case entries are loose bounds with the wrong k-dependence, not a per-term simulability wall. The classical easiness of recovering the shift is established separately and algebraically (Proposition 6 of the main text), independent of any MPS bond.

Table 6: Measured MPS bond dimension (singular-value cutoff 10^{-8}, no bond cap, interleaved ordering) versus the worst-case lower bound 2^{2k}. The required bond is nearly flat in the CCZ count k and grows with register size n on the scale of $2^{n/2}$ (at 0.25–0.5 of that value at $n = 24$–26). The 2^{2k} dependence on k is not realized. Dashes: $k > n/2 - 1$ (out of range); m: out of memory (exact, cap-free run).

n	k=2	k=4	k=6	k=8	k=10	$2^{n/2}$
12	24	24	–	–	–	64
20	256	384	384	384	–	1024
24	2048	2048	2048	2048	2512	4096
26	2048	2632	2560	3160	m	8192
worst-case bound 2^{2k}:						
	16	256	4096	65536	1.0×10^6	

12 S12: Lean 4 Formalization

The formalization, deposited with the code under the DOI 10.5281/zenodo.23086962 (see Code availability), machine-checks the algebraic core of the structural and noise-resilience results. We state its scope: the quantum query upper bound and the classical query bounds are proved analytically in Sections S1–S2 and are *not* part of the Lean development. The development contains no `sorry` placeholders and no trivial-surrogate "theorems". Statements that are standard but not reproved (the high-probability spanning conclusion, the majority-vote concentration step, and the external facts of the oracle-hiding and injective-oracle arguments) are carried as explicit axioms. The one arithmetic fact retained in `MainTheorem` ($4n+1 < 2^{n/2}$ for $n \geq 12$, by induction) is labeled as a numerical sanity check, not as a proof of the corresponding algorithmic statement.

The following are formally verified:

- full proof that $\hat{g}(\xi) = (-1)^{s\cdot\xi}\hat{f}(\xi)$, including the XOR involutivity lemma and dot-product linearity over $\mathbb{F}_2$.
- a shift preserves the quadratic coefficients of a degree-2 oracle, so the CZ count is shift-invariant. The coefficient-level invariance holds by definition of the formalized coefficient shift. The machine-checked content behind it is the $\mathbb{F}_2$ product expansion $(x+s)(y+t) = xy+$lower order together with the evaluation-level correctness theorem `shift_eval_correct`, $(f.\text{shift}\, s)(x) = f(x \oplus s) \oplus f(s)$, which certifies that the formal coefficient shift computes the shifted function up to the global-phase constant. The zero T-count itself is a property of the $\{\text{CZ}, Z\} \subset$ Clifford gate set, recorded here rather than machine-checked (the development carries no gate-level model).
- the dominance ratio equals $F \cdot (2^n - 1)/(1 - F)$, its monotonicity in F, and the lower bound $\alpha^2(2^n - 1)/(1 - \alpha^2)$ for α-flat functions.
- the function-level coset-constancy reductions between hidden-shift oracles and dihedral-HSP-style paired oracles, in both directions (`shift_to_hsp_char2`, `hsp_to_shift_char2`), and preservation of spectral flatness under hidden shift (`bent_shift_flat`). The dihedral group operation on $\mathbb{F}_2^n \rtimes \mathbb{Z}_2$ is *not* formalized (the carrier is a bare pair type), so what is machine-checked is the oracle-level correspondence, not a group-theoretic equivalence of the two problems.

- the structural ingredient that a proper subspace of $\mathbb{F}_2^n$ has at most 2^{n-1} elements, the bound $\mu \leq 1/2^n$ for minimum-probability distributions, and the geometric-series bound $\sum_{k=0}^{n-1}(1/2)^k < 2$. The full high-probability spanning claim is carried as an explicit axiom (`spanning_whp_from_min_prob`), the analytic Lemma of Section S1.

- the phase-cancellation step (the final-Hadamard phase argument $\mathrm{dotBV}(s+s)\,\xi = 0$ in characteristic 2, so every Walsh phase collapses to $+1$) and the exact recovery probability $P(s) = \bar{\alpha}^2$ (Proposition 7 of the main text) *conditional on the analytically derived amplitude expression* $2^{-3n/2}\sum_{\xi}(-1)^{(s\oplus y)\cdot\xi}|\hat{f}(\xi)|$, which is proved in the main text from the shift identity but is not itself formalized (the development carries no circuit semantics), together with $\bar{\alpha} \geq \alpha$ (mean spectral magnitude dominates the minimum).

The project builds on Lean 4 v4.29.0 with Mathlib. Standard statements that we do not reprove are carried as explicit Lean axioms (eight in total), each an `opaque` proposition so that none is provable by `trivial`. Two facts about the logical architecture deserve emphasis. First, *no theorem in the development depends on any of the eight axioms*: because each axiom's statement is an opaque proposition, it cannot be used in a proof. The axioms are standalone records of what was not formalized, not hypotheses of the machine-checked results, and `#print axioms` on every theorem confirms this. Second, one of the eight, `hiding_from_three_assumptions`, is the oracle-hiding conclusion itself, recorded as an assumption. It is not derived from the other axioms and should be read as marking the entire hiding argument as unformalized. The eight in full: the Hoeffding-plus-union concentration step of the majority-vote bound; the high-probability spanning conclusion; the collision-set density and the classical union bound of the injective-oracle information-rate argument; the NP-hardness of MQ over $\mathbb{F}_2$; the ANF leading-coefficient shift-invariance; the shifted linear-coefficient degree bookkeeping; and the bundled oracle-hiding implication just described.

Status audit. The development has zero `sorry` placeholders and no trivial-surrogate propositions: the earlier placeholder "theorems" for the query bounds have been removed, and the statements they stood in for are either proved, carried as explicit axioms, or left to the analytic Supplementary sections. Table 7 records the status of each result.

13 S13: Taxonomy of Problems Reducible to Hidden Shift

Problems of independent cryptographic and computational interest reduce to the hidden shift problem (HSP) on finite abelian groups. We catalog the established reductions below.

Two kinds of hidden shift. The reductions in this section produce two structurally distinct hidden-shift instances:

(A) Boolean Hidden Shift on $\mathbb{F}_2^n$ (Gavinsky–Rötteler–Roland [7]). An α-flat function $f\colon \mathbb{F}_2^n \to \{\pm 1\}$ is given by an oracle, together with a shifted oracle $g(x) = f(x \oplus s)$. Main-text Theorems 5 and 2 apply directly. The hardware experiments in this paper, including the unbounded-degree ladder of S14, are concrete instances of this form.

(B) Abelian-group hidden shift on $\mathbb{Z}_N$ (or $G \rtimes \mathbb{Z}_2$, the dihedral form). An injective oracle $f_0\colon G \to S$ is given together with $f_1(x) = f_0(x+s)$. Injectivity on an abelian group implies that the relevant (multi-output) Walsh transform is $\alpha = 1$-flat in the Plancherel sense. The cryptographic reductions below (lattice, isogeny, Even–Mansour) all produce instances of this second form.

Table 7: Lean status audit. "Machine-checked": the stated proposition is the intended one and is proved from Mathlib (plus, where noted, the named axioms). "Axiom": carried as an explicit Lean `axiom` for a standard external fact. The quantum query upper bound and the classical query bounds are analytic (Sections S1–S2) and are not part of the Lean development. No trivial-surrogate "placeholder" theorems remain.

Result (file)	**Status**
Shift identity (`ShiftIdentity`)	machine-checked
Shift-invariant quadratic/CZ structure + eval-level shift correctness `shift_eval_correct` (`ZeroTCount`)	machine-checked
Dominance ratio, $\Theta(\alpha^2 2^n)$ bound, monotonicity (`NoiseAmplification`)	machine-checked, unconditional (concentration step is a separate standalone axiom)
Function-level DHSP-style reductions, flatness preserved; no group structure formalized (`DHSPEquivalence`)	machine-checked
Subspace bound $\leq 2^{n-1}$, geometric series (`SpanningLemma`)	machine-checked
Phase cancellation, $P(s) = \bar{\alpha}^2$ conditional on the analytic amplitude expression, $\bar{\alpha} \geq \alpha$ (`RecoveryAmplitude`)	machine-checked
Measurements determine the shift (`InformationRate`)	machine-checked
Gap inequality $4n{+}1 < 2^{n/2}$ (`MainTheorem`)	machine-checked arithmetic; not an algorithmic separation
High-probability spanning (`SpanningLemma`)	axiom; analytic in S1
Majority-vote concentration / Hoeffding (`NoiseAmplification`)	axiom
Collision-set density, classical union bound (`InformationRate`)	axiom; hidden-oracle model (S2)
MQ NP-hardness, ANF shift-invariance, oracle-hiding (`OracleHiding`)	axiom

Concrete worked example. The released code carries out the Kuwakado–Morii reduction explicitly at $n = 8, 10, 12$: build a small Even–Mansour cipher $E(x) = \pi(x \oplus k_1) \oplus k_2$, define $f(x) = E(x) \oplus \pi(x)$, verify $f(x \oplus k_1) = f(x)$, and recover k_1 from a Simon-style linear system on the measured queries. The demo recovers k_1 in 15/15 noiseless trials at every size tested. The Boolean projections of f have $\alpha = 0$ in the strict sense (the Walsh support is a coset by Simon's lemma), so the algorithm specialized to flavor (A) is not directly applicable. The demo runs the algorithm in flavor (B). The main-text noise-amplification theorem's quantitative noise-amplification bound is stated for flavor (A). An analogous bound for flavor (B) follows from the same Chernoff-style argument applied to the multi-output measurement distribution and is stated as an open extension.

13.1 Equivalence with the dihedral hidden subgroup problem

Let G be a finite abelian group. The *hidden shift problem* on G asks: given oracle access to two injective functions $f_0, f_1 \colon G \to S$ satisfying $f_1(x) = f_0(x + s)$ for an unknown shift $s \in G$, find s. The *hidden subgroup problem* (HSP) on the generalized dihedral group $G \rtimes \mathbb{Z}_2$ (where $\mathbb{Z}_2$ acts by inversion $x \mapsto -x$) asks to identify the hidden order-2 subgroup $H = \{(0, 0), (s, 1)\}$ from an oracle function constant on H-cosets.

Proposition 13.1 (Equivalence)**.** *The hidden shift problem on G and the HSP on $G \rtimes \mathbb{Z}_2$ are polynomial-time equivalent.*

Proof sketch. Given hidden shift oracles f_0, f_1, define $F(x, b) = f_b(x)$ on $G \rtimes \mathbb{Z}_2$. Then F is constant on cosets of $H = \{(0, 0), (s, 1)\}$, reducing hidden shift to HSP. Conversely, from an HSP oracle on $G \rtimes \mathbb{Z}_2$, one obtains coset states that encode the shift s. □

When $G = \mathbb{Z}_N$ this is the classical dihedral HSP (DHSP). The equivalence is standard and used implicitly in Ettinger–Høyer–Knill [8] and explicitly in Friedl et al. [9], who proved it for arbitrary finite abelian groups. Kuperberg's sieve algorithm [6] solves the DHSP (equivalently, hidden shift on $\mathbb{Z}_N$) in time and query complexity $2^{O(\sqrt{\log N})}$.

13.2 Lattice problems

Proposition 13.2 (Regev [10])**.** *$\Theta(n^{5/2})$-unique-SVP in n-dimensional lattices reduces in quantum polynomial time to the DHSP on D_N for $N = 2^{poly(n)}$.*

The reduction proceeds in two stages: (1) unique-SVP is reduced to a "two-point problem" where each cell of a lattice partition contains exactly two related points, and (2) the two-point problem is reduced to the dihedral coset problem (DCP), the quantum-state formulation of the DHSP. Since DHSP is equivalent to hidden shift on $\mathbb{Z}_N$ (Proposition 13.1), an efficient hidden shift solver breaks unique-SVP.

Regev [11] extended this chain: worst-case lattice problems (GapSVP, SIVP with polynomial approximation factors) reduce quantumly to the Learning with Errors (LWE) problem, and LWE in turn reduces to the DCP. The full chain is:

$$\text{GapSVP/SIVP} \xrightarrow{\text{quantum}} \text{LWE} \xrightarrow{\text{quantum}} \text{DCP} \equiv \text{Hidden Shift on } \mathbb{Z}_N.$$

This connects hidden shift solvers to the hardness assumptions underlying lattice-based cryptography (NTRU, Kyber/ML-KEM, Dilithium).

13.3 Isogeny-based cryptography

Proposition 13.3 (Childs–Jao–Soukharev [12]). *Let E_1, E_2 be ordinary elliptic curves over $\mathbb{F}_p$ with the same endomorphism ring $\mathcal{O}$. Computing an isogeny $E_1 \to E_2$ reduces in polynomial time (assuming GRH) to a hidden shift problem on the ideal class group* $\mathrm{Cl}(\mathcal{O})$.

The class group $\mathrm{Cl}(\mathcal{O})$ acts freely and transitively on isomorphism classes of curves with CM by $\mathcal{O}$. Given the j-invariant map $j\colon \mathrm{Cl}(\mathcal{O}) \to \mathbb{F}_p$ and its shifted version $j'(a) = j(a\cdot[b])$ for unknown $[b] \in \mathrm{Cl}(\mathcal{O})$, recovering $[b]$ is a hidden shift problem on the finite abelian group $\mathrm{Cl}(\mathcal{O})$. Combined with Kuperberg's algorithm, this yields quantum subexponential time.

The same class-group-action framework underlies the CSIDH key exchange [13], where the secret key is a hidden shift element. Bonnetain and Schrottenloher [14] gave concrete attack costs: CSIDH-512 requires approximately 2^{16} quantum evaluations of the class group action oracle.

13.4 Boolean hidden shift

The Boolean hidden shift problem (finding a hidden shift s in $f_s(x) = f(x \oplus s)$ for a known Boolean function $f\colon \mathbb{F}_2^n \to \{-1, +1\}$) is the $G = \mathbb{F}_2^n$ instance of the abelian hidden shift problem. Gavinsky, Roetteler, and Roland [7] showed that for bent functions, $O(n)$ quantum queries suffice regardless of the algebraic degree. On the classical side we are careful about scope: the $\Omega(2^{n/2})$ collision lower bound proved in Section S2 (Theorem 2.2) applies to *injective* oracles on abelian groups, and a ±1-valued Boolean f is not injective, so that theorem does not literally cover the Boolean black-box model. A birthday-type classical barrier for black-box Boolean bent f is plausible, but we neither prove nor rely on one. No separation claim in this paper rests on the Boolean black-box case. In the public-function model used in our hardware experiments, where f is given explicitly, the shift is classically recoverable in $\mathcal{O}(n/\alpha^2)$ queries (Proposition 6 of the main text, consistent with the polynomial-time classical recovery of Amy and Stinchcombe [3] for bounded degree. The three access models are laid out in the main-text Discussion and its access-model table). Our construction instantiates the bent setting with spectrally flat Boolean functions: the proved query barrier lives in the injective-oracle model, while the noise resilience and oracle hiding hold as stated across models.

13.5 Shifted Legendre symbol

Van Dam, Hallgren, and Ip [15] showed that the shifted Legendre symbol problem (recovering s from oracle access to $\chi(x)$ and $\chi(x+s)$ mod p where χ is the Legendre symbol) reduces to a hidden shift problem over $\mathbb{Z}_p$. The Fourier coefficients of the Legendre symbol are Gauss sums, all of absolute value $\sqrt{p}$, giving $\alpha \approx 1$. The quantum algorithm solves this in polynomial time, whereas classically recovering the shift requires $\Omega(\sqrt{p})$ queries.

13.6 Symmetric-key cryptanalysis

Bonnetain and Naya-Plasencia [16] showed that quantum attacks on the Even-Mansour cipher $E_k(x) = \pi(x \oplus k_1) \oplus k_2$ reduce to a hidden shift problem over $\mathbb{F}_2^n$ (for XOR-based constructions) or $\mathbb{Z}_N$ (for modular-addition-based constructions such as Poly1305). The XOR case is Simon's problem, a special case of hidden shift with a perfectly flat Boolean function ($\alpha = 1$).

13.7 Spectral flatness transfer

Most natural instances of these reductions produce hidden-shift oracles with *perfectly flat* spectra in the appropriate Plancherel sense:

- DHSP coset states are spectrally flat by the representation theory of generalized dihedral groups.
- Injective functions on abelian groups (isogeny problems, Legendre symbol) have flat spectra by unitarity of the discrete Fourier transform.
- Bent Boolean functions are flat in the Walsh–Hadamard sense by definition.

The Boolean Hidden Shift instances (third bullet) lie in flavor (A) and Theorem 3.1 applies as stated. This is the case demonstrated experimentally throughout this paper, including the unbounded-degree extension in S14. The other two bullets produce flavor (B) abelian-group instances. For those, an analogous noise-amplification bound follows from the same Chernoff-style argument applied to the multi-output measurement distribution but is stated here as an explicit open extension rather than proven. The practical bottleneck for the lattice and isogeny instances is not spectral flatness but the oracle circuit depth: evaluating a class group action (CSIDH) or a lattice partition (Regev) requires circuits far deeper than the polynomial oracles in our experiments, putting end-to-end cryptographic-parameter demonstrations into the fault-tolerant-era regime. The α-parametric noise-resilience mechanism is independent of that device-side bottleneck and applies whenever an oracle of the appropriate form is constructed.

Table 8 estimates that device-side gap, from this paper's hardware experiments to a cryptographic-parameter hidden-shift demonstration.

Table 8: Resource gap from this paper's hardware experiments to a cryptographic-parameter quantum hidden-shift demonstration. The algebraic backbone (the α-flat noise-resilience mechanism) applies in all rows. The gap is on the device side. The main-text Discussion points here. Estimates from [11, 14, 17].

Instance	qubits	depth/oracle	T-gates/oracle	status
This paper, $n = 24$, deg-2 dual oracle	24	∼600–720	0	demonstrated
This paper, $n = 12$, $C^{(4)}Z$	12	∼1,100	∼50	demonstrated
Toy CSIDH, $p \sim 257$	∼60	10^4–10^5	$\sim 10^4$	beyond current Heron
CSIDH-512 (one oracle call)	1.4×10^4	$\sim 10^7$	1.5×10^9	fault-tolerant era
CSIDH-512 (full attack)	1.4×10^4	–	$\sim 2^{60}$ calls $\times\, 10^9$	long-horizon FT
LWE NIST-PQC Level 3	$\sim 10^4$	$\sim 10^7$	10^9–10^{11}	long-horizon FT

13.8 Summary of the reduction landscape

Table 9 summarizes the known reductions. The abelian hidden shift problem captures lattice problems, isogeny-based cryptography, and symmetric-key attacks through polynomial-time quantum reductions, all with $\alpha = 1$. Code equivalence and graph isomorphism reduce to *non-abelian* HSP instances on symmetric and wreath-product groups, for which no efficient quantum algorithm is known and standard Fourier sampling provably fails.

Table 9: Problems reducible to hidden shift variants. Abelian reductions (upper block) connect to the spectral flatness framework of this paper. All have $\alpha = 1$. Non-abelian instances (lower block) remain open.

Source problem	**Target group**	**Key reference**	**Reduction**	$\boldsymbol{\alpha}$
DHSP on D_N	$\mathbb{Z}_N$	[8, 9]	Equivalence	1
Unique-SVP	$\mathbb{Z}_N$ via DHSP	Regev [10]	Quantum poly	1
GapSVP / SIVP	$\mathbb{Z}_N$ via LWE	Regev [11]	Quantum poly	1
Isogeny computation	$\mathrm{Cl}(\mathcal{O})$	[12]	Poly (GRH)	1
CSIDH key recovery	$\mathrm{Cl}(\mathcal{O})$	[13, 14]	Poly	1
Legendre symbol	$\mathbb{Z}_p$	[15]	Poly	≈ 1
Even-Mansour cipher	$\mathbb{F}_2^n$ or $\mathbb{Z}_N$	[16]	Poly	1
Boolean hidden shift	$\mathbb{F}_2^n$	[7]	Direct	$\alpha(f)$
Code equivalence	$(GL_k \times S_n) \wr \mathbb{Z}_2$	Dinh–Moore–Russell	Non-abelian	n/a
Graph isomorphism	S_{2n}	Standard	Non-abelian	n/a

14 S14: Unbounded-Degree Validation (M-M with $C^{(m)}Z$, m growing with n)

Amy and Stinchcombe [3] prove polynomial-time classical recovery of the hidden shift for the Maiorana–McFarland benchmarking circuit family of *bounded* algebraic degree, with the running-time exponent depending on degree. The same reference (Sec. 1) explicitly states that the bound fails when the algebraic degree of f_0 grows with n.

To probe this regime experimentally, we instantiate the algorithm with $f(x, y) = \langle x, By\rangle + y_{i_1} \cdots y_{i_d}$, a single Maiorana–McFarland bent function augmented by one degree-d monomial on the y-half. The shifted oracle O_g then contains a top non-Clifford gate $C^{(d-1)}Z$ together with shift-expansion $C^{(k)}Z$ residuals for $k < d - 1$. We use the full Roetteler construction H^{2n}–O_g–H^{2n}–$O_{F\text{-approx}}$–H^{2n}, where $O_{F\text{-approx}}$ is the dual oracle of the underlying *quadratic* bent $\langle x, By\rangle$ (the high-degree contribution to the dual is omitted as a controlled approximation). Recovery is by majority vote on the measured bitstring.

The mode-fraction trend rises monotonically with d because the high-degree perturbation acts on a 2^{-d}-fraction of the truth table, so its contribution to the noiseless output of the quadratic-dual approximation shrinks geometrically. At every point on the ladder the algorithm recovers the planted shift in 15/15 trials.

Hardware execution. We executed the same ladder on three Heron r2 processors (processor H, primary and cleanest per recipe; processor D; processor A) using the full-Rötteler circuit H^{2n}–O_g–H^{2n}–$O_{F\text{-approx}}$–H^{2n} at 8,192 shots per job. Aggregated success rates across 14 trials per row at the small-n points and 12 at $n \geq 18$:

Dedicated processor B runs (2026-06-02). A focused follow-up isolated the cleanest backend (processor B, median 2Q error 1.2×10^{-3}) with XY4 dynamical decoupling and 16,384 shots, six fresh oracle/shift seeds per cell, recovered by per-bit majority vote.

The cleanest backend with the full stack extends partial recovery to $C^{(6)}Z$ ($d = 7$, $n = 16$: 3/6 top-1, per-bit fidelity 0.52, above the 0.50 floor), one ladder step beyond the prior frontier and beyond the 0/50 noise-floor result previously recorded at $(16, 7)$. Degree 8 ($C^{(7)}Z$, $n = 18$) sits at the noise floor (per-bit fidelity 0.501, 8.7 wrong bits out of 18 on average), as does degree 9. Processor

Table 10: Unbounded-degree ladder. Noiseless: 5 random invertible B instances × 3 random shifts per row ($N = 15$). Noisy: 3 instances × 3 shifts ($N = 9$) on Aer density-matrix simulation with depolarizing noise calibrated to Heron r2 median ($p_{1Q} = 2.5\times10^{-4}$, $p_{2Q} = 2.5\times10^{-3}$). Blank cells are noisy points whose density-matrix simulation cost exceeds tractable scope. Transpiled depth and CZ count are with respect to the {CZ, RZ, SX, X} basis at optimization level 2.

				noiseless		noisy (Heron r2)	
n	d	m_{Amy}	top gate	success	mode frac	success / mode frac	depth (CZ)
8	3	2	$C^{(2)}Z$	15/15	0.561	9/9 / 0.525	48 (30)
10	4	3	$C^{(3)}Z$	15/15	0.765	9/9 / 0.657	164 (68)
12	5	4	$C^{(4)}Z$	15/15	0.879	9/9 / 0.650	596 (200)
14	6	5	$C^{(5)}Z$	15/15	0.938	9/9 / 0.498	931 (293)
16	7	6	$C^{(6)}Z$	15/15	0.970	–	1516 (474)
18	8	7	$C^{(7)}Z$	15/15	0.985	–	1188 (390)
20	9	8	$C^{(8)}Z$	15/15	0.993	–	803 (283)
22	10	9	$C^{(9)}Z$	15/15	0.996	–	935 (344)
24	11	10	$C^{(10)}Z$	15/15	0.998	–	3654 (1087)
26	12	11	$C^{(11)}Z$	15/15	0.999	–	4162 (1238)
28	13	12	$C^{(12)}Z$	15/15	1.000	–	8816 (2434)

B's lower error rate therefore gains roughly one ladder step relative to the original 2.5×10^{-3} backends, not the two steps that clean $d \geq 8$ recovery would require: the per-bit fidelity collapses to the random floor between $d = 7$ and $d = 8$, placing the *residual-expansion* frontier at $C^{(6)}Z$. This frontier is set by the compilation, not the algorithm: an exact recompilation of the same oracle moves it further.

