BEYOND ERROR RATES: SEPARATING DIFFICULTY FROM DECISION-RULE SHIFT IN NEURAL QUANTUM ERROR CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantum error correction (QEC) uses measured error signatures to infer and correct physical faults. Neural decoders learn a syndrome-to-prediction mapping from samples generated under one noise model, but real quantum hardware does not operate under a fixed noise distribution. Existing evaluations typically report how the logical error rate changes with noise, but this cannot tell whether the problem has become harder or the optimal decoding decisions have changed. We introduce NSBench, a circuit-level benchmark that evaluates neural QEC decoders on shared syndrome streams across 18 streaming scenarios, a 75-configuration noise grid, and mechanism-specific strength sweeps. In tractable settings, exact source- and target-optimal references separate target difficulty from the cost of retaining old decoding decisions, which we call the shift cost. Broad multiplier-grid perturbations substantially increase the logical error rate of seven trained neural decoders at distance three and two released pre-decoders. In the shallow exact settings, the grid's shift cost remains small relative to target difficulty, whereas sufficiently strong local hotspots and correlated pairs have much larger relative shift costs. In absolute terms, the hardest grid targets can cost as much as structured faults. Adaptation recovers part of the shift cost when syndrome history reveals the change, but can also degrade performance when little shift exists. An exact Bayesian filter further shows that syndrome history alone cannot identify some noise changes. The logical error rate alone is therefore insufficient to determine the need for adaptation, and neural QEC decoders should be evaluated by how physical noise changes the decoding problem itself.

1 INTRODUCTION

Quantum error correction (QEC) protects a logical qubit by encoding it across many physical qubits. In the surface code (Dennis et al., 2002; Fowler et al., 2012), these qubits form a lattice, and local stabilizer checks on neighboring qubits are measured repeatedly to detect errors without directly measuring the encoded logical state. The resulting sequence of check outcomes forms a syndrome that contains information about physical faults. A decoder uses this syndrome to predict whether the physical faults have flipped the logical state. Decoder accuracy directly affects the logical error rate. Neural decoders learn this syndrome-to-prediction mapping from samples generated under a particular noise model, and recent models achieve strong performance across several code settings and architectures (Bausch et al., 2024; Yan et al., 2026).

However, the noise seen on a real device is not stationary. Calibration parameters drift over time, individual qubits can become unusually noisy, and correlated error events can affect several qubits together (Klimov et al., 2018; McEwen et al., 2022; Google Quantum AI and Collaborators, 2025). A decoder trained under one fixed noise condition may therefore face a different distribution after deployment. When its logical error rate increases, the increase alone does not explain what went wrong.

A decoder must choose a logical prediction for each observed syndrome. We refer to this syndrome-to-prediction mapping as the decoder's decision rule. A noise shift can make the target problem harder without changing the optimal decision rule. Another shift can change the optimal prediction

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

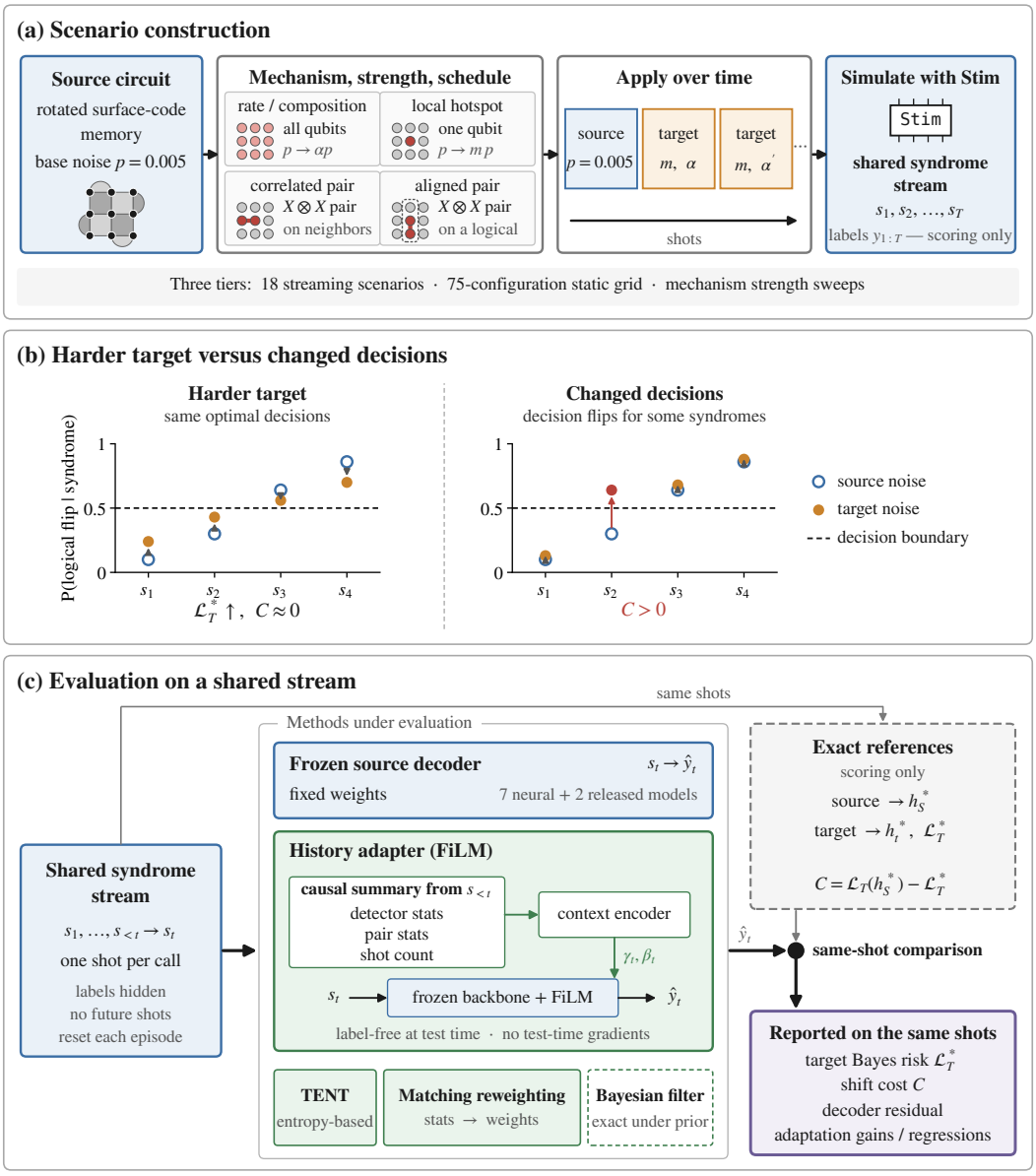


Figure 1: **Overview of the NSBench evaluation pipeline.** (a) Controlled changes in noise mechanism, strength, and temporal schedule define source-to-target scenarios and generate shared syndrome streams. (b) A schematic illustration of the exact source- and target-optimal references used in tractable settings to separate target difficulty from changes in optimal decoding decisions. (c) Frozen and adaptive decoders process the same stream causally. The decoders never receive the exact references, and we use them only for scoring.

for some syndromes, making the rule learned under the training noise systematically wrong for those syndromes.

Standard robustness evaluations largely hide this distinction. Varying physical error rates and plotting decoder accuracy shows whether performance degrades, but not why. A large rise in logical error rate can leave the optimal decisions almost unchanged relative to the target's own error rate, while a structured fault can change them after only a modest rise in difficulty. An error-rate curve alone cannot tell whether a benchmark tests a harder target, a changed decision rule, or both.

We therefore measure the two effects separately. The target Bayes risk $\mathcal{L}_T^*$ is the error rate of the best possible decoder for the new (target) noise, so it measures how hard the target problem is even for a decoder that knows that noise. The shift cost C is the extra error a decoder makes on the target noise by keeping the rule that was optimal for the training (source) noise (Figure 1). The distinction matters for adaptation. No decoder can reduce the error below $\mathcal{L}_T^*$, and C can be recovered only when deployment observations show how the optimal decisions have changed.

We introduce NSBench to measure these effects. NSBench generates controlled circuit-level noise shifts and evaluates every decoder on shared syndrome streams under a causal protocol. The benchmark contains 18 streaming scenarios, a 75-configuration noise grid, and separate strength sweeps for local and correlated faults. In tractable settings, exact source- and target-optimal references provide $\mathcal{L}_T^*$ and C on the same syndrome samples. We use NSBench to study trained neural decoders, released pre-decoders, and adaptive methods.

Our contributions are threefold.

- We introduce NSBench, a circuit-level benchmark that measures decoding difficulty and decision shift separately, on the same syndrome streams.
- We show that a rise in logical error rate does not by itself predict decision shift. Across the settings with exact dense references, the evaluated noise grids have shift costs of at most 3.8% of target Bayes risk, while sufficiently strong local hotspots and correlated pairs reach 14–87%. Exact per-shot inference extends this separation through code distance five.
- We show that this distinction changes how decoder adaptation should be evaluated. Decision shift explains one component of frozen-decoder failure. In our experiments, a label-free history adapter recovers 34–69% of the static-hotspot shift cost but significantly worsens 79 of 98 small-cost scenario–implementation pairs, by a mean of 0.64×10^{-3} LER. An exact Bayesian filter also identifies shifts that syndrome history alone cannot resolve.

2 Related Work

2.1 Neural QEC Decoding and Evaluation

Researchers have studied neural decoders with recurrent, convolutional, graph-based, and message-passing architectures, with recent models showing strong performance across several surface-code settings (Bausch et al., 2024; Yan et al., 2026; Sayedsalehi et al., 2026; Zenati & Nachmani, 2026; Peled et al., 2026). As these models have grown more capable, shared benchmarks and hardware datasets have made it easier to compare decoder performance under common conditions (Maurya et al., 2025; Google Quantum AI and Collaborators, 2025; Manor et al., 2026). Most evaluations, however, still measure performance under a specified noise model or a fixed hardware dataset. They can show that one decoder has a lower logical error rate than another, but not how much a change in physical noise alters the optimal decoding decisions themselves. NSBench focuses on this question by evaluating controlled source-to-target noise changes on shared syndrome streams (Appendix C, Table C.1).

2.2 Noise Shift and Adaptive Decoding

Prior work has examined how decoders behave when the noise differs from the model they assume. Darmawan (2025) studies how surface-code decoders should be adapted to local noise, while Hammar et al. (2022) show that decoding can operate without knowing the physical error rate, provided the noise bias is known. Classical adaptive decoders adjust to changing noise by estimating noise parameters from new observations or by updating the decoding graph (Huo & Li, 2017; Spitz et al., 2018; Nickerson & Brown, 2019; Wang et al., 2023; Kobori & Todo, 2025; Bhardwaj et al., 2026). Other work characterizes which aspects of the noise are identifiable from syndrome statistics alone (Wagner et al., 2021; 2022; 2023) and, therefore, what information syndrome-based adaptation can use. More recently, QAdapt trains a neural pre-decoder sequentially across noise regimes and evaluates it on a multiplier-based out-of-distribution noise grid (Miao et al., 2026). In machine learning, Wang et al. (2021) showed that test-time adaptation can improve performance under distribution shift, while Niu et al. (2023) and Press et al. (2023) later showed that such updates can become

unstable or degrade performance. NSBench connects these questions by measuring separately how much the target becomes harder and how much its optimal decision rule changes.

3 Problem Definition

3.1 Streaming Decoding Task

We consider a decoder trained under a source noise distribution and then deployed on a target stream. At shot t, the decoder receives a syndrome s_t and must predict whether the stored logical value has flipped. The syndrome is a vector of detector bits, where a detector is a parity of measurement outcomes that is deterministic without errors. We denote this binary label by y_t. The decoder is called once per shot and may use earlier syndromes from the same episode, a block of consecutive shots with one noise trajectory. A deployable decoder cannot access future shots, test labels, or the true target noise parameters, and its state is cleared at every episode boundary. We measure performance by the logical error rate (LER), the fraction of shots for which the predicted logical value is incorrect.

3.2 Difficulty and Shift Cost

Formally, let h_S^* be the optimal source rule and h_t^* the optimal rule for the true noise at target shot t. For T shots, define target difficulty and shift cost as

$$\mathcal{L}_T^* = \frac{1}{T}\sum_{t=1}^{T} \Pr\{h_t^*(s_t) \neq y_t\}, \tag{1}$$

$$C = \mathcal{L}_T(h_S^*) - \mathcal{L}_T^* \geq 0. \tag{2}$$

Here $\mathcal{L}_T(g)$ is the target error of a rule g. C is the target regret of the source-optimal rule, and we call it the shift cost throughout.

For any decoder g, adding and subtracting the source-rule error gives

$$\mathcal{L}_T(g) - \mathcal{L}_T^* = C + \underbrace{\mathcal{L}_T(g) - \mathcal{L}_T(h_S^*)}_{\text{decoder-specific residual}}. \tag{3}$$

The residual can be positive or negative because a learned rule may already make some target decisions correctly when the source rule does not. C is therefore the shift-related error of the exact source rule, not an upper bound on how much an imperfect decoder can improve.

We report $C/\mathcal{L}_T^*$ in the main text so that settings with different baseline error rates can be compared. Values above 100% mean that retaining the source rule more than doubles the target-optimal error. We also report absolute C. Whether an adapter can remove this error depends on what the syndrome history reveals, so we measure recovery relative to C using only unlabeled syndromes available at deployment. Appendix B collects the notation, and Appendix E derives the identity and its sampled estimators.

4 NSBench

NSBench has three parts: a library of controlled noise shifts, exact reference calculations for $\mathcal{L}_T^*$ and C in tractable settings, and a shared-stream protocol that scores every decoder on the same shots.

4.1 Noise Shifts

The shifts center on one contrast. Broad perturbations change noise strength across much of the circuit, while structured perturbations concentrate faults on particular qubits or introduce correlations between them. For example, a global multiplier changes the overall fault rate across the circuit, whereas a row hotspot changes where faults are most likely to occur. These two families let us test whether changes in noise strength and noise structure affect the optimal decoding decisions differently.

All shifts start from one source: a rotated surface-code memory circuit with a uniform base error probability $p = 0.005$ on gates, idling, measurement, and reset, simulated with Stim (Gidney, 2021). A shift modifies this circuit, so the noise model for every target is known exactly, as required by the references in Section 4.2.

The shifts form three tiers. The streaming tier has 18 scenarios presented as episode-ordered target streams. Some are stationary within the episode, such as a rate change or a static hotspot, and others change during it, such as drifts, bursts, switches in noise composition, and a moving hotspot (Appendix D). It functions as our deployment-style setting, and in the nonstationary scenarios an adaptive decoder must respond as the noise changes. Four of the 18—the two hotspots and the two pairs—are the structured tests, and the other fourteen serve as regression controls for adaptive decoders. The static grid has 75 configurations that multiply nonempty subsets of four noise axes (measurement, after-Clifford depolarization, idling, and Z bias) by factors from 1.2 to 3. It provides a broad multiplier-based perturbation comparable to the grid used in QAdapt (Miao et al., 2026). The strength sweeps vary one structured mechanism at a time from weak to strong. These mechanisms are single-qubit, row, and moving-row hotspots, XX pairs on neighboring qubits or along a logical operator, and a leakage-like control. In the control, one data qubit leaks at a random round boundary with small probability and stays fully depolarized for the rest of the shot, so most shots contain no leaked qubit. The sweeps are used to identify the strength at which each mechanism begins to change the optimal decisions.

For the structured scenarios and sweeps, the target circuit uses a base rate of 0.004 before the structured fault is applied, while the source remains at 0.005. This rate difference alone costs less than 0.02% of $\mathcal{L}_T^*$ (Appendix G). The benchmark strengths are stress settings, chosen so that their effects are measurable at $d = 3$: a fivefold hotspot and pair probabilities of 0.03 and 0.025 per round. Section 5.2 compares them with device observations, and the sweeps cover weaker settings.

4.2 Exact References

Equations 1 and 2 require optimal decisions under the source and target noise models. Because each target circuit is known, NSBench can compute these decisions exactly in tractable settings. For small circuits, a dense engine enumerates the full joint distribution of syndrome and logical label, which is exact maximum-likelihood decoding (Bravyi et al., 2014), giving population-exact values of $\mathcal{L}_T^*$ and C with no sampling error. It covers $d = 3$ with up to three rounds and $d = 5$ with one round (Appendix F). For deeper circuits, a per-shot engine computes the exact posterior and optimal decision for each sampled syndrome by eliminating fault variables in time order, which keeps the working set bounded as rounds are added. The decisions are exact, while averages over the sampled stream retain sampling uncertainty. It extends the references to $d = 5$ with 25 rounds and $d = 7$ with two rounds (Section 5.5). The per-shot engine reproduces the dense tables to within 10^{-14}, and every run checks that the source and target rules coincide when the two noise models are identical (Appendix H).

4.3 Evaluation Protocol

Every decoder at a setting receives the same shots, so all comparisons are paired. Each streaming scenario and each grid configuration has 64 independent episodes of 512 shots, and adaptive decoders are reset at every episode boundary. A causality audit perturbs future syndromes and checks that earlier predictions do not change, testing the streaming contract of Section 3.1 on the decoder's behavior rather than its code. We estimate statistical variability by resampling or permuting whole episodes, which preserves within-episode dependence, and we correct regression tests for multiple comparisons (Benjamini & Hochberg, 1995). We always report gains next to the frozen baseline, and we report recovery relative to C only when the paired estimate of C is distinguishable from zero (Appendix M).

5 Experiments and Results

5.1 Decoders and Setup

The main controlled experiments use code distance $d = 3$ with $r = 2$ measurement rounds. We train seven neural implementations on the same six million source shots with the same training protocol, repeated across three seeds. We additionally evaluate classical decoders and the released Ising Accurate (Chamberland et al., 2026) and QAdapt HTNet (Miao et al., 2026) pre-decoders with their published weights, without retraining. Because these released models were trained under different conditions, we use them as external deployment comparisons rather than architecture-matched baselines. Appendix K (Table K.1) summarizes the methods, training budgets, and source performance.

To test whether causal syndrome history can recover shift-related error, the main adaptation study adds an identity-initialized feature-modulation (FiLM) layer (Perez et al., 2018) to each neural backbone. A context encoder receives causal detector statistics, selected detector-pair statistics, and the elapsed shot count. The adapter is trained on labeled episodes drawn from a noise family and is deployed without labels or test-time gradients. Comparisons include entropy-based adaptation, matching reweighting, and an exact Bayesian filter. The Bayesian filter is optimal only under its specified prior and is used as a causal reference.

5.2 When Does the Optimal Decision Change?

In the evaluated settings, structured faults change the optimal decisions much more than the multiplier-grid perturbations. Across the four dense-integration settings, every grid configuration has a shift cost below 3.8% of target Bayes risk, while the hotspot and pair scenarios cost 14–87% (Table F.2). The strength sweeps show that both mechanism and strength matter (Figure 2). Every mechanism starts with low shift cost, but the cost grows at different rates depending on fault type and location. At the same multiplier of three, the center row costs 9.3% of its target Bayes risk, and a global multiplier costs 0.3% (Table F.5). A row hotspot crosses the 4% reference level at about $2.1\times$, a corner hotspot at about $4.5\times$, and the leakage-like control stays below it over the evaluated range.

The aligned pair is the extreme case. It lies on a logical operator, so at $d = 3$ it shares a syndrome with a single boundary error while carrying the opposite logical label (Appendix G.1). Its cost also grows with code size. At event probability 0.004, it rises from 0.11% at $d = 3$, $r = 2$ to 13.4% at $d = 5$, $r = 1$ and 22.4% at $d = 5$, $r = 5$ (Table G.3).

