Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Mean-Matched Controls Fail to Preserve Logical Risk in Rotated Surface-Code Memories with Spatially Resolved Noise

Aditya Kumar Singh Pundir
Poornima Institute of Engineering & Technology
Pallavi Singh
Poornima University Jaipur
Ramji Gupta
ramji.gupta270086@paruluniversity.ac.in

Parul Institute of Engineering & Technology, Faculty of Engineering & Technology, Parul University
Ramesh Kumar
Manipal University Jaipur
Kamal Saini
Poornima Institute of Engineering & Technology

Research Article

Keywords: quantum error correction, rotated surface code, spatially heterogeneous noise, minimum-weight perfect matching, modular quantum computing, statistical validation

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11279713/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Mean-Matched Controls Fail to Preserve Logical Risk in Rotated Surface-Code Memories with Spatially Resolved Noise

Aditya Kumar Singh Pundir[1], Pallavi Singh[2], Ramji Gupta[3a*, 3b], Ramesh Kumar[4], Kamal Saini[5]

[1]Department of Artificial Intelligence and Data Science, Poornima Institute of Engineering & Technology, Jaipur, Rajasthan, India, aditya.pundir@gmail.com

[2]Faculty of Public Health, Poornima University Jaipur, India, pallavi.singh@poornima.edu.in

Corresponding author*: Aditya Kumar Singh Pundir aditya.pundir@gmail.com

[3a*]Department of Electronics & Communication Engineering, Parul Institute of Engineering & Technology, Faculty of Engineering & Technology, Parul University, Vadodara, Gujarat-391760, India, ramji.gupta270086@paruluniversity.ac.in

[3b]Computer Science and Engineering, Lincoln University College, Petaling Jaya, Selangor Darul Ehsan 47301, Malaysia. pdf.ramji@lincoln.edu.my

[4*]Department of Electronics and Communication Engineering, Manipal University Jaipur, Jaipur,India, ramesh.meena@jaipur.manipal.edu

[5]Department of Artificial Intelligence and Data Science, Poornima Institute of Engineering & Technology, Jaipur, Rajasthan, India, kamal08saini@gmail.com

Abstract

Spatially concentrated noise is a central concern in modular quantum error correction, yet its effect is often estimated by replacing a heterogeneous device with a uniform effective error rate. We show that matching the expected number of physical faults does not preserve the logical failure probability in rotated surface-code memories. First, we reanalyse a published 2.6-million-shot proxy benchmark and identify six specific implementation issues that constrain its interpretation: an undocumented seam fraction of 0.5 instead of the configured 0.3, noise injection outside the repeated circuit body, an X error appended after measurement that cannot flip the recorded outcome, an unused p_ebit field, an algebraically generated decay curve presented as data, and a resource annotation not matching any circuit layout. Second, we perform 14.4 million additional sampled memory experiments across 72 conditions, covering code distances three to nine, both logical memory bases, spatially resolved seam noise, a channel-wise mean-matched control, and decoder mismatch. All numerical comparisons use recorded failure counts, with binomial confidence intervals and multiplicity-adjusted hypothesis tests. At code distance three, X-basis memory, and the largest tested noise severity, the geometric seam produces 977 failures while a channel-wise mean-matched homogeneous control produces 1430 failures in 200,000 shots, giving a risk ratio of 0.683 with a 95% confidence interval of 0.630 to 0.741 (Holm-adjusted $p = 4.6 \times 10^{-19}$). Only this $d = 3$ X-basis comparison is statistically decisive at the available shot count; the larger-distance observations are consistent with the same effect but are individually under-powered. Decoding the same geometric-noise samples with calibrated matching

weights reduces failures from 2060 to 977 relative to baseline weights. Z-basis point estimates show the opposite geometry trend, although the selected largest-severity comparisons do not survive familywise adjustment. A simple repetition-code calculation explains why matching the mean physical error probability cannot generally preserve logical risk. The results establish a reproducible method for separating spatial exposure, temporal exposure, and decoder assumptions in seam-noise studies. They support geometry-aware modeling and calibrated decoding within the tested memory circuits; they do not implement a distributed gate, a Bell-pair channel, or classical feedforward, and they do not establish a distributed-gate threshold or hardware-level resource advantage.

Keywords: quantum error correction; rotated surface code; spatially heterogeneous noise; minimum-weight perfect matching; modular quantum computing; statistical validation.

1 Introduction

Quantum error correction turns unreliable physical operations into a protected logical representation, provided that the noise, code structure, and recovery procedure satisfy suitable conditions. Shor established a concrete method for protecting quantum memory against decoherence [1]. Steane showed how quantum error correction can exploit classical coding structure [2], while Calderbank and Shor established the existence of good quantum codes [3]. Knill and Laflamme gave a rigorous definition of the circumstances under which an error set is correctable [4]. These results make an essential distinction between the probability of an individual physical fault and the likelihood of a fault that does not get corrected. This is particularly significant if the device has a loud interface.

The topological construction of Kitaev is a relationship between protection and support of nontrivial error operators. [5]. Dennis et al. related surface-code recovery to error chains in spacetime [6]. Therefore, knowledge of fault type is not just as important as knowledge of fault marginal probability. A narrow The sum of the total of a noisy strip, a uniformly degraded array, and a short temporal burst can be the same. presented various decoding problems, with the expected number of faults.

Modular architectures make this issue practical. A local processor can operate with relatively reliable gates while a communication link introduces larger or more variable errors. Surface-code implementation principles were developed in detail by Fowler et al. [7]. Lattice surgery provides a route to logical interactions through joint boundary measurements, as introduced by Horsman et al. [8] and developed into resource-oriented layouts by Litinski [9]. The attraction of such architectures does not remove the need to specify which checks cross the interface, how link faults propagate, or how the decoder represents them.

This paper asks a narrow, directly answerable question: does matching the expected number of physical faults between a heterogeneous and a homogeneous noise model preserve the logical failure probability of a rotated surface-code memory? We examine an existing two-distance proxy benchmark and construct a separate comparison in which a geometric seam and a homogeneous model have exactly the same expected fault count within each modified channel class. Both are decoded with noise-matched minimum-weight perfect matching. The geometric samples are also decoded with baseline weights, enabling a paired assessment of decoder mismatch.

Scope boundaries. This is a numerical methodology and sensitivity study. The surface code, matching decoder, and stabilizer simulation are established components. We do not

implement a distributed gate, a Bell-pair channel, or classical feedforward. The geometric model G is a synthetic localized-noise circuit, not a calibrated teleportation implementation. All claims are confined to rotated surface-code memory experiments under Pauli noise with minimum-weight perfect matching.

The contributions are fourfold. First, we provide an implementation-level interpretation of the source benchmark, retaining its sampled counts while correcting the meaning of its model labels and uncertainties. Second, we introduce a reproducible, channel-wise mean-matching protocol that separates noise location from mean exposure. Third, we quantify basis-dependent geometry effects, decoder sensitivity, and temporal exposure over 72 new conditions. Fourth, we place these observations against established distributed-code studies without treating incompatible memory, merge, gate-fidelity, and threshold metrics as interchangeable. The work is a numerical methodology and sensitivity study; the seam-rate formulas, surface code, simulator, and matching decoder are established components.

2 Critical reviews of related work

2.1 Teleportation and the physical meaning of an interface

Bennett et al. introduced quantum teleportation using shared entanglement and classical communication [10]. Gottesman and Chuang showed how teleportation can support quantum computation [11]. Eisert et al. analyzed the local resources required to implement nonlocal gates [12]. Together, these works establish why a remote interaction must be represented by a protocol with resource states, measurements, and feedforward, rather than solely by increasing a gate-error parameter. An effective noise channel can approximate such a protocol only after its propagation rules and approximation conditions are specified.

Entanglement purification changes that effective channel and incurs additional resources. Bennett et al. showed how noisy pairs can be purified for more faithful teleportation [13]. Deutsch et al. developed a related purification approach using bilateral operations [14]. Dür and Briegel considered purification in the presence of imperfect computational operations [15]. Krastanov et al. optimized purification protocols and emphasized that protocol choice affects performance and resource use [16]. A memory experiment with a fixed seam probability cannot capture the yield, latency, or extra local operations of these procedures. We therefore do not interpret the severity parameter in this paper as a complete interconnect specification.

2.2 Modular architecture and noisy network links

When building blocks are interconnected, there are various forms of noisy network links. Jiang et al. have suggested distributed computation via small quantum registers [17]. Monroe et The architecture of al. entailed combination of atomic memories and photonic interconnects [18]. Li and Benjamin studied the three qubit nodes distributed computation [19]. Nickerson It has been shown theoretically by et al. that such strongly noisy network links can coexist with Significantly improved local operations [20]. Later, they added lossy photonic links. Approximately five to fifty qubits in and will be stored in cells [21]. These studies demonstrate the effect of "local gate" The fidelity, link fidelity and production rate of entanglements are a coupled design problem. Fowler et al. have discussed surface-code quantum communication [22] and Li and Benjamin have discussed the use of ad hoc codes [23]. developed a hierarchical surface-code architecture of modules of different size [23].

Nigmatullin et al. studied modules with a minimum of complexity of ion traps [24]. The shared lesson is The overhead associated with the architecture is determined by the partitioning/scheduling of the code. Counting while data and check qubits can be used to determine the cost of a remote operation, Idle time, repeated link attempts and classical control may dominate communication qubits. the additional burden. Ramette et al. considered the interaction of a noisy seam and cleaner bulk, both analytically and numerically [25]. Their model still possesses the low dimensional structure of the interface and reports tolerance to interface noise significantly higher than the bulk noise. Its The conclusion will be different based on the topology of the decoding graph. A fixed-fraction homogeneous-noise approximation changes that topology of exposure and cannot be assumed to preserve the same threshold behavior. This observation directly motivates the spatial controls used here.

2.3 Recent distributed error correction studies

Jacinto et al. linked distributed-memory simulations to end-to-end architecture estimates using measurement teleportation [26]. Their published model jointly specifies gate errors, cycle time, Bell-pair availability, and pair fidelity. Marqversen et al. compared lattice surgery, transversal interfaces, and grow-and-distill strategies across hardware constraints [27]. These studies provide a stronger architectural basis than a single scalar seam fraction, because they assess the communication primitive and its resource requirements together.

Chandra et al. studied the logical XX merge operation in an H-shaped spacetime geometry [28]. Their phenomenological model distinguishes bulk data, bulk syndrome, seam data, and seam syndrome probabilities. The two seam-rate expressions used in the source benchmark are the same expressions given in that study. However, the benchmark does not reproduce its merge geometry, bulk effective probabilities, or boundary-parity success criterion. Direct numerical agreement should therefore not be expected. The present work attributes the expressions to Chandra et al. and treats their reuse as a parameterized stress test.

Schmidt et al. introduced distributed stabilizer-measurement constructions, which allow to trade Bell-pair resources for local space and time [29]. Thakur et al. extended seam scheduling and detector annotation to support arbitrary rotated-code layouts while explicitly considering the issues of hook propagation [30]. Both are preprints (as indicated in the references) and are new ones. Stack et al. studied transversal distributed fault-tolerant operations [31]. The current gap is now exact: A useful seam is one that maintains the connection between circuit scheduling, detector construction and logical observables. Our memory controls deal with the implications of that relationship, without applying the distributed primitives of these studies.

2.4 Interconnects in experiments and encoded demonstrations

Cavity based interfaces between atoms and optical photons were discussed by Reiserer and Rempe [32]. Stephenson et al. showed high rate, high fidelity entanglement in an elementary network [33]. Hensen et al. have shown a Bell test between distant electron spins that has no loophole in it [34] and Pompili et al. have achieved a multi-node network by realizing remote solid-state qubits [35]. These results do not directly provide the physical connectivity and entanglement capabilities and do not directly produce a surface code logical failure probability. In a distributed computation, main et al. has shown teleported gates across an optical link [36]. Magnard et al. established a microwave link between superconducting circuits in separate cryogenic systems [37]. Burkhart et al. studied error-detected state transfer

and entanglement in a superconducting network [38], while Zhong et al. demonstrated deterministic multiqubit entanglement across a network [39]. Their metrics include state fidelity, gate fidelity, heralding, and error detection. Comparing any of those values directly with an unconditional seven-round memory failure probability would mix different operational questions.