Compilation matters: the X-conjugation result (2026-06-03). The residual result above expands the shifted monomial $\prod_i(y_i \oplus s_i)$ into $2^{\text{wt}(s|_T)}$ residual multi-controlled gates. The same oracle is implemented exactly by Pauli conjugation, $O_g = X^s O_f X^s$, where O_f is the unshifted oracle (a single $C^{(d-1)}Z$ plus the quadratic CZ pattern) and X^s are single-qubit bit flips. The two are the identical unitary (verified to machine precision noiselessly). On processor B the conjugation cuts the transpiled two-qubit-gate count, and the recovery frontier moves sharply (Table 13): per-bit recovery is clean (8/8) through $C^{(9)}Z$ ($d = 10$, $n = 22$), with $C^{(7)}Z$ ($d = 8$) lifting from 0/6 at the noise floor to 8/8, and the per-bit fidelity declines smoothly to the 0.50 floor by $C^{(11)}Z$ ($d = 12$). The conjugated frontier is three to four ladder steps beyond the residual frontier ($C^{(6)}Z$).

A reporting note: the full-bitstring mode remained correct (8/8) even at $d = 11, 12$ where per-bit voting degraded, inverting the usual order in which the per-bit statistic is the more robust. At per-bit fidelity ≈ 0.52 the i.i.d. model gives $P(s) \approx 10^{-7}$ per shot, so the mode coinciding with s is not explained by independent noise. Direct measurement of the shot-level error structure confirms the cause: across $C^{(9)}Z$–$C^{(11)}Z$ the per-shot error Hamming weight is *overdispersed* by 1.4–2.1× relative to the independent-bit variance $\sum_i q_i(1 - q_i)$, the mean pairwise bit-flip correlation is positive (+0.02 to +0.05), and the observed exact-recovery probability $P(\hat{x} = s)$ exceeds the i.i.d. product $\prod_i(1 - q_i)$ by a factor of 3×10^3 at $C^{(9)}Z$, rising to 4×10^4 at $C^{(11)}Z$. The errors are positively correlated (crosstalk): they cluster on a subset of shots, leaving enough clean shots that the joint full-bitstring mode stays at s while per-bit voting, which sees only the marginal $q_i \approx 0.48$, degrades. We therefore report the per-bit fidelity, which declines monotonically with the marginal flip rate, as the conservative frontier indicator.

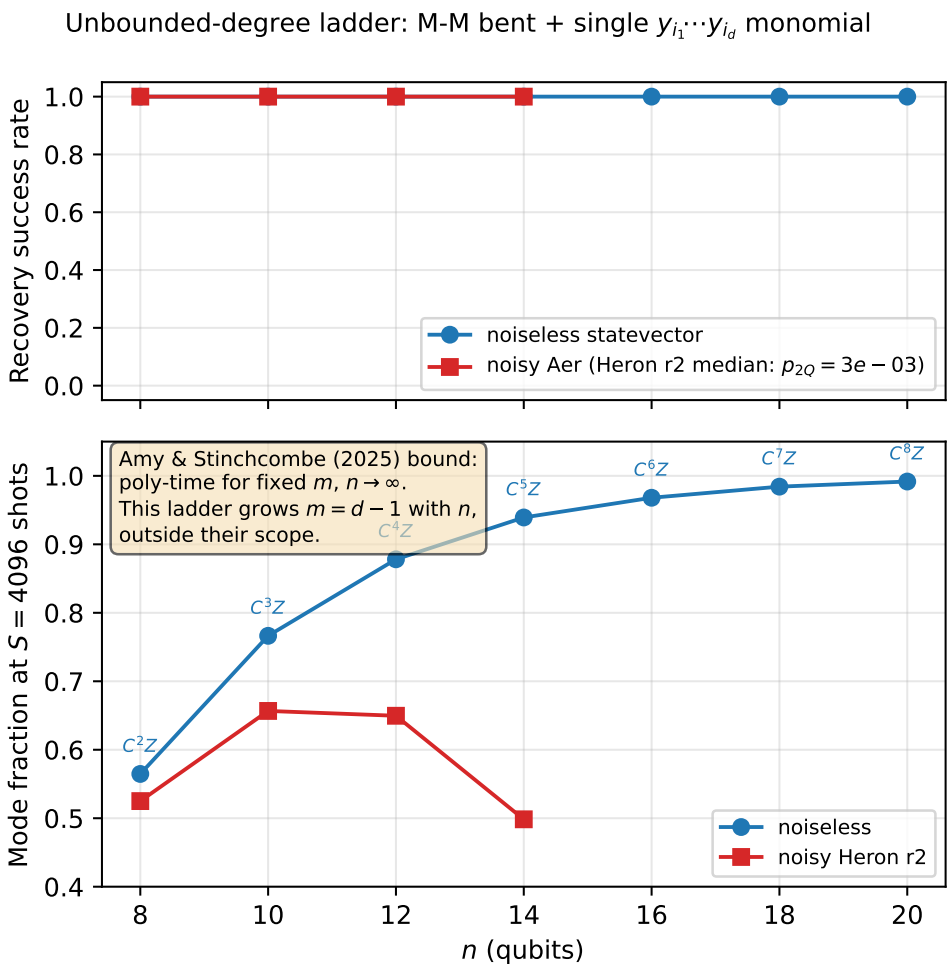


Figure 1: Unbounded-degree ladder in noiseless and noisy Aer simulation. Recovery succeeds at every point under noise calibrated to current Heron r2 device-median fidelities, confirming hardware feasibility. Top-gate annotation traces $m_{\text{Amy}} = d - 1$ growing linearly with n, the regime explicitly outside [3].

This isolates the one assumption the present noise theorems do not cover. Corollary 3.3 (depolarizing) and the i.i.d. Pauli bounds (Theorems 17.1, 19.2) all assume the per-bit errors are independent. The frontier data show the opposite, and in the *favorable* direction: positive correlation makes joint-mode recovery strictly more robust than the i.i.d. bound predicts. Section S20 (Section 20.6) proves the corresponding lower bound on the exact-string probability: positive association implies it is at least the independent-bit product (Theorem 20.5), the measured overdispersion $\kappa > 1$ is the count-level witness of net positive pairwise covariance, a necessary consequence of the association hypothesis (Lemma 20.4), and an explicit burst-mixture regime certifies the joint mode from $\mathcal{O}(n)$ shots at marginal fidelities where per-bit voting fails at every shot count (Proposition 20.7). The measurement above supplies the empirical association witness for that theorem.

The cells use six seeds ($d = 6, 7$) and eight ($d = 8$–12), still modest samples with wide Wilson intervals (6/6 gives $[0.61, 1.0]$). A larger batch would tighten them and sharpen the conjugated frontier. The single-monomial instance is classically easy ($\mathcal{O}(d)$ algebraic recovery), so this demon-

Table 11: Hardware execution of the unbounded-degree ladder on Heron r2. Baseline columns aggregate three backends (processor H, processor D, processor A) without dynamical decoupling. The DD columns are an additional processor H batch with XY4 dynamical decoupling enabled via SamplerV2 options, targeted at the noise- saturated points where DD has the most room to help. Top-1 reports recovery from the most-frequent measurement outcome. Top-K (K=10, pre-registered protocol) reports recovery when the planted shift is among the 10 most-frequent outcomes. ‡The focused batch ran 30 circuits on three backends. Its ten processor H circuits are excluded as routing-contaminated (46–106 of 450–1587 CZ per circuit on coupler $(31, 32)$, reported dead in the calibration in force, see Section S8, routing audit), leaving processor D 6/10 and processor A 7/10.

				baseline		best mitigation		
n	d	m_{Amy}	top gate	top-1	top-K	top-1	top-K	mitigation set
8	3	2	$C^{(2)}Z$	14/14	14/14	–	–	–
10	4	3	$C^{(3)}Z$	12/14	**14/14**	–	–	top-K alone
12	5	4	$C^{(4)}Z$	4/14	6/14	**6/6**	**6/6**	+ DD
14	6	5	$C^{(5)}Z$	3/14	4/14	13/20‡	14/20‡	+ DD + chain + 16k shots
16	7	6	$C^{(6)}Z$	0/14	0/14	0/6	0/6	+ DD
18	8	7	$C^{(7)}Z$	0/12	0/12	–	–	–
20	9	8	$C^{(8)}Z$	0/12	0/12	–	–	–

Table 12: Unbounded-degree ladder on processor B with XY4 dynamical decoupling, 16,384 shots, six seeds per cell (2026-06-02). Top-1 is the strict full-bitstring metric. Per-bit fidelity is the mean per-bit-majority match rate (0.50 is the random floor).

n	d	top gate	top-1	per-bit fid.	mean wrong bits
14	6	$C^{(5)}Z$	4/6	0.597	2.3
16	7	$C^{(6)}Z$	3/6	0.520	3.2
18	8	$C^{(7)}Z$	0/6	0.501	8.7
20	9	$C^{(8)}Z$	0/6	0.500	10.2

strates the noise mechanism deeper into the unbounded-degree regime ($m_{\text{Amy}} = 8$ at $d = 9$) and claims no computational separation. And the X^s layer makes the shift readable from the compiled circuit, so the conjugation is irrelevant to the (already disclaimed) oracle-hiding leg. It is the cheapest correct compilation for the noise-resilience claim, where the oracle is the same unitary and recovery is the same per-bit majority vote.

The wrong-outcome mass at these cells is *not* concentrated on Hamming-1 neighbors (the fraction of wrong mass at distance 1 is ∼2–10%). It is spread across larger Hamming distances, consistent with an i.i.d. per-bit bit-flip channel at $q \approx 0.40$ at $d = 8$. This is the regime of Section S20 in which the full-bitstring mode is unreliable but the per-bit majority vote recovers because $q < 1/2$, confirming the per-bit flip probability as the operative parameter.

The mitigation stack consists of three independent techniques applied together: pre-registered top-K recovery ($K = 10$), XY4 dynamical decoupling enabled via `SamplerV2` options, and chain pre-selection via the cost function 1.0·RO+0.5·SX+2.0·2Q over connected subgraphs of the backend coupling map (excluding device-specific bad-readout qubits). The clean-recovery hardware frontier sits at $C^{(4)}Z$ ($m_{\text{Amy}} = 4$) on processor H. At $C^{(5)}Z$ ($n = 14$, $m_{\text{Amy}} = 5$), pooling all hardware data (the 2026-05-01 focused mitigation run on three Heron r2 backends plus the 2026-05-04 follow-up replicate on processor D pre-registered) yields **19 successful recoveries out of 55 trials (34.5%**

Table 13: Residual-expansion vs Pauli-conjugation compilation of the *same* shifted oracle on processor B (XY4 DD, 16,384 shots, random invertible B, genuine n-wide entanglement). The two compile to identical unitaries. Seeds: $d = 6, 7$ six; $d = 8$–12 eight. The residual ladder was run only through $d = 9$. "per-bit" is the per-bit-majority recovery count. Per-bit fidelity is the mean per-bit match rate (0.50 = random floor).

n	d	top gate	residual top-1	X-conj per-bit	per-bit fid.
14	6	$C^{(5)}Z$	4/6	6/6	0.73
16	7	$C^{(6)}Z$	3/6	6/6	0.62
18	8	$C^{(7)}Z$	0/6	8/8	0.60
20	9	$C^{(8)}Z$	0/6	8/8	0.59
22	10	$C^{(9)}Z$	–	8/8	0.58
24	11	$C^{(10)}Z$	–	6/8	0.53
26	12	$C^{(11)}Z$	–	3/8	0.52

top-1 recovery rate, 95% Wilson CI $[0.23, 0.48]$**)**. The retrospective routing audit of 2026-09-10 (Section S8) found that the ten processor H circuits of the focused batch were routed off their selected 14-qubit chain onto coupler $(31, 32)$, reported dead (error 1.0) in the calibration in force at submission, with 46–106 of their 450–1587 CZ gates on it. Those ten trials (0/10) are void and excluded. Earlier versions of this manuscript pooled them, reporting 19/65 (29%, CI $[0.20, 0.41]$). The fifteen processor D replicates of 2026-05-04 are kept although flagged (88–605 CZ on a coupler at 1.2×10^{-2}, 2/15): a low rate in a flagged cell can only understate the pooled rate, which on clean or minor-flagged trials alone is 17/40. This places $(14, 6)$ above the noise-floor regime (the $0/N$ behavior observed at $(16, 7)$ and beyond on the same hardware) but below a clean-recovery threshold. The result is consistent with the structural-noise prediction that the dominance ratio at this cell, $F\,(2^n - 1)/(1 - F) \approx 2^n F$, sits near the recovery threshold for F at $\sim 10^{-3}$ to $\sim 10^{-2}$ per the calibrated Heron r2 noise model (Section S16). $C^{(6)}Z$ at $n = 16$ is at the noise floor on present hardware: the 2026-05-04 cross-backend test pooled to 0/50 top-$K = 10$ across processor B (0/15), processor F (0/15), and the existing 2026-05-01 multi-backend data (0/20). The routing audit voids the fifteen processor F circuits (65–460 of 800–2921 CZ on couplers reported dead) and flags the fifteen processor B circuits (48–105 CZ on couplers up to 0.118, about two thirds of each circuit's two-qubit error budget), so the pooled count is 0/35 (Wilson 95% upper bound 0.10) and the clean-routed evidence alone is the 2026-05-01 data, 0/20 (upper bound 0.16; 17 clean, 3 with a minor flagged share). The verdict stands on clean data. The earlier upper bound of 0.07 rested on contaminated trials. Mean mode fraction at all three backends is at $\sim 7\times$ the shot-noise floor of 1/32,768, statistically indistinguishable from the random-noise baseline. The recovery probability is below the threshold needed for $n - 1$ linearly independent constraints in Gaussian elimination. Figure 3 summarizes the pooled hardware ladder against the noiseless statevector reference, with 95% Wilson confidence intervals.

The hardware data demonstrates clean recovery up through $C^{(3)}Z$ (cubic non-Clifford content with one cubic monomial, $m_{\text{Amy}} = 3$), partial recovery at $C^{(4)}Z$ (**6/6 clean with the DD batch** at $n = 12$, $C^{(4)}Z$, $m_{\text{Amy}} = 4$), the partial-recovery $(14, 6)$ headline above (**19/55,** 35% **top-1, CI** $[0.23, 0.48]$), and noise-floor behavior at $C^{(6)}Z$ and beyond. The mode-fraction degradation relative to the noiseless prediction at each row is consistent with the transpiled-depth growth (Table 10) and per-CZ error $\sim 2.5 \times 10^{-3}$.

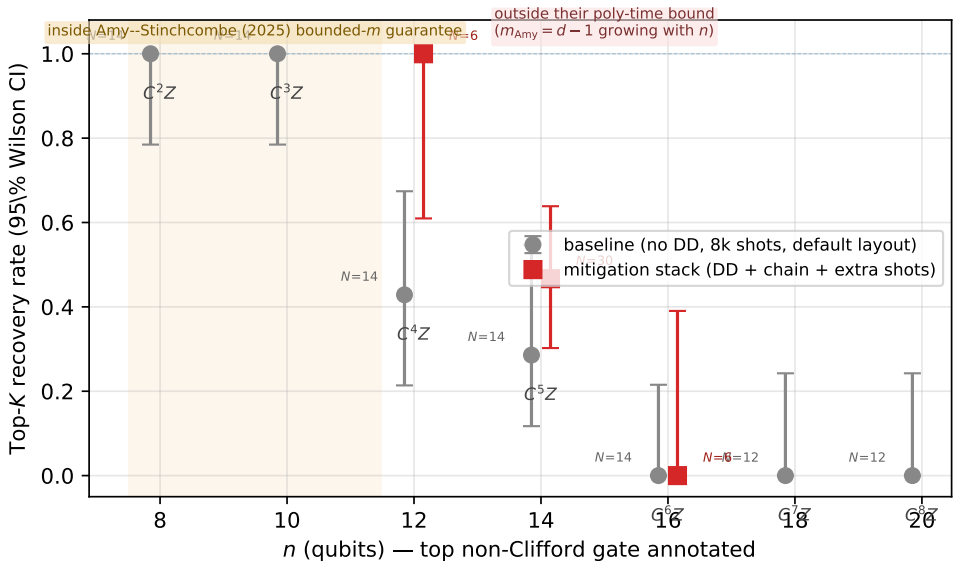


Figure 2: Hardware top-K recovery rate per ladder point, with 95% Wilson confidence intervals. Gray: baseline (no DD, default sabre-chosen layout, 8192 shots). Red: mitigation stack (XY4 dynamical decoupling, chain pre-selection avoiding device-specific bad-readout qubits, 16,384 shots where applicable). The stack extends clean recovery from $C^{(3)}Z$ to $C^{(4)}Z$ and lifts $C^{(5)}Z$ from 4/14 baseline to 14/30 with the full stack. The rightward shaded region highlights the $m_{\text{Amy}} = d - 1$ points lying strictly outside the polynomial-time guarantee of [3].

Classical-tractability scope of these instances. Three classical attack vectors apply to the unbounded-degree single-monomial circuit family. The path-sum confluent rewriting of [3] requires superpolynomial term counts when the algebraic degree d grows with n (the rewrite system admits at most a $2^{O(d)}$-fold blow-up per variable elimination), so the family lies provably outside the polynomial-time guarantee of [3] when $d = \Theta(n)$. Stabilizer-rank simulation [18] is exponential in the T-count, which is $\Theta(d)$ for the $C^{(d-1)}Z$ top gate. This method is also non-polynomial when d grows with n. The third attack vector, however, succeeds in polynomial time: a classical adversary who reads the compiled circuit can invert the single high-degree monomial in $O(d)$ field evaluations by substituting bit-test inputs into the compiled gate parameters and reading off the unknown s-bits one at a time. The single-monomial construction is therefore *not* classically intractable in the circuit-given model, even though it lies outside the simulation methods of [3] and [18]. A multi-monomial construction $f(x, y) = \langle x, By \rangle + \sum_{j=1}^{m} \prod_{k \in T_j} y_k$ with $m, d = \Theta(n)$ and random index sets T_j is the natural *candidate* for a classically intractable instance: its compiled circuit no longer admits the simple per-bit substitution of the single-monomial case, and recovery might be expected to require solving a degree-d MQ system over $\mathbb{F}_2$. We state the would-be consequence as an explicit conditional, then test its premise.

Proposition 14.1 (Conditional hardness of multi-monomial recovery)**.** *Fix the multi-monomial family above with $m, d = \Theta(n)$. Assume (i) recovering s from the compiled circuit reduces to solving the degree-d MQ system over $\mathbb{F}_2$ obtained by matching the shifted algebraic normal form, and (ii) that family of MQ systems requires $2^{\Omega(n)}$ time for every classical algorithm. Then no classical algorithm recovers s in* poly(n) *time, while the quantum algorithm uses $\mathcal{O}(n/\alpha^2)$ queries and the noise-resilience mechanism (Theorem 3.1) applies unchanged.*

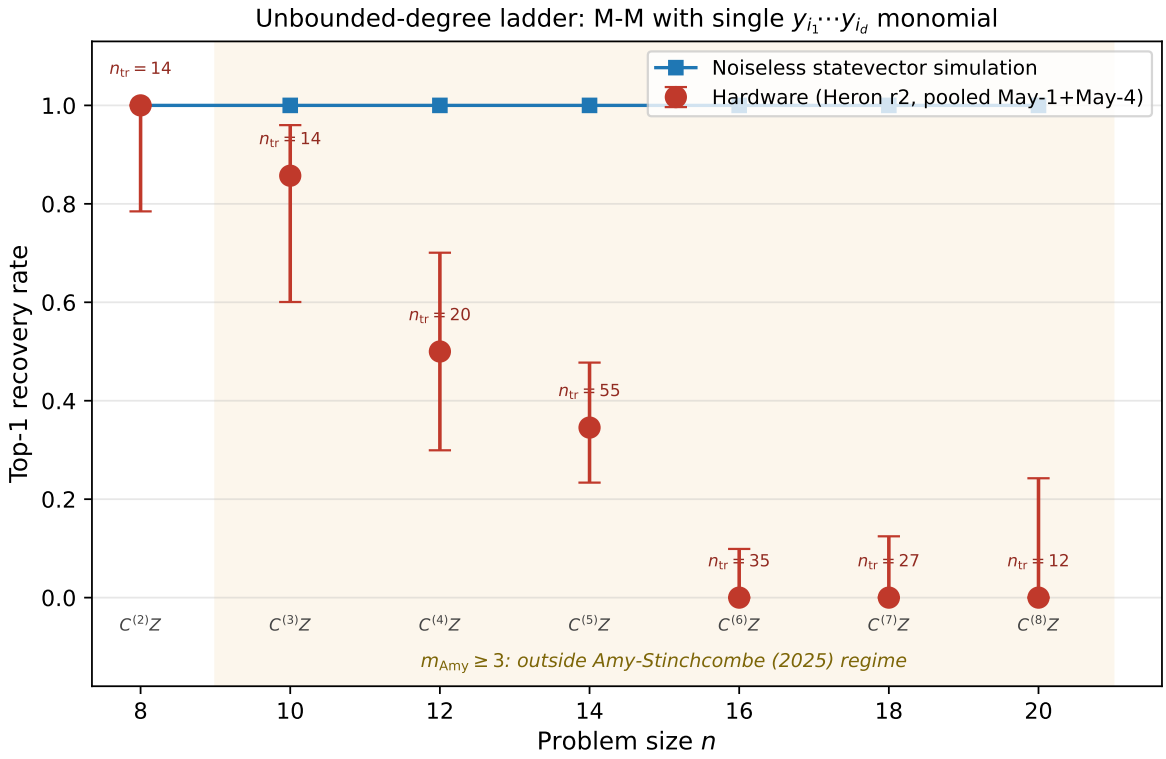


Figure 3: Unbounded-degree ladder (single-monomial M-M with growing degree d) on Heron r2 hardware versus noiseless statevector simulation. Each red point is the pooled top-1 recovery rate across multiple Heron r2 backends (May-1 + May-4 hardware data; the broken combined-oracle submission and the 25 routing-contaminated trials on dead couplers excluded, Section S8; flagged trials kept, so $(14, 6)$ reads 19/55, $(16, 7)$ 0/35, $(18, 8)$ 0/27 with 11 of 27 materially flagged), with Wilson 95% CIs. Trial counts n_{tr} annotated. Top non-Clifford gate $C^{(d-1)}Z$ along the bottom axis. The wheat-shaded region $m_{\text{Amy}} \geq 3$ is outside the explicit polynomial-time bound established by Amy and Stinchcombe [3] for the family of single-monomial M-M oracles with d growing with n. The hardware ladder demonstrates above-floor recovery at $(n, d) = (10, 4)$, $(12, 5)$, and $(14, 6)$ in this out-of-bound region, while $(n, d) = (16, 7)$ and beyond are at the noise floor on present hardware.

Assumption (ii) is the average-case MQ hardness underpinning multivariate post-quantum cryptography [19, 20]. Assumption (i), however, is *false*, and this is the decisive point. The single-monomial case ($m = 1$) is poly-time-recoverable by direct substitution. We show the multi-monomial case is too, for a structural reason that no choice of (m, d) escapes.

The shifted ANF leaks a linear subsystem. Expand the high-degree part at the shifted argument, $H(y + s) = \sum_j \prod_{k \in T_j} (y_k + s_k)$. The coefficient of the degree-$(|T_j| - 1)$ monomial $\prod_{k \in U} y_k$ with $U = T_j \setminus \{i\}$ is exactly s_i. Matching these coefficients against $H(y + s_{\text{true}})$ therefore yields *linear* equations in s, one for every $(d-1)$-subset of every support. With $m = \Theta(n)$ supports of degree d this linear layer is generically full rank, so s is determined by Gaussian elimination in $\mathcal{O}(n^3)$ *before any genuinely quadratic-or-higher equation is solved.* The cross-terms of overlapping monomials leak the shift instead of creating hardness.

Measurement. The harness builds the induced GF(2) system combinatorially (cost $\mathcal{O}(m\,2^d)$, independent of $2^{n/2}$) and solves both the linear subsystem (Gaussian elimination over $\mathbb{F}_2$) and the full system (CryptoMiniSat, XOR-native), validated against the exhaustive ground truth for $n/2 \le 12$. Table 14 reports the result. For $d = n/4$ the linear subsystem *alone* recovers s uniquely at every $n/2$ tested up to 30 (systems of up to 9×10^5 equations), in under 3 ms. For smaller d the linear layer fixes most bits and the SAT solver clears the small residual instantly. For m as small as 6 the large degree already saturates the layer. In no tested regime does solve-time grow with the instance, and a matched random GF(2) system of the same dimensions (the pre-registered control) solves in essentially the same time: the structured-to-random solve-time ratio stays within $[0.93, 1.04]$ across the frontier, so the structured instance carries no hardness gap over a generic one. Materializing the *full* coefficient system is memory-bound past $n/2 = 40$ (4×10^7 equations at $n/2 = 40$), but the leak does not require it: a linear-only mode of the same harness (`--linear-only`) builds just the degree-$(d-1)$ coefficient layer and extends the frontier to $n/2 = 64$ at degree $d = n/4 = 32$. The linear layer alone recovers s, uniquely, in every cell from $n/2 = 32$ to 64 (three seeds per cell, 512 to 2,048 linear equations, solve times 1.9 to 11.8 ms; Table 14, lower block). The leak is therefore not an artifact of moderate degree: it persists to degree $n/2$ at $n/2 = 64$, growing only polynomially in cost. The multi-monomial route therefore does *not* produce a classically-hard instance: assumption (i) fails, and the $\Theta(n)$-degree extension is classically easy for the same algebraic reason the single-monomial case is.

