Academia Quantum

10.20935/AcadQuant8648

Research Article

Gauge-dependent post-selection in quantum spectral estimation on trapped-ion hardware

David Vesterlund[1,2,*]

[1] Vesterlund Ventures Holding AB, Jönköping, Sweden

[2] Ramanujan Machine Team, Technion—Israel Institute of Technology, Haifa, Israel

* Correspondence: david@vesterlundventures.se

Received: 19 September 2026 | **Accepted:** 10 November 2026 | **Published:** 25 November 2026

Abstract

Introduction: A constant change of basis (gauge) applied to every step matrix of a Conservative Matrix Field (CMF) product is a similarity transformation of the product: it leaves the eigenvalues, trace and determinant unchanged but alters the block-encoding subnormalization and hence the post-selection success probability. Ideal, noiseless simulation predicts that a finite-horizon optimized gauge U_N outperforms the identity gauge I. We tested this prediction on a 20-qubit trapped-ion processor.

Materials and methods: Frozen CMF circuits for N = 2, 4, 6, 8, 10, 12 (two gauges by two initial states) were executed on the PIAST-Q trapped-ion quantum processing unit (QPU) in randomized calibration blocks. The single pre-specified primary hypothesis compared the calibrated-state U_{12} and I circuits at $N = 12$ over 30 blocks (planned pass: 1171 of 1200 scheduled 200-shot jobs returned; 234,200 shots across four circuit variants, of which 116,200 enter the primary comparison). Inference uses the block as the unit: paired block differences, cluster-robust job-level tests, a block bootstrap and a random-intercept/random-slope logistic model. An N-sweep ($N = 2$–10; 15 blocks, 562 of 600 scheduled jobs) provides secondary, Holm-adjusted comparisons. Unplanned weekly re-executions (2171 further $N = 12$ jobs) serve as a replication set.

Results: At $N = 12$, the identity gauge succeeded in 24,838 of 58,400 shots ($p = 0.4253$) and the optimized gauge succeeded in 30,142 of 57,800 shots ($p = 0.5215$). The block-mean difference is +9.55 percentage points (pp; 95% CI 8.04 to 11.07; paired $t(29) = 12.9$, $p < 1 \times 10^{-12}$; 30 of 30 blocks positive), consistent with the simulated +9.61 pp. The shot-level z-test ($z = 32.8$) is anti-conservative because the job-level variance exceeds the binomial expectation by a factor of 4–5. The advantage was reproduced in all five weekly re-executions (block means +6.8 to +10.5 pp). In the N-sweep the point estimate is positive at every N; after block-level inference and Holm adjustment it is significant at $N = 2$ and $N = 10$ but not at $N = 4$, 6 or 8. The primary comparison evaluates each encoding at its own optimal input state; a frame-related control with the same input vector in both representations reproduces the predicted change of sign of the advantage across N (sign agreement at five of six sizes, r = 0.96).

Conclusions: Within the tested CMF family and hardware platform, a representation choice that is invisible to the spectral invariants produces a reproducible, simulation-predicted gain in post-selection probability at $N = 12$, while the ideal simulation does not predict the absolute success probabilities, which are reduced by a roughly gauge-independent absolute loss of about 0.38. Claims of generality across N and platforms, and of an intrinsic noise-robustness of the optimized gauge, remain hypotheses.

Keywords: quantum computing; gauge optimization; post-selection; Conservative Matrix Fields; trapped-ion quantum processor; spectral estimation; block encoding

1. Introduction

Conservative Matrix Fields (CMFs) [1, 2], which emerged from the Ramanujan Machine programme of algorithm-assisted discovery of formulas for mathematical constants [3], provide a unified framework for Diophantine approximation formulas by encoding them as orbits in a lattice of commuting matrix shifts. The arithmetic quality of a formula—governed by Lyapunov exponents and irrationality measures δ—is determined by spectral data of the orbit products $P_N = M(1)\cdots M(N)$. A natural quantum question arises: at what cost can this spectral data be estimated on a quantum computer?

The quantum algorithm studied here block-encodes [4–7] each step matrix $M(n)$ into a two-qubit unitary acting on one system qubit and one fresh ancilla, chains N such encodings, and post-selects on the ancilla register. The success probability p_{succ} is set by the subnormalization of the chained block encoding, i.e., by the ratio between $\|P_N v\|$ and the product of the per-step normalization constants $\alpha_n \geq \sigma_1(M(n))$. Subnormalization is the central cost parameter of block-encoding-based quantum linear algebra [5, 7–10], and its dependence on how a matrix is presented to the encoder is well documented in the theory of block encodings and quantum singular value transformation (QSVT) [8, 9]. Post-selection failures must otherwise be paid for by repetition or amplitude amplification [11], or avoided by algorithm designs that sidestep post-selection altogether [12].

Crucially, the subnormalization is a gauge-dependent quantity. Under a coboundary transformation $M(n) \to V(n)^{-1}M(n)V(n+1)$ the finite product becomes $\tilde{P}_N = V(1)^{-1}P_N V(N+1)$. This is a similarity transformation of P_N—preserving its eigenvalues, trace, determinant and characteristic polynomial—when the endpoint gauges coincide, $V(N+1) = V(1)$; for general position-dependent gauges the finite-product invariants are not preserved, and only asymptotic quantities such as Lyapunov exponents survive under suitable boundedness conditions. In this work we use only constant gauges, $V(n) \equiv U$, for which $\tilde{P}_N = U^{-1}P_N U$ is similar to P_N for every N. Singular values are not similarity invariants, so the per-step norms $\sigma_1(\tilde{M}(n))$, the normalization constants and the subnormalization of the chain can change substantially even though the spectrum of the unnormalized matrix product does not. This gauge freedom is intrinsic to the CMF structure [1] and offers a design parameter that adds no entangling gates: the transformed circuits have the same number of R_{XX} gates and qubits as the original ones, and differ from them only in rotation angles (and, on the present hardware, by at most one effective single-qubit pulse; Section 2.3). Two gauges are compared in this study:

- Identity gauge (I): The raw matrix $M(n)$ is encoded directly. Along the trajectory considered here the subnormalization gap accumulates with N and p_{succ} decreases from 0.98 at $N = 2$ to 0.80 at $N = 12$ (**Table 1**)
- Finite-horizon optimized gauge (U_N): A constant gauge $V(n) \equiv U_N$ obtained by numerically minimizing the cumulative subnormalization gap of the finite chain of length N (Section 2.1)

Ideal simulation predicts $\Delta_{\text{sim}}(N) = p_{\text{succ}}(U_N) - p_{\text{succ}}(I) > 0$ at all N considered, with $\Delta_{\text{sim}}(12) = +9.61$ percentage points (pp). The question addressed experimentally is whether this difference survives on a noisy intermediate-scale [13, 14] trapped-ion processor, where both circuits lose roughly half of their ideal success probability.

Hardware demonstrations of post-selected block-encoding circuits exist—for example, a quantum-signal-processing protocol run end-to-end on a trapped-ion computer [15]—and subnormalization-aware block-encoding constructions have been analysed theoretically [8, 9]. We are not aware, however, of a hardware

experiment that compares two encodings of similar matrix products—identical entangling-gate count and qubit count, identical spectra of the underlying unnormalized matrix products, differing only in the representation of the factors—by their best achievable post-selection probabilities, with each encoding evaluated at its own optimal input state, together with a frame-related control in which the same input vector is used in both representations; to our knowledge the present experiment is the first of this kind, and the claim is restricted to that specific sense. The initial state is therefore not a factor that is held fixed in the primary comparison (Section 2.2.1). The results should be read as three separate statements with different levels of support: (i) the optimized gauge yields a larger measured post-selection probability at $N = 12$; (ii) the measured difference at $N = 12$ is numerically compatible with the ideal prediction; (iii) the optimized gauge may interact differently with hardware noise. Statements (i) and (ii) are supported directly by the data; statement (iii) is a hypothesis that the present data do not establish.

Table 1. Circuit resources by problem size N for the calibrated-state cells.

N	Qubits	R_{XX}	Depth	p_{ideal} (I)	p_{ideal} (U_N)	Pulses (I)	Pulses (U_N)
2	3	4	18	0.9772	1.0000	11 (0)	11 (0)
4	5	8	34	0.9740	0.9877	22 (1)	23 (2)
6	7	12	50	0.9578	0.9660	34 (3)	34 (3)
8	9	16	66	0.9190	0.9423	44 (3)	45 (4)
10	11	20	82	0.8647	0.9193	54 (3)	55 (4)
12	13	24	98	0.8020	0.8981	64 (3)	65 (4)

N: number of encoded steps; Qubits: $N + 1$ qubits; R_{XX}: native two-qubit entangling gate, with $2N$ gates per circuit; Depth: native circuit depth; p_{ideal}: ideal post-selection probability; I: identity gauge; U_N: optimized gauge; AQT: Alpine Quantum Technologies; Pulses: effective single-qubit pulses under the AQT two-pulse rule; parentheses: number of subthreshold R rotations.

2. Materials and methods

2.1. CMF trajectory and gauge optimization

All circuits encode the π-CMF of Raz et al. ([2], Eq. 6) along the trajectory of unit steps in the x direction starting at $(x, y, z) = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$, so that $M(n) = M_x(\frac{1}{2} + n - 1, \frac{1}{2}, \frac{1}{2})$ is a real 2×2 matrix and $P_N = M(1) \cdots M(N)$. The identity gauge encodes $M(n)$ directly. The finite-horizon gauge is a constant real matrix U_N applied as $\tilde{M}(n) = U_N^{-1} M(n) U_N$ for every n, i.e., the coboundary $V(n) \equiv U_N$ with equal endpoints, so that $\tilde{P}_N = U_N^{-1} P_N U_N$ is similar to P_N and the two gauges encode operators with identical eigenvalues, trace and determinant (verified from the artifacts in Section 2.2.1). No position-dependent gauge is used anywhere in this study.