Encoded experiments provide a separate benchmark for the meaning of distance scaling. Krinner et al. demonstrated repeated error correction in a distance-three surface code [40]. The 2023 surface-code experiment by Google Quantum AI measured an improvement from distance three to distance five under a detailed device noise model [41]. Its later below-threshold experiment demonstrated stronger distance-dependent suppression [42]. Bluvstein et al. implemented logical processing with reconfigurable atom arrays [43]. These experiments show why a claim of fault tolerance requires more than monotonic performance in one synthetic noise sweep. We use them to define evidential standards, rather than to claim superior simulated performance over hardware.

2.5 Simulation and decoding under nonuniform noise

Aaronson and Gottesman developed efficient stabilizer-circuit simulation methods [44]. Gidney's Stim supports high-throughput sampling and detector-error-model extraction for Clifford circuits with compatible noise [45]. Higgott's PyMatching provides minimum-weight perfect matching for quantum-code decoding [46], and Higgott and Gidney's Sparse Blossom improves the practical processing of sparse detector graphs [47]. These tools enable large Monte Carlo studies, but computational speed does not validate an incorrectly specified channel or observable.

Fowler et al.'s Autotune demonstrated the value of constructing decoding information from detailed circuit faults [48]. Bravyi et al. studied efficient maximum-likelihood decoding methods for surface codes [49], and Delfosse and Nickerson developed an almost-linear-time topological-code decoder [50]. These approaches address different accuracy and complexity trade-offs. We use standard matching throughout, so a difference between two conditions is not confounded by switching decoder families. Matching on decomposed detector errors is nevertheless not a claim of exact maximum-likelihood decoding for correlated circuit faults.

A significant improvement in code performance was demonstrated by Tuckett et al. by taking advantage of noise bias [51]. The XZZX construction by Bonilla Ataides et al. also showed the need to consider together noise structure and code structure [52]. The researchers of Chen et al. investigated calibrated decoders for experimental error correction [53]. Wagner et al. set up conditions to estimate Pauli channels using syndrome information [54]. Strikis et al. investigated the scalability of arrays with fabrication defects [55] and Auger et al. examined an architecture for a fault-tolerant Rydberg-atom array [56]. These works justify spatially resolved models, but are not intended to suggest that a specific seam geometry is always a good thing. Positioning note. None of the above cited studies would do the following experiment: a channel-wise mean-matched control for a rotated surface-code memory using a geometrical seam. The closest methodological comparison is Ramette et al. [25], who preserve the lower-dimensional structure of the interface in a spacetime decoding model. Our construction differs in three ways: (i) we match expected exposure within each channel class rather than preserving the interface shape, (ii) we apply the comparison to a memory circuit rather than a merge, and (iii) we pair the noise-placement experiment with a decoder-

mismatch experiment on shared samples. The present work is therefore a methodology paper, not a competing threshold study.

2.6 Review synthesis and comparison boundaries

The literature supports three distinctions. A noisy interface is a geometric object, a remote gate is an operational protocol, and a decoder is a statistical model of the resulting detector events. Replacing all three by a single effective error rate can obscure the cause of a performance change. Table 1 summarizes the nearest methodological comparisons. The review is a focused technical synthesis of foundational and directly relevant studies, including work available by 4 October 2026; it is not a systematic-review prevalence estimate or a ranking of all published architectures.

Table 1. Closest methodological comparisons and the scope of their conclusions

Study	Main construction	Evidence or metric	Relation to this study
Horsman et al. [8]	Merge and split of surface-code patches	Logical operations by joint checks	Defines a gate protocol absent from memory proxies
Nickerson et al. [21]	Purified network resource states	Network noise and link-loss tolerance	Includes communication resources
Li and Benjamin [23]	Hierarchical modular encoding	Module-size and overhead trade-offs	Requires explicit partitioning
Ramette et al. [25]	Lower-dimensional seam and bulk	Analytical and numerical threshold behavior	Preserves spatial interface support
Chen et al. [53]	Experimentally calibrated matching	Decoder performance from measured events	Motivates calibration comparison
Jacinto et al. [26]	Measurement teleportation	Network and algorithm resources	Includes pair rate and fidelity
Marqversen et al. [27]	Comparison of interface protocols	Target logical error and hardware costs	Addresses architecture selection
Chandra et al. [28]	Distributed XX merge	H-shaped graph and effective channels	Source of the seam-rate formulas
Schmidt et al. [29]	Distributed stabilizer measurements	Circuit noise and resource trade-offs	Explicit distributed primitives
Thakur et al. [30]	Seam scheduling and annotation	Hook avoidance and detector validity	Explicit intermodule circuit layout
Present study	Rotated-code memory controls	Counts, intervals, paired decoding	Separates spatial and temporal exposure

3 Core theory and model definitions

3.1 Rotated surface-code memory

A distance-d rotated patch encodes one logical qubit using d^2 data qubits. With one ancilla per stabilizer, the elementary patch contains $d^2 - 1$ check ancillas and

$$N_{\mathrm{patch}}(d) = 2d^2 - 1. \tag{1}$$

Boundary checks have weight two and bulk checks have weight four. The X-type and Z-type stabilizers commute, and a Pauli fault produces a syndrome according to its anticommutation

with those checks. Logical operators commute with the stabilizers but act nontrivially on the encoded qubit [6]. An X-basis memory prepares and reads out a logical X eigenstate, so failure is an incorrect logical X measurement prediction, commonly associated with an effective logical Z fault. A Z-basis memory tests the complementary observable. Neither experiment measures a full logical-gate process matrix.

Repeated syndrome extraction distinguishes persistent data faults from transient measurement faults. For a repeated check with binary outcomes $m_{a,t}$, an interior detector has the form

$$D_{a,t} = m_{a,t} \oplus m_{a,t-1}. \tag{2}$$

Initial and final detectors require boundary-specific expressions. The generated circuits provide these detectors and one logical observable. We preserve that structure when modifying probabilities, rather than manually constructing a graph unrelated to the sampled circuit.

3.2 Pauli channel conventions

The single-qubit depolarizing channel used by Stim assigns total nonidentity probability p:

$$\mathcal{D}_{1,p}(\rho) = (1-p)\rho + \frac{p}{3}(X\rho X + Y\rho Y + Z\rho Z). \tag{3}$$

Its bit-flip marginal and phase-flip marginal are each $2p/3$, because Y contributes to both. A two-qubit depolarizing channel distributes total probability p over the fifteen nonidentity two-qubit Paulis. The probability that it anticommutes with a specified nonidentity single-qubit Pauli on one participant is $8p/15$. Therefore, a probability derived for an effective phase-error process is not automatically a total depolarizing probability of the same magnitude.

Independent error locations also compose nonlinearly. Two binary flips with probabilities p and q have odd-parity probability $p + q - 2pq$. Two single-qubit depolarizing channels combine as

$$\mathcal{D}_{1,p} \circ \mathcal{D}_{1,q} = \mathcal{D}_{1,p+q-4pq/3}. \tag{4}$$

These identities explain why repeatedly appending a noise channel is different from replacing one channel probability. They also explain why the number and timing of appended instructions must be reported.

3.3 Inherited seam parametrization and its domain

The severity sweep uses the parameter pair reported by Chandra et al. [28]:

$$p_s = \frac{\varepsilon}{2}(k+9), \qquad q_s = \varepsilon(k+7). \tag{5}$$

In that work, the corresponding effective bulk values are $p_b = 3\varepsilon$ and $q_b = 4\varepsilon$, under a first-order mapping from gate, readout, idle, and Bell-pair faults. The assumptions include a Bell-pair error scale $k\varepsilon$ and a particular distributed merge schedule. Those assumptions are part of the model, rather than universal properties of a surface-code seam.

Here, $\varepsilon = 10^{-3}$. At $k = 1$, the two values are 0.005 and 0.008; at $k = 11$, they are 0.010 and 0.018. We retain this numerical parametrization to connect the source benchmark and the validation study. In the validation circuit, p_s explicitly denotes a total data depolarization probability at a selected round-boundary location, and q_s denotes a syndrome readout-flip probability. This convention defines a hybrid circuit stress test. It is not a rederivation of the

effective phase-error channel in the merge study. In particular, $k = 1$ is not a no-seam baseline.

3.4 Source proxies and a geometric control

The source benchmark has a standard generated-memory baseline B and two modified circuits. We call the localized instruction-injection circuit A and the homogeneous effective-noise circuit H. These neutral labels avoid attributing distributed operations that the circuits do not contain. Model H replaces all four generator noise parameters by

$$p_{\mathrm{eff}} = \varepsilon + f\left(\frac{p_s+q_s}{2} - \varepsilon\right) = \varepsilon\left[1 + f\left(\frac{3k}{4} + \frac{19}{4}\right)\right], \quad f = 0.3. \tag{6}$$

This yields $p_{\mathrm{eff}} = 0.00265$ at $k = 1$ and 0.00490 at $k = 11$. It is a well-defined uniformly noisy memory circuit, but the average mixes two different channel roles and has no explicit spatial seam.

The new geometric model G selects the central data column $x = d$ and the adjacent check columns $|x - d| = 1$ in the generated lattice coordinates. It replaces the start-of-round data noise with p_s on the d selected data qubits and the immediately premeasurement syndrome-flip probability with q_s on the $2d$ selected ancillas. All other channels retain ε. The selected fractions are

$$f_D = \frac{1}{d}, \qquad f_A = \frac{2d}{d^2-1}. \tag{7}$$

At small distance, this is a substantial fraction of the check ancillas: $f_A = 0.75$ at $d = 3$. At larger distance, both fractions decrease as the interface becomes relatively narrow. This finite-size effect is one reason not to infer asymptotic thresholds from the smallest patches.

The mean-matched control M distributes the same expected number of modified-channel faults uniformly across the appropriate qubit class:

$$\overline{p}_D = \varepsilon + f_D(p_s - \varepsilon), \qquad \overline{q}_A = \varepsilon + f_A(q_s - \varepsilon). \tag{8}$$

G and M thus satisfy separate data-channel and syndrome-channel exposure identities:

$$\sum_{i\in D} p_i = d^2\overline{p}_D, \qquad \sum_{a\in A} q_a = (d^2 - 1)\overline{q}_A. \tag{9}$$

Equation (9) matches expected nonidentity events at the modified locations. It does not match the distribution of error chains, their higher moments, or the logical-failure distribution. Figure 1 shows the spatial supports; Figure 2 summarizes the parameterization and changing fractions.