Two boundaries apply. The comparison is relative to target Bayes risk, and in absolute LER, the harder grid targets can cost as much as the structured faults (Table 1). The structured strengths are stress settings: published single-qubit relaxation fluctuations reach about a factor of two, and an

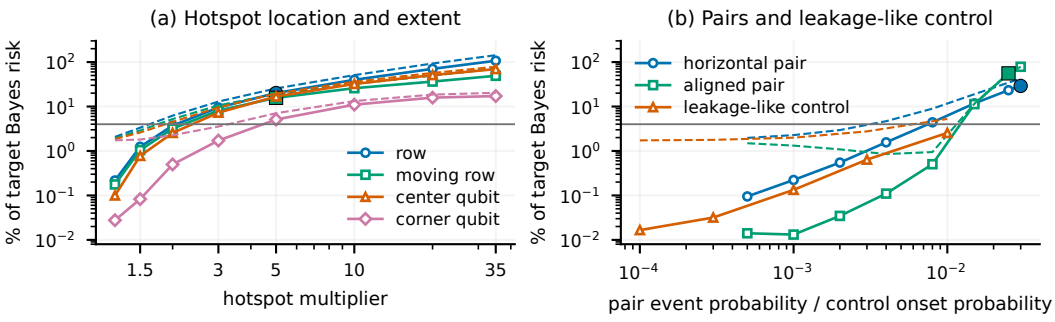


Figure 2: **Shift cost versus fault strength for each mechanism.** Solid lines with open markers show the exact population $C/\mathcal{L}_T^*$ at the reference setting, and dashed lines show the mean normalized excess error of the six source-matched neural decoders. Both quantities are expressed as a percentage of target Bayes risk. Filled markers indicate the strengths used in the streaming benchmark. The horizontal line marks a 4% reference level chosen just above the dense-grid maximum.

order of magnitude at the documented extreme, where a tenfold single-qubit hotspot costs 26.6% at $d = 5$, $r = 5$ (Appendix U).

5.3 Decoder Failure Reflects Both Rule Shift and Decoder-Specific Error

Shift cost explains much of the excess error of source-matched frozen decoders, but not all of it. For the six implementations that fit the source rule well, excess error tracks C across the strength sweep (Spearman $\rho = 0.96$, Figure 3a). They stay near the target optimum on the small-cost shifts and fall behind on the hotspot and pair scenarios. Neural-BP, with a larger source error, sits above the others (Table K.3, Appendix L).

The released pre-decoders show the same component despite different training distributions. On the aligned pair, Ising Accurate and QAdapt exceed the target-optimal error by 200% and 278%. On the same shots, Ising performs no better than the exact source rule, and QAdapt performs clearly worse than it (Figure 3b, Table I.1). Shift cost does not explain all of their error. Under the strongest multi-axis grid change, both models trail the source rule, and even without a shift their point estimates sit above the optimum, which is the decoder-specific residual of Equation 3. QAdapt is evaluated below its training distance and outside its task family, so these numbers describe deployment under shift rather than the method at its design point.

5.4 Can Adaptation Recover the Shift?

Recovery depends on whether the information available to the method identifies the change. The learned history adapter can also add error when little shift is available to recover. Where the fault is identifiable and stationary within the episode, recovery is substantial. The history adapter recovers 34–69% of the static-hotspot cost from detector statistics that reveal the hot row. The moving hotspot, which the filter identifies only transiently and recovers at 48%, is not recovered by the adapter (-29% to $+3\%$, with no significant gain on any backbone, Table L.1). A matching adapter with a circuit-enumerated fault library recovers 92% of the correlated-pair cost measured against its own frozen matching decoder, or 71% of the larger true-model matching headroom (Appendix L, Table R.2).

Syndrome history can also place a ceiling on recovery. At $d = 3$, the aligned pair and a single boundary error produce the same syndrome with opposite logical labels. The exact Bayesian filter, under a prior that contains the deployed configuration, recovers about half of the aligned-pair cost, and longer episodes do not remove the ambiguity. At $d = 5$, the pair has its own footprint, and the ambiguity disappears (Appendices G.1 and P).

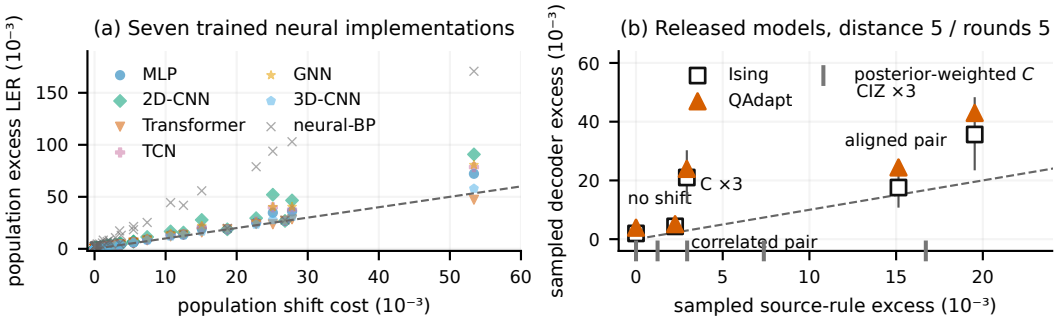


Figure 3: **Frozen-decoder error versus shift cost.** (a) Each point represents one neural decoder at one strength-sweep setting, averaged over training seeds. The diagonal shows the excess error expected from the exact source-optimal rule. (b) Released pre-decoders evaluated on shared distance-five shots. Error bars show paired 95% intervals for Ising Accurate, drawn at the shared source-rule excess (QAdapt's run recorded aggregate rates only). Grey ticks on the horizontal axis mark the posterior-weighted estimate of C. Full values are reported in Tables I.1 and 1, and Appendix H describes the two estimators.

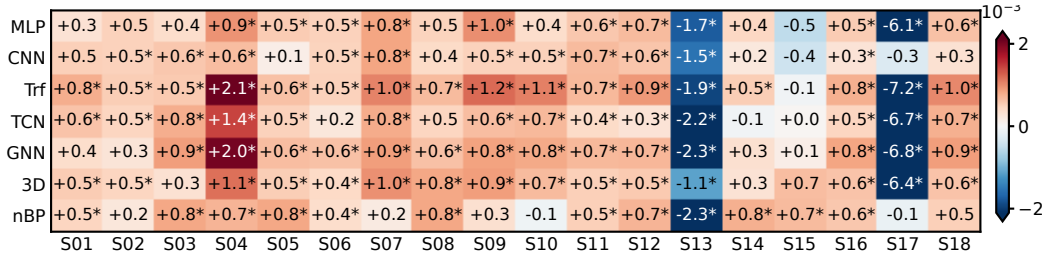


Figure 4: **History-adapter change in LER on all 18 scenarios.** Each cell is the history adapter's change in LER relative to its frozen backbone, in units of 10^{-3}, and negative values indicate improvement. Stars denote significance after multiple-comparison correction. Columns follow the scenarios in Table D.1. CNN, Trf, 3D, and nBP denote 2D-CNN, Transformer, 3D-CNN, and neural-BP.

Table 1: **Shift cost from exact decisions at deeper circuits.** For each setting, "rel." reports $C/\mathcal{L}_T^*$ in percent and "abs." reports C in units of 10^{-3}. Brackets give 95% block-bootstrap intervals where available. "Grid" is the largest grid point evaluated at that depth, not the full grid maximum. "Unresolved" means at most one differing decision in 1,024 shots, and dashes mean not computed. The † row has one 512-shot block per shifted case and so no block-based interval. The $(3, 8)$ row reports label-based estimates from a separate 32,768-shot sample, whereas the other rows report posterior-weighted estimates (Appendix H). "Static row" and "moving row" are scenarios S13 and S14 of Table D.1.

(d, r)		Grid	Static row	Moving row	Correlated pair	Aligned pair
$(3, 8)$	rel.	4.4 [3.3, 5.8]	23.9 [20.9, 27.6]	19.0	45.8 [40.3, 51.4]	94.4 [90.0, 100.4]
	abs.	8.21	12.15	9.92	11.44	82.76
$(5, 5)$	rel.	5.8 [5.2, 6.6]	24.2 [16.6, 31.4]	–	67.0 [37.0, 109.3]	173.2 [135.1, 220.2]
	abs.	7.37	2.89		2.95	16.71
$(5, 25)^\dagger$	rel.	2.2	–	–	81.5	216.1
	abs.	3.74			16.30	83.75
$(7, 2)$	rel.	9.1 [5.7, 12.8]	unresolved	–	unresolved	unresolved
	abs.	2.70				

Adaptation also adds error where there is little to recover. In small-cost scenarios, the history adapter worsens 97 of 98 scenario–implementation cells, 79 of them significantly, by a mean of 0.64×10^{-3} LER, including cells inside its training support and on the no-shift control (Figure 4, Table L.5). An identity gate that delays modulation halves the number of significant regressions, from 58 to 30 on the in-house cells, so they reflect this adapter's calibration at least in part (Appendix O).

Recovery also depends on what the adapter was trained on. Removing the aligned-pair class from meta-training leaves its recovery intact, but this is not an out-of-family test, because the retained pair class contains the physically identical XX channel on vertical pairs, including the benchmark pair. Removing the hotspot class turns static-hotspot recovery negative and worsens the moving hotspot for all seven backbones (Table O.1).

5.5 Scaling Beyond the Small Exact Setting

The separation holds with exact decisions through distance five, and only weaker evidence exists beyond it. At $d = 5$ with five rounds, the largest evaluated grid point costs 5.8% of target Bayes risk, while the aligned pair costs 173% and the row hotspot and correlated pair 24% and 67% (Table 1). Single-qubit hotspots at multiplier 5 are not resolved at this depth (Table U.1).

Beyond that, the exact evidence thins out. The 25-round results rest on a single 512-shot block per case. At $d = 7$, the grid point is estimated from only two blocks, so it is indicative only, and the structural costs there remain unresolved. The released pre-decoders keep their gains over source-model matching under rate changes at $d = 7, r = 7$, but lose or reverse them under strong multi-axis shifts and a strong hot row (Table I.2). A matching surrogate follows the structural sweep to $d = 11$,

where a tenfold hotspot and the aligned pair still read about 20% and 100% (Appendix R). The surrogate underestimates grid costs, so it cannot compare the two families beyond $d = 5$.

6 Discussion

6.1 Error-Rate Robustness Is Not Decision Robustness

In the shallow exact settings, the multiplier grid raises the logical error rate mainly by making the target harder, and the decision changes it induces stay small relative to that harder target's Bayes risk. Sufficiently strong structured faults change the optimal decisions themselves. Neither a strength grid nor a mechanism label alone determines whether an evaluation contains substantial decision changes, because the shift cost depends on the mechanism and its strength together. A noise-shift benchmark should therefore report target difficulty and shift cost beside decoder error. It should measure adaptation gains and regressions against an explicit frozen baseline on the same shots.

6.2 Adaptation Depends on What the Decoder Can Observe

Decision shift does not guarantee that adaptation can recover it. The hotspot exposes its location through detector statistics, so the history adapter recovers part of its shift cost. In our correlated-pair setting, recovery requires different information, such as pair statistics or a circuit fault library. The aligned pair goes further: at $d = 3$, two physical explanations produce the same syndrome with opposite logical effects, so syndrome history alone cannot completely resolve the decision. Recovery can also depend on which fault classes were represented during training. The observation interface and the training family are therefore part of the adaptation method, and an evaluation should state both.

6.3 Scope and Limitations

Our strongest conclusions are limited to settings where optimal decisions are exact and shift costs are statistically resolved. Population-exact references cover the small circuits, exact per-shot decisions provide resolved structural comparisons through distance five, and beyond that the evidence weakens (Section 5.5). The separation between grid and structured costs holds at the evaluated operating points, not as a general property of rate scaling. The deep grid comparisons also lie outside the regime where increasing the distance suppresses the logical error per round. For CIZ$\times$3, the target Bayes risk per round is 0.023 at $d = 3$ with eight rounds and 0.025 at $d = 5$ with five rounds.

The benchmark does not reproduce the full noise process of a specific device. It uses synthetic Pauli stress mechanisms on a uniform source model and evaluates only Z-memory circuits. The structured strengths were chosen as stress tests rather than calibrated device models, so their ranking should not be read as a prediction of hardware decoder rankings. Time is also compressed. An episode has at most 8,192 shots, about a second of device time at microsecond cycle times. The adaptation results therefore concern sub-second changes, such as burst recovery and fast telegraph switching, not hour-scale drift.

Our adaptation results concern one controlled setting: a single adapter design, one meta-training family, and $d = 3$ with $r = 2$. The meta-training family never varies the measurement multiplier, the axis that hardware recalibration most often targets, and the adapter's largest regressions fall on the measurement-dominated scenario. Appendix V lists further limits.

7 Conclusion

Changes in logical error rate do not by themselves reveal what has changed in a decoding problem. NSBench separates target difficulty from changes in the optimal decision rule and measures both on shared syndrome streams. These two effects can diverge substantially, and shift-related error can be recovered only when the relevant decision change is visible in the information available at deployment. This makes observability, not error-rate degradation alone, central to evaluating adaptive QEC decoders. Extending the same separation to larger codes and hardware data remains an open problem.

Reproducibility Statement

The appendices specify scenario parameters, decoder training, exact inference, paired statistical analysis, and implementation checks. An anonymized code archive is provided as supplementary material: the benchmark package with its scenario specifications and dataset digests, the exact dense and per-shot reference engines, the strength-sweep and surrogate scripts, the neural training and adaptation code, the released-model adapter, the case builders for the long exact runs, unit tests, and compact result records behind the tables and figures; its README maps each part of the paper to the code and records. The archive includes the per-episode error rates behind the episode-level paired tests, with a script that recomputes them; per-shot probability arrays are needed only to recompute those episode statistics from the decoders. Dataset archives are regenerated from the scenario files on the pinned Stim version and checked against the digests. Appendix T describes the additional resources needed to retrain models and to run the long exact references.

Ethics Statement

This work uses simulated quantum-error-correction circuits and a publicly released hardware corpus. It involves no human subjects or personal data. The controlled stress scenarios do not represent measured reliability claims about a deployed device.

AI Use Statement

Generative AI tools assisted with manuscript drafting and editing, code development, and checks of mathematical definitions, statistical calculations, and result interpretation. Numerical claims are linked to saved simulation and analysis records. Responsibility for the final submission, including verification of AI-assisted content, rests with the authors.

References

Abduxoliq Ashuraliyev. DAQEC-Benchmark: Drift-aware quantum error correction dataset with IBM hardware validation. Zenodo dataset, version 1.0.0, 2025. Self-deposited dataset record, not a refereed publication.

Johannes Bausch, Andrew W. Senior, Francisco J. H. Heras, Thomas Edlich, Alex Davies, Michael Newman, Cody Jones, Kevin Satzinger, Murphy Yuezhen Niu, Sam Blackwell, George Holland, Dvir Kafri, Juan Atalaya, Craig Gidney, Demis Hassabis, Sergio Boixo, Hartmut Neven, and Pushmeet Kohli. Learning high-accuracy error decoding for quantum processors. *Nature*, 635: 834–840, 2024.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Devansh Bhardwaj, Evangelia Takou, Yingjia Lin, and Kenneth R. Brown. Adaptive estimation of drifting noise in quantum error correction. *PRX Quantum*, 7:033024, 2026. arXiv:2511.09491.

Sergey Bravyi, Martin Suchara, and Alexander Vargo. Efficient algorithms for maximum likelihood decoding in the surface code. *Physical Review A*, 90:032326, 2014. doi: 10.1103/PhysRevA.90.032326.

Jonathan J. Burnett, Andreas Bengtsson, Marco Scigliuzzo, David Niepce, Marina Kudra, Per Delsing, and Jonas Bylander. Decoherence benchmarking of superconducting qubits. *npj Quantum Information*, 5:54, 2019. doi: 10.1038/s41534-019-0168-5.

Tiffany Tianhui Cai, Hongseok Namkoong, and Steve Yadlowsky. Diagnosing model performance under distribution shift. *Operations Research*, 74(2):898–916, 2026. doi: 10.1287/opre.2023.0217. arXiv:2303.02011.

Malcolm Carroll, Sami Rosenblatt, Petar Jurcevic, Isaac Lauer, and Abhinav Kandala. Dynamics of superconducting qubit relaxation times. *npj Quantum Information*, 8:132, 2022.

Christopher Chamberland, Jan Olle, Muyuan Li, Scott Thornton, and Igor Baratta. Fast and accurate AI-based pre-decoders for surface codes. *arXiv preprint arXiv:2604.12841*, 2026.

Edward H. Chen, Theodore J. Yoder, Youngseok Kim, Neereja Sundaresan, Srikanth Srinivasan, Muyuan Li, Antonio D. Córcoles, Andrew W. Cross, and Maika Takita. Calibrated decoders for experimental quantum error correction. *Physical Review Letters*, 128:110504, 2022. doi: 10.1103/PhysRevLett.128.110504.

Andrew S. Darmawan. Optimal adaptation of surface-code decoders to local noise. *Physical Review A*, 112:042431, 2025. arXiv:2403.08706.

Dripto M. Debroy, Matt McEwen, Craig Gidney, Noah Shutty, and Adam Zalcman. LUCI in the surface code with dropouts. *Quantum*, 9:1936, 2025. doi: 10.22331/q-2025-12-11-1936.

Nicolas Delfosse and Naomi H. Nickerson. Almost-linear time decoding algorithm for topological codes. *Quantum*, 5:595, 2021. doi: 10.22331/q-2021-12-02-595.

Eric Dennis, Alexei Kitaev, Andrew Landahl, and John Preskill. Topological quantum memory. *Journal of Mathematical Physics*, 43(9):4452–4505, 2002. doi: 10.1063/1.1499754.

Austin G. Fowler, Matteo Mariantoni, John M. Martinis, and Andrew N. Cleland. Surface codes: Towards practical large-scale quantum computation. *Physical Review A*, 86:032324, 2012. doi: 10.1103/PhysRevA.86.032324.

Craig Gidney. Stim: a fast stabilizer circuit simulator. *Quantum*, 5:497, 2021.

Google Quantum AI. Exponential suppression of bit or phase errors with cyclic error correction. *Nature*, 595:383–387, 2021. doi: 10.1038/s41586-021-03588-y.

Google Quantum AI. Suppressing quantum errors by scaling a surface code logical qubit. *Nature*, 614:676–681, 2023. arXiv:2207.06431.

Google Quantum AI and Collaborators. Quantum error correction below the surface code threshold. *Nature*, 638:920–926, 2025. arXiv:2408.13687; data: Zenodo doi:10.5281/zenodo.13273331.

Manuel Guatto, Francesco Preti, Michael Schilling, Tommaso Calarco, Francisco Andrés Cárdenas-López, and Felix Motzoi. Real-time adaptive quantum error correction by model-free multi-agent learning. *arXiv preprint arXiv:2509.03974*, 2025. v1 September 2025; v2 revised 14 July 2026.

Karl Hammar, Alexei Orekhov, Patrik Wallin Hybelius, Anna Katariina Wisakanto, Basudha Srivastava, Anton Frisk Kockum, and Mats Granath. Error-rate-agnostic decoding of topological stabilizer codes. *Physical Review A*, 105:042616, 2022. arXiv:2112.01977.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

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. doi: 10.1145/3505637.

Ming-Xia Huo and Ying Li. Learning time-dependent noise to reduce logical errors: real time error rate estimation in quantum error correction. *New Journal of Physics*, 19:123032, 2017. arXiv:1710.03636.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

P. V. Klimov, J. Kelly, Z. Chen, et al. Fluctuations of energy-relaxation times in superconducting qubits. *Physical Review Letters*, 121:090502, 2018.

