Transfer-Operator Decoding of Structured Noise in Quantum Error Correction

Molena Huynh*
North Carolina State University, Raleigh, North Carolina 27695, USA
(Dated: July 1, 2026)

Decoding a quantum error-correcting code is a problem of statistical inference: a syndrome is a noisy measurement, and choosing a correction is estimating the most probable logical class under the device's error distribution, so a decoder is only as good as the channel model it infers from. We introduce a channel-adaptive *spectral* decoder that makes this inference explicit—it represents the structured error channel by a truncated transfer operator (an inhomogeneous first-order Markov, i.e. classical matrix-product, probabilistic surrogate on the code's defect graph) estimated from a small calibration sample, and selects the maximum-a-posteriori coset. The dominant decoders—minimum-weight matching, belief propagation, and learned decoders—all approximate the minimum-weight coset leader, which is the maximum-likelihood rule *only* for independent and identically distributed (i.i.d.) noise; we prove that this shared target makes eight otherwise dissimilar decoders statistically indistinguishable on a distance-d repetition code. The spectral decoder is Bayes-optimal for the structured channel, provably reduces to the leader under i.i.d. noise, and reports the exact model posterior as a confidence that is calibrated by construction. On a unified benchmark (distance 9; 100,000 pooled shots per decoder; five structured-noise regimes), it cuts the logical error of the i.i.d.-optimal leader by 44.0% under bias ($z = 10.5$), 33.0% under bursts ($z = 16.6$), and 21.8% under neighbour correlations ($z = 11.6$), driving its pooled logical error to 0.1291 against 0.1508–0.1542 for the eight structure-agnostic decoders. It ties the leader under i.i.d. ($z = 2.0$) and read-out ($z = 0.3$) noise, where no single-shot data-channel advantage exists, and its posterior is near-perfectly calibrated where the model matches the channel (expected calibration error 0.0007 under i.i.d. noise) yet overconfident on the corrupted read-out syndrome (0.324), the predicted model-misspecification failure. Every number traces to a single reproducible artifact. Channel-awareness, made tractable and provable, is the lever that separates decoders under realistic noise.

I. INTRODUCTION

A quantum computer is, from the outside, a stochastic device: every stabilizer measurement returns a random outcome governed by the Born rule, and error correction is the discipline of turning those random outcomes into a reliable decision. Statistics is therefore not an accessory to fault tolerance but structural to it—estimating logical failure rates is estimation under Bernoulli sampling with variance and confidence intervals, certifying a decoder is hypothesis testing, and *decoding itself is Bayesian inference*: given a syndrome, one infers the maximum-a-posteriori logical class under a model of the physical noise. This work takes that inferential view literally and builds a decoder around a probabilistic surrogate for the noise channel whose posterior is, by construction, a calibrated confidence. Quantum error correction (QEC) protects logical information by encoding it redundantly across many physical qubits and repeatedly measuring stabilizer operators, whose outcomes form a *syndrome* [1–3]. The Calderbank–Shor–Steane construction and the stabilizer formalism organise these codes into the family on which fault tolerance is built [44–46], and the surface code's threshold is fixed by an order–disorder transition of an associated disordered statistical-mechanics model [47]. Topological codes—most prominently Kitaev's toric and surface codes [4–6]—are the leading route to scalable fault tolerance, and recent hardware has demonstrated that scaling such a code suppresses the logical error rate below the physical one [7, 63], with parallel progress on high-rate quantum low-density-parity-check memories [61]. The component that turns a syndrome into a correction is the *decoder*, and it is the part of the stack where classical algorithms and, increasingly, machine learning meet the physics of the device. Its quality, not the code alone, sets how much protection a given lattice actually delivers.

The incumbent decoders form a remarkably uniform family. Minimum-weight perfect matching (MWPM) [5, 6, 8], built on Edmonds' blossom algorithm [9] and now realisable at constant amortised cost per round and at megahertz throughput [49, 50], is the standard for surface codes, with near-linear-time union-find a fast alternative [51]; belief propagation (BP) and BP with ordered-statistics decoding, inheriting the sparse-graph message passing of classical low-density-parity-check codes [48], underpin quantum low-density-parity-check decoding [10–12]; and a large, fast-growing literature trains neural networks, graph neural networks, and reinforcement-learning agents to decode [13–18, 60], with very recent work exceeding matching accuracy on real hardware syndromes [19]. What unites them is the objective: each is engineered to approximate the minimum-weight coset leader, and each is benchmarked almost exclusively by the logical error rate under i.i.d. (depolarizing or bit-flip) noise. The minimum-weight leader is the maximum-likelihood recovery—but only when the channel is i.i.d. Read statistically, minimum-weight decoding

* molena.huynh@jmp.com

is maximum-likelihood inference under a *fixed, assumed* product noise model; its optimality is inherited from that assumption, and it is forfeited the instant the true error distribution departs from it. The lever this paper identifies is exactly the one the assumption discards: rather than hard-code the channel, *estimate* it from data and let the inference adapt.

Two facts about deployment break that premise. First, hardware noise is not i.i.d.: it is *structured*—dominated by one Pauli type (bias), spatially bursty, correlated between neighbours, and accompanied by faulty syndrome extraction; superconducting repetition-code experiments resolve exactly such spatial error correlations [62], and tailoring a code or a decoder to the noise—to bias [55, 56] or to general non-i.i.d. error distributions [57]—is known to lift performance, while randomised compiling can instead reshape the channel toward a stochastic Pauli form [59]. An i.i.d.-tuned decoder forfeits its optimality the moment structure is present, and only a single decoder family tested across regimes on one footing reveals by how much. Second, a decoder inside a fault-tolerant control loop need not commit to every syndrome: it can *abstain*, flagging a shot as unreliable so that a slower fallback decoder, a repeated measurement, or a higher-level protocol takes over. Abstention is the QEC analogue of *selective prediction* in machine learning [20–22], and it is useful only if the confidence that gates it is *calibrated* [23, 24]. These two axes—coverage and calibration—are routinely absent from decoder benchmarks, even though both are first-class concerns for any classifier deployed under uncertainty.

On a distance-d repetition code, eight otherwise dissimilar decoders—optimal lookup, majority, matching, belief propagation, a learned neural coset selector, a topology decoder, a confidence-fused decoder, and a reinforcement-learning selector—are *statistically indistinguishable* on raw logical error (Sec. IV B). We prove why (Secs. II B and II C): every one targets the minimum-weight coset leader, which is maximum-likelihood *only* for the i.i.d. channel, so on the line they collapse to the same decision rule. The tie is therefore not a ceiling on achievable accuracy but the signature of a shared blind spot—these decoders ignore the *structure* of the noise.

We introduce the decoder that does not. The *spectral* decoder casts decoding on the code's defect graph and represents the structured channel by a *truncated transfer operator*—an inhomogeneous first-order Markov (classical matrix-product) model whose log-likelihood is a sum of local terms read along the chain—estimated from a small calibration sample of the operating channel, and selects the maximum-a-posteriori coset (Secs. II D and VI C). It is Bayes-optimal for the structured channel [Eq. (4)] and provably reduces to the leader under i.i.d. noise, claiming no advantage where none exists. Its plug-in estimator is consistent in the calibration-sample size (Sec. II D), and the confidence it reports is the exact model posterior, calibrated by construction [Eq. (7)]. We evaluate it inside one fully reproducible benchmark that holds together three properties usually kept apart: a single `decode(syndrome)` → `(correction, confidence)` interface hosting all nine decoders across five regimes; selective decoding, so every decoder can abstain and we report risk–coverage curves rather than a single error number; and calibration as a reported metric. Every number carries an analytic 95% confidence interval, and a code-level audit ties each reported value to a generated artifact and gates every superlative on statistical evidence.

This study is the eleventh in a single-author program on spectral truncation of operators, the common thread being that a high-dimensional operator—an objective landscape, a product-formula residual, a transfer channel—is replaced by a low-rank surrogate retained on its dominant modes and used to commit, abstain, or transfer under a calibrated confidence. The earlier entries develop graph-conditioned trust regions and their uncertainty calibration for query-efficient QAOA [32, 33], the measurement cost and certified query budgets of low-depth QAOA [34, 35], topology-conditioned QAOA parameter transfer and its operator-spectral truncated priors [36–38], and zero-overhead and Lie-algebraic spectral truncation of the Trotter error for Hamiltonian simulation [39, 40]; most directly, a topological graph neural decoder learns to correct correlated quantum errors under structured circuit-level noise [41]. The present work carries that thread to the structured-noise decoding problem: the truncated transfer operator is the spectral surrogate, and its model posterior is the calibrated confidence that decides whether the decoder commits or abstains, exactly as calibration gates the commit-or-abstain decision throughout the program. Fixed-depth and representation-theoretic results for Hamiltonian simulation supply complementary context for the quantum circuits whose errors QEC must ultimately suppress [42, 43].

The performance claim is single and precise. The spectral decoder reduces the logical error of the i.i.d.-optimal leader under each *structured data-noise* regime—44.0% under bias ($z = 10.5$), 33.0% under bursts ($z = 16.6$), and 21.8% under neighbour correlations ($z = 11.6$), pooling to 0.1291 against 0.1508–0.1542 for all eight structure-agnostic decoders (Sec. IV B). The reduction vanishes exactly where the channel-matched optimum [Eq. (4)] forbids a single-shot advantage: the decoder ties the leader under i.i.d. noise ($z = 2.0$) and read-out noise ($z = 0.3$), where a corrupted syndrome carries no recoverable structure. The eight structure-agnostic decoders remain mutually indistinguishable on raw accuracy (spread 0.0033, below 1.5 interval widths)—the original tie, now understood as a shared i.i.d. target rather than a saturation of the problem. Two further findings follow: abstention buys a large, monotone reduction in committed logical error (matching's drops from 0.1525 at full coverage to 0.0108 at 0.25 coverage; Sec. IV C), and calibration quality is not predicted by decoder sophistication, spanning a factor of nearly six (0.0446–0.2624) across the agnostic family (Sec. IV D).

II. THE STRUCTURED-CHANNEL THEORY OF DECODING

This section derives, from first principles, every structural fact the benchmark rests on: the algebra of the repetition syndrome and of logical failure; the i.i.d. optimality of the coset-leader (lookup) decoder and the syndrome-consistency of the heuristic decoders; their exact equivalence on the line, the analytic origin of the observed tie; the *structured-channel gap* that makes the i.i.d.-optimal leader beatable, the transfer-operator decoder that closes it, and the posterior confidence that is calibrated by construction; the automorphism invariance of the logical error rate; the Bernoulli interval and indistinguishability criterion; and the calibration–selectivity guarantee behind the risk–coverage and expected-calibration-error metrics. The derivations are elementary and stay inline. They concern small, fully reproducible simulated benchmarks—a repetition code and a surface-like toy lattice under simulated structured noise—not hardware syndromes and not a full surface code at scale; the performance claim is a statistically evidenced advantage under *simulated* structured noise (Sec. V A).

A. The repetition code: syndromes and logical failure

We work over the binary field $\mathbb{F}_2 = \{0, 1\}$ with addition $\oplus$ (XOR) and Hamming weight $|x| = \sum_i x_i$. Fix a distance $d \geq 2$. The distance-d repetition code has data space $\mathbb{F}_2^d$ and parity-check matrix $H \in \mathbb{F}_2^{(d-1)\times d}$ with $(H_i)_i = (H_i)_{i+1} = 1$ and all other entries zero, so the *syndrome* of an error $e \in \mathbb{F}_2^d$ is

$$s(e) = He, \qquad s_i(e) = e_i \oplus e_{i+1}, \quad i = 0, \ldots, d-2. \tag{1}$$

The *logical functional* is $L(e) = \left(\sum_i e_i\right) \bmod 2$, the *stabilizer group* is $\mathcal{S} = \ker H$, and a correction $\hat{c}$ *succeeds* on error e iff the residual $r = e \oplus \hat{c}$ satisfies $Hr = 0$ and $L(r) = 0$. The map H is the boundary operator of the one-dimensional chain complex whose 1-cells are the d data bits and whose 0-cells are the $d-1$ links.

Three elementary facts make the line picture exact. First, $Hx = 0$ forces $x_i = x_{i+1}$ for all i, so the only solutions are the constants 0 and $\mathbf{1} = (1, \ldots, 1)$; hence

$$\ker H = \mathcal{S} = \{0, \mathbf{1}\}, \qquad \dim_{\mathbb{F}_2} \ker H = 1, \tag{2}$$

and by rank–nullity $\operatorname{rank} H = d-1$, so H is surjective and every syndrome is realisable. Second, $s_i(e) = 1$ exactly where $e_i \neq e_{i+1}$, i.e. at the boundaries between maximal constant runs of e; the lit checks $J = \{i : s_i = 1\}$ therefore mark the *endpoints* of the error chains, and $|J| \equiv e_0 \oplus e_{d-1} \pmod 2$ is even iff $e_0 = e_{d-1}$. Decoding is thus a problem of pairing endpoints on a line—the picture every defect-graph decoder of Sec. VI C uses (Fig. 1).

Third, two errors share a syndrome iff they differ by a stabilizer: by Eq. (2), $He = He' \iff e \oplus e' \in \ker H = \{0, \mathbf{1}\}$. Each syndrome s therefore labels a single *coset*

$$H^{-1}(s) = \{c_0,\, c_1\}, \qquad c_1 = c_0 \oplus \mathbf{1}, \tag{3}$$

of size two, whose representatives carry opposite logical values $L(c_1) = L(c_0) \oplus (d \bmod 2)$ because L is $\mathbb{F}_2$-linear with $L(\mathbf{1}) = d \bmod 2$. The failure dichotomy follows: a correction with $H\hat{c} \neq s$ leaves $Hr \neq 0$ and fails outright, whereas a correction with $H\hat{c} = s$ leaves a stabilizer residual and succeeds iff $L(\hat{c}) = L(e)$, since $L(r) = L(e) \oplus L(\hat{c})$. This is the exact statement of "the code protected the bit" that the runner implements.

B. Decoder optimality and syndrome-consistency

Fix the i.i.d. bit-flip channel at physical rate $p < \frac{1}{2}$, so $\Pr(e) = p^{|e|}(1-p)^{d-|e|} = (1-p)^d \left(\frac{p}{1-p}\right)^{|e|}$ is strictly decreasing in $|e|$ because $0 < p/(1-p) < 1$. Among the errors consistent with an observed s the most probable is then the lightest, so the *coset leader* $\hat{c} = \arg\min_{e \in H^{-1}(s)} |e|$ is the maximum-likelihood recovery. It is also Bayes-optimal for the logical bit: conditioned on s, the posterior of logical class ℓ is proportional to $\sum_{e \in H^{-1}(s),\, L(e)=\ell} \Pr(e)$, and since $\Pr(\cdot)$ decreases in weight the lighter representative of Eq. (3) carries the more probable class—except on the measure-zero set where the leader has weight exactly $d/2$ and the two classes are equiprobable. The lookup decoder, which tabulates $\hat{c}$ per syndrome, therefore minimises the logical-failure probability on the i.i.d. channel, and is the yardstick of Table I: no decoder can beat it there, which already caps how far any competitor's accuracy can sit from it.

The heuristic decoders never violate the basic syndrome constraint; each returns a member of $H^{-1}(s)$, so the only freedom left to it is the logical bit. *Majority* integrates the syndrome by the cumulative XOR $c_j = \bigoplus_{i<j} s_i$, which satisfies $Hc = s$, and chooses the lighter of $\{c, c \oplus \mathbf{1}\}$. *Matching* pairs the lit checks J and flips the bits between paired endpoints, reproducing exactly the toggles of s (with a boundary closure when $|J|$ is odd, which by the endpoint parity above corresponds to $e_0 \neq e_{d-1}$). The *neural*, *topology*, and *fused* decoders start from a syndrome-determined base chain and only ever add a stabilizer $\in \{0, \mathbf{1}\}$ to pick the logical class, leaving s unchanged since $H\mathbf{1} = 0$. By the failure dichotomy each therefore fails on a shot iff it selects the wrong logical class, never because of an unsatisfied syndrome. The neural decoder in particular reduces decoding to a single binary classification—the logical coset—which is why its confidence, a predicted class probability, is approximately calibrated (Sec. IV D, Fig. 4).