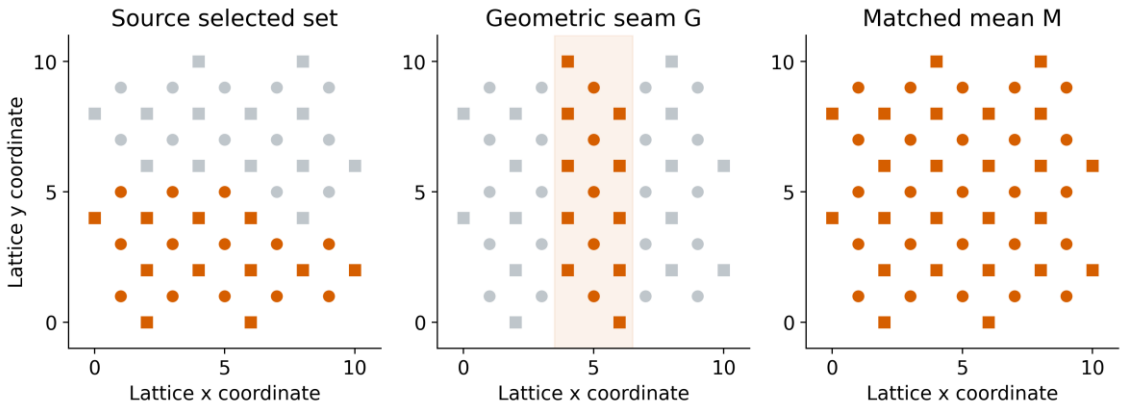


Figure 1. Spatial supports obtained from the distance-five circuit coordinates. Circles denote data qubits and squares denote check ancillas; orange denotes selected or averaged support. The source injection set consists of the first 24 active identifiers. G uses a central data column and adjacent check columns. M distributes the same class-specific expected exposure over each full qubit class.

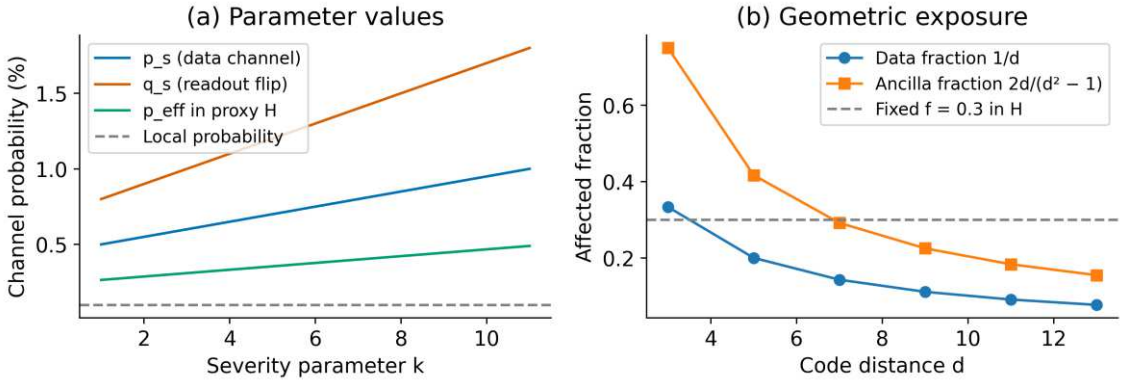


Figure 2. Probabilities of noise and geometric fractions. Panel a separates the seam parameters from the homogeneous proxy probability. The reason for the different distance dependence in both panel b is also shown. 3.5 Why averaging is not a logical way to determine risk.

3.5 Why averaging cannot determine logical risk

It is not necessary to run a surface-code simulation to see the dependence on spatial structure. Let's take a three-bit repetition code of probabilities p_1,p_2,p_3 of bit flips and majority decoding. It has a definite probability of failure, which is

$$P_{\mathrm{rep}} = p_1p_2 + p_1p_3 + p_2p_3 - 2p_1p_2p_3. \tag{10}$$

Compare the heterogeneous vector $(r+\delta, r, r)$ with the uniform vector $(r+\delta/3, r+\delta/3, r+\delta/3)$. The mean and expected number of physical flips are identical. Substituting into (10) gives

$$P_{\mathrm{hetero}} - P_{\mathrm{uniform}} = \frac{\delta^2}{27}(2\delta + 18r - 9). \tag{11}$$

For $0 \le r < 1/2$ and $0 < \delta \le 1/2 - r$, this difference is negative. Thus, even a simple error-correcting code supplies a counterexample to the assumption that the mean error probability determines logical risk. The calculation is an explanatory example, not a new threshold theorem. A surface code contains many overlapping paths and boundaries, so changing where the elevated probabilities lie can also change the sign of the effect for different logical observables. Figure 3 illustrates the repetition-code counterexample showing that equal mean physical exposure does not imply equal logical failure probability.

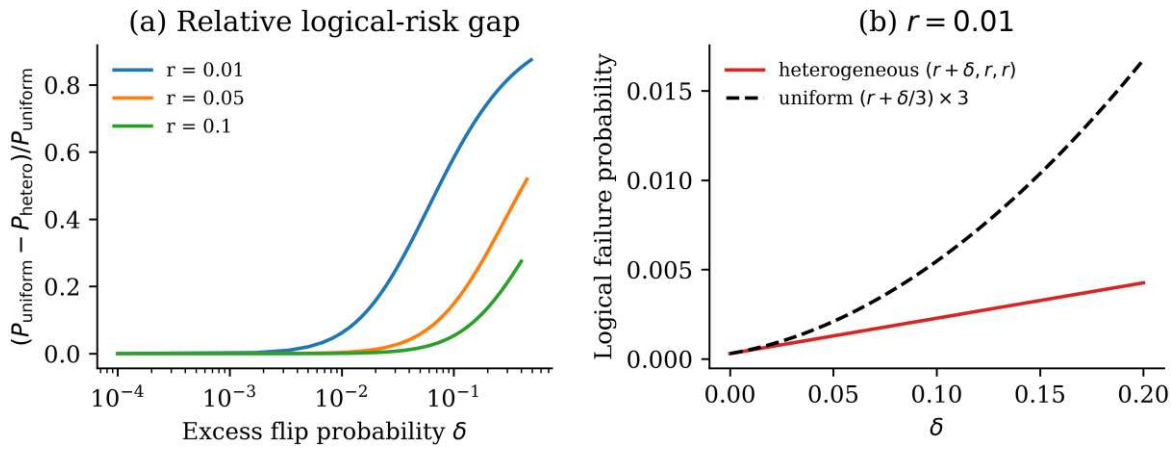


Figure 3. Three-bit repetition code with majority decoding, using the notation of Equation (11). The heterogeneous vector (r + δ, r, r) is compared with the uniform vector (r + δ/3, r + δ/3, r + δ/3), which has the same mean and the same expected number of physical flips. Panel a shows the relative reduction in logical failure probability of the heterogeneous vector for three values of r; panel b shows both logical failure probabilities at r = 0.01. Equal expected exposure does not imply equal logical risk.

3.6 Detector models and matching weights

A detector error model describes fault mechanisms by their detector flips and logical-observable flips. In binary notation, a fault vector e gives

$$s = He \pmod 2, \qquad \mathrm{l} = i \pmod 2. \tag{12}$$

For graphlike independent mechanisms, matching uses log-likelihood weights of the form

$$w_j = \log\frac{1-p_j}{p_j}. \tag{13}$$

Each condition is decoded using a detector error model extracted from its own circuit, with decomposition enabled. The matching prediction is compared with the sampled logical observable. Standard matching does not exploit every correlation preserved in a decomposed circuit model, so it should not be described as exact maximum-likelihood decoding. The mismatch experiment changes only the decoder model: the same G samples are decoded with weights obtained from the baseline circuit. This isolates a practical consequence of ignoring the seam during decoder construction.

4 Experimental and statistical methods

4.1 Source benchmark and implementation audit

The source archive contains the simulator, summary, figures, and aggregate result file. It reports five trials of 20,000 shots for each of 26 conditions: two baseline distances and two proxy models at six severity values for each distance. Each listed mean is the integer failure count divided by 100,000 to the precision of a floating-point number. Trial-level counts were not archived, and nor were sampler seeds and software versions, but these counts support a reanalysis.

The reanalysis identifies six specific implementation issues; Figure 4 summarizes the audit and Table 2 documents the resulting model interpretations.

(i) Undocumented seam fraction. Model A chooses qubits in order, without any kind of geometric seam rule. Its caller doesn't supply the seam-fraction argument, and therefore the default value of 0.5 is hardcoded in the builder instead of the value in the configuration (which is 0.3). It selects 8 and 24 qubits with 17 and 49 active qubits, respectively.

(ii) Temporal concentration outside the repeated body. The injection loop visits top-level instructions but does not go into the repeated body with the remaining six extraction rounds. This additional noise is consequently concentrated at its beginning and end, not spread evenly across the seven rounds.

(iii) X error after measurement. The X error which comes after a measurement does not change its outcome. The source's post-measurement injection is therefore partially ineffective.

(iv) Unused p_ebit field. The source's separate p_ebit parameter is stored in the configuration but is not consumed by either circuit builder. The variables named rounds_before and rounds_after are only summed to seven; no remote CNOT is inserted between them.

(v) Algebraically generated decay curve. A displayed decay curve was generated algebraically from endpoint values rather than from measurements at successive times.

(vi) Contradictory resource annotation. The source resource figure mixed single-patch and three-patch counts and associated the computed 339-qubit bar with a 373-versus-521 annotation. No matching circuit layout or reliability target accompanies that annotation.

The decay curve, unsupported published-reference points, and the resource annotation are excluded from the numerical evidence in this paper; the sampled failure counts are retained and reinterpreted.

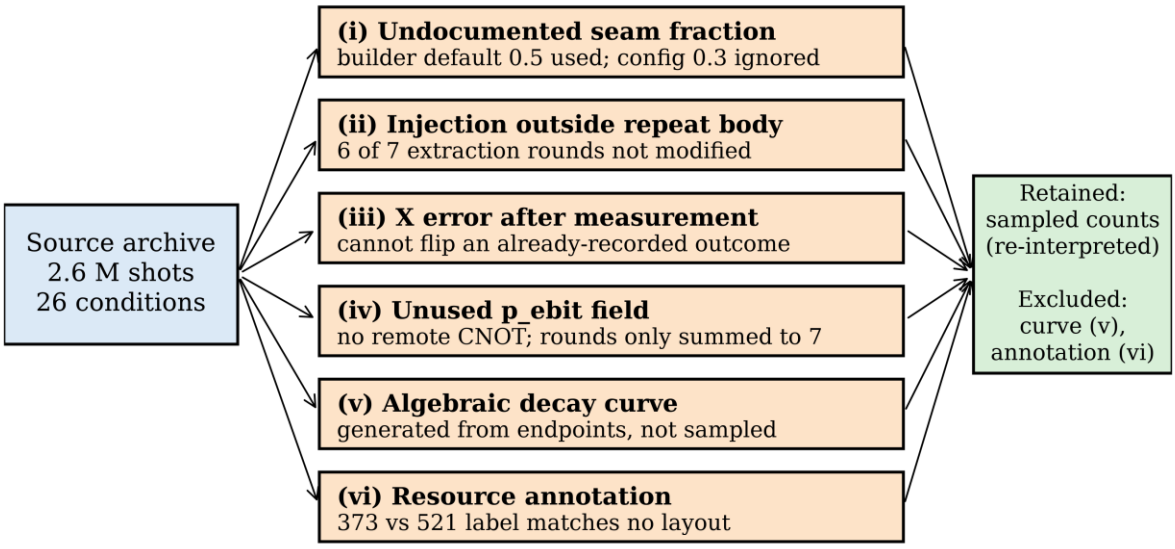


Figure 4. Implementation audit of the source benchmark. Six issues were identified; sampled failure counts were retained and reinterpreted as elevated-noise memory results, whereas the algebraic decay curve (v) and the resource annotation (vi) were excluded from the evidence.

Table 2. Implemented models and the interpretation supported by their circuits

Model	Implemented exposure	Interpretation
B	Generated memory; four local noise classes at 0.001	Reference memory; one logical observable
A	First 8 or 24 active identifiers; extra instructions outside repeated body; post-measurement X error ineffective	Boundary-concentrated injection proxy; no remote gate; six implementation issues documented in Section 4.1
H	All generator classes at p_eff; f = 0.3	Homogeneous circuit-noise scaling; no explicit seam
G	Central data column and adjacent check columns; replacements every active round	Spatial circuit stress test with a geometric seam
M	Uniform class-specific averages matched to G	Control for expected physical exposure
Decoder mismatch	G samples decoded using B detector weights	Paired test of assumed noise model
Time ablation	Elevated G channels active in first 1, 4, or 7 rounds	Duration sensitivity at fixed total time

4.2 New controlled experiments

The principal validation grid contains 56 conditions: two bases, four distances, and, for each combination, one baseline plus G and M at $k \in \{1,5,11\}$. All use seven extraction rounds. Four additional G conditions at distance five and $k = 11$ activate elevated seam probabilities for the first one or first four rounds, in both bases. Twelve further conditions set $T = d$ for distances three, five, and nine at $k = 11$, using both models and bases. The distance-seven cases are already in the principal grid and don't need to be counted again. The total is 72 conditions and 14,400,000 shots.