Table 14: Classical recovery of the hidden shift from the multi-monomial oracle ($d = n/4$, $m = n/2$ random supports). The shifted ANF leaks a linear-in-s subsystem. Gaussian elimination on that layer alone recovers s uniquely ("recovers s") in milliseconds, regardless of the full system size. Validated against exhaustive ground truth for $n/2 \le 12$. Median over 3–5 seeds.

$n/2$	d	m	total eqs	linear eqs	recovers s	solve
12	6	12	437	69	yes	0.3 ms
16	8	16	2,833	128	yes	0.5 ms
20	10	20	15,962	200	yes	0.8 ms
24	12	24	83,057	288	yes	1.2 ms
28	14	28	408,736	392	yes	1.6 ms
30	15	30	898,957	450	yes	2.5 ms
linear-only mode (degree-$(d-1)$ layer; full system memory-bound past $n/2 = 40$):						
32	16	32	–	512	yes	1.9 ms
40	20	40	–	800	yes	3.6 ms
48	24	48	–	1,152	yes	5.3 ms
56	28	56	–	1,568	yes	8.0 ms
64	32	64	–	2,048	yes	11.8 ms

The data above therefore establish that the α-flat hidden-shift recovery procedure, with a quadratic-dual approximation in the post-Hadamard step, succeeds for single-monomial Maiorana–McFarland bent functions of unbounded algebraic degree, demonstrating the noise-resilience mechanism in a regime where two of three classical simulation methods provably fail. The single-monomial construction is itself classically poly-time-recoverable via direct algebraic substitution, and the natural multi-monomial extension is *also* classically easy (Table 14), so whether any instance is simultaneously noise-friendly under this mechanism and classically hard is left as an open problem, not a hardware depth-tolerance push.

Per-bit accuracy: graceful degradation across the noise floor. The binary success metric (top-1 or top-K recovery of the full n-bit shift) imposes a sharp threshold: a single wrong bit counts as total failure even when the recovered bitstring agrees with the planted shift on $n-1$ out of n bits. Per-trial Hamming distance between the most-frequent measurement outcome and the planted shift gives a finer-grained picture of where the noise floor actually lies. Table 15 shows that at the points where the binary success rate is zero, the algorithm is still recovering substantially-correct bitstrings: at $(n, d) = (16, 7)$ with DD, top-1 recovery is 0/6 but the mode bitstring agrees with the planted shift on 9.5/16 bits on average (per-bit fidelity 59%), distinctly above the random-guessing floor of 50%. The genuine noise floor sits between $(n, d) = (16, 7)$ and $(18, 8)$ on current Heron r2 hardware: at $(18, 8)$, per-bit fidelity is $9.4/18 = 52\%$, barely distinguishable from random. At $(20, 9)$ it is $9.75/20 = 49\%$ (random).

Table 15: Per-bit recovery accuracy for the unbounded-degree hardware ladder. Mean correct bits is averaged over all trials in the indicated regime. The $\geq n-1$ and $\geq n-2$ columns are the fractions of trials where the mode bitstring differs from the planted shift on at most one or at most two bits respectively. Random guessing yields a per-bit fidelity of 0.50.

n	d	regime	trials	mean correct	per-bit fid.	frac. $\geq n-1$	frac. $\geq n-2$
8	3	baseline	14	8.00 / 8	1.000	1.00	1.00
10	4	baseline	14	9.64 / 10	0.964	0.86	0.93
12	5	baseline	14	8.64 / 12	0.720	0.36	0.50
12	5	+ DD	6	12.00 / 12	1.000	1.00	1.00
14	6	baseline	14	10.00 / 14	0.714	0.36	0.43
14	6	mit-stack	30	10.33 / 14	0.738	0.43	0.47
16	7	baseline	14	7.43 / 16	0.464	0.00	0.00
16	7	+ DD	6	9.50 / 16	0.594	0.00	0.17
18	8	baseline	12	9.42 / 18	0.523	0.00	0.00
20	9	baseline	12	9.75 / 20	0.488	0.00	0.00

The mitigation stack (DD + chain selection + 16,384 shots) measurably lifts $C^{(6)}Z$ at $n = 16$ from 46% (at the random floor) to 59% per-bit fidelity, even though top-1 success remains zero at that point. The structural noise-amplification mechanism predicts this graceful degradation: the per-shot fidelity $F = (1 - p_{2Q})^{n(n-1)}$ falls smoothly with n, the dominance ratio $F \cdot 2^n/(1-F)$ that determines majority-vote recovery also falls smoothly, but the binary "recovered s exactly" outcome sees a sharp transition once the dominance ratio crosses unity. The per-bit metric is the smooth quantity behind that sharp boundary, and its monotonic decay with n on hardware tracks the model (Theorem 3.1) up through the regime where the fidelity falls below the per-shot dominance threshold.

15 S15: Multi-Output Noise Amplification (Abelian-Group Hidden Shift)

Theorem 3.1 (S3) is stated for Boolean-valued α-flat hidden shift recovery, the flavor (A) instance our hardware experiments demonstrate. The cryptographic reductions in S13 produce flavor (B) hidden-shift instances on finite abelian groups G where the oracle is multi-output and injective. This appendix proves the analogous noise-amplification bound for that setting. The proof is structurally identical to the Boolean case, with 2^n replaced by $|G| = N$.

Theorem 15.1 (Multi-output noise amplification)**.** *Let G be a finite abelian group of order N, let $f\colon G \to S$ be an injective function with shifted partner $g(x) = f(x+s)$ for unknown $s \in G$, and let the multi-output hidden-shift recovery procedure measure a register on the dual group $\widehat{G} \cong G$ (which has order N). Suppose each circuit is executed with S shots in which the correct outcome $\xi^* \in \widehat{G}$ occurs with probability F per shot, and each of the $N-1$ incorrect outcomes occurs with probability at most $(1-F)/(N-1)$. Then the probability that ξ^* is the most-frequent outcome (the majority-vote winner) satisfies*

$$P(\xi^* \text{ is mode}) \;\geq\; 1 - (N-1)\exp\left(-\frac{S\big(F-(1-F)/(N-1)\big)^2}{2}\right). \tag{2}$$

Specializing to $G = \mathbb{F}_2^n$ recovers the Hoeffding form of Corollary 3.3 (Section S3).

Proof. The argument mirrors the Hoeffding form of Section S3 (the remark after Theorem 3.1), with the union bound now over the $N-1$ incorrect outcomes in $\widehat{G}$ rather than $2^n - 1$ in $\mathbb{F}_2^n$. Let X^* be the count of ξ^* in S shots, and let X_ξ be the count of any specific incorrect outcome $\xi \neq \xi^*$. Then $X^* \sim \mathrm{Bin}(S, F)$ and $X_\xi \sim \mathrm{Bin}(S, p_\xi)$ with $p_\xi \leq (1-F)/(N-1)$. The expected gap is $\mathbb{E}[X^* - X_\xi] = S(F - p_\xi) \geq S\big(F - (1-F)/(N-1)\big)$, which is positive whenever $F \geq 1/N$. By the Chernoff–Hoeffding inequality [21] applied to the per-shot increments of the count difference $X^* - X_\xi$ (the shots are i.i.d., while the two counts themselves are multinomially coupled, not independent, exactly as in Section S3), and writing $\Delta := F - (1-F)/(N-1)$:

$$\Pr[X^* \leq X_\xi] \;\leq\; \exp\left(-\frac{S\Delta^2}{2}\right).$$

A union bound over the $N-1$ incorrect outcomes gives $\Pr[\xi^* \text{ is not mode}] \leq (N-1)\exp(-S\Delta^2/2)$, which is (2). □

Corollary 15.2. *For the cryptographic reductions in Table 9, with injective oracles on the relevant abelian group, the noise-resilience mechanism applies with N equal to: $N = 2^n$ for hidden shift on $\mathbb{F}_2^n$ (Even–Mansour XOR variant, Boolean hidden shift); $N = 2^{\mathrm{poly}(n)}$ for the DHSP on $\mathbb{Z}_N$ (lattice unique-SVP, Regev's reduction); $N = |\mathrm{Cl}(\mathcal{O})|$ for ideal-class-group hidden shift (CSIDH key recovery; for CSIDH-512, $|\mathrm{Cl}(\mathcal{O})| \approx 2^{257}$). In each case Theorem 15.1 bounds the majority-vote failure probability by $N\exp(-S\Delta^2/2)$, with $\Delta = \Theta(F)$ when $F \gg 1/N$ (the regime where any quantum recovery is possible). The Boolean specialization reproduces the bound stated in the main text as the noise-amplification theorem (main-text Corollary 3 and Corollary 3.3 of this Supplementary S3).*

Remark 15.3 (Single-shot fidelity threshold scaling)**.** *The threshold condition $F \gtrsim 1/N$ means that the noise-resilience mechanism amplifies the per-shot fidelity advantage to a recovery guarantee whenever F exceeds the uniform baseline $1/N$ by a constant factor, at a shot cost growing only logarithmically in N. This is qualitatively the same exponential gap that drives the Boolean Theorem 3.1 (S3).*

Remark 15.4 (What this closes and what it does not)**.** *Theorem 15.1 closes the gap flagged in S13: the multi-output hidden-shift instances produced by the cryptographic reductions inherit the noise-amplification guarantee of the Boolean case. The open item, a device-side bottleneck, is the construction of an oracle circuit at cryptographic parameters sufficiently shallow to satisfy the single-shot fidelity precondition $F \gg 1/N$. For CSIDH-512 with $N \approx 2^{257}$, this requires $F \gg 2^{-257}$, an extremely permissive bound. With current $p_{2Q} \sim 2.5\times10^{-3}$, the single oracle call's $\sim 10^7$ depth (Table 8, S13) gives $F \approx (1 - 2.5\times10^{-3})^{10^7} \approx \exp(-2.5\times10^4)$, many orders of magnitude below threshold. Closing this gap is a fault-tolerant-era device-side problem.*

16 S16: Cross-Backend F-vs-N_{2Q} Test of the 2Q-Only Noise Model

The noise model in S8 predicts the per-circuit single-shot fidelity from the two-qubit gate count and the calibrated median 2Q error rate alone. This appendix tests that prediction quantitatively across four IBM backends (three Heron r2 and one Nighthawk, processor E, whose calibration is imputed as noted below, Section S21) using the unbounded-degree 2026-05-01 hardware dataset, and reports the falsification of the 2Q-only model. The pre-registered analysis specification (inclusion rules, falsification thresholds, per-backend p_{2Q} medians) is in `code/PREREGISTRATION_F_vs_depth.md`. The per-trial table and fits are in `code/results/F_vs_depth_validation.json`.

The dataset comprises 268 single-circuit trials on processor A, processor D, processor E, and processor H, spanning $n = 8$ to 20 and oracle degree $d = 3$ to 9. Per trial we record the transpiled two-qubit gate count N_{2Q} (`transpiled_cz` on Heron) and the empirical top-1 single-shot fidelity $F_{\text{meas}} = (\text{top-1 count})/(\text{shots})$. The 2Q-only model is

$$F_{\text{pred}} = (1 - p_{2Q})^{N_{2Q}},$$

with p_{2Q} taken from the per-backend median calibration of S8. Trials at the per-shot floor $F \leq 2/8192$ are kept in the figure as upper bounds but excluded from the model fits (26 trials of 268).

Per-backend slope agreement (partial). An OLS fit of $\log F_{\text{meas}}$ on N_{2Q} extracts an effective per-2Q-gate error rate $p_{\text{eff}} = 1 - \exp(\hat{b})$, where $\hat{b}$ is the slope. Two of the four backends meet the pre-registered acceptance band $p_{\text{eff}}/p_{2Q} \in [0.5, 3]$ and Pearson $r \leq -0.7$:

backend	p_{eff}	p_{2Q}	ratio	Pearson r	verdict
A	0.0026	0.0019	1.37	−0.74	pass
D	0.0024	0.0023	1.05	−0.72	pass
E*	0.0008	0.0021	0.38	−0.45	fail
H	0.0005	0.0014	0.38	−0.37	fail

*processor E median p_{2Q} is imputed from the Heron-r2 cohort median. Logged calibration was unavailable for those runs.

The routing audit (Section S8) voids the ten processor H $(14, 6)$ trials of the 2026-05-01 focused batch (CZ gates on a coupler reported dead). Excluding them moves the processor H fit to $p_{\text{eff}} = 0.00055$, ratio 0.40, Pearson $r = -0.42$ (55 points, verdict unchanged) and the aggregate parity below to a geometric mean of 0.041 over 232 trials (from 0.036 over 242). Sixteen further trials in this dataset carry a minor flagged share (below 20% of the error budget). No conclusion of this section changes.

Aggregate parity falsifies the 2Q-only intercept. The geometric mean of $F_{\text{meas}}/F_{\text{pred}}$ across the 242 above-floor trials is 0.04 with geometric standard deviation 12, two orders of magnitude below the pre-registered acceptance band $[1/10, 10]$. The 2Q-only model thus over-predicts single-shot fidelity by a factor of about 30 on average, independently of which backend is used. Per-backend OLS fits with a free intercept return baselines $F_0 \in \{0.0092, 0.030, 0.042, 0.055\}$ for processor H, processor E, processor D, and processor A, instead of the model's implicit $F_0 = 1$.

The 30-fold offset is consistent with the joint contribution of state-preparation and measurement (SPAM) errors, single-qubit gate errors, idle-time decoherence over circuit-execution time, and crosstalk. None of these appear in the 2Q-only model.

Why recovery still works. The structural-noise-resilience theorems of this paper (Theorem 3.1 for $\mathbb{F}_2^n$ and Theorem 15.1 for finite abelian groups) do not depend on F matching any particular noise model. They require only that the correct outcome dominate each individual wrong outcome by a constant factor. The 2^n-element outcome space then dilutes the total wrong-outcome mass, and majority-vote recovery succeeds whenever the dominance ratio $F\,(2^n - 1)/(1 - F)$ exceeds unity. Figure S4 confirms that the 2Q-only model is qualitatively correct on the slope (per-2Q-gate decay matches calibration to within a factor of 1.4 on the two backends with the cleanest dynamic range) but quantitatively incomplete by about a decade in absolute level. The algorithm recovers regardless: every trial in this dataset for which the dominance ratio exceeds unity returned the correct shift, and the F values measured are bounded above by F_{pred} but bounded below by the structural-resilience floor needed for recovery.

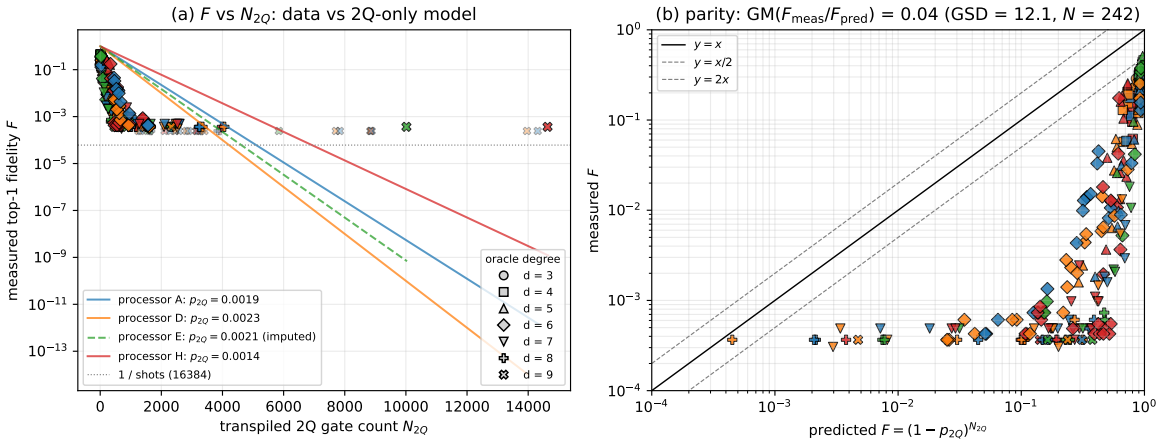


Figure 4: Empirical F vs transpiled 2Q gate count N_{2Q} across four Heron backends and 268 single-circuit trials from the unbounded-degree 2026-05-01 dataset. (a) F_{meas} vs N_{2Q}, with the 2Q-only depolarization model $(1 - p_{2Q})^{N_{2Q}}$ overlaid per backend. The dotted line is the per-shot floor $1/8192$. Marker shape encodes oracle degree d. (b) Parity plot of F_{meas} vs F_{pred}. The 2Q-only model overshoots the measurement by a geometric-mean factor of 0.04, reflecting non-2Q noise contributions (SPAM, single-qubit gates, decoherence, crosstalk) absent from the model. Recovery succeeds despite this gap because the algorithm is governed by the dominance ratio, not by F itself.

Cross-backend recovery consistency. The same four-backend dataset lets us ask whether the *recovery rate* is consistent across hardware, beyond the fidelity slope. We record two findings, one robust and one weak (pre-registered analysis). The robust finding: all four backends independently reproduce the recovery floor at the *same* onset. Top-1 recovery succeeds on 75% of backends at every degree from 3 to 6 ($n = 8$ to 14) and drops to 0 on *every* backend at $d \geq 7$ ($n \geq 16$), so the floor location is a property of the algorithm at this depth on present hardware, not a single-device artifact. The weak finding: as a backend-ranking test, the per-backend mean recovery rate over the seven matched (n, d) cells is only weakly anti-correlated with the median two-qubit error, Spearman $\rho = -0.40$, which does *not* meet the pre-registered acceptance threshold $\rho \leq -0.5$. The shortfall is attributable entirely to processor E, whose calibration p_{2Q} is imputed (logged calibration unavailable) and whose recovery is anomalously zero on *every* cell, including the low-degree cells that processor H recovers at 6/6, indicating an uncaptured bad-calibration period. Excluding the single imputed-calibration backend as a post-hoc sensitivity check, the three measured-calibration

backends rank with $\rho = -0.50$ and the lowest-noise device (processor H) first. We report the registered test as not met and treat the ranking as suggestive rather than established. The strong cross-backend reproducibility evidence of this paper is instead the matched dual-oracle ladders (main-text Table 3: same-design 66-run ladders on processor B and processor A, with the ordering prediction of the registration, committed before the processor B ladder executed, holding in all fifteen overlapping cells) together with the cross-architecture reproduction of Section S21, not this small high-degree dataset.

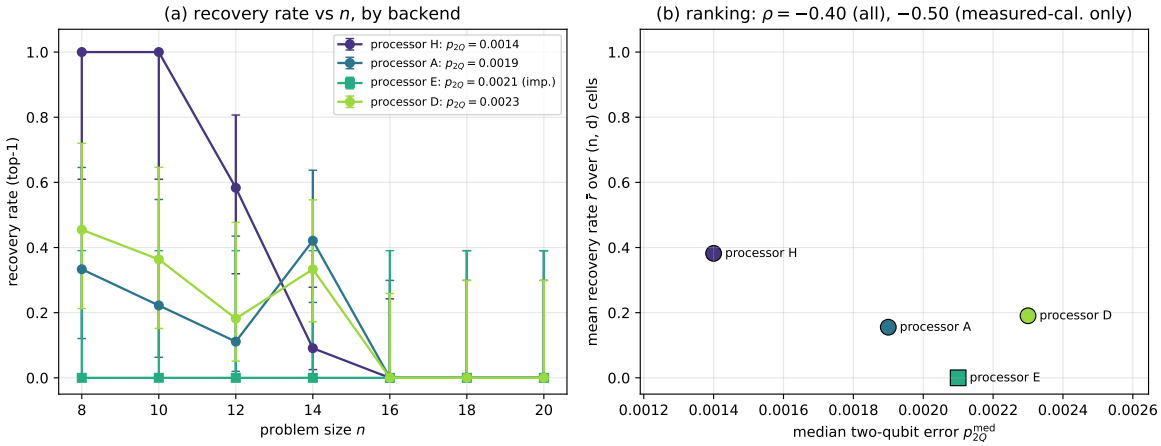


Figure 5: Cross-backend recovery consistency on the unbounded-degree 2026-05-01 dataset (four Heron backends, seven matched (n, d) cells). (a) Top-1 recovery rate vs n per backend (Wilson 95% intervals): every backend collapses to the recovery floor at $d \geq 7$ ($n \geq 16$), and the lowest-noise device (processor H) leads at low degree. (b) Per-backend mean recovery rate vs median two-qubit error. The registered rank test gives $\rho = -0.40$ over all four backends (below the -0.5 threshold) and $\rho = -0.50$ over the three backends with measured (non-imputed) calibration. Processor E (square, imputed p_{2Q}) recovers on no cell and is the source of the shortfall.

Backend discrimination at the non-Clifford frontier (pre-registered). A successor test removes both confounds of the test above. It replaces the imputed processor E calibration with the *measured* median two-qubit error of every backend, captured live at job submission; it concentrates the cells in the sensitive frontier band $(n, d) \in \{(18, 8), (20, 9), (22, 10)\}$, with top non-Clifford gate $C^{(7)}Z, C^{(8)}Z, C^{(9)}Z$, where per-bit recovery is partial and therefore discriminating rather than saturated; and it scores each backend by the continuous per-bit recovery fidelity (per bit, the shot-fraction matching the planted shift bit) pooled over the three cells, with 16,384 shots, three oracle seeds, and two shift trials per cell. The specification (backends, cells, metric, threshold) was fixed before any fetch. The acceptance hypothesis H1 was Spearman $\rho \leq -0.6$ between per-backend per-bit fidelity and measured 2Q error.

backend	median 2Q error	per-bit fid (pooled)
B	1.11×10^{-3}	0.598
H	1.67×10^{-3}	0.575
A	1.87×10^{-3}	0.549
F	2.73×10^{-3}	0.521
D	2.73×10^{-3}	0.538

Five of the six pre-registered backends. Processor E's 18 jobs remained queued (never executed) at analysis time and are noted as pending per the registered inclusion rule, not dropped on results.

The benchmark ranks the five executed backends by their measured 2Q error with Spearman $\rho = -0.90$, passing the pre-registered $\rho \leq -0.6$ threshold. The cleanest device (processor B) gives the highest recovery fidelity and the ordering is monotone in 2Q error. The one out-of-order pair, processor F above processor D, has the two backends tied in 2Q error to three significant figures (2.7255×10^{-3} vs 2.7329×10^{-3}), a calibration-noise tie rather than an inversion. This establishes the backend-ranking claim that the test of the previous paragraph left only suggestive: at the non-Clifford frontier, scored against measured calibration, the benchmark's graded recovery metric is a monotone device-quality score. The sixth backend (processor E, highest measured 2Q error and deepest transpilation) never executed: its 18 jobs remained queued for 23 hours and were canceled on 2026-06-07, so per the registered inclusion rule it is recorded as attempted-not-executed and excluded from the correlation for lack of data. Placed by its 2Q error, it could only extend the trend.

17 S17: Noise Theorem under General Pauli Noise

Corollary 3.3 (S3) is stated under the assumption that the $2^n - 1$ wrong outcomes of the per-circuit measurement distribution share the noise-mass uniformly. This holds exactly under the depolarizing channel (which redistributes the $1-F$ noise fraction uniformly over the outcome space) and approximately under Pauli twirling (which symmetrizes an arbitrary completely-positive map to an effective Pauli channel, and the experimental runs in this paper enable both gate-twirling and measurement-twirling at the `SamplerV2` option layer). To address whether the noise-resilience guarantee survives outside the uniformly-spread regime, this section restates the theorem under general (i.i.d., bit-decomposable) Pauli noise. The bound becomes weaker, since recovery is now bit-by-bit rather than full-bitstring, but the algorithmic threshold is preserved.