Takumi Kobori and Synge Todo. Bayesian inference of general noise-model parameters from the syndrome statistics of surface codes. *Physical Review A*, 112:052448, 2025. arXiv:2406.08981.

Xuegang Li, Junhua Wang, Yao-Yao Jiang, Guang-Ming Xue, Xiaoxia Cai, Jun Zhou, Ming Gong, Zhao-Feng Liu, Shuang-Yu Zheng, Deng-Ke Ma, Mo Chen, Wei-Jie Sun, Shuang Yang, Fei Yan, Yi-Rong Jin, S. P. Zhao, Xue-Feng Ding, and Hai-Feng Yu. Cosmic-ray-induced correlated errors in superconducting qubit array. *Nature Communications*, 16:4677, 2025. doi: 10.1038/s41467-025-59778-z.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning (ICML)*, 2018.

Shay J. Manor, Leila S. Erhili, and Yassine Jebbouri. A sim-to-real study of surface-code decoder benchmarking. *arXiv preprint arXiv:2609.04557*, 2026.

Satvik Maurya, Joshua Viszlai, Nithin Raveendran, Poulami Das, and Swamit Tannu. decoderbench: Benchmarking decoders for quantum error correction. In *IEEE International Symposium on Workload Characterization (IISWC)*, 2025.

M. McEwen, D. Kafri, Z. Chen, et al. Removing leakage-induced correlated errors in superconducting quantum error correction. *Nature Communications*, 12:1761, 2021.

Matt McEwen, Lara Faoro, Kunal Arya, et al. Resolving catastrophic error bursts from cosmic rays in large arrays of superconducting qubits. *Nature Physics*, 18:107–111, 2022.

Ran Miao, Rui Luo, Xiaohan Shan, and Xiaoming Sun. QAdapt: A noise-adaptive neural predecoding framework for quantum error correction. *arXiv preprint arXiv:2607.28422*, 2026.

Naomi H. Nickerson and Benjamin J. Brown. Analysing correlated noise on the surface code using adaptive decoding algorithms. *Quantum*, 3:131, 2019. arXiv:1712.00502.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations (ICLR)*, 2023.

Yotam Peled, David Zenati, and Eliya Nachmani. Neural minimum weight perfect matching for quantum error codes. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2601.00242.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671.

Ori Press, Steffen Schneider, Matthias Kümmerer, and Matthias Bethge. RDumb: A simple approach that questions our progress in continual test-time adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Timothy Proctor, Melissa Revelle, Erik Nielsen, Kenneth Rudinger, Daniel Lobser, Peter Maunz, Robin Blume-Kohout, and Kevin Young. Detecting and tracking drift in quantum information processors. *Nature Communications*, 11:5396, 2020. doi: 10.1038/s41467-020-19074-4.

Joschka Roffe, David R. White, Simon Burton, and Earl Campbell. Decoding across the quantum low-density parity-check code landscape. *Physical Review Research*, 2:043423, 2020. doi: 10.1103/PhysRevResearch.2.043423.

Samira Sayedsalehi, Nader Bagherzadeh, Maxim Shcherbakov, and Jean-Luc Gaudiot. Sparse Mamba decoder for quantum error correction: Efficient defect-centric processing of surface code syndromes, 2026.

Volodymyr Sivak, Michael Newman, and Paul Klimov. Optimization of decoder priors for accurate quantum error correction. *Physical Review Letters*, 133:150603, 2024. doi: 10.1103/PhysRevLett.133.150603.

Stephen T. Spitz, Brian Tarasinski, Carlo W. J. Beenakker, and Thomas E. O'Brien. Adaptive weight estimator for quantum error correction in a time-dependent environment. *Advanced Quantum Technologies*, 1:1800012, 2018. arXiv:1712.02360.

Evangelia Takou and Kenneth R. Brown. Estimating decoding graphs and hypergraphs of memory QEC experiments. *Physical Review A*, 112:052414, 2025. arXiv:2504.20212.

David K. Tuckett, Stephen D. Bartlett, and Steven T. Flammia. Ultrahigh error threshold for surface codes with biased noise. *Physical Review Letters*, 120:050505, 2018. doi: 10.1103/PhysRevLett.120.050505.

Thomas Wagner, Hermann Kampermann, Dagmar Bruß, and Martin Kliesch. Optimal noise estimation from syndrome statistics of quantum codes. *Physical Review Research*, 3:013292, 2021.

Thomas Wagner, Hermann Kampermann, Dagmar Bruß, and Martin Kliesch. Pauli channels can be estimated from syndrome measurements in quantum error correction. *Quantum*, 6:809, 2022.

Thomas Wagner, Hermann Kampermann, Dagmar Bruß, and Martin Kliesch. Learning logical Pauli noise in quantum error correction. *Physical Review Letters*, 130:200601, 2023.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Hanrui Wang, Pengyu Liu, Yilian Liu, Jiaqi Gu, Jonathan Baker, Frederic T. Chong, and Song Han. DGR: Tackling drifted and correlated noise in quantum error correction via decoding graph re-weighting. *arXiv preprint arXiv:2311.16214*, 2023. revised 2024.

C. D. Wilen, S. Abdullah, N. A. Kurinsky, et al. Correlated charge noise and relaxation errors in superconducting qubits. *Nature*, 594:369–373, 2021. arXiv:2012.06029.

Ge Yan, Shanchuan Li, and Yuxuan Du. Rethink the role of neural decoders in quantum error correction. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2605.12046.

David Zenati and Eliya Nachmani. SAQ: Stabilizer-aware quantum error correction decoder. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2512.08914.

Han Zhao, Rémi Tachet des Combes, Kun Zhang, and Geoffrey J. Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning (ICML)*, 2019.

A GUIDE TO THE APPENDICES

The appendices hold the definitions, derivations, complete tables and scope statements behind the main text. Each appendix opens with the question it answers, then gives the method, the evidence and the boundary of what that evidence supports. Table A.1 maps each section of the main text to the appendices that support it. The appendices fall into five groups. Definitions and exact references are in Appendices B, E, F and H. The construction of the benchmark is in Appendices D, G and U. The decoder results are in Appendices K, L, I and J. The adaptation studies are in Appendices N, O, P, Q and M. Scope, artifacts and limitations are in Appendices C, R, S, T and V.

Table A.1: Where each part of the main text is supported.

Main text	Supporting appendices
Section 2, related work	C comparison with prior benchmarks and adaptive decoders
Section 3.2, difficulty and shift cost	B notation, E the corrected-error identity and its sampled estimators
Section 4, noise shifts	D scenario definitions, G strength sweeps, U physical scope of the strengths
Section 4.2, exact references	F population integration, H per-shot inference at deeper circuits
Section 4.3, evaluation protocol	M statistical estimands and tests
Section 5.1, decoders and setup	K training and implementation details
Section 5.2, when the decision changes	F, G (the aligned pair in G.1), J static grid and axis drift, U
Section 5.3, frozen-decoder error	L complete frozen results, I released pre-decoders
Section 5.4, adaptation	L adapted results, N meta-training support, O interface interventions, P the Bayesian reference, Q episode length
Section 5.5, scaling	H, R larger codes and the matching surrogate, S hardware corpus
Section 6.3, scope and limitations	T artifacts, V additional limitations

Three levels of evidence appear throughout, and they differ in what is exact and where sampling uncertainty enters (Table A.2). Dense integration enumerates every syndrome and gives population values of $\mathcal{L}_T^*$ and C with no sampling error. Per-shot inference gives the exact posterior and the optimal decision for each sampled syndrome, so only the averages over the sampled stream carry uncertainty. Beyond these settings the evidence is behavioral. Decoder error rates on sampled streams are compared with one another, or with a matching surrogate, rather than with an exact reference, and the text says so wherever such evidence is used.

Table A.2: Levels of evidence. The exact quantity and the source of sampling uncertainty differ by level.

Level	Exact quantity	Sampling uncertainty	Settings	Appendices
Dense	population $\mathcal{L}_T^*$ and C over all syndromes	none	d=3 with $r\leq$3, d=5 with r=1	F, G
Per-shot	posterior $P(y \mid s)$ and the optimal decision for each sampled syndrome	stream averages only, posterior-weighted with block-bootstrap intervals	d=3 with r=8 (label-based), d=5 with r=5 and r=25, d=7 with r=2	H, I
Behavioral	none, decoder error rates on sampled streams	yes, episode-level resampling	all streaming and grid settings, released models at d=7, r=7, hardware corpus	L, R, S
Surrogate	none, a matching decoder's excess under the shift	yes	structural sweep to d=11	R

B GLOSSARY

term	meaning in this paper
syndrome s	the vector of detector bits of one shot; a detector is a parity of measurement outcomes that is deterministic without errors (two consecutive measurements of one stabilizer in the bulk; the prepared state or the final data readout at the boundaries in time)
logical observable y	whether the logical qubit flipped during the shot; the decoder predicts y from s
LER	logical error rate: the fraction of shots on which the prediction of y is wrong
DEM	detector error model: the list of independent fault mechanisms of a circuit, each with a probability and the set of detectors and observables it flips
source / target	the training noise model ($p = 0.005$ on four channels) / the noise of the scenario being decoded
source-optimal rule h_S^*	the decision rule that is optimal under the source model; the "training rule"
target-optimal rule h_t^*	the decision rule that is optimal under the true noise of shot t; its LER is the target Bayes risk
shift cost C	target regret of h_S^*: the extra error of keeping the training rule under the target noise (Eq. 2)
episode	512 consecutive shots with one noise trajectory in the reference suite (2,048 and 8,192 in the episode-length tests); adaptive decoders are reset at its start
segment	an interval of an episode with constant noise parameters; one exact table per segment
MWPM, BP+OSD, UF	minimum-weight perfect matching, belief propagation with ordered-statistics decoding, and Union-Find: classical decoders built from a DEM
FiLM	feature-wise linear modulation: a learned per-channel scale and shift applied to a hidden layer

C RELATION TO PRIOR WORK AND SCOPE OF THE COMPARISON

Table C.1: Complementary roles of decoder evaluation resources. Exact references in NSBench apply to the tractable settings specified in Section 4.2.

Resource	Main variation	Reference for interpreting error
Decoder benchmarks (Maurya et al., 2025; Yan et al., 2026)	Architecture, code, noise level	Accuracy under specified noise
Hardware data (Google Quantum AI and Collaborators, 2025; Manor et al., 2026)	Device observations	Measured device performance
QAdapt (Miao et al., 2026)	Noise-axis multipliers; sequential tasks	Pre-decoder and matching baselines
NSBench	Mechanism, strength, and time	Target Bayes risk and source-rule shift cost

To place the shift cost among the quantities that prior work already uses, this appendix compares it with the target regret of decoder-mismatch studies and with the estimands of domain adaptation, and then states which released baselines could be rerun. The target regret in Eq. 2 is the quantity that Darmawan (2025) measures for decoding under a mischaracterized noise model. Our use of target regret follows prior work on decoder mismatch. The contribution here is not a new regret functional but the benchmark around it: circuit-level temporal and structural scenarios, an exact reference under each tractable segment, common-shot evaluation, and a causal streaming contract.

Domain-adaptation theory makes a related distinction with a different estimand. If the hypothesis class contains the source-optimal rule, and $\lambda = \inf_h[\mathcal{L}_S(h) + \mathcal{L}_T(h)]$, then

$$0 \le \lambda - (\mathcal{L}_S^* + \mathcal{L}_T^*) \le C,$$

where the upper inequality evaluates the sum at h_S^*. This neither identifies λ with C nor bounds every fixed rule's error below by C (Ben-David et al., 2010). Results on invariant representations apply

to that method class (Zhao et al., 2019). DISDE separates performance changes across source and target distributions; our residual compares two rules on the same target distribution (Cai et al., 2026), so its covariate, conditional and support terms are not the terms of Eq. 4. Label-shift correction is one identifiable special case (Lipton et al., 2018).

Baseline coverage has three boundaries. QAdapt's public model card releases a pre-decoder checkpoint, a baseline, configurations and an inference patch, but not its training orchestration or optimizer and Fisher state; we evaluate that frozen release at $d = 5$ and $d = 7$ through an interface adapter (Appendix I), but its adaptation itself cannot be rerun (Miao et al., 2026). DGR motivates graph reweighting, and our classical tiers implement related estimation mechanisms rather than reproducing that system (Wang et al., 2023). The 3D-CNN and neural-BP backbones are ports of two designs of Yan et al. (2026), retrained on our data; our in-house MLP, TCN and Transformer differ from theirs.

Static severity grids are standard in machine-learning robustness evaluation (Hendrycks & Dietterich, 2019); our result concerns their decision-shift content at specified QEC operating points. Circuit adaptation (Guatto et al., 2025) changes a different object from the decoder. The DAQEC record is a self-deposited drift-aware dataset (Ashuraliyev, 2025). Hardware datasets and decoder benchmarks are complementary to controlled mechanism tests (Maurya et al., 2025; Manor et al., 2026).

D SCENARIO DEFINITIONS AND PHYSICAL ABSTRACTIONS

To make every scenario reproducible from its identifier, this appendix fixes the scenario registry (Table D.1), the episode structure and the physical abstraction behind each structured fault. Table D.1 fixes the identifiers used in every numerical table. All scenarios use the source circuit family; circuit transformations, code-portable selectors, episode seeds and sampled trajectories are in the code archive. The registry uses distinct episode seeds for evaluation and training. Each reference scenario has 64 independent episodes of 512 shots with a 32-shot reporting window. A segment is an interval with fixed noise parameters; the simulator samples the corresponding circuit jointly for syndrome and logical label.

Relaxation-time fluctuations support rate variability and telegraph-like transitions (Klimov et al., 2018; Carroll et al., 2022); radiation and correlated charge observations motivate bursts and spatially correlated mechanisms (McEwen et al., 2022; Li et al., 2025; Wilen et al., 2021); local quality loss and defective qubits motivate hotspot tests (Google Quantum AI and Collaborators, 2025; Debroy et al., 2025). None of these references fixes our strengths or timescales: the compressed 512-shot trajectories, the global burst approximation, the pair probabilities and a row moving deterministically every 96 shots are benchmark schedules. The periodic scenario has the weakest direct physical support and is labeled a stress test.

Both XX scenarios use the same memoryless stochastic channel at every round boundary and differ in orientation and strength. Its identifier in the code archive contains "persistent" because the fault is present at a fixed site at every boundary; the channel carries no state and does not model leakage occupation, spreading or recovery (McEwen et al., 2021). Its motivation is a fixed-site correlated fault, such as a residual coupling or a leaked qubit, whose effect on the decision the exact reference can integrate; a leakage-like fault with within-shot persistence is swept in Appendix G. The scenario's decision content comes from its placement: the vertical pair straddles the column carrying X_L, so in the true DEM it adds probability 0.025 to the boundary footprints $\{1\}$ and $\{13\}$ without a logical flip, which the source model carries only with a flip (0.0163 and 0.0109 after merging); the horizontal pair creates the new footprints $\{1, 3\}$ and $\{13, 15\}$ instead. The ZZZ cluster and the ZZ overlay are Z-type faults in a Z-memory experiment: they cannot flip Z_L and are controls of the unmeasured sector; in an X-memory experiment they would play the roles of the XX scenarios. Bursts are abstracted as global rate steps rather than the localized, near-saturating events reported on hardware. The always-present ZZZ cluster abstracts event-like correlated noise, and its role as a low-cost control rests on its computed small C, not on the absence of a direct flip of the measured observable.

Strength, support and source distribution enter every quantitative contrast, so we report Bayes risk beside regret rather than treating syndrome-density matching as equal difficulty. Broader source

Table D.1: Reference scenarios. Rates are physical channel probabilities; multiplier and overlay factors are identified separately. IDs are stable within this paper. The JSON specifications give segment boundaries and selector resolution.

ID	short name	specified mechanism
S01	matched	stationary $p = 0.005$
S02	rate $1.4\times$	stationary $p = 0.007$
S03	rate $2\times$	stationary $p = 0.010$
S04	measurement-dominated	$p = 0.0031$, measurement-flip multiplier 4
S05	linear drift	$p : 0.0025 \to 0.0095$ over the episode
S06	OU drift	bounded Ornstein–Uhlenbeck trajectory around 0.0055
S07	periodic stress	sinusoidal $p = 0.0055 \pm 0.0025$, period 200 shots
S08	abrupt switch	$p : 0.0035 \to 0.0085$ at shot 160
S09	short bursts	$p = 0.004 \to 0.011$ for 32 shots at 128 and 320
S10	long burst	$p = 0.004 \to 0.010$ for 160 shots from 170
S11	composition ramp	measurement multiplier $0.25 \to 4$, 16 stages, approximately matched syndrome density
S12	composition switch	same endpoints, switch at shot 160
S13	static hotspot	$p = 0.004$; every noise channel touching a center-row data qubit (8, 10, 12), i.e. its idle depolarization, the two-qubit depolarization of every gate acting on it and its readout flip, multiplied by 5
S14	moving hotspot	same strength, data rows swept every 96 shots
S15	correlated pair	$p = 0.004$, XX on the horizontal pair (10,12), probability 0.03, sampled independently at every round boundary; adds two detector-pair footprints absent from the source DEM
S16	correlated cluster	$p = 0.004$, ZZZ on center column (3,10,17), probability 0.02; small measured decision regret
S17	aligned pair	code identifier `temporally_persistent_error`; $p = 0.004$, XX on the vertical pair (10,17), which lies on a representative of X_L, probability 0.025, sampled independently at every round boundary; at $d = 3$, $X_{10}X_{17} = X_3 \cdot X_L$, so its footprint is that of a single X on boundary qubit 3 with the opposite logical parity
S18	drift + correlated	$p : 0.003 \to 0.008$ plus ZZ on (8,10); overlay factor $0 \to 1$ multiplies probability 0.02

models, other fault sites and strengths, X-memory experiments and continuous operation would test external validity rather than add samples from the same conditions.

E EXACT REFERENCE AND THE CORRECTED-ERROR IDENTITY

To justify reading the shift cost as one component of a decoder's excess error, this appendix derives the corrected-error identity, shows that the residual is signed, and defines the population and sampled estimators of $\mathcal{L}_T^*$ and C. Let $\mathcal{L}_T(f)$ be the expected target LER of a decoder f averaged over the episode. For a frozen decoder f the excess over the target-optimal rule splits into the shift cost of Eq. 2 and a residual,

$$H(f) = \mathcal{L}_T(f) - \mathcal{L}_T^* = C + E(f), \qquad E(f) = \mathcal{L}_T(f) - \mathcal{L}_T(h_S^*). \tag{4}$$

$C \geq 0$, while the residual $E(f)$ is *signed*: a learned decoder can already get right some of the decisions on which h_S^* and h_t^* disagree, so C is a reference, not a floor under every frozen decoder. The identity below separates those corrected decisions from errors on syndromes where the two Bayes rules agree.