New condition has 40,000 shots for each seeded batch. This seed is 202610040000+10i+j, where i is the saved condition index and j is the batch index, which is the number of the batch from 0 to 4. The recorded environment is Python 3.12.14, Stim 1.16.0, PyMatching 2.4.0, and NumPy 2.3.5. The paper is accompanied by Circuit files, Circuit hashes, Aggregate and Batch counts, and Exact parameter choices. Seeds help repeatability in a single environment but there is no assumption of cross-version and cross-platform bitwise identity.

The baseline extraction circuit, Clifford schedule, detector definitions, and logical observable remain as they were. The circuit is flattened first and then probabilities are changed, ensuring that every intended round is visited. Data channels and premeasurement channels will be replaced, not by blindly adding additional faults. Data readout after a reset, faults after reset and the probability of gate noise are kept at their baseline values. In 8000 shots sampled, there were no detector events or flips from a noiseless check performed over eight combinations of distance and basis. The generated G and M circuits are also tested directly in equation (9).

The number of failures, NOT shots, is used to calculate the statistical power of each comparison. Table 3 contains the number of failures recorded for each of the cells of the principal grid. As distance increases, fewer failures are recorded in a number of cells and these are presented as descriptive trends, not as decisive results, at distances 7 and 9. Table 4 summarizes the simulation settings and reproducibility information for the source benchmark and the new validation experiments.

Table 3. Failures per cell for the principal seven-round grid (each cell = 200,000 shots). Baseline B does not depend on k.

d	Basis	Model	k = 1	k = 5	k = 11
3	X	Baseline B	340	—	—
3	X	Geometric G	660	780	977
3	X	Matched M	780	1090	1430
3	Z	Baseline B	362	—	—
3	Z	Geometric G	780	1010	1540
3	Z	Matched M	720	950	1392
5	X	Baseline B	22	—	—
5	X	Geometric G	24	60	102
5	X	Matched M	40	80	150
5	Z	Baseline B	35	—	—
5	Z	Geometric G	35	90	192
5	Z	Matched M	30	70	141
7	X	Geometric G	3	6	9
7	X	Matched M	5	8	11
7	Z	Geometric G	6	15	24
7	Z	Matched M	8	20	37
9	X	Geometric G	0	1	2
9	X	Matched M	1	2	3
9	Z	Geometric G	1	2	4
9	Z	Matched M	2	3	6

Table 4. Simulation settings and reproducibility information

Quantity	Source benchmark	New validation
Code and observable	Rotated X memory	Rotated X and Z memories
Distances	3 and 5	3, 5, 7, 9
Severity k	1, 3, 5, 7, 9, 11	1, 5, 11; ablations at 11
Local probability	0.001	0.001
Rounds	7	7; additional T = d control
Shots per condition	100000	200000
Batches	5 × 20000	5 × 40000
Distinct conditions	26	72
Total sampled shots	2600000	14400000
Sampler seeds	Not archived	Saved for every batch
Decoder	Noise-matched MWPM	Noise-matched and baseline-model MWPM
Interval reported here	Pooled Wilson 95%	Pooled Wilson 95%
Versions	Not archived	Stim 1.16.0; PyMatching 2.4.0

4.3 Probability estimates and uncertainty

For E failures in N independent sampled shots, we estimate the complete-experiment failure probability by

$$\hat{P}_L = E/N. \tag{14}$$

We report two-sided 95% Wilson intervals [57], consistent with the small-probability concerns discussed by Brown et al. [58]. With $z = 1.959964$, the interval center and half-width are

$$c = \frac{\hat{P}_L + z^2/(2N)}{1 + z^2/N}, \qquad h = \frac{z\sqrt{\hat{P}_L(1-\hat{P}_L)/N + z^2/(4N^2)}}{1 + z^2/N}. \tag{15}$$

The interval is $[c - h, c + h]$, restricted to $[0,1]$. For zero failures in 200,000 shots, its upper endpoint is approximately 1.92×10^{-5}. Zero observed failures therefore do not imply a zero physical failure probability. Sampling intervals are conditional on the specified simulator and noise model; they do not account for uncertainty in the physical validity of that model.

For nonzero independent counts, the displayed risk-ratio interval uses the log approximation

$$R = \frac{E_1/N_1}{E_0/N_0}, \qquad \mathrm{SE}(\log R) = \sqrt{\frac{1}{E_1} - \frac{1}{N_1} + \frac{1}{E_0} - \frac{1}{N_0}}. \tag{16}$$

Ratios involving zero counts are left undefined. Intervals for very small nonzero counts are treated as descriptive, and the principal ratio figure is restricted to distances three and five. These pointwise intervals are distinct from the familywise-adjusted hypothesis tests.

4.4 Hypothesis tests and exposure metrics

We use two-sided Fisher exact tests for the 24 independent G-versus-M comparisons in the seven-round grid. Holm's procedure controls the familywise error rate across that family [59]. Decoder mismatch uses the same shots for both predictions. Its relevant counts are the discordant outcomes: matched success with mismatched failure, and matched failure with mismatched success. We apply the exact binomial version of McNemar's test [60] and a separate Holm adjustment across the 24 G decoder comparisons. No optional stopping is used; each condition has the fixed shot count stated above.

The source's significance function uses a standard-error expression associated with unequal variances but assigns $2(n-1)$ degrees of freedom. That is not the Welch–Satterthwaite calculation [61]. Its five-trial bootstrap intervals also estimate a different sampling object from the pooled binomial interval. We retain the source standard deviations only in the archived data and do not label them as confidence intervals or reuse its significance stars.

Distance benefit is summarized descriptively by $\Lambda_{3,5} = P_L(d=3)/P_L(d=5)$ at the same model and exposure. The reported failure probabilities concern complete memory experiments. Under a separate stationary independent logical-flip approximation, one could write

$$P_L(T) = \frac{1-(1-2r_L)^T}{2}. \tag{17}$$

We do not use (17) to convert the observations to per-round rates, because boundaries, nonstationary injections, and decoder history can violate its assumptions. This restriction is especially important for source model A and the temporal ablation.

Power note. For a two-sided log-risk-ratio test at 200,000 shots per cell, 80% power, and α = 0.05/24, the minimum detectable risk ratio (MDRR) depends on the failure count of the comparator cell: approximately 0.59 at 150 failures, 0.77 at 500 failures, and 0.86 at 1430 failures (Figure 5). Cells with fewer than 100 failures (all d = 7 and d = 9 cells, and several d = 5 cells) cannot resolve effects of the size observed here and are flagged as under-powered for ratio-based inference. This supports the descriptive-trend framing applied to those cells in Section 5.

Minimum detectable risk ratio. For the two-sided log-risk-ratio test of Equation (16) with significance level α and power $1 - \beta$, let E_0 be the expected failure count of the comparator cell and $R < 1$ the true ratio, so that the other cell expects RE_0 failures. Neglecting the small $1/N$ terms, the minimum detectable risk ratio R_{MDRR} solves

$$|\ln R_{\mathrm{MDRR}}| = (z_{1-\alpha/2} + z_{1-\beta})\,[(1 + 1/R_{\mathrm{MDRR}})/E_0]^{1/2}. \qquad (18)$$

Here z_q is the standard normal quantile. Equation (18) is solved numerically; when R_{MDRR} is close to unity it reduces to the familiar approximation $R_{\mathrm{MDRR}} \approx \exp[-(z_{1-\alpha/2} + z_{1-\beta})(2/E_0)^{1/2}]$. Figure 5 plots the exact solution of (18) at α = 0.05/24 and 80% power.

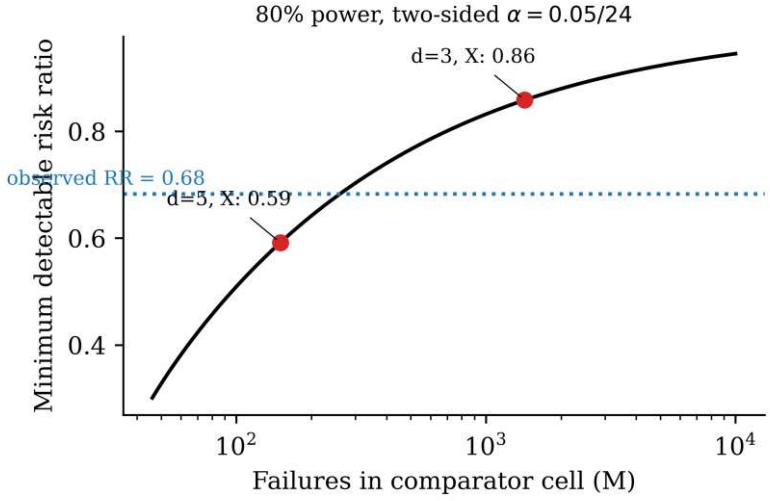


Figure 5. Minimum detectable risk ratio from Equation (18) as a function of the comparator-cell failure count at 80% power and two-sided α = 0.05/24. Red points mark the matched-control counts of the two X-basis k = 11 comparisons; the dotted line is the observed ratio 0.68.

5 Results

5.1 Reanalysis of the source observations

The baseline contains 169 failures at distance three and 22 at distance five, each in 100,000 shots. The corresponding probabilities are 0.001690 and 0.000220. Their ratio is 7.68, with an approximate 95% interval of 4.93 to 11.98. The wide interval is driven primarily by the 22-failure denominator. Thus, a large degradation factor relative to the distance-five baseline should not be presented with the apparent precision of its unrounded quotient.

Both source proxies increase monotonically across the six recorded severity values. Model A rises from 0.01531 to 0.04055 at distance three and from 0.00835 to 0.03574 at distance five. Model H rises from 0.01315 to 0.03989 and from 0.00365 to 0.02208, respectively. Table 5 reports all original proxy counts through their probabilities, and Figure 6 adds recomputed

Wilson intervals. These are genuine sampled outcomes of the implemented circuits. Their interpretation is elevated-noise memory performance, rather than remote-gate failure.