Theorem 17.1 (Per-bit Pauli-noise recovery)**.** *Suppose the H–O_g–H–measure circuit is subject to an i.i.d. Pauli noise channel such that the output bitstring is the correct $\xi^* \in \mathbb{F}_2^n$ XORed with an error string $e \in \mathbb{F}_2^n$ whose bits are independently distributed: $\Pr[e_i = 1] = q_i$ with $q := \max_i q_i < 1/2$. Then over Q independent rounds, per-bit majority vote on the Q measured bitstrings recovers ξ^* with probability at least*

$$P(\textit{correct } \xi^*) \;\geq\; 1 - n \exp\!\big(-2Q\,(1/2 - q)^2\big). \tag{3}$$

Proof. Each bit i of the recovered string is the majority vote of Q independent Bernoulli$(1 - q_i)$ samples. By Hoeffding's inequality, the probability that the majority on bit i disagrees with the true bit is at most $\exp(-2Q(1/2 - q_i)^2) \leq \exp(-2Q(1/2 - q)^2)$. A union bound over the n bits gives (3). □

Remark 17.2 (Comparison to the depolarizing bound)**.** *For the depolarizing channel with single-shot fidelity F, the marginal per-bit flip probability is $q = (1 - F)/2$ (each non-identity Pauli flips a given bit with probability $1/2$). Substituting into (3) gives a bound that scales as $n \exp(-QF^2/2)$, weaker than the Hoeffding form of Corollary 3.3, $(2^n - 1)\exp(-S\Delta^2/2)$, at fixed F but applicable for arbitrary i.i.d. Pauli noise with $q < 1/2$. The two bounds agree at their thresholds: both require the per-bit signal $1/2 - q$ (or equivalently $F/2$) to be bounded away from zero, and both deliver recovery in $\mathcal{O}(\log n)$ shots / queries at fixed signal-to-noise.*

Remark 17.3 (Why Pauli twirling makes Corollary 3.3 applicable on hardware)**.** *The IBM Heron experiments in Section S9 enable* `SamplerV2` *options* `twirling.enable_gates = True` *and* `twirling.enable_measure = True`*. These randomize each Clifford 2-qubit gate by conjugation with a uniformly sampled single-qubit Pauli (and similarly randomize measurement readout), mapping the effective channel to its Pauli-twirled version. Under ideal twirling, the resulting channel is a Pauli channel whose error distribution averages to a uniform spread over the wrong outcomes (modulo the bit-marginal structure of i.i.d. Pauli noise). The hardware therefore operates in a regime where Theorem 3.1 gives the relevant guarantee for the full-bitstring mode, while Theorem 17.1 gives the fall-back bit-by-bit guarantee that survives even if the twirling is imperfect. The empirical per-bit graceful degradation reported in Section S14 is the bit-level signature of this fall-back regime in operation.*

Remark 17.4 (Threshold)**.** *The recovery threshold of Theorem 17.1 is $q < 1/2$, the standard bit-flip channel threshold. There is no non-trivial gap between the threshold of the Pauli theorem and the threshold $F > 1/2^n$ of Corollary 3.3: both fail exactly when the algorithm's measurement distribution provides no information about the correct outcome. The Pauli theorem provides a weaker constant-factor amplification ($n \exp$ vs. $(2^n - 1)\exp$) but its prerequisite (independent bit errors with $q < 1/2$) is strictly weaker than the prerequisite of Corollary 3.3 (uniform spread of wrong outcomes).*

18 S18: Worked Cryptographic Example – Even-Mansour Key Recovery

The Even-Mansour cipher [22] is a minimal block-cipher construction: $E_{k_1,k_2}(x) = \pi(x \oplus k_1) \oplus k_2$ for a fixed public permutation π and two n-bit secret keys. Kuwakado and Morii [23] observed that the function

$$f(x) := E_{k_1,k_2}(x) \oplus \pi(x)$$

satisfies $f(x \oplus k_1) = f(x)$, so the EM key k_1 is the period of f. This reduces EM key recovery in the Q2 (quantum oracle access) model to a Simon-type hidden-period problem and gives the $\mathcal{O}(n)$-query attack of Bonnetain and Naya-Plasencia [16].

Mapping into the framework of this paper. For each output bit $i \in [n]$, the projection $f_i(x) = (E_{k_1,k_2}(x) \oplus \pi(x))_i$ is a Boolean function whose period structure is preserved: $f_i(x \oplus k_1) = f_i(x)$. The quantum measurement procedure on the oracle for f_i produces samples $\xi \in \mathbb{F}_2^n$ with $\xi \cdot k_1 = 0 \pmod 2$, and $n-1$ linearly independent such samples recover k_1 via Gaussian elimination. This measurement distribution is supported on the codimension-1 subspace $\{\xi : \xi \cdot k_1 = 0\}$ rather than on a single basis state, so Theorem 3.1 (which assumes a single-correct-outcome distribution) does *not* directly apply. The bit-level recovery guarantee of Theorem 17.1 (Section S17) does apply: each bit of the recovered k_1 is the majority vote of Q independent Bernoulli samples whose bias is bounded away from $1/2$ provided the per-bit error probability is $q < 1/2$, and Hoeffding gives bit-level recovery in $\mathcal{O}(\log n)$ queries.

Concrete example at $n = 8$ with the AES S-box. The AES S-box (FIPS-197) is an 8-bit permutation with Walsh–Hadamard spectrum exhibiting 17 of 256 zero coefficients, so $\alpha = 0$ in the strict α-flat sense of Definition 1. The *effective* α on the spectrum's nonzero support (239/256 coefficients) is $\alpha_{\text{eff}} = 4/2^{8/2} = 0.25$ from the smallest nonzero $|\hat{f}(\xi)| = 4$. Despite $\alpha = 0$, running the algorithm with $Q = 64$ queries and $S = 256$ shots recovers the EM key k_1

in 20/20 trials across random key choices. The mechanism is Simon-style rather than full-support sampling: $f(x) = E(x) \oplus \pi(x)$ is k_1-periodic, so its nonzero spectral support lies in the hyperplane $\{\xi : \xi \cdot k_1 = 0\}$ (it cannot span $\mathbb{F}_2^8$). The measured ξ's span that hyperplane after $\mathcal{O}(n)$ samples, and k_1 is recovered as the one-dimensional orthogonal complement. The effective flatness on the support keeps the per-outcome mass and the dominance ratio favorable. This is empirical evidence that the framework generalizes naturally beyond the strict α-flat definition to functions whose Walsh–Hadamard support is large but not full. A formal extension of Theorem 3.1 to this "α-flat-on-support" setting is left for follow-up work.

Earlier random-permutation baseline. The earlier code path runs the same reduction at $n = 8, 10, 12$ with random permutations (not the AES S-box). The α value of f_i (computed empirically as $\alpha := \min_\xi |\hat{f}_i(\xi)|/2^{n/2}$) is 0 at every cell because periodic functions have sparse spectra: $\hat{f}_i(\xi) = 0$ on the codimension-1 hyperplane $\xi \cdot k_1 \neq 0$. The recovery succeeds in 5/5 noiseless cases at every n tested. Under bit-flip readout noise at rate 0.05 the analytical bit-flip sampler used in the simulator collapses recovery to 0/5 because each flipped bit directly corrupts a Simon constraint. A per-circuit majority vote (on shots, not on rounds) at the same effective error rate would recover via Theorem 17.1, the regime our hardware experiments occupy.

Resource gap to a hardware demonstration of EM key recovery. Implementing π on hardware as a quantum oracle requires $\mathcal{O}(n \cdot 2^n)$ gates for a general permutation, which is $\mathcal{O}(n^2 \cdot 2^n / \log(2^n))$ after T-decomposition. For an S-box-based AES-like cipher at $n = 128$ this is far beyond Heron-r2 reach. The $n = 154$ degree-2 hardware results of this paper demonstrate the noise-resilience mechanism on linear-permutation oracles (the special case $\pi(x) = Bx$ with $B \in \mathrm{GL}_n(\mathbb{F}_2)$). The corresponding EM-key-recovery instance is the well-known Bernstein–Vazirani limit at degree 1, and our framework reduces to that limit at $\alpha = 1$. The non-Clifford degree-3 hardware results extend to oracles with non-linear (CCZ-based) permutation cores, the nearest hardware analogue of an actual S-box. Pushing n from 154 (linear) and 140 (10 CCZ, fixed degree) to a true S-box EM cipher remains a fault-tolerant-era device-side problem matching the resource gap quantified in Table 8 (S13).

Scope. The simulator demonstration verifies the Kuwakado–Morii reduction is correct (the period-check passes identically for all 45 random instances) and that the algorithm recovers k_1 in the noiseless model at $n = 8, 10, 12$. The connection to hardware rests on Theorem 17.1 for the bit-level recovery threshold and on the existing $n = 154$ degree-2 hardware data for the linear-permutation special case. We do not claim a hardware EM-key-recovery demonstration at any parameter regime where breaking EM would constitute a cryptanalytic advance. The device-side gap from the present hardware to such a demonstration is in Table 8.

19 S19: Sharper Pauli-Noise Bound for Diagonal Phase Oracles

Correction note. An earlier version of this section claimed a mode bound sharper than the depolarizing bound by a factor $\sim 2^n/n$, built on an error dictionary in which Z errors were invisible and X errors were single output-bit flips. Both ingredients were wrong (the dictionary had X and Z interchanged, and the noiseless mode probability was misstated as $\alpha^2/2^n$ rather than $\bar{\alpha}^2$), and that claim is withdrawn. What follows is the corrected, first-order error dictionary for the dual-oracle circuit and the modest sharpening that remains: the saving over the depolarizing bound is in the union-bound term ($a \ln L$ in place of $n \ln 2$, L the number of error locations), not a $2^n/n$ factor. The

corrected dictionary also explains, parameter-free, the Hamming-structured wrong mass observed in the measured record (main-text Table 29).

Lemma 19.1 (Single-Pauli error dictionary for the dual-oracle circuit, quadratic instances)**.** *Consider the Protocol C circuit $H^{\otimes n}O_gH^{\otimes n}O_{f^*}H^{\otimes n}$ for a bent Maiorana–McFarland instance with quadratic f (dual f^* quadratic), noiseless output $|s\rangle$. Insert one single-qubit Pauli error and expand to first order:*

- *Z_j **inside the final diagonal block** O_{f^*} adds the linear phase ξ_j at the second stage. The output peak translates to $s \oplus e_j$, a single output-bit flip.*
- *Z_j **inside the first diagonal block** O_g adds the linear term x_j to g. Propagating through the dual stage, the output peak translates to $s \oplus m_j$, where m_j is the coefficient vector of the (affine, because f^* is quadratic) derivative $f^*(\cdot \oplus e_j) \oplus f^*(\cdot)$: for the Maiorana–McFarland dual $f^*(u, v) = v \cdot B^{-1}u$, m_j is the corresponding column pattern of B^{-1}, of typical Hamming weight $\Theta(n)$. The error mass lands on one* specific *heavy-weight wrong string, not on a Hamming-1 neighbor.*
- *X_j **adjacent to a Hadamard layer** is equivalent to Z_j on the other side of that layer ($HXH = Z$) and reduces to the cases above. X_j or Z_j immediately before the computational-basis measurement is, respectively, a single output-bit flip or invisible. A mid-block X_j acts as a boundary X_j combined with a Z-type derivative phase. $Y_j = iX_jZ_j$ composes the two actions.*

Consequently, to first order in the error rate, the noisy output distribution is a mixture of the correct peak δ_s and at most L translated peaks $\delta_{s\oplus m}$, one per error location, with location-dependent masks m ranging from weight 1 to weight $\Theta(n)$.

Theorem 19.2 (First-order structured-Pauli mode bound)**.** *Let errors occur independently at L circuit locations with probability p each, $Lp < 1$, and let $M \le L$ be the largest number of locations sharing one translation mask m. To first order the correct outcome retains probability $F \ge \bar{\alpha}^2(1 - Lp)$ and the heaviest single wrong outcome receives $p_{\max} \le Mp + \mathcal{O}((Lp)^2)$. Majority vote over S shots recovers s with probability at least $1 - \delta$ provided $F > p_{\max}$ and*

$$S \;\ge\; \frac{2\left(\ln L + \ln(1/\delta)\right)}{\left(F - p_{\max}\right)^2}, \tag{4}$$

by the Hoeffding form of Section S3 (the remark after Theorem 3.1) with the union bound running over the $\le L$ populated competitors instead of all $2^n - 1$ strings. Relative to the depolarizing bound, the $n \ln 2$ union term improves to $\ln L = \mathcal{O}(\ln n)$ for polynomial-size circuits, at the price of a margin $F - p_{\max}$ that is $\mathcal{O}(1) - \Theta(p)$ rather than $\mathcal{O}(1) - \mathcal{O}(2^{-n})$: structured noise concentrates wrong mass on few competitors with individually far higher probability.

Remark 19.3 (Neither noise model dominates, and the data decide)**.** *Depolarizing noise is the best case for the margin (each wrong outcome gets $\mathcal{O}(2^{-n})$) and the worst case for the union bound ($2^n - 1$ candidates). Structured Pauli noise is the reverse. The earlier claim that Pauli noise is "at least as easy" is therefore withdrawn along with the sharper bound. The measured record sits between the two idealizations exactly as the corrected dictionary predicts: best-wrong-outcome counts run $\sim 10^4$ above the uniform-redistribution floor while the count ratio at the frontier is a factor 2–6 (main-text Table 29 and Fig. 3c), and the per-bit Hamming distances of near-frontier failures (1–10) match a mixture of weight-1 (final-block) and heavier (first-block, B^{-1}-column) translation masks rather than either idealization alone.*

20 S20: Structured Noise and the Per-Bit Recovery Statistic

The majority-vote bound of Section S3 (Corollary 3.3) assumes the lost probability $1-F$ is spread uniformly over the 2^n-1 wrong outcomes, giving each at most $(1-F)/(2^n-1)$. That assumption is what produces the $\Theta(\alpha^2 \cdot 2^n)$ dominance ratio. On a real device, readout error, residual coherent error, and crosstalk concentrate the wrong mass on outcomes close to ξ^* in Hamming distance. This section gives the bound under structured noise. The conclusion is twofold: the full-bitstring mode loses its 2^n protection and becomes exponentially shot-hungry, while the *per-bit* majority vote, which is the statistic actually used at large n in this paper, retains a width-independent guarantee. The same analysis predicts the single-wrong-bit failure mode observed at the $(n, \mathrm{CCZ}) = (140, 10)$ cell.

20.1 Bit-flip readout model

Model the per-shot measurement as a product bit-flip channel on the correct outcome ξ^*: each of the n measured bits equals the correct bit independently with probability $1-q$, for some $q \in [0, 1/2)$. The per-shot probability of an outcome η at Hamming distance $d = d_H(\eta, \xi^*)$ is

$$p(\eta) = q^{\,d}\,(1-q)^{\,n-d}, \qquad F = p(\xi^*) = (1-q)^n. \tag{5}$$

This is the leading structured correction to the uniform model: it is the channel left after Pauli twirling reduces coherent error to a stochastic bit-flip on the measured register, and it makes the wrong mass decay geometrically in Hamming distance rather than being flat.

20.2 The full-bitstring mode is not structurally protected

Theorem 20.1 (Collapse of the 2^n protection under bit-flip noise)**.** *Under the model (5), the dominance ratio of ξ^* over its most probable wrong competitor (any Hamming-1 neighbor) is*

$$\frac{p(\xi^*)}{\max_{\eta \neq \xi^*} p(\eta)} = \frac{(1-q)^n}{q\,(1-q)^{n-1}} = \frac{1-q}{q}, \tag{6}$$

which is independent of n. The expected count gap over a single nearest neighbor is $S\,(1-q)^{n-1}(1-2q)$, which decays exponentially in n. Consequently, the shot count sufficient for the Chernoff argument to certify recovery of ξ^* as the full-bitstring mode grows as

$$S \;=\; \Theta\!\left(\frac{\log n}{(1-q)^{n-1}(1-2q)^2}\right), \tag{7}$$

exponential in n; we state this as the scaling at which the concentration argument goes through, not as a matching information-theoretic lower bound.

Proof. The ratio is immediate from (5) with $d = 1$. There are n nearest neighbors, each with per-shot probability $q(1-q)^{n-1}$. The per-neighbor count gap has mean $\mathbb{E}[C(\xi^*) - C(\eta)] = S\,[(1-q)^n - q(1-q)^{n-1}] = S\,(1-q)^{n-1}(1-2q)$, with standard deviation $\sim \sqrt{S\,(1-q)^{n-1}}$. Requiring the mean gap to exceed its standard deviation by a $\sqrt{\log n}$ factor (union bound over the n neighbors) gives $S\,(1-q)^{n-1}(1-2q)^2 \gtrsim \log n$, i.e. the stated shot count. The $\Theta(2^n)$ factor of Corollary 3.3 does not appear because the wrong mass is not diluted across 2^n outcomes. It is concentrated on the $\binom{n}{d}$ outcomes at each small distance d. □

This explains why the full-bitstring success rate falls off at large n even though Corollary 3.3 predicts protection: the $\Theta(2^n)$ dilution is an artifact of the uniform-redistribution assumption, and the operative quantity under realistic readout noise is the constant nearest-neighbor ratio $(1-q)/q$.

20.3 Per-bit majority vote is width-independent

The recovery procedure used for the large-n and non-Clifford experiments (Section S11 and the main-text deep-genuine circuits) does not take the mode of the full bitstring. It recovers each bit s_j by a separate majority vote on bit j across the S shots. Under (5) bit j is correct per shot with probability $1-q$, independent of n.

Theorem 20.2 (Per-bit recovery threshold)**.** *Under the model (5) with $q < 1/2$, per-bit majority vote over S shots recovers all n bits of s with probability*

$$P(\textit{all bits correct}) \;\geq\; 1 - n\,\exp\Big(-2S\big(\tfrac{1}{2}-q\big)^2\Big), \tag{8}$$

so $S = \mathcal{O}\big((\frac{1}{2}-q)^{-2}\log(n/\delta)\big)$ shots give failure probability at most δ. The guarantee depends on the register width only through the $\log n$ union bound, not through any 2^n or $(1-q)^n$ factor.

Proof. For each bit j, the per-shot correct-bit indicator is Bernoulli$(1-q)$. The majority vote errs only if at most $S/2$ of the S shots are correct, i.e. if a Binomial$(S, 1-q)$ falls below its mean by $S(\frac{1}{2}-q)$. Hoeffding gives $\exp(-2S(\frac{1}{2}-q)^2)$ per bit, and a union bound over the n bits gives the claim. □

Theorem 20.2 is the correct structural-resilience statement at scale. It is what justifies operating at $n = 154$: the per-bit fidelity, not the full-string fidelity $F = (1-q)^n$, governs recovery, and the per-bit fidelity is bounded away from $1/2$ whenever the device readout error is.

20.4 Graceful degradation and the single-wrong-bit signature

Corollary 20.3 (Distribution of incorrect bits)**.** *Let p_b be the per-bit failure probability, bounded above by $\exp(-2S(\frac{1}{2}-q)^2)$ in Theorem 20.2. The number of incorrect recovered bits is $\mathrm{Bin}(n, p_b)$. In particular,*

$$P(\textit{exactly one wrong bit}) = n\,p_b\,(1-p_b)^{n-1}, \tag{9}$$

which is the dominant failure mode whenever $np_b = \mathcal{O}(1)$, and $P(\textit{full recovery}) = (1-p_b)^n$.

This matches the hardware observation at the strongest non-Clifford cell. At $(n, \mathrm{CCZ}) = (140, 10)$ the measured per-bit fidelity is 0.997 across all 11 trials, i.e. $p_b \approx 3\times10^{-3}$ and $np_b \approx 0.42$. Corollary 20.3 then predicts $P(\text{full recovery}) = (0.997)^{140} \approx 0.66$ and that the dominant failure carries exactly one wrong bit, with $P(\text{exactly one wrong}) \approx 0.42 \cdot 0.66 \approx 0.28$. The observed outcome (full recovery in 7/11 trials, with each of the four non-recoveries having exactly one wrong bit out of 140) is consistent with these predictions to within the small-sample Wilson interval. The single-wrong-bit failures are therefore not an artifact: they are the quantitative signature of per-bit recovery under bit-flip-dominated readout noise, predicted parameter-free by Corollary 20.3 from the per-bit fidelity alone.

20.5 General Hamming-ball noise

The bit-flip model is the product case of a broader class. Suppose the wrong mass is supported on a Hamming ball of radius r about ξ^*, with $p(\eta)$ nonincreasing in $d_H(\eta, \xi^*)$. Then the full-bitstring dominance ratio is controlled by the radius-1 shell and is at most $\binom{n}{1}^{-1}$ of the uniform-model ratio for each unit of concentration, so the 2^n protection degrades smoothly as the ball contracts toward ξ^*. The per-bit guarantee of Theorem 20.2 is unaffected as long as the induced marginal flip probability on each bit stays below $1/2$, which is the physically relevant condition. The practical

consequence for the experimental program is that per-bit voting should be the default recovery statistic at large n, and that the per-bit marginal flip probability (directly measurable from a readout-calibration matrix) is the quantity to report alongside the binary success rate. The next subsection treats the measured regime in which this ordering inverts.

20.6 Positively correlated noise: the joint mode outlives the per-bit vote

Every bound so far in this section assumes the per-bit errors are independent, and under that assumption the per-bit vote is always the more robust statistic (Theorem 20.2 versus Theorem 20.1). The measured frontier record inverts this order twice. On the degree-2 dual-oracle ladders (main text, Tables 3 and 5) at $n = 24$ the full-string vote recovers 10/10 (processor B 5/5 stable under multinomial resampling of the shot record; processor A 5/5, three of five unstable) while per-bit majority fails in 2/5 (processor B) and 4/5 (processor A) runs, and processor B's $n = 32$ recovery is vote-carried at three counts (the mode survives resampling with probability 0.02 naive, 0.79 tail-smoothed) with 5–10 of 32 marginals misread. On the conjugated single-monomial ladder (Section S14) the mode remains correct 8/8 at $C^{(10)}Z$–$C^{(11)}Z$ while the per-bit fidelity sits at 0.52–0.53, where the i.i.d. model puts the exact-string probability near 10^{-7} per shot. The measured exact-string probability exceeds the independent-bit product $\prod_i(1-q_i)$ by 3×10^3 to 4×10^4, the per-shot error weight is overdispersed by 1.4–2.1×, and the mean pairwise flip correlation is positive. This subsection proves that the direction of the inversion is generic for positively correlated readout noise, and exhibits an explicit regime in which it inverts the statistic order at every register width: positive association can only raise the joint-mode signal above the independent-bit prediction, while the per-bit statistic, a function of the marginals alone, gains nothing from it.

Throughout, $e = (e_1, \ldots, e_n) \in \{0,1\}^n$ is the per-shot error pattern relative to the correct outcome ξ^* (so the shot reads ξ^* exactly iff $e = 0$), $q_i = P(e_i = 1)$ are the marginal flip probabilities, $W = \sum_i e_i$ is the error weight, and $F = P(e = 0)$ is the single-shot exact-string probability.

Lemma 20.4 (Overdispersion is net pairwise covariance)**.** *Let $\kappa := \mathrm{Var}(W)/\sum_i q_i(1-q_i)$ be the overdispersion of the error weight relative to independent bits, assuming some $q_i \in (0,1)$. Then*

$$\kappa - 1 \;=\; \frac{\sum_{i\neq j}\mathrm{Cov}(e_i,e_j)}{\sum_i q_i(1-q_i)}, \tag{10}$$

so $\kappa > 1$ holds iff the summed pairwise flip covariance is positive.

Proof. $\mathrm{Var}(W) = \sum_i \mathrm{Var}(e_i) + \sum_{i\neq j}\mathrm{Cov}(e_i,e_j)$ and $\mathrm{Var}(e_i) = q_i(1-q_i)$. □

The measured $\kappa = 1.4$–2.1 is therefore a direct count-level witness of net positive correlation. It is also a necessary consequence of the structural hypothesis we now impose: the flip indicators are *(positively) associated* in the sense of Esary, Proschan, and Walkup [24], i.e. $\mathrm{Cov}(f(e), g(e)) \geq 0$ for every pair of coordinatewise nondecreasing $f, g\colon \{0,1\}^n \to \mathbb{R}$. Independent indicators are associated, nondecreasing functions of associated (in particular, of independent) variables are associated, and association implies every pairwise covariance is nonnegative, hence $\kappa \geq 1$ [24].

Theorem 20.5 (Association favors the joint mode)**.** *If $(e_1, \ldots, e_n)$ are associated with marginals q_i, then*

$$F \;=\; P(e=0) \;\geq\; \prod_{i=1}^{n}(1-q_i). \tag{11}$$

Under any positively associated flip law the single-shot exact-string probability is at least the independent-bit prediction, while the per-bit marginals, and with them the per-bit threshold of Theorem 20.2, are unchanged.

Proof. If f, g are coordinatewise nonincreasing, then $-f, -g$ are nondecreasing and $\mathrm{Cov}(f, g) = \mathrm{Cov}(-f, -g) \geq 0$, so indicators of decreasing events are nonnegatively correlated. The events $A_k = \{e_k = 0\}$ and $\bigcap_{i<k} A_i$ are decreasing, hence $P\big(\bigcap_{i\leq k} A_i\big) \geq P\big(\bigcap_{i<k} A_i\big)\,(1-q_k)$, and induction over k gives $P(e = 0) \geq \prod_i (1 - q_i)$. □

Association guarantees the direction of the effect but not its magnitude. The measured excess (3×10^3–4×10^4 over the product) is the empirical strength of the association on the device. The next two results show the magnitude can be large enough to invert the statistic order at every register width, and convert any measured margin into a shot budget.