For one target segment, let $\eta(s) = P_T(y = 1 \mid s)$, $h^*(s) = \mathbf{1}\{\eta(s) > 1/2\}$, $w(s) = |2\eta(s) - 1|$ and $A = \{s : h_S^*(s) \neq h^*(s)\}$; ties are assigned label zero. Conditional binary zero-one loss gives

$$H(f) = \mathbb{E}_T[w(s)\mathbf{1}\{f(s) \neq h^*(s)\}], \qquad C = \mathbb{E}_T[w(s)\mathbf{1}_A].$$

Table E.1: Posterior-weighted identity on observed syndromes, in 10^{-3} LER, averaged over three seeds. Tildes denote averages over sampled syndromes, rather than population integration or empirical label errors. $\widetilde{U}$ is corrected cost on changed decisions and $\widetilde{V}$ is error on stable decisions. Each row satisfies $\widetilde{H} = \widetilde{C} - \widetilde{U} + \widetilde{V}$ and $\widetilde{E} = \widetilde{V} - \widetilde{U}$.

scenario	model	$\widetilde{C}$	$\widetilde{U}$	$\widetilde{V}$	$\widetilde{H}$	signed $\widetilde{E}$
S13	MLP	3.2771	0.3902	1.0870	3.9738	+0.6967
S13	2D-CNN	3.2771	0.2639	1.9712	4.9844	+1.7073
S13	Transf.	3.2771	0.3244	0.9434	3.8962	+0.6190
S13	TCN	3.2771	0.2415	0.9301	3.9657	+0.6886
S13	GNN	3.2771	0.4571	1.2974	4.1175	+0.8403
S13	3D-CNN	3.2771	0.2891	0.8119	3.7999	+0.5228
S13	nBP	3.2771	0.3769	13.8340	16.7342	+13.4571
S15	MLP	2.1834	0.3710	1.5875	3.4000	+1.2166
S15	2D-CNN	2.1834	0.2107	1.7314	3.7041	+1.5208
S15	Transf.	2.1834	0.2500	0.6524	2.5858	+0.4024
S15	TCN	2.1834	0.2756	1.5221	3.4299	+1.2465
S15	GNN	2.1834	0.3816	0.8606	2.6623	+0.4789
S15	3D-CNN	2.1834	0.2499	0.6071	2.5406	+0.3572
S15	nBP	2.1834	0.1568	3.9573	5.9839	+3.8005
S17	MLP	19.2019	0.1060	0.0651	19.1610	-0.0409
S17	2D-CNN	19.2019	0.1089	0.1126	19.2055	+0.0036
S17	Transf.	19.2019	0.0503	0.0841	19.2357	+0.0338
S17	TCN	19.2019	0.1003	0.0947	19.1963	-0.0056
S17	GNN	19.2019	0.1119	0.2043	19.2943	+0.0924
S17	3D-CNN	19.2019	0.1004	0.0363	19.1378	-0.0641
S17	nBP	19.2019	0.4624	1.4808	20.2204	+1.0185
S14	MLP	2.7776	0.2848	1.0192	3.5120	+0.7344
S14	2D-CNN	2.7776	0.2209	1.6762	4.2329	+1.4553
S14	Transf.	2.7776	0.2123	0.7396	3.3049	+0.5273
S14	TCN	2.7776	0.2173	0.8200	3.3802	+0.6026
S14	GNN	2.7776	0.2626	1.4385	3.9535	+1.1759
S14	3D-CNN	2.7776	0.1863	0.7024	3.2937	+0.5161
S14	nBP	2.7776	0.5895	9.0619	11.2500	+8.4724

On A a binary rule agrees either with h^*_S or with h^*. Splitting the first expectation over A and its complement yields

$$U(f) = \mathbb{E}_T[w(s)\mathbf{1}_A\mathbf{1}\{f(s) = h^*(s)\}], \tag{5}$$

$$V(f) = \mathbb{E}_T[w(s)\mathbf{1}_{A^c}\mathbf{1}\{f(s) \neq h^*(s)\}], \tag{6}$$

$$H(f) = C - U(f) + V(f), \qquad E(f) = V(f) - U(f). \tag{7}$$

The identity holds after averaging over segments. $U, V \geq 0$ while E can be negative: a fixed rule equal to h^* on the target has $H = 0$ even when $C > 0$, so $H \geq C$ does not follow for a source-trained approximation without further assumptions.

The stable-region contribution is the *unconditional* expectation V, not the conditional mean on A^c. Table E.1 verifies Eq. 7 with saved probabilities for every frozen implementation; tildes denote posterior-weighted averages over observed syndromes. For the static-hotspot Transformer, in LER units, $\widetilde{C} = 0.0032771$, $\widetilde{U} = 0.0003244$, $\widetilde{V} = 0.0009434$ and $\widetilde{H} = 0.0038962$; its signed residual is 0.0006190. Some aligned-pair models have small negative residuals.

The exact engine constructs the distribution of the detector and observable parity vector by convolving independent fault mechanisms, summing the probabilities of degenerate configurations rather than selecting a minimum-weight path. It is exact for the supplied DEM up to floating-point error. Brute-force checks on small mechanisms and the causality tests are in the repository test suite. Feasibility is exponential in the number of detectors plus observables; the 2^{26} cap is an implementation limit.

Table F.1: Population decision regret, in 10^{-3} LER, from full outcome-table integration. Each completed row contains 75 grid configurations and the same four structural mechanisms. In absolute LER the weakest structural case is the correlated pair; normalized by Bayes risk it is the moving hotspot (Table F.3). The $d = 3$, $r = 1$ and $d = 5$, $r = 1$ grid maxima agree to three digits by coincidence (0.2637 and 0.2644). Grid axes are defined in Appendix J; the maximizing multiplier is 3. No Monte Carlo confidence interval is needed for these sums.

setting	bits	max. grid C (axes)	min. structural C	ratio
$d = 3, r = 1$	9	0.264 (CI)	1.026	3.89 ×
$d = 3, r = 2$	17	0.964 (MCZ)	2.212	2.29 ×
$d = 3, r = 3$	25	2.419 (CIZ)	3.730	1.54 ×
$d = 5, r = 1$	25	0.264 (C)	0.304	1.15 ×

Table F.2: Population $C/\mathcal{L}_T^*$, separately from the finite-shot extrema of Table F.3. The structural column spans the four specified mechanisms.

setting	max. grid	structural range
$d = 3, r = 1$	1.12%	14.32–24.05%
$d = 3, r = 2$	2.42%	15.69–55.72%
$d = 3, r = 3$	3.79%	16.89–70.67%
$d = 5, r = 1$	2.54%	19.90–86.53%

E.1 Population and sampled metrics

For $P_t(s, y)$, the exact joint probability table at shot t, binary zero–one loss gives

$$C = \frac{1}{T} \sum_t \sum_s |P_t(s, 1) - P_t(s, 0)| \, \mathbf{1}\{h_S^*(s) \neq h_t^*(s)\}. \tag{8}$$

The sum weights each changed decision by the target margin, so a change close to a tie carries little cost. On sampled shots we write $\widehat{C}$ for the paired source-rule minus target-rule error. For adapted decoder a and its own frozen backbone f, we report

$$\Delta\text{LER} = \widehat{\mathcal{L}}(a) - \widehat{\mathcal{L}}(f), \qquad \widehat{R} = \frac{-\Delta\text{LER}}{\widehat{C}}. \tag{9}$$

The recovery ratio is descriptive and can exceed one because of decoder-specific error and sampling noise. It is not an estimate of a universally bounded fraction. We report it only on resolved-cost scenarios, alongside absolute LER changes and paired intervals. No ratio is formed from unresolved or negative sampled costs.

F Population integration, sampled estimates and the multiplier sweep

To compute $\mathcal{L}_T^*$ and C without sampling error wherever the outcome table is small, this appendix explains the population integration, states how schedules and random trajectories are handled, compares the finite-shot estimates with the population values, and reports the multiplier sweep. Equation 8 integrates C and $\mathcal{L}_T^* = \sum_s \min(P_T(s, 0), P_T(s, 1))$ directly. Deterministic piecewise-constant schedules are weighted by segment length, including a short final segment; separate noise configurations are averaged with their evaluation weights; random trajectories (the Ornstein–Uhlenbeck scenario, S06) are not integrated, and their costs are reported from sampled shots only. The population analysis records identify the point, scenario, number of distinct tables and exact sums.

Table F.1 gives the absolute costs behind the shallow-setting comparison in Section 5.2. In absolute LER the ratio of the smallest structural cost to the largest grid cost falls from $3.9\times$ at $d = 3$, $r = 1$ to $1.15\times$ at $d = 5$, $r = 1$. Two things move. A fixed-width defect occupies a smaller fraction of a larger code, so the correlated pair's cost drops from 2.2×10^{-3} at $d = 3$, $r = 2$ to 0.3×10^{-3} at

Table F.3: Finite-shot $\widehat{C}/\widehat{\mathcal{L}}_T^*$ at the four exactly tractable settings, 512,000 shots per scenario.

setting	max. grid	static hotspot	moving hotspot	correlated pair	aligned pair
$d = 3, r = 1$	1.42%	20.6%	14.7%	23.9%	24.5%
$d = 3, r = 2$	2.69%	19.6%	16.5%	29.0%	55.2%
$d = 3, r = 3$	3.76%	23.4%	16.8%	35.6%	68.7%
$d = 5, r = 1$	3.22%	21.2%	16.5%	28.4%	91.5%

Table F.4: Integrated and empirical shift costs on the reference settings, in 10^{-3} LER. Empirical intervals use paired episode resampling; population sums have no shot-sampling uncertainty.

scenario	population C	empirical $\widehat{C}$ [95% CI]
Static hotspot	3.533	3.357 [2.106,4.608]
Correlated pair	2.212	1.953 [1.312,2.625]
Aligned pair	18.683	16.998 [14.069,19.989]
Moving hotspot	2.944	2.563 [1.526,3.601]

$d = 5$, $r = 1$; and the grid maximum is always attained at multiplier 3, on two or three axes at $d = 3$ and on the after-Clifford axis alone at $d = 5$, $r = 1$, at a Bayes risk five to eight times the pair's. Relative to that risk the separation stays at least fourfold.

The finite-shot estimates of the main text use 512,000 shots per configuration and range over 15–92% for the structural scenarios, from the moving hotspot at $d = 3$, $r = 1$ to the aligned pair at $d = 5$, $r = 1$ (Table F.3). At $2\times$ the local arm of Table F.5 has 18 times the regret of the global arm at 40% of its risk; at $5\times$ the row costs 0.51 percentage points. By mechanism, the finite-shot costs over the four settings are 19.6–23.4% for the static hotspot, 15–17% for the moving hotspot, 24–36% for the correlated pair and 25–92% for the aligned pair; the aligned pair's range spans the $d = 3$ parity twin and the identifiable $d = 5$ mechanism of Appendix G.1. The finite-shot estimates agree with the population values of Table F.2: the grid maximum stays below 4% and the structural values span 15–92%. They are a different evaluation from the 32,768-shot reference suite of Tables F.4 and L.3, which covers $d = 3, r = 2$ only, so the two must not be compared cell by cell.

Deeper circuits are covered by exact per-shot decisions (Appendix H); larger codes use a true-model versus source-model matching surrogate that separates sensitivity to the evaluated mechanisms but cannot be read as C (Appendix R).

Population integration removes sampling error from the reference costs without making a sampled decoder LER exact. At $d = 5$, $r = 1$, for example, $C = 0.0002644$ for the Clifford-only multiplier 3 and 0.0002139 for MCZ at 3; the latter can win a finite-shot sweep without being the population maximizer. At $d = 3$, $r = 2$ all 75 grid configurations have strictly positive integrated C, so a shot-based interval covering zero means an unresolved effect, not an absent one (Table F.2).

The reference streaming evaluation uses the same finite episodes for decoder errors and $\widehat{C}$. Four raw estimates in the 18-scenario suite are negative through sampling error, S02, S09, S11 and S18; they are retained in the analysis records and never used as denominators. Population regret is nonnegative; the estimator need not be (Table F.4).

In the local/global multiplier experiment both arms start at $p = 0.005$ and use the same multiplier; the local arm affects only the center data row. Table F.5 reports population risk and regret. Uniform rate scaling changes a mechanism's log-odds by

$$w(mp) - w(p) = -\log m + \log \frac{1 - mp}{1 - p},$$

whose second term depends on p; this holds per circuit fault, and DEM mechanisms that merge several circuit faults with one footprint combine their probabilities as $p_1 + p_2 - 2p_1p_2$, so the multiplier does not act uniformly even on mechanism log-odds. A common additive shift can still reorder paths of different lengths, and exact decoding sums over degeneracy. The small grid costs are numerical results, not an invariance theorem.

Table F.5: Global and center-row multiplier sweep at $d = 3, r = 2$. The suffix identifies the multiplier; Bayes risks are population sums in plain LER and C is in units of 10^{-3}. The source rate is 0.005 for both arms.

configuration	target Bayes risk	C
global $\times 1$	0.01067	0.0000
global $\times 1.2$	0.01506	0.0002
global $\times 1.5$	0.02281	0.0027
global $\times 2$	0.03839	0.0320
global $\times 2.5$	0.05664	0.1024
global $\times 3$	0.07684	0.2241
global $\times 4$	0.12070	0.6675
global $\times 5$	0.16604	1.3096
global $\times 7$	0.25138	2.8420
local $\times 1$	0.01067	0.0000
local $\times 1.2$	0.01176	0.0197
local $\times 1.5$	0.01326	0.1564
local $\times 2$	0.01554	0.5630
local $\times 2.5$	0.01762	1.1290
local $\times 3$	0.01953	1.8192
local $\times 4$	0.02304	3.3885
local $\times 5$	0.02621	5.1480
local $\times 7$	0.03187	8.9371

G STRENGTH SWEEP OF THE STRUCTURAL MECHANISMS

To find the strength at which each structural mechanism begins to change the optimal decisions, this appendix sweeps one mechanism at a time at $d = 3$, $r = 2$ with population-exact references, and explains why the aligned pair is the extreme case (Section G.1). The sweep behind Figure 2 keeps the base noise at $p = 0.004$ and the source rule at $p = 0.005$, and varies the strength of one structural mechanism at a time at $d = 3$, $r = 2$, with every quantity integrated over the exact outcome table. With no structural mechanism, the base circuit at $p = 0.004$ against the source rule at $p = 0.005$ has $C = 1.1 \times 10^{-6}$, or 0.016% of $\mathcal{L}_T^*$.

Three mechanisms are swept over seven footprints. Hotspot footprints are the benchmark center row (S13), the benchmark row sweep of the moving hotspot (S14; rows weighted by their time shares 192/192/128 of 512 shots), one center data qubit and one corner data qubit, at multipliers 1.2 to 35. The XX pairs are the benchmark horizontal pair (S15) and the aligned vertical pair (S17) at event probabilities 0.0005 to 0.03 per round boundary. The leakage-like fault has within-shot persistence: one data qubit, chosen uniformly, enters a leaked state at a random round boundary and is fully depolarized at every later boundary of the shot; the target model is the exact mixture over the 18 (qubit, onset) components with onset probability 10^{-4} to 10^{-2} per qubit and boundary, and the surrogate models neither spreading to neighbors nor leakage-specific readout.

Two checks apply to the sweep. Each frozen decoder of the paper is queried once on all 2^{16} syndromes, so its LER under any target table is a population value as well. The benchmark points reproduce the population values of Appendix F to 10^{-9}, and every decoder's decisions are the same whether queried in one batch or on a subset.

Tables G.1 and G.2 list the sweep. Four predictions were written down before the run. Monotone cost in strength held for every mechanism, but the row footprints already cost 0.18–0.22% at $1.2\times$, above the 0.1% predicted for the weakest strength. The prediction that the aligned pair would cross the 4% comparison line at a lower event probability than the horizontal pair failed: the horizontal pair crosses at 0.0075 and the aligned pair at 0.012, where its cost then rises steeply, because the pair's decision flips once its probability exceeds that of the single error it mimics. The rank correlation between C and the six source-matched decoders' excess (predicted ≥ 0.8) is 0.96, and the leakage-like fault stays below 4% up to 10^{-3} onset probability as predicted, and up to 10^{-2} in fact. Crossing strengths are interpolated between adjacent sweep points in log-strength and log-cost coordinates; they are descriptive crossings of 4%, not exact zero-cost thresholds. The crossing strengths are $2.1\times$ for the center row, $2.2\times$ for the moving row, $2.4\times$ for one center qubit and $4.5\times$ for one

Table G.1: Strength sweep of the hotspots at $d = 3$, $r = 2$: exact population $C/\mathcal{L}_T^*$ in percent (first line of each row) and the mean population excess LER over the target-optimal decoder of the six source-matched frozen decoders, also in percent of $\mathcal{L}_T^*$ (second line). Base $p = 0.004$, source $p = 0.005$; the benchmark multiplier is 5.

footprint	$\times 1.2$	$\times 1.5$	$\times 2$	$\times 3$	$\times 5$	$\times 10$	$\times 20$	$\times 35$
hot center row (S13)	0.22	1.24	3.71	9.49	20.1	39.9	70.5	107.1
frozen excess	2.07	3.34	6.24	12.9	25.6	50.9	92.9	142.9
moving row (S14)	0.18	1.05	3.18	8.11	15.7	25.9	36.4	49.1
frozen excess	1.93	2.90	5.23	10.6	19.2	32.3	49.4	71.3
one center qubit	0.10	0.77	2.55	7.27	16.8	32.2	51.0	69.6
frozen excess	1.85	2.61	4.52	9.53	19.6	36.2	57.0	78.3
one corner qubit	0.03	0.08	0.50	1.72	5.10	11.0	15.9	17.2
frozen excess	1.74	1.81	2.26	3.55	7.08	13.3	18.5	20.4

Table G.2: Strength sweep of the XX pairs and of a leakage-like fault at $d = 3$, $r = 2$, as in Table G.1. Pairs: event probability per round boundary; the benchmark values are 0.03 (correlated, S15) and 0.025 (aligned, S17). Leakage-like: one data qubit, uniformly chosen, is fully depolarized at every boundary from a random onset boundary to the end of the shot; the strength is the onset probability per qubit and boundary.

pair, event probability	0.0005	0.001	0.002	0.004	0.008	0.015	0.025	0.03
correlated pair (S15)	0.09	0.22	0.55	1.57	4.41	11.8	23.3	29.1
frozen excess	1.97	2.28	2.96	4.68	8.96	18.9	34.5	42.4
aligned pair (S17)	0.01	0.01	0.03	0.11	0.50	11.3	55.7	78.4
frozen excess	1.49	1.32	1.09	0.85	0.94	11.6	55.9	78.5
leakage-like, onset probability	0.0001	0.0003	0.001	0.003	0.01			
any data qubit	0.02	0.03	0.13	0.64	2.55			
frozen excess	1.73	1.78	1.99	2.76	5.26			

corner qubit, so the benchmark strengths exceed their thresholds by $2.4\times$, $2.3\times$, $4.0\times$ (horizontal pair) and $2.1\times$ (aligned pair). No hardware range is drawn on Figure 2: measured relaxation-time ratios, residual couplings and leakage rates have no one-to-one calibration to these Pauli multipliers and event probabilities.

G.1 Why the aligned pair costs what it does

At distance three, one representative logical operator is $X_L = X_3X_{10}X_{17}$. Consequently $X_{10}X_{17} = X_3X_L$. Multiplication by a logical operator leaves every stabilizer syndrome unchanged and reverses the relevant logical parity. The pair and the single error thus give the same observation with different logical labels. If both noise explanations are admitted, syndrome observations alone cannot determine their relative weights beyond the assumptions encoded in a prior. This proves the observation ambiguity, not a universal one-half recovery bound.

At $d = 3$, $r = 2$ the pair fires at two round boundaries, with DEM footprints $\{1\}$ and $\{13\}$ of even logical parity, which the source model carries only with odd parity: a parity twin. Of the target mass on syndromes whose decision moved, 73% sits on the pair's own footprint syndromes (39% on $\{1\}$, 34% on $\{13\}$). At $d = 5$, $r = 1$ the pair's footprint $\{4, 5\}$ is absent from the source model, the source rule decodes it correctly, and the pair's own syndrome is not among the moved decisions; 78% of the moved mass comes from shots in which the pair fired, and these syndromes carry one (54%), two (22%) or three (16%) further detectors beyond the pair's footprint. The cost at $d = 5$ is the source rule's preference for fewer errors of the opposite parity when the pair combines with further errors, and it is tied to a footprint the source model lacks, the same class as the correlated pair at $d = 3$, which the library method recovers. The two settings reproduce the population values $C = 1.868 \times 10^{-2}$ (55.7% of the Bayes risk) and 2.135×10^{-3} (86.5%) of Table F.2.