U_N is restricted to the three-parameter family $U(\phi, \psi, s) = R(\phi)\,\mathrm{diag}(e^s, e^{-s})\,R(\psi)$, where $R(\cdot)$ is a planar rotation; the condition number $\kappa(U) = e^{2|s|}$ is bounded by $\kappa_{\max} = 20$ ($|s| \le \frac{1}{2} \ln 20$) and $\phi, \psi \in [-\pi, \pi]$. The objective is the η-regularized subnormalization gap in **Equation (1)**:

$$\gamma_\eta(U) = \frac{1}{N}\Big(\sum_{n=1}^{N} \ln \alpha_n - \ln \sigma_1(\tilde{P}_N)\Big), \qquad \alpha_n = (1 + \eta)\,\sigma_1\big(\tilde{M}(n)\big), \tag{1}$$

with margin $\eta = 10^{-6}$, so that p_{succ} for the calibrated initial state equals $e^{-2N\gamma_\eta}$ exactly (see below). γ_η was minimized with SciPy (exact installed version not recorded in the experiment logs; project requirements specify

SciPy ≥ 1.13), using its `differential_evolution` optimizer: a population of 12 candidates per parameter (36 in total) initialized by Latin-hypercube sampling of the bounded parameter box with seed 123, evolved for at most 60 generations or until the standard deviation of the population objective values fell below 10^{-12} times their mean, followed by limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm with box constraints (L-BFGS-B) polishing of the best candidate; parameter sets yielding singular matrices were penalized with a large objective value. For $N = 12$ the optimizer returned $U_{12} \approx \left(\begin{smallmatrix} 0.8990 & 0.0137 \\ 0.3943 & 1.1184 \end{smallmatrix}\right)$ ($\phi = -0.445$, $\psi = 0.632$, $s = -0.230$, $\kappa = 1.58$), reducing γ_η from 9.20×10^{-3} (identity) to 4.48×10^{-3}. The per-N gauges $U_2, \ldots, U_{10}$ were generated with the same code path and settings. All optimizer outputs are stored as JavaScript Object Notation (JSON) next to the circuit artifacts (data availability statement).

2.2. Block encoding and circuit construction

Each normalized step $A_n = \tilde{M}(n)/\alpha_n$ (spectral norm $\leq 1/(1+\eta) < 1$) is embedded in the exact unitary dilation of **Equation (2)** [16]:

$$W_n = \begin{pmatrix} A_n & \sqrt{\mathbb{1} - A_n A_n^\dagger} \\ \sqrt{\mathbb{1} - A_n^\dagger A_n} & -A_n^\dagger \end{pmatrix}, \tag{2}$$

a 4×4 unitary acting on the system qubit and the n-th ancilla qubit, whose top-left block is A_n when the ancilla is in $|0\rangle$. The circuit prepares the system qubit in a real state ψ, applies $W_N, W_{N-1}, \ldots, W_1$ (each with a fresh ancilla) and measures the N ancillas; the run is successful when all ancillas read 0, in which case the system qubit holds $\tilde{P}_N\psi/\|\tilde{P}_N\psi\|$. The ideal success probability is given by **Equation (3)**:

$$p_{\text{succ}} = \frac{\|\tilde{P}_N\psi\|^2}{\prod_{n=1}^{N} \alpha_n^2}. \tag{3}$$

Two initial states are used: the calibrated state $\psi = v_1(\tilde{P}_N)$, the top right-singular vector, for which $p_{\text{succ}} = (\sigma_1(\tilde{P}_N)/\prod_n \alpha_n)^2 = e^{-2N\gamma_\eta}$, and the asymptotic state $\psi = v_1(\tilde{M}_\infty)$, the dominant eigenvector of the gauge-transformed limit matrix $M_\infty = \lim_{n\to\infty} M(n)/\|M(n)\|$, which is N-independent and serves as a control for state-preparation effects. The primary comparison uses the calibrated state in both gauges; the asymptotic-state cells are reported as a control.

2.2.1. Task equivalence of the two encodings

The two circuits compared at a given N block-encode the operators $P_N/\prod_n \alpha_n$ and $\tilde{P}_N/\prod_n \tilde{\alpha}_n$ with $\tilde{P}_N = U_N^{-1} P_N U_N$. Because the unnormalized products are similar, their eigenvalues, trace and determinant coincide; a downstream spectral-estimation routine can recover the same product-level spectral data from either block encoding after accounting for its known scalar normalization. The normalized block-encoded operators themselves need not have identical eigenvalues, trace or determinant. Two things do differ, and should be kept apart. First, the unnormalized products P_N and $\tilde{P}_N$ are similar but not equal, and their singular values differ ($\sigma_1(\tilde{P}_{12}) = 83{,}530$ versus $\sigma_1(P_{12}) = 81{,}431$). Second, the scalar normalizations $\prod_n \alpha_n$ and $\prod_n \tilde{\alpha}_n$ differ ($90{,}931$ versus $88{,}142$ at $N = 12$), so the block-encoded operators $P_N/\prod_n \alpha_n$ and $\tilde{P}_N/\prod_n \tilde{\alpha}_n$ are similar only up to the scalar factor $\prod_n \tilde{\alpha}_n/\prod_n \alpha_n = 0.969$. Both differences enter the maximal post-selection probability $p_{\max} = (\sigma_1/\prod_n \alpha_n)^2$: writing the ideal gain as $\ln[p_{\max}(U_N)/p_{\max}(I)] = 2\ln[\sigma_1(\tilde{P}_N)/\sigma_1(P_N)] + 2\ln[\prod_n \alpha_n/\prod_n \tilde{\alpha}_n]$, the two terms at $N = 12$ are $+0.051$ (singular-value change of the similar product, 45% of the total 0.113) and $+0.062$ (normalization, 55%).

At $N = 4$, 6 and 8 the normalization term is negative (-0.001, -0.022, -0.014) and the whole ideal gain comes from the change of σ_1; at $N = 2$ the reverse holds (-0.024 versus $+0.047$). The gauge therefore does not act on the normalization alone, and we describe the compared quantity as the best achievable post-selection probability of each encoding rather than as a pure subnormalization effect. The input and output transformations are as follows. For the asymptotic-state cells the inputs are related by the frame change, $\tilde{\psi} \propto U_N^{-1}\psi$, so the post-selected outputs satisfy $\tilde{P}_N\tilde{\psi} \propto U_N^{-1}(P_N\psi)$: the two circuits perform the same linear map on the same vector expressed in two frames, and the output of one is obtained from the other by the fixed, classically known matrix U_N^{-1} followed by renormalization. For the calibrated-state cells each gauge is run with its own optimal input, the top right-singular vector of its own product; these inputs are not frame-related (their overlap after the frame change is 0.91–0.9996 depending on N), and the compared quantity is the maximal post-selection probability $p_{\max} = (\sigma_1/\prod_n \alpha_n)^2 = e^{-2N\gamma_\eta}$ of each encoding, a property of the encoded operator and its normalization alone. The primary comparison is therefore between the best achievable post-selection probabilities of two encodings of similar unnormalized products, and the asymptotic-state control is a comparison of the same computation in two frames.

These statements were verified directly from the frozen artifacts (script `verify_task_equivalence.py`, output `paper/data/task_equivalence.json`). For every N, the trace, determinant and eigenvalues of P_N and of $U_N^{-1}P_NU_N$ reconstructed from the stored U_N and the CMF step matrices agree to a relative precision of 2.4×10^{-12} or better (at $N = 12$: tr $= 81{,}304.23$, det $= 25{,}412.54$, eigenvalues $81{,}303.92$ and 0.3126 in both gauges), whereas $\sigma_1(P_{12}) = 81{,}431$ versus $\sigma_1(\tilde{P}_{12}) = 83{,}530$ and $\prod_n \alpha_n = 90{,}931$ versus $88{,}142$, giving $p_{\max} = 0.8020$ versus 0.8981. Exact statevector simulation of the 24 frozen native circuits that were executed on the hardware reproduces the manifest success probabilities to within 9×10^{-14} and the post-selected system-qubit output states $\tilde{P}_N\tilde{\psi}/\|\tilde{P}_N\tilde{\psi}\|$ with fidelity $1 - 2 \times 10^{-16}$; for the asymptotic-state cells the input and output states of the two gauges are related by U_N^{-1} with the same fidelity.

Each W_n and the state preparation were decomposed with Qiskit 1.4.6 [17] (optimization level 1, seed 7) into the PIAST-Q native gate set $\{R_z(\lambda), R(\theta,\phi), R_{XX}(\theta)\}$; every W_n compiles to exactly two R_{XX} gates, so a circuit for problem size N uses $N + 1$ qubits, $2N$ entangling gates and native depth $8N + 2$. Compiled circuits were frozen as Open Quantum Assembly Language (OpenQASM) 2.0 files with Secure Hash Algorithm 256-bit (SHA-256) hashes, gate-angle lists and pulse-level metadata, and the identical native circuit was submitted in every job (no re-transpilation at run time). The ideal p_{succ} of each frozen circuit was verified on a noiseless simulator (Qiskit Aer, 8192 shots) to within 0.006 before any hardware use. **Table 1** summarizes the circuit resources.

2.3. Hardware platform

PIAST-Q is a 20-qubit trapped-ion quantum processing unit (QPU) [18, 19] built on the Alpine Quantum Technologies (AQT; Innsbruck, Austria) architecture [20, 21] with all-to-all connectivity, accessed through the PCSS-QAPI interface (the quantum application programming interface of the Poznań Supercomputing and Networking Center, PCSS; Qiskit 1.4.6, AQT provider 1.14). The native gate set is $\{R_z, R(\theta,\phi), R_{XX}(\theta)\}$, the standard single-qubit rotation and Mølmer–Sørensen-type entangling gate set of ion traps [22]. Operational constraints were a maximum of 200 shots per execution, a throughput of roughly 250 jobs per hour and a few hours of access per week. Shot-by-shot outcomes (memory) were requested for every job.

A hardware-specific consideration communicated by the hardware team is the two-pulse rule: single-qubit rotations $R(\theta,\phi)$ with $|\theta| < 0.2\pi \approx 36^\circ$ are implemented as two large pulses, $R(\pi,\phi+\pi)\,R(\theta+\pi,\phi)$, which is mathematically identical but physically longer. At $N = 4$, 8, 10 and 12 the U_N circuits contain one more

subthreshold rotation than the identity circuits, and hence one more effective pulse (23 vs. 22, 45 vs. 44, 55 vs. 54 and 65 vs. 64); at $N = 2$ and $N = 6$ the counts are equal (**Table 1**). Where it occurs, the asymmetry makes the physical single-qubit cost of the optimized gauge slightly higher. The direction of this asymmetry works against the hypothesis under test; its quantitative effect on the measured probabilities was not isolated experimentally.