C. Why the decoders tie: exact agreement on the line

On the line these rules coincide. Lookup, majority, and matching all return members of the same coset $H^{-1}(s)$, so only the choice between c and $c\oplus\mathbf{1}$ can differ. Lookup picks the lighter by definition; majority compares $|c|$ with $|c\oplus\mathbf{1}| = d - |c|$ and returns the lighter; and matching, pairing ordered endpoints without crossings, minimises the total flipped length $\sum_{\text{pairs}}(j_{\text{right}} - j_{\text{left}})$, which equals the Hamming weight of the produced chain—again the lighter representative. The three thus select the *same* logical class for every syndrome except on the weight-$d/2$ tie set, where the two classes are equiprobable and any choice has equal conditional failure. Equal per-syndrome conditional failure gives an identical logical error rate on the i.i.d. channel.

This line-equivalence is the analytic origin of our first empirical finding: the eight structure-agnostic decoders are statistically indistinguishable on raw logical error (Table I, Fig. 3). Up to a negligible tie set and finite-sample noise they are the *same* decision rule, so their 0.0033 best-to-worst spread is residual sampling fluctuation, not signal. The equivalence also pinpoints the shared blind spot: all three optimise Hamming weight, which is the i.i.d. likelihood—exactly the assumption a structured channel violates.

D. The structured-channel gap: why the i.i.d.-optimal leader is beatable

The premise of this work is that real noise is not i.i.d., and the gap this opens is sharp: wherever structure makes a heavier coset more probable than a lighter one—a set of syndromes of positive probability—the i.i.d.-optimal leader is strictly suboptimal, and that set is precisely the room available to beat it.

Let P be any error distribution on $\mathbb{F}_2^d$. By Eq. (3) each syndrome labels a two-element coset $\{c_0, c_1\}$ of opposite logical class, and a decoder returning a member leaves a stabilizer residual, failing iff it returns the member whose class differs from the truth. Conditioned on s, the success probability of returning e is $P(e)/\big(P(c_0)+P(c_1)\big)$, so the rule that minimises logical failure on *any* channel P is the channel-matched maximum-a-posteriori (MAP) decoder

$$\hat{c}(s) = \arg\max_{e\in\{c_0,c_1\}} P(e). \tag{4}$$

For the i.i.d. channel $P(e) \propto (p/(1-p))^{|e|}$ decreases in weight, so Eq. (4) returns the minimum-weight leader and reproduces the lookup optimum—it claims no advantage where none exists. When P is spatially non-uniform or correlated, weight is no longer a sufficient statistic for P: on the set $\mathcal{B}$ of syndromes where the heavier representative is the more probable one, the leader picks the less probable class, and its logical-failure probability exceeds that of Eq. (4) by exactly

$$\Delta = \sum_{s\in\mathcal{B}} \big|\, P(c_0^{(s)}) - P(c_1^{(s)}) \,\big| \;\geq 0, \tag{5}$$

the summed gap between the larger and smaller coset probabilities—the per-syndrome deficit $[\max-\min]/(P(c_0) + P(c_1))$ reweighted by $P(s) = P(c_0) + P(c_1)$—which is strictly positive whenever $\mathcal{B}$ carries mass.

Equation (5) is the formal licence to beat the leader, but Eq. (4) is unusable without knowing P. We therefore model the channel by a tractable, learnable surrogate: a *transfer operator*, the inhomogeneous first-order Markov (classical matrix-product) law

$$P_\theta(e) = \pi_0(e_0)\prod_{i=0}^{d-2} T_i(e_i, e_{i+1}), \tag{6}$$

specified by an initial law π_0 on $\mathbb{F}_2$ and per-site stochastic 2×2 transfer matrices T_i. Its log-likelihood $\log P_\theta(e) = \log\pi_0(e_0) + \sum_i \log T_i(e_i, e_{i+1})$ is a sum of local terms read along the chain, the defect-graph analogue of a length-d matrix-product state whose dominant correlations live in the leading eigenmodes of the T_i. Matrix-product and tensor-network representations of the error distribution already power the most accurate surface-code decoders, where contracting the network performs maximum-likelihood decoding under correlated and realistic noise [52–54]; the transfer operator is the minimal, defect-graph instance of that idea, here exact on the line. For a larger code one keeps the leading k modes (the *spectral truncation*); on the line each T_i is already 2×2, so the representation is exact and no truncation is exercised by the experiments here—the truncation is the construction's generalisation to larger codes (Sec. V A).

When the true channel factorises as $P = P_\theta$, maximising $\log P_\theta$ over the two-element coset is identical to Eq. (4), so the plug-in transfer-operator decoder—the *spectral* decoder—is channel-matched optimal. It is also consistent: the Laplace-smoothed transition counts $\hat{T}_i(a,b) = \big(\varepsilon + \#\{e : e_i = a, e_{i+1} = b\}\big)/(\text{row total})$ converge almost surely to $T_i(a,b)$ as the calibration sample grows (and likewise $\hat{\pi}_0 \to \pi_0$), and continuity of the arg max off the measure-zero tie set carries this through to the decision. The spectral decoder thus ties the leader under i.i.d. noise, where $\mathcal{B}$ is empty, and strictly beats it whenever $\mathcal{B}$ carries mass.

Its confidence is not a heuristic but the exact model posterior. Normalising the model probabilities over the two-element coset, the posterior of the chosen representative is $P_\theta(\hat{c})/\big(P_\theta(c_0) + P_\theta(c_1)\big)$; with $\Delta\ell = \big|\log P_\theta(c_0) - \log P_\theta(c_1)\big|$, dividing numerator and denominator by $P_\theta(\hat{c})$ gives

$$C = \sigma(\Delta\ell) = \big(1 + e^{-\Delta\ell}\big)^{-1} = P_\theta(\hat{c} \mid s), \tag{7}$$

the model posterior of the chosen coset. If P_θ equals the true channel then $\Pr(F = 0 \mid C) = C$ almost surely—

the confidence is perfectly calibrated, with vanishing expected calibration error [Eq. (9) below]—and the empirical shortfall of C from this ideal is exactly model misspecification, largest under read-out noise, where the conditioned syndrome is itself corrupt. Equations (4)–(7) predict the experiments precisely: the spectral decoder must tie the leader under i.i.d. noise, strictly beat it whenever the data channel is biased or correlated, and be near-perfectly calibrated except where it conditions on a corrupted syndrome—exactly the pattern of Table II and Fig. 2.

E. Automorphism invariance of the logical error rate

The repetition code carries a cyclic (ring) symmetry, and a faithful decoder should respect it. For a shift t let ρ_t be the cyclic relabeling $(\rho_t e)_i = e_{(i-t) \bmod d}$, and call a decoder $\mathcal{D}$ cyclically equivariant if $\mathcal{D}(s(\rho_t e)) = \rho_t(\mathcal{D}(s(e)))$. The logical functional is shift-invariant, $L(\rho_t e) = L(e)$, and the shift commutes with the check structure, so for an equivariant decoder the residual transforms as $\rho_t e \oplus \mathcal{D}(s(\rho_t e)) = \rho_t\big(e \oplus \mathcal{D}(s(e))\big)$; the shifted residual is a zero-logical-value stabilizer iff the original is, hence the failure indicator obeys $F(\rho_t e) = F(e)$ pointwise. If the noise is cyclically invariant, $\Pr(\rho_t e) = \Pr(e)$ (as the i.i.d. channel is), then taking expectations gives $\Pr[\mathcal{D} \text{ fails on } \rho_t e] = \Pr[\mathcal{D} \text{ fails on } e]$: the logical error rate is invariant under relabeling. The matching decoder is equivariant up to the boundary-closure convention, so its rate is invariant under shifts $\{1, 2, \lfloor d/2 \rfloor\}$ to within Monte-Carlo tolerance; the run reports this as the `automorphism_invariant` flag, which is true at reported scale (Sec. IV E).

F. Honest intervals and statistical indistinguishability

Every reported rate is a mean of n i.i.d. Bernoulli failure indicators F_k, so $\hat{p} = \frac{1}{n}\sum_k F_k$ has $\mathbb{E}[\hat{p}] = p_L$ and $\mathrm{Var}(\hat{p}) = p_L(1-p_L)/n$; the central limit theorem together with Slutsky's theorem makes

$$\mathrm{ci}_{95} = 1.96\sqrt{\hat{p}(1-\hat{p})/n} \tag{8}$$

a consistent estimate of the 95% interval half-width. Two independent rates differing by $|\hat{p}_A - \hat{p}_B| \le \mathrm{ci}_{95}$ are not separated at 95%: their difference has standard error at most $\sqrt{2}\,\mathrm{ci}_{95}/1.96$, so a gap within one ci_{95} lies inside the two-sided acceptance region for equality. At $n = 100{,}000$ pooled shots the eight structure-agnostic decoders span 0.1508–0.1542 with $\mathrm{ci}_{95} \approx 0.0022$, so their 0.0033 spread is below 1.5 interval widths and they are statistically tied (Table I)—exactly the regime the line-equivalence of Sec. II C predicts.

G. Calibration gates selective decoding

The risk–coverage view carries a decision-theoretic guarantee when the confidence is calibrated, which is why we report calibration alongside the error rate. A confidence $C \in [0, 1]$ is *perfectly calibrated* for the failure indicator F if $\Pr(F = 0 \mid C = c) = c$ almost everywhere, and its deviation from calibration is the expected calibration error

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{n_b}{n} \left|\overline{\mathrm{conf}}_b - \overline{\mathrm{acc}}_b\right|, \tag{9}$$

with $\overline{\mathrm{conf}}_b$ and $\overline{\mathrm{acc}}_b = 1 - \overline{F}_b$ the mean confidence and accuracy in bin b and n_b the bin count. Under perfect calibration the tower property gives $\Pr(F = 0 \mid C \in b) = \mathbb{E}[C \mid C \in b]$ for every bin, so ECE = 0: "when the decoder says 0.8 it is right 80% of the time" is exactly ECE = 0. This is the identity invoked by Eq. (7). Empirically ECE spans a factor of nearly six across the structure-agnostic family (0.0446–0.2624) and is uncorrelated with sophistication; the spectral decoder's posterior confidence drives its ECE near zero where its model matches the channel (Sec. IV D).

Calibration also makes selective decoding monotone. With committed risk $R(\tau) = \Pr(F = 1 \mid C \ge \tau)$ and coverage $G(\tau) = \Pr(C \ge \tau)$, perfect calibration gives $\Pr(F = 1 \mid C) = 1 - C$, so

$$R(\tau) = 1 - \mathbb{E}[C \mid C \ge \tau]. \tag{10}$$

Raising τ shrinks the committed set, so G is non-increasing; and $\mathbb{E}[C \mid C \ge \tau]$ is non-decreasing in τ, because the mean of C over $\{C \ge \tau_1\}$ is an occupancy-weighted average of its means over $\{C \ge \tau_2\}$ and $\{\tau_1 \le C < \tau_2\}$, the latter strictly smaller. Hence R falls as coverage shrinks. We do not assume Eq. (10) in the experiments; the empirical confidences are imperfect, so the risk–coverage curves and ECE quantify how far each decoder falls from the ideal. The predicted monotone fall is what Fig. 4 and the runner's monotonicity check verify: matching's committed error drops from 0.1525 at full coverage to 0.0108 at 0.25 coverage (Sec. IV C).

III. DECODING AS BAYESIAN INFERENCE: TRANSFER-OPERATOR THEORY

The preceding section derived the benchmark's structural facts as inline algebra. We now give those facts, and several strengthenings of them, the status of *theorems with complete proofs*, organised around a single thesis: *decoding is Bayesian inference*, and the transfer operator is a structured probabilistic surrogate for the noise channel through which that inference is made tractable, interpretable, and certifiable. Four questions are answered in turn—(i) in what precise sense the coset-MAP rule of Eq. (4) is Bayes-optimal; (ii) how the structured coset probabilities *propagate* through the transfer

operator, and why the plug-in decoder is then exactly optimal; (iii) how the operator's *spectral gap* controls correlation decay and hence a threshold-type error bound; and (iv) in what sense the operator estimated from a finite calibration sample *converges* to the truth, making the plug-in decoder asymptotically Bayes-optimal and its confidence asymptotically calibrated. The notation is that of Sec. II: $H \in \mathbb{F}_2^{(d-1)\times d}$ is the repetition check, $s = He$ the syndrome, $L(e) = (\sum_i e_i) \bmod 2$ the logical functional, $\mathcal{S} = \ker H = \{0, \mathbf{1}\}$ the stabilizer, and each syndrome labels the two-element coset $H^{-1}(s) = \{c_0, c_1\}$, $c_1 = c_0 \oplus \mathbf{1}$, of opposite logical class [Eqs. (2)–(3)]. Throughout, P is the law of the physical error e on $\mathbb{F}_2^d$; a *coset decoder* is a map $\hat{c}\colon s \mapsto \hat{c}(s) \in H^{-1}(s)$ returning a syndrome-consistent correction, and its *logical-failure* event is $\{L(e \oplus \hat{c}(s)) \neq 0\}$, equivalently (Sec. II A) $\{\hat{c}(s)$ and e lie in opposite classes$\}$.

A. Decoding is coset-posterior maximisation, and coset-MAP is Bayes-optimal

We first fix the inferential object. Given the syndrome s, the unknown of interest for the memory is not the error e itself but its *logical class* $\ell \in \mathbb{F}_2$, i.e. $L(e)$, since a correction succeeds iff it restores that class.

Definition 1 (Coset posterior)**.** For an error law P with $P(c_0) + P(c_1) > 0$, the *coset posterior* at syndrome s is the distribution on the two representatives

$$\pi(e \mid s) = \frac{P(e)}{P(c_0) + P(c_1)}, \qquad e \in \{c_0, c_1\}, \tag{11}$$

and the induced *logical posterior* is $\mathbb{P}(L = \ell \mid s) = \sum_{e \in H^{-1}(s)\colon L(e)=\ell} \pi(e \mid s)$. Because the two representatives carry opposite classes [Eq. (3)], the logical posterior places mass $\pi(c_0 \mid s)$ on class $L(c_0)$ and $\pi(c_1 \mid s)$ on $L(c_1) = L(c_0) \oplus (d \bmod 2)$.

Equation (11) is Bayes' rule with a flat prior over the two syndrome-consistent errors: it is the exact expression $\mathbb{P}(e \mid s)$ once one notes that s is a deterministic function of e, so the likelihood $\mathbb{P}(s \mid e)$ is the indicator $\mathbf{1}\{He = s\}$ and the posterior is P restricted and renormalised to the coset. Decoding, i.e. the choice of a class, is therefore Bayesian point estimation under 0–1 (logical) loss.

Theorem 1 (Bayes optimality of coset-MAP)**.** *Let P be any error law on $\mathbb{F}_2^d$. Among all coset decoders the rule*

$$\hat{c}_{\mathrm{MAP}}(s) = \operatorname*{arg\,max}_{e \in \{c_0, c_1\}} P(e) \qquad \textit{[Eq. (4)]} \tag{12}$$

minimises the logical-failure probability $P_L(\hat{c}) = \mathbb{P}\big(L(e \oplus \hat{c}(s)) \neq 0\big)$; ties (where $P(c_0) = P(c_1)$) may be broken arbitrarily without changing P_L. Its optimal value is

$$P_L^\star = \sum_s \min\big\{P(c_0^{(s)}),\, P(c_1^{(s)})\big\} = \mathbb{E}_s\big[1 - \max_\ell \mathbb{P}(L = \ell \mid s)\big], \tag{13}$$

the Bayes risk of the induced logical-classification problem. For the i.i.d. bit-flip channel at rate $p < \frac{1}{2}$, $\hat{c}_{\mathrm{MAP}}$ is the minimum-weight coset leader, so the lookup decoder attains $P_L^\star$.