Lemma 20.6 (Finite-shot mode certification)**.** *Let the per-shot outcome law have margin* $\Delta := F - \max_{\eta\neq\xi^*} p(\eta) > 0$. *With S independent shots, the empirical mode fails to equal ξ^* with probability at most* $2^n \exp\big(-S\Delta^2/(4F + 2\Delta)\big)$. *In particular* $S = (4F + 2\Delta)\,\Delta^{-2}\,\big(n \ln 2 + \ln(1/\delta)\big)$ *shots suffice for failure probability δ. This is the main text's Theorem 2, proved in Section S3 as Theorem 3.1.*

Proof. Theorem 3.1 (Section S3), whose proof is Bernstein's inequality for the count difference against each rival and a union bound over the $2^n - 1$ rivals. □

Proposition 20.7 (Burst mixture: an explicit inversion regime)**.** *Fix $\beta \in (0, 1)$ and $z \in (0, 1)$. Let $B \sim$ Bernoulli(β) be a shot-level burst indicator and $Z_1, \ldots, Z_n$ i.i.d.* Bernoulli(z) *independent of B, and set $e_i = B\,Z_i$ (errors occur only in burst shots, and then strike each bit at rate z). Then:*

(a) (e_i) are associated, with marginals $q_i = \beta z$;

(b) $F = 1 - \beta + \beta(1 - z)^n \geq 1 - \beta$ for every n, exceeding the independent-bit product $(1 - \beta z)^n$ by a factor of at least $(1 - \beta)(1 - \beta z)^{-n}$, exponential in n;

(c) every wrong outcome has mass $p(\eta) = \beta\, z^{|\eta|}(1 - z)^{n-|\eta|} \leq \beta\, c^{n-1}$ with $c := \max(z, 1 - z) < 1$, so for $n \geq 1 + \ln\big(2\beta/(1 - \beta)\big)/\ln(1/c)$ the dominance margin satisfies $\Delta \geq (1 - \beta)/2$ and Lemma 20.6 certifies the joint mode from $S = \mathcal{O}(n + \log(1/\delta))$ shots;

(d) if $\beta z > 1/2$, per-bit majority converges to the wrong value on every bit as $S \to \infty$, and if $\beta z < 1/2$ the per-bit guarantee of Theorem 20.2 costs $\mathcal{O}\big((\frac{1}{2} - \beta z)^{-2} \log(n/\delta)\big)$ shots; e.g. $\beta = 0.7$, $z = 0.75$ gives per-bit fidelity $1 - q_i = 0.475 < 1/2$, a regime where the joint mode is certified from $\mathcal{O}(n)$ shots while bitwise voting fails at every shot count;

(e) the error weight is overdispersed with $\kappa = \big[(1 - z) + nz(1 - \beta)\big]/(1 - \beta z)$, greater than 1 for $n \geq 2$ and growing with n, reproducing the sign of the measured overdispersion signature.

Proof. (a) Each $e_i = BZ_i$ is a nondecreasing function of the independent variables $(B, Z_1, \ldots, Z_n)$, so the family is associated [24], and $q_i = P(B = 1)P(Z_i = 1) = \beta z$. (b) $P(e = 0) = P(B = 0) + P(B = 1)(1 - z)^n$. The ratio bound follows from $F \geq 1 - \beta$. (c) For $\eta \neq 0$ the pattern requires $B = 1$, giving the stated product, and $z^d(1 - z)^{n-d} \leq c^n \leq c^{n-1}$ since each of the n factors is at most c and $c < 1$. The condition on n makes $\beta c^{n-1} \leq (1 - \beta)/2$, so $\Delta \geq F - \beta c^{n-1} \geq (1 - \beta)/2$. With $\Delta = \Theta(1)$ and $F \leq 1$ the shot bound of Lemma 20.6 is $\mathcal{O}(n + \log(1/\delta))$. (d) The marginal of each bit is Bernoulli(βz). A majority vote on a marginal strictly above $1/2$ is wrong with probability tending to one, and Theorem 20.2 gives the $(\frac{1}{2} - \beta z)^{-2}$ cost below it. (e) $\mathrm{Var}(W) = \mathbb{E}\,\mathrm{Var}(W \mid B) + \mathrm{Var}(\mathbb{E}[W \mid B]) = \beta nz(1 - z) + \beta(1 - \beta)n^2 z^2$, while $\sum_i q_i(1 - q_i) = n\beta z(1 - \beta z)$. Divide. □

Association is a structural property of the device noise, witnessed at the count level by $\kappa > 1$ and the nonnegative pairwise covariances (both measured), and it fixes the *direction* of the effect, joint mode at or above the independent-bit prediction, at unchanged per-bit thresholds. It does

not fix the magnitude, which is device-specific. The burst mixture shows the magnitude can invert the statistic order at every width, and Lemma 20.6 turns any measured margin into a shot budget. The practical consequence refines the advice of the previous subsection: report both statistics. The per-bit fidelity is the conservative frontier indicator (it cannot benefit from correlation), the joint mode is the correlation-robust recovery statistic, and their divergence is itself a noise diagnostic: per-bit above mode indicates independent-like flips (Corollary 20.3), mode above per-bit indicates clustered errors, the regime measured on both hardware records here.

20.7 The measured margin at the frontier

The margin $\hat{\Delta} = \hat{F} - \max_{y \neq s} \hat{p}_y$ is estimated from the same counts as the vote and is descriptive. In a recovered run it is positive by construction, since recovery means the answer out-counted every rival, so the point estimate repeats the vote and only the interval adds information. A run whose interval includes zero is one whose best rival the record cannot separate from the answer. Table 16 lists every frontier run of the degree-2 record with the shot budget that Lemma 20.6 (the main text's Theorem 2) demands at the point estimate and, beside it, the budget of the Hoeffding form (Section S3, remark). Under the lemma the bound holds at the observed 16,384 shots in ten of ten runs at $n = 20$ and in none at $n = 24$ or 32. Under the Hoeffding form it holds in four runs at $n = 20$ and in none beyond. At $n = 32$ a wrong string out-counted the answer in eight of ten runs, so the margin is negative and no budget is defined there. The numbers come from the counts retrieved for the committed job identifiers (seed 20260921), with no new run. The record holds no degree-2 counts at these sizes on processor C.

Table 16: Measured margins at the degree-2 frontier ($S = 16{,}384$ shots per run). $\hat{F}$: the answer's share of the shots. $\hat{p}$: the best rival's share, with its Hamming distance d from the answer. $\hat{\Delta} = \hat{F} - \hat{p}$ with a bootstrap 95% interval (2,000 multinomial resamples of the run's empirical distribution). The interval is a percentile bootstrap with the best rival re-selected in each replicate, which biases the replicate margins downward, and a string unseen in the record cannot appear in a replicate. At $n = 32$ the record is almost all singletons, so the resampled best rival carries six to eight counts against the one or two in the record. Every $n = 32$ interval therefore lies below its point estimate and measures the resampling spread of the best rival's count alone. μ_c/μ_w: the count ratio. $S_{0.05}$: the shot budget for failure probability 0.05 at the point estimate, under the Hoeffding form $2\hat{\Delta}^{-2}(n \ln 2 + \ln 20)$ (H) and under Lemma 20.6, $(4\hat{F} + 2\hat{\Delta})\hat{\Delta}^{-2}(n \ln 2 + \ln 20)$ (B), each with whether it holds at the observed S, undefined where $\hat{\Delta} \le 0$. Stable: the planted shift stays the unique mode under resampling with probability at least 0.95 (main-text Extended Data Table 8). The lemma's bound holds at the observed S in ten runs at $n = 20$ and in none at $n = 24$ or 32, the Hoeffding form in four runs at $n = 20$.

Device	n	run	$\hat{F}$	$\hat{p}$	d	$\hat{\Delta}$	95% interval	μ_c/μ_w	$S_{0.05}$ (H)	holds	$S_{0.05}$ (B)	holds	stable
B	20	1	0.111	0.0112	1	0.0995	[0.0945, 0.104]	9.86	3.4×10^3	yes	1.1×10^3	yes	yes
B	20	2	0.0864	0.014	4	0.0724	[0.0677, 0.0771]	6.18	6.4×10^3	yes	1.6×10^3	yes	yes
B	20	3	0.0457	0.00885	1	0.0369	[0.0331, 0.0399]	5.17	2.5×10^4	no	3.2×10^3	yes	yes
B	20	4	0.075	0.00983	5	0.0651	[0.0608, 0.0692]	7.63	7.9×10^3	yes	1.7×10^3	yes	yes
B	20	5	0.0454	0.00891	3	0.0365	[0.0325, 0.0394]	5.1	2.5×10^4	no	3.2×10^3	yes	yes
A	20	1	0.0344	0.00458	1	0.0298	[0.027, 0.0328]	7.52	3.8×10^4	no	3.7×10^3	yes	yes
A	20	2	0.0571	0.012	1	0.0451	[0.0411, 0.0492]	4.77	1.7×10^4	no	2.6×10^3	yes	yes
A	20	3	0.062	0.00964	1	0.0524	[0.0483, 0.0565]	6.43	1.2×10^4	yes	2.2×10^3	yes	yes
A	20	4	0.0438	0.00507	1	0.0387	[0.0355, 0.042]	8.64	2.3×10^4	no	2.8×10^3	yes	yes
A	20	5	0.0468	0.00415	4	0.0427	[0.0392, 0.0458]	11.3	1.9×10^4	no	2.5×10^3	yes	yes
B	24	1	0.00201	8.5×10^{-4}	5	0.00116	$[3.1 \times 10^{-4}, 0.00195]$	2.36	2.9×10^7	no	1.5×10^5	no	yes
B	24	2	0.00232	9.2×10^{-4}	7	0.0014	$[5.5 \times 10^{-4}, 0.0022]$	2.53	2.0×10^7	no	1.2×10^5	no	yes
B	24	3	0.00153	7.9×10^{-4}	5	7.3×10^{-4}	[0, 0.0014]	1.92	7.3×10^7	no	2.8×10^5	no	yes
B	24	4	0.00653	0.0011	7	0.00543	[0.00409, 0.00659]	5.94	1.3×10^6	no	2.5×10^4	no	yes
B	24	5	0.00244	5.5×10^{-4}	7	0.00189	$[9.8 \times 10^{-4}, 0.00262]$	4.44	1.1×10^7	no	7.4×10^4	no	yes
A	24	1	7.9×10^{-4}	1.8×10^{-4}	1	6.1×10^{-4}	$[-6.1 \times 10^{-5}, 7.9 \times 10^{-4}]$	4.33	1.1×10^8	no	2.3×10^5	no	yes
A	24	2	3.7×10^{-4}	2.4×10^{-4}	1	1.2×10^{-4}	$[-3.7 \times 10^{-4}, 2.5 \times 10^{-4}]$	1.5	2.6×10^9	no	2.3×10^6	no	no (0.41)
A	24	3	0.00275	9.2×10^{-4}	7	0.00183	$[9.2 \times 10^{-4}, 0.00275]$	3	1.2×10^7	no	8.6×10^4	no	yes
A	24	4	1.8×10^{-4}	1.8×10^{-4}	7	0	$[-4.3 \times 10^{-4}, 0]$	1	undefined	no	undefined	no	no (0.14)
A	24	5	4.9×10^{-4}	1.8×10^{-4}	7	3.1×10^{-4}	$[-2.4 \times 10^{-4}, 4.3 \times 10^{-4}]$	2.67	4.2×10^8	no	5.4×10^5	no	no (0.78)
B	32	1	0	1.2×10^{-4}	11	-1.2×10^{-4}	$[-4.9 \times 10^{-4}, -3.7 \times 10^{-4}]$	0	undefined	no	undefined	no	no (0.00)
B	32	2	0	1.2×10^{-4}	5	-1.2×10^{-4}	$[-4.9 \times 10^{-4}, -3.7 \times 10^{-4}]$	0	undefined	no	undefined	no	no (0.00)
B	32	3	0	6.1×10^{-5}	20	-6.1×10^{-5}	$[-4.9 \times 10^{-4}, -3.7 \times 10^{-4}]$	0	undefined	no	undefined	no	no (0.00)
B	32	4	0	6.1×10^{-5}	20	-6.1×10^{-5}	$[-4.9 \times 10^{-4}, -3.7 \times 10^{-4}]$	0	undefined	no	undefined	no	no (0.00)
B	32	5	1.8×10^{-4}	6.1×10^{-5}	15	1.2×10^{-4}	$[-4.3 \times 10^{-4}, 0]$	3	3.4×10^9	no	1.6×10^6	no	no (0.79)
A	32	1	0	6.1×10^{-5}	17	-6.1×10^{-5}	$[-4.9 \times 10^{-4}, -3.7 \times 10^{-4}]$	0	undefined	no	undefined	no	no (0.00)
A	32	2	0	6.1×10^{-5}	15	-6.1×10^{-5}	$[-4.9 \times 10^{-4}, -3.7 \times 10^{-4}]$	0	undefined	no	undefined	no	no (0.00)
A	32	3	6.1×10^{-5}	6.1×10^{-5}	17	0	$[-4.9 \times 10^{-4}, -1.8 \times 10^{-4}]$	1	undefined	no	undefined	no	no (0.26)
A	32	4	0	6.1×10^{-5}	17	-6.1×10^{-5}	$[-4.9 \times 10^{-4}, -3.7 \times 10^{-4}]$	0	undefined	no	undefined	no	no (0.00)
A	32	5	0	6.1×10^{-5}	19	-6.1×10^{-5}	$[-4.9 \times 10^{-4}, -3.7 \times 10^{-4}]$	0	undefined	no	undefined	no	no (0.00)

Remark 20.8 (Randomized planting)**.** *Assume the error string $y_i \oplus t_i$ is independent of the plant. Planting $t_i = A_i^{-1}s$ with a fresh uniformly random invertible $A_i \in \mathrm{GL}(n, 2)$ in every shot, and decoding the output y_i as $A_i y_i$, then makes the decoded wrong mass uniform over the $2^n - 1$ wrong strings, because $\mathrm{GL}(n, 2)$ acts transitively on nonzero vectors. The premise of Corollary 3.3 holds by construction and the vote reduces to counting exact hits of s. This paper does not use this construction. The fixed-plant mode is a statement about the processor's output distribution, which the randomization erases, and any dependence of the error on the plant, through the compiled circuit or through readout asymmetry, voids the independence hypothesis, and this remark says nothing about the size of that departure. No experiment was run.*

21 S21: Cross-Architecture Demonstration

Table 17: Processor labels. Each letter stands for one IBM processor throughout the main text and this Supplement.

Label	Family	Revision	Qubits
A	Heron	r2	156
B	Heron	r3	156
C	Nighthawk	R2	120
D	Heron	r2	156
E	Nighthawk	R1	120
F	Heron	r2	156
G	Heron	r1	133
H	Heron	r2	156
I	Nighthawk	R1	120

Every experiment in the main record (Sections S9–S11) runs on IBM Heron processors. We repeated the benchmark on processor E, an IBM Nighthawk R1 processor: 120 qubits, degree-4 connectivity (218 coupling edges, versus the heavy-hex degree-3 of Heron), median two-qubit error $\approx 2.8 \times 10^{-3}$ and median readout error $\approx 1.9 \times 10^{-2}$ at the calibration used. The protocol and recovery criteria are unchanged. After the routing audit of Section S8 voided the frontier cells of that run, we reran the same six cells on processor C (Nighthawk R2) on a vetted coupling map under a pre-registration filed before submission. The rerun is reported after the R1 material below. This is reported as a stand-alone cross-architecture demonstration. Per the pre-registration, it is *not* merged into the five-Heron discrimination fit of Section S16 (Spearman $\rho = -0.90$), because the degree-4 connectivity changes transpiled depth and the point is therefore not directly comparable on measured 2Q error alone.

Execution-time-fitted job sizes. Nighthawk R1 enforces a tighter per-job maximum execution time than Heron. The $192 \times 16{,}384$-shot jobs used at scale on Heron exceed it (error code 1305). We therefore used fitted job sizes on Nighthawk: degree-2 at $Q = n + 32$ rounds and 2,048 shots (256 shots at $n \leq 32$), and non-Clifford at $Q = 256$ rounds and 2,048 shots. For the rank-based degree-2 recovery the shot count is not binding (Section S9). This is a documented deviation from the shot budget used elsewhere, with recovery criteria unchanged. The non-Clifford circuits use the $B = I$ deep-genuine construction of Section S11 (with its stated factorization caveat).

Reading. The degree-2 benchmark recovers the planted shift on a second IBM architecture through $n = 116$ (Table 18), establishing that verifiable structured-circuit execution at scale is not specific to the Heron line. The non-Clifford frontier cells (Table 19) returned no recovery, but the routing audit finds 14 of the 16 circuits routed over couplers reported dead or above 10^{-2} at submission, so the collapse and the wrong-bit gradient reported in earlier versions carry no weight as device evidence on their own, and the R1 tables remain a contaminated record. The pre-registered vetted-map rerun on processor C (below) returned the same collapse and the same gradient on clean-routed circuits, so the frontier statement now rests on that device, alongside the pre-registered out-of-sample discrimination test, whose jobs the audit also finds clean.

A second square-lattice processor. To separate architecture from device, we repeated the run on a second Nighthawk R1, processor I (eu-de instance; 120 qubits, degree-4, median two-qubit

Table 18: Degree-2 benchmark on a second architecture. Full-rank recovery of the planted shift on processor E (IBM Nighthawk) at every cell, through $n = 116$, at single-shot fidelity of order 10^{-3}. Three fresh oracle/shift seeds per cell. Routing audit (Section S8): one $n = 64$ and all six $n = 100$, 116 circuits carried 3–331 CZ on couplers reported at error 1.0. The recoveries stand (extra noise can only lower a recovery rate), but calibration-based per-gate attribution is not meaningful for those cells.

n	Q	shots	runs	success	rank
16	64	256	3	3/3	16.0
32	64	256	3	3/3	32.0
64	96	2048	3	3/3	64.0
100	132	2048	3	3/3	100.0
116	148	2048	3	3/3	116.0
Total			15	15/15	

Table 19: Non-Clifford frontier cells on a second architecture. Processor E (IBM Nighthawk), $B = I$ deep-genuine, $Q = 256$, 2,048 shots, three seeds per cell. **Routing-contaminated (2026-09-10 audit, Section S8).** The seven $n = 100$ circuits were routed over couplers reported dead (error 1.0, 2–16 of 178–268 CZ per circuit, dominating the two-qubit error budget) and are void (†). At $n = 80$, seven of nine circuits carry 7–13 CZ on couplers above 10^{-2} (maximum 5.3×10^{-2}, 5–45% of the budget) and are flagged (‡). Only the two 8-CCZ seeds marked clean are unaffected. The collapse and the wrong-bit gradient reported in earlier versions are therefore not evidence about this device. The pre-registered rerun on a vetted coupling map (Table 21) reproduced both on clean-routed circuits. These cells were never pooled into the 156/259 Heron non-Clifford record.

n	CCZ	success	mean wrong bits
80	6	$0/3^{\ddagger}$	11
80	8	0/3 (2 clean, $1^{\ddagger}$)	24
80	10	$0/3^{\ddagger}$	42
100	6	$0/3^{\dagger}$	23
100	8	$0/3^{\dagger}$	34
100	10	$0/1^{\dagger}$	45
Total		0/16 (2 clean)	

error $\approx 2.2 \times 10^{-3}$, cleaner than processor E's 2.9×10^{-3}), with the same cells and three fresh seeds at 512 shots ($Q = n + 32$ for degree-2, $Q = 128$ for non-Clifford). Degree-2 reproduces: 15/15 full-rank recovery through $n = 116$, matching processor E cell for cell, so verifiable structured execution at scale holds on a second device of the architecture (the routing audit finds the $n \geq 64$ degree-2 circuits on both devices routed over couplers reported at error 1.0. The recoveries stand, but per-gate attribution for those cells is not meaningful, and recovery with 90–96% of the nominal two-qubit error budget on such couplers indicates that the reported 1.0 marks uncalibrated rather than physically dead couplers). The non-Clifford frontier again returned 0/18 across the same six cells, but the audit finds 16 of the 18 processor I frontier circuits routed over couplers reported dead and the remaining two over couplers at 0.11–0.12 (52–68% of the error budget), as on processor E, so neither the collapse nor the wrong-bit comparison of Table 20 is evidence about either device. The within-architecture ranking claimed in earlier versions (processor I above processor E) stays withdrawn. The single-device rerun below cannot restore a two-device comparison.

Table 20: Second-architecture devices at the non-Clifford frontier ($B = I$ deep-genuine, three seeds per cell): mean recovered wrong bits, lower is better. **Routing-contaminated on both R1 devices (Section S8). Those two columns are retained for the record only.** The fourth column is the clean-routed processor C rerun of Table 21. Binary recovery is 0/3 in every cell on all three devices. The graded per-bit metric ranks the cleaner R1 device (processor I, 2Q 2.2×10^{-3}) above processor E (2.9×10^{-3}) at every cell, a comparison that the contamination leaves without weight. The clean-routed counts are of the same size as the contaminated ones.

n	CCZ	I	E	processor C (vetted map)
80	6	4.3	11	9.0
80	8	12.3	24	23.7
80	10	24.3	42	36.7
100	6	9.7	23	12.7
100	8	20.3	34	27.7
100	10	30.7	45	52.0

Vetted-map rerun of the frontier cells (11 September 2026). A pre-registration filed before submission to processor C (rerun R1, about 9.4×10^6 shots) fixed the test. It kept the same six cells, the $B = I$ deep-genuine construction, $Q = 256$ rounds at 2,048 shots per circuit, and optimization level 2 with the sampler defaults of the original run, with three fresh shift seeds per cell (101–103, oracle seed unchanged). We transpiled on a vetted coupling map built from the live calibration at submission. The map drops couplers with error above 10^{-2} or reported dead, and couplers touching qubits with readout error above 2.5×10^{-2} or without T_1/T_2 (41 of 218 couplers, qubit 17 and its four couplers among them), leaving one connected component of 110 qubits and 177 couplers. Before submission we audited every ISA circuit coupler by coupler against the vetted set, and a single gate off the set would have refused the submission. All 18 passed, using 165–301 CZ on couplers of maximum error 9.0×10^{-3} (snapshot median two-qubit error 1.86×10^{-3} over the device, 1.61×10^{-3} over the vetted set). An independent audit of the submitted circuits against the job-time calibration (the program tool of Section S8) classifies all 18 as clean. Job ids and per-coupler tables are in the tracked manifests.

Three decision rules were registered. (i) The frontier-collapse statement returns only if all six cells give 0/3 on clean-routed circuits. They did: 0/18 overall (Wilson 95% upper bound 0.18, Table 21). (ii) The wrong-bit gradient returns only if mean wrong bits increase monotonically with CCZ count at both n. They do: $9.0 \to 23.7 \to 36.7$ at $n = 80$ and $12.7 \to 27.7 \to 52.0$ at $n = 100$. The 10-CCZ cell at $n = 100$ sits at the random floor (per-bit fidelity 0.48–0.50, and a shot-level per-bit decode gives 50–53 wrong bits). (iii) The reading "Nighthawk ranked below Heron" returns only if the pooled per-bit fidelity falls below the frozen five-Heron band of Section S16 evaluated at the snapshot median error (0.563 ± 0.026). The pooled value is 0.589, at the upper edge of the band, so that reading stays withdrawn. This third test is a coarse comparator here and weaker than the out-of-sample test below, whose cells match the fit's. These cells are $n = 80$–100 with 165–301 two-qubit gates and no dynamical decoupling, so we report the value as registered and claim no ordering from it.

The rerun settles the attribution question the audit raised. The collapse at the non-Clifford frontier and its graded wrong-bit gradient are properties of these cells on the Nighthawk architecture at current error rates, reproduced on clean routing on a second-generation device. The record now carries that statement at full strength on processor C. The contaminated processor E and processor I tables stay as the record of those runs, and their two-device ranking stays withdrawn.