2.4. Experimental design and pre-specification

The campaign followed a written testing plan fixed before hardware execution (data availability statement), which pre-specified (a) the single primary hypothesis H1: $p_{\text{succ}}(U_{12}) > p_{\text{succ}}(I)$ at $N = 12$, $\eta = 10^{-6}$, calibrated state, with effect size $\Delta = p_{U_{12}} - p_I$ estimated from raw counts without readout mitigation; (b) the block structure with interleaved calibration circuits and randomized job order within blocks; (c) the hierarchy shots ⊂ jobs ⊂ blocks and a binomial mixed-effects logistic model with a block random effect as the intended inferential model; and (d) the N-sweep as a secondary analysis. The plan was not deposited in a public registry; it is included verbatim in the repository with its version history.

Phase Q2 (N = 12, primary): Of thirty blocks each block consists of a pre-calibration suite, 40 CMF jobs and a post-calibration suite. The 40 CMF jobs are the four $N = 12$ variants (two gauges by two initial states) repeated 10 times, in an order drawn by a seeded pseudo-random shuffle (seed 20260818) of the 40-job pool; gauges were therefore interleaved within every block but not strictly paired back-to-back. Each job used 200 shots. Blocks 0 and 29 used the full calibration suite (five circuits), the others the mini suite (four circuits). Planned size: 1200 CMF jobs and 244 calibration jobs, i.e., 600 jobs (120,000 shots) per gauge in total and 300 jobs (60,000 shots) per gauge for the primary calibrated-state comparison.

Phase Q3 (N-sweep, secondary): Fifteen blocks of 40 CMF jobs covering $N = 2, 4, 6, 8, 10$ (20 circuits, each twice per block, shuffled as above), i.e., 30 planned jobs (6000 shots) per circuit, with the same calibration structure. The $N = 12$ point of the sweep is taken from Q2.

Calibration suite: Because the error rates of a quantum processor fluctuate in time [23] and limit the reproducibility of circuit outcomes [24], four small reference circuits on three qubits were run before and after every block (SPAM denotes state preparation and measurement): $\mathcal{C}_0$ (ground-state SPAM, ideal outcome $|000\rangle$), $\mathcal{C}_1$ (excited-state SPAM, $R(\pi, 0)$ on each qubit, ideal $|111\rangle$), $\mathcal{C}_8$ (a single R_{XX} at the dominant CMF angle, ideal $P(000) = \cos^2(\theta/2) = 0.945$) and $\mathcal{C}_9$ (the $N = 2$ CMF circuit followed by its inverse, ideal $|000\rangle$). The full suite adds the CMF sentinel: the 13-qubit $N = 12$ identity-gauge calibrated-state circuit itself, scored against a 4096-outcome reference distribution that places the ideal success probability 0.802 on the all-zero ancilla string and spreads the remaining 0.198 uniformly over the other 4095 strings. For every calibration job the total variation distance $d_{TV} = \frac{1}{2}\sum_x |\hat{P}(x) - P_{\text{ref}}(x)|$ between the 200-shot empirical distribution and the reference distribution is recorded.

2.5. Execution history and job accounting

Q2 was executed as a single planned pass: blocks 0–7 on 5–6 September 2026 and blocks 8–29 on 7–8 September 2026, the latter as three concurrently running schedulers (blocks 8–15, 16–22 and 23–29) submitting to the same device. Of the 1200 scheduled CMF jobs, 1171 returned results and 29 submissions failed with client-side network errors or time-outs before a result was obtained (no QPU result was produced for these, and they were not re-attempted within the pass); 241 of 244 calibration jobs returned. The resulting design is

slightly unbalanced (292 identity and 289 U_{12} calibrated-state jobs, 293 and 297 asymptotic-state jobs). No job, shot or block was excluded on the basis of its outcome.

Because the scheduler agents were not deactivated after the planned pass, they re-fired on five subsequent Mondays (14, 21, 28 September and 5, 12 October 2026) and re-executed the schedules for blocks 8–29. Each re-execution wrote its result to the same per-step file, overwriting the earlier record, while failed submissions left separate error records. The complete execution history was therefore reconstructed from the run logs, which echo every job with its block, step, circuit identifier and 200-shot success fraction (four decimals; the integer count k is recovered exactly). The reconstruction was validated against all 1759 result files present on disk (the most recent execution of each slot; 0 mismatches) and reproduced the per-variant counts of the originally submitted manuscript exactly when restricted to the files that existed on 17 September 2026. The submitted manuscript's $N = 12$ figures ($\hat{p} = 0.4216$ vs. 0.5172, $\Delta = +9.56$ pp, "237,600 shots") were computed from that on-disk state, which unknowingly mixed the planned pass with the 14 September re-execution; they are superseded by the analysis below.

Three datasets are distinguished throughout (**Table 2**). The primary dataset is the planned pass (first execution of every slot, 5–8 September). The replication dataset comprises all later successful re-executions of already measured slots (2171 CMF jobs); it was not planned, but every job in it was generated by the same frozen circuits and the same randomized schedule, so it is used as an independent check of temporal stability and as a pooled sensitivity analysis. For Q3 the planned schedule was completed across several weekly windows because the device was unreachable for most of 28 September: the Q3 primary dataset is the first successful execution of every scheduled slot (562 of 600 CMF jobs filled, 117 of 124 calibration jobs; 38 slots never returned a result), and later re-executions (467 jobs) form the Q3 replication set. All dataset definitions are outcome-blind.

Table 2. Job accounting for the Q2 and Q3 hardware campaigns.

Phase	Dataset	CMF ok	CMF failed	Cal ok	Cal failed
Q2 ($N = 12$)	Planned pass, 5–8 September	1171	29	241	3
	Re-executions, 14 September–13 October	2171	2229	425	465
	All executions	3342	2258	666	468
Q3 ($N = 2$–10)	First success per slot, 17 September–13 October	562	–	117	–
	Re-executions of filled slots	467	–	78	–
	All executions	1029	651	195	149

Q2: $N = 12$ primary campaign, 1200 scheduled Conservative Matrix Field (CMF) jobs and 244 calibration jobs; Q3: $N = 2$–10 sweep, 600 scheduled CMF jobs and 124 calibration jobs; Cal: calibration; ok: jobs that returned quantum processing unit (QPU) results; failed: submissions without results because of network, Domain Name System (DNS), Transport Layer Security (TLS), or time-out errors; each returned job has 200 shots. Dashes indicate counts not reported separately.

2.6. Statistical analysis

All analyses use raw counts; no readout or other error mitigation [25] is applied. For each circuit variant and dataset we report the integer success count k (all-ancilla-zero outcomes), the total number of shots n, $\hat{p} = k/n$ and a Wilson 95% interval [26]. Three levels of inference are then distinguished.

Shot level (descriptive): The difference $\hat{\Delta} = \hat{p}_U - \hat{p}_I$ with the Newcombe hybrid-score 95% interval for a difference of two independent proportions [27] and the pooled two-proportion z statistic. These quantities treat the shots

as independent and are reported for comparability only; they are not used for confirmatory inference because shots within a job share the same hardware state.

Job level (cluster-robust): Each 200-shot job contributes one success fraction. The difference of means between the U and I jobs is tested with Welch's t test, which is robust to arbitrary within-job correlation. The design effect DEFF $= s^2_{\text{jobs}}/[\bar{p}(1-\bar{p})/200]$ quantifies the excess of the observed between-job variance over the binomial expectation.

Block level (confirmatory): For each block b the paired difference $\Delta_b = \hat{p}_{U,b} - \hat{p}_{I,b}$ is formed from all jobs of that block (Q2: 10 jobs, 2000 shots per gauge; Q3: 2 jobs, 400 shots per gauge). The block is the unit at which the job order was randomized and at which hardware drift is expected to act, so the Δ_b are treated as exchangeable observations: we report their mean $\bar{\Delta}_{\text{hw}}$, a t-based 95% interval, the paired t test, the Wilcoxon signed-rank and sign tests, and a block bootstrap (resampling blocks with replacement, 20,000 resamples) of the pooled difference. This analysis makes no assumption about the correlation structure within blocks. Its assumptions are that the block differences Δ_b are independent across blocks and drawn from a common distribution (exchangeable blocks), and—for the t interval—that the mean of 30 bounded differences is approximately normal; the Wilcoxon, sign and bootstrap procedures relax the normality assumption. Because the planned $N = 12$ pass was executed in two sessions (Section 2.5), the common-distribution assumption is examined in Section 3.1 by reporting the two sessions separately and by a session-stratified bootstrap. None of the procedures model temporal autocorrelation between consecutive blocks beyond what the between-block variance already absorbs.

Mixed-effects model: As pre-specified, we also fit logit $p_{gb} = \beta_0 + \beta_G\, g + u_b + v_b\, g$, with $g \in \{0,1\}$ indicating the gauge, $u_b \sim \mathcal{N}(0,\sigma_u^2)$ a block random intercept and $v_b \sim \mathcal{N}(0,\sigma_v^2)$ a block-specific random gauge effect, to the per-block binomial counts by maximum likelihood (two-dimensional Gauss–Hermite quadrature, 24 nodes per dimension; standard errors from the numerical Hessian; likelihood-ratio tests for β_G and for the random effects).

Multiplicity and reporting: $N = 12$ is the single pre-specified primary hypothesis. The five N-sweep comparisons are secondary and their block-level p values are Holm-adjusted [28]; shot-level p values are shown for completeness but are not adjusted or interpreted. p values smaller than 10^{-3} are reported as powers-of-ten upper bounds computed from the log survival function; no p value is reported as zero. Hardware retention is $\mathcal{R} = \hat{p}/p_{\text{ideal}}$ and the absolute loss is $\mathcal{L} = p_{\text{ideal}} - \hat{p}$; comparisons of $\mathcal{R}$ or of $\Delta_{\text{hw}} - \Delta_{\text{sim}}$ between gauges are exploratory. All analyses are implemented in three scripts (`reconstruct_jobs.py`, `analysis_hierarchical.py`, `generate_figures.py`) that regenerate every number and figure in this article from the raw logs; a fourth (`verify_task_equivalence.py`) performs the artifact checks of Section 2.2.1 (data availability statement).