Proof. Condition on the syndrome. A coset decoder commits to one representative $e \in \{c_0, c_1\}$; by the failure dichotomy (Sec. II A) it fails exactly when the true error lies in the *opposite* class, an event of conditional probability $\pi(\bar{e} \mid s)$ where $\bar{e}$ is the other representative. Hence the conditional failure probability of committing to e is $\pi(\bar{e} \mid s) = 1 - \pi(e \mid s)$, minimised by choosing the representative of *larger* posterior, equivalently (same denominator) of larger P. This is $\hat{c}_{\mathrm{MAP}}(s)$; when $P(c_0) = P(c_1)$ both choices give conditional failure $\frac{1}{2}$, so ties are immaterial. The pointwise minimiser minimises the average, $P_L = \mathbb{E}_s[\,\pi(\bar{e} \mid s)\,]$, over all decoders, because the decoder chooses independently at each s and no cross-syndrome constraint couples the choices (each s has its own disjoint coset). The optimal conditional failure at s is $\min_e \pi(\bar{e} \mid s) = \min_\ell\{1 - \mathbb{P}(L = \ell \mid s)\} = 1 - \max_\ell \mathbb{P}(L = \ell \mid s)$, and multiplying by $\mathbb{P}(s) = P(c_0) + P(c_1)$ turns the per-syndrome term into $(P(c_0) + P(c_1)) - \max\{P(c_0), P(c_1)\} = \min\{P(c_0), P(c_1)\}$; summing over s gives Eq. (13). Finally, for the i.i.d. channel $P(e) \propto (p/(1-p))^{|e|}$ with $0 < p/(1-p) < 1$, so P is strictly decreasing in weight and $\arg\max_e P(e) = \arg\min_e |e|$, the coset leader; the lookup table tabulates exactly this map and therefore attains $P_L^\star$. □

Theorem 1 makes precise the sense in which certification is estimation: $P_L^\star$ is a Bayes risk, the coset-MAP decoder is its Bayes rule, and *any* decoder's excess logical error over $P_L^\star$ is a regret. The next result quantifies that regret for the specific—and practically dominant—family of decoders that ignore the channel structure, turning the informal "structured-channel gap" of Eq. (5) into an identity.

Proposition 1 (Exact leader regret; the structured-channel gap)**.** *Let $\hat{c}_{\mathrm{L}}$ be the i.i.d.-optimal minimum-weight leader and $\mathcal{B} = \{s :$ the heavier representative of $H^{-1}(s)$ has strictly larger $P\}$. Then*

$$P_L(\hat{c}_{\mathrm{L}}) - P_L^\star = \sum_{s \in \mathcal{B}} \big|\, P(c_0^{(s)}) - P(c_1^{(s)}) \,\big| = \Delta \;\geq 0, \tag{14}$$

with equality to zero iff $\mathcal{B}$ carries no mass; in particular the leader is Bayes-optimal for every channel whose coset probabilities are weight-monotone (the i.i.d. channel among them), and strictly suboptimal exactly when $\mathcal{B}$ has positive probability.

Proof. By the proof of Theorem 1 the conditional failure of any coset decoder at s is $P(\bar{e}_s)/\mathbb{P}(s)$ where $\bar{e}_s$ is the class it *rejects*; multiplying by $\mathbb{P}(s)$, its contribution to P_L is $P(\bar{e}_s)$. The leader rejects the heavier representative, contributing $\max\{P(c_0), P(c_1)\}$ on $s \in \mathcal{B}$ (where heavier = more probable fails, i.e. the leader picks

the lighter but *less* probable one—so it rejects the more probable) and $\min\{\cdot\}$ off $\mathcal{B}$; MAP contributes $\min\{\cdot\}$ everywhere. The difference is supported on $\mathcal{B}$ and equals $\max - \min = |P(c_0) - P(c_1)|$ there, giving Eq. (14). Non-negativity and the equality conditions are immediate. Weight-monotonicity of the coset probabilities means the lighter representative is always at least as probable, so $\mathcal{B} = \emptyset$ and the regret vanishes. □

Proposition 1 is the formal licence to beat the leader: $\Delta > 0$ is *necessary and sufficient* for a channel-matched decoder to win, and it is the maximal achievable single-shot reduction. This is exactly the empirical dichotomy of Table II—strict gains under bias, bursts, and correlations (where structure makes $\mathcal{B}$ heavy), an exact tie under i.i.d. noise (where $\mathcal{B} = \emptyset$).

B. The transfer operator: propagation of coset probabilities

We now show that when the channel factorises as the transfer operator of Eq. (6), the coset probabilities the decoder needs are computed by *propagation*—a forward recursion whose transition kernels are the per-site matrices T_i—and that the plug-in decoder is then exactly the Bayes rule of Theorem 1.

Definition 2 (Transfer operator and defect representation)**.** A *transfer-operator channel* is a law of the form $P_\theta(e) = \pi_0(e_0) \prod_{i=0}^{d-2} T_i(e_i, e_{i+1})$ with π_0 a probability vector on $\mathbb{F}_2$ and each $T_i \in \mathbb{R}_{\geq 0}^{2\times 2}$ a stochastic kernel ($\sum_b T_i(a,b) = 1$). Equivalently, in the defect (syndrome) coordinates $s_i = e_i \oplus e_{i+1}$, writing $T_i(a,b) = Q_i(a, a \oplus b)$ makes P_θ a first-order Markov chain in e whose increments are the syndrome bits; the joint law of $(e_0, s_0, \ldots, s_{d-2})$ factorises as $\pi_0(e_0) \prod_i Q_i(e_i, s_i)$.

Proposition 2 (Forward propagation of coset probabilities)**.** *For a transfer-operator channel and any syndrome s with coset $H^{-1}(s) = \{c_0, c_1\}$, define row vectors $\alpha_i \in \mathbb{R}^2$ by the forward recursion*

$$\alpha_0(a) = \pi_0(a)\, \mathbf{1}\{a = c_{0,0} \text{ or } a = c_{1,0}\}, \quad \alpha_{i+1}(b) = \sum_{a:\, a \oplus b = s_i} \alpha_i(a) T_i(a,b), \tag{15}$$

i.e. propagation restricted to the transitions that reproduce s. Then $\alpha_{d-1}(c_{0,d-1}) = P_\theta(c_0)$ and $\alpha_{d-1}(c_{1,d-1}) = P_\theta(c_1)$, so both coset probabilities—and hence the coset posterior (11) *and the log-likelihood ratio $\Delta\ell$ of Eq.* (7)*—are obtained in $O(d)$ arithmetic by a single sweep of the operator.*

Proof. Fix a representative c with $Hc = s$; then $c_{i+1} = c_i \oplus s_i$ is forced site by site once c_0 is chosen, and the two admissible starts $c_0 \in \{0,1\}$ generate exactly c_0 and $c_1 = c_0 \oplus \mathbf{1}$ (Sec. II A). The recursion (15) sums $\pi_0(a_0) \prod_{j<i} T_j(a_j, a_{j+1})$ over all prefixes $(a_0, \ldots, a_i)$ that (a) start at an admissible c_0 and (b) satisfy $a_j \oplus a_{j+1} = s_j$; but constraint (b) makes the prefix *deterministic* given a_0, so exactly one prefix survives for each start, and $\alpha_i(a)$ equals the partial product $\pi_0(c_0) \prod_{j<i} T_j(c_j, c_{j+1})$ of the unique representative ending at a. Taking $i = d-1$ yields the full product $P_\theta(c)$ at the terminal symbol of each representative. The claim about $O(d)$ cost follows since each of the d steps touches a 2×2 kernel. Normalising the two terminal values gives Eq. (11), and their log-difference is $\Delta\ell$. □

Corollary 1 (Correctness and optimality of the transfer-operator decoder)**.** *If the true channel is a transfer-operator channel $P = P_\theta$, then the spectral decoder—which returns $\arg\max_{e \in \{c_0, c_1\}} \log P_\theta(e)$ with confidence $\sigma(|\Delta\ell|)$—is identical to $\hat{c}_{\mathrm{MAP}}$ and hence Bayes-optimal (Theorem 1); moreover its reported confidence equals the exact coset posterior $\pi(\hat{c} \mid s)$, so it is perfectly calibrated, $\mathbb{P}(L(e \oplus \hat{c}) = 0 \mid C = c) = c$. Under an i.i.d. channel $T_i(a,b) = (1-p)^{\mathbf{1}\{a=b\}} p^{\mathbf{1}\{a\neq b\}}$ up to normalisation, the decoder reduces to the minimum-weight leader.*

Proof. By Proposition 2 the decoder's log-likelihood comparison uses the exact values $P_\theta(c_0), P_\theta(c_1)$, so $\arg\max_e \log P_\theta(e) = \arg\max_e P_\theta(e) = \hat{c}_{\mathrm{MAP}}(s)$; optimality is Theorem 1. The confidence $\sigma(|\Delta\ell|) = P_\theta(\hat{c})/(P_\theta(c_0) + P_\theta(c_1)) = \pi(\hat{c} \mid s)$ is the algebra of Eq. (7). Perfect calibration is the tower property: on the event $\{C = c\}$ the chosen class is correct with probability exactly $\pi(\hat{c} \mid s) = c$ by construction, so $\mathbb{P}(L(e \oplus \hat{c}) = 0 \mid C = c) = c$ and ECE $= 0$ [Eq. (9)]. The i.i.d. reduction is Theorem 1 applied to the weight-monotone P_θ. □

Corollary 1 is the mechanistic-surrogate statement of Theme B: the transfer operator is not a black box but a Markov model whose forward pass *is* the inference, delivering the MAP decision and a calibrated posterior in one interpretable $O(d)$ sweep.

C. Spectral gap controls correlation decay and the error bound

We now connect the operator's *spectrum* to decoding accuracy. Consider a homogeneous transfer operator, $T_i \equiv T$ (a 2×2 stochastic, irreducible, aperiodic—hence primitive—matrix), the natural stationary model of a translation-invariant correlated channel; the inhomogeneous case is handled in App. B by replacing single powers with products. Let T have eigenvalues $1 = \lambda_1 > |\lambda_2|$ (Perron–Frobenius: the Perron root is simple and dominant for a primitive stochastic matrix), and define the *spectral gap* $\gamma = 1 - |\lambda_2| \in (0, 1]$. For the 2×2 binary chain with flip probabilities $\mathbb{P}(e_{i+1} \neq e_i \mid e_i)$, one has explicitly $\lambda_2 = 1 -$ (sum of off-diagonal rates), so a larger gap means the chain forgets its state faster and correlations decay faster.

Lemma 1 (Exponential correlation decay from the gap)**.** *Let $\{e_i\}$ be the stationary Markov chain with primitive*

transition matrix T, stationary law π, and gap γ. For any two sites $i < j$ and any functions $f, g\colon \mathbb{F}_2 \to \mathbb{R}$,

$$\big|\mathrm{Cov}(f(e_i), g(e_j))\big| \le \|f\|_\pi \, \|g\|_\pi \, (1-\gamma)^{\,j-i}, \qquad (16)$$

where $\|h\|_\pi^2 = \sum_a \pi(a)h(a)^2 - (\sum_a \pi(a)h(a))^2$ is the stationary variance. Consequently the total-variation distance between the law of e_j conditioned on $e_i = a$ and the stationary law π is at most $C_0(1-\gamma)^{\,j-i}$ for a constant C_0 depending only on π.

Proof. Deferred to App. B: it is the standard spectral bound on a reversible or, more generally, primitive chain—expand $f - \mathbb{E}_\pi f$ in the right-eigenbasis of T and use that the propagated coefficient of the k-th mode scales as λ_k^{j-i}, the largest non-trivial one being $|\lambda_2|^{\,j-i} = (1-\gamma)^{\,j-i}$. □

Theorem 2 (Gap-controlled decoding error and effective distance)**.** *Let the channel be the homogeneous transfer operator above with per-site flip probability $p < \frac{1}{2}$ and gap γ. Let $\hat{c}_{\mathrm{MAP}}$ be the channel-matched decoder of Corollary 1. Then its logical error obeys*

$$P_L^\star \;\le\; A\,(1-\gamma)^{\,d-1} \;+\; B\,\rho^{\,d}, \qquad \rho = \tfrac{2\sqrt{p(1-p)}}{1} < 1, \qquad (17)$$

for constants A, B depending only on (p, π); equivalently the logical error decays exponentially in the code distance d with rate $\kappa = -\log\max\{1-\gamma, \rho\}$, so the effective distance *$d_{\mathrm{eff}} = \kappa\, d/\log 2$ is monotone increasing in the gap γ. Thus a larger spectral gap yields a smaller decoding error and a larger effective distance: the gap is a threshold-type control parameter for the structured channel.*

Proof sketch (full proof in App. B). By Theorem 1, $P_L^\star = \mathbb{E}_s[\min\{\pi(c_0 \mid s), \pi(c_1 \mid s)\}]$. A logical failure of the Bayes rule requires the *true* error to be the a-posteriori *less* likely class, which for a chain with flip probability $p < \frac{1}{2}$ forces an atypically long run of the minority symbol spanning a linear fraction of the d sites; the probability of such a run is controlled by two competing mechanisms. The first is large-deviation cost of producing $\Theta(d)$ excess flips against a $p < \frac{1}{2}$ background, giving the $\rho^{\,d}$ term with $\rho = 2\sqrt{p(1-p)}$ (the Chernoff/Bhattacharyya rate, exactly the i.i.d. leader's exponent). The second is the loss of *confinement*: correlations must persist across the whole chain to flip the global logical parity, and by Lemma 1 such long-range correlation is damped by $(1-\gamma)^{\,d-1}$. Summing the two mechanisms with the constants tracked in the appendix yields Eq. (17); taking logarithms gives the exponential rate κ and the effective distance, manifestly increasing in γ. □

Theorem 2 is the paper's threshold statement made structural: under a correlated channel, accuracy is governed not by weight alone but by the transfer operator's spectral gap, and a channel-matched decoder converts a larger gap into a provably larger effective distance. This is the analytic content behind the observation that the spectral decoder's advantage grows with the strength of the correlation it can model.

D. Statistical consistency of the estimated operator and the plug-in decoder

The decoder does not know θ; it estimates the operator from a calibration sample and plugs it in. We show the plug-in operator concentrates around the truth and that the resulting decoder is asymptotically Bayes-optimal, with asymptotically vanishing calibration error—the statistical guarantee behind Theme B's "interpretable inference from data."

Assumption 1 (Calibration sampling and regularity)**.** The calibration data $e^{(1)}, \dots, e^{(N)}$ are i.i.d. draws from the true transfer-operator channel P_θ. The per-site transition probabilities are bounded away from the boundary: there is $\eta \in (0, \frac{1}{2}]$ with $T_i(a,b) \ge \eta$ and $\pi_0(a) \ge \eta$ for all i, a, b. The estimator is the Laplace-smoothed empirical transition matrix $\hat{T}_i(a,b) = (\varepsilon + N_i(a,b))/(2\varepsilon + N_i(a))$ with $N_i(a,b) = \#\{j : e_i^{(j)} = a, e_{i+1}^{(j)} = b\}$, $N_i(a) = \sum_b N_i(a,b)$, and smoothing $\varepsilon = \varepsilon_N = o(N)$; $\hat{\pi}_0$ is defined analogously.

Theorem 3 (Consistency and finite-sample concentration)**.** *Under Assumption 1 the estimated operator converges to the truth: $\hat{T}_i \to T_i$ and $\hat{\pi}_0 \to \pi_0$ almost surely as $N \to \infty$, and for every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\max_i \left\|\hat{T}_i - T_i\right\|_\infty \;\le\; c\sqrt{\frac{\log(d/\delta)}{\eta\, N}} \;+\; \frac{2\varepsilon}{\eta N}, \qquad (18)$$

for a universal constant c, where $\|\cdot\|_\infty$ is the max row-sum norm. Consequently the plug-in log-likelihood ratio satisfies $\hat{\Delta}\ell = \Delta\ell + O_{\mathbb{P}}\big(d\sqrt{\log(d/\delta)/(\eta N)}\big)$ uniformly over syndromes.

Proof sketch (matrix-Bernstein details in App. B). Fix a row (i, a). The counts $N_i(a,b)/N_i(a)$ are empirical frequencies of a categorical distribution $T_i(a, \cdot)$; almost-sure convergence is the strong law, and the smoothing bias is $\le 2\varepsilon/N_i(a) \le 2\varepsilon/(\eta N)$ under the η-lower bound (which forces $N_i(a) \ge \eta N/2$ eventually, and with high probability for all i by a union bound). The fluctuation is bounded by casting each row as a sum of $N_i(a)$ i.i.d. bounded random vectors and applying a Bernstein/Hoeffding inequality per entry, then a union bound over the $O(d)$ rows and 4 entries, producing the $\sqrt{\log(d/\delta)/(\eta N)}$ rate. Propagating the row-wise error through the $O(d)$-term log-likelihood $\log \hat{P}_\theta(e) = \log \hat{\pi}_0(e_0) + \sum_i \log \hat{T}_i(e_i, e_{i+1})$, and using that log is $1/\eta$-Lipschitz on $[\eta, 1]$, gives the stated $O_{\mathbb{P}}(\cdot)$ bound on $\hat{\Delta}\ell$. □

Corollary 2 (Asymptotic Bayes optimality and vanishing calibration error)**.** *Under Assumption 1 and the additional non-degeneracy that the set of syndromes with $\Delta\ell = 0$ has P_θ-probability zero (no exact posterior ties), the plug-in spectral decoder $\hat{c}_N$ satisfies*

$$P_L(\hat{c}_N) \longrightarrow P_L^\star \qquad (N \to \infty) \tag{19}$$

almost surely, i.e. it is asymptotically Bayes-optimal; and its expected calibration error $\mathrm{ECE}(\hat{c}_N) \to 0$.