Thresholds across code size. Table G.3 repeats the weak end of the sweep at the other exact settings. The hotspot and the horizontal pair keep their thresholds: one center qubit crosses 4% at

Table G.3: Weak-strength structural faults across code size: $C/\mathcal{L}_T^*$ in percent. $d = 3$, $r \leq 3$ and $d = 5$, $r = 1$ are population sums over the dense table; $d = 5$, $r = 5$ values are posterior-weighted estimates from exact per-shot decisions on sampled shots (2,048–4,096 shots; 95% block-bootstrap interval in brackets). A dash marks a value not computed. An interval of [0.0, 0.0] means that no differing decision was sampled; it is not a bound that the cost is zero (with no differing decision in 2,048 shots, the rate of differing decisions is below about 1.5×10^{-3} at 95%).

	aligned pair, event prob.			correlated pair	one center qubit		
setting	0.002	0.004	0.008	0.004	$\times 1.5$	$\times 2$	$\times 3$
$d = 3, r = 2$	0.03	0.11	0.50	1.57	0.77	2.55	7.27
$d = 3, r = 3$	0.17	0.39	2.65	2.47	1.11	2.89	7.59
$d = 5, r = 1$	7.00	13.4	25.2	1.45	0.48	1.68	4.69
$d = 5, r = 5$	1.6	22.4	81.6	0.7	0.0	0.6	0.0
95% interval	[0.0, 4.0]	[9.6, 27.8]	[61.9, 100.8]	[0.0, 1.6]	[0.0, 0.0]	[0.0, 1.5]	[0.0, 0.0]

$2.4\times$ at $d = 3$, $r = 2$ and near $3\times$ at $d = 5$, $r = 1$, and the horizontal pair at 0.0075 and 0.008. The aligned pair does not: at $d = 5$, $r = 1$ it costs 1.4% of the Bayes risk at event probability 0.0005 and 7.0% at 0.002, against 0.01% and 0.03% at $d = 3$, $r = 2$, so its interpolated crossing falls by about an order of magnitude, to near 0.001. In these evaluated settings the aligned pair crosses the comparison level at a lower strength, while local-rate mechanisms show a different dependence on distance and round count. At $d = 5$, $r = 5$ the aligned pair crosses 4% between event probabilities 0.002 and 0.004 (1.6% and 22%), a threshold about five times lower than at $d = 3$, $r = 2$ but above the $d = 5$, $r = 1$ value, while one center qubit stays below 1% up to $\times 3$ and the horizontal pair at 0.004 costs 0.7%: the measured response depends on the mechanism, distance, and round count together.

H EXACT REFERENCES AT DEEPER CIRCUITS

To extend exact decoding beyond the dense-table regime, this appendix describes the per-shot inference engine, its validation against the dense tables, the posterior-weighted estimators and the stopping rule, and then reads the values at $d = 5$ and $d = 7$. The dense engine rejects requests beyond its configured table-size limit instead of approximating them. Dense tables stop at 25 outcome bits, but exact maximum-likelihood decisions do not: contracting the detector error model one mechanism at a time, in an order that sweeps time, keeps only the detectors that are still open, and at $d = 5$ that frontier stays at 26 detectors for every round count from 5 to 25, so the frontier, and with it the memory, does not grow with r; the measured time per shot grows faster than r (about 0.9 s at $r = 5$ and 22 s at $r = 25$). At $d = 7$ the frontier is 20 detectors at $r = 2$ and 50 at $r = 7$, which is out of reach.

Two checks guard every run. The engine reproduces the dense tables to 6.5×10^{-15} with identical decisions at $d = 3$, $r \leq 3$ and $d = 5$, $r = 1$, and every run below includes an identity check in which source and target model coincide and no decision may differ.

Sampling uncertainty enters only through the stream average. On sampled shots it gives the exact posterior $P(y \mid s)$ under both models, and we report the posterior-weighted estimates $C_{rb} = \mathbb{E}[\mathbf{1}\{h_S^*(s) \neq h^*(s)\}\,|2P(y{=}1 \mid s) - 1|]$ and $\mathcal{L}_{rb}^* = \mathbb{E}[\min_y P(y \mid s)]$, which remove the label-sampling noise, with block-bootstrap intervals over 512-shot blocks. Runs stop once the interval half-width falls below $\max(10^{-3},\, 0.25\, C_{rb})$ or at a shot cap, so the intervals rest on 2 to 13 blocks; at $d = 7$, $r = 2$ the runs used tighter thresholds (10^{-4} and 0.1), stopped at their 1,024-shot cap with two blocks, and those intervals are indicative only. At $(7, 2)$ the label-based estimate of the grid cost has the opposite sign (-2.9×10^{-3}) to the posterior-weighted one.

One row predates this estimator. The $d = 3$, $r = 8$ row of Table 1 is the exception: it predates the posterior-weighted estimator and reports label-based $\widehat{C}/\widehat{\mathcal{L}}^*$, from a separate 32,768-shot sample for the grid, static-row and pair columns and from the 32,768 reference-suite shots for the moving row, whose ratio interval was not recorded.

Table 1 lists the values. At $d = 3$, $r = 8$ and $d = 5$, $r = 5$ the measured costs remain mechanism-dependent: the largest evaluated grid point reaches 4.4% and 5.8% of its Bayes risk, slightly above the dense-setting maximum of 3.8% and growing with circuit depth, as the composition axes do in the round sweep of the dense tier, while the structural scenarios cost 19–173%. The aligned pair costs more than its own Bayes risk at $d = 5$, $r = 5$. The $d = 5$, $r = 25$ row is the deepest setting of the Sycamore memory experiment on which neural decoders have been evaluated (Bausch et al., 2024), with noise of our uniform source model rather than that device's; there the grid point is $\times 2$ on the measurement, Clifford and Z-bias axes, because $\times 3$ saturates the code at 25 rounds (matching LER 0.47). At $d = 7$, $r = 2$ the grid rises further: the configurations with all four axes at $\times 2$ and $\times 3$ cost 3.9% and 6.6%, the Clifford axis alone at $\times 3$ 1.4%, and the worst point, CIZ$\times 3$, 9.1%. The structural scenarios there are not resolved: at their benchmark strengths the source and target rules disagree on at most one of 1,024 sampled syndromes. Both their population cost and its ratio to Bayes risk remain unresolved by this sample. Resolving it needs far more shots or sampling stratified by the number of fault events, which we leave open.

I RELEASED PRE-DECODERS AGAINST THE EXACT REFERENCES

To test a decoder that is neither ours nor a lookup table, we evaluate NVIDIA's released Ising pre-decoders (Chamberland et al., 2026), a 3D convolutional network that removes local errors before PyMatching, with the published weights and inference only (the model named Accurate in the public release unless stated). Two alignments are needed. Of the released circuit orientations only one uses the same two-qubit gate order as our circuits (its bulk detector error model coincides with ours), and the model was trained on circuits whose final round is noiseless, so an r-round NSBench shot is presented as $r + 1$ rounds with the final data measurement in the noiseless slot. With these, the adapter reproduces the released evaluation module bit for bit, a zeroed network reproduces PyMatching on 10,000 of 10,000 shots at every distance and orientation (the QAdapt adapter repeats this check on 2,048 shots, in the supplied records), and in separate validation samples on our source noise the pre-decoder lowers matching's LER to 0.90, 0.74 and 0.71 of its value at $d = 3$, $r = 3$, $d = 5$, $r = 5$ and $d = 7$, $r = 7$; on the 512 no-shift shots of Table I.1 the ratio is 0.88, and on the 32,768 of Table I.2 0.73.

Table I.1 places it on the shots of Appendix H, where the exact rules are known. Without shift it sits 17% above the optimum, closer than matching. Under the aligned and correlated pairs it stays at or behind the exact source rule (26.4 against 23.9 and 8.7 against 6.6 per thousand), so the network recovers none of the shift cost, and under CIZ$\times 3$ it is worse than both the source rule and plain matching. It also blocks the classical correction: re-weighting the matching graph to the true model lowers plain matching's LER on the aligned pair by 19.5×10^{-3} but the pre-decoder's by 3.9×10^{-3}, a fifth. At $d = 7$, $r = 7$ (Table I.2), where no exact reference exists, the same pattern holds on 32,768 shots per scenario: gains under rate and single-axis shifts (0.87–0.90 of matching in the table, 0.66–0.93 on the other single axes and 0.79–0.82 under the rate drifts), losses under three-axis composition at $\times 3$, a hot row whose gain reverses with strength (0.71 of matching at $\times 2$, 1.34 at $\times 10$, while true-model matching improves to 0.67), and a fifth of the aligned-pair re-weighting gain surviving behind the network (0.49 for true-model matching alone, 0.70 behind the pre-decoder). A frozen pre-decoder trained on uniform noise therefore behaves like the source rule under structural shift and removes the syndrome evidence that model-aware matching would use.

QAdapt (Miao et al., 2026) is the published adaptive system closest to this benchmark, and its release lets us test it directly. Its adaptation is offline: an HTNet pre-decoder trained sequentially over five noise tasks with elastic weight consolidation (Kirkpatrick et al., 2017), released as one frozen checkpoint (receptive field 9, released configuration at $d = 9$, 25-parameter Pauli model with 1% measurement error) together with a patch to the pinned Ising revision; training code is not released.

At test time it is a stateless pre-decoder, so under our contract it is a frozen decoder, and it runs through the same adapter (same orientation; the pinned revision reproduces the recorded Ising LERs exactly on every $d = 5$ case and to one shot in 32,768 at $d = 7$).

At $d = 7$, $r = 7$ it tracks Ising under no shift, rate and single-axis shifts (0.74–0.90 of matching), is worse than matching under three-axis composition at $\times 3$ and under the strong hot row, and is weaker than Ising on the structural pairs: 0.94 and 0.96 of matching on the correlated and aligned

Table I.1: The released Ising (Accurate model) and QAdapt pre-decoders, each followed by matching, against exact decisions at $d = 5$, $r = 5$, on the same sampled shots: logical error rate $\times 10^3$, with the excess over $\widehat{\mathcal{L}}^*$, the empirical error of the exact target rule on the same shots, in parentheses. Source-optimal is the exact rule of the training model. MWPM, Ising and QAdapt use the source matching graph, and the columns marked T the target's.

scenario	shots	$\widehat{\mathcal{L}}^*$	source-optimal	MWPM	Ising	QAdapt	MWPM$_T$	Ising$_T$	QAdapt$_T$
no shift	512	11.7	11.7 (+0%)	15.6 (+33%)	13.7 (+17%)	15.6 (+33%)	15.6 (+33%)	13.7 (+17%)	15.6 (+33%)
grid C×3	2,048	79.6	82.5 (+4%)	110.4 (+39%)	100.6 (+26%)	103.5 (+30%)	112.3 (+41%)	100.1 (+26%)	104.5 (+31%)
grid CIZ×3	2,048	110.8	130.4 (+18%)	136.7 (+23%)	146.5 (+32%)	153.8 (+39%)	135.3 (+22%)	147.0 (+33%)	153.8 (+39%)
correlated pair	6,656	4.4	6.6 (+52%)	12.3 (+183%)	8.7 (+100%)	9.5 (+117%)	9.2 (+110%)	8.6 (+97%)	9.3 (+114%)
aligned pair	2,048	8.8	23.9 (+172%)	35.2 (+300%)	26.4 (+200%)	33.2 (+278%)	15.6 (+78%)	22.5 (+156%)	26.9 (+206%)

Table I.2: The Ising and QAdapt pre-decoders at $d = 7$, $r = 7$ (32,768 shots per scenario): matching LER from the source graph, and each decoder's LER as a ratio to it, with the 95% episode-bootstrap interval of each pre-decoder's LER difference from MWPM, in 10^{-3}. Columns marked T use the target's matching graph and are defined for static scenarios; the hot row $\times 10$ comes from the second, independently sampled round.

scenario	MWPM (10^{-3})	Ising	Δ [95%]	QAdapt	Δ [95%]	MWPM$_T$	Ising$_T$	QAdapt$_T$
no shift	9.8	0.73	[-3.8, -1.6]	0.74	[-3.6, -1.4]	1.00	0.73	0.74
rate ×2	99.3	0.90	[-13.6, -7.0]	0.90	[-12.7, -6.3]	1.01	0.90	0.91
grid C×3	147.5	0.87	[-23.8, -14.6]	0.88	[-21.5, -14.5]	1.01	0.88	0.89
grid MCZ×3	172.3	1.10	[+12.4, +21.2]	1.04	[+3.5, +11.7]	1.01	1.11	1.05
grid CIZ×3	207.1	1.04	[+2.9, +14.4]	1.05	[+5.6, +15.2]	1.02	1.05	1.06
static hotspot	14.7	1.04	[-0.9, +2.1]	1.09	[-0.2, +2.7]	0.76	1.00	1.05
hot row ×10	29.4	1.34	[+7.9, +12.5]	1.32	[+7.3, +11.7]	0.67	1.20	1.13
correlated pair	7.0	0.83	[-2.0, -0.5]	0.94	[-1.3, +0.4]	0.72	0.83	0.95
aligned pair	14.3	0.80	[-4.1, -1.6]	0.96	[-1.9, +0.9]	0.49	0.70	0.84

pairs, intervals including no gain. At $d = 5$, $r = 5$, below its training size, it gains nothing without shift, stays behind the exact source rule under both pairs (33.2 against 23.9 and 9.5 against 6.6 per thousand) and is the worst decoder under CIZ×3; behind it, true-model re-weighting recovers a third ($d = 5$) and a quarter ($d = 7$) of the aligned-pair gain of plain matching.

Its training tasks change noise-axis multipliers by 1.5; the structural scenarios are outside that family, which is the regime in which the shift cost is large. We read these as two released decoders at their training distributions, not a statement about pre-decoding in general.

Three remarks apply to Table I.1. Every cell is an empirical rate on the listed shots, so the source-optimal excess is the label-based shift cost. The posterior-weighted estimate on the same shots (Table 1) is close for the two pairs (2.95 against 2.25 and 16.7 against 15.1×10^{-3}, each inside the label-based interval) and is lower for CIZ×3 (7.4 against 19.5×10^{-3}), a label-noise difference of about 2.6 binomial standard errors of the 88 differing decisions (2.2 under the block bootstrap). The percentages are relative to the label-based target-optimal error, which on the CIZ×3 shots is 110.8×10^{-3} where the posterior-weighted Bayes risk is 126.4×10^{-3}.

J STATIC GRID AND AXIS-DRIFT EVALUATION

To state what the static grid and the axis-drift scenarios contain, this appendix defines the four noise axes and their multipliers and summarizes the evaluation on both. The four axes are measurement flips (M), after-Clifford depolarization (C), data-idle depolarization (I) and a Z bias (Z) that grows the pure-Z components of every depolarizing channel as if they were multiplied (a phase-flip and ZZ-crosstalk bias); the letters identify channels, not Pauli operators. Every nonempty subset is scaled by each of five multipliers.

The Z axis needs one remark. Pure-Z noise cannot flip Z_L, but the exact rule reads both detector types, and X-type detections are evidence for Y errors, which do flip it; changing the Z share re-weights that evidence. The Z axis alone therefore moves the decision slightly (0.4% and 1.5% of Bayes risk at $\times 3$ for $d = 3$ with $r = 2$ and $r = 3$) and adds to the other axes in the largest grid points; the Z-type overlays of Table D.1, which act on data qubits only, remain small-cost controls.

The 75 configurations are the 15 nonempty subsets times five multipliers: QAdapt's 11 multi-axis subsets, which that work evaluates at two larger distances (Miao et al., 2026), plus the four single axes; distance-portable selectors retarget local fault supports for within-point comparisons.

One boundary applies. The grid varies the strength of four unbiased channels; a change of Pauli bias, which is known to change the optimal decoder (Tuckett et al., 2018), enters only as extra Z noise on the Z axis, which in Z memory lies in the unmeasured sector, so decision preservation is claimed for multiplier grids, not for bias shifts.

The grid evaluation includes four adapter variants, 21 backbone/seed combinations and 75 configurations: 6,300 streamed evaluations with no causality-audit failure. The three history contexts have median LER about 0.0198–0.0199. The known-noise adapter is unstable outside its training support: 529 of 1,575 cells exceed the largest LER of the history contexts, across 40 measurement-axis configurations, reaching 0.8844 at the worst cell, because its measurement-multiplier field is constant in meta-training. These failures stay in the record; an in-support summary does not turn the known-noise adapter into an upper bound. A causal audit passes such a decoder because it checks future dependence, not calibration; the audit and the accuracy tests answer separate questions.

Five additional trajectories vary the same axes through time. Their sampled source-rule cost is at most 0.00025 while the largest sampled target-optimal LER reaches 0.0214: a small-cost result at the reference code. Schedules are in the code archive; per-configuration metrics are kept in the research repository.

K Training and implementation details

Table K.1: Decoder families and evaluation-time information access. All deployable adaptive methods use unlabeled syndromes only.

Group	Method family	Implementations	Evaluation-time behavior
Neural decoders	Trained on source noise	MLP, 2D-CNN, Transformer, TCN, GNN; 3D-CNN and neural-BP (Yan et al., 2026)	Frozen
	Released pre-decoder, then matching	Ising Accurate (Chamberland et al., 2026); QAdapt HTNet (Miao et al., 2026)	Frozen, published weights
Classical	Graph-based decoding	MWPM (Higgott, 2022), Union-Find (Delfosse & Nickerson, 2021), BP+OSD (Roffe et al., 2020)	Fixed source noise model
Adaptive	Neural	History FiLM adapter; TENT	Causal updates
	Classical	Matching reweighting	Causal updates
Adaptive reference	Bayesian filtering	Exact Bayesian filter	Causal; specified prior
Exact reference	Optimal rules	Source-optimal h_S^*; target-optimal h_t^*	Known noise; not deployable

To make the decoder comparison reproducible, this appendix reports source performance, training budgets and the implementation details of every neural and classical decoder. Six implementations reach source validation LERs of 0.0108–0.0111; neural-BP reaches 0.0137 and is retained in every table, so its structural-scenario excess includes a source-fitting deficit. Trained for ten times the steps (150,000, three seeds, same shots) it reaches 0.0116 ± 0.0001, 8.5% above the exact optimum and 7% above the best matched-budget implementation, down from 26% above that implementation (28% above the optimum) at the matched budget: most of the deficit is optimization cost, and the tables keep the matched-budget runs. The benchmark's structural scenarios are in-family instances of the meta-training family: their sites are among the family's random sites and their strengths lie inside the family's continuous ranges, so the benchmark draws are unused episodes rather than unseen configurations. Leave-one-class-out runs remove a class, with the caveat of Appendix O for the aligned pair. A *known-noise* adapter with the same mechanism receives the true per-shot parameters as a diagnostic.

The in-house models are an MLP over the detector vector, a 2D-CNN with time as channels, a Transformer over detector tokens, a causal TCN over time slices, and a detector-graph GNN. The ported 3D-CNN uses a stabilizer-type/time/row/column volume; neural-BP passes messages between check nodes and DEM fault mechanisms for $N = d$ iterations, one of the two settings of Yan et al. (2026). Equal parameter or step budgets do not equalize computational cost or training adequacy (Table K.3).