2.7. Disclosure of AI tools use

During preparation of this manuscript, GLM-5.2 was used to assist with figure generation and drafting of the original submission. For the revised analysis, Devin (Cognition; web service, specific version/build not recorded; powered by Anthropic's Claude) was used under the author's direction to assist with execution-log reconstruction, statistical-analysis code, figure regeneration and drafting. During final production proofreading, ChatGPT (OpenAI GPT-6) was used to check presentation details and adjust figures without changing the underlying experimental data. All AI-assisted outputs were critically reviewed and verified by the author, who takes full responsibility for the accuracy, integrity, and final content of the manuscript.

3. Results

3.1. Primary result: N = 12 gauge advantage

Table 3 gives the integer counts and **Table 4** the inference at all three levels for the primary dataset; **Figure 1** places the result in the context of the N-sweep. The finite-horizon gauge succeeded in $30{,}142$ of $57{,}800$ shots ($\hat{p} = 0.5215$) against $24{,}838$ of $58{,}400$ for the identity gauge ($\hat{p} = 0.4253$).

Table 3. Primary comparison at $N = 12$ (Q2 planned pass, calibrated initial state): integer counts.

Gauge	Jobs	Shots n	Successes k	$\hat{p}$ (Wilson 95% CI)	$\mathcal{R}$ (%)
Identity (I)	292	58,400	24,838	0.4253 [0.4213, 0.4293]	53.0
Finite-opt (U_{12})	289	57,800	30,142	0.5215 [0.5174, 0.5256]	58.1

N: number of encoded steps; I: identity gauge; U_{12}: optimized gauge; n: shots; k: successful shots; $\hat{p} = k/n$: observed success probability; CI: confidence interval (Wilson 95%); $\mathcal{R}$: retention as a percentage of the ideal success probability; pp: percentage points. Of 234,200 total shots, 116,200 enter this comparison and 118,000 enter the asymptotic-state control. Ideal prediction: $\Delta_{\text{sim}} = +9.61$ pp.

Table 4. Primary $N = 12$ inference for $\Delta = p_{\text{succ}}(U_{12}) - p_{\text{succ}}(I)$ at shot, job and block levels, with the pre-specified mixed-effects model.

Level	$\hat{\Delta}$ (pp)	95% CI (pp)	Test
Shot (descriptive)	$+9.62$	$[+9.05, +10.19]$ (Newcombe)	$z = 32.8$, $p < 10^{-235}$; assumes independent shots
Job (cluster-robust)	$+9.62$	$[+8.39, +10.85]$ (Welch)	$t_{574} = 15.3$, $p < 10^{-44}$; DEFF 4.3 (I), 4.9 (U_{12})
Block (confirmatory)	$+9.55$ (mean of 30 Δ_b)	$[+8.04, +11.07]$ (t); $[+8.16, +11.05]$ (block bootstrap)	paired $t_{29} = 12.9$, $p < 10^{-12}$; Wilcoxon $p = 1.9 \times 10^{-9}$; sign test 30/30, $p = 1.9 \times 10^{-9}$
Mixed model	$+9.47$ (marginal)	odds ratio 1.47 $[1.38, 1.56]$	$\beta_G = 0.383 \pm 0.030$, Wald $z = 12.6$, $p < 10^{-35}$; LR = 55.3; $\sigma_u = 0.14$, $\sigma_v = 0.14$

N: number of steps; I: identity gauge; U_{12}: optimized gauge; $\hat{\Delta}$: empirical difference, in percentage points (pp); CI: confidence interval; z and t: test statistics; DEFF: variance-based design effect of 200-shot jobs; LR: likelihood-ratio statistic; β_G: fixed gauge coefficient; σ_u, σ_v: random-effect standard deviations. The block analysis is confirmatory; the shot-level analysis is descriptive.

The three inferential levels agree on the point estimate ($+9.5$ to $+9.6$ pp) but not on its uncertainty. The job-level standard deviation of the 200-shot success fractions is 0.072 (I) and 0.078 (U_{12}), against a binomial expectation of 0.035; the design effect of 4–5 means that each 200-shot job carries roughly the information of 40–50 independent shots, and the shot-level interval is correspondingly too narrow by a factor of about two. The standard deviation of the 30 block differences, 4.1 pp, is 2.5 times the value expected from within-block binomial noise (1.6 pp), and the mixed model attributes this to block-specific variation in both the baseline ($\sigma_u = 0.14$ on the logit scale) and the gauge effect ($\sigma_v = 0.14$); the likelihood-ratio statistic for the random effects is 740. The confirmatory block-level interval, $[+8.0, +11.1]$ pp, is therefore the uncertainty we attach to the result. $\Delta_{\text{sim}} = +9.61$ pp lies inside it; the difference $\bar{\Delta}_{\text{hw}} - \Delta_{\text{sim}} = -0.06$ pp has 95% CI $[-1.58, +1.45]$ pp, so

the data are compatible with the simulated difference but cannot distinguish it from values about 1.5 pp away. Both gauges retain only 53–58% of their ideal success probability, so the agreement concerns the difference, not the absolute probabilities.

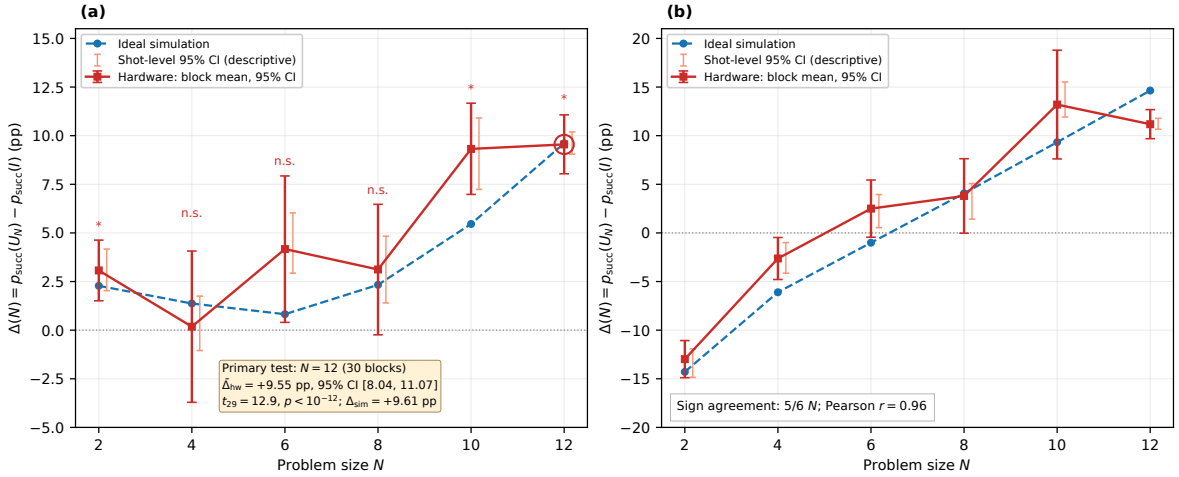


Figure 1. Gauge advantage $\Delta(N) = p_{\mathrm{succ}}(U_N) - p_{\mathrm{succ}}(I)$ in percentage points (pp) versus number of encoded steps N. (**a**) Calibrated-state comparison at each gauge's optimal input. (**b**) Asymptotic-state comparison of frame-related inputs; the ideal simulation predicts a sign change between $N = 6$ and $N = 8$. In both panels, blue dashed lines with circles denote ideal simulation; red solid lines with squares denote hardware block-mean differences and thick vertical bars their block-level 95% confidence intervals (CIs); thin peach-colored offset bars denote descriptive shot-level Newcombe 95% CIs, which understate uncertainty under between-job correlation. In (**a**), the open red circle highlights the primary $N = 12$ result; asterisks mark the significant Holm-adjusted secondary tests ($N = 2, 10$) and the primary test ($N = 12$), while n.s. means not significant. Panel (**b**) also reports Pearson $r = 0.96$ across the six N values. The thin grey dashed horizontal reference line denotes zero advantage ($\Delta = 0$). Here Δ_{hw} is the block-mean hardware gauge difference, Δ_{sim} is the noiseless predicted difference, and t_{29} denotes the paired Student t statistic with 29 degrees of freedom (30 paired blocks). Figure created using AI-assisted Python plotting with ChatGPT (OpenAI GPT-6).

Figure 2a shows the 30 paired block differences. Every block favours the optimized gauge (minimum $\Delta_b = +0.3$ pp, maximum $+17.2$ pp). The eight blocks executed sequentially on 5–6 September have a higher mean ($+12.9$ pp, SD 3.1) than the 22 blocks executed by three concurrent schedulers on 7–8 September ($+8.3$ pp, SD 3.7; Welch $p = 0.004$); the identity-gauge probability was nearly identical in the two sessions (0.427 vs. 0.425) and the difference arises from the optimized gauge (0.556 vs. 0.509). We have no independent calibration evidence that discriminates between the two sessions (Section 3.5) and treat this as between-session variability of the effect, which is exactly the component the block-level analysis is designed to absorb. Two qualifications apply. First, the two sessions differed in more than one respect—different days, and sequential versus three concurrently submitting schedulers whose jobs were interleaved on the device—so the difference cannot be attributed to any single cause. Second, the pooled block-level interval treats the 30 blocks as draws from one population; if the sessions are instead analysed separately, the effect is $+12.9$ pp $[+10.3, +15.5]$ (8 blocks) and $+8.3$ pp $[+6.7, +10.0]$ (22 blocks), both excluding zero, and a session-stratified block bootstrap of the overall mean gives $[+8.3, +10.8]$ pp, close to the pooled t interval $[+8.0, +11.1]$ pp. The size of the advantage at $N = 12$ is thus established in both sessions, while its exact magnitude varies between sessions by more than the within-session uncertainty.