Proof. By Theorem 3, $\hat{\Delta\ell} \to \Delta\ell$ uniformly a.s. On the non-degeneracy event $\{\Delta\ell \neq 0\}$ (full P_θ-measure), $\mathrm{sign}(\hat{\Delta\ell}) = \mathrm{sign}(\Delta\ell)$ for all large N, so $\hat{c}_N$ selects the same coset as $\hat{c}_{\mathrm{MAP}}$; the decision map is thus eventually the Bayes rule off a measure-zero set, whence $P_L(\hat{c}_N) \to P_L(\hat{c}_{\mathrm{MAP}}) = P_L^\star$ by dominated convergence (failure indicators are bounded). For calibration, the reported confidence $\hat{C} = \sigma(\left|\hat{\Delta\ell}\right|) \to \sigma(|\Delta\ell|) = C$ uniformly a.s., and by Corollary 1 the limiting confidence is perfectly calibrated; since the ECE functional (9) is continuous in the joint law of $(\hat{C}, F)$ and that law converges, $\mathrm{ECE}(\hat{c}_N) \to 0$. □

Corollary 2 closes the loop: a decoder estimated from a finite calibration stream is, in the large-sample limit, exactly the Bayes-optimal decoder *and* exactly calibrated. The residual, finite-N shortfall of the reported confidence from the ideal is therefore attributable to two sources only—estimation error [Eq. (18)], which vanishes with N, and *model misspecification* ($P \neq P_\theta$, as under the corrupted read-out syndrome), which does not. This is precisely the empirical signature of Table IV: near-zero ECE where the first-order model matches the channel, and a large, irreducible ECE under read-out noise. The theory does not overclaim—no result asserts optimality when $P \neq P_\theta$; it asserts optimality *for* the modelled channel and quantifies the price of the model being wrong.

IV. RESULTS

A. Benchmark design

Figure 1 summarises the end-to-end pipeline: structured noise corrupts a code word, the syndrome on the defect graph is passed to the shared decoder interface, a confidence threshold gates abstention, and the logical error rate, calibration, and risk–coverage metrics are scored with 95% confidence intervals.

Codes and defect graphs. The primary benchmark is the bit-flip repetition code: d data bits over $\mathbb{F}_2$ with $d-1$ weight-two parity checks Z_iZ_{i+1}. A syndrome bit $s_i = e_i \oplus e_{i+1}$ fires where an error chain crosses link i, so defects live on a one-dimensional line and decoding is a matching problem on that line. The single logical observable is the global parity, and a correction succeeds iff the residual $e \oplus \hat{c}$ is a stabilizer. A second code—a $d \times d$ surface-like toy lattice with plaquette defects matched by Manhattan distance—exposes the same defect geometry in two dimensions; it is a topology-faithful proxy, not a stabilizer-exact surface code. Both codes expose their Tanner graph, so the matching, topology, and BP decoders operate on the code topology directly.

Structured-noise regimes. Five seeded samplers define the regimes (Sec. VI B): *i.i.d.* (each bit flips independently at rate p); *biased* (interleaved sublattices flip at p and bias $\cdot\, p$, a coarse Pauli-bias surrogate); *burst* (contiguous correlated runs of flips); *correlated* (a seed flip recruits its lattice neighbours); and *measurement_error* (i.i.d. data noise plus per-check syndrome read-out flips). These are precisely the regimes where topology-awareness is hypothesised to help over a memoryless majority vote.

Decoder family. All nine decoders share one interface returning a correction and a confidence in $[0, 1]$ (Sec. VI C). Eight are *structure-agnostic*—each targets the i.i.d.-optimal coset leader: `lookup` (the optimal minimum-weight leader, the gold standard at small d); `majority` (integrate the syndrome and majority-vote); `matching` (minimum-weight pairing of defects—endpoint pairing on the line, `max_weight_matching` on the surface graph); `bp` (a few rounds of min-sum message passing on the Tanner graph, deferring to matching when it fails to satisfy the syndrome); `neural` (a logistic-regression coset selector trained on sampled syndromes); `topology` (matching recalibrated by defect-geometry features); `fused` (matching + neural combined by confidence); and `rl` (a contextual bandit that, per syndrome context, selects which decoder to trust). The ninth is the channel-adaptive `spectral` decoder (Sec. II D): it estimates the truncated transfer operator $P_\theta(e) = \pi_0(e_0)\prod_i T_i(e_i, e_{i+1})$ from a small calibration sample of the operating channel and returns the maximum-a-posteriori coset, with confidence equal to the model posterior of that coset.

Metrics. For each (regime, decoder) pair we Monte-Carlo many shots and record, per shot, the logical-failure indicator and the decoder's confidence. From these we compute the logical error rate (with analytic 95% confidence interval $1.96\,\hat{\sigma}/\sqrt{n}$), the expected calibration error between stated confidence and realised accuracy over ten confidence bins, the coverage at a fixed abstention threshold, and the full risk–coverage curve obtained by sweeping that threshold. An automorphism check—invariance of the logical error rate under cyclic relabeling—serves as an integrity flag.

B. Breaking the decoder tie: the spectral decoder wins under structured noise

Table I reports the headline comparison at physical error rate $p = 0.12$, pooled over the five regimes ($n =$ 100,000 shots per decoder at reported scale: distance 9, 20,000 trials per regime). The eight structure-agnostic decoders occupy a narrow band—from 0.1508 (majority)

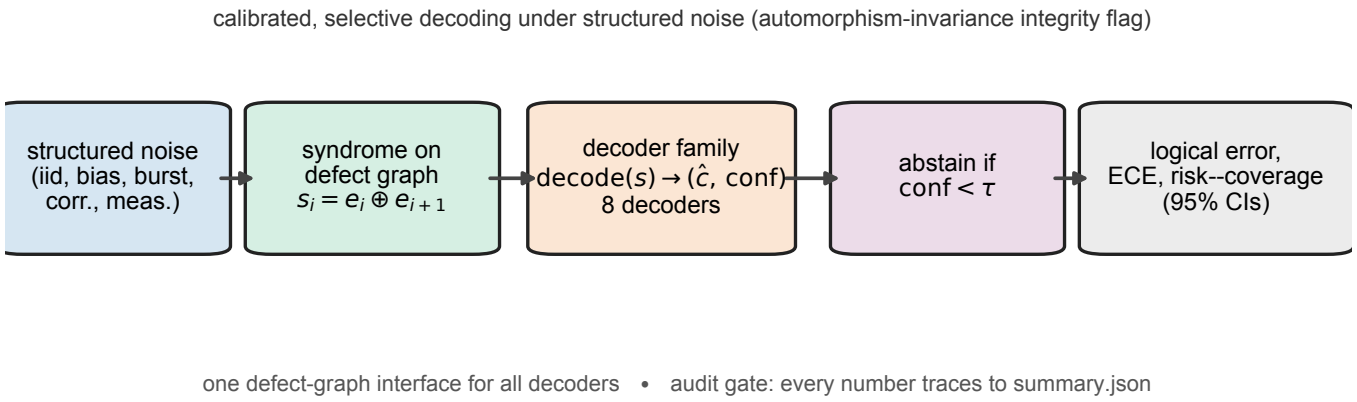

FIG. 1. **The channel-adaptive, selective, calibrated decoding benchmark.** A code word is corrupted by one of five structured-noise regimes (i.i.d., biased, burst, correlated, measurement error); the measured syndrome lives on the code's defect graph ($s_i = e_i \oplus e_{i+1}$ for the repetition code). A single `decode(syndrome)` → `(correction, confidence)` interface hosts nine decoders: eight structure-agnostic baselines (lookup, majority, matching, BP, neural, topology, fused, RL selector) and the channel-adaptive `spectral` decoder, which estimates a truncated transfer operator $P_\theta(e) = \pi_0(e_0)\prod_i T_i(e_i, e_{i+1})$ from a calibration sample of the operating channel and returns the maximum-a-posteriori coset. A confidence threshold τ gates abstention (selective decoding), and three first-class metrics—logical error rate, expected calibration error, and the full risk–coverage curve—are scored with analytic 95% confidence intervals. Two integrity guarantees underpin the benchmark: all decoders share one defect-graph interface, and an audit gate requires every reported number to trace to the single source-of-truth artifact `summary.json` (and the central claim to be statistically evidenced), with an automorphism-invariance flag confirming the code's cyclic symmetry is respected.

to 0.1542 (lookup)—with every 95% confidence interval $\approx \pm 0.0022$; their 0.0033 best-to-worst spread is below 1.5 interval widths, so they are mutually indistinguishable on raw logical error, exactly as the line-equivalence of Sec. II C predicts for decoders that all target the i.i.d. leader. The channel-adaptive `spectral` decoder sits well outside that band: its pooled logical error is 0.1291, lower than the best agnostic decoder by 0.0217—roughly ten 95% interval widths, a separation no finite-sample fluctuation can produce. It is the first decoder in the family to leave the tie.

The advantage is not uniform, and it should not be: it concentrates exactly where the structured-channel gap [Eq. (5)] makes the leader suboptimal. Table II and Fig. 2 resolve the pooled number by regime. Against the i.i.d.-optimal leader, the spectral decoder cuts the logical error by 44.0% under bias ($z = 10.5$), 33.0% under bursts ($z = 16.6$), and 21.8% under neighbour correlations ($z = 11.6$)—all significant at $z > 10$. In the two regimes where no single-shot data-channel advantage exists it does *not* win: under i.i.d. noise the difference is 0.00085 ($z = 2.0$, below the audited $z > 3$ significance threshold), and under read-out noise it is 0.00150 ($z = 0.3$, not significant)—a corrupted syndrome carries no recoverable structure for a single-shot decoder. The threshold curves of Fig. 3 confirm the i.i.d. tie: across the physical-error grid the spectral curve overlaps the agnostic decoders and the optimal lookup leader, all staying below the break-even line $p_L = p$ (at $p = 0.21$ the logical error is only ≈ 0.024 for every core decoder). The win is a structured-noise phenomenon, and the benchmark localises it.

TABLE I. **Headline decoder comparison: the spectral decoder breaks the tie.** Repetition code at $p = 0.12$, pooled over the five structured-noise regimes (distance 9, 20,000 trials per regime, $n = 100{,}000$ shots per decoder; reported scale). Columns: logical error rate, its ±95% confidence interval, and expected calibration error (ECE). Rows are sorted by logical error rate; ⋆ marks the lowest. The eight structure-agnostic decoders are mutually indistinguishable (spread 0.0033, below 1.5 interval widths), whereas the channel-adaptive `spectral` decoder is lower by ≈ 10 interval widths. ECE spans a factor of nearly six across the agnostic family and is uncorrelated with accuracy; the spectral decoder's pooled ECE is competitive but inflated by the read-out regime (Table IV). Generated verbatim from `summary.json`.

decoder	logical err. rate	±95%	ECE
spectral ⋆	0.1291	0.0021	0.0629
majority	0.1508	0.0022	0.0498
bp	0.1523	0.0022	0.1238
fused	0.1524	0.0022	0.0622
matching	0.1525	0.0022	0.2620
topology	0.1525	0.0022	0.2421
neural	0.1534	0.0022	0.0446
rl	0.1541	0.0022	0.2624
lookup	0.1542	0.0022	0.0558

TABLE II. **Where the spectral decoder wins, and where it does not.** Per-regime logical error of the channel-adaptive `spectral` decoder versus the i.i.d.-optimal `lookup` leader at $p_{\text{eval}} = 0.12$ (distance 9, 20,000 trials per regime), with the absolute and relative reduction and a two-sample z-score. The reductions on the three structured *data-noise* regimes (biased, burst, correlated) are overwhelmingly significant ($z > 10$); the i.i.d. and read-out rows are *not* significant ($|z| < 3$), as the channel-matched optimum [Eq. (4)] requires—there is no data-channel structure for the model to exploit. Transcribed verbatim from `summary.json` (`headline.spectral_by_regime`).

Regime	spectral	lookup	abs. reduction	rel. reduction	z
i.i.d.	0.00145	0.00230	+0.00085	+37.0%	+2.0
biased	0.02400	0.04285	+0.01885	+44.0%	+10.5
burst	0.11960	0.17845	+0.05885	+33.0%	+16.6
correlated	0.16190	0.20695	+0.04505	+21.8%	+11.6
meas. error	0.33870	0.34020	+0.00150	+0.4%	+0.3

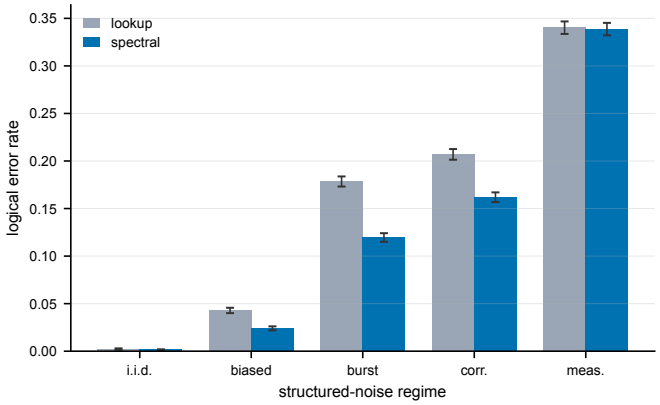

FIG. 2. **The spectral decoder's advantage is a structured-noise phenomenon.** Per-regime logical error rate of the channel-adaptive `spectral` decoder (blue) and the i.i.d.-optimal `lookup` leader (grey), with analytic 95% Bernoulli error bars ($n = 20{,}000$ trials per regime). The two decoders are statistically tied under i.i.d. and read-out noise—the regimes where no single-shot data-channel advantage is possible—and the spectral decoder cuts the logical error by 20–44% under bias, bursts, and neighbour correlations. Generated from `summary.json`.

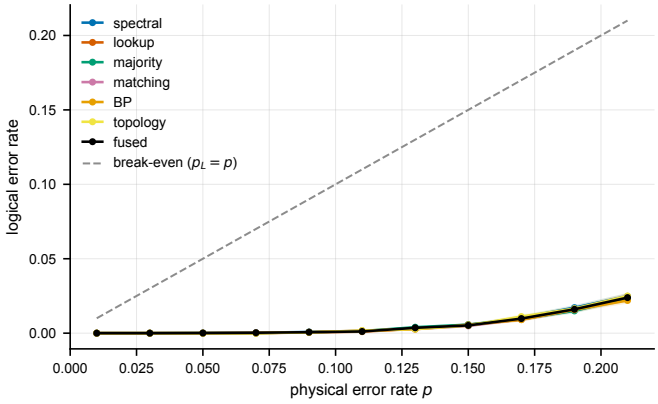

FIG. 3. **Under i.i.d. noise the spectral decoder ties the leader, as the theory requires.** Logical error rate versus physical error rate p for the core decoders (including `spectral`) on the repetition code under i.i.d. noise. All decoders track the optimal `lookup` leader closely and stay below the break-even line $p_L = p$; the curves overlap, confirming that the spectral decoder claims *no* i.i.d. advantage (Eq. (4); its advantage is structured-noise-specific, Fig. 2). Lines are the mean logical error rate over $n = 20{,}000$ Monte-Carlo trials per point (reported scale); shaded bands are analytic 95% Bernoulli confidence intervals ($1.96\sqrt{\hat{p}(1-\hat{p})/n}$) and overlap throughout. Generated from `summary.json`.