Table K.2: Methods and the information each may read at prediction time, beyond the current syndrome.

method	reads	what changes
frozen neural / classical	source training or calibration only	nothing
history FiLM; TENT (Wang et al., 2021)	prior syndromes of the episode; unlabeled test syndromes	feature modulation; normalization or bias parameters
adaptive matching	syndrome statistics, candidate fault library	matching weights and edges
known-noise FiLM (diagnostic)	true per-shot noise parameters	feature modulation
Bayesian filter; target-optimal Bayes	syndrome history and a specified prior; the true noise segment	posterior over noise hypotheses; the exact optimal decision

Table K.3: Implementation budgets and source validation LER at $d = 3, r = 2$. CPU inference time per shot: batched uses batches of 2,048 and all threads; streaming uses one thread and includes causal context updates. These are software measurements, not hardware latency claims. Input representations and implementation origins are described in the text.

implementation	parameters	adapter	source LER	batch ms	stream ms
MLP	95,701	23,950	0.0109	0.002	2.26
2D-CNN	79,873	23,994	0.0111	0.007	2.89
Transformer	105,985	23,994	0.0109	0.016	4.36
TCN	89,601	24,064	0.0109	0.006	4.03
GNN (detector graph)	94,849	23,976	0.0111	0.011	3.43
3D-CNN	121,929	24,012	0.0108	0.404	4.56
neural-BP GNN	75,395	23,994	0.0137	0.670	5.64

The backbones share 6M training shots at $p = 0.005$, three model seeds, and 15,000 optimizer steps at batch 2,048. FiLM sites: the first hidden layer for the MLP, convolutional features before spatial pooling for the 2D-CNN, tokens after the first Transformer layer, the first TCN convolution, detector features after the second GNN layer, the second 3D-CNN residual block, and variable-node states after the first neural-BP iteration. A standardized context passes through two SiLU hidden layers and an output layer producing scale and shift corrections, initialized to the identity.

The history context is 26 summaries, 16 running detector means, 16 long detector EWMAs ($\alpha = 0.01$), 50 running detector-pair co-firing means, and the elapsed shot count divided by 512. The all-pairs variant replaces the 50 pair features by 120, for 179 in total. The known-noise input contains true rates and position-specific fault parameters; it is a privileged-information comparison through the same modulation mechanism and is subject to the same optimization and support failures.

Meta-training uses 4,096 labeled episodes of 512 shots, 4,000 steps at batch 2,048 with AdamW, and an identical batch sequence across implementations; checkpoints are selected by loss on 512 separate validation episodes. Family training, validation and mixed-backbone seeds occupy disjoint 300000, 310000 and 320000 blocks, and the grid-family arm 330000 and 340000, distinct from benchmark seeds. Leave-one-class-out selection excludes the held-out class from validation as well as training. Benchmark episodes are never training examples. The research repository contains the full configurations.

Recipe sensitivity was measured. At $d = 3$, $r = 2$ the published 3D-CNN recipe yields validation LER 0.0108–0.0109 and similar full-suite performance. Neural-BP improves with more steps and its released recipe while remaining less accurate under these settings; the remaining deficit is not attributable to iteration count or architecture without the corresponding intervention.

L Complete frozen and adapted results

To give the complete results behind Sections 5.3 and 5.4, this appendix lists frozen and adapted error rates for every decoder on all 18 scenarios, with the paired statistics used in the main text. Figure L.2 compares the frozen implementations against a common scenario-specific reference; Ta-

Table L.1: The four appreciable-cost scenarios on the 512-shot reference episodes; $\widehat{C}$ in 10^{-3} LER. History adapter: range of $\widehat{R}$ over the seven implementation means and the number with a BH-significant gain. Bayesian filter: $\widehat{R}$ relative to the source-optimal rule with a paired 95% episode-bootstrap interval, then its posterior mass on the true mechanism class at shot 512 and the first shot at which that mass exceeds 0.9 (*: not sustained, see the text).

scenario	$\widehat{C}$	history $\widehat{R}$ (%)	sig. gains	filter $\widehat{R}$ (%) [CI]	$P(\text{class})_{512}$	shots to 0.9
Static hotspot	3.36	34 to 69	7/7	103 [94, 114]	1.00	51
Correlated pair	1.95	−34 to 28	0/7	98 [91, 106]	0.52	never
Aligned pair	17.00	1 to 42	5/7	49 [38, 60]	0.57	never
Moving hotspot	2.56	−29 to 3	0/7	48 [20, 73]	0.60	191*

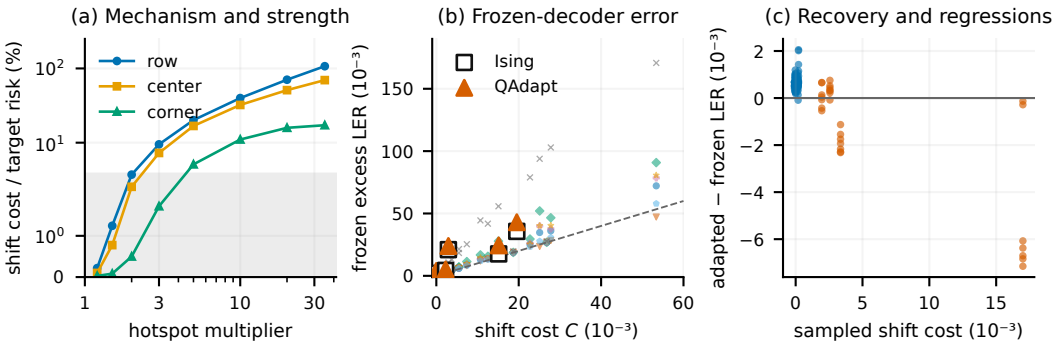


Figure L.1: **Three questions for noise-shift evaluation.** (a) How do mechanism and strength change the optimal rule? The band marks the largest dense-setting grid cost, 3.8% at $d = 3$, $r = 3$ (at the plotted $d = 3$, $r = 2$ it is 2.4%). (b) How much frozen-decoder error accompanies that change? The seven trained implementations use population values at the reference setting: the six source-matched ones as colored dots, neural-BP, whose source error is larger, as grey crosses. The two released pre-decoders use sampled distance-five values (Figure 3b). The diagonal marks equal excess error and shift cost. (c) Does adaptation recover error or introduce it? Positive values are regressions; orange marks the four appreciable-cost scenarios, blue the others. Panels use separate experiments, detailed in Figures 2–4 of the main text.

bles L.2 and L.3 list the values. All frozen LERs are estimated on the same reference episodes, so an empirical decoder value can fall slightly below the empirical target-optimal value without violating population optimality; comparisons use paired differences.

Three remarks apply to Table L.1. The Bayesian filter runs under a prior that contains the benchmark configurations, and values above 100% arise because the filter and target-optimal LERs are both estimates on the same shots. Adapter and filter columns have different references, and same-baseline gaps are in Table P.3. For the moving hotspot the averaged mass exceeds 0.9 only at shot 191 (peak 0.905), just before the row moves at shot 192, and is not sustained (trace in the supplied record identification_switching_v11.json).

Two cautions apply to Table L.5. Counts are not comparable across rows: the variants ran on different implementation sets, TENT rows count per-seed cells (backbone × seed × scenario) while the history rows count seed-averaged cells, and on the 90 in-house cells they share the three history variants regress on 58, 64 and 59. TENT cannot reach the cells marked §, four backbones having no normalization layer for it to modulate. It gets 256–896 parameters on the other three, the bias generalization reaches all seven, and neither exceeds 1.4×10^{-3} either way. Figure 4 names the history-adapter outcomes.

On the four structural scenarios the six source-matched frozen implementations carry 17–25% excess LER over the target-optimal decoder on the hotspots, 35–54% on the correlated pair and about 47% on the aligned pair; neural-BP adds a source-fitting deficit. Differences between scenarios

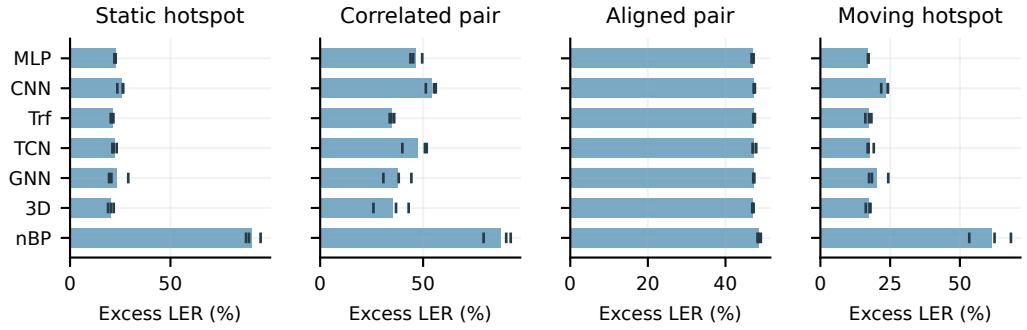


Figure L.2: Frozen-decoder excess LER on the four structural scenarios, relative to each scenario's empirical target-optimal LER. Bars are means over three training seeds; ticks are individual seeds.

Table L.2: Frozen neural LER at $d = 3, r = 2$, averaged over three seeds on shared shots. Scenario IDs are defined in Table D.1. These are Monte Carlo LER estimates.

scenario	MLP	2D-CNN	Transf.	TCN	GNN	3D-CNN	nBP
S01	0.0105	0.0109	0.0105	0.0105	0.0107	0.0105	0.0130
S02	0.0198	0.0201	0.0200	0.0200	0.0203	0.0196	0.0245
S03	0.0386	0.0392	0.0384	0.0383	0.0390	0.0383	0.0468
S04	0.0094	0.0093	0.0092	0.0092	0.0093	0.0092	0.0121
S05	0.0166	0.0167	0.0164	0.0165	0.0167	0.0164	0.0202
S06	0.0144	0.0147	0.0143	0.0144	0.0145	0.0143	0.0179
S07	0.0159	0.0159	0.0157	0.0159	0.0161	0.0158	0.0193
S08	0.0223	0.0222	0.0218	0.0223	0.0224	0.0217	0.0268
S09	0.0114	0.0115	0.0113	0.0113	0.0116	0.0113	0.0149
S10	0.0173	0.0171	0.0170	0.0171	0.0175	0.0169	0.0214
S11	0.0114	0.0115	0.0114	0.0113	0.0116	0.0113	0.0140
S12	0.0106	0.0107	0.0104	0.0106	0.0107	0.0105	0.0132
S13	0.0227	0.0232	0.0224	0.0226	0.0228	0.0223	0.0353
S14	0.0213	0.0224	0.0213	0.0214	0.0218	0.0213	0.0293
S15	0.0102	0.0107	0.0094	0.0103	0.0096	0.0094	0.0130
S16	0.0073	0.0073	0.0072	0.0073	0.0074	0.0073	0.0094
S17	0.0529	0.0530	0.0530	0.0530	0.0530	0.0529	0.0535
S18	0.0130	0.0129	0.0130	0.0128	0.0131	0.0128	0.0160

exceed differences between architectures within a scenario, and three seeds leave the architectures unranked.

Figure 4 shows every history-adapter cell. On the 14 small-cost scenarios the adapter worsens 97 of the 98 cells, 79 of them significantly, by a mean $+0.64 \times 10^{-3}$ and a median $+0.60 \times 10^{-3}$ LER. The 82 significant regressions are these 79 cells and three on the structural scenarios; 12 cells are significantly better. Of the 79, all 21 cells on the three scenarios whose measurement multiplier the training family never varies regress, and so do 58 of the 77 cells inside its support, including 4 of the 7 backbones on the no-shift control. One of the 79 (ou_global_drift, 3D-CNN; reported $p = 0.0370$) lies at the Benjamini–Hochberg cutoff, so an independent Monte Carlo draw of the sign-flip test can place it on either side and give 78; the supplied recheck script reports this cell. The largest regressions, up to about $+2 \times 10^{-3}$, fall on the measurement-dominated scenario, whose measurement multiplier the meta-training family never varies (Appendix J). The regression on the stationary control has the same size as on the drift scenarios, and the grid-trained variant of Appendix N worsens 56 of the same 98 cells. TENT leaves every cell unchanged at the significance threshold.

For the classical tiers, the circuit-enumerated library improves the correlated pair by 1.80×10^{-3} (paired 95% CI $[1.07, 2.56] \times 10^{-3}$), which is 92% of $\widehat{C}$ against its own frozen MWPM and 71% of the larger true-model matching headroom used in Table R.2, and worsens the static hotspot by

Table L.3: Frozen source-model classical decoders and target-optimal exact-Bayes decisions, evaluated on the same sampled shots as Table L.2. The exact column is an empirical LER, not a population risk.

scenario	MWPM	BP+OSD	Union-Find	target-optimal
S01	0.01169	0.01059	0.01590	0.01038
S02	0.02237	0.01987	0.02829	0.01956
S03	0.04288	0.03900	0.05200	0.03781
S04	0.01004	0.00912	0.01480	0.00879
S05	0.01855	0.01627	0.02267	0.01593
S06	0.01608	0.01465	0.02075	0.01422
S07	0.01767	0.01611	0.02231	0.01562
S08	0.02374	0.02185	0.03195	0.02170
S09	0.01334	0.01147	0.01718	0.01117
S10	0.01947	0.01712	0.02454	0.01660
S11	0.01273	0.01129	0.01694	0.01141
S12	0.01157	0.01062	0.01605	0.01025
S13	0.02972	0.02368	0.03259	0.01852
S14	0.02542	0.02213	0.02985	0.01819
S15	0.01111	0.00964	0.01382	0.00696
S16	0.00848	0.00751	0.01065	0.00708
S17	0.05377	0.05310	0.05481	0.03598
S18	0.01395	0.01285	0.01892	0.01285

1.56×10^{-3}. Four of the six deployable matching tiers have no significant regression on the 14 small-cost scenarios; raw correlation estimation and blind edge search regress significantly on 18 and 16 of the 18 scenarios. Neural and classical methods start from different frozen decoders, so their $\widehat{R}$ values compare each to its own reference.

Table L.4 gives every history-adapter cell on the structural scenarios, and Table L.5 the full-suite counts for every variant. Variants differ in the evaluated model sets, so the totals of Table L.5 are not one paired comparison. TENT's normalization-only form reaches three backbones with 256–896 parameters and the bias variant reaches all seven; neither form has a significant change under this protocol, which does not test the batched or fully trainable settings TENT was designed for. The known-noise adapters have 21 off-support cells on the measurement-dominated and composition scenarios; they are listed separately (‡) and excluded from the better/worse counts, and their regressions reach $+583 \times 10^{-3}$. The classical p_{ij} and library rows use one deterministic decoder per scenario (no model seeds) against frozen source-model MWPM; every tier estimates causally within the episode, as Appendix R specifies.

M Statistical estimands and reproducible inference

To make every reported gain and regression a paired, episode-level statement, this appendix defines the estimands, the sign-flip and bootstrap procedures, the multiple-comparison correction, and the reanalysis that reproduces the main-text numbers. For architecture a, seed k and episode e, let d_{ake} be adapted minus frozen mean error over the shots of that episode; the primary estimate is $\overline{d}_a$. Sign-flip testing uses the seed average of each episode and flips episode signs under a symmetric paired null, so dependent shots within an adapted stream are never treated as independent units. The reanalysis reproduces all 126 history-adapter mean differences from per-shot predictions and digest-checked labels.

Confidence intervals for a three-seed model mean resample seed and episode indices, sharing the episode draw across every quantity in a contrast; this preserves the covariance among adapted, frozen, source-optimal and target-optimal errors. For $\widehat{R}$ each draw computes its own numerator and $\widehat{C}$ before forming the ratio. The four reported denominators are large enough for percentile intervals; no ratio inference is attempted on the small-cost controls. For the deterministic Bayesian filter only episodes are resampled; for a filter–adapter gap the adapter seed dimension is resampled as well. Intervals are 95% percentile intervals from 20,000 draws with analysis seed 20260919 and are not simultaneous.

Table L.4: History-adapter LER and $\widehat{R}$ (%) for all seven implementations, three seeds. Intervals jointly resample episodes and seeds, including $\widehat{C}$. A star marks a significant LER change under the 126-test BH family (episode sign-flip test on ΔLER); the $\widehat{R}$ intervals also resample $\widehat{C}$ and are unadjusted, so a starred cell can have an $\widehat{R}$ interval that touches zero. Scenario IDs refer to Table D.1.

scenario	model	frozen LER	adapted LER	$\widehat{R}$ [95% CI]
S13	MLP	0.02268	0.02094	52.1 [33, 79]⋆
S13	2D-CNN	0.02323	0.02169	46.1 [24, 75]⋆
S13	Transf.	0.02240	0.02047	57.6 [37, 84]⋆
S13	TCN	0.02260	0.02044	64.5 [46, 89]⋆
S13	GNN	0.02279	0.02047	69.1 [51, 96]⋆
S13	3D-CNN	0.02230	0.02117	33.6 [19, 53]⋆
S13	nBP	0.03530	0.03301	68.2 [22, 134]⋆
S15	MLP	0.01017	0.00963	27.6 [-29, 83]
S15	2D-CNN	0.01073	0.01030	21.9 [-13, 53]
S15	Transf.	0.00938	0.00931	3.6 [-38, 44]
S15	TCN	0.01026	0.01030	-2.1 [-52, 40]
S15	GNN	0.00958	0.00971	-6.8 [-47, 28]
S15	3D-CNN	0.00941	0.01006	-33.3 [-83, 2]
S15	nBP	0.01304	0.01370	-33.9 [-80, -2]⋆
S17	MLP	0.05291	0.04683	35.7 [26, 45]⋆
S17	2D-CNN	0.05304	0.05275	1.7 [-9, 13]
S17	Transf.	0.05303	0.04588	42.1 [32, 52]⋆
S17	TCN	0.05302	0.04632	39.4 [29, 50]⋆
S17	GNN	0.05300	0.04616	40.2 [28, 51]⋆
S17	3D-CNN	0.05292	0.04654	37.5 [25, 48]⋆
S17	nBP	0.05349	0.05335	0.8 [-2, 4]
S14	MLP	0.02131	0.02174	-16.7 [-53, 8]
S14	2D-CNN	0.02244	0.02266	-8.7 [-41, 12]
S14	Transf.	0.02133	0.02185	-20.2 [-55, 0]⋆
S14	TCN	0.02142	0.02134	3.2 [-28, 26]
S14	GNN	0.02184	0.02218	-13.1 [-50, 12]
S14	3D-CNN	0.02133	0.02164	-11.9 [-44, 11]
S14	nBP	0.02935	0.03010	-29.4 [-82, 4]⋆

Table L.5: Every variant on all 18 scenarios; better and worse are BH-significant at $q = 0.05$ against the frozen backbone of the same seed. ‡ = known-noise context outside the family's support, excluded, regressions reaching $+583$. The last two rows are classical matching against frozen MWPM, two representative tiers of the six deployable ones (Appendix R). § = cells TENT cannot reach. Counts are not comparable across rows (see the text).

variant	cells	better	worse	worst regression, ΔLER ($\times 10^{-3}$)
history, 50 model pairs	126	12	82	meas.-dominant, $+2.05$
history, all 120 pairs	108	13	75	meas.-dominant, $+2.41$
history + per-position term	126	14	79	meas.-dominant, $+2.26$
known-noise (reference)	105 ($+21$ ‡)	25	7	rate $\times 2$ (TCN), $+0.55$
TENT, normalization affine	162 ($+216$ §)	0	0	rate $\times 2$ (3D-CNN, seed 2), $+0.55$
TENT, bias generalization	378	0	0	rate $\times 2$, $+1.28$
classical, p_{ij} re-estimation	18	0	0	static hotspot, $+0.98$
classical, circuit library	18	1	1	static hotspot, $+1.56$

Episode-length tables estimate the mean of the fixed five architectures at their three trained seeds and resample episodes only; each length has its own seed block, so length contrasts are unpaired. Selecting the best architecture on the same data introduces a selection effect, so every architecture is shown and any selected comparison is marked descriptive; a small point gap to the filter is not an equivalence result. The sign-only counts of the meta-training-family experiment are not multiplicity-

Table N.1: Grid-family minus structural-family adapter LER on the 75-point grid (10^{-3} units). Each mean pairs 1,575 configuration/model/seed cells. The cell-IID interval reproduces the original aggregate summary but ignores shared models; the block-bootstrap sensitivity interval resamples configurations and architecture/seed blocks. Neither is a formal equivalence test for a random population of circuits. The fixed margin is ± 1.0 in these units.

variant	mean	cell-IID 95% CI	block 95% interval	grid better
history (50 pairs)	-0.577	[-0.601, -0.553]	[-0.693, -0.472]	1463/1575
history (all 120 pairs)	-0.652	[-0.682, -0.621]	[-0.827, -0.506]	1476/1575
history + position term	-0.686	[-0.717, -0.655]	[-0.854, -0.544]	1489/1575

adjusted regression counts. Episode sufficient statistics and the ratio scripts are in the research repository.