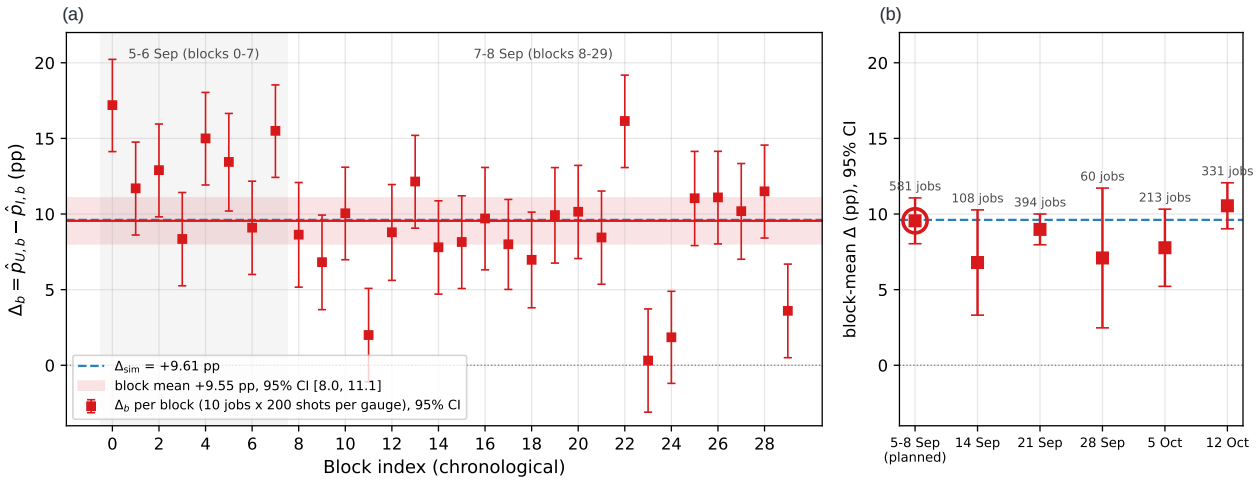


Figure 2. (**a**) Paired gauge difference Δ_b for all 30 blocks in the planned $N = 12$ run (ten 200-shot jobs per gauge per block); vertical red bars are single-block Newcombe 95% confidence intervals (CIs), the red horizontal line and translucent band show the block mean and its 95% CI, and the dashed blue line is the ideal simulated difference. Shading marks the sequential session of 5–6 September 2026. (**b**) Block-mean gauge difference in percentage points (pp) across the planned run and five unplanned weekly repetitions; red squares and red vertical bars show block means and t-based 95% CIs, the dashed blue line is the noiseless prediction, and the open red circle highlights the planned primary result. Labels give the number of contributing calibrated-state jobs. The thin grey dashed horizontal reference lines in both panels denote zero gauge difference. In panel (**a**), $\Delta_b = \hat{p}_{U,b} - \hat{p}_{I,b}$, with b denoting a calibration block, $\hat{p}_{U,b}$ and $\hat{p}_{I,b}$ the observed success fractions for optimized and identity gauges, respectively, and Δ_{sim} the noiseless predicted gauge difference. In panel (**b**), Δ denotes the weekly block-mean gauge difference; horizontal labels indicate execution dates and job-count annotations specify contributing calibrated-state jobs. Figure created using ChatGPT (OpenAI GPT-6).

3.2. Temporal stability: unplanned weekly re-executions

The five re-executions of blocks 8–29 constitute 2171 further $N = 12$ jobs (434,200 shots) spread over five weeks, of which 1106 jobs (221,200 shots) are calibrated-state cells (**Figure 2b**, **Table 5**). The block-mean difference was positive and significant in every weekly pass, ranging from $+6.8$ pp (14 September, 6 blocks) to $+10.5$ pp (12 October, 22 blocks). The two weeks with the lowest differences (14 and 28 September) were also the weeks with the lowest absolute success probabilities for both gauges and with the highest submission-failure rates, consistent with degraded device or network conditions on those days. Pooling all 2171 replication jobs gives $\hat{p}_I = 44{,}999/110{,}400 = 0.4076$, $\hat{p}_U = 55{,}098/110{,}800 = 0.4973$ and a block-level mean of $+9.22$ pp $[+8.11, +10.33]$ (22 blocks); pooling all 3342 executions of the campaign gives $+10.01$ pp $[+8.88, +11.14]$ over 30 blocks (shot-level $\hat{\Delta} = +9.18$ pp). The replication data thus reproduce the primary result independently and indicate that the week-to-week variation of Δ (roughly ± 2 pp) exceeds the within-week statistical uncertainty.

Table 5. Unplanned weekly re-executions of the calibrated-state $N = 12$ schedule (blocks 8–29).

Launch	Jobs (I/U)	$\hat{p}_I$	$\hat{p}_U$	$\hat{\Delta}$ (pp)	Newcombe CI	B	$\bar{\Delta}_b$ [CI]
14 September 2026	53/55	0.3684	0.4374	$+6.90$	$[+5.59, +8.20]$	6	$+6.79$ $[+3.31, +10.27]$
21 September 2026	198/196	0.4264	0.5175	$+9.10$	$[+8.41, +9.80]$	22	$+8.98$ $[+7.97, +10.00]$
28 September 2026	32/28	0.3887	0.4596	$+7.09$	$[+5.32, +8.86]$	4	$+7.09$ $[+2.47, +11.72]$
5 October 2026	106/107	0.4085	0.4857	$+7.72$	$[+6.78, +8.66]$	12	$+7.77$ $[+5.22, +10.32]$
12 October 2026	163/168	0.4006	0.5069	$+10.64$	$[+9.88, +11.39]$	22	$+10.54$ $[+9.02, +12.07]$
Re-executions pooled	552/554	0.4076	0.4973	$+8.97$	$[+8.55, +9.38]$	22	$+9.22$ $[+8.11, +10.33]$
Planned pass	292/289	0.4253	0.5215	$+9.62$	$[+9.05, +10.19]$	30	$+9.55$ $[+8.04, +11.07]$

N: number of encoded steps; I: identity gauge; U: optimized gauge; Jobs: successful 200-shot jobs by gauge; $\hat{p}$: empirical success fraction; $\hat{\Delta}$: difference in percentage points (pp); CI: 95% confidence interval; Newcombe: descriptive shot-level interval; B: number of blocks; $\bar{\Delta}_b$: block-mean paired difference, with t-based 95% CI. Each weekly pass ended when the device became unreachable or the scheduled run finished.

3.3. Multi-*N* sweep

The planned 15-block N-sweep is now complete (562 of 600 scheduled jobs returned; 25–30 jobs, i.e., 5000–6000 shots, per circuit). **Table 6** and **Figure 1** present the calibrated-state comparison for each N. The point estimate of Δ_{hw} is positive at all six problem sizes. With block-level inference and Holm adjustment over the five secondary comparisons, the advantage is significant at $N = 2$ ($p_{\text{Holm}} = 0.004$) and $N = 10$ ($p_{\text{Holm}} = 5\times10^{-6}$) in addition to the primary $N = 12$ result, but not at $N = 4$ ($p_{\text{Holm}} = 0.92$), $N = 6$ (0.10) or $N = 8$ (0.13). The shot-level z tests would declare $N = 6$ and $N = 8$ significant as well; the discrepancy reflects design effects of 3–7 at those sizes and illustrates why shot-level tests are not used for inference here. At $N = 4$ the predicted effect is small ($\Delta_{\text{sim}} = +1.4$ pp) and one U_4 job returned an anomalously low success fraction (0.41, all other 50 jobs ≥ 0.75); it was retained because no exclusion rule was pre-specified, and it accounts for the very large design effect at $N = 4$. In the Q3 replication set (467 later re-executions) the differences were $+3.0$, $+4.2$, $+1.9$, $+3.4$ and $+9.4$ pp for $N = 2, 4, 6, 8, 10$ (shot-level), again positive throughout. We therefore regard the direction of the effect as consistently supported across N, its magnitude as established at $N = 2$, 10 and 12, and its magnitude at $N = 4$–8 as not yet resolved by the available shots.

Frame-related control across N: The asymptotic-state cells use the same input vector in both frames ($\tilde{\psi} \propto U_N^{-1}\psi$, Section 2.2.1) and therefore compare the same computation in two representations. For these cells the simulation predicts that the optimized gauge is worse at $N = 2$, 4 and 6 ($\Delta_{\text{sim}} = -14.3, -6.1, -1.0$ pp) and better at $N = 8$, 10 and 12 ($+4.1$, $+9.3$, $+14.6$ pp). The hardware reproduces this sign change (**Table 7**, **Figure 1b**): block-mean differences of -13.0 pp $[-14.9, -11.1]$ at $N = 2$ and -2.6 pp $[-4.8, -0.5]$ at $N = 4$, $+13.2$ pp $[+7.6, +18.8]$ at $N = 10$ and $+11.2$ pp $[+9.7, +12.7]$ at $N = 12$; at $N = 6$ ($+2.5$ pp $[-0.4, +5.4]$, predicted -1.0) and $N = 8$ ($+3.8$ pp $[0.0, +7.6]$, predicted $+4.1$) the intervals include small values of either sign. The sign agrees at five of six N and the Pearson correlation between predicted and measured block-mean differences across N is $r = 0.96$ (calibrated cells: $r = 0.85$). These cells were not part of the pre-specified hypothesis and are reported descriptively; they show that the hardware follows the ideal prediction—including its reversal of sign—for a comparison in which

the input state is held fixed up to the frame change, and they complement the primary comparison, in which each encoding is evaluated at its own optimal input.

Table 6. Calibrated-state N-sweep, primary datasets (Q3 first success per slot for $N \leq 10$, Q2 planned pass for $N = 12$).