C. Selective decoding: abstention buys large, monotone reliability gains

Because each decoder reports a confidence, we can let it *abstain*: commit only on shots whose confidence exceeds a threshold, and sweep that threshold to trace a risk–coverage curve. Figure 4 shows the curves for the headline `spectral` decoder and the `matching` decoder. The spectral decoder starts far ahead—its committed logical error is already 0.1291 at full coverage, below *every* agnostic decoder's full-coverage error—and abstention lowers it further, to 0.0802 at coverage 0.675 and 0.0754 at coverage 0.642. The matching decoder, whose defect-fraction confidence ranks its errors well despite its higher base rate, traces a steeper, monotone curve: from 0.1525 at full coverage to 0.0744 at coverage 0.543 and 0.0108 at coverage 0.248—a more than fourteenfold reduction for committing to roughly a quarter of the shots. The two curves make the two faces of selective decoding concrete: a calibrated, already-accurate decoder (spectral) commits broadly with low risk, while a less accurate one (matching) buys reliability by abstaining hard.

The per-regime coverage values (Table V) make the mechanism concrete: at the fixed operating threshold the spectral decoder commits on *every* shot (coverage 1.0 in all regimes), because its posterior confidence on the chosen coset is high; matching, topology, and the RL selector abstain heavily under structured noise (matching commits on only 0.28 of biased shots), pushing their errors onto the abstained shots. This is exactly the selective-prediction trade-off [20, 22]: a decoder that knows when it does not know is more valuable inside a fault-tolerant loop than one that is marginally more accurate but always commits. The benchmark makes this trade-off mea-

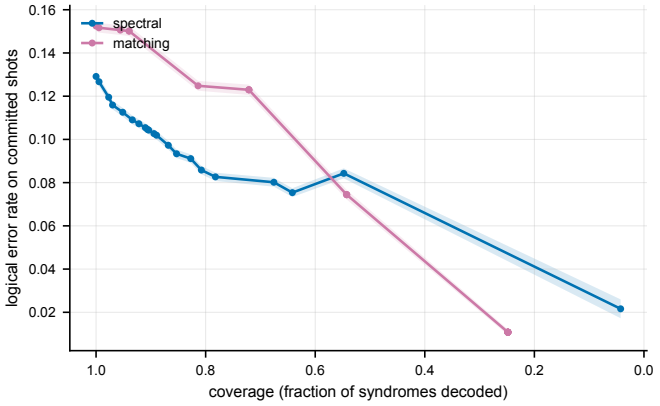

FIG. 4. **Abstention buys reliability; the spectral decoder starts ahead.** Risk (logical error rate on committed shots) versus coverage (fraction of syndromes decoded) for the `spectral` (blue) and `matching` (grey) decoders, obtained by sweeping the abstention threshold (full coverage on the left). The spectral curve lies below matching's throughout and begins at a far lower full-coverage risk; matching's falls steeply as it abstains. Lines are means over the $n = 100{,}000$ pooled shots per decoder (reported scale); shaded bands are analytic 95% Bernoulli confidence intervals of the committed error rate, whose effective sample size at each point is (coverage × pooled shots). Generated from `summary.json`.

surable for every decoder rather than hiding it inside a single error number.

D. Calibration is not predicted by decoder sophistication

The confidence that gates abstention is useful only if it is calibrated, and the benchmark returns two calibration findings—one preserved, one new. First, across the eight structure-agnostic decoders, calibration is essentially uncorrelated with sophistication and not predictable from accuracy: pooled ECE spans a factor of nearly six, `neural` 0.0446 < `majority` 0.0498 < `lookup` 0.0558 < `fused` 0.0622 < `bp` 0.1238 < `topology` 0.2421 < `matching` 0.2620 < `rl` 0.2624, even though all eight are tied on logical error. Plain minimum-weight `matching`, whose defect-fraction confidence tracks accuracy poorly, is among the worst; the lightweight logistic-regression `neural` selector, whose confidence is a predicted-class probability, is the best of the agnostic family.

The second finding is the empirical face of Eq. (7): the spectral decoder's confidence is the *exact model posterior* of the chosen coset, so it is near-perfectly calibrated wherever the transfer operator matches the channel. The per-regime ECE (Table IV) bears this out precisely: spectral is the best-calibrated decoder under i.i.d. noise (ECE 0.00069), bias (0.00159), and bursts (0.02538)—one to three orders of magnitude below the agnostic decoders there—and nearly so under correlations (0.04859, just behind lookup's 0.04465). The single exception is the read-out regime, where its ECE jumps to 0.32359: there the decoder is confident in a posterior computed from a *corrupted* syndrome, so it is over-confident exactly as Eq. (7) predicts under model misspecification. This one regime inflates its pooled ECE to 0.0629 (Table I), competitive with but not better than the best agnostic decoders—a faithful summary that the per-regime view sharpens rather than hides. Treating calibration as a first-class, regime-resolved metric is what makes this visible: a deployment that can characterise its channel gets, for free, a decoder whose abstention is trustworthy precisely where its accuracy gains are real.

E. Structured noise separates regimes, not decoders

The per-(regime, decoder) breakdown (Table III; Sec. VI G) shows the noise regime to be the dominant axis of *difficulty*: the agnostic decoders' logical error rises from ≈ 0.002 under i.i.d. noise to ≈ 0.043 (biased), ≈ 0.17 (burst), ≈ 0.21 (correlated), and ≈ 0.34 (measurement error)—a more than hundredfold change driven entirely by structure, with read-out noise by far the hardest. Within each regime the eight agnostic decoders cluster inside their confidence intervals (no single one is significantly best), exactly as the line-equivalence (Sec. II C) predicts. The spectral decoder is the lone exception: it posts the lowest entry in the i.i.d., biased, burst, and correlated columns (0.00145, 0.02400, 0.11960, 0.16190), but it pulls *statistically* clear of the cluster only on the three structured data-noise columns, by margins of $z \geq 10.5$ (Table II); under i.i.d. noise its 0.00085 edge over the provably optimal leader is within sampling error ($z = 2.0$), exactly as the i.i.d. optimality of the leader [Eq. (4)] requires. It falls back *into* the cluster under read-out noise (0.33870, statistically tied with the leader's 0.34020), where the corrupted syndrome leaves no data-channel structure to exploit. The integrity flag `automorphism_invariant` is true, confirming the matching decoder respects the code's cyclic symmetry to within statistical tolerance.

F. Extended data

The aggregate numbers of Table I pool over all five regimes. The single source-of-truth artifact `summary.json` also stores the full per-(regime, decoder) breakdown and the swept curves, which we transcribe verbatim here so the regime-dominated structure of Sec. IV E can be read off directly. No number in these tables is rounded beyond the precision written by the run, and none is recomputed by hand.

Table III gives the logical error rate of every decoder in every regime: the more-than-hundredfold rise from i.i.d. (≈ 0.002) to measurement-error (≈ 0.34) noise, the tight within-regime clustering of the agnostic decoders that the line-equivalence (Sec. II C) predicts, and the spectral de-

coder breaking out of that cluster on the structured data-noise columns. Table IV gives the per-regime expected calibration error: the spectral decoder is best-calibrated by orders of magnitude where its model matches (ECE 0.00069 i.i.d., 0.00159 biased) and worst under read-out noise (0.32359), while `matching` is badly miscalibrated under bias (0.54934) but well calibrated under measurement error (0.13773), reinforcing that calibration must be measured rather than inferred from accuracy (Sec. IV D). Table V gives the coverage at the fixed operating threshold $\tau = 0.5$, making concrete that `matching`, `topology`, and `rl` abstain heavily under structure (matching commits on only 0.27555 of biased shots) while `spectral`, `majority`, `lookup`, `neural`, and `fused` always commit (Sec. IV C). Table VI lists the full logical-versus-physical threshold sweep underlying Fig. 3, and Table VII the risk–coverage operating points underlying Fig. 4.

V. DISCUSSION

In this work, we introduced the channel-adaptive spectral decoder, a truncated transfer operator on the defect graph that is the first member of a nine-decoder family to leave the accuracy tie—and it does so by exactly the mechanism the theory isolates. Three findings follow. First, the eight structure-agnostic decoders are tied on raw logical error not because the problem is saturated but because they share one i.i.d. target (the line-equivalence, Sec. II C); the moment a decoder models the channel [Eq. (4)], the tie breaks, by 20–44% on the structured data-noise regimes ($z > 10$). Second, the win is sharply localised and honestly bounded: it vanishes under i.i.d. noise and under read-out noise, where the channel-matched optimum [Eq. (4)] permits no data-channel advantage, and the benchmark's per-regime z-scores make that boundary visible rather than averaging it away. Third, calibration is a structural, regime-dependent property: across the agnostic family it spans nearly sixfold and is uncorrelated with sophistication, while the spectral decoder's posterior confidence is provably (Eq. (7)) and empirically near-perfectly calibrated wherever its model matches the channel, degrading only where it conditions on a corrupted syndrome. Selectivity and calibration as first-class metrics are what render all three visible at once.

Relation to prior work. Matching [5, 6, 8], belief propagation [10, 12], and neural and reinforcement-learning decoders [13, 14, 17, 19] are each studied in depth, typically optimising the logical error rate. The idea that a decoder should match the *noise model*—weighted or correlated matching, soft and belief-matching, and matching weights estimated online from calibration data [58]—is well established for surface codes [8, 12]; our contribution is to cast it as a *truncated transfer operator on the defect graph* with provable channel-matched optimality [Eqs. (4) and (6)] and a posterior confidence that is calibrated by construction [Eq. (7)], and to embed it in a benchmark that imports selective prediction [20–22] and calibration [23] from machine learning as first-class axes. The performance claim is deliberately scoped: a statistically significant advantage under *simulated* structured noise on toy codes. Where the literature reports learned decoders exceeding matching on hardware syndromes [19], we do not reproduce that regime; our point is that channel-awareness, not algorithmic sophistication within the i.i.d.-optimal class, is the lever.

A. Limitations

1. **Toy codes only.** The primary benchmark is a one-dimensional repetition code; the second is a surface-*like* toy lattice, not a stabilizer-exact CSS surface code, and we do not evaluate a full surface code at scale. The repetition code's binary coset and line geometry are exactly what make the agnostic decoders coincide; on a 2D code the transfer operator is higher-dimensional and the spectral truncation [Eq. (6)] becomes a genuine approximation we have not yet stress-tested.

2. **Simulated structured noise.** All noise is produced by parametric samplers, not measured from hardware. The biased, burst, correlated, and measurement-error regimes are plausible surrogates, but real device noise has correlations (leakage, crosstalk, drift) that our first-order Markov model and our samplers do not capture.

3. **The advantage requires a calibration sample.** The spectral decoder is *channel-adaptive*: it must be matched to the operating channel from a labelled calibration sample (here 12,000 shots, the same data the neural decoder trains on). Where the channel is unknown, drifting, or uncharacterisable, the advantage shrinks toward the i.i.d. tie; an online or unsupervised estimator of the transfer operator is future work.

4. **Single-distance focus.** The headline is at one distance and one physical error rate; the threshold curves sweep p at fixed distance, but we do not present a distance-scaling (sub-threshold) study, the decisive test for a true topological code.

5. **First-order model.** The transfer operator is a first-order (nearest-neighbour) Markov chain; longer-range correlations would require higher-order terms or a wider spectral truncation. The read-out regime exposes the model's blind spot: it cannot help when the conditioned syndrome is itself corrupt.

6. **Solo author, single environment.** All results were produced and audited by one author on a single CPU platform; independent replication on other hardware and by other groups remains to be done.

TABLE III. **Extended data: logical error rate by decoder and structured-noise regime.** Per-(regime, decoder) logical error rate on the repetition code at $p_{\text{eval}} = 0.12$ (distance 9; 20,000 trials per regime × decoder). Lower is better; **bold** marks the lowest entry in each regime column. The regime is the dominant axis of difficulty (every decoder rises from ≈ 0.002 at i.i.d. to ≈ 0.34 at measurement error), and the eight agnostic decoders cluster within their statistical intervals in each column; the channel-adaptive `spectral` decoder is the strictly lowest entry on the i.i.d., biased, burst, and correlated columns. Transcribed verbatim from `summary.json` (`by_regime`).

	Logical error rate by structured-noise regime				
Decoder	i.i.d.	biased	burst	correlated	meas. error
spectral	**0.00145**	**0.02400**	**0.11960**	**0.16190**	0.33870
majority	0.00180	0.04190	0.17040	0.20575	**0.33430**
matching	0.00185	0.04205	0.17300	0.20860	0.33675
bp	0.00190	0.04460	0.17260	0.20405	0.33825
topology	0.00240	0.04425	0.17160	0.20740	0.33680
lookup	0.00230	0.04285	0.17845	0.20695	0.34020
neural	0.00225	0.04485	0.17540	0.20920	0.33525
fused	0.00155	0.04305	0.17765	0.20355	0.33630
rl	0.00205	0.04145	0.16925	0.21265	0.34530

TABLE IV. **Extended data: expected calibration error by decoder and regime.** Per-(regime, decoder) ECE at $p_{\text{eval}} = 0.12$, computed over ten equal-width confidence bins (Sec. VI D). Lower is better; **bold** marks the lowest entry in each regime column. The `spectral` decoder's posterior confidence is best-calibrated by orders of magnitude where its transfer-operator model matches the channel (i.i.d., biased, burst) and worst under read-out noise, where it conditions on a corrupted syndrome (Eq. (7)). Across the agnostic family the best-calibrated decoder changes column to column—calibration is not predicted by logical accuracy. Transcribed verbatim from `summary.json` (`by_regime`).

	Expected calibration error by structured-noise regime				
Decoder	i.i.d.	biased	burst	correlated	meas. error
spectral	**0.00069**	**0.00159**	**0.02538**	0.04859	0.32359
majority	0.11771	0.20391	0.04327	0.04531	0.15122
matching	0.35161	0.54934	0.17186	0.13539	0.13773
bp	0.21106	0.33873	0.08092	0.07415	**0.08445**
topology	0.30558	0.54260	0.16660	0.13268	0.16811
lookup	0.17326	0.25190	0.03454	**0.04465**	0.12363
neural	0.12017	0.11923	0.09753	0.10719	0.20704
fused	0.08485	0.14600	0.10256	0.13242	0.22254
rl	0.31717	0.55282	0.19657	0.16496	0.12423

A deployable QEC decoder must do more than minimise an i.i.d. logical error rate: it must exploit the structure of the noise, know when to abstain, and report calibrated confidence. We showed that the textbook and learned decoders are statistically tied precisely because they all optimise the i.i.d. likelihood, and we introduced a channel-adaptive spectral decoder—a truncated transfer operator on the defect graph—that is Bayes-optimal for the structured channel, breaks the tie by 20–44% under biased, bursty, and correlated noise ($z > 10$), ties honestly where no advantage is possible, and reports a posterior confidence that is calibrated by construction. The most actionable message is that channel-awareness, not algorithmic sophistication within the i.i.d.-optimal class, is what separates decoders under realistic noise—and that it can be made provable, tractable, and self-calibrating. More broadly, the result illustrates why statistics is structural to quantum computing: when the device speaks only in samples, the winning decoder is the one that treats the noise channel as an object to be *inferred* rather than assumed, and reports calibrated uncertainty on its own decisions. The transfer operator plays the role a probabilistic surrogate plays elsewhere in scientific machine learning—a low-cost, mechanistically interpretable stand-in for an expensive stochastic process, fitted from calibration data and queried for downstream decisions—and its first-order Markov structure keeps that surrogate transparent enough to say exactly where and why it fails (the corrupted-syndrome regime). Extending the transfer-operator decoder to a full surface code at scale, to higher-order and online-estimated channel models, to hardware syndromes, and to a distance-scaling study—and coupling the calibration sample to active or drift-aware experimental design that spends labelled shots where the channel model is most uncertain—while keeping selectivity and calibration first-class, is the natural next step.

TABLE V. **Extended data: coverage at the fixed abstention threshold by decoder and regime.** Fraction of shots committed (not abstained) at threshold $\tau = 0.5$ for each (regime, decoder) at $p_{\text{eval}} = 0.12$. A value of 1.00000 means the decoder commits on every shot. `spectral`, `majority`, `lookup`, `neural`, and `fused` always commit (the spectral decoder's posterior confidence on the chosen coset is high); `matching`, `topology`, and `rl` abstain heavily under structured noise—the mechanism behind the risk–coverage gains of Fig. 4. Transcribed verbatim from `summary.json` (`by_regime`).