N META-TRAINING SUPPORT AND BROADER OFFLINE TRAINING

To test whether the adapter's behavior depends on what it was meta-trained on, this appendix replaces the structural meta-training family by an axis-multiplier family, and reports leave-one-class-out and broader offline training. Replacing the structural meta-training family by 16 classes of axis subsets, including the empty subset, with continuous multipliers in $[1.2, 3]$ changes the exposure and nothing else: backbone, FiLM budget and evaluation episodes are fixed. Three contexts, seven architectures and three seeds were evaluated on all 18 scenarios with no causality-audit failure. For the 50-pair history context, structural-family versus grid-family mean recoveries over the seven architectures are 28.2% versus 0.1% on the aligned pair, 55.9% versus about 0% on the static hotspot, -13.8% versus -5.3% on the moving hotspot, and -3.3% versus -10.0% on the pair. The sign-only counts of worse-than-frozen outcomes on the 14 small-cost scenarios are 97/98 and 56/98.

The completed grid evaluation has 4,725 audited grid-family runs, three interfaces by 75 configurations by seven architectures by three seeds, each paired with its structural-trained counterpart by configuration, architecture and seed. Grid-family adapters improve 92.9–94.5% of the cells, with median LER about 0.0193 versus 0.0198–0.0199 (Table N.1). A local analysis plan fixed a ± 0.001 LER margin before these effect sizes were available, from the rounded-up maximum population grid regret at $d = 3$, $r = 2$; it applies to the per-interface aggregate. Both the cell-IID summary intervals and the block intervals that resample whole configurations and architecture/seed blocks lie inside it. The differences are consistently nonzero and favor grid training; they are small on the specified LER scale.

The planned per-cell interval widths have median about 0.00425 and maximum 0.00835 LER under an independent-binomial approximation; these are full widths that omit the covariance of paired correctness indicators and are retained only as a resolution diagnostic.

Mixed-noise training asks whether a broader frozen backbone can absorb the changes: 6M shots from the structural family at the same optimization budget. Table N.2 reports paired changes over the six source-matched implementations, with all-seven records in the code archive. It helps on the correlated pair and is worse than the history adapter on the static hotspot.

O INTERFACE INTERVENTIONS AND TEMPORAL BEHAVIOR

To separate the effect of the observation interface from the adapter itself, this appendix reports three controlled modifications of the context, the FiLM interface and its activation timing, and the temporal behavior of the adapter within an episode. Three controlled modifications change the context, the FiLM interface or its activation timing. Tracking all 120 detector pairs instead of 50 reference-graph pairs improves correlated-pair recovery by 16–43 percentage points across the five in-house architectures while often worsening hotspot recovery; at fixed parameter budget the larger input also narrows the encoder, so this is a budget-constrained interface change. A position-specific FiLM term raises the 2D-CNN's aligned-pair recovery from about 35% to 84% given true noise parameters, and from about 2% to 20% under the history context. A 64-shot identity gate reduces

Table N.2: Mixed-noise training against the stationary backbone and against adaptation ($\Delta\mathrm{LER} \times 10^{-3}$ = mixed − comparator, negative = mixed better; range over the six implementations with a matched source excess; $\star$ = BH-significant at $q = 0.05$).

scenario	vs stationary frozen (R_{mixed})	vs stationary + history adapter	vs stationary + known-noise adapter
correlated pair	-0.95 to -2.05, all $\star$ (48–105%)	-1.08 to -2.05, all $\star$	-0.01 to -1.41 ($\star$ for four)
static hotspot	-0.28 to -0.82 (8–25%; $\star$ GNN only)	+0.84 to +1.85, all $\star$	+0.79 to +2.22 ($\star$ for five)
moving hotspot	-0.03 to -0.82 (1–32%; $\star$ 2D-CNN only)	-0.02 to -1.05 ($\star$ for four)	+0.97 to +2.11 ($\star$ for five)
aligned pair	-0.03 to +0.17 (-1 to 0%)	+0.44 to +7.13 ($\star$ for five)	+6.06 to +15.70, all $\star$
14 rate / composition / cluster / control scenarios (84 cells)	mean +0.12, -0.13 to +0.53 (1 $\star$ worse)	mean -0.54; $\star$ better on 39, worse on 0	≈ 0 where the context is in support

significant regressions from 58 to 30 on the 90 in-house cells at a cost of 3–7 points of static-hotspot recovery. Per-cell contrasts are in the analysis records.

Independent probes show that static-hotspot identification and adaptation both improve as evidence accumulates: across twelve additional hotspot configurations, identification time and recovery time have Spearman correlation 0.80 on the nine appreciable-cost conditions. Leave-one-class-out adapters can respond with the wrong correction, which is the training-support dependence in action. For moving hotspots, a period sweep changes the success of the switching Bayesian reference while the history adapter stays weak; the context contains a 100-shot EWMA as well as cumulative means, so the weakness is not the absence of a time constant, and separating the roles of cumulative statistics, finite-memory features, training support and the counter needs the interventions of Appendix Q.

Table O.1 gives the leave-one-class-out recoveries of the deployable adapter. Only the aligned pair keeps its recovery when its class is removed, and it does so because the retained pair class draws the same XX channel on the same vertical pairs at overlapping strengths. The hotspot and pair recoveries depend on seeing their class.

Table O.1: Leave-one-class-out generalization of the history adapter (deployable context). Each cell is the recovery share $\widehat{R}$ in percent, meta-trained on the full family → meta-trained without the scenario's fault class, mean over three seeds; the held-out class is excluded from validation as well. Bold: every seed improves on its frozen backbone with $p < 0.05$. Intervals and the known-noise context are in the supplied records. The aligned-pair row is not an out-of-family test, because the retained pair class contains the same XX channel (Appendix O).

scenario (left-out class)	MLP	CNN	Trf	TCN	GNN	3D	nBP
static hotspot	52→-12	46→-29	58→-12	65→-34	69→-20	34→-5	68→-68
moving hotspot	-17→-158	-9→-88	-20→-162	3→-194	-13→-135	-12→-144	-29→-47
correlated pair	28→13	22→18	4→-24	-2→-33	-7→-31	-33→-45	-34→-25
aligned pair	36→**43**	2→4	42→**46**	39→**42**	40→**41**	38→**43**	1→1

P BAYESIAN REFERENCE: DEFINITION, IDENTIFICATION AND LIMITS

To place a causal reference beside the learned adapter, this appendix defines the exact Bayesian filter over a discrete noise prior, shows what syndrome history can and cannot identify, and states the limits of that reference. Let z be a discrete noise hypothesis and $H_{t-1} = (s_1, \ldots, s_{t-1})$. Given predictive weights $q_t(z)$, the filter computes

$$w_t(z) = \frac{q_t(z)P(s_t \mid z)}{\sum_{z'} q_t(z')P(s_t \mid z')}, \qquad P(y_t = 1 \mid H_{t-1}, s_t) = \sum_z w_t(z)P(y_t = 1 \mid s_t, z),$$

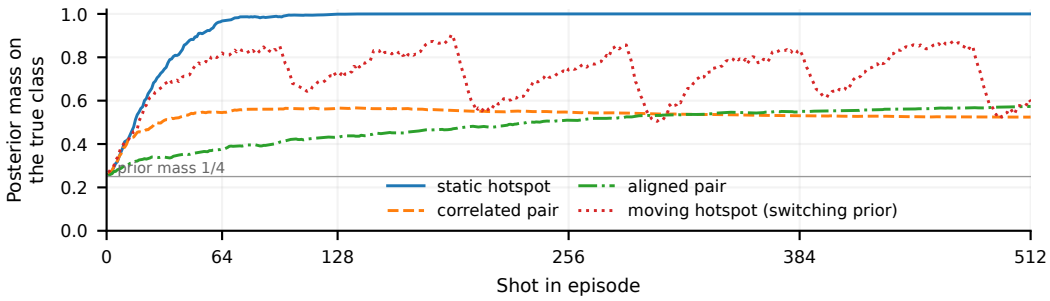


Figure P.1: Identification along the episode (Section 5.4): posterior mass that the Bayesian filter places on the true mechanism class, averaged over the 64 reference episodes. The static hotspot is identified within about 50 shots; the aligned pair is not identified within 512 shots; the correlated pair is decoded at 98% recovery although its class holds only about half of the posterior mass (0.51–0.57).

Table P.1: Posterior mass that the Bayesian filter places on the true mechanism class after 32, 64, 128 and 512 shots, averaged over the 64 reference episodes, and the first shot at which it exceeds 0.9. The moving hotspot uses the switching variant ($\epsilon = 1/96$) designated by the scenario schedule, and the other rows use the static prior. Prior mass is $1/4$ per class (*: not sustained, see the text).

scenario	32	64	128	512	shots to 0.9
Static hotspot	0.75	0.97	1.00	1.00	51
Correlated pair	0.51	0.54	0.57	0.52	never
Aligned pair	0.34	0.38	0.43	0.57	never
Moving hotspot	0.68	0.82	0.72	0.60	191*

and predicts by thresholding at $1/2$. The static model sets $q_{t+1} = w_t$; the switching model sets $q_{t+1} = (1-\epsilon)w_t + \epsilon\pi$. The recursion is exact for its finite latent-state model and minimizes prior-averaged zero-one loss; it uses no logical labels.

There are 1,497 hypotheses: 21 stationary rates, 648 hotspots, 240 correlated-pair settings and 588 aligned-pair settings, each class with prior mass $1/4$ spread over its entries. Stationary rates run from 0.002 to 0.012 in steps of 0.0005, whereas the continuous adapter-training family uses 0.003 to 0.008; the other classes also use discrete grids. The hypothesis grids contain the benchmark's own sites and strengths (base rate 0.004, multiplier 5 on the center row, pair probabilities 0.03 and 0.025), so the filter is evaluated under a well-specified prior, which is its most favorable case. The stationary-rate grid and the adapter family differ in range and discretization. The recoveries are a well-specified-prior reference, not a bound for a filter whose prior misses the deployed configuration. The switching family uses $\epsilon \in \{1/96, 1/512\}$, with the static model at zero. The reference rule selects the static variant for constant-parameter scenarios and $1/96$ for the moving hotspot's 96-shot schedule; it is a fixed schedule-based choice, not a best-of-three on evaluation losses, and all variants are archived with their audits.

For the moving hotspot in Table P.1, the averaged mass exceeds 0.9 only at shot 191 (peak 0.905), just before the row moves at shot 192, and is not sustained (trace in the supplied record identification_switching_v11.json).

Figure P.1 and Table P.1 show the identification behavior behind the recoveries of Table L.1, and Table P.2 the absolute LERs. Class mass is a label-level quantity: the pair class contains the physically identical XX channel on the same vertical pair, and the aligned-pair class contains the single-cell parity twins of Section 5.4, so mass on the true class understates identification of the mechanism and overstates identification of its logical parity. For any three decoders the empirical identity

$$\widehat{C} = [\widehat{\mathcal{L}}(h_S^*) - \widehat{\mathcal{L}}(b)] + [\widehat{\mathcal{L}}(b) - \widehat{\mathcal{L}}(h^*)]$$

holds as accounting; we call the terms source-rule improvement and filter–target-optimal gap. Prior-average optimality does not prohibit another causal rule from improving on this filter on a fixed

Table P.2: Absolute empirical LER on identical reference episodes. The moving-hotspot filter uses the switching variant ($\epsilon = 1/96$); the other rows use the static variant. Values are rounded only for presentation.

scenario	source-optimal	Bayesian filter	target-optimal
Static hotspot	0.02188	0.01843	0.01852
Correlated pair	0.00891	0.00699	0.00696
Aligned pair	0.05298	0.04462	0.03598
Moving hotspot	0.02075	0.01953	0.01819

Table P.3: Absolute comparison on identical episodes. The adapter with lowest final LER is selected descriptively within each scenario. Positive gap means worse than the Bayesian filter; gap units are 10^{-3} LER. Intervals are paired episode/seed bootstrap intervals conditional on the selected model, not selection-adjusted superiority tests.

scenario	selected model	adapted	filter	gap [95% CI]
Static hotspot	TCN	0.02044	0.01843	2.00 [1.09,2.96]
Correlated pair	Transf.	0.00931	0.00699	2.32 [1.41,3.27]
Aligned pair	Transf.	0.04588	0.04462	1.26 [-0.15,2.72]
Moving hotspot	TCN	0.02134	0.01953	1.81 [0.96,2.70]

scenario, so a formal information lower bound would need matched prior-average evaluation or a separate minimax argument. The original 512-shot reference gives 49.2% aligned-pair recovery with a paired interval of about 38–60%.

On a common baseline, the adapter with the lowest final LER trails the filter by 2.00×10^{-3} and 2.32×10^{-3} on the static hotspot and pair and by 1.26×10^{-3} $[-0.15, 2.72]$ on the aligned pair (Table P.3); on the aligned pair the gap is unresolved at this sample size, and both stop near the parity limit of Section 5.4, a limit under the filter's prior rather than a bound on every causal rule. On the aligned pair the filter's class mass climbs from 0.34 at shot 32 to 0.57 at shot 512 and never reaches 0.9; on the static hotspot it passes 0.9 at shot 51.

Baseline choice affects the numerical adapter–filter comparison. Selecting GNN and MLP by their largest own-baseline recoveries gives gaps of 60.6 and 135.4 points of $\widehat{C}$ on the static hotspot and pair; selecting by lowest final adapted LER chooses TCN and Transformer (Table P.3). Both selections are descriptive, and subtracting an own-baseline recovery from the filter's source-optimal recovery would be wrong. The aligned-pair Transformer gap is about 7.4 points and its paired interval includes zero.

Q Episode-length sensitivity

To check whether the adaptation results depend on the episode length, this appendix repeats the evaluation at 2,048 and 8,192 shots per episode with the same trained adapters. The length experiment changes only the number of shots and the episode seeds of the scenario specification: 64/32/16 episodes at 512/2,048/8,192 shots, the same trained adapters, and total shots increasing by $1/2/4$ times; the moving-hotspot period stays at 96 shots. The 512-shot column is a fresh draw over the five in-house architectures, so its $\widehat{C}$ need not equal the original estimate. Its static-hotspot filter recovery of 86.0% (Table Q.2) lies below the reference-set interval of Table L.1: the two draws have different $\widehat{C}$ denominators, and the reference-set interval does not describe draw-to-draw variation of $\widehat{C}$.

The long-episode evaluation assigns dataset seeds by a fixed scenario order and evaluates all fifteen trained models on each common input. A reproduction audit found that an earlier subset evaluation assigned seeds by subset index: a static-hotspot seed-1 record matched seed base 710000 instead of 713000, so those records cannot be paired by episode with the full-order reference. We reran frozen and adapted predictions together and report the corrected common-input evaluation in Table Q.1; the original records and the identifying reproduction test are in the analysis provenance. The defect

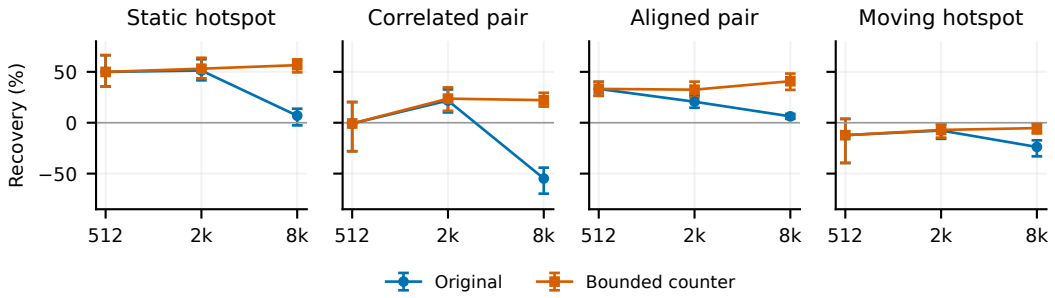


Figure Q.1: Episode length and the counter intervention (this appendix). Means over fifteen trained models (five in-house architectures, three seeds); bars are paired episode-bootstrap 95% intervals. Recovery uses each adapter's own frozen baseline. All three lengths are fresh episode sets, so the 512-shot values differ from Table L.1; 2k and 8k are 2,048 and 8,192 shots.

Table Q.1: Fresh common-input length evaluation and exploratory counter intervention. All five in-house models and three seeds use identical episodes within each row. Recovery is in percent of the same empirical source-rule cost; intervals jointly resample episodes and the denominator, conditional on the 15 trained models. The last column is bounded minus original adapter LER in 10^{-3} units; negative is better. No retraining is used.

scenario	shots	original $\widehat{R}$ [CI]	bounded $\widehat{R}$ [CI]	LER change [CI]
Static hotspot	512	49.9 [36, 66]	49.9 [36, 66]	+0.00 [+0.00, +0.00]
Static hotspot	2,048	51.3 [42, 62]	53.1 [44, 64]	-0.07 [-0.14, -0.01]
Static hotspot	8,192	6.9 [-3, 14]	56.6 [49, 62]	-1.62 [-1.84, -1.43]
Correlated pair	512	-0.8 [-28, 20]	-0.8 [-28, 20]	+0.00 [+0.00, +0.00]
Correlated pair	2,048	21.8 [10, 33]	23.7 [12, 35]	-0.04 [-0.10, +0.02]
Correlated pair	8,192	-54.8 [-70, -44]	22.2 [16, 29]	-1.55 [-1.68, -1.42]
Aligned pair	512	33.2 [26, 40]	33.2 [26, 40]	+0.00 [+0.00, +0.00]
Aligned pair	2,048	20.6 [15, 27]	32.5 [25, 40]	-2.37 [-3.03, -1.75]
Aligned pair	8,192	6.3 [4, 9]	40.8 [32, 48]	-5.96 [-7.24, -4.71]
Moving hotspot	512	-12.2 [-40, 4]	-12.2 [-40, 4]	+0.00 [+0.00, +0.00]
Moving hotspot	2,048	-7.7 [-16, -1]	-7.1 [-15, -1]	-0.02 [-0.07, +0.03]
Moving hotspot	8,192	-23.8 [-33, -17]	-5.3 [-11, -1]	-0.62 [-0.72, -0.52]

is specific to the length extension and does not touch the central 126-cell benchmark, which uses stored shared datasets.

The filter uses the schedule-based variant at every length, and all twelve scenario/length rows record a passing causality audit (Table Q.2). Its moving-hotspot recovery of 24/42/32% is the switching variant's ($\epsilon = 1/96$); the static variant is negative there.