N	k_I/n_I (jobs)	k_U/n_U (jobs)	Δ_{sim}	$\hat{\Delta}$ shot [CI]	B	$\bar{\Delta}_b$ [CI]	p_{Holm}	Sig.
2	4662/5200 (26)	5194/5600 (28)	$+2.28$	$+3.10$ $[+2.03, +4.17]$	14	$+3.07$ $[+1.51, +4.63]$	0.0038	Yes
4	4223/5000 (25)	4410/5200 (26)	$+1.37$	$+0.35$ $[-1.05, +1.75]$	14	$+0.18$ $[-3.71, +4.07]$	0.92	No
6	4145/5600 (28)	4710/6000 (30)	$+0.82$	$+4.48$ $[+2.93, +6.03]$	15	$+4.17$ $[+0.40, +7.93]$	0.097	No
8	3756/6000 (30)	3943/6000 (30)	$+2.33$	$+3.12$ $[+1.40, +4.83]$	15	$+3.12$ $[-0.24, +6.47]$	0.13	No
10	2692/5400 (27)	3418/5800 (29)	$+5.46$	$+9.08$ $[+7.24, +10.91]$	14	$+9.32$ $[+6.98, +11.67]$	5×10^{-6}	Yes
12	24,838/58,400 (292)	30,142/57,800 (289)	$+9.61$	$+9.62$ $[+9.05, +10.19]$	30	$+9.55$ $[+8.04, +11.07]$	$<10^{-12}$ (primary)	Yes

N: number of encoded steps; I and U: identity and optimized gauge; k/n: successes/shots (successful 200-shot jobs in parentheses); Δ_{sim}: ideal predicted difference in percentage points (pp); $\hat{\Delta}_{shot}$: observed shot-level difference, with descriptive Newcombe 95% confidence interval (CI); B: paired blocks; $\bar{\Delta}_b$: block-mean difference with t-based 95% CI; p_{Holm}: Holm-adjusted block-level p for five secondary comparisons; Sig.: significance after adjustment. $N = 12$ is the pre-specified, unadjusted primary test.

Table 7. Asymptotic-state (frame-related) control for the same input vector expressed in two frames, primary datasets.

N	k_I/n_I (jobs)	k_U/n_U (jobs)	Δ_{sim}	$\hat{\Delta}$ shot [CI]	B	$\bar{\Delta}_b$ [CI]
2	4890/5600 (28)	3992/5400 (27)	-14.29	-13.40 $[-14.85, -11.93]$	14	-12.98 $[-14.89, -11.07]$
4	4350/5600 (28)	4206/5600 (28)	-6.10	-2.57 $[-4.14, -1.00]$	15	-2.63 $[-4.79, -0.47]$
6	3844/5800 (29)	3974/5800 (29)	-1.01	$+2.24$ $[+0.54, +3.95]$	15	$+2.50$ $[-0.44, +5.44]$
8	3081/5600 (28)	3263/5600 (28)	$+4.07$	$+3.25$ $[+1.41, +5.08]$	15	$+3.80$ $[-0.03, +7.63]$
10	2508/5800 (29)	3304/5800 (29)	$+9.33$	$+13.72$ $[+11.92, +15.52]$	15	$+13.20$ $[+7.61, +18.79]$
12	21,318/58,600 (293)	28,272/59,400 (297)	$+14.64$	$+11.22$ $[+10.66, +11.78]$	30	$+11.18$ $[+9.69, +12.68]$

N: number of encoded steps; I and U: identity and optimized gauges; k/n: successes/shots (successful 200-shot jobs in parentheses); Δ_{sim}: ideal predicted difference in percentage points (pp); $\hat{\Delta}_{shot}$: observed shot-level difference, with descriptive Newcombe 95% confidence interval (CI); B: blocks; $\bar{\Delta}_b$: paired block-mean difference with t-based 95% CI. These are descriptive comparisons; no multiplicity adjustment is applied.

3.4. Hardware retention and the noise-interaction hypothesis

Figure 3 shows the retention $\mathcal{R} = \hat{p}/p_{ideal}$. Retention falls from about 92% at $N = 2$ to 53–58% at $N = 12$ for both gauges, as expected from the accumulation of gate errors with depth. The optimized gauge has higher retention than the identity gauge at $N = 2$, 6, 8, 10 and 12 and lower retention at $N = 4$ (85.9% vs. 86.7%); the differences at $N = 2$ and $N = 4$ (about ± 1 pp) are within the uncertainty, so the previously stated "consistently higher at every N" does not hold in the completed sweep.

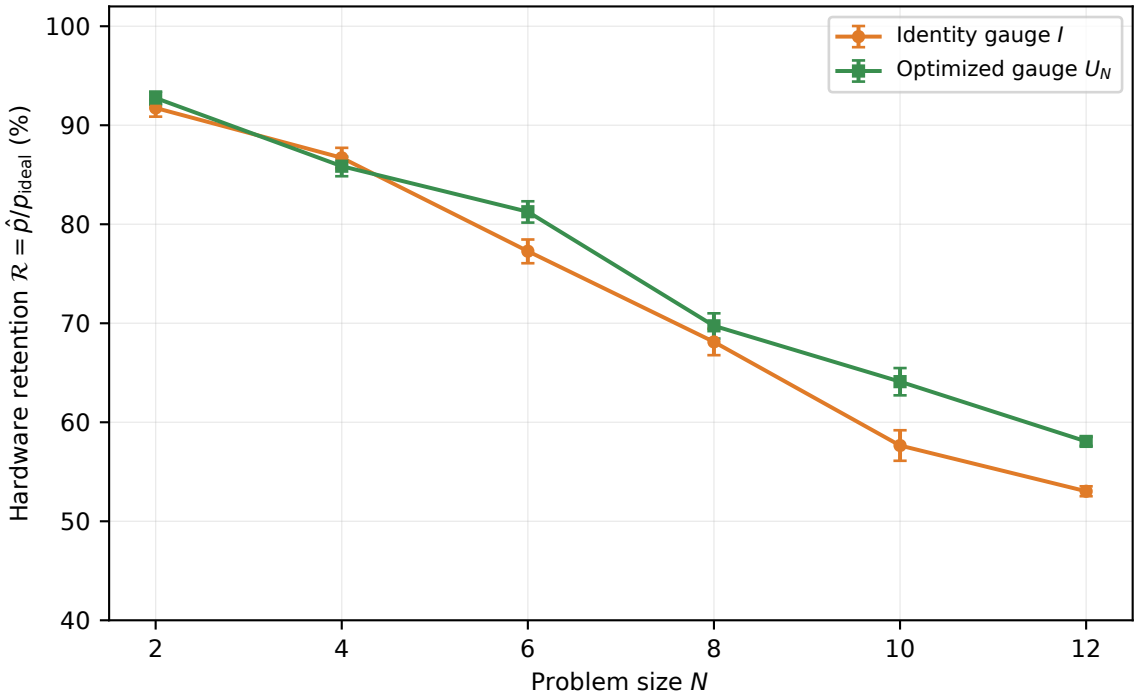


Figure 3. Hardware retention $\mathcal{R} = \hat{p}/p_{\text{ideal}}$ (percentage of ideal success probability) by number of encoded steps N in primary datasets. Orange circles and green squares denote identity gauge I and optimized gauge U_N, respectively; vertical bars denote Wilson 95% confidence intervals of $\hat{p}$, divided by p_{ideal}. These shot-level intervals do not account for correlation between jobs and are therefore optimistic. The ordering at $N = 2$ and $N = 4$ is not resolved. Figure created using AI-assisted Python plotting with ChatGPT (OpenAI GPT-6).

Whether the optimized gauge interacts differently with hardware noise can be assessed by comparing Δ_{hw} with Δ_{sim}. If noise acted as a gauge-independent multiplicative factor, $\hat{p} \approx \mathcal{R}\, p_{\text{ideal}}$ with a common $\mathcal{R} \approx 0.55$, one would expect $\Delta_{\text{hw}} \approx 0.55 \times 9.61 = 5.3$ pp at $N = 12$; the observed block-level interval $[+8.0, +11.1]$ pp excludes this value. The absolute losses $\mathcal{L} = p_{\text{ideal}} - \hat{p}$ of the two calibrated $N = 12$ circuits are instead equal to within 0.01 pp (0.3767 for I, 0.3766 for U_{12}), which is what preserves the ideal difference; both circuits have identical gate counts (24 R_{XX}, 61 R, 96 R_z) and differ only in rotation angles. The asymptotic-state control at $N = 12$ does not show the same pattern: there $\Delta_{\text{sim}} = +14.64$ pp but $\bar{\Delta}_{\text{hw}} = +11.18$ pp $[+9.69, +12.68]$ (losses 0.326 for I and 0.360 for U_{12}), so the hardware difference is smaller than the ideal one for that state. Across the sweep, the excess $\bar{\Delta}_{\text{hw}} - \Delta_{\text{sim}}$ with block-level intervals is $+0.8$ $[-0.8, +2.4]$, -1.2 $[-5.1, +2.7]$, $+3.4$ $[-0.4, +7.1]$, $+0.8$ $[-2.6, +4.1]$, $+3.9$ $[+1.5, +6.2]$ and -0.1 $[-1.6, +1.5]$ pp for $N = 2, 4, 6, 8, 10, 12$; only $N = 10$ excludes zero before any multiplicity adjustment. We conclude that the noise-induced loss is not a common multiplicative factor, that it is of similar absolute size for circuits of identical gate structure, and that the data provide no consistent evidence for an intrinsic noise-robustness of the optimized gauge. The mechanism previously suggested (a more favourable distribution of rotation angles relative to the two-pulse threshold) remains a hypothesis that would require a noise-model study or a controlled experiment to test.

3.5. Calibration stability

Figure 4a shows d_{TV} for the four three-qubit calibration circuits over the 30 blocks of the Q2 planned pass (59–60 executions each; 241 calibration jobs in total). All four series are low and show no trend: $\mathcal{C}_0$ mean 0.0007

(SD 0.0017, maximum 0.005), $\mathcal{C}_1$ mean 0.018 (SD 0.010, maximum 0.045), $\mathcal{C}_8$ mean 0.072 (SD 0.015, maximum 0.110) and $\mathcal{C}_9$ mean 0.053 (SD 0.018, maximum 0.090). With 200 shots, the sampling contribution to d_{TV} for these few-outcome distributions is of order $0.01–0.03$, so the $\mathcal{C}_8$ and $\mathcal{C}_9$ values reflect genuine gate error at the level of a few per cent per circuit. Because these circuits act on three qubits, they monitor SPAM, single-qubit and one entangling gate but cannot resolve the session difference observed in the 13-qubit CMF circuits (**Figure 2a**); the stability conclusion is accordingly limited to what these four series measure.