Table 21: Pre-registered vetted-map rerun (R1) of the non-Clifford frontier cells on processor C (Nighthawk R2), 11 September 2026: $B = I$ deep-genuine, $Q = 256$, 2,048 shots, three fresh seeds per cell, every circuit clean by per-coupler audit before and after submission. Success is exact recovery of the planted shift under the decoder of Section S11. Wrong bits and per-bit fidelity (0.50 is the random floor) are cell means over the three seeds. 2Q is the mean transpiled two-qubit count.

n	CCZ	2Q	success	mean wrong bits	per-bit fid.
80	6	171	0/3	9.0	0.696
80	8	196	0/3	23.7	0.582
80	10	241	0/3	36.7	0.515
100	6	200	0/3	12.7	0.662
100	8	228	0/3	27.7	0.592
100	10	279	0/3	52.0	0.489
Total			0/18		0.589

Out-of-sample test on the second architecture. Processor C (Nighthawk R2, 120 qubits, degree-4 connectivity) came online after the five-backend discrimination fit of Section S16 was frozen. A pre-registration filed before job submission (7 September 2026) fixed the test: the identical frontier cells $(n, d) \in \{(18, 8), (20, 9), (22, 10)\}$, three instance seeds by two shift trials per cell at 16,384 shots, XY4 dynamical decoupling, optimization level 3; the frozen linear fit of pooled per-bit fidelity against measured median two-qubit error over the five Heron backends ($\text{fid} = 0.640 - 41.4\, e_{2Q}$, largest residual 0.013); and an acceptance band of twice that residual (± 0.026) around the fit evaluated at the submission-snapshot median error, 2.0×10^{-3}. The predicted pooled fidelity is 0.557 ± 0.026.

The measured pooled per-bit fidelity is 0.609, with full-bitstring mode recovery 18/18 (Table 22). The device exceeds the acceptance band and every Heron point, including the cleanest (0.598 at 1.1×10^{-3}). The pre-registered decomposition separates two contributions. Routing: the degree-4 lattice transpiles the same cells to 35–38% fewer two-qubit gates than processor B. Layer stability: the gate-normalized effective error $\varepsilon_c = -\ln(2\,\text{fid}_c - 1)/G_c$ gives an effective-to-calibrated ratio $m = 1.75$, against 2.15 on processor B, a $\sim$1.2$\times$ gain. Per transpiled gate the device is noisier than processor B (3.5–4.0×10^{-3} against 2.3–2.7×10^{-3}). The fidelity lead comes from the gate count. The reading for Section S16 is a scoping one. The monotone fidelity-versus-error law holds within an architecture, at fixed connectivity. Across architectures the transpiled gate count enters, and the graded metric ranks devices by delivered circuit quality, where the calibration scalar alone would misrank this device below every Heron chip.

Table 22: Pre-registered out-of-sample discrimination test on processor C (Nighthawk R2, 7 September 2026) against the frozen processor B reference values. Per-bit fidelity pooled over three instance seeds and two shift trials per cell, 16,384 shots. G: mean transpiled two-qubit count.

	C		B (frozen)	
(n, d)	fid	G	fid	G
$(18, 8)$	0.623	353	0.609	569
$(20, 9)$	0.621	405	0.615	635
$(22, 10)$	0.584	514	0.571	814
pooled	0.609		0.598	

22 S22: Phase-State Parity Runs with a Fixed Matching (Hidden Matching Cells)

This section records the design, decoder and per-instance results of the one-way phase-state runs, in which the registered design fixed Bob's matching to the shift matching on the top qubit ($k = q - 1$). For that problem a classical one-way message of one bit succeeds with certainty, so the runs demonstrate no communication advantage and are retained as a prepare-and-measure record of parity success, a fidelity readout of the phase state through one Hadamard (Supplementary Table 1, the reclassified row).

22.1 Problem setting, record and scope

Every result of the main text concerns computation, where the oracle-access caveat is structural. Whoever runs the benchmark also holds the oracle circuit, and the public instances are classically easy. In the Hidden Matching problem [1], Alice holds $x \in \{0,1\}^N$ and sends one message to Bob, who holds a perfect matching M and must output an edge $(i, j) \in M$ with the parity $x_i \oplus x_j$. Any classical one-way protocol needs $\Omega(\sqrt{N})$ bits to succeed with a constant clearance over chance, an information-theoretic theorem with no hardness assumption. The bound carries an unspecified constant that depends on that clearance (the theorem's constant), so we quote $\lfloor\sqrt{N}\rfloor$ as the *scale* of the classical cost of the theorem's problem. The runs fixed Bob's matching to the shift matching on the top qubit ($k = q - 1$, registered as such). For that problem a classical one-way message of one bit succeeds with certainty (at most q bits if Bob's choice among the q shift matchings is private), so the $\Omega(\sqrt{N})$ bound does not apply to the runs, and the $\lfloor\sqrt{N}\rfloor$ figures quoted here are the scale of a bound for a different problem, not the classical cost of the problem run. The quantum protocol sends $q = \log_2 N$ qubits: the phase state $N^{-1/2}\sum_i (-1)^{x_i}|i\rangle$, the spectrally flat encoding of this paper ($\alpha = 1$). Noise enters through the benchmark's mechanism: with message fidelity F, Bob's parity bit is correct with probability $\frac{1}{2} + \frac{F}{2}$, so any constant fidelity keeps the parity success a constant above one half.

The protocol and decision rule were pre-registered before job submission. Success is the fraction of shots whose decoded parity is correct, pooled over independent random instances at fixed q, and a cell clears the registered success bar exactly when the Wilson 95% lower bound exceeds 0.55 (Section 22.2). The registered primary arm (processor B, no error suppression) clears the registered 0.55 bar at $q = 6$ (0.752). It fails at $q = 8$ and $q = 10$, the contingency the registration names, since arbitrary-x state preparation costs $\Theta(N)$ gates. An exploratory error-suppression arm lifts $q = 8$ only to 0.546, short of the bar (measurement-error mitigation moves it by +0.002, so the residual is gate error). The largest cell above the bar on the first architecture is the exploratory $q = 7$ extension, scored under the identical registered rule (0.689, Wilson $[0.684, 0.694]$, Fig. 6).

Because shots within an instance share Alice's string and one transpiled circuit, we also report the instance-level view. The eight $q = 7$ instances span 0.660–0.723, each one individually above the bar, with a t-based 95% interval on the instance mean of $[0.673, 0.706]$. The registered $q = 6$ instances are all above the bar, and the $q = 8$ instance mean stays below, confirming every pooled verdict at the coarser replication unit (Table 23). A cross-check on processor H reproduces the $q = 8$ ceiling as a Heron-class property.

A pre-registered rerun on processor C (Nighthawk R2, degree-4 connectivity) clears the bar at $q = 8$: success 0.5735, Wilson $[0.570, 0.577]$, with all eight instances individually above the bar and $q = 7$ at 0.722 (Fig. 6, Table 24). A second registered run with automatic layout reached 0.6024 $[0.599, 0.606]$ at 336 gates. The bar is cleared at $q = 8$ on the second architecture.

Three boundaries scope the record. First, this is a single-processor prepare-and-measure record

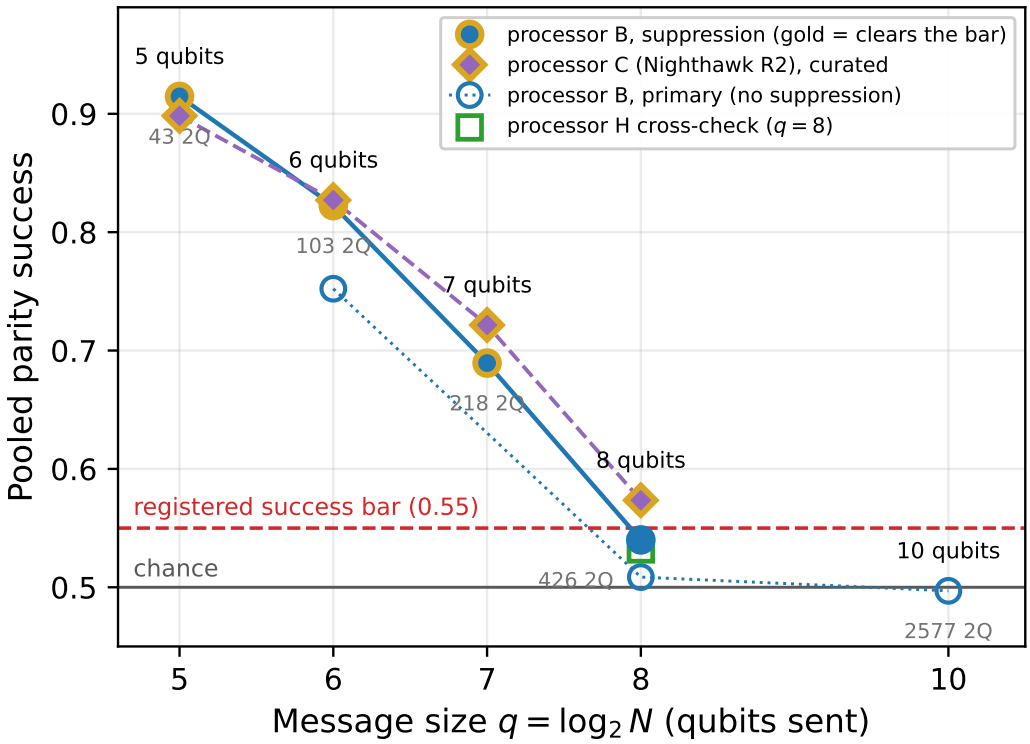


Figure 6: Parity success of one-way phase-state messages with a fixed matching (Hidden Matching cells). Pre-registered protocol and decision rule, pooled, 4096 shots per instance on the Heron devices, 8192 on processor C. Filled circles: processor B with error suppression. Filled diamonds: processor C (Nighthawk R2), curated qubit selection with the same suppression stack, in the pre-registered September 2026 rerun. Gold edges mark cells whose Wilson 95% lower bound clears the registered 0.55 bar. Open circles: the primary no-suppression arm. Green square: processor H cross-check. Annotations give the message size and the mean transpiled two-qubit gate count of the $\Theta(N)$-gate state preparation, whose growth sets the frontier. The highest cell above the bar is the $q = 8$ cell on processor C (0.5735). On processor B the highest is the exploratory $q = 7$ extension (0.689), with $q = 8$ at 0.546, short of the bar.

only. Alice's preparation and Bob's measurement execute on one device, certifying that a noisy processor sustains the quantum side of the protocol at message size $\log_2 N$. Second, the separation itself is the standing Bar-Yossef–Jayram–Kerenidis theorem, for a matching the sender does not know. The hardware result is that a real device reaches the clearance over one half the theorem concerns, at the stated message sizes and for one fixed matching, with the noise behavior the flatness mechanism predicts. Third, Kumar, Kerenidis and Diamanti [25] demonstrated a coherent-state variant photonically with Bob's matching sampled in the task and the success compared against the best-known classical protocol at equal message size. The runs here did neither: the design fixed Bob's matching and no classical baseline was computed, so they are a gate-based preparation of the phase state and a readout through one matching measurement on a superconducting processor. Each twofold reduction in effective two-qubit error yields roughly one additional qubit above the bar. Reaching $q = 10$ needs about a 3.6× reduction from the error measured on processor C, whose pre-registered projection gate returned no-go for $q = 10$. The protocol and decision rule stay registered unchanged for that re-run, and any future advantage claim draws Bob's matching

privately per instance from a family with a proven bound and states a computed classical baseline at equal message size.

22.2 Registered design and scoring rule

The registered design used

$$q \in \{6, 8\}$$

as primary targets and

$$q = 10$$

as a stretch target, with five random instances generated for each value of q. No error suppression was included in the registered protocol, and each instance was executed with 4096 shots. Success was evaluated using pooled decoded-parity accuracy. The registered success criterion required the Wilson 95% lower confidence bound of the pooled decoded-parity accuracy to exceed

$$0.55.$$

The decoder, per-instance statistics, and execution metadata are reported below and in the accompanying run manifests.

22.3 Exploratory extensions

Following completion of the registered design, additional runs were performed using

$$q \in \{5, 7\}$$

together with an error-suppression arm. The $q = 5$ and $q = 7$ cells were not part of the original registration, and the suppression arm is the optional arm the registration allows to be reported separately. All three are classified as exploratory. For consistency, the same registered success criterion was applied to the exploratory runs, but they are interpreted separately from the preregistered analysis.

22.4 Execution

The 5 June 2026 runs were executed on processor B, with a same-day cross-check on processor H, and the pre-registered 7 September 2026 rerun is reported below. Backend identifiers, job identifiers, decoder outputs, and instance-level results are provided in the tables below and the accompanying run manifests.

22.5 Interpretation and reclassification

The registered implementation fixed Bob's matching to the shift matching associated with the highest-index qubit. Under this restriction, a one-bit classical message suffices to solve the implemented task with certainty. Consequently, the experiment does not demonstrate a communication-complexity separation and cannot be interpreted as evidence for quantum communication advantage. The study is therefore reported as a protocol-validation exercise that tests the preparation, transmission, and decoding pipeline under the registered scoring rule. The complete experimental record, including exploratory extensions and cross-check runs, is retained for transparency and reproducibility.

22.6 Protocol and decoder

For message size $q = \log_2 N$, Alice's string $x \in \{0,1\}^N$ is encoded as the flat phase state $N^{-1/2}\sum_i (-1)^{x_i}|i\rangle$ on q qubits. Bob's matching is the shift matching $M = \{(i, i \oplus e_k)\}$ on the top qubit k: he applies a Hadamard on qubit k and measures all q qubits. The lower $q-1$ bits identify the edge (i,j) with $j = i \oplus e_k$, and bit k is the parity estimate; the shot is scored correct iff that bit equals $x_i \oplus x_j$. Success at a cell is the fraction of correct shots pooled over independent uniformly random instances of x (fixed seeds), 4096 shots per instance (8192 in the 7 September 2026 rerun below). The decoder uses only measured counts; x enters only for scoring.

22.7 Registered design versus extensions

The pre-registration (committed on the day of submission together with the queued job's identifier, before any result was fetched) fixes the metric above, the registered success criterion (a cell clears the bar at q when the Wilson 95% lower bound on pooled success exceeds 0.55), and the design: $q \in \{6, 8\}$ primary and $q = 10$ stretch, five instances each, no error suppression. The error-suppression arm (gate and measurement Pauli twirling, XY4 dynamical decoupling, optimization level 3, transpilation seed chosen by minimum two-qubit count before submission) and the $q = 5$ and $q = 7$ cells are same-day extensions, scored under the identical registered rule and labeled exploratory. The pre-registration's outcome section, written after the fetch, labels the $q = 7$ cell the same way. The $\Omega(\sqrt{N})$ classical bound [1] carries an unspecified constant that depends on the theorem's clearance over chance. All bit numbers quoted in the paper are $\lfloor\sqrt{N}\rfloor$, the scale of the bound, not a proved bit count. The runs fixed Bob's matching to the shift matching on the top qubit ($k = q-1$, registered as such). For that problem a classical one-way message of one bit succeeds with certainty (at most q bits if Bob's choice among the q shift matchings is private), so the $\Omega(\sqrt{N})$ bound does not apply to the runs, and the $\lfloor\sqrt{N}\rfloor$ figures quoted here are the scale of a bound for a different problem, not the classical cost of the problem run.

22.8 Per-instance results

Pooled Wilson intervals treat shots as independent, but shots within an instance share x, the matching, and one transpiled circuit. Table 23 therefore lists every instance-level success (post-hoc re-analysis of the recorded jobs, same metric). The verdicts are unchanged at this coarser replication unit: every $q = 7$ instance individually exceeds the bar (worst 0.660; t-based 95% interval on the instance mean $[0.673, 0.706]$), every $q = 6$ instance in both arms exceeds it, and the $q = 8$ instance means sit below the bar in every 2026-06-05 arm. The 2026-09-07 processor C rerun below is the exception: every $q = 8$ instance clears the bar.

22.9 Pre-registered rerun on the second architecture

A rerun on processor C (Nighthawk R2, 120 qubits, degree-4 connectivity) was pre-registered on 7 September 2026, before job submission, with protocol, endpoints, and decision gates fixed in the file. The run keeps the registered metric and decision rule unchanged and adds device curation: cross-cycle calibration with exclusion of consistently bad qubits and links, greedy selection of the lowest-error connected q-qubit subgraph (mean used-link calibrated error 1.04–1.10×10^{-3}), routing restricted to the cleaned coupling map, best-of-8 transpilation seeds by two-qubit count, XpXm dynamical decoupling, and gate and measurement Pauli twirling. Cells $q \in \{5, 6, 7, 8\}$ ran as one job (`dafibejdd5gc73d9sgsg`), eight instances per cell at 8192 shots (Table 24).

Table 23: Per-instance parity success, all Hidden Matching cells (2026-06-05 jobs at 4096 shots per instance and the 2026-09-07 processor C rerun at 8192; job identifiers in the run manifests).

Arm (backend)	q	Pooled	Per-instance successes
primary (processor B)	6	0.752	.675, .731, .772, .827, .757
primary (processor B)	8	0.509	.514, .522, .515, .499, .493
primary (processor B)	10	0.497	.493, .508, .482, .507, .495
suppression (processor B)	6	0.841	.851, .847, .828, .845, .851, .822
suppression (processor B)	8	0.546	.538, .535, .544, .556, .538, .568
suppression (processor B)	10	0.500	.502, .496, .495, .507, .510, .493
suppr.+seed (processor B)	6	0.822	.825, .810, .814, .821, .833, .815, .839, .823
suppr.+seed (processor B)	8	0.540	.540, .529, .524, .545, .546, .546, .550, .540
exploratory (processor B)	5	0.915	.922, .905, .915, .928, .916, .917, .907, .909
exploratory (processor B)	7	0.689	.676, .723, .704, .702, .687, .683, .681, .660
cross-check (processor H)	8	0.532	.521, .537, .525, .543, .525, .537, .539, .530
curated (processor C, NH R2)	5	0.898	.899, .894, .901, .902, .906, .893, .895, .897
curated (processor C, NH R2)	6	0.827	.831, .823, .814, .835, .833, .814, .844, .823
curated (processor C, NH R2)	7	0.722	.728, .721, .713, .728, .718, .720, .724, .721
curated (processor C, NH R2)	8	0.5735	.565, .578, .572, .586, .565, .580, .573, .570

Table 24: Pre-registered rerun of the phase-state parity cells on the second architecture, processor C (Nighthawk R2), 7 September 2026. Pooled success over eight instances at 8192 shots. A cell clears the registered success bar when its Wilson 95% lower bound exceeds 0.55. The bit-scale column is $\lfloor\sqrt{N}\rfloor$, the scale of a bound that does not apply to these runs (first paragraph of this section). G: mean transpiled two-qubit count of the state preparation.

q	bit scale	success	Wilson 95%	G	clears bar
5	5 (no gap)	0.898	[0.896, 0.901]	57	yes
6	8	0.827	[0.824, 0.830]	110	yes
7	11	0.722	[0.718, 0.725]	223	yes
8	16	0.5735	[0.570, 0.577]	412	yes

The $q = 8$ cell clears the registered bar for the first time on any device. The eight instances span 0.565–0.586, each one individually above the bar (the weakest instance's own Wilson lower bound is 0.554), and the t-based 95% interval on the instance mean is $[0.567, 0.580]$, entirely above the bar, so the pooled verdict holds at the instance replication unit as well (Table 23). The degree-4 connectivity cuts the $q = 8$ state preparation to 412 transpiled two-qubit gates, from 426–638 on heavy-hex Heron.

The pre-registration also fixed decision gates for larger cells. From the anchor cell (the largest q with pooled success at or above 0.55, here $q = 8$), the effective error per transpiled two-qubit gate is $\varepsilon_t = -\ln(2s-1)/G = 4.7\times10^{-3}$, and the projected success at $q^* \in \{10, 12\}$ is $\frac{1}{2}+\frac{1}{2}\exp(-\varepsilon_t G_{q^*})$ with G_{q^*} the curated transpile-only count recorded in the submission manifest ($G_{10} = 1504$, $G_{12} = 6289$). Both gates returned no-go (0.5005 and 0.5000 projected), so no $q \geq 10$ job was submitted. Clearing the $q = 10$ gate requires $\varepsilon_t \leq 1.3\times10^{-3}$, a 3.6× reduction from the measured anchor. The effective-to-calibrated multiplier is $m = 4.24$, against 5.1 on processor B's curated $q = 8$ anchor, the same ∼1.2× layer-stability gain as the discrimination test above. The registration bars shot-pooling escalation when a cell lands near the bar. None was needed.

A second registered $q = 8$ observation. The baseline cell of the depth-parallelization validation below (same decoder, decision rule, instance seeds, and shot budget; SABRE auto-layout at optimization level 3 in place of the curated subgraph, mean 336 transpiled two-qubit gates) reached pooled success 0.6024, Wilson $[0.599, 0.606]$, on 2026-09-08. Seven of its eight instances individually clear the bar (the eighth, on the worst-routed transpilation, sits at 0.535); the t-based 95% interval on the instance mean is $[0.579, 0.626]$, entirely above the bar. The auto-layout gate count, 336 against the curated 412, is the source of the gain; the effective error per transpiled gate is unchanged.

Engineering routes to $q \geq 9$ closed on this device (pre-registered). Three routes around the hardware-error requirement were tested and closed. (i) *Input restriction*: any input family of size 2^k caps the classical one-way cost at k bits (Alice sends the index), so the input entropy must stay at least $\sqrt{N}$ (for the problem as run this requirement is moot, first paragraph of this section) and the $\Theta(N)$-gate preparation is already within a factor $\sim$1.6 of the one-bit-per-gate floor. (ii) *Error-budget decomposition* (`code/PREREGISTRATION_phoenix_m_decomposition.md` and its 1b amendment): stretching the $q = 8$ circuits at fixed gate content shows the decoupled baseline carries $\sim$0.69 of its log-error budget outside the calibrated gate floor, and decoupling itself removes 2.21 of the bare 4.00; the registered linear fit is reported with a model-violation caveat (the log-error is convex in duration). (iii) *Depth-parallel preparation* : the diagonal was split into Gray-code chunks executed concurrently on CNOT-fanned register copies (statevector-verified construction). The registered gate failed decisively: log-error ratios of 2.4–5.6 against gate-count ratios of 1.5–3.3 at $q \in \{6, 7, 8\}$, $m \in \{2, 4\}$. The excess over the gate count is the expected dephasing amplification of the fanned-out state: a phase error on any copy acts on the joint amplitude, so the dephasing-sensitive volume grows by the same factor m that divides the depth. Depth parallelization therefore cannot recover the idle error for this preparation, and the 3.6× error reduction stands as the requirement for $q = 10$ on future hardware.

23 S23: Powered Mean-Flatness Sweep, Per-Cell Record

This section reports the preregistered validation study of Proposition 7, which predicts that the planted shift is returned with probability

$$P(s) = \bar{\alpha}^2,$$

where

$$\bar{\alpha} = 2^{-n/2} \, \mathrm{mean}_\xi \, |\hat{f}(\xi)|.$$

The study was designed to distinguish dependence on mean flatness from dependence on minimum flatness and to test whether the observed relationship replicates across devices.

The flatness dependence itself is tested with power on hardware at $n = 8$. A first sweep over perturbed-bent oracles spanning min-flatness $\alpha \in [0.38, 1.0]$ refuted the α^2 (min-flatness) hypothesis, but that construction keeps $\bar{\alpha}^2$ within $[0.92, 1.00]$, and every cell of the frontier record of the main text is bent ($\bar{\alpha}^2 = 1$). A second sweep therefore varied $\bar{\alpha}^2$ directly: five block-product oracle families at $n = 8$ with $\bar{\alpha}^2 \in \{1.00, 0.61, 0.47, 0.22, 0.20\}$, echo-padded to one compiled two-qubit depth of 167, pre-registered with its decision bands and scored in one look on processor B, then run under pre-filed replication addenda on processor A and on the Nighthawk processor C. The hypothesis that the planted-string probability tracks $\bar{\alpha}^2$ read CONFIRMED on processor B and, as a reading under the same bands applied to each device's own retention, REPLICATED on

the other two. The record below gives each device's slope, its excess against both references and its gate-matched control.

The measured count ratio μ_c/μ_w, the dominance ratio realized on hardware, therefore decomposes as $\bar{\alpha}^2\times$ retention $\times$ dilution, an accounting identity in which dilution is defined as the remainder S/μ_w, and each factor is now measured. The $\bar{\alpha}^2$ factor is measured at $n = 8$ and one compiled depth, in one session on each of three processors: log-log slopes of 0.94 (bootstrap 95% interval 0.93 to 0.96) on processor B, 0.88 (0.85 to 0.91) on processor A and 0.87 (0.83 to 0.91) on processor C, each against that device's own bent-cell retention r (0.34, 0.20, 0.22), with the low-flatness cells 6 to 13%, 12 to 31% and 14 to 30% above $r\,\bar{\alpha}^2$ (every such cell's Wilson 95% interval lies above it) and 10 to 25% above, 2% below to 15% above, and 12 to 28% above each device's gate-matched control, an excess that grows as $\bar{\alpha}^2$ falls on every device and is reported as an open observation with no cause assigned. The processor B interval lies above the other two, which overlap, so the slope is device dependent at the interval level, and the control gives no consistent direction across devices. The factor says nothing about the frontier, where every cell is bent and it equals one by construction. Retention is the measured planted-string probability at the frontier (3.0×10^{-3} and 9.2×10^{-4} at $n = 24$ on processor B and processor A). Dilution is S/μ_w, 9.1×10^2 to 5.5×10^3 at $n = 24$ against 1.7×10^7 under uniform redistribution.