Table 5. Complete source proxy results with Wilson 95% intervals, in units of 10^{-3}

d	k	Proxy A probability [95% CI]	Proxy H probability [95% CI]
3	1	15.310 [14.567, 16.090]	13.150 [12.462, 13.875]
3	3	19.160 [18.329, 20.028]	17.780 [16.979, 18.618]
3	5	24.420 [23.481, 25.395]	23.020 [22.109, 23.968]
3	7	28.740 [27.722, 29.794]	27.680 [26.681, 28.715]
3	9	35.540 [34.410, 36.705]	34.360 [33.249, 35.507]
3	11	40.550 [39.345, 41.790]	39.890 [38.695, 41.121]
5	1	8.350 [7.805, 8.933]	3.650 [3.295, 4.043]
5	3	11.730 [11.081, 12.416]	6.110 [5.646, 6.612]
5	5	15.980 [15.221, 16.776]	8.950 [8.385, 9.553]
5	7	22.170 [21.276, 23.101]	12.330 [11.665, 13.033]
5	9	27.920 [26.917, 28.959]	17.110 [16.325, 17.933]
5	11	35.740 [34.607, 36.909]	22.080 [21.187, 23.009]

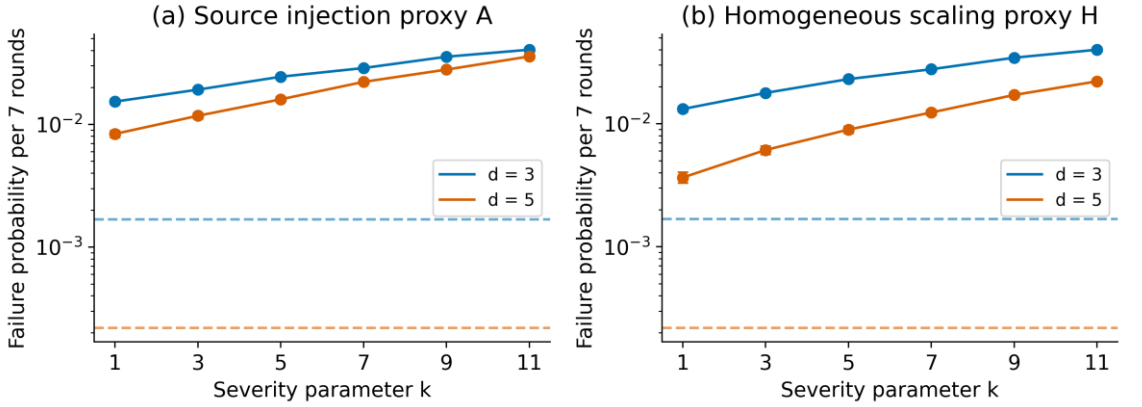


Figure 6. Reanalyzed source observations. All points are 100,000 shots, two-sided 95% Wilson intervals are shown as error bars. The corresponding baseline probabilities are indicated by dashed horizontal lines. A is a different spatial/temporal exposure than H, and is not a calibrated version of the same channel.

The ratio of A to distance-three baseline is about 24.0 and the ratio of A to distance-five baseline is about 162.5 at k=11. For H, the factors are 23.6 and 100.4. The large relative failure probability for distance-five doesn't imply that distance five has the large absolute failure probability. It has a much lower level. This is an example of why benefit from absolute probabilities, ratios and distance benefits should be reported together.

5.2 Cater to the diminished apparent advantage

This source baseline distance advantage is reduced to 1.83 at k=1 and 1.13 at k=11 for A. Over this range, it drops from 1.81 to 0.11 for H. This decrease is from the relative increase above baseline (Figure 7). In A the exposed set takes about half of the active qubits and the active qubits are highly localized at circuit edges. All generator noise channels are boosted in H. Both mechanisms can be detrimental to the distance benefit but neither one or the other removes the effect of a fixed-width physical interface.

A two-distance result does not result in a threshold. Rather than sweeping local noise for several distances across a crossing, it is the source that varies the severity at one local error probability. A threshold formula is not a good fit for these observations, because it would collapse finite-size response, the definition of the channel, and the number of rounds (seven). The physical error rate is not therefore considered critical and distance advantage is described.

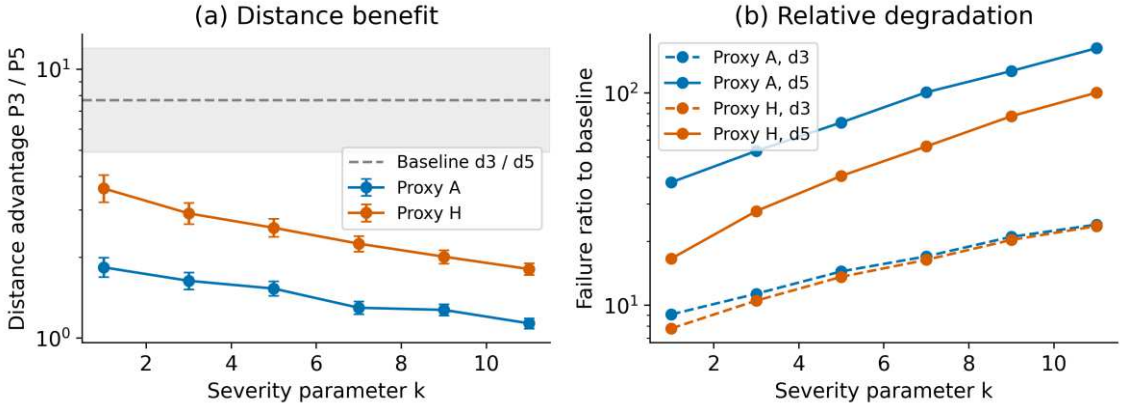


Figure 7. Degradation of source models: interpretation. In Panel a, the distance advantage is reported with approximate pointwise risk-ratio intervals with baseline ratio interval shown in gray band. In each case, point estimates of degradation to respective distance baselines are displayed in panel b. It has a large distance-five ratios thanks to a small baseline denominator.

5.3 Geometry at matched mean exposure

The new G and M models are different, controlled comparison. They both have the same class by class expected exposure at all distances and severity as described in (9). The logical consequences of their actions do not, however, coincide. In Figure 8 the distance is scaled at k=11 and 7 rounds. The probabilities of failure decrease significantly between the distances three and seven in both bases. For many of the cells at distance nine, there have only been one or two observed failures, which doesn't guarantee the precision of the ranking.

The final answer is at distance 3, basis X, and k = 11: G has 977 failures, and M has 1430 failures. The probabilities are 0.004885 and 0.007150. Their risk ratio is 0.683, with a pointwise 95% interval of 0.630 to 0.741; the Holm-adjusted Fisher probability is 4.56×10^{-19}. This single comparison is a valid counterexample to the universal claim that matching expected physical exposure preserves logical risk. A single counterexample is sufficient to refute a universal claim; the larger-distance observations below are consistent with the same effect but are individually under-powered. It does not demonstrate an improvement in the physical device: the models have different spatial distributions by design.

In the Z basis, the largest-severity point estimates instead favor M: 1540 versus 1392 failures at distance three and 192 versus 141 at distance five. The respective ratios are 1.106 and 1.362. Their Holm-adjusted probabilities are 0.116 and 0.115, so these selected opposite-direction differences are descriptive trends rather than familywise-significant findings. Table 6 and Figure 9 present the effect sizes and uncertainties explicitly. The observed basis dependence is consistent with different relationships between the seam and logical error paths, but the present study does not isolate every boundary and schedule contribution.

The Z-basis sign reversal. In the rotated surface code, an X-basis memory fails when an undetected chain of Z-type errors connects one pair of opposite boundaries, whereas a Z-basis

memory fails when a chain of X-type errors connects the orthogonal pair. The geometric seam G is a vertical strip: it is aligned with the logical failure paths of one basis and transverse to those of the other. Elevated noise concentrated on a strip transverse to the failure paths is crossed by every path but occupies only one site per path, so concentrating exposure there can lower logical risk relative to spreading the same exposure uniformly, as in the repetition-code example of Section 3.5. Elevated noise along a strip parallel to the failure paths can instead supply most of the support of a single logical chain, which raises logical risk relative to the uniform control. Under this reading, the sign of the G/M ratio follows the relative orientation of the seam and the logical operator of the measured basis.

This asymmetry between the two logical operator supports is a known feature of rotated surface codes, and the XZZX variant [51, 52] is constructed in part to symmetrize the code's response to biased noise. We do not claim to have isolated every boundary and schedule contribution, and the Z-basis differences are not familywise significant. The orientation argument is therefore stated as a hypothesis whose direct test is a repetition of the experiment with the seam rotated by 90 degrees, which should exchange the signs of the X- and Z-basis ratios.

Table 6. Geometry comparisons at k = 11, with 200,000 shots per model

Basis	d	Failures G / M	Risk ratio G / M [95% CI]	Holm p	Familywise significant
X	3	977 / 1430	0.683 [0.630, 0.741]	4.56×10^{-19}	Yes
X	5	102 / 150	0.680 [0.529, 0.874]	0.0597	No (descriptive)
Z	3	1540 / 1392	1.106 [1.029, 1.189]	0.1156	No (descriptive)
Z	5	192 / 141	1.362 [1.096, 1.692]	0.1148	No (descriptive)

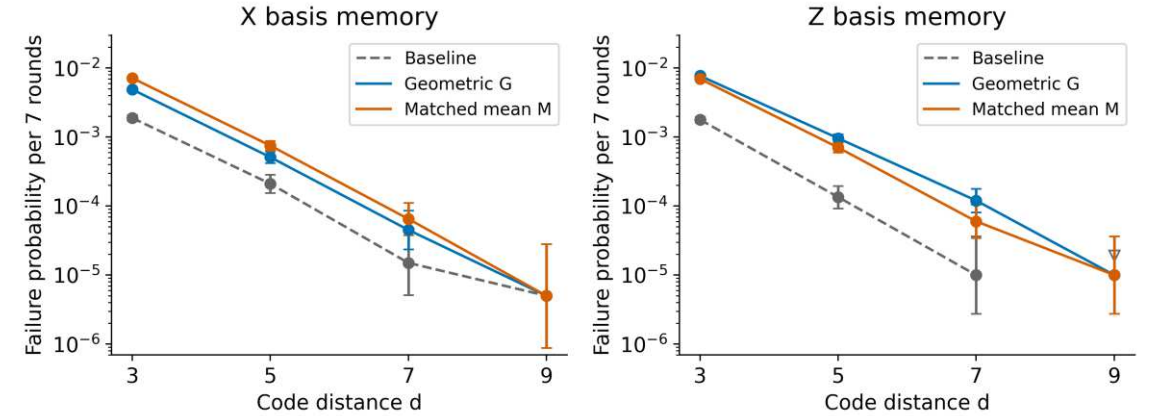


Figure 8. Controlled distance scaling at k equal to 11. All conditions use 200,000 shots and seven rounds. G and M match expected fault counts separately within the modified data and syndrome channels. Open downward triangles mark the upper Wilson endpoint for zero observed failures, not a nonzero point estimate.

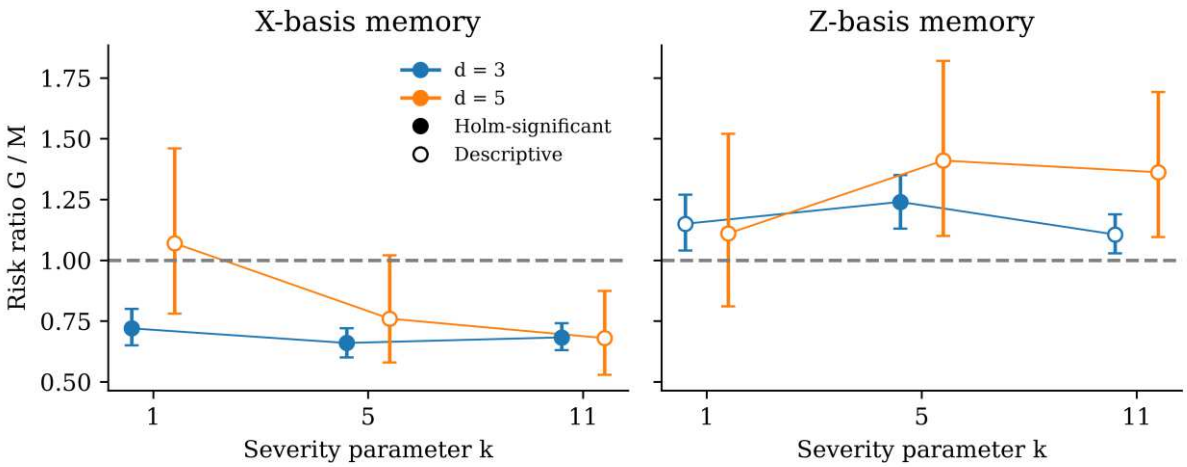


Figure 9. Geometric-to-mean-matched risk ratios at distances three and five. Error bars are approximate pointwise 95% intervals; the reference line is unity. Filled markers indicate points that survive the 24-comparison Holm family; open markers indicate descriptive trends. Because statistical decisions use the familywise-adjusted tests, an interval excluding unity need not correspond to a familywise-significant result.