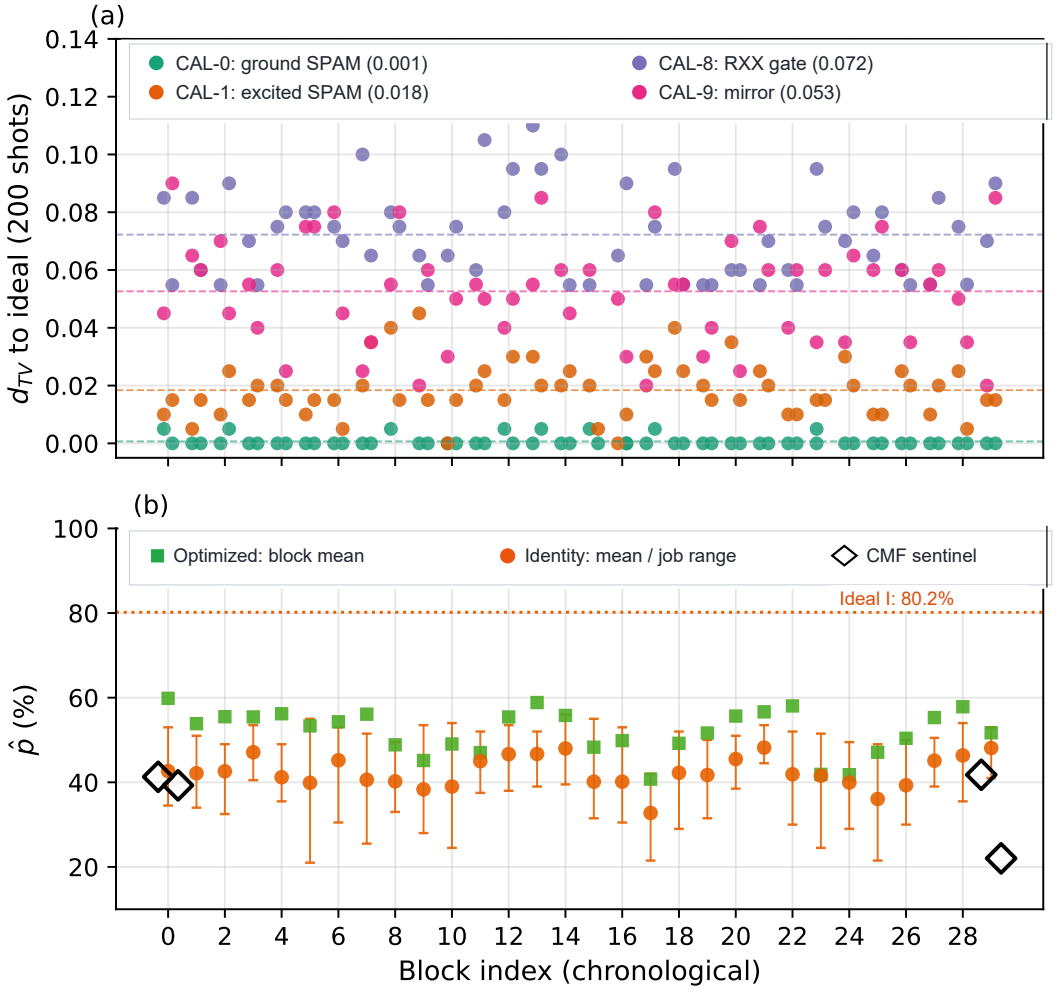


Figure 4. (**a**) Total variation distance d_{TV} from the ideal distribution for four three-qubit calibration circuits, before and after the 30 Q2 blocks: green circles, CAL-0 (ground-state SPAM); orange circles, CAL-1 (excited-state SPAM); purple circles, CAL-8 (single R_{XX} entangling gate); pink circles, CAL-9 (circuit followed by its inverse). Dashed lines in corresponding colors denote series means. CAL: calibration; SPAM: state preparation and measurement; R_{XX}: native two-qubit entangling gate. (**b**) Post-selection success probabilities for the 13-qubit Conservative Matrix Field (CMF) circuits at $N = 12$: green squares, optimized-gauge block means; orange circles with vertical bars, identity-gauge block means with job-level minima/maxima; black open diamonds, four CMF sentinel runs; orange dotted horizontal line, ideal identity-gauge success probability (0.802). Both panels use chronological block index 0–29 as their horizontal axis; pp: percentage points. Figure created using ChatGPT (OpenAI GPT-6).

The fifth series in the originally submitted figure, labelled `CMF_SENTINEL`, is not a calibration circuit in the above sense. It is the 13-qubit $N = 12$ identity-gauge circuit itself, run twice in block 0 and twice in block 29 (the two

full-suite blocks), with d_{TV} computed against the 4096-outcome reference distribution defined in Section 2.4 (0.802 on the all-zero string, the remaining 0.198 spread uniformly over 4095 strings, 4.8×10^{-5} each). For such a reference a 200-shot sample has an unavoidable sampling floor of $d_{TV} \approx 0.21$ even from a perfect device, and in the presence of hardware loss $d_{TV} \approx 1 - \hat{P}(0^{12})$ to within 0.005. The four sentinel values 0.587, 0.607, 0.582 and 0.780 thus correspond to $\hat{P}(0^{12}) \approx 0.41$, 0.39, 0.42 and 0.22. The first three agree with the campaign estimate for the identical circuit (0.425); the last, the post-calibration job at the very end of the 7–8 September session, is low but lies within the range of the 292 identity-gauge jobs of the pass (minimum 0.21, **Figure 4b**). The sentinel therefore reports the same hardware loss that the primary experiment measures, not an additional instability of the processor, and we no longer describe it as a calibration series.

Figure 5 summarizes the frozen-circuit resource scaling and the two-pulse rule.

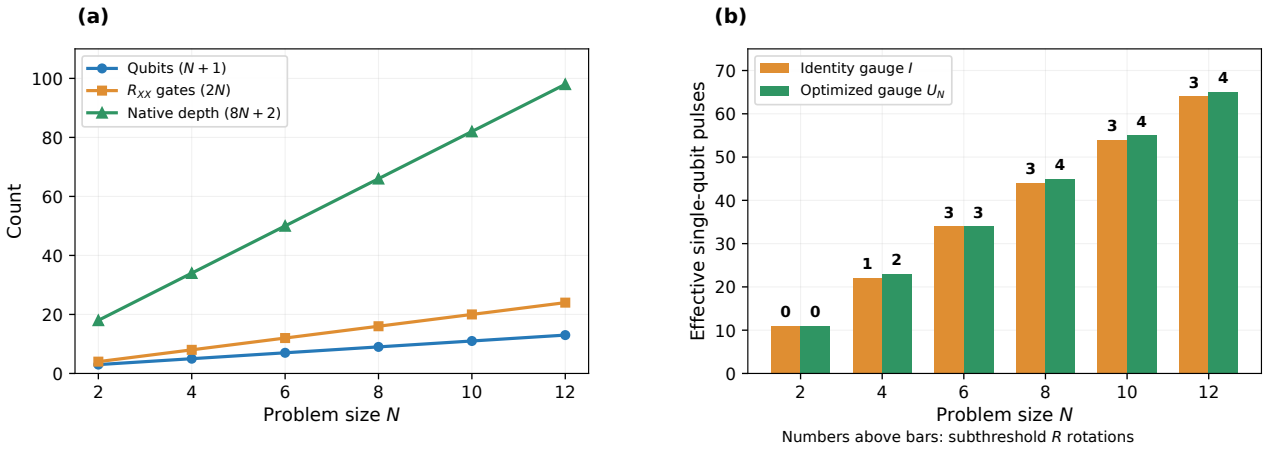


Figure 5. Circuit resources versus encoded step count N. (**a**) Qubits (blue circles), R_{XX} entangling gates (orange squares) and native gate depth (green triangles) of the frozen circuits. (**b**) Effective single-qubit pulses under the Alpine Quantum Technologies (AQT) two-pulse rule for the identity gauge I (orange bars) and optimized gauge U_N (green bars); numbers above each individual bar give the counts of subthreshold R rotations for the corresponding gauge. Figure created using AI-assisted Python plotting with ChatGPT (OpenAI GPT-6).

4. Discussion

The primary finding is that a constant change of basis that leaves the encoded matrix product similar to the original one (identical eigenvalues, trace and determinant, verified from the circuit artifacts), and that adds no entangling gates, changes the measured post-selection probability on a trapped-ion processor by $+9.6$ pp at $N = 12$, with a confirmatory 95% interval of $[+8.0, +11.1]$ pp that accounts for the block structure of the experiment. The noiseless simulation—with no noise model and no free parameters—predicted $+9.61$ pp, which lies inside this interval. The result was reproduced in five unplanned weekly re-executions over the following five weeks. These are the two statements that the data establish.

The second-order statements require more caution than in the originally submitted version. First, the agreement between $\bar{\Delta}_{\mathrm{hw}}$ and Δ_{sim} should not be read as "to within 0.05 pp": the point estimates happen to coincide, but the confidence interval is ± 1.5 pp wide, the week-to-week replication values range from $+6.8$ to $+10.5$ pp, and the asymptotic-state control at the same N shows a hardware difference about 3.4 pp below its ideal value. What the data support is that the ideal difference is a good predictor of the hardware difference for the calibrated-state

circuits at $N = 12$, to within roughly ± 2 pp. Second, the N-sweep is complete but the effect is established only at $N = 2$, 10 and 12; at $N = 4$, 6 and 8 the point estimates are positive but the block-level intervals include zero after multiplicity adjustment. A claim that the advantage "increases with N" is consistent with the data but is not something the sweep demonstrates. Third, the retention and excess-over-simulation analyses do not provide consistent evidence for an intrinsic noise-robustness of the optimized gauge. The clearest empirical regularity is that the two calibrated $N = 12$ circuits, which share gate counts and differ only in angles, lose the same absolute probability mass (0.377); an additive-loss description fits those cells, whereas it does not fit the asymptotic-state cells. We offer no mechanism for this and regard the earlier suggestion involving rotation-angle distributions and the two-pulse decomposition as an untested hypothesis. In particular, we withdraw the statement that the advantage would be "at least as large" on hardware without the two-pulse rule; the one additional subthreshold rotation in the U_N circuits at $N = 4$, 8, 10 and 12 is a documented asymmetry, but its effect was not measured.