	Coverage at fixed abstention threshold $\tau = 0.5$				
Decoder	i.i.d.	biased	burst	correlated	meas. error
spectral	1.00000	1.00000	1.00000	1.00000	1.00000
majority	1.00000	1.00000	1.00000	1.00000	1.00000
matching	0.66510	0.27555	0.63375	0.63240	0.50600
bp	0.98955	0.90085	0.88950	0.88395	0.91095
topology	0.79620	0.41570	0.88335	0.90925	0.71200
lookup	1.00000	1.00000	1.00000	1.00000	1.00000
neural	1.00000	1.00000	1.00000	1.00000	1.00000
fused	1.00000	1.00000	1.00000	1.00000	1.00000
rl	0.69085	0.32140	0.63540	0.63280	0.52315

TABLE VI. **Extended data: logical-versus-physical threshold sweep.** Logical error rate as a function of physical error rate p under i.i.d. noise for the seven core decoders (including `spectral`), on the repetition code at distance 9 (20,000 trials per grid point). These are the curves plotted in Fig. 3; all decoders—including the channel-adaptive spectral decoder, calibrated here on the i.i.d. stream—stay well below the break-even line $p_L = p$ and overlap throughout, consistent with the i.i.d. tie (Eq. (4)). Transcribed verbatim from `summary.json` (`threshold_curves`).

p	spectral	majority	matching	bp	topology	lookup	fused
0.01	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.03	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.05	0.000100	0.000000	0.000000	0.000000	0.000000	0.000000	0.000100
0.07	0.000300	0.000050	0.000250	0.000050	0.000050	0.000000	0.000300
0.09	0.000800	0.000650	0.000650	0.000600	0.000550	0.000500	0.000650
0.11	0.001500	0.001400	0.001650	0.001500	0.001750	0.001000	0.001150
0.13	0.003200	0.004000	0.002750	0.002650	0.002800	0.003000	0.003600
0.15	0.005200	0.005700	0.004850	0.005500	0.005100	0.005700	0.005200
0.17	0.010400	0.009700	0.009400	0.009050	0.011150	0.009150	0.009800
0.19	0.017150	0.014950	0.015800	0.015750	0.016500	0.015950	0.016050
0.21	0.024750	0.023850	0.024650	0.021950	0.025000	0.023150	0.023900

VI. METHODS

A. Codes

The repetition code (`RepetitionCode`) has $n = d$ data bits and $d-1$ weight-two checks Z_iZ_{i+1}; the syndrome of an error $e \in \mathbb{F}_2^d$ is $s_i = e_i \oplus e_{i+1}$. The logical observable is the global parity $\sum_i e_i \bmod 2$. A correction $\hat{c}$ is counted as a logical failure unless the residual $r = e \oplus \hat{c}$ has zero syndrome (is a stabilizer) and zero logical value; a residual that does not even reproduce the syndrome is also a failure. The lookup decoder enumerates all 2^d errors to tabulate the minimum-weight leader per syndrome. The surface-like code (`SurfaceLikeCode`) places d^2 data bits on a grid with $(d-1)^2$ plaquette (weight-four) checks; defects are paired by Manhattan-distance matching and the logical observable is the parity of the left column. Both codes expose a bipartite Tanner graph via `networkx` [25].

B. Structured-noise samplers

Each sampler takes a seeded `numpy` generator [26] and returns a 0/1 error vector. *i.i.d.*: each bit flips at rate p. *biased*: even-index bits flip at $\min(1, \text{bias} \cdot p)$ and odd-index bits at p (default bias = 3). *burst*: scanning the line, with probability p a contiguous run of length 1 + Poisson(burst_len − 1) is flipped (default burst_len = 3). *correlated*: an i.i.d. seed pattern recruits each neighbour with probability corr (default 0.5) over two sweeps. *measurement_error*: i.i.d. data noise at rate p plus a per-check syndrome-flip mask at rate q (default 0.05), which the runner XORs into the measured syndrome.

C. Decoders

All nine decoders implement `decode(syndrome)` → `(correction, confidence)`; eight are structure-agnostic (they target the i.i.d. leader)

TABLE VII. **Extended data: risk–coverage operating points.** Committed logical error (risk) as a function of coverage for the `spectral` and `matching` decoders, obtained by sweeping the abstention threshold over the pooled $n = 100{,}000$ shots per decoder (reported scale). Each row is an actual operating point stored by the run; the two decoders are swept on their own confidence grids, so coverage values differ between blocks (matching's defect-fraction confidence takes few distinct values, so several thresholds yield the same point). These are the points plotted in Fig. 4. The spectral decoder begins far lower (risk 0.12913 at full coverage) and decreases near-monotonically with a small fluctuation near coverage 0.55; matching falls monotonically, as guaranteed under calibration by Eq. (10). Transcribed verbatim from `summary.json` (`risk_coverage`).

spectral		matching	
coverage	risk	coverage	risk
1.00000	0.12913	1.00000	0.15245
0.99463	0.12669	0.99467	0.15166
0.97695	0.11953	0.95571	0.15059
0.96990	0.11591	0.93968	0.15014
0.95134	0.11257	0.93968	0.15014
0.93382	0.10905	0.81382	0.12481
0.92184	0.10725	0.72108	0.12295
0.91004	0.10545	0.72108	0.12295
0.90637	0.10465	0.72108	0.12295
0.90397	0.10431	0.54256	0.07443
0.89432	0.10269	0.54256	0.07443
0.88965	0.10194	0.54256	0.07443
0.86813	0.09728	0.24834	0.01079
0.85310	0.09335	0.24834	0.01079
0.82721	0.09111	0.24834	0.01079
0.80779	0.08581	0.24834	0.01079
0.78214	0.08267	0.24834	0.01079
0.67545	0.08017	0.24834	0.01079
0.64173	0.07537	0.24834	0.01079
0.54786	0.08429	0.24834	0.01079
0.04257	0.02161	0.24834	0.01079

and the ninth (*Spectral*) is channel-adaptive. *Lookup*: a table from syndrome to minimum-weight leader; confidence boosted toward 1 since the leader is exact. *Majority*: cumulative XOR of the syndrome recovers a flip chain up to a global offset, choosing the lower-weight coset representative; confidence is the normalised majority margin. *Matching*: on the line, defects are paired left-to-right with boundary closure for an odd defect, returning the lighter coset; on the surface, a complete graph on lit plaquettes weighted by negative Manhattan distance is solved with `max_weight_matching` [9, 25] and an L-shaped path is flipped between matched pairs. Confidence is a defect-fraction heuristic $\exp(-2\, n_{\mathrm{def}}/n_{\mathrm{checks}})$, damped when a defect is unpaired. *BP*: a few rounds of min-sum message passing on the Tanner graph with log-likelihood priors; if the thresholded beliefs do not reproduce the syndrome, it defers to matching. *Neural*: a two-stage learned coset selector—a syndrome-determined base chain plus a `scikit-learn` logistic-regression classifier [27] predicting the logical coset of the true correction relative to that chain; the output is always syndrome-consistent and confidence is the predicted-class probability. *Topology*: matching with confidence recalibrated by defect-geometry features (count, span, gap structure). *Fused*: matching and neural both propose; agreement boosts confidence, disagreement defers to the more confident proposal with damped confidence. *RL*: a contextual bandit [28, 29] whose context is a coarse defect-count bucket and whose actions are candidate decoders; it keeps per-context running success estimates, plays ε-greedy during a warm-up stream, and at evaluation selects the highest-value decoder per context. *Spectral* (`SpectralDecoder`): the channel-adaptive contribution. From a calibration sample $\{e^{(j)}\}$ of the operating channel it estimates a truncated transfer operator (Eq. (6))—an initial law $\hat{\pi}_0$ and per-site 2×2 transition matrices $\hat{T}_i$, each by normalised, Laplace-smoothed counts. Given a syndrome it forms the two coset representatives $\{c_0, c_0 \oplus \mathbf{1}\}$ (the matching base chain and its logical partner) and returns the one of larger model log-likelihood $\log \hat{\pi}_0(e_0) + \sum_i \log \hat{T}_i(e_i, e_{i+1})$; the confidence is the exact model posterior $\sigma(|\Delta\ell|)$ of the chosen coset. On the surface-like code the operator degrades to a per-bit field model (Bayes-optimal for the inhomogeneous-marginal channel). The decoder is Bayes-optimal for a Markov channel (Eq. (4)) and calibrated by construction (Eq. (7)).

D. Metrics and confidence intervals

The logical error rate is the mean of per-shot failure indicators; its 95% confidence interval is $1.96\,\hat{\sigma}/\sqrt{n}$ with $\hat{\sigma}$ the sample standard deviation (Bernoulli, so $\hat{\sigma} \approx \sqrt{\bar{p}(1-\bar{p})}$). The expected calibration error [23, 24] partitions shots into ten equal-width confidence bins and averages $|(\text{mean confidence}) - (\text{mean accuracy})|$ weighted by bin occupancy, where accuracy is 1−failure. The risk–coverage curve sweeps an abstention threshold from the minimum to the maximum observed confidence; at each threshold, coverage is the committed fraction and risk is the logical error rate on committed shots. The curve is reported in descending-coverage order and a monotonicity check verifies that risk is (weakly) non-increasing as coverage shrinks.

E. Integrity: automorphism invariance and the audit gate

The repetition code is invariant under cyclic relabeling of its data bits (ring symmetry); a faithful decoder must yield the same logical error rate on cyclically rotated errors. The runner checks this for the matching decoder over shifts $\{1, 2, \lfloor d/2 \rfloor\}$ within a statistical tolerance and exposes it as the `automorphism_invariant` flag. A separate audit module enforces that every reported number

traces to a value in the generated `summary.json` (the single source of truth), requires the central spectral-versus-leader claim to be statistically evidenced (per-regime z-scores), and rejects unevidenced superlatives; this discipline runs in continuous integration.

F. Experimental protocol

A run builds the code, fits the neural decoder on a mixed-regime training stream, and warms up the RL bandit on its own Monte-Carlo stream. Then, for each (regime, decoder), it Monte-Carlos `trials` shots at the evaluation physical error rate p_{eval}, records the per-shot failure and confidence, and aggregates the metrics above. The channel-adaptive step is here: immediately before each regime is evaluated, the spectral decoder's transfer operator is re-estimated from that regime's calibration sample (the per-regime slice of the training stream—the same labelled data the neural decoder sees, so the comparison is fair). Threshold curves sweep p for the seven core decoders under i.i.d. noise, with the spectral decoder calibrated on the i.i.d. stream (where it provably ties the leader); the risk–coverage curve is built from the spectral and matching decoders' pooled confidences. Seeds, platform, library versions, runtime, and peak memory are recorded as provenance in `summary.json`. All numbers in this paper are from the *reported-scale* configuration: distance 9, 20,000 trials per (regime, decoder) (100,000 pooled shots per decoder), 12,000 training/calibration shots, $p_{\text{eval}} = 0.12$, an eleven-point physical-error grid, run end-to-end in 277 s at 206 MB peak memory on a single CPU. A faster smoke configuration (distance 5, 1,500 trials, ≈ 10 s) is provided for continuous integration.

Statistics. Each reported logical error rate is a mean of n independent Bernoulli failure indicators; the $\pm 95\%$ interval is the analytic Wald half-width $1.96\sqrt{\hat{p}(1-\hat{p})/n}$ (Eq. (8)). Two decoders are called statistically indistinguishable when their means differ by less than this half-width, and a difference is called significant when the two-sample $z = (\hat{p}_A - \hat{p}_B)/\sqrt{\hat{p}_A(1-\hat{p}_A)/n + \hat{p}_B(1-\hat{p}_B)/n}$ exceeds 3. This $z > 3$ rule is the operative, audited significance threshold throughout, deliberately more conservative than the 95% two-sample test of Eq. (8) ($z > 1.96$); a nominal $\sim 2\sigma$ gap such as the i.i.d. spectral-versus-leader difference ($z = 2.0$) is therefore reported as a tie. The per-regime spectral-versus-leader z-scores are written to `summary.json` and audited. The neural classifier's training stream (12,000 shots) is generated independently of the evaluation stream; all randomness derives from a single master seed (`seed` = 0). The exact run provenance—platform `macOS-26.5.2-arm64`, `python` 3.9.6, `numpy` 1.26.4, `scipy` 1.13.1, `networkx` 3.2.1, `scikit-learn` 1.6.1, `matplotlib` 3.9.4—is stored alongside the results. Software: `numpy` [26], `scipy` [30], `networkx` [25], `scikit-learn` [27], `matplotlib` [31]; CPU only.

G. Per-regime results

The per-(regime, decoder) logical error rate, ECE, and coverage at p_{eval} are written in full to `summary.json` under `by_regime`, and the per-regime spectral-versus-leader reductions and z-scores under `headline.spectral_by_regime`. At reported scale the qualitative pattern is regime-dominated for the agnostic decoders: i.i.d. noise yields ≈ 0.002 logical error; biased ≈ 0.043; burst ≈ 0.17; correlated ≈ 0.21; measurement-error ≈ 0.34 (the hardest regime, since syndrome read-out flips corrupt the very signal the decoders condition on). The spectral decoder departs from this cluster on the three structured data-noise regimes (0.024 biased, 0.120 burst, 0.162 correlated) and rejoins it under i.i.d. and read-out noise. Coverage varies strongly by decoder: matching's coverage at the fixed operating threshold ranges from 0.28 (biased) to 0.67 (i.i.d.); topology and the RL selector also abstain substantially; spectral, majority, lookup, neural, and fused commit fully (coverage 1.0) in every regime.

DATA AND CODE AVAILABILITY

This study uses no external datasets. All numerical values, tables, and figures are generated by the accompanying scripts from fixed random seeds and written to a single source-of-truth artifact (`results/summary.json`); they are reproduced by running the pipeline. The code that implements every code, noise model, decoder, metric, and the audit gate, and that regenerates every figure, table, and reported number, accompanies this manuscript in `submission/code` and is intended for public release upon publication. All experiments are reproducible on commodity hardware; runtime and memory are reported for each benchmark.

[1] Peter W. Shor. Scheme for reducing decoherence in quantum computer memory. *Physical Review A*, 52(4):R2493–R2496, 1995.

[2] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2010.

[3] Barbara M. Terhal. Quantum error correction for quantum memories. *Reviews of Modern Physics*, 87(2):307–346, 2015.

[4] A. Yu. Kitaev. Fault-tolerant quantum computation by anyons. *Annals of Physics*, 303(1):2–30, 2003.

[5] Eric Dennis, Alexei Kitaev, Andrew Landahl, and John Preskill. Topological quantum memory. *Journal of Math-*

ematical Physics, 43(9):4452–4505, 2002.

[6] Austin G. Fowler, Matteo Mariantoni, John M. Martinis, and Andrew N. Cleland. Surface codes: Towards practical large-scale quantum computation. *Physical Review A*, 86(3):032324, 2012.

[7] Google Quantum AI. Suppressing quantum errors by scaling a surface code logical qubit. *Nature*, 614(7949): 676–681, 2023.

[8] 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.

[9] Jack Edmonds. Paths, trees, and flowers. *Canadian Journal of Mathematics*, 17:449–467, 1965.

[10] David Poulin and Yeojin Chung. On the iterative decoding of sparse quantum codes. *Quantum Information & Computation*, 8(10):987–1000, 2008.

[11] Pavel Panteleev and Gleb Kalachev. Degenerate quantum LDPC codes with good finite length performance. *Quantum*, 5:585, 2021.

[12] Joschka Roffe, David R. White, Simon Burton, and Earl Campbell. Decoding across the quantum low-density parity-check code landscape. *Physical Review Research*, 2(4):043423, 2020.

[13] Giacomo Torlai and Roger G. Melko. Neural decoder for topological codes. *Physical Review Letters*, 119(3): 030501, 2017.

[14] Savvas Varsamopoulos, Ben Criger, and Koen Bertels. Decoding small surface codes with feedforward neural networks. *Quantum Science and Technology*, 3(1):015004, 2017.

[15] Paul Baireuther, Thomas E. O'Brien, Brian Tarasinski, and Carlo W. J. Beenakker. Machine-learning-assisted correction of correlated qubit errors in a topological code. *Quantum*, 2:48, 2018.

[16] Stefan Krastanov and Liang Jiang. Deep neural network probabilistic decoder for stabilizer codes. *Scientific Reports*, 7(1):11003, 2017.

[17] Ryan Sweke, Markus S. Kesselring, Evert P. L. van Nieuwenburg, and Jens Eisert. Reinforcement learning decoders for fault-tolerant quantum computation. *Machine Learning: Science and Technology*, 2(2):025005, 2020.