At distance five in the X basis, the ratio at $k = 11$ is 0.680, but its adjusted probability is 0.0597. It should therefore not be grouped with the distance-three comparison as an equally established effect. The ability to distinguish these cases is a benefit of reporting counts and family definitions rather than only percentage changes. At larger distances, more shots would be required to distinguish modest geometry effects once failures become rare.

5.4 Decoder calibration on shared samples

The decoder comparison is more decisive at low distance. For G in the X basis at $d = 3$ and $k = 11$, the baseline decoder fails on 2060 shots, whereas the matched decoder fails on 977 of the same 200,000 shots. This is a 52.6% reduction in observed failures. There are 1160 shots corrected only by the matched decoder and 77 corrected only by the baseline decoder. The adjusted exact paired probability is approximately 1.73×10^{-247}. Calibration helps on many more shots than it harms, even though it is not uniformly better on every individual syndrome.

At distance five in the X basis, the corresponding counts are 137 and 102, a 25.5% reduction, with adjusted paired probability 6.62×10^{-5}. In the Z basis at distance three, matched decoding reduces failures from 1843 to 1540, with adjusted probability 1.74×10^{-24}. At distance five in that basis, the counts decrease from 221 to 192, but the adjusted probability is 0.0718. Table 7 reports the selected paired comparisons, and Figure 10 includes distance seven to show how the event counts become small.

Table 7. Paired decoder results at k equal to 11, on 200000 shared shots

Basis	d	Failures matched / baseline	Discordant matched-only / baseline-only successes	Holm p
X	3	977 / 2060	1160 / 77	1.73e-247
X	5	102 / 137	46 / 11	6.62e-05
Z	3	1540 / 1843	574 / 271	1.74e-24
Z	5	192 / 221	63 / 34	0.0718

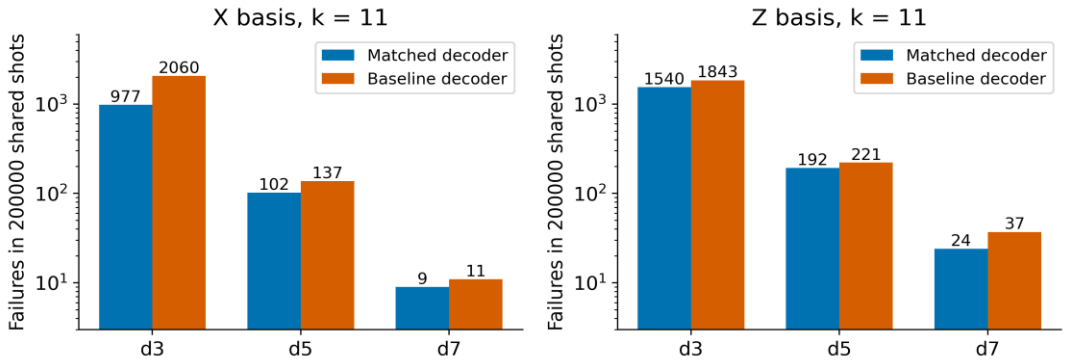


Figure 10. Decoder comparison on identical geometric-noise samples at k equal to 11. Counts are shown for distances three, five, and seven. The log scale accommodates the different event frequencies; significance is evaluated from the paired discordant outcomes, not from independent error bars on these bars.

This result concerns knowledge of the simulated channel. It does not include the experimental cost of estimating that channel, uncertainty in calibration, or time variation. The matched decoder is given the correct circuit model by construction. Practical systems would require an estimation procedure such as those considered by Chen et al. [53] and Wagner et al. [54], followed by validation against held-out syndrome data.

5.5 Temporal exposure and round-count sensitivity

The temporal ablation keeps distance five, seven total rounds, and $k = 11$ fixed. In the X basis, activating the seam for the first one, first four, or all seven rounds produces 50, 72, and 102 failures in 200,000 shots. The Z-basis counts are 35, 122, and 192. Figure 11a reports these observations with Wilson intervals. The results support an exposure-duration effect within this schedule. They do not measure a continuous relaxation curve, and the intermediate points are not generated from an assumed exponential function.

The $T = d$ ablation changes the temporal aspect ratio. Its G failure counts are 399, 63, 9, and 2 in the X basis and 637, 108, 24, and 4 in the Z basis, for distances three, five, seven, and nine, respectively. The distance-seven entries are the same data as in the seven-round grid. Figure 11b shows suppression with distance, while the few-event endpoints again limit quantitative ranking. Since only one local physical error probability is used, even this additional aspect-ratio check does not establish a threshold.

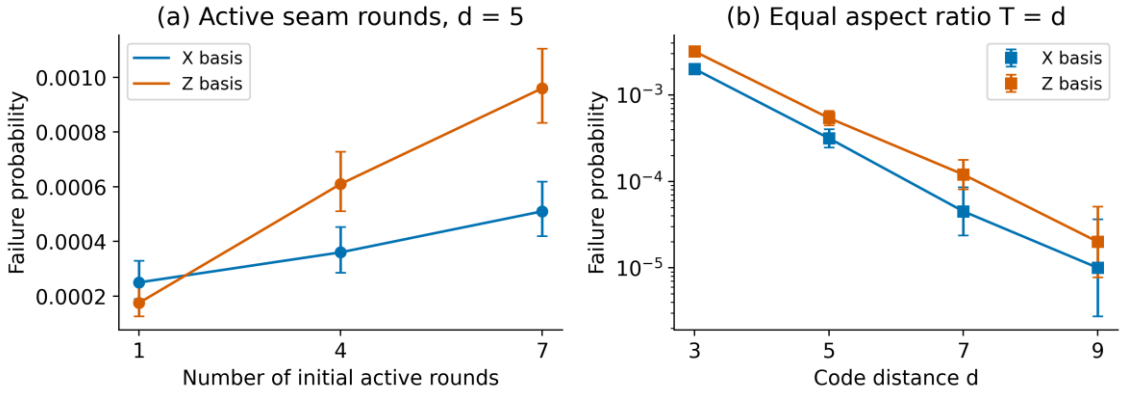


Figure 11. Temporal controls. Panel a varies the number of initial rounds exposed to elevated seam noise while keeping seven total rounds. Panel b uses T equal to d at k equal to 11. Each distinct condition contains 200,000 shots; error bars are Wilson intervals. Distance-seven results are reused from the principal grid.

5.6 Qubit counts and the limits of resource comparisons

For independent rotated patches, the elementary qubit count is the sum of (1). A three-patch arrangement with distances 3, 7, and 3 contains 131 data-and-check qubits. Distances 5, 11, and 5 require 339, while three distance-nine patches require 483. Figure 12 and Table 8 compare like-for-like numbers of patches. The 5–11–5 count is 29.8% below the 9–9–9 count, but that arithmetic is not evidence of an equal-reliability resource saving.

The source resource figure mixed single-patch and three-patch counts and associated the computed 339-qubit bar with a 373-versus-521 annotation. No matching circuit layout or reliability target accompanies that annotation. The corrected comparison therefore uses only the explicit patch-count formula. Communication ancillas, merge bridges, Bell-pair factories, spare resources, and timing overhead are excluded. A hardware resource claim would require all of those quantities and an equal logical-error target.

Table 8. Data-and-check qubit counts for three independent rotated patches

Patch distances	Calculation	Total	Scope
3–3–3	3 × 17	51	Three encoded patches
3–7–3	17 + 97 + 17	131	Unequal patch distances
5–5–5	3 × 49	147	Three encoded patches
7–7–7	3 × 97	291	Three encoded patches
5–11–5	49 + 241 + 49	339	Unequal patch distances
9–9–9	3 × 161	483	Three encoded patches

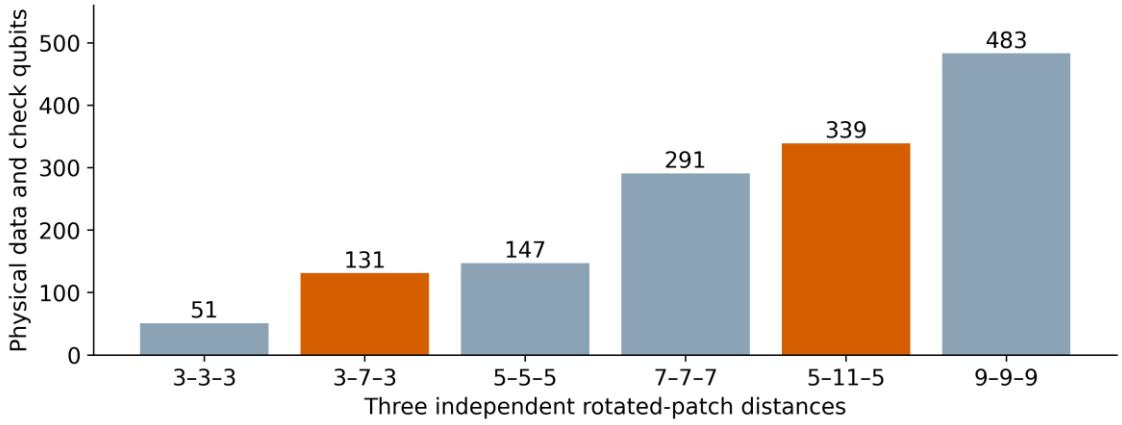


Figure 12. Consistent data-and-check qubit counts for three independent rotated patches. These are geometric counts, not complete distributed architectures or comparisons at equal logical reliability.

6 Comparison with published studies

6.1 Quantitative context with distinct metrics

Table 9 lists selected published numerical results alongside their experimental or modeling conditions. Nickerson et al.'s network studies use purification and explicit local-operation assumptions [20] [21]. Ramette et al. preserve a lower-dimensional noisy interface in a spacetime decoding model [25]. Chandra et al. use a distributed XX merge geometry and effective phase-error probabilities [28]. Their threshold estimates therefore answer questions that the present fixed-noise memory sweep does not answer.

The latest architectural estimates are even more conditional. Jacinto et al.'s published result specifies Bell-pair availability per cycle as well as fidelity [26]. Main et al.'s experimental teleported-gate fidelity is a physical gate metric [36]. The below-threshold surface-code

experiment reports logical error per cycle under a calibrated device model [42]. Our complete-experiment probabilities cannot be subtracted from any of these values to obtain a performance improvement. The comparison instead identifies what additional evidence each stronger claim requires.

Table 9. Published numerical context and the conditions that prevent direct ranking

Study	Reported result	Metric and essential condition	Comparison boundary
Nickerson et al. [20]	Local errors below 0.82% with network noise at least 10%	Protocol-specific tolerance with purification	Not a seven-round memory rate
Nickerson et al. [21]	13.3% network threshold; 98% photon loss example	Purification and entanglement attempts at 2 MHz	Loss and rate modeled explicitly
Ramette et al. [25]	Approximately 1% bulk and 10% interface threshold scales	Spatially resolved seam and bulk	Different noise convention and spacetime graph
Chandra et al. [28]	0.8683% to 0.8339% as k goes from 1 to 11	XX merge; d = 5 to 13; 200000 shots per point	Preprint threshold estimates; different observable
Main et al. [36]	86% teleported CZ fidelity; 71% Grover success	Two trapped-ion modules and optical interconnect	Physical gate and algorithm metrics
Google Quantum AI and Collaborators [42]	0.143% ± 0.003% per cycle at d = 7	101-qubit experimental memory; calibrated hardware	Per-cycle metric under device-specific noise
Jacinto et al. [26]	70 Bell pairs per cycle with fidelity above 98.1%	One published architecture at local error 0.001 and 1 microsecond cycle	Conditional rate-and-fidelity requirement
Present G versus M	X basis, d = 3, k = 11: 0.004885 versus 0.007150	Seven-round memory; equal expected channel exposure	Controlled model comparison, not a gate threshold

6.2 Why the large source degradation does not contradict seam tolerance

Ramette et al.'s conclusion that a seam can tolerate elevated noise relies on its lower-dimensional support [25]. Source proxy H instead raises the error probabilities throughout the entire array. Proxy A changes an identifier-selected subset, uses a different fraction, and does not inject throughout the repeated body. Neither is a faithful implementation of the seam model in that reference. Large degradation in A or H therefore does not falsify the published seam-tolerance result.