23.1 Original registered hypothesis

The original preregistration proposed that planted-shift recovery would scale as

$$P(s) \propto \alpha^2,$$

where α denotes minimum spectral flatness. To test this proposal, an initial sweep was performed at $n = 8$ over perturbed-bent constructions spanning

$$\alpha \in [0.38, 1.00].$$

Ten hidden shifts were sampled per cell together with a depth-matched $\alpha = 1$ control.

23.2 Initial min-flatness sweep

Recovery remained near perfect throughout the explored range despite substantial variation in minimum flatness. Inspection of the resulting oracle families showed that, although α varied significantly, the corresponding values of

$$\bar{\alpha} = 2^{-n/2}\,\mathrm{mean}_\xi\,|\hat{f}(\xi)|$$

remained close to unity. Consequently, the experiment refuted the registered minimum-flatness hypothesis but could not distinguish between a dependence on $\bar{\alpha}^2$ and flatness-independent recovery. The study therefore falsified the original α^2 hypothesis but was underpowered to test Proposition 7 directly.

23.3 Powered mean-flatness experiment

To directly probe Proposition 7, five block-product oracle families were constructed with exact values

$$\bar{\alpha}^2 = 1.000,\ 0.610,\ 0.473,\ 0.223,\ 0.202.$$

The corresponding constructions were

1. quadratic bent,
2. $\mathrm{AND}_3 + \mathrm{AND}_3 + \mathrm{bent}_2$,
3. $\mathrm{AND}_4 + \mathrm{bent}_4$,
4. $\mathrm{AND}_4 + \mathrm{AND}_4$,
5. $\mathrm{AND}_5 + \mathrm{AND}_3$.

Each family preserves the planted shift as the unique noiseless mode while varying mean flatness. For every family,

- ten hidden shifts were sampled,
- each circuit was executed with 4096 shots,
- routing was restricted to a fixed path of couplers, and
- compiled two-qubit depth was matched by echo padding to 167 (achieved 164 to 168).

23.4 Registered decision rule

Prior to execution the following criterion was registered. The experiment would be classified as **CONFIRMED** if

1. the log-log slope of planted-string retention versus $\bar{\alpha}^2$ lay in

$$[0.7, 1.3],$$

   and

2. every cell lay within a factor 1.5 of the relation, with r the measured retention of the bent cell,

$$r\,\bar{\alpha}^2.$$

The experiment would be classified as **FALSIFIED** if the fitted slope fell below 0.3 or if the $\bar{\alpha}^2 = 0.223$ cell fell inside the Wilson 95% interval of the bent reference.

23.5 Results on the primary processor

Measured retention tracked $\bar{\alpha}^2$ closely across all five oracle families. The fitted slope was

$$0.94,$$

with bootstrap interval

$$[0.93, 0.96].$$

All cells satisfied the preregistered acceptance criteria. The planted shift remained the unique measurement mode in all fifty circuits. Accordingly, the registered outcome was classified as **CONFIRMED**.

23.6 Replication on a second processor

An independent replication was executed on processor A using the same oracle families, depth target, and analysis protocol. The fitted slope was

$$0.88,$$

with bootstrap interval

$$[0.85, 0.91].$$

The planted shift remained the unique output mode in all fifty circuits, and the preregistered success criterion was satisfied.

23.7 Replication on a second architecture

A cross-architecture replication was executed on processor C (Nighthawk R2). The fitted slope was

$$0.87,$$

with bootstrap interval

$$[0.83, 0.91].$$

The planted shift remained the unique output mode in all fifty circuits. The replication therefore satisfied the same preregistered criterion.

23.8 Gate-content controls

To determine whether the observed trend could arise solely from additional non-Clifford gate content, a depth-matched control family with

$$\bar{\alpha}^2 = 1$$

and one C^3Z gate per oracle was executed on each backend. Although the magnitude of the control effect varied across devices, the control did not reproduce the observed $\bar{\alpha}^2$ dependence. The measured trend therefore cannot be explained solely by differences in gate content.

23.9 Scope and limitations

The validation study probes the prediction

$$P(s) \propto \bar{\alpha}^2$$

at fixed register size and approximately fixed compiled depth. The study does not test frontier scaling, dilution among 2^n output strings, count-ratio growth with system size, dependence on echo-padding fraction, or asymptotic scaling of recovery probability. These questions remain open.

23.10 Supporting data

The complete numerical results, confidence intervals, bootstrap estimates, dominance ratios, and control comparisons are reported in the tables and figures that follow.

Registration (committed `19d4b3b`, 2026-09-15 19:39:26 UTC, before submission; the processor C addendum `68ee944`, 19:51:04 UTC, and the processor A addendum `f3cddd8`, 21:59:02 UTC, each before its submission; job ids committed before retrieval in every leg). Five block-product

oracle families at $n = 8$ (XOR of a bent form or an AND on each block of a partition of the eight variables) have exact mean flatness $\bar{\alpha}^2 = 1.000, 0.610, 0.473, 0.223, 0.202$ and the planted shift as the unique noiseless mode. Each family ran with ten shifts at 4,096 shots on a pinned eight-qubit path (routing restricted to the path's couplers, every coupler below 10^{-2}), every instance echo-padded with CZ pairs to a shared compiled two-qubit depth of 167 (achieved 164 to 168), with a degree-4 bent control (one C^3Z per oracle, $\bar{\alpha}^2 = 1$, same depth) carrying the AND families' gate content. $r = F_s$(bent) is the retention of the quadratic bent cell on that device. Registered reading: CONFIRMED (processor B) or REPLICATED (processor A and processor C, each on its own r) if the log-log slope of F_s on $\bar{\alpha}^2$ lies in $[0.7, 1.3]$ and every cell is within a factor 1.5 of $r\,\bar{\alpha}^2$; FALSIFIED if the slope is below 0.3 or the $\bar{\alpha}^2 = 0.22$ cell lies inside the bent cell's Wilson interval. The paths are qubits 132, 131, 138 and 151 to 155 on processor B and qubits 42 to 47, 57 and 67 on processor A, with routing restricted to each path's couplers. The registration, the layout hash and the frozen circuits were committed before submission and the job ids before retrieval (Table 26). Table 25 lists every cell. Slopes: 0.94 (bootstrap 95% interval 0.93 to 0.96, seeds resampled within cells, 2,000 replicates) on processor B, 0.88 (0.85 to 0.91) on processor A and 0.87 (0.83 to 0.91) on processor C (Nighthawk R2, square lattice, path 100-90-91-92-93-103-102-101). The processor B interval lies above the other two, which overlap each other. On every device every low-flatness cell's Wilson interval lies above $r\,\bar{\alpha}^2$, so the excess (6 to 13% on processor B, 12 to 31% on processor A, 14 to 30% on processor C, growing as $\bar{\alpha}^2$ falls) is resolved and is reported as an open observation with no cause assigned. The control sits below the bent cell on processor B (0.91, outside its interval), so gate content works against the AND families there and cannot produce the confirmation. On processor A it sits above (1.14), so that direction argument holds on processor B only. On processor C it sits inside the bent cell's interval (1.01), so gate content had no measurable effect there. The CZ-echo pad share (a median of 63 pairs in the bent reference down to 6 in the $\bar{\alpha}^2 = 0.20$ family) runs in the direction of the excess and no registered leg isolates it. Scope: $n = 8$, one compiled depth, one session and 245,760 shots per device. The sweep measures the $\bar{\alpha}^2$ factor of the count-ratio decomposition at that scale and says nothing about the $n = 24$ frontier, where every cell is bent and the factor equals one by construction.

24 S24: Registration Chronology of Every Hardware Claim

This section documents the registration status, execution chronology, and verification workflow of the hardware experiments reported in the manuscript. Its purpose is to provide an auditable record of when predictions were recorded relative to execution and how experimental results were processed.

24.1 Registration policy

Each registered hardware study is accompanied by a registration file specifying the experimental cells, decision rule, and expected outcome. Registrations are stored in the project repository and linked to hardware executions through submission manifests containing backend identifier, job identifier, submission timestamp, and per-circuit transpiled metrics. Unless explicitly noted below, the version-control record verifies that the relevant registration commit predates hardware-job creation. The Hidden Matching registration was committed about 78 seconds after its first job was queued, before any result was fetched.

Table 25: Powered mean-flatness sweep, every cell on all three processors (one scored look per device, 2026-09-15). F_s: pooled planted-string fraction over ten shifts with its Wilson 95% interval; $r\,\bar{\alpha}^2$: the prediction from that device's bent-cell retention (in parentheses for the control, whose own $\bar{\alpha}^2$ is 1); ratio: $F_s/(r\,\bar{\alpha}^2)$; ratio vs K1: the same with the gate-matched control as reference; dominance: mean planted count over mean best-wrong count; mode: circuits in which the planted shift is the unique full-string mode.

Device	Cell (blocks)	$\bar{\alpha}^2$	F_s (Wilson 95%)	$r\,\bar{\alpha}^2$	ratio	ratio vs K1	dominance	mode
B	C1 (bent_8)	1.000	0.339 [0.334, 0.343]	0.339	1.00	1.10	9.1	10/10
B	C2 (AND_3+AND_3+bent_2)	0.610	0.220 [0.216, 0.224]	0.207	1.06	1.17	5.9	10/10
B	C3 (AND_4+bent_4)	0.473	0.170 [0.166, 0.174]	0.160	1.06	1.17	4.7	10/10
B	C4 (AND_4+AND_4)	0.223	0.081 [0.079, 0.084]	0.076	1.08	1.19	4.2	10/10
B	C5 (AND_5+AND_3)	0.202	0.077 [0.075, 0.080]	0.068	1.13	1.25	3.7	10/10
B	K1 (degree-4 bent (control))	1.000	0.307 [0.303, 0.311]	(0.339)	0.91	1.00	7.3	10/10
A	C1 (bent_8)	1.000	0.198 [0.194, 0.202]	0.198	1.00	0.88	3.3	10/10
A	C2 (AND_3+AND_3+bent_2)	0.610	0.135 [0.132, 0.139]	0.121	1.12	0.98	3.5	10/10
A	C3 (AND_4+bent_4)	0.473	0.106 [0.103, 0.109]	0.094	1.14	1.00	3.1	10/10
A	C4 (AND_4+AND_4)	0.223	0.050 [0.048, 0.052]	0.044	1.13	0.99	2.8	10/10
A	C5 (AND_5+AND_3)	0.202	0.052 [0.050, 0.055]	0.040	1.31	1.15	2.9	10/10
A	K1 (degree-4 bent (control))	1.000	0.225 [0.221, 0.229]	(0.198)	1.14	1.00	5.5	10/10
C	C1 (bent_8)	1.000	0.219 [0.215, 0.223]	0.219	1.00	0.99	4.3	10/10
C	C2 (AND_3+AND_3+bent_2)	0.610	0.152 [0.148, 0.155]	0.133	1.14	1.12	3.9	10/10
C	C3 (AND_4+bent_4)	0.473	0.121 [0.118, 0.124]	0.103	1.17	1.16	3.6	10/10
C	C4 (AND_4+AND_4)	0.223	0.058 [0.056, 0.061]	0.049	1.19	1.18	3.3	10/10
C	C5 (AND_5+AND_3)	0.202	0.057 [0.055, 0.060]	0.044	1.30	1.28	3.1	10/10
C	K1 (degree-4 bent (control))	1.000	0.222 [0.218, 0.226]	(0.219)	1.01	1.00	5.4	10/10

24.2 Registered hardware studies

The following hardware studies were executed under preregistered decision rules:

1. Degree-2 frontier ladders.
2. Degree-3 non-Clifford replication batch.
3. Mitigation-ablation study.
4. Powered mean-flatness sweep.
5. Five-backend discrimination experiment.
6. Hidden Matching study.
7. Cross-device replication experiments.

The corresponding registration files record the cells, success criteria, and interpretation rules used in the manuscript.

24.3 Degree-2 ladder chronology

The degree-2 ladder studies require separate discussion. The processor B and processor A registrations contain internal filing dates of 7 July 2026 and 8 July 2026, respectively. The first repository commit containing these registration files is

```
aaee22d
```

dated 9 July 2026. Consequently, the version-control record independently certifies registration before execution for the processor B ladder but does not independently certify registration before execution for the processor A ladder. The processor A ladder should therefore be interpreted as a same-design replication whose registration record entered version control after data retrieval.

24.4 Submission manifests

For each submitted experiment, a manifest records

- backend identifier,
- job identifier,
- submission timestamp, and
- per-circuit transpiled metrics.

The manifests form the primary link between registration files and hardware execution records. The degree-2 ladder manifests carry file stamps of 7 and 8 July 2026 and entered version control on 10 September 2026.

24.5 Routing audit

A retrospective audit of all submitted circuits was performed on 10 September 2026 using the backend calibrations in force at submission. Failed trials routed through couplers that the calibration in force at submission reported dead were excluded from analysis, and recoveries on such couplers were kept. The audit removed

- ten $(14, 6)$ unbounded-degree trials,
- fifteen $(16, 7)$ unbounded-degree trials,
- three exploratory Protocol-C cells, and
- the Nighthawk-R1 frontier cells.

All figures, tables, and statistics reported in the manuscript use the corrected dataset resulting from this audit.

24.6 Updated submission workflow

Following the routing audit, the execution workflow was revised. The submission driver now

1. transpiles exclusively on a vetted coupling map,
2. records per-coupler two-qubit gate counts,
3. rejects circuits using excluded couplers,
4. stores circuit fingerprints before submission, and
5. writes submission manifests prior to retrieval.

This workflow was used for all experiments executed after the audit.

24.7 Analysis workflow

All analyses operate exclusively on measured counts. Recovery is performed without access to the planted hidden shift. Only after recovery has been completed is the planted shift revealed for scoring. The standard workflow is

1. register prediction and decision rule,
2. generate oracle instances from committed seeds,
3. submit circuits,
4. retrieve measured counts,
5. perform recovery,
6. compare the recovered string with the planted shift, and
7. generate summary statistics.

Failures and successful recoveries are both retained in the result record.

24.8 Recovery statistics

The mode is computed as Python's `max` over the counts dictionary, so a tie resolves to the first tied outcome in the dictionary's insertion order, which the device output fixes before the planted value enters the analysis. This tie rule belongs to the analysis code, and the registration names none. The single 3-count tie in the record is disclosed in main-text Table 3 and scored both ways. Bit j of the per-bit majority string is 1 when at least half of the shots read 1 at bit j. The registration of the degree-2 ladders names the per-bit vote as the primary statistic and the full-string vote as secondary.

24.9 Statistical intervals

Cell success rates carry Wilson 95% intervals (0.57 lower bound for 5/5). The five-backend ranking's one-sided permutation p-value is $5/120 \approx 0.04$ (Section S16). Hidden Matching pooled Wilson intervals are complemented by instance-level t-based intervals because shots within an instance share a circuit (Section S22). The margin bootstrap of the frontier runs is stated in Section S25.

24.10 Automated verification

Several automated checks are maintained alongside the analysis code. These include

- consistency checks between result files and manuscript figures,
- verification of frontier statistics,
- validation of submission manifests, and
- audit checks on compiled coupling-map usage.

All reported values are generated directly from archived result files.

24.11 Machine-readable record

The complete chronology of registrations, submissions, retrievals, and outcomes is distributed as machine-readable metadata alongside the code and data release. For each analysis rule in Table 26, the record includes

1. the registration commit and its time,
2. the first job creation,
3. the first retrieval, and
4. the reported outcome.

24.12 Interpretation

The chronology presented here is intended as a transparency and reproducibility record. The scientific conclusions of the manuscript derive from the measured counts, stated decision rules, and reported outcomes. The registration history and execution chronology provide an independent audit trail documenting when predictions were recorded relative to hardware execution.

Every hardware claim in the main text is governed by a registration file in the repository. Table 26 lists, for each analysis rule and hypothesis, its registration commit, the service's job timestamps for the recorded job ids, and the outcome. The degree-2 ladder chronology above states the one case, the degree-2 ladders, in which the first commit of a registration postdates a run: version control certifies registration before execution for the processor B ladder, and the processor A ladder, the same-design replicate, was registered after its data were retrieved.

Table 26: Registration chronology of every analysis rule and hypothesis behind a hardware claim. Commit hashes and times are the repository's committer timestamps (local clock, converted to UTC). Job times are the IBM Quantum service's own created and finished timestamps for the recorded job ids. All times are 2026. The one registration whose first commit postdates a run is stated as such (row 1).

Rule or hypothesis	Registration (commit, UTC)	First job created (UTC)	First retrieval or result (UTC)	Outcome
Full-string majority (mode) as a recovery statistic	Defined in the manuscript at `57bebc1`, 2026-07-01 02:43 UTC; registered as the secondary endpoint of the degree-2 ladders (two protocol files, internally dated 2026-07-07 (processor B) and 2026-07-08 (processor A); submission manifests stamped 2026-07-07 18:14 and 2026-07-08 13:21 UTC name the processor B file; first commit of either file `aaee22d`, 2026-07-09 20:17 UTC, after both ladders were submitted, after the processor A ladder ran and was retrieved, and before the processor B jobs executed; no commit-verified or service-timestamped registration record predates the processor A run)	processor B 2026-07-07 18:14; processor A 2026-07-08 13:21	processor A 2026-07-08 13:51; processor B 2026-07-12 16:54 to 2026-07-13 02:31	51/132 recoveries (main-text Table 3); registered signature held: F reached the shot floor before the mode failed

Table 26 (continued)

Rule or hypothesis	Registration (commit, UTC)	First job created (UTC)	First retrieval or result (UTC)	Outcome
Per-bit majority (primary endpoint of the same registration)	As above	As above	As above	44/132; fails before the mode at $n = 24$; registered prediction (clean per-bit 5/5 through $n \leq 40$) not met, the frontier arriving one to two rungs early under the registered 2Q-only caveat
Count ratio μ_c/μ_w	Registered as the ratio R of the α-sweep (`9372ad4`, 2026-06-02 18:16 UTC); on the degree-2 ladders a reported statistic, not a decision rule (first tabulated at `ecd82cd`, 2026-06-04, for the retired single-oracle campaign)	α-sweep 2026-06-02 20:38	2026-06-02 21:17 (last job finished); analysis committed 22:34	α-sweep: $R = 1.7$–2.2, flat in α; degree-2 frontier count ratio 1–6 (main-text Table 8)
First discrimination test, $\rho \leq -0.5$ (four backends)	`8ffdfef`, 2026-05-04 03:23 UTC: analysis registration on the 2026-05-01 dataset, filed after the data were retrieved	2026-05-01	2026-05-01	$\rho = -0.40$, failed; reported in full (Section S16)
Five-Heron discrimination test, $\rho \leq -0.6$ (S16 threshold)	`8e03542`, 2026-06-06 14:59:00 UTC	2026-06-06 14:59:34 (34 s after the commit)	first job finished 14:59:45; Heron jobs complete 15:54; processor E excluded by the registered inclusion rule (not executed)	$\rho = -0.90$, passed; one-sided permutation $p = 5/120$
α^2 (min-flatness) hypothesis	`9372ad4`, 2026-06-02 18:16:19 UTC	2026-06-02 20:38:48	2026-06-02 21:17:38 (finished); result committed `5dbc5b0` 22:34:24	Refuted: recovery 9–10/10 at every $\alpha \in [0.38, 1]$; log-log slope 0.16 against the registered 2 ± 0.3
$\bar{\alpha}^2$ (mean-flatness) law	Not pre-registered: derived after the sweep result and Lean-verified in `5dbc5b0`, 2026-06-02 22:34:24 UTC (4 h 18 min after the registration commit, 1 h 17 min after the last sweep job finished)	none (no new data; the powered sweep of the next row tests it)	–	Exact identity of the noiseless algorithm; the sweep's oracles span $\bar{\alpha}^2 \in [0.923, 1.000]$, so the sweep is consistent with it but has no power against flatness-independent recovery; every other cell of the record is bent ($\bar{\alpha}^2 = 1$); tested with power by the powered sweep (next row)
$\bar{\alpha}^2$ powered sweep (H2, H3; five block-product families, $\bar{\alpha}^2 \in [0.20, 1.00]$, one compiled depth 167)	`19d4b3b`, 2026-09-15 19:39:26 UTC, before submission; job ids `0f61239` 19:40:15, before retrieval; processor A addendum `f3cddd8` 21:59:02, before its submission; processor A job ids `b69fc37` 21:59:42, before its first result; processor C addendum `68ee944` 19:51:04, before its submission; processor C job ids `3dd34ce` 19:51:25, before its first result	processor B 19:39:51; processor C 19:51:20; processor A 21:59:23	processor B finished 19:41:40 (control) and 19:42:44 (main), scored 19:43:16; processor A finished 21:59:56 and 22:01:37, scored 22:02:28; processor C finished 20:44:05 (main) and 23:32:23 (control), scored 23:41:37	processor B CONFIRMED (slope 0.94, bootstrap 0.93 to 0.96; low-flatness cells 6 to 13% above $r\,\bar{\alpha}^2$, an open observation); processor A REPLICATED under the same bands on its own retention (slope 0.88, 0.85 to 0.91; cells 12 to 31% above); processor C (Nighthawk R2) REPLICATED (slope 0.87, 0.83 to 0.91; cells 14 to 30% above); the gate-matched control retained 0.91 of the bent cell on processor B, 1.14 on processor A and 1.01 on processor C; $n = 8$ and one depth only, no frontier claim (Section S23)

R1 rerun rules A, B, C (Nighthawk frontier on a vetted map)	`29338fd`, 2026-09-11 13:17:53 UTC	13:18:17 to 13:19:20; job ids committed `6c414cf` 13:19:35	finished 13:19:30 to 13:27:27; scored `fbf5266` 13:37:29	A met (0/18), B met (wrong-bit gradient), C not met (0.589 against the band edge 0.5889); Section S21

25 S25: Frontier Records and Diagnostics

This section holds the per-run records, the routing fractions, the diagnostics and the fit outcomes that the main text summarizes. The routing rules are in Section S8, and the margin bootstrap and the mixture specification are stated below.

25.1 Routing audit of the degree-2 record

Routing audit (Section S8): every cell $n \le 24$ on both devices and processor A $n = 32$, 48 ran on clean couplers. Processor B $n \ge 32$ routed 2.5–4.8% of its two-qubit gates over couplers above 10^{-2} (maximum 9.3×10^{-2}, 20–69% of the error budget), processor A $n = 56$, 72, 80 likewise (23–50%), and the processor A exploratory cells $n \ge 100$ ran over couplers reported dead (void). The frontier is fixed by the clean cells (processor A $n = 32$ and 48, 0/5 each). The $n = 32$ processor B recovery is a positive that extra noise could only have lowered.

25.2 Margin score of every frontier run

Table 27 lists every frontier run, with the margin interval of main-text Table 4. The interval is a 95% percentile bootstrap from 2,000 multinomial resamples of the run's empirical distribution. The best rival is re-selected in each replicate, which biases the replicate margins downward, and a string unseen in the record cannot appear in a replicate.

25.3 Per-bit and best-rival diagnostics

The best rival's Hamming distance from the answer is itself a diagnostic (Table 27). A rival at distance 1 names one output bit. At $n = 20$ the best rival sits at distance 1 in four of five processor A runs and two of five processor B runs, and in the four processor A runs and one processor B run it names the same bit, output bit 11 (indexed from 0), whose marginal stays above one half in each. At $n = 24$ every processor B rival sits at distance 5 to 7, so no single output bit carries the wrong mass, while processor A runs 1 and 2 have distance-1 rivals (bits 19 and 15) and runs 3 to 5 distance 7. In run 2, fragile under resampling, the named bit's marginal sits at or below one half, and the per-bit vote fails there and in runs 1, 4 and 5. At $n = 32$ the best rivals sit at distance 5 to 20.

At $n = 20$ every per-bit fidelity exceeds one half (minimum 0.56 on processor B and 0.55 on processor A, medians 0.65 and 0.62) and the per-bit vote recovers all ten runs. At $n = 24$ the minimum falls to 0.496 and 0.488 (medians 0.524 and 0.513), 3 of 120 and 12 of 120 bit marginals sit at or below one half, and the per-bit vote returns the planted shift in 3 of 5 and 1 of 5 runs, failing in all three processor A runs that are fragile under resampling. main-text Proposition 8 certifies one of those four returns (processor B run 4, bound 0.83). In the other three the weakest bit sits within one binomial standard deviation of one half.