Input-range diagnostics identify the extrapolation: the normalized shot count reaches nearly 16 at 8,192 shots against less than 1 in meta-training, an excursion of 15 training-range widths, while the largest excursion among the other 108 features is about 2. The intervention clamps only this raw feature to $[0, 1]$, without retraining or model selection, on all five in-house architectures, three seeds, four structural scenarios and three lengths; at 512 shots it is exactly the identity. At every length we check that future-syndrome perturbations preserve the feature prefix, and the deployed wrapper is audited on training-length episodes for each scenario. Windowed statistics and length-mixed training were not tested.

R LARGER CODES, TRAINING BUDGETS AND THE MATCHING SURROGATE

To say what the evidence is beyond the exact settings, this appendix reports the training-budget and larger-code results and the matching surrogate, together with where the surrogate fails. Exact tables

Table Q.2: Bayesian-reference recovery (%) at three episode lengths. The same designated variant is used and audited at all lengths. These are point estimates on independently generated episode sets, not universal bounds.

scenario	512	2,048	8,192
Static hotspot	86.0	96.5	100.2
Correlated pair	101.4	100.7	100.0
Aligned pair	55.2	54.1	52.5
Moving hotspot	24.4	41.7	32.0

have 17 bits at $d = 3$, $r = 2$, 25 at $d = 3$, $r = 3$ and $d = 5$, $r = 1$, and 49 at $d = 5$, $r = 2$, where a dense double-precision table would need 4 PiB.

Beyond the exact settings, a matching surrogate stands in. At $d = 3$, $r = 8$, $d = 5$, $r = 5$ and $d = 7$, $r = 7$, source-model versus true-model MWPM is the available *matching-model mismatch* diagnostic. Its error on the multiplier grid has no fixed sign: it over-reads source-optimal regret by up to about $16\times$ in the $d = 3$, $r = 2$ calibration, reads CIZ$\times$3 at 0.2 times its exact cost on the $d = 5$, $r = 5$ exact-decision shots, and reads the multi-axis grid near zero in the $r = d$ sweep of Table R.1, including $d = 3$ where the exact CIZ$\times$3 cost at $r = 3$ is 3.1%; its value depends on matching approximations as well as model mismatch, and large-code confidence intervals do not carry that systematic error.

The exact per-shot decisions calibrate the surrogate where both exist. At $d = 5$, $r = 5$ the exact per-shot decisions of Appendix H calibrate it on the same shots: over the 11 shifted cases its rank correlation with the exact C is 0.77; where C is resolved on the structural side it reads 0.4–1.2 times the exact value (correlated pair 1.07, aligned pair 1.17, aligned pair at event probability 0.008 0.71, and 0.38 at 0.004, where the surrogate rests on one net shot); where C is below 10^{-4} its value is noise around zero; and on the two grid points it reads 0.2 times the exact cost (CIZ$\times$3) and a negative value within noise (C$\times$3).

The calibration fixes how the larger-code surrogate values may be read. At this size the surrogate therefore does not over-read the grid, it can under-read it, which favours the separation, so the $d = 7$, $r = 7$ surrogate values are not a bound on the grid's cost. Larger-code structural/control separation under the surrogate shows sensitivity to these mechanisms; it does not place the complete grid below every structural population regret there.

Table R.1 repeats the strength sweep with this surrogate at $r = d$ up to $d = 11$. The pattern of the exact sweep persists: uniform rate changes stay within sampling noise of zero at every distance, each structural mechanism grows with strength, the center-qubit hotspot weakens with distance but still reads about 20% at $\times 10$ and $d = 11$, and the aligned pair remains the largest (about 100% at event probability 0.03 and $d = 11$). These are trends in a calibrated but approximate diagnostic, not exact costs. The surrogate does not measure the multi-axis grid: it reads CIZ$\times$3 and MCZ$\times$3 near zero at every distance, including $d = 5$ and $d = 7$ where the exact cost of CIZ$\times$3 is 5.8% ($r = 5$) and 9.1% ($r = 2$). Beyond $d = 5$ the table therefore supports the strength dependence of the structural mechanisms, not a comparison between grid and structural costs.

Two limits of Table R.1 matter for reading it. The surrogate is not C. At $d = 5$, $r = 5$ it reads 0.4–1.2 times the exact cost on resolved structural cases and can under-read grid costs, and at $d = 3$ it fails on the aligned pair, which there is a parity twin that matching resolves differently from the Bayes rule. The two grid rows read near zero at every distance, although the exact cost of CIZ$\times$3 is 5.8% at $d = 5$, $r = 5$ and 9.1% at $d = 7$, $r = 2$ (Table 1), so the surrogate does not see these grid costs and the table speaks to the structural mechanisms only.

Classical graph updates include rate rescaling, shrunk and raw p_{ij} estimates, blind pair search, and a circuit-enumerated candidate library (Spitz et al., 2018; Google Quantum AI, 2021; Chen et al., 2022; Google Quantum AI, 2023; Sivak et al., 2024; Takou & Brown, 2025); guided updates and true-model matching receive extra information and are diagnostic references (Table R.2). All deployable tiers estimate causally within the episode: statistics start at shot 64, the graph is rebuilt every 32 shots, and nothing is carried across episodes. The shrunk p_{ij} tier weights its running estimate by $n/(n + 1024)$, so by shot 512 it has moved a third of the way from the source DEM

Table R.1: Strength sweep under the matching surrogate at $r = d$: H_{rel}, the extra LER of source-model matching over true-model matching, in percent of the latter, 50,000–400,000 shots per point (more at larger d), base $p = 0.004$ (uniform rows: the target rate itself), source $p = 0.005$. Values within about $\pm 3\%$ are sampling noise. *: at $d = 3$ the aligned pair is a parity twin that matching resolves differently from the Bayes rule.

mechanism	$d = 3$	$d = 5$	$d = 7$	$d = 9$	$d = 11$
uniform $p = 0.006$	0.0	-0.4	0.0	0.6	-1.5
uniform $p = 0.008$	-0.1	-2.5	-0.3	1.0	-0.8
center qubit $\times 2$	0.1	4.1	5.7	-2.8	2.8
center qubit $\times 3$	9.9	2.7	3.1	4.1	3.4
center qubit $\times 5$	20.2	14.6	13.1	9.0	7.9
center qubit $\times 10$	53.2	28.3	22.9	25.6	19.5
horizontal pair 0.003	0.2	3.9	5.0	4.2	3.9
horizontal pair 0.01	6.4	17.7	12.2	12.1	15.2
horizontal pair 0.03	32.1	43.3	41.0	37.1	36.1
aligned pair 0.001	-53.0*	4.4	4.8	4.8	3.3
aligned pair 0.003	-42.3*	14.4	15.1	10.3	9.2
aligned pair 0.01	-15.4*	59.1	51.2	37.6	37.3
aligned pair 0.03	8.7*	148.2	128.9	115.8	97.3
grid CIZ$\times 3$	-0.9	-1.0	-0.4	-0.9	-0.1
grid MCZ$\times 3$	-0.3	-1.8	-0.4	-0.5	-0.1

Table R.2: Matching tiers: percent of source-model versus true-model MWPM headroom recovered. This denominator is a matching surrogate, not exact C. Headroom is absolute LER. "Discover" adds edge discovery to p_{ij}; "guided" receives extra information and is a diagnostic reference, not a deployable tier.

point	scenario	headroom	rescale	p_{ij}	discover	blind	library	guided
d3r2	pair	0.0025	-1%	-12%	-5%	+41%	+71%	+71%
d3r8	pair	0.0121	-18%	-8%	+8%	+38%	+66%	+67%
d5r5	pair	0.0033	-7%	-8%	-2%	-28%	+49%	+70%
d5r5	aligned pair	0.0182	-1%	-8%	-2%	+47%	+74%	+77%
d3r2	aligned pair	0.0157	-1%	-2%	-2%	-2%	-1%	-1%
d3r8	aligned pair	0.0610	-3%	-1%	-1%	-2%	-4%	-3%
d3r8	moving	0.0179	+5%	+3%	+3%	-26%	+3%	+5%
d5r5	moving	0.0042	-4%	-17%	-17%	-404%	-53%	-4%

toward the data; the raw tier uses pseudo-count 64 with edge discovery. These tiers therefore run far below the shot counts at which p_{ij} has been used on hardware (7.6×10^4 experiments in Google Quantum AI, 2021), and their regressions on the small-cost scenarios are a statement about 512-shot causal estimation, not about p_{ij} as a calibration method. Both matchings keep the mechanisms of their undecomposed DEMs that touch at most two detectors, so the surrogate's mismatch is in edge probabilities and in the true model's added edges, not in a decomposition. Candidate libraries help on new mechanisms and introduce false positives elsewhere as the candidate set grows.

Large-code neural comparisons depend on optimization. The six evaluated backbones at $d = 5$, $r = 5$ have very different source errors: the ported 3D-CNN is near the classical line, the small in-house models are not (Table R.3). The matched-protocol Transformer includes an interrupted run; the completed recipe Transformer reaches mean LER 0.0636 against 0.0837 for the matched run, with one seed where several others use two.

At $d = 7$, $r = 7$ the 3D-CNN reaches 0.0090 at the 6M-shot budget, below source-model MWPM's 0.0106, and 0.0062 at 25M unique shots and 150,000 steps (a longer run not listed in Table R.4, which keeps the 6M-shot budget). The 0.0061 value cited from Yan et al. (2026) is their TCN, not their 3D-CNN, so this is numerical proximity between two different architectures and not a reproduction. Our detector-graph GNN and TCN at the larger budget reach 0.0232 and 0.0498; for the TCN, longer optimization on the same 6M shots moves LER from 0.0870 to 0.0548 before the additional data moves it to 0.0498. Optimization matters, and these runs do not uniquely separate

Table R.3: Scale points. Spread = worst/best -1 over the six implementations; the last column is the best neural decoder against BP+OSD on the matched control. *One seed, interrupted.

scale point	unseen-syndrome share	spread	ranking	best neural vs BP+OSD
$d = 3$, $r = 2$	0.05–0.6%	0.2–14% (3% on the control)	unstable, at the edge of 32k-shot resolution	0.0105 vs 0.0106
$d = 3$, $r = 8$	21–71%	2–48%	3D-CNN > GNN/TCN > Transformer > MLP > 2D-CNN	0.0338 vs 0.0336 (3D-CNN); best in-house 0.0377
$d = 5$, $r = 5$	70–99%	118–546%	3D-CNN > TCN > GNN > MLP > 2D-CNN > Transformer*	0.0110 vs 0.0108 (3D-CNN); best in-house 0.0273

Table R.4: Frozen validation LER at $d = 7, r = 7$, means over two seeds except neural-BP (†: one seed) and the longer-step and wider-model rows, which are single runs. Matched runs use 15,000 steps and remain improving at the budget limit. Longer-step and wider-model rows vary one axis. Times are recorded on an A100, except neural-BP on a 96 GB RTX PRO 6000 VM with micro-batch 128 and 16-step accumulation; these platform-dependent times are not matched-compute comparisons. The extended-step sweep excludes the best-performing 3D-CNN and is not an exhaustive architecture optimization.

implementation	parameters	wall-clock min	LER	× MWPM
3D-CNN (ported)	1,163,145	16.1	**0.0090**	**0.9**
GNN	125,569	20.1	0.0665	6.3
neural-BP GNN†	75,395	669.8	0.0567	5.4
TCN	86,529	0.8	0.0870	8.2
MLP	191,701	0.4	0.1202	11.3
Transformer	126,465	107.0	0.1233	11.6
2D-CNN	82,753	0.9	0.1327	12.5
TCN, 10× steps	86,529	7.7	0.0548	5.2
MLP, 10× steps	191,701	4.2	0.0885	8.3
2D-CNN, 10× steps	82,753	9.4	0.0511	4.8
TCN, 2× width	652,545	1.0	0.0637	6.0
TCN, 3× width	1,421,121	1.7	0.0609	5.7

architecture, optimization and data effects. The $d = 3$, $r = 2$ data-budget experiment changes unique training shots at fixed steps and shows little change in architecture spread across a large coverage change (Table R.5); parameter matching is not compute matching.

S EXTERNAL HARDWARE CORPUS

To check the evaluation questions against hardware data, this appendix analyzes the Willow release for movement within and between its acquisition blocks. The Willow release (Google Quantum AI and Collaborators, 2025) provides an external check on the evaluation questions, not a device fit of NSBench; drift detection on hardware in general has its own methodology (Proctor et al., 2020). Detection-event derivation was compared byte-for-byte with the release for 420 configurations. Each configuration has 50,000 shots in ten initialization-bitstring blocks of 5,000 repetitions; if the blocks are sequential, time and initialization identity are confounded.

Under an exchangeable-shot null using the measured variance of per-shot detector rate, 370 of 420 configurations reject at $p < 0.001$, with median 500-shot-bin overdispersion 2.08; a binomial detector-independence null would be inappropriate because one physical fault fires several detectors. Initialization blocks explain a median 38% of bin-to-bin variation, and 149 configurations keep at least 70% of their variation within blocks (283 keep more than half). After removing the common detector-rate series, median residual per-detector overdispersion is about 1.01, and an audited causal matching decoder shows no aggregate gain from global rate tracking and small pooled gains from

Table R.5: Architecture spread against training-data budget at $d = 3/r = 2$. Spread is the range over the six *architecture means* of the matched implementations; seed spread is the widest range across the 3 seeds of any one of them. Uncovered mass is the worst of the 18 scenarios.

unique shots	passes over data	uncovered test mass	arch spread	worst seed spread
500,000	61.4	2.670%	0.00029	0.00026
6,000,000	5.1	0.635%	0.00029	0.00019
25,000,000	1.2	0.247%	0.00020	0.00016

local re-weighting at $d = 5$ and $d = 7$. Static noise-model misfit and acquisition structure remain competing explanations; there is no device oracle and no neural deployment test on this corpus.

T Reproducibility and artifact scope

To state what can be reproduced from the released material, this appendix lists what the supplementary archive contains, what remains in the research repository, and what each artifact regenerates. The anonymized supplementary archive contains the code and the compact result records behind the tables and figures, and its README maps each paper element to a record folder. The research repository contains the benchmark source, scenario definitions, reference dataset manifests, compact reference data and prediction-derived episode statistics, the reanalysis, result records supporting every reported comparison, and the table and figure generators. Per-shot probability arrays, trained weights, dataset archives and job logs stay in the research repository, and statements in the appendices that cite the research repository refer to these.

The compact analysis records support reproducing the central comparisons and the exact small-code diagnostics without retraining every model. The datasets are regenerated from the scenario files and checked against the shipped digests. The dependency record pins the simulator to the Stim 1.15 series (1.15.0 was used), because a seed alone does not define bitwise samples across simulator versions, so regenerating with another release is a new dataset version. Training configurations and code permit further reproduction at the stated budgets. Reported costs depend on hardware: the five in-house frozen sweeps take about two hours on an 8 GB laptop GPU, streamed adaptation runs about 40 minutes per variant on twelve CPU threads, and the finite Bayesian table construction with the six-scenario evaluation about four minutes. Dense-table population sweeps at 25 outcome bits are substantially more expensive than at 17. Batched CPU throughput and one-shot streaming latency in Table K.3 are software measurements.

Absent optional checkpoints and external data are listed in the archive's README. One appendix statement can be checked only in the research repository: the Ising adapter validation. The analysis records keep file digests and distinguish measured results from proposed interventions. Dated local records improve provenance but are not an independently timestamped preregistration, which is why the schedule-based reference selection is described directly with all variants retained, and why the counter intervention is labeled exploratory. Windowed-statistic ablations, broader held-out sites and realistic source noise are future work.

U Physical scope of the hotspot strengths

To relate the benchmark's hotspot strengths to measured device fluctuations, this appendix collects published relaxation-time fluctuations and evaluates a tenfold single-qubit hotspot at $d = 5$, $r = 5$. Relaxation fluctuations motivate a sweep of local fault strength, but do not define an all-channel error multiplier. Burnett et al. find that "T1 can vary by more than a factor 2 for both qubits" and that "the dominant mechanism for T1 fluctuations is local to each qubit" (Burnett et al., 2019). Klimov et al. report that "T1 can vary by up to an order of magnitude" and describe "deep but sparse relaxation resonances" (Klimov et al., 2018). Long-term records show similar variation (Carroll et al., 2022). These observations place single-qubit degradations of about $2\times$ in the common range and about $10\times$ at the documented extreme.

Radiation-induced bursts provide evidence for events affecting several qubits. On an earlier processor, radiation events were initially localized but spread across the chip, with recovery over tens of

milliseconds (McEwen et al., 2022). In the high-distance repetition-code runs of the Willow study, bursts are "spatially localized to neighbourhoods of roughly 30 qubits", occur "approximately once an hour" and "decay with an exponential time constant of around 400 μs" (Google Quantum AI and Collaborators, 2025). These observations support our mechanism choices, but relaxation time, gate-error probability, and detector firing probability are different observables. The studies provide neither a conversion to our multipliers nor a persistent analogue of a whole hot row. The row-wide multiplier is therefore a stress setting.

Table U.1 reports the distance-five, five-round costs computed from exact decisions. The rule remains nearly unchanged through $\times 3$, with costs of at most 0.6% (Table G.3). At $\times 5$, the two single-qubit variants yield only three and five differing decisions. These samples cannot distinguish the variants from each other or from the 4% comparison level. The larger point estimate comes from scaling only the gate and idle depolarizing channels, a closer proxy for a relaxation defect than scaling all channels. At $\times 10$, a single qubit costs about a quarter of the target Bayes risk, comparable to the whole row at $\times 5$.

In Table U.1 the base probability is 0.004 and the source references use 0.005. Intervals resample 512-shot blocks. Three runs ended under the driver's precision rule, and the $\times 10$ run reached its 4,096-shot cap.

Table U.1: Distance-five, five-round hotspot costs from exact decisions (final runs). "Depolarizing only" scales the qubit's gate and idle depolarizing channels while leaving reset and readout flips unchanged.

Configuration	Shots	Differing decisions	$C/\mathcal{L}_T^*$ (%)
Whole row $\times 5$ (benchmark)	3,072	29	24.2 [16.6, 31.4]
Center qubit $\times 5$	2,048	3	4.1 [0, 9.7]
Center qubit $\times 5$, depolarizing only	2,048	5	10.2 [0, 17.2]
Center qubit $\times 10$	4,096	25	26.6 [11.9, 38.3]

Each distance-five, 25-round result in Table 1 uses one completed 512-shot block. A single block cannot estimate variability between blocks, so the identical bootstrap endpoints in the saved records are not valid uncertainty intervals. We therefore report these results as preliminary point estimates. At the same depth, the source-equals-target check found no differing decisions across 128 shots; identity checks are shorter than the shifted runs because they only need to show that no decision differs.

V Additional Limitations

This appendix collects the limitations that apply across the study rather than to one appendix. The exact reference is exact for the supplied detector error model, up to numerical precision. It does not establish that this model captures every physical correlation in a device. Dense integration has no sampling uncertainty, whereas posterior-weighted estimates over sampled syndromes do. Unobserved decision changes remain possible even when no sampled decision differs.

The main architecture comparison uses a shared data and optimization budget, not equal wall-clock compute. Three seeds support the reported model averages but do not establish a stable ranking of architectures. At larger codes, optimization and data coverage differ in their effects. The released pre-decoders retain different original training distributions, so their comparison tests deployment under controlled shifts rather than the effect of architecture alone.

The Bayesian filter is optimal for its discrete prior-average problem. Its recovery on a particular scenario is not an information-theoretic upper bound for all causal policies. A parity twin prevents syndrome-only distinction when both explanations are admitted, but the numerical recovery depends on their prior weights. The shot-counter clamp and the identity gate are diagnostic controls designed after observing failures. Broader held-out fault sites, continuous operation, and device-calibrated noise models remain separate tests of external validity.