Several limitations should be noted. The experiment used one CMF trajectory, one family of constant gauges, one hardware platform and one compilation path; the generality of the effect across these dimensions is untested. The planned pass executed blocks 8–29 through three concurrent schedulers rather than sequentially, so those blocks were interleaved in time; the block-level analysis is valid under this execution, but the between-session difference in the size of the effect (Section 3.1) is unexplained. The scheduler fault that produced the replication data also overwrote result files, so the primary dataset relies on the run logs; the reconstruction is exact for success counts but shot-by-shot memory is available only for the last execution of each slot. The Q3 schedule was completed across several weeks, so its blocks are not contiguous in time. The task-equivalence argument of Section 2.2.1 relies on the gauge being constant along the chain; position-dependent gauges with unequal endpoints would change the finite-product invariants and would require a separate equivalence argument before their post-selection probabilities could be compared in the same way. Finally, inference on the N-sweep is limited by the design effects, which at some N reduce the effective information of 200-shot jobs by a factor of 3–7; future campaigns should plan sample sizes at the block level rather than at the shot level.

With these qualifications, the results have three implications for near-term quantum algorithm design. (1) Representation choices that are classically free can produce measurable, predictable performance differences on a real processor. (2) For the circuits studied, the noiseless simulation predicted the relative performance of two representations even though it predicts neither absolute probability; whether this holds more generally is an empirical question that the block-level methodology used here can address at modest cost. (3) Gauge optimization adds no entangling gates and is computed classically before compilation; it is not entirely free, however. Its costs are classical (the gauge search and, for the calibrated cells, the computation of the gauge-specific optimal input state, which presupposes classical knowledge of the product) and, on this platform, at most one additional effective single-qubit pulse at $N = 4$, 8, 10 and 12 (Section 2.3); the asymptotic-state cells, whose input is N-independent and gauge-covariant, show that the effect does not depend on the gauge-specific input choice.

5. Conclusions

On a 20-qubit trapped-ion processor, the finite-horizon optimized gauge U_{12} achieved a post-selection success probability 9.6 pp higher than the identity gauge for the same Conservative Matrix Field product at $N = 12$ (block-level 95% CI $[8.0, 11.1]$ pp; paired $t_{29} = 12.9$, $p < 10^{-12}$; 30 of 30 blocks positive), in agreement with the noiseless prediction of 9.61 pp, and the effect reproduced in five subsequent weekly re-executions.

The completed N-sweep shows positive point estimates at all N and significant, Holm-adjusted advantages at $N = 2$ and $N = 10$. Within the tested CMF implementation and hardware platform, gauge choice is therefore a practically relevant and simulation-predictable design parameter for post-selected block-encoding circuits. Its generality across problem sizes, platforms and compilation strategies, and the hypothesis that the optimized gauge is intrinsically more robust to hardware noise, remain to be established.

Acknowledgments

Quantum hardware access was provided by the PIAST-Q trapped-ion facility through the PCSS-QAPI interface. The author thanks the PCSS team for technical support and access. The author acknowledges an in-kind allocation of 25 QPU-hours on PIAST-Q made available through the European High Performance Computing Joint Undertaking (EuroHPC JU); no direct financial grant was associated with this allocation. The author acknowledges the use of GLM-5.2 for original figure generation and drafting, Devin (Cognition; version/build not recorded; using Anthropic Claude) for execution-log reconstruction, analysis scripts, revised figures and drafting, and ChatGPT (OpenAI GPT-6) for final proofreading and figure adjustments, as declared in Section 2.7. The author fully supports Academia.edu Journals' adherence to COPE guidelines on AI in publication ethics and confirms that AI-assisted material has been critically reviewed and used responsibly.

Funding

This research received no external funding.

Conflict of interest

The author is the founder of Vesterlund Ventures Holding AB. The PIAST-Q/PCSS facility provided hardware access and technical support, but played no role in the study's conception, design, execution, or interpretation, or with the writing of this manuscript. The author declares no other financial or personal competing interests.

Data availability statement

The data supporting the findings of this publication (code and selected project artifacts) has been made available within a publicly accessible repository at https://github.com/VesterlundCoder/QuantumCMF. The additional experimental data and analysis materials supporting the findings of this publication can be made available upon request.

Publisher's note

Academia.edu Journals stays neutral regarding jurisdictional claims in published maps and institutional affiliations. All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright

© 2026 copyright by the author. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

References

1. Elimelech R, David O, De la Cruz Mengual C, Kalisch R, Berndt W, Shalyt M, et al. Algorithm-assisted discovery of an intrinsic order among mathematical constants. Proc Natl Acad Sci USA. 2024; 121 (25): e2321440121. doi: 10.1073/pnas.2321440121
2. Raz T, Shalyt M, Leibtag E, Kalisch R, Weinbaum S, Hadad Y, et al. From Euler to AI: unifying formulas for mathematical constants. In: Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS 2025); 2–7 December 2025; San Diego, CA, USA, and 30 November–5 December 2025; Mexico City, Mexico. San Diego, CA, USA: Neural Information Processing Systems Foundation; 2025. p. 137390–471. doi: 10.52202/085713-4132
3. Raayoni G, Gottlieb S, Manor Y, Pisha G, Harris Y, Mendlovic U, et al. Generating conjectures on fundamental constants with the Ramanujan Machine. Nature. 2021; 590 (7844): 67–73. doi: 10.1038/s41586-021-03229-4
4. Low GH, Chuang IL. Optimal Hamiltonian simulation by quantum signal processing. Phys Rev Lett. 2017; 118 (1): 010501. doi: 10.1103/PhysRevLett.118.010501
5. Gilyén A, Su Y, Low GH, Wiebe N. Quantum singular value transformation and beyond: exponential improvements for quantum matrix arithmetics. In: Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing (STOC); 23–26 June 2019; Phoenix, AZ, USA. New York, NY, USA: Association for Computing Machinery; 2019. p. 193–204. doi: 10.1145/3313276.3316366
6. Low GH, Chuang IL. Hamiltonian simulation by qubitization. Quantum. 2019; 3: 163. doi: 10.22331/q-2019-07-12-163
7. Martyn JM, Rossi ZM, Tan AK, Chuang IL. Grand unification of quantum algorithms. PRX Quantum. 2021; 2 (4): 040203. doi: 10.1103/PRXQuantum.2.040203
8. Camps D, Lin L, Van Beeumen R, Yang C. Explicit quantum circuits for block encodings of certain sparse matrices. SIAM J Matrix Anal Appl. 2024; 45 (1): 801–27. doi: 10.1137/22M1484298
9. Sünderhauf C, Campbell E, Camps J. Block-encoding structured matrices for data input in quantum computing. Quantum. 2024; 8: 1226. doi: 10.22331/q-2024-01-11-1226
10. Lin L. Lecture notes on quantum algorithms for scientific computation. arXiv. 2022. doi: 10.48550/arXiv.2201.08309
11. Brassard G, Høyer P, Mosca M, Tapp A. Quantum amplitude amplification and estimation. Contemp Math. 2002; 305: 53–74.
12. Dong Y, Lin L, Tong Y. Ground-state preparation and energy estimation on early fault-tolerant quantum computers via quantum eigenvalue transformation of unitary matrices. PRX Quantum. 2022; 3 (4): 040305. doi: 10.1103/PRXQuantum.3.040305
13. Preskill J. Quantum computing in the NISQ era and beyond. Quantum. 2018; 2: 79. doi: 10.22331/q-2018-08-06-79
14. Bharti K, Cervera-Lierta A, Kyaw TH, Haug T, Alperin-Lea S, Anand A, et al. Noisy intermediate-scale quantum algorithms. Rev Mod Phys. 2022; 94 (1): 015004. doi: 10.1103/RevModPhys.94.015004
15. Kikuchi Y, Mc Keever C, Coopmans L, Lubasch M, Benedetti M. Realization of quantum signal processing on a noisy quantum computer. npj Quantum Inf. 2023; 9: 93. doi: 10.1038/s41534-023-00762-0

16. Nielsen MA, Chuang IL. Quantum computation and quantum information. 10th anniversary ed. Cambridge, UK: Cambridge University Press; 2010. doi: 10.1017/CBO9780511976667
17. Javadi-Abhari A, Treinish M, Krsulich K, Wood CJ, Lishman J, Gacon J, et al. Quantum computing with Qiskit. arXiv. 2024. doi: 10.48550/arXiv.2405.08810
18. Bruzewicz CD, Chiaverini J, McConnell R, Sage JM. Trapped-ion quantum computing: progress and challenges. Appl Phys Rev. 2019; 6 (2): 021314. doi: 10.1063/1.5088164
19. Moses SA, Baldwin CH, Allman MS, Ancona R, Ascarrunz L, Barnes C, et al. A race-track trapped-ion quantum processor. Phys Rev X. 2023; 13 (4): 041052. doi: 10.1103/PhysRevX.13.041052
20. Pogorelov I, Feldker T, Marciniak CD, Postler L, Jacob G, Krieglsteiner O, et al. Compact ion-trap quantum computing demonstrator. PRX Quantum. 2021; 2 (2): 020343. doi: 10.1103/PRXQuantum.2.020343
21. Postler L, Heußen S, Pogorelov I, Rispler M, Feldker T, Meth M, et al. Demonstration of fault-tolerant universal quantum gate operations. Nature. 2022; 605 (7911): 675–80. doi: 10.1038/s41586-022-04721-1
22. Maslov D. Basic circuit compilation techniques for an ion-trap quantum machine. New J Phys. 2017; 19 (2): 023035. doi: 10.1088/1367-2630/aa5e47
23. Proctor T, Revelle M, Nielsen E, Rudinger K, Lobser D, Maunz P, et al. Detecting and tracking drift in quantum information processors. Nat Commun. 2020; 11: 5396. doi: 10.1038/s41467-020-19074-4
24. Dasgupta S, Humble TS. Characterizing the reproducibility of noisy quantum circuits. Entropy. 2022; 24 (2): 244. doi: 10.3390/e24020244
25. Cai Z, Babbush R, Benjamin SC, Endo S, Huggins WJ, Li Y, et al. Quantum error mitigation. Rev Mod Phys. 2023; 95 (4): 045005. doi: 10.1103/RevModPhys.95.045005
26. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc. 1927; 22 (158): 209–12. doi: 10.1080/01621459.1927.10502953
27. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med. 1998; 17 (8): 873–90. doi: 10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I
28. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979; 6 (2): 65–70.