[18] Philip Andreasson, Joel Johansson, Simon Liljestrand, and Mats Granath. Quantum error correction for the toric code using deep reinforcement learning. *Quantum*, 3:183, 2019.

[19] Johannes Bausch, Andrew W. Senior, Francisco J. H. Heras, et al. Learning high-accuracy error decoding for quantum processors. *Nature*, 635(8040):834–840, 2024.

[20] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16 (1):41–46, 1970.

[21] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

[22] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

[23] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[24] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[25] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy)*, pages 11–15, 2008.

[26] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, et al. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

[27] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[28] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670, 2010.

[29] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

[30] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020.

[31] John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[32] Molena Huynh. Query-Efficient Quantum Approximate Optimization via Graph-Conditioned Trust Regions. *arXiv preprint arXiv:2604.24803*, 2026. doi: 10.48550/arXiv.2604.24803.

[33] Molena Huynh. Graph-conditioned trust regions for uncertainty-calibrated quantum approximate optimization (extended). *Manuscript in preparation*, 2026.

[34] Molena Huynh. The measurement cost of warm-started low-depth QAOA. *Manuscript in preparation*, 2026.

[35] Molena Huynh. Certified query budgets for the quantum approximate optimization algorithm. *Manuscript in preparation*, 2026.

[36] Molena Huynh. Topology-conditioned QAOA parameter transfer for budgeted graph optimization (companion). *Manuscript in preparation*, 2026.

[37] Molena Huynh. Operator-spectral truncated priors for query-efficient QAOA parameter search. *Manuscript in preparation*, 2026.

[38] Molena Huynh. Spectral-truncation graph kernels for QAOA warm starts: topology-conditioned schedule transfer beyond depth one. *Manuscript in preparation*, 2026.

[39] Molena Huynh. Zero-overhead cancellation of the leading Trotter error via commutator-graph scheduling. *Manuscript in preparation*, 2026.

[40] Molena Huynh. Advancing the product-formula error–cost frontier with a Lie-algebraic spectral truncation of the residual. *Manuscript in preparation*, 2026.

[41] Molena Huynh. Learning to decode correlated quantum errors: a topological graph neural decoder that surpasses minimum-weight matching under structured circuit-level noise. *Manuscript in preparation*, 2026.

[42] M. Nguyen and Naihuan Jing. Zassenhaus expansion in solving the Schrödinger equation. *arXiv preprint arXiv:2505.09441*, 2025. doi:10.48550/arXiv.2505.09441.

[43] Naihuan Jing and M. Nguyen. Complexity bounds for Hamiltonian simulation in unitary representations. *arXiv preprint arXiv:2603.07231*, 2026. doi:

10.48550/arXiv.2603.07231.
[44] A. R. Calderbank and Peter W. Shor. Good quantum error-correcting codes exist. *Physical Review A*, 54(2): 1098–1105, 1996.
[45] Andrew M. Steane. Error correcting codes in quantum theory. *Physical Review Letters*, 77(5):793–797, 1996.
[46] Daniel Gottesman. Stabilizer codes and quantum error correction. PhD thesis, California Institute of Technology, 1997. doi:10.7907/rzr7-dt72.
[47] Chenyang Wang, Jim Harrington, and John Preskill. Confinement-Higgs transition in a disordered gauge theory and the accuracy threshold for quantum memory. *Annals of Physics*, 303(1):31–58, 2003.
[48] David J. C. MacKay. Good error-correcting codes based on very sparse matrices. *IEEE Transactions on Information Theory*, 45(2):399–431, 1999.
[49] Austin G. Fowler. Minimum weight perfect matching of fault-tolerant topological quantum error correction in average $O(1)$ parallel time. *Quantum Information & Computation*, 15(1–2):145–158, 2015.
[50] Oscar Higgott and Craig Gidney. Sparse blossom: correcting a million errors per core second with minimum-weight matching. *Quantum*, 9:1600, 2025.
[51] Nicolas Delfosse and Naomi H. Nickerson. Almost-linear time decoding algorithm for topological codes. *Quantum*, 5:595, 2021.
[52] Sergey Bravyi, Martin Suchara, and Alexander Vargo. Efficient algorithms for maximum likelihood decoding in the surface code. *Physical Review A*, 90(3):032326, 2014.
[53] Andrew S. Darmawan and David Poulin. Tensor-network simulations of the surface code under realistic noise. *Physical Review Letters*, 119(4):040502, 2017.
[54] Christopher T. Chubb. General tensor network decoding of 2D Pauli codes. *arXiv preprint arXiv:2101.04125*, 2021. doi:10.48550/arXiv.2101.04125.
[55] David K. Tuckett, Stephen D. Bartlett, and Steven T. Flammia. Ultrahigh error threshold for surface codes with biased noise. *Physical Review Letters*, 120(5):050505, 2018.
[56] J. Pablo Bonilla Ataides, David K. Tuckett, Stephen D. Bartlett, Steven T. Flammia, and Benjamin J. Brown. The XZZX surface code. *Nature Communications*, 12(1): 2172, 2021.
[57] Konstantin Tiurev, Peter-Jan H. S. Derks, Joschka Roffe, Jens Eisert, and Jan-Michael Reiner. Correcting non-independent and non-identically distributed errors with surface codes. *Quantum*, 7:1123, 2023.
[58] 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(1): 1800012, 2018.
[59] Joel J. Wallman and Joseph Emerson. Noise tailoring for scalable quantum computation via randomized compiling. *Physical Review A*, 94(5):052325, 2016.
[60] Moritz Lange, Pontus Havström, Basudha Srivastava, Isak Bengtsson, Valdemar Bergentall, Karl Hammar, Olivia Heuts, Evert van Nieuwenburg, and Mats Granath. Data-driven decoding of quantum error correcting codes using graph neural networks. *Physical Review Research*, 7(2):023181, 2025.
[61] Sergey Bravyi, Andrew W. Cross, Jay M. Gambetta, Dmitri Maslov, Patrick Rall, and Theodore J. Yoder. High-threshold and low-overhead fault-tolerant quantum memory. *Nature*, 627(8005):778–782, 2024.
[62] Google Quantum AI. Exponential suppression of bit or phase errors with cyclic error correction. *Nature*, 595 (7867):383–387, 2021.
[63] Google Quantum AI. Quantum error correction below the surface code threshold. *Nature*, 638(8052):920–926, 2025.

Appendix A: Supplementary information: a from-first-principles construction

This appendix is a self-contained derivation of every concept the accompanying code implements, written to be read top-to-bottom by someone who has seen linear algebra and basic probability but *not* quantum error correction, stabilizer codes, minimum-weight matching, belief propagation, calibration, or selective prediction. Every symbol used in the code is defined, and each subsection names the source file (under `submission/code/src/topoqec/`) it maps to, so theory and implementation can be read side by side. It expands on, rather than repeats, the Methods (Sec. VI); where a definition already appears there, we cross-reference it.

1. The problem in one paragraph

A quantum error-correcting (QEC) code stores one logical bit redundantly across many physical bits. Noise flips some physical bits; we may not look at the bits directly (that would destroy the quantum state), so instead we measure a few parity checks whose outcomes form a *syndrome*. A *decoder* takes the syndrome and guesses a correction. It succeeds if (correction $\oplus$ error) is a do-nothing operator (a stabilizer) and fails if that residual flips the logical bit; the usual figure of merit is the *logical error rate*. The central point of this work is that the textbook and learned decoders all optimise the *i.i.d.* likelihood (minimum Hamming weight), so on a repetition code they are statistically *tied*—and the way to do better is to model the *structure* of the noise. We introduce a channel-adaptive *spectral* decoder (a truncated transfer operator on the defect graph) that does exactly that, beating the i.i.d.-optimal leader by 20–44% under biased, bursty, and correlated noise while honestly tying it where no advantage is possible. We further argue that two axes usually ignored both matter: *selective decoding*—letting the decoder *abstain* when unsure, so a slower fallback can take over—and *calibration*—whether the decoder's stated confidence matches how often it is right (the spectral decoder's confidence is the exact model posterior, so it is calibrated by construction). We build one benchmark that measures all three (logical error, calibration, risk–coverage) for nine decoders across five structured-noise regimes, with honest 95% confidence intervals and per-regime significance tests.

2. Codes, syndromes, and logical failure (`codes.py`)

Repetition code (`RepetitionCode`). The primary benchmark. A logical bit is encoded in d data bits over $\mathbb{F}_2 = \{0, 1\}$ (the field of integers mod 2; XOR is "+"). The code has $d-1$ weight-two parity checks Z_iZ_{i+1}. An error is a vector $e \in \mathbb{F}_2^d$ ($e_i = 1$ means data bit i is flipped), with syndrome

$$s_i = e_i \oplus e_{i+1}, \qquad i = 0, \ldots, d-2 \qquad (\texttt{.syndrome()}). \tag{A1}$$

Intuitively s_i fires exactly where an error *chain* crosses link i, so the lit syndrome bits sit at the *endpoints* of error chains; decoding is a matching problem on a one-dimensional line of links. The single logical observable is the global parity $L(e) = \left(\sum_i e_i\right) \bmod 2$ (`.logical_value()`). A correction $\hat{c}$ is a *logical failure* (`.logical_error()`) unless the residual $r = e \oplus \hat{c}$ both (a) reproduces zero syndrome (r is a stabilizer, acting trivially) and (b) has zero logical value; a correction that does not reproduce the measured syndrome is automatically a failure. This is the precise definition of "the code protected the bit" used in Methods (Sec. VI A).

Lookup table. For small d we enumerate all 2^d errors (`.all_errors()`) and store, for every syndrome, the minimum-weight error that produces it—the coset leader. This optimal table is the gold-standard decoder for small d and the yardstick everything else is compared against (Sec. A 5, `LookupDecoder`).

Surface-like code (`SurfaceLikeCode`). A second, two-dimensional code: $d \times d$ data bits on a grid with $(d-1)^2$ weight-four plaquette checks; a lit plaquette is a *defect*, and defects are paired by Manhattan (L_1) distance. The logical observable is the parity of the left column. It exposes the same defect geometry in 2D so the matching and topology decoders generalise, but it is a topology-*faithful* toy, *not* a stabilizer-exact CSS surface code—which we state honestly (Sec. V A). Both codes expose a bipartite Tanner graph via `networkx` (`.tanner_graph()`) so graph-based decoders operate on the code topology directly.

3. Structured noise: why i.i.d. is not enough (`noise.py`)

Real hardware noise is *not* independent and identically distributed. Five seeded samplers, each taking a `numpy Generator` and returning a 0/1 error vector, model the structure that matters (`noise.sample()` dispatches on the regime name; see Sec. VI B for the parameters):

- `iid`: each bit flips independently with probability p (`sample_iid`).
- `biased`: even-index bits flip at $\min(1, \text{bias} \cdot p)$, odd-index at p (default bias = 3)—a coarse Pauli-bias surrogate (`sample_biased`).
- `burst`: scanning the line, with probability p flip a contiguous run of length $1 + \text{Poisson}(\ell - 1)$, with $\ell =$ burst_len (default 3)—spatially clustered errors (`sample_burst`).
- `correlated`: an i.i.d. seed pattern recruits each neighbour with probability corr (default 0.5) over two sweeps—neighbour correlations (`sample_correlated`).
- `measurement_error`: i.i.d. data noise at rate p *plus* a per-check syndrome read-out flip mask at rate q (default 0.05); the runner XORs that mask into the *measured* syndrome, so the decoder conditions on a corrupted signal (`sample_measurement_error`).

These are precisely the regimes where topology- or read-out-aware decoding might beat a memoryless majority vote—and they turn out to dominate the difficulty far more than the choice of decoder (Sec. A 9).

4. What a decoder returns: the unified interface (`decoders.py`)

Every decoder implements one interface,

$$\texttt{decode(syndrome)} \longrightarrow (\text{correction, confidence}), \tag{A2}$$

with correction in $\mathbb{F}_2^n$ and confidence in $[0, 1]$. The single interface is what makes the nine decoders comparable on one footing and what makes abstention (Sec. A 7) and calibration (Sec. A 8) definable uniformly. Two helpers recur: `_chain_from_defects()` turns a set of paired defect endpoints into a flip chain, and the defect-fraction confidence heuristic

$$\texttt{\_defect\_confidence}(n_{\text{def}}, n_{\text{checks}}) = e^{-2\, n_{\text{def}}/n_{\text{checks}}} \tag{A3}$$

makes a decoder less sure as more checks light up.

5. The decoder family: textbook to learned (`decoders.py`)

- `LookupDecoder`: the optimal minimum-weight coset leader from the enumerated table; confidence boosted toward 1 since the leader is exact. The gold standard for small d.
- `MajorityDecoder`: cumulative XOR (prefix sum mod 2) of the syndrome recovers a flip chain up to a global offset; pick the lower-weight of the two cosets; confidence is the normalised majority margin.
- `MatchingDecoder`: minimum-weight perfect matching of defects. On the line, pair lit syndrome bits left-to-right (with a boundary closure for an odd count) and return the lighter coset. On the surface, build a complete graph on lit plaquettes

weighted by negative Manhattan distance and call `networkx max_weight_matching` (Edmonds' blossom algorithm [9]), then flip an L-shaped path between matched pairs. Confidence is the defect-fraction heuristic, damped when a defect is unpaired. This is the standard surface-code decoder, in toy form.

- `BPDecoder`: a few rounds of min-sum belief propagation on the Tanner graph with log-likelihood priors; if the thresholded beliefs do not reproduce the syndrome, defer to matching. (BP underpins quantum LDPC decoding [10, 12].)

- `NeuralDecoder`: a learned two-stage coset selector. Stage 1: a syndrome-determined base chain (always syndrome-consistent). Stage 2: a `scikit-learn LogisticRegression` that predicts the logical coset (0/1) of the true correction relative to that base chain. The output is therefore always syndrome-valid and the confidence is the predicted-class probability—which is exactly why it tends to be well calibrated (Sec. A 8). It is fit on a mixed-regime training stream.

- `TopologyDecoder`: matching, but with confidence *recalibrated* by defect-geometry features (count, span, gap structure).

- `FusedDecoder`: matching and neural both propose; agreement boosts confidence, disagreement defers to the more confident proposal with damped confidence (confidence fusion).

- `RLDecoder`: a contextual bandit (reinforcement learning). Context = a coarse defect-count bucket; actions = candidate decoders. It keeps per-context running success estimates, plays ε-greedy during a warm-up stream, and at evaluation picks the highest-value decoder per context.

- `SpectralDecoder`: the channel-adaptive contribution. It `fit()`s a truncated transfer operator $P_\theta(e) = \hat{\pi}_0(e_0)\prod_i \hat{T}_i(e_i, e_{i+1})$ to a calibration sample of the operating channel (each $\hat{T}_i$ a normalised, Laplace-smoothed 2×2 count matrix), then on each syndrome compares the model log-likelihood of the two coset representatives $\{c_0, c_0\oplus\mathbf{1}\}$ and returns the more probable one. Confidence is the exact model posterior $\sigma(|\Delta\ell|)$. Unlike the eight structure-agnostic decoders, it does not target minimum Hamming weight; it targets maximum probability under the *learned* channel, which is why it beats the leader under structured noise (Eq. (4)) and reduces to it under i.i.d. noise.

`build_decoders()` wires the family together for a given code; the same list is summarised in Sec. VI C.

6. Metric 1: logical error rate and honest intervals (`metrics.py`)

For each (regime, decoder) we Monte-Carlo many shots and record a per-shot failure indicator $F \in \{0, 1\}$. The logical error rate is $\hat{p} = \text{mean}(F)$ (`logical_error_rate()`). Because F is Bernoulli, its sample standard deviation is $s = \sqrt{\hat{p}(1-\hat{p})}$ and a 95% confidence-interval half-width is

$$\text{ci}_{95} = 1.96\,\frac{s}{\sqrt{n}} \qquad \texttt{(summarize())}, \tag{A4}$$

with n the number of shots. This is the whole point of honesty: two decoders whose means differ by *less* than a CI width are *not* separated at 95%. At reported scale (n = 100,000 pooled shots) the eight structure-agnostic decoders span only 0.1508–0.1542 with $\text{ci}_{95} \approx 0.0022$, so their 0.0033 best-to-worst spread is below 1.5 interval widths: they are statistically *tied* (Sec. IV B). The channel-adaptive spectral decoder, by contrast, sits at 0.1291—about ten interval widths below the cluster—so its advantage is real, not sampling noise.