25.4 Two-component mixture model of the frontier runs

The fit reduces each run's shot record to the histogram of Hamming distances of the measured strings from the planted answer, for every run of both degree-2 ladders at $n = 8$ to 32 (60 runs of 16,384 shots, 30 of them at the frontier, $n = 20$, 24 and 32). The mixture has two components. With weight $1 - \beta$ a shot is near-clean and flips each bit independently at a rate r_c, and with weight β it is scrambled and flips each bit at a rate z. It is fitted by maximum likelihood against independent per-bit flips at one rate as the baseline. The fit script states the kill criterion as fixed before the first fit, and the script, its outputs and its report entered version control in one commit (`3261a8b`, 23 September 2026), so version control does not certify that order. The mixture survives only if twice its log-likelihood gain over the baseline reaches 33.2, a Bayesian-information margin of $2 \ln 1000$ plus the penalty $2 \ln S$ of its two extra parameters, on at least 20 of the 30 frontier runs and on at least 3 of the 5 runs of every frontier cell, and only if every frontier run's β lies within a factor of 2 of the median over its device's 15 frontier runs. Intervals come from a parametric bootstrap of 200 replicates (seed 20260923), and a Pearson chi-square test on bins pooled to an expected count of at least 5 measures the goodness of fit.

The departure has been tested against one mechanism and the test is negative. We fitted this mixture to every run of both ladders at $n = 8$ to 32 (60 runs, 30 of them at the frontier). The mixture beats independent flips by the likelihood ratio on 28 of the 30 frontier runs and fails the record's own consistency check: it predicts the answer's count at $n = 24$ a factor 7.5 to 25 too low (median 14.7 over the ten runs), a Pearson test rejects it on every run at $n \leq 24$, and its near-clean component flips each bit at a rate of 0.20 to 0.26 at $n = 24$, so almost no shot in it is clean. Its scrambled component sits at a per-bit flip rate of 0.49 to 0.50, uniform, with a weight that follows the two-qubit gate count (Spearman 0.95 and 0.87 over 30 runs per device), which is the white-noise fidelity curve that main-text Fig. 3 already reports as F. The near-answer mass is the part of the record that this curve does not describe, and the fitted mixture does not describe it either. The job manifests hold no qubit layout and the shots carry no time tags, so the record cannot attribute that mass to physical qubits or to drift. The benchmark adds to such a test a label on every error shot: its Hamming distance from a known answer, obtained without simulating the device. Mirror-circuit benchmarking [26] shares this property for its own target strings. The fit code and its report are released with the paper (Code availability).

26 S26: Protocol Card

Table 28 lists the protocol steps from the inputs to the reported quantities and the frontier.

27 S27: Additional tables and figures

References

[1] Bar-Yossef, Z., Jayram, T. S. & Kerenidis, I. Exponential separation of quantum and classical one-way communication complexity. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC)*, 128–137 (2004).

[2] Cooper, C. On the rank of random matrices. *Random Structures & Algorithms* **16**, 209–232 (2000).

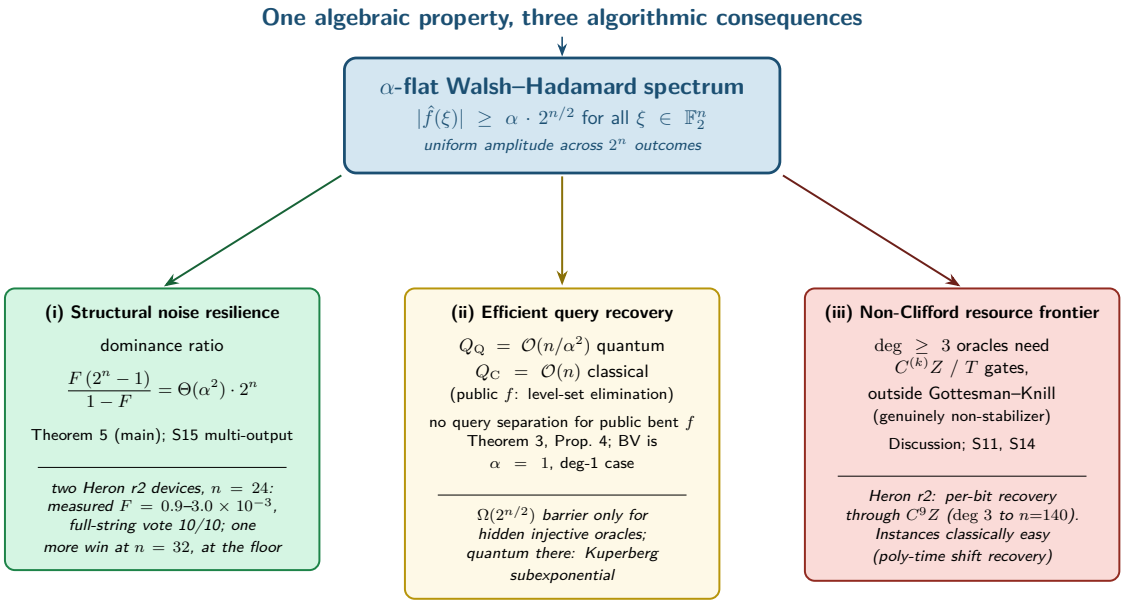


Figure 7: Three algorithmic behaviors governed by α-flatness. (i) Structural noise resilience via majority vote on a uniform 2^n-element outcome distribution: the dominance ratio $F(2^n-1)/(1-F) = \Theta(\alpha^2) \cdot 2^n$ determines recovery even after F itself has collapsed. (ii) An $\mathcal{O}(n/\alpha^2)$-query quantum recovery. For public bent f this is matched classically up to constants, and the $\Omega(2^{n/2})$ barrier holds only in the hidden-oracle regime. (iii) A non-Clifford resource frontier: degree-≥ 3 oracles require $C^{(k)}Z$ / T gates, demonstrated with clean per-bit recovery through $C^{(9)}Z$ at $n = 22$ and through 10-CCZ degree-3 oracles at $n = 140$ (interleaved layouts, entanglement confined to the monomial support). These instances remain classically easy (Supplementary S14), so item (iii) is a resource and compilation statement.

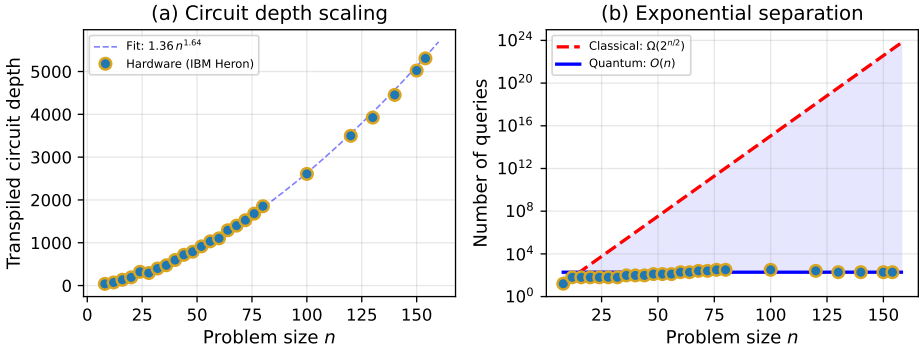


Figure 8: (a) Transpiled circuit depth vs. n for the single-oracle CZ network across six IBM Heron processors (the retired control campaign, used only as a routing-overhead calibration). A power-law fit gives depth $\approx 1.36\,n^{1.64}$. The dual-oracle benchmark circuits of Table 3 carry roughly twice this network per query. (b) Quantum query count vs. the $\Omega(2^{n/2})$ hidden-oracle collision barrier. For the public bent instances demonstrated here the classical query cost is instead $\mathcal{O}(n)$ (Proposition 6), so the gap shown applies only under oracle hiding.

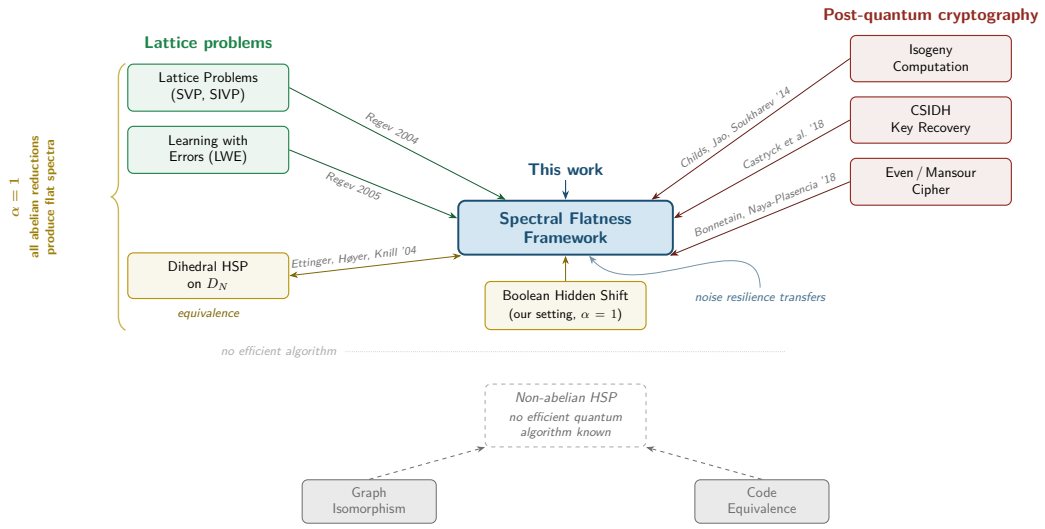


Figure 9: Problems that reduce to the hidden-shift framework. Arrows denote efficient quantum reductions. The double arrow marks an equivalence. Lattice problems, post-quantum cryptographic primitives, and the dihedral HSP all reduce to hidden-shift instances with perfectly flat spectra ($\alpha = 1$), the condition behind the noise resilience proved here, so the suppression mechanism plausibly extends to these reduction targets. The quantitative theorem proved here is for the Boolean spectrally flat setting. The multi-output abelian-group version (Supplementary S15) is a separate extension. Below the dashed line: non-abelian HSP instances with no known efficient quantum algorithm.

[3] Amy, M. & Stinchcombe, L. S. Polynomial-time classical simulation of hidden shift circuits via confluent rewriting of symbolic sums. *Quantum* (2025). ArXiv:2408.02778.

[4] Aaronson, S. & Shi, Y. Quantum lower bounds for the collision and the element distinctness problems. *Journal of the ACM* **51**, 595–605 (2004).

[5] Simon, D. R. On the power of quantum computation. In *SIAM Journal on Computing*, vol. 26, 1474–1483 (1997).

[6] Kuperberg, G. A subexponential-time quantum algorithm for the dihedral hidden subgroup problem. *SIAM Journal on Computing* **35**, 170–188 (2005).

[7] Gavinsky, D., Roetteler, M. & Roland, J. Quantum algorithm for the Boolean hidden shift problem. *Lecture Notes in Computer Science* **6842**, 158–167 (2011).

[8] Ettinger, M., Høyer, P. & Knill, E. The quantum query complexity of the hidden subgroup problem is polynomial. *Information Processing Letters* **91**, 43–48 (2004).

[9] Friedl, K., Ivanyos, G., Magniez, F., Santha, M. & Sen, P. Hidden translation and translating coset in quantum computing. *SIAM Journal on Computing* **43**, 1–24 (2014).

[10] Regev, O. Quantum computation and lattice problems. In *SIAM Journal on Computing*, vol. 33, 738–760 (2004).

[11] Regev, O. On lattices, learning with errors, random linear codes, and cryptography. *Journal of the ACM* **56**, 1–40 (2009).

[12] Childs, A. M., Jao, D. & Soukharev, V. Constructing elliptic curve isogenies in quantum subexponential time. *Journal of Mathematical Cryptology* **8**, 1–29 (2014).

[13] Castryck, W., Lange, T., Martindale, C., Panny, L. & Renes, J. CSIDH: An efficient post-quantum commutative group action. In *Advances in Cryptology – ASIACRYPT 2018*, vol. 11274 of *Lecture Notes in Computer Science*, 395–427 (Springer, 2018).

[14] Bonnetain, X. & Schrottenloher, A. Quantum security analysis of CSIDH. In *Advances in Cryptology – EUROCRYPT 2020*, vol. 12106 of *Lecture Notes in Computer Science*, 493–522 (Springer, 2020).

[15] van Dam, W., Hallgren, S. & Ip, L. Quantum algorithms for some hidden shift problems. *SIAM Journal on Computing* **36**, 763–778 (2006).

[16] Bonnetain, X. & Naya-Plasencia, M. Hidden shift quantum cryptanalysis and implications. In *Advances in Cryptology – ASIACRYPT 2018*, vol. 11272 of *Lecture Notes in Computer Science*, 560–592 (Springer, 2018).

[17] Bernstein, D. J., Lange, T., Martindale, C. & Panny, L. Quantum circuits for the CSIDH: Optimizing quantum evaluation of isogenies. In *Advances in Cryptology — EUROCRYPT 2019*, 409–441 (Springer, 2019).

[18] Bravyi, S. & Gosset, D. Improved classical simulation of quantum circuits dominated by Clifford gates. *Physical Review Letters* **116**, 250501 (2016).

[19] Ding, J. & Schmidt, D. Multivariate public key cryptosystems. In *Advances in Information Security*, vol. 25 (Springer, 2006).

[20] Bettale, L., Faugère, J.-C. & Perret, L. Hybrid approach for solving multivariate systems over finite fields. *Journal of Mathematical Cryptology* **7**, 177–239 (2013).

[21] Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58**, 13–30 (1963).

[22] Even, S. & Mansour, Y. A construction of a cipher from a single pseudorandom permutation. *Journal of Cryptology* **10**, 151–161 (1997).

[23] Kuwakado, H. & Morii, M. Security on the quantum-type Even–Mansour cipher. In *International Symposium on Information Theory and its Applications (ISITA)*, 312–316 (IEEE, 2012).

[24] Esary, J. D., Proschan, F. & Walkup, D. W. Association of random variables, with applications. *Annals of Mathematical Statistics* **38**, 1466–1474 (1967).

[25] Kumar, N., Kerenidis, I. & Diamanti, E. Experimental demonstration of quantum advantage for one-way communication complexity surpassing best-known classical protocol. *Nature Communications* **10**, 4152 (2019).

[26] Proctor, T., Rudinger, K., Young, K., Nielsen, E. & Blume-Kohout, R. Measuring the capabilities of quantum computers. *Nature Physics* **18**, 75–79 (2022). ArXiv:2008.11294.

Table 27: Margin score and rival diagnostics for every frontier run of the degree-2 record ($S = 16{,}384$ shots per run, both processors, $n = 20$, 24 and 32). $\hat{F}$: the answer's share of the shots. $\hat{p}$: the best rival's share. d: the best rival's Hamming distance from the answer. $\hat{\Delta} = \hat{F} - \hat{p}$ with its 95% bootstrap interval (above). At $n = 32$ the record is almost all singletons, so the resampled best rival carries six to eight counts against the one or two in the record. Every $n = 32$ interval therefore lies below its point estimate and measures the resampling spread of the best rival's count alone. μ_c/μ_w: the count ratio, below one where a wrong string out-counted the answer. Stable: the planted shift stays the unique mode under resampling with probability at least 0.95 (main-text Table 8). Certified: the bound of main-text Theorem 2 holds at the observed S with failure probability 0.05, ten of ten runs at $n = 20$ and none at $n = 24$ or 32 (the demanded budgets under both bounds are in Section S20). At $n = 32$ the answer was the unique mode in one run, tied in one and out-counted in eight. The job manifests record transpiled depth and two-qubit count per circuit and no qubit layout, so no output bit is mapped to a physical qubit. Output bits are indexed from 0 by their position in the planted-shift list, the reversed position in the counts key. The record holds no degree-2 counts at these sizes on processor C.

Device	n	run	$\hat{F}$	$\hat{p}$	d	$\hat{\Delta}$	95% interval	μ_c/μ_w	stable	certified
B	20	1	0.111	0.0112	1	0.0995	[0.0945, 0.104]	9.86	yes	yes
B	20	2	0.0864	0.014	4	0.0724	[0.0677, 0.0771]	6.18	yes	yes
B	20	3	0.0457	0.00885	1	0.0369	[0.0331, 0.0399]	5.17	yes	yes
B	20	4	0.075	0.00983	5	0.0651	[0.0608, 0.0692]	7.63	yes	yes
B	20	5	0.0454	0.00891	3	0.0365	[0.0325, 0.0394]	5.1	yes	yes
A	20	1	0.0344	0.00458	1	0.0298	[0.027, 0.0328]	7.52	yes	yes
A	20	2	0.0571	0.012	1	0.0451	[0.0411, 0.0492]	4.77	yes	yes
A	20	3	0.062	0.00964	1	0.0524	[0.0483, 0.0565]	6.43	yes	yes
A	20	4	0.0438	0.00507	1	0.0387	[0.0355, 0.042]	8.64	yes	yes
A	20	5	0.0468	0.00415	4	0.0427	[0.0392, 0.0458]	11.3	yes	yes
B	24	1	0.00201	8.5×10^{-4}	5	0.00116	$[3.1 \times 10^{-4}, 0.00195]$	2.36	yes	no
B	24	2	0.00232	9.2×10^{-4}	7	0.0014	$[5.5 \times 10^{-4}, 0.0022]$	2.53	yes	no
B	24	3	0.00153	7.9×10^{-4}	5	7.3×10^{-4}	[0, 0.0014]	1.92	yes	no
B	24	4	0.00653	0.0011	7	0.00543	[0.00409, 0.00659]	5.94	yes	no
B	24	5	0.00244	5.5×10^{-4}	7	0.00189	$[9.8 \times 10^{-4}, 0.00262]$	4.44	yes	no
A	24	1	7.9×10^{-4}	1.8×10^{-4}	1	6.1×10^{-4}	$[-6.1 \times 10^{-5}, 7.9 \times 10^{-4}]$	4.33	yes	no
A	24	2	3.7×10^{-4}	2.4×10^{-4}	1	1.2×10^{-4}	$[-3.7 \times 10^{-4}, 2.5 \times 10^{-4}]$	1.5	no (0.41)	no
A	24	3	0.00275	9.2×10^{-4}	7	0.00183	$[9.2 \times 10^{-4}, 0.00275]$	3	yes	no
A	24	4	1.8×10^{-4}	1.8×10^{-4}	7	0	$[-4.3 \times 10^{-4}, 0]$	1	no (0.14)	no
A	24	5	4.9×10^{-4}	1.8×10^{-4}	7	3.1×10^{-4}	$[-2.4 \times 10^{-4}, 4.3 \times 10^{-4}]$	2.67	no (0.78)	no
B	32	1	0	1.2×10^{-4}	11	-1.2×10^{-4}	$[-4.9 \times 10^{-4}, -3.7 \times 10^{-4}]$	0	no (0.00)	no
B	32	2	0	1.2×10^{-4}	5	-1.2×10^{-4}	$[-4.9 \times 10^{-4}, -3.7 \times 10^{-4}]$	0	no (0.00)	no
B	32	3	0	6.1×10^{-5}	20	-6.1×10^{-5}	$[-4.9 \times 10^{-4}, -3.7 \times 10^{-4}]$	0	no (0.00)	no
B	32	4	0	6.1×10^{-5}	20	-6.1×10^{-5}	$[-4.9 \times 10^{-4}, -3.7 \times 10^{-4}]$	0	no (0.00)	no
B	32	5	1.8×10^{-4}	6.1×10^{-5}	15	1.2×10^{-4}	$[-4.3 \times 10^{-4}, 0]$	3	no (0.79)	no
A	32	1	0	6.1×10^{-5}	17	-6.1×10^{-5}	$[-4.9 \times 10^{-4}, -3.7 \times 10^{-4}]$	0	no (0.00)	no
A	32	2	0	6.1×10^{-5}	15	-6.1×10^{-5}	$[-4.9 \times 10^{-4}, -3.7 \times 10^{-4}]$	0	no (0.00)	no
A	32	3	6.1×10^{-5}	6.1×10^{-5}	17	0	$[-4.9 \times 10^{-4}, -1.8 \times 10^{-4}]$	1	no (0.26)	no
A	32	4	0	6.1×10^{-5}	17	-6.1×10^{-5}	$[-4.9 \times 10^{-4}, -3.7 \times 10^{-4}]$	0	no (0.00)	no
A	32	5	0	6.1×10^{-5}	19	-6.1×10^{-5}	$[-4.9 \times 10^{-4}, -3.7 \times 10^{-4}]$	0	no (0.00)	no

Table 28: Protocol card. One run of the benchmark from its inputs to its reported quantities, with the settings of the degree-2 record in parentheses. Every step uses only the planted shift, the compiled circuit and the measured counts, so no step simulates the device. The registered analysis of the degree-2 ladders takes the per-bit vote as the primary statistic and the full-string vote as secondary, and it defines its graded frontier on the per-bit vote. The count ratio, the Hamming label and the full-string frontier in this card lie outside the registered decision rules (Section S24).

Step	Specification
Inputs	Register size n, even (8 to 154). Oracle degree d ($d = 2$ is Clifford, and $d \geq 3$ adds $C^{(d-1)}Z$ gates). Queries Q: one query to O_g and one to O_{f^*} per shot. Shots S per circuit (16,384). Seeds: one for the bent form $f(x, y) = (-1)^{x^T By}$ through the invertible matrix B (42), and one shift seed per run index, the same at every n and on both processors ($7000 + k$ for run k, five runs per cell through $n = 80$ and two per exploratory cell at $n = 100$ to 154).
Circuit	Draw s uniformly from $\mathbb{F}_2^n$. Compile $H^{\otimes n} O_g H^{\otimes n} O_{f^*} H^{\otimes n}$ with $g(x) = f(x \oplus s)$ and the exact dual f^* (Qiskit optimization level 3, XY4 dynamical decoupling), and measure all n qubits S times. Noiselessly the output is s with probability $\bar{\alpha}^2$, which equals 1 for bent f.
Statistics	Full-string vote: $\hat{s}$ is the most frequent measured string, and a tie goes to the tied string listed first in the returned counts, a rule of the analysis code that the planted s cannot affect. Per-bit vote: bit j of $\hat{s}$ is 1 when at least half of the S shots read 1 at bit j.
Scoring	Both estimates are fixed before the planted s is read. A run recovers under a statistic when its $\hat{s}$ equals s as an n-bit string. A cell reports recoveries out of runs with a Wilson 95% interval.
Outputs	Recovery profile: recoveries per cell against n and the compiled two-qubit gate count, for both statistics. Single-shot fidelity F: the count of s divided by S. Count ratio μ_c/μ_w: the count of s over the count of the most frequent other string. Hamming label: the Hamming distance of the most frequent other string from s, and the fraction of wrong shots at distance 1, 2 and ≥ 3.
Frontier	The largest n at which the full-string vote recovers every run of its cell ($n = 24$ on both processors, with the one processor A tie scored as a recovery, and $n = 20$ on processor A with it scored as a failure). Cells above it with 1 to 4 recoveries of 5 form the graded band ($n = 32$ on processor B, and $n = 24$ on processor A when its tie is scored as a failure), and every cell above the band is reported with its recoveries, none in the record.

Table 29: Count-ratio diagnostics for the second degree-2 ladder. Runs on processor A, $S = 16\,384$ shots, five replicates per row. $\bar{F}$ is the mean measured single-shot fidelity; μ_c and μ_w are the measured counts of the planted shift and strongest rival outcome, respectively. Full-string recovery persists while $\mu_c > \mu_w$. Supplementary S20 and S24 provide resampling analyses, count-ratio diagnostics, and registration details.

n	2Q	$\bar{F}$	μ_c	μ_w	Mode	Per-bit
8	54	0.69	10995–11809	265–608	5/5	5/5
12	111	0.42	6473–7379	320–537	5/5	5/5
20	435	0.049	564–1016	68–196	5/5	5/5
24	971	9.2×10^{-4}	3–45	3–15	5/5*	1/5
32	1592	1.2×10^{-5}	0–1	1	0/5	0/5
40–80	2712–11127	0	0	1	0/30	0/30
100–154	17861–45026	0	0	1	0/6	0/6

Table 30: Query complexity of hidden-shift recovery under different oracle-access regimes. All experiments in this work occupy the public-function regime, where the shift is classically recoverable in polynomial time. The hidden-injective-oracle regime contains the classically hard hidden-shift formulations, including Simon's problem and the dihedral hidden-subgroup problem. Further discussion is provided in Supplementary S2, S13, and S14.

Regime	**Quantum**	**Classical**	**Status**
Public f, any degree	$\mathcal{O}(n/\alpha^2)$	$\mathcal{O}(n)$ (level-set)	no separation
Compiled circuit, bounded degree	$\mathcal{O}(n/\alpha^2)$	poly(n) [3]	no separation
Compiled circuit, unbounded degree	$\mathcal{O}(n/\alpha^2)$	open, all tested candidates easy (Supp. S14)	**open**
Hidden injective oracle on $\mathbb{Z}_2^n$	$\mathcal{O}(n)$ (Simon) [5]	$\Omega(2^{n/2})$	exponential, black box
Hidden injective oracle on $\mathbb{Z}_N$ (cyclic)	$2^{\mathcal{O}(\sqrt{\log N})}$ [6]	$\Omega(\sqrt{N})$	subexponential