The same caution applies to the formulas inherited from Chandra et al. [28]. Their bulk effective probabilities, H-shaped boundaries, and success criterion are essential to their reported merge thresholds. Reusing only p_s and q_s, while retaining a different circuit and converting one parameter to a total depolarizing probability, defines another experiment. The present study makes that distinction explicit and uses a control that preserves mean exposure within its own channel convention.

6.3 What the controlled comparison adds

The numerical contribution is the separation of three mechanisms within one reproducible memory framework. G versus M changes spatial placement while matching channel-wise expected exposure. Matched versus baseline decoding changes the inference model while keeping sampled syndromes fixed. The temporal ablation changes how long the seam is active while preserving the total circuit duration. These controls make the explanation of a logical-error difference more specific than a comparison between two differently parameterized noisy circuits.

The analytical repetition-code example establishes why an average-rate identity is insufficient in principle; the surface-code runs show a concrete consequence in the tested implementation. The study does not introduce a new code or decoder, and geometry-aware modeling is already established in the literature. Its contribution is a transparent comparison protocol and an auditable dataset demonstrating the risks of interpreting an effective-noise proxy as a distributed operation.

7 Discussion and limitations

7.1 Implications for model construction

A seam model should state the geometric selection rule, the affected operation classes, the temporal activation rule, and the meaning of each channel probability. An apparent interface fraction is not enough. At distance three, the selected ancilla fraction in G is much larger than the data fraction; a single common fraction would obscure that distinction. At increasing distance, a fixed-width seam occupies a decreasing portion of the patch, whereas a fixed-fraction model does not. The appropriate scaling depends on the intended physical architecture.

Decoder construction must be reported with equal care. A low logical-error probability under a matched decoder is a joint property of the noisy circuit and its assumed likelihood model. The paired results show that ignoring the elevated seam can impose a substantial penalty at small distance. Nevertheless, exact knowledge of synthetic probabilities is easier to obtain than a stable calibration of an experimental link. Future work should test robustness to estimated probabilities, calibration drift, and correlated faults rather than assuming that a matched decoder is available at no cost.

The reanalysis of the source benchmark yields a methodology that is itself generalizable. The six issue categories (undocumented parameter values, injection outside repeated circuit bodies, ineffective post-measurement operations, unused configuration fields, algebraic curve substitution, and mismatched resource annotations) are common failure modes in QEC simulation studies. We recommend that future QEC simulation papers report (i) the exact circuit-modification pipeline, (ii) the number of times each noise channel is applied, (iii) whether injection occurs inside or outside repeated circuit bodies, and (iv) the failure counts per cell rather than only aggregate probabilities. These four practices would have made the six issues above visible without a reanalysis.

7.2 Statistical and physical limits

The principal study uses one local error scale and four small-to-moderate distances. It provides no threshold crossing, no extrapolated logical error target below the sampling resolution, and no proof that an observed sign persists asymptotically. At distance nine, the

shot budget resolves probabilities only coarsely. Increasing the number of conditions cannot substitute for collecting more failures when estimating a rare-event ratio. The opposite-direction Z-basis geometry trend therefore remains a result to investigate with additional samples and orientations.

The noise is stochastic, Markovian, and Pauli-based. Leakage, coherent over-rotation, crosstalk, burst correlations, failed entanglement attempts, and finite communication latency are absent. G is a synthetic localized-noise circuit, not a calibrated teleported-CNOT implementation. It does not prepare a Bell pair, condition on heralding, or execute classical feedforward. The source benchmark also lacks a remote logical operation. Consequently, no remote-CNOT fidelity, Bell-pair threshold, or distributed-computation speedup follows from the numerical results.

The two logical bases are sampled separately. Their probabilities cannot be combined into a full logical channel without additional joint information. The fixed vertical seam also couples differently to the code boundaries and extraction schedule. A more complete orientation study would rotate the seam and examine whether the basis-dependent trends transform as expected. The present results identify that issue but do not replace such an experiment.

Power asymmetry. The $d = 3$ X-basis counterexample is decisive (Holm-adjusted $p = 4.56 \times 10^{-19}$), while the $d = 5$ X-basis comparison, with an almost identical point estimate (0.680), is not familywise significant (Holm-adjusted $p = 0.0597$). This asymmetry is expected from binomial statistics. With 150 failures in the matched cell, the minimum detectable ratio from Equation (18) is approximately 0.59 (Figure 5); a true ratio of 0.68 would reach 80% power with approximately 1.7 times as many shots (about 340,000 per model). We therefore report the $d = 5$ comparison as consistent with the counterexample but not independently decisive, and identify this shot count for a future replication.

7.3 Requirements for an architectural extension

A distributed extension should implement an explicit remote stabilizer or merge circuit with its Bell-pair channel, local operations, measurements, and feedforward. The logical observable must correspond to the intended operation, and fault propagation must preserve the claimed distance. The circuit should then be tested over several local and link error rates, multiple distances, both logical error types, and a stated communication schedule. The recent constructions of Schmidt et al. [29] and Thakur et al. [30] provide relevant starting points, while the interface comparisons of Marqversen et al. [27] help frame hardware-dependent trade-offs.

Resource comparisons should include communication qubits, expected entanglement attempts, buffering, measurement latency, and decoder throughput. Heterogeneous-distance layouts need an optimization criterion tied to a common logical-error target. Until those conditions are specified, the qubit arithmetic in Section 5.6 is informative geometry rather than an architectural advantage. Keeping that boundary explicit makes the current findings useful as a modeling baseline for later distributed implementations.

8 Conclusion

This study examined how seam-noise modeling choices affect rotated surface-code memory results. The primary contributions are twofold. First, the reanalysis of the source benchmark preserved its 2.6 million sampled observations while identifying six specific implementation

issues (undocumented parameter values, injection outside repeated circuit bodies, ineffective post-measurement operations, unused configuration fields, algebraic curve substitution, and mismatched resource annotations) that constrain the interpretation of its spatial support, temporal exposure, channel probabilities, confidence intervals, and resource plots. The audit methodology is reproducible and generalizable to other QEC simulation studies.

Second, a separate 14.4-million-shot validation study compared a geometric seam with a channel-wise mean-matched homogeneous control across two memory bases and four distances.

The main result is that matching expected physical exposure does not generally match logical risk. In the distance-three X-basis experiment at the largest severity, the geometric-to-uniform risk ratio was 0.683 with a 95% interval of 0.630 to 0.741. Decoding the same geometric samples with matched rather than baseline weights reduced failures from 2060 to 977. Opposite-direction Z-basis point estimates and rare-event results at larger distances require more restrained interpretation after accounting for multiplicity and sampling precision. The Z-basis point estimate at the same distance shows the opposite trend (ratio 1.106), consistent with a physical hypothesis based on the different orientations of the logical X and Z failure paths relative to the seam; this hypothesis is stated for future work and is not confirmed by the present data.

Geometry, activation duration, and decoder calibration should therefore be treated as separate variables in seam-noise studies. The supplied reproducibility package supports that separation through explicit circuits, recorded seeds, failure counts, and regenerated figures. The evidence establishes a controlled memory-model comparison; extension to fault-tolerant distributed gates requires an explicit communication protocol and a broader physical-noise study.

Declarations

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

No external funding was received for this study.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable. This study is based entirely on computational simulations and numerical analysis and did not involve human participants, human tissues, animals, or identifiable personal data.

Consent for publication

Not applicable. This manuscript does not contain any individual person's identifiable data or images.

Data availability

The data and analysis artifacts that support the findings of this study are not publicly available because they contain simulation outputs and internal analysis pipelines that are subject to institutional review. They are available from the corresponding author upon reasonable authenticated request and are intended for academic research use only.

Author contributions

AKSP: Conceptualization, methodology, software, investigation, data analysis, and manuscript preparation.
PS: Methodology, analysis, interpretation, and manuscript review.
RG: Supervision, conceptualization, research coordination, interpretation, and critical revision of the manuscript.
RK: Technical review, analysis, and manuscript revision.
KS: Computational analysis, validation, and manuscript review.

References

[1] Peter W. Shor. Scheme for reducing decoherence in quantum computer memory. *Physical Review A* 52(4), R2493-R2496 (1995). https://doi.org/10.1103/physreva.52.r2493

[2] A. M. Steane. Error Correcting Codes in Quantum Theory. *Physical Review Letters* 77(5), 793-797 (1996). https://doi.org/10.1103/physrevlett.77.793

[3] A. R. Calderbank, Peter W. Shor. Good quantum error-correcting codes exist. *Physical Review A* 54(2), 1098-1105 (1996). https://doi.org/10.1103/physreva.54.1098

[4] Emanuel Knill, Raymond Laflamme. Theory of quantum error-correcting codes. *Physical Review A* 55(2), 900-911 (1997). https://doi.org/10.1103/physreva.55.900

[5] A.Yu. Kitaev. Fault-tolerant quantum computation by anyons. *Annals of Physics* 303(1), 2-30 (2003). https://doi.org/10.1016/s0003-4916(02)00018-0

[6] Eric Dennis, Alexei Kitaev, Andrew Landahl, John Preskill. Topological quantum memory. *Journal of Mathematical Physics* 43(9), 4452-4505 (2002). https://doi.org/10.1063/1.1499754

[7] Austin G. Fowler, Matteo Mariantoni, John M. Martinis, Andrew N. Cleland. Surface codes: Towards practical large-scale quantum computation. *Physical Review A* 86(3), 032324 (2012). https://doi.org/10.1103/physreva.86.032324

[8] Dominic Horsman, Austin G Fowler, Simon Devitt, Rodney Van Meter. Surface code quantum computing by lattice surgery. *New Journal of Physics* 14(12), 123011 (2012). https://doi.org/10.1088/1367-2630/14/12/123011

[9] Daniel Litinski. A Game of Surface Codes: Large-Scale Quantum Computing with Lattice Surgery. *Quantum* 3, 128 (2019). https://doi.org/10.22331/q-2019-03-05-128

[10] Charles H. Bennett, Gilles Brassard, Claude Crépeau, Richard Jozsa, Asher Peres, William K. Wootters. Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Physical Review Letters* 70(13), 1895-1899 (1993). https://doi.org/10.1103/physrevlett.70.1895

[11] Daniel Gottesman, Isaac L. Chuang. Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations. *Nature* 402(6760), 390-393 (1999). https://doi.org/10.1038/46503

[12] J. Eisert, K. Jacobs, P. Papadopoulos, M. B. Plenio. Optimal local implementation of nonlocal quantum gates. *Physical Review A* 62(5), 052317 (2000). https://doi.org/10.1103/physreva.62.052317

[13] Charles H. Bennett, Gilles Brassard, Sandu Popescu, Benjamin Schumacher, John A. Smolin, William K. Wootters. Purification of Noisy Entanglement and Faithful Teleportation via Noisy Channels. *Physical Review Letters* 76(5), 722-725 (1996). https://doi.org/10.1103/physrevlett.76.722

[14] David Deutsch, Artur Ekert, Richard Jozsa, Chiara Macchiavello, Sandu Popescu, Anna Sanpera. Quantum Privacy Amplification and the Security of Quantum Cryptography over Noisy Channels. *Physical Review Letters* 77(13), 2818-2821 (1996). https://doi.org/10.1103/physrevlett.77.2818

[15] W. Dür, H.-J. Briegel. Entanglement Purification for Quantum Computation. *Physical Review Letters* 90(6), 067901 (2003). https://doi.org/10.1103/physrevlett.90.067901

[16] Stefan Krastanov, Victor V. Albert, Liang Jiang. Optimized Entanglement Purification. *Quantum* 3, 123 (2019). https://doi.org/10.22331/q-2019-02-18-123

[17] Liang Jiang, Jacob M. Taylor, Anders S. Sørensen, Mikhail D. Lukin. Distributed quantum computation based on small quantum registers. *Physical Review A* 76(6), 062323 (2007). https://doi.org/10.1103/physreva.76.062323

[18] C. Monroe, R. Raussendorf, A. Ruthven, K. R. Brown, P. Maunz, L.-M. Duan, et al. Large-scale modular quantum-computer architecture with atomic memory and photonic interconnects. *Physical Review A* 89(2), 022317 (2014). https://doi.org/10.1103/physreva.89.022317

[19] Ying Li, Simon C Benjamin. High threshold distributed quantum computing with three-qubit nodes. *New Journal of Physics* 14(9), 093008 (2012). https://doi.org/10.1088/1367-2630/14/9/093008

[20] Naomi H. Nickerson, Ying Li, Simon C. Benjamin. Topological quantum computing with a very noisy network and local error rates approaching one percent. *Nature Communications* 4(1), 1756 (2013). https://doi.org/10.1038/ncomms2773

[21] Naomi H. Nickerson, Joseph F. Fitzsimons, Simon C. Benjamin. Freely Scalable Quantum Technologies Using Cells of 5-to-50 Qubits with Very Lossy and Noisy Photonic Links. *Physical Review X* 4(4), 041041 (2014). https://doi.org/10.1103/physrevx.4.041041

[22] Austin G. Fowler, David S. Wang, Charles D. Hill, Thaddeus D. Ladd, Rodney Van Meter, Lloyd C. L. Hollenberg. Surface code quantum communication. *Physical Review Letters* 104(18), 180503 (2010). https://doi.org/10.1103/physrevlett.104.180503

[23] Ying Li, Simon C. Benjamin. Hierarchical surface code for network quantum computing with modules of arbitrary size. *Physical Review A* 94(4), 042303 (2016). https://doi.org/10.1103/physreva.94.042303

[24] Ramil Nigmatullin, Christopher J Ballance, Niel de Beaudrap, Simon C Benjamin. Minimally complex ion traps as modules for quantum communication and computing. *New Journal of Physics* 18(10), 103028 (2016). https://doi.org/10.1088/1367-2630/18/10/103028

[25] Joshua Ramette, Josiah Sinclair, Nikolas P. Breuckmann, Vladan Vuletić. Fault-tolerant connection of error-corrected qubits with noisy links. *npj Quantum Information* 10(1), 58 (2024). https://doi.org/10.1038/s41534-024-00855-4

[26] Hugo Jacinto, Élie Gouzien, Nicolas Sangouard. Network requirements for distributed quantum computation. *Physical Review Research* 8(1), 013205 (2026). https://doi.org/10.1103/v9ln-c4v2

[27] Frederik K. Marqversen, Gefen Baranes, Maxim Sirotin, Johannes Borregaard. Fault-tolerant interfaces for modular quantum computing on diverse qubit platforms. *Physical Review Research* 8(2), 023097 (2026). https://doi.org/10.1103/ds45-fm9n

[28] Nitish Kumar Chandra, Reza Nejabati, Eneet Kaur. Towards the Characterization of Logical Errors in Distributed Lattice Surgery. *arXiv preprint* arXiv:2607.29186 (2026). https://arxiv.org/abs/2607.29186

[29] Moritz Schmidt, Martin Moureau, Benjamin Rodatz, Boldizsár Poór, Elie Mounzer, Linnea Grans-Samuelsson. Resource-adaptive distributed fault tolerance with very noisy Bell pairs. *arXiv preprint* arXiv:2609.03048 (2026). https://arxiv.org/abs/2609.03048

[30] Pritesh Thakur, Daniel Dilley, Zain Saleem. A Syndrome-Extraction Framework for Distributed Lattice Surgery on Arbitrary Rotated Surface-Code Layouts. *arXiv preprint* arXiv:2609.26736 (2026). https://arxiv.org/abs/2609.26736

[31] John Stack, Ming Wang, Frank Mueller. Transversal fault tolerant distributed quantum computing operations. *Nature Communications* 17(1), 8838 (2026). https://doi.org/10.1038/s41467-026-75693-3

[32] Andreas Reiserer, Gerhard Rempe. Cavity-based quantum networks with single atoms and optical photons. *Reviews of Modern Physics* 87(4), 1379-1418 (2015). https://doi.org/10.1103/revmodphys.87.1379

[33] L. J. Stephenson, D. P. Nadlinger, B. C. Nichol, S. An, P. Drmota, T. G. Ballance, et al. High-rate, high-fidelity entanglement of qubits across an elementary quantum network. *Physical Review Letters* 124(11), 110501 (2020). https://doi.org/10.1103/physrevlett.124.110501

[34] B. Hensen, H. Bernien, A. E. Dréau, A. Reiserer, N. Kalb, M. S. Blok, et al. Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature* 526(7575), 682-686 (2015). https://doi.org/10.1038/nature15759

[35] M. Pompili, S. L. N. Hermans, S. Baier, H. K. C. Beukers, P. C. Humphreys, R. N. Schouten, et al. Realization of a multinode quantum network of remote solid-state qubits. *Science* 372(6539), 259-264 (2021). https://doi.org/10.1126/science.abg1919

[36] D. Main, P. Drmota, D. P. Nadlinger, E. M. Ainley, A. Agrawal, B. C. Nichol, et al. Distributed quantum computing across an optical network link. *Nature* 638(8050), 383-388 (2025). https://doi.org/10.1038/s41586-024-08404-x

[37] P. Magnard, S. Storz, P. Kurpiers, J. Schär, F. Marxer, J. Lütolf, et al. Microwave Quantum Link between Superconducting Circuits Housed in Spatially Separated Cryogenic Systems. *Physical Review Letters* 125(26), 260502 (2020). https://doi.org/10.1103/physrevlett.125.260502

[38] Luke D. Burkhart, James D. Teoh, Yaxing Zhang, Christopher J. Axline, Luigi Frunzio, M.H. Devoret, et al. Error-Detected State Transfer and Entanglement in a Superconducting Quantum Network. *PRX Quantum* 2(3), 030321 (2021). https://doi.org/10.1103/prxquantum.2.030321

[39] Youpeng Zhong, Hung-Shen Chang, Audrey Bienfait, Étienne Dumur, Ming-Han Chou, Christopher R. Conner, et al. Deterministic multi-qubit entanglement in a quantum network. *Nature* 590(7847), 571-575 (2021). https://doi.org/10.1038/s41586-021-03288-7

[40] Sebastian Krinner, Nathan Lacroix, Ants Remm, Agustin Di Paolo, Elie Genois, Catherine Leroux, et al. Realizing repeated quantum error correction in a distance-three surface code. *Nature* 605(7911), 669-674 (2022). https://doi.org/10.1038/s41586-022-04566-8

[41] Google Quantum AI. Suppressing quantum errors by scaling a surface code logical qubit. *Nature* 614(7949), 676-681 (2023). https://doi.org/10.1038/s41586-022-05434-1

[42] Google Quantum AI and Collaborators. Quantum error correction below the surface code threshold. *Nature* 638(8052), 920-926 (2025). https://doi.org/10.1038/s41586-024-08449-y Figure-label correction: https://doi.org/10.1038/s41586-026-10559-8 (2026).

[43] Dolev Bluvstein, Simon J. Evered, Alexandra A. Geim, Sophie H. Li, Hengyun Zhou, Tom Manovitz, et al. Logical quantum processor based on reconfigurable atom arrays. *Nature* 626(7997), 58-65 (2024). https://doi.org/10.1038/s41586-023-06927-3

[44] Scott Aaronson, Daniel Gottesman. Improved simulation of stabilizer circuits. *Physical Review A* 70(5), 052328 (2004). https://doi.org/10.1103/physreva.70.052328

[45] Craig Gidney. Stim: a fast stabilizer circuit simulator. *Quantum* 5, 497 (2021). https://doi.org/10.22331/q-2021-07-06-497

[46] Oscar Higgott. PyMatching: A Python Package for Decoding Quantum Codes with Minimum-Weight Perfect Matching. *ACM Transactions on Quantum Computing* 3(3), 1-16 (2022). https://doi.org/10.1145/3505637

[47] Oscar Higgott, Craig Gidney. Sparse Blossom: correcting a million errors per core second with minimum-weight matching. *Quantum* 9, 1600 (2025). https://doi.org/10.22331/q-2025-01-20-1600

[48] Austin G. Fowler, Adam C. Whiteside, Angus L. McInnes, Alimohammad Rabbani. Topological Code Autotune. *Physical Review X* 2(4), 041003 (2012). https://doi.org/10.1103/physrevx.2.041003

[49] Sergey Bravyi, Martin Suchara, Alexander Vargo. Efficient algorithms for maximum likelihood decoding in the surface code. *Physical Review A* 90(3), 032326 (2014). https://doi.org/10.1103/physreva.90.032326

[50] Nicolas Delfosse, Naomi H. Nickerson. Almost-linear time decoding algorithm for topological codes. *Quantum* 5, 595 (2021). https://doi.org/10.22331/q-2021-12-02-595

[51] David K. Tuckett, Stephen D. Bartlett, Steven T. Flammia. Ultrahigh Error Threshold for Surface Codes with Biased Noise. *Physical Review Letters* 120(5), 050505 (2018). https://doi.org/10.1103/physrevlett.120.050505

[52] J. Pablo Bonilla Ataides, David K. Tuckett, Stephen D. Bartlett, Steven T. Flammia, Benjamin J. Brown. The XZZX surface code. *Nature Communications* 12(1), 2172 (2021). https://doi.org/10.1038/s41467-021-22274-1

[53] Edward H. Chen, Theodore J. Yoder, Youngseok Kim, Neereja Sundaresan, Srikanth Srinivasan, Muyuan Li, et al. Calibrated Decoders for Experimental Quantum Error Correction. *Physical Review Letters* 128(11), 110504 (2022). https://doi.org/10.1103/physrevlett.128.110504

[54] Thomas Wagner, Hermann Kampermann, Dagmar Bruß, Martin Kliesch. Pauli channels can be estimated from syndrome measurements in quantum error correction. *Quantum* 6, 809 (2022). https://doi.org/10.22331/q-2022-09-19-809

[55] Armands Strikis, Simon C. Benjamin, Benjamin J. Brown. Quantum Computing is Scalable on a Planar Array of Qubits with Fabrication Defects. *Physical Review Applied* 19(6), 064081 (2023). https://doi.org/10.1103/physrevapplied.19.064081

[56] James M. Auger, Silvia Bergamini, Dan E. Browne. Blueprint for fault-tolerant quantum computation with Rydberg atoms. *Physical Review A* 96(5), 052320 (2017). https://doi.org/10.1103/physreva.96.052320

[57] Edwin B. Wilson. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association* 22(158), 209-212 (1927). https://doi.org/10.1080/01621459.1927.10502953

[58] Lawrence D. Brown, T. Tony Cai, Anirban DasGupta. Interval Estimation for a Binomial Proportion. *Statistical Science* 16(2), 101–133 (2001). https://doi.org/10.1214/ss/1009213286

[59] Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6(2), 65–70 (1979). https://www.jstor.org/stable/4615733

[60] Quinn McNemar. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika* 12(2), 153-157 (1947). https://doi.org/10.1007/bf02295996

[61] B. L. Welch. The generalization of Student's problem when several different population variances are involved. *Biometrika* 34(1-2), 28-35 (1947). https://doi.org/10.1093/biomet/34.1-2.28