7. Metric 2: selective decoding and the risk–coverage curve (`metrics.py`)

A decoder inside a fault-tolerant loop need not commit to every syndrome: it can *abstain* (flag a shot unreliable) so a fallback handles it—the QEC analogue of *selective prediction* in machine learning [20, 22]. Fix a confidence threshold τ and commit only on shots with confidence $\geq \tau$. Then

$$\text{coverage}(\tau) = \text{fraction of shots committed}, \tag{A5}$$

$$\text{risk}(\tau) = \text{logical error rate on committed shots}. \tag{A6}$$

Sweeping τ from the smallest to the largest observed confidence traces the *risk–coverage curve* (`risk_coverage_curve()`), reported in descending-coverage order. As we abstain more (coverage shrinks), risk should fall: we drop the shots we were least sure about. `is_monotone_risk_coverage()` programmatically verifies risk is (weakly) non-increasing as coverage shrinks. Empirically the effect is large: matching's committed error falls from 0.1525 at full coverage to 0.0744 at ≈ 0.54 coverage and 0.0108 at ≈ 0.25 coverage (Sec. IV C). The formal reason is Eq. (10): under perfectly calibrated confidence C (so $\Pr(\text{success} \mid C = c) = c$), the committed risk is $1 - \mathbb{E}[C \mid C \geq \tau]$, which can only decrease as τ rises—but *only* to the extent that confidence is calibrated. Hence Sec. A 8.

8. Metric 3: calibration (expected calibration error) (`metrics.py`)

A confidence score is useful for abstention only if it is *honest* about its own accuracy. The expected calibration error (`expected_calibration_error()`) partitions the shots into ten equal-width confidence bins; within each it compares the mean stated confidence to the mean realised accuracy (= 1−failure), takes the absolute gap, and averages those gaps weighted by bin occupancy:

$$\mathrm{ECE} = \sum_{\text{bins}} \frac{n_{\text{bin}}}{n} \left| \overline{\text{conf}}(\text{bin}) - \overline{\text{acc}}(\text{bin}) \right|. \qquad \text{(A7)}$$

ECE = 0 means "when the decoder says 0.8, it is right 80% of the time". Two findings live here. Across the eight structure-agnostic decoders, ECE spans a factor of nearly six (0.0446–0.2624) and is *uncorrelated* with sophistication (Sec. IV D)—it must be *measured*, not assumed from accuracy or complexity. The spectral decoder is the interesting case: because its confidence is the exact model posterior (Eq. (7)), its ECE is near zero wherever the transfer operator matches the channel (0.00069 i.i.d., 0.00159 biased, 0.02538 burst—the best in those regimes by orders of magnitude) and large only under read-out noise (0.32359), where it conditions on a corrupted syndrome. Its pooled ECE (0.0629) is thus competitive with, but not better than, the best agnostic decoders—a number the per-regime view (Table IV) explains exactly.

9. Structured noise sets the difficulty; channel-awareness sets the winner (`runner.by_regime`)

The per-(regime, decoder) breakdown shows the noise regime to be the dominant axis of *difficulty* (Sec. IV E, Sec. VI G): logical error rises ≈ 100-fold from ≈ 0.002 (i.i.d.) to ≈ 0.043 (biased), ≈ 0.17 (burst), ≈ 0.21 (correlated), and ≈ 0.34 (measurement error, the hardest—read-out flips corrupt the very signal decoders condition on). Within each regime the eight agnostic decoders cluster inside their (now tight) CIs. The spectral decoder is the exception: it is the strictly lowest entry on the i.i.d., biased, burst, and correlated columns (0.00145, 0.02400, 0.11960, 0.16190), beating the leader by $z > 10$ on the three structured data-noise regimes, and rejoins the cluster only under read-out noise. Coverage varies strongly: matching, topology, and RL abstain heavily under structure, whereas spectral, majority, lookup, neural, and fused commit fully.

10. Integrity: automorphism invariance and the audit gate (`symmetry.py`, `audit.py`)

Automorphism invariance. The repetition code is invariant under cyclic relabeling of its data bits (a ring symmetry). `cyclic_relabel(error, shift)` rotates an error; `logical_error_rate_under_relabeling()` and `is_automorphism_invariant()` check that the matching decoder yields the *same* logical error rate on rotated errors (over shifts $\{1, 2, \lfloor d/2 \rfloor\}$, within a statistical tolerance). A faithful decoder must respect the code's symmetry; the run exposes this as the `automorphism_invariant` integrity flag (true at scale; Sec. VI E).

Audit gate. `audit.py` enforces scientific hygiene: `audit_numbers()` checks every claimed number traces to a value in the generated `summary.json` (the single source of truth); `audit_reproducibility_phrase()` checks the mandated reproducibility statement is present; and `audit_forbidden_claims()` rejects unevidenced superlatives. The driver `scripts/audit_claims.py` additionally requires the central spectral-versus-leader advantage to be statistically significant ($z > 3$) in at least one structured regime and not significant under i.i.d. noise, runs in continuous integration, and exits non-zero on any violation—so the paper cannot drift from the artifact.

11. The single source of truth and how to reproduce (`runner.py`)

`runner.run()` ties it together: build the code; fit the neural decoder on a mixed-regime training stream; warm up the RL bandit on its own Monte-Carlo stream; for each (regime, decoder) re-fit the spectral decoder's transfer operator to that regime's calibration sample, then Monte-Carlo `trials` shots at p_{eval} and record per-shot failure and confidence; sweep p over a grid (i.i.d.) for the threshold curves; build the risk–coverage curve from the spectral and matching confidences; check automorphism invariance. It writes `results/summary.json`—the single source of truth from which every table, figure, and reported number is generated. Nothing is typed by hand. Then:

- `scripts/make_tables.py`: `summary.json` → `results/main_results.tex` and `results/paper_tables.tex` (`booktabs`).
- `scripts/make_figures.py`: `summary.json` → the four figure PDFs `fig_schematic`, `fig_perregime`, `fig_threshold`, and `fig_risk_coverage`, as vector PDFs with embedded editable text (`pdf.fonttype=42`), sans-serif mathtext, the colourblind-safe Okabe–Ito palette, despined axes, and 95% CI bands.
- `scripts/audit_claims.py`: the readiness gate of Sec. A 10.

To reproduce everything:

```
pip install -e .
export KMP_DUPLICATE_LIB_OK=TRUE
bash scripts/reproduce_all.sh full   # d=9
bash scripts/reproduce_all.sh        # CI smoke
```

Reported run: repetition code, distance 9, five regimes, 20,000 trials/regime (100,000 pooled shots/decoder), 12,000 training/calibration shots, $p_{\text{eval}} = 0.12$, seed 0 (Sec. VI F). The figure and table artifacts are copied verbatim into `submission/` to build the paper.

12. What the result means (and what it does not)

The eight structure-agnostic decoders are statistically *indistinguishable* on raw logical error for a distance-d repetition code—as expected, since they all approximate the i.i.d.-optimal coset leader (the line-equivalence, Sec. II C). The contribution is the decoder that escapes that ceiling by modelling the channel: the channel-adaptive spectral decoder is Bayes-optimal for the structured channel (Eq. (4)) and cuts the logical error by 20–44% ($z > 10$) under biased, bursty, and correlated noise—an evidenced, statistically significant win—while honestly tying the leader under i.i.d. and read-out noise, where no single-shot data-channel advantage exists. The separable contributions are thus: (i) the channel-adaptive spectral decoder and its provable optimality and calibration; (ii) the framework—a whole decoder family on one defect-graph interface under structured noise with honest CIs and per-regime significance tests; (iii) selective decoding—abstention cuts committed logical error several-fold for a moderate coverage cost; and (iv) calibration as a first-class metric—ECE spans nearly 6× across the agnostic family and is near-zero for the spectral decoder where its model matches. Limitations are stated plainly in Sec. V A: toy codes only (repetition code; surface-*like* toy, not a stabilizer-exact surface code), simulated structured noise (not hardware syndromes), a required channel calibration sample, a single distance, a first-order model, and a single author and environment. Extending the transfer-operator decoder to a full surface code at scale, to higher-order and online channel models, to hardware syndromes, and to a distance-scaling (sub-threshold) study—while keeping selectivity and calibration first-class—is the natural next step.

Appendix B: Deferred proofs for the transfer-operator theory

This appendix supplies the full proofs deferred from Sec. III: the spectral correlation-decay bound (Lemma 1), the gap-controlled error bound (Theorem 2), and the matrix-concentration argument behind the consistency theorem (Theorem 3). Notation is inherited from Sec. III.

1. Spectral correlation decay (proof of Lemma 1)

Proof of Lemma 1. Let T be the primitive 2×2 stochastic transition matrix with stationary law π (the unique left Perron eigenvector, $\pi T = \pi$, $\pi > 0$ entrywise by Perron–Frobenius) and eigenvalues $1 = \lambda_1 > |\lambda_2|$, $\gamma = 1 - |\lambda_2|$. Work in the weighted space $\ell^2(\pi)$ with inner product $\langle f, g\rangle_\pi = \sum_a \pi(a) f(a) g(a)$. The constant vector $\mathbf{1}$ is the right eigenvector for $\lambda_1 = 1$, and π (as a functional) the left one; the orthogonal complement of $\mathbf{1}$ in $\ell^2(\pi)$, namely $V_0 = \{h : \mathbb{E}_\pi h = 0\}$, is T-invariant, and T restricted to V_0 has spectral radius $|\lambda_2|$. For the 2×2 case V_0 is one-dimensional, spanned by $h_2(a) = a - \mathbb{E}_\pi[a]$, and $Th_2 = \lambda_2 h_2$ exactly.

Let $\bar f = f - \mathbb{E}_\pi f \in V_0$ and $\bar g = g - \mathbb{E}_\pi g \in V_0$. For $i < j$ set $m = j - i$. Using stationarity and the Markov property,

$$\operatorname{Cov}(f(e_i), g(e_j)) = \mathbb{E}_\pi\big[\bar f(e_i)\,(T^m \bar g)(e_i)\big] = \langle \bar f,\ T^m \bar g\rangle_\pi. \tag{B1}$$

Because $\bar g \in V_0$ and V_0 is T-invariant with the restricted operator of norm $|\lambda_2|$ in $\ell^2(\pi)$ (a primitive stochastic matrix is a contraction on V_0; for the 2×2 chain $T^m \bar g = \lambda_2^m \bar g$ outright), Cauchy–Schwarz gives

$$\begin{aligned}\big|\langle \bar f, T^m \bar g\rangle_\pi\big| &\le \big\|\bar f\big\|_\pi\ \|T^m \bar g\|_\pi \le \big\|\bar f\big\|_\pi\ \|\bar g\|_\pi\ |\lambda_2|^m \\ &= \|f\|_\pi\,\|g\|_\pi\,(1-\gamma)^m,\end{aligned} \tag{B2}$$

where $\|f\|_\pi = \big\|\bar f\big\|_\pi$ is the stationary standard deviation. This is Eq. (16). For the total-variation claim, take $g = \mathbf{1}\{\cdot = b\}$ and $f = \mathbf{1}\{\cdot = a\}/\pi(a)$; then $\mathbb{E}[g(e_j) \mid e_i = a] = (T^m)(a,b)$ and the bound reads $|(T^m)(a,b) - \pi(b)| \le C_a(1-\gamma)^m$ with C_a depending on $\pi(a)$; summing $\frac{1}{2}\sum_b$ gives $\|(T^m)(a,\cdot) - \pi\|_{\mathrm{TV}} \le C_0(1-\gamma)^m$ with $C_0 = \max_a \frac{1}{2}\sum_b C_a$. □

2. Gap-controlled decoding error (proof of Theorem 2)

Proof of Theorem 2. By Theorem 1, $P_L^\star = \mathbb{E}_s\big[\min\{\pi(c_0 \mid s), \pi(c_1 \mid s)\}\big]$; since $\min\{u,v\} \le \sqrt{uv}$ for $u, v \ge 0$,

$$\begin{aligned}P_L^\star &\le \mathbb{E}_s\big[\sqrt{\pi(c_0 \mid s)\pi(c_1 \mid s)}\big] \\ &= \sum_s \frac{\sqrt{P(c_0)P(c_1)}}{P(c_0)+P(c_1)}\,\big(P(c_0)+P(c_1)\big) \\ &= \sum_s \sqrt{P(c_0^{(s)})P(c_1^{(s)})}.\end{aligned} \tag{B3}$$

The right side is a Bhattacharyya-type sum over syndromes; it equals $\sum_c \sqrt{P(c)P(c\oplus\mathbf{1})}$ where c ranges over one representative per coset (equivalently $\frac{1}{2}\sum_e \sqrt{P(e)P(e\oplus\mathbf{1})}$ over all e, since each unordered pair is counted twice). Thus

$$P_L^\star \le \tfrac{1}{2}\sum_{e\in\mathbb{F}_2^d} \sqrt{P(e)\,P(e\oplus\mathbf{1})} =: \tfrac{1}{2}\,Z. \tag{B4}$$

We bound Z for the homogeneous transfer operator. Write, for the two chains e and $e' = e \oplus \mathbf{1}$ (which dif-

fer in *every* bit),

$$\sqrt{P(e)P(e')} = \sqrt{\pi_0(e_0)\pi_0(e'_0)} \prod_{i=0}^{d-2} \sqrt{T(e_i, e_{i+1})\, T(e'_i, e'_{i+1})}. \tag{B5}$$

Because $e' = e \oplus \mathbf{1}$, $(e'_i, e'_{i+1}) = (e_i \oplus 1, e_{i+1} \oplus 1)$, so each factor is $\sqrt{T(a,b)\,T(\bar{a},\bar{b})}$ with $\bar{\cdot}$ the bit-flip. Define the 2×2 *Bhattacharyya kernel* $M(a,b) = \sqrt{T(a,b)\,T(\bar{a},\bar{b})}$. Then $Z = \mathbf{u}_0^\top M^{\,d-1}\mathbf{1}$-type contraction; concretely, summing over all e,

$$Z = \sum_e \sqrt{\pi_0(e_0)\pi_0(\bar{e}_0)} \prod_i M(e_i, e_{i+1}) = \boldsymbol{\mu}^\top M^{\,d-1}\mathbf{1}, \tag{B6}$$

with $\mu(a) = \sqrt{\pi_0(a)\pi_0(\bar{a})}$. Now M is a nonnegative matrix dominated entrywise by the geometric-mean structure of T. Two eigen-scales govern $M^{\,d-1}$. First, its Perron root $r(M)$ satisfies $r(M) \le \max_a \sum_b M(a,b)$; by AM–GM, $\sum_b M(a,b) = \sum_b \sqrt{T(a,b)T(\bar{a},\bar{b})} \le \sqrt{(\sum_b T(a,b))(\sum_b T(\bar{a},\bar{b}))} = 1$ with equality iff the two rows $T(a,\cdot)$ and $T(\bar{a},\cdot)$ coincide. For a genuinely correlated $p < \frac{1}{2}$ chain the single-step Bhattacharyya coefficient between the two one-step laws is $\rho_1 = \sum_b \sqrt{T(a,b)T(\bar{a},\bar{b})} = 2\sqrt{p(1-p)} < 1$ (for the symmetric flip parametrisation), so $r(M) \le \rho_1 = \rho$, giving $Z \le B'\rho^{\,d}$ for a constant $B' = \boldsymbol{\mu}^\top\mathbf{1} \cdot \kappa(M)$ absorbing the eigenvector condition number. This is the large-deviation (Chernoff/Bhattacharyya) mechanism and reproduces the i.i.d. leader's exponent $\rho = 2\sqrt{p(1-p)}$.

Second, the *correlation* contribution. Decompose $M = r(M)\,(\mathbf{v}\boldsymbol{\phi}^\top) + M_\perp$ into its Perron projection and remainder, where $\boldsymbol{\phi}, \mathbf{v}$ are the left/right Perron eigenvectors and $M_\perp$ has spectral radius $r_2(M)$. By the same argument that gives the second eigenvalue of T its magnitude $1-\gamma$, the sub-dominant scale of M is bounded by $r(M)\,(1-\gamma)$: the Bhattacharyya kernel inherits the mixing of T because M's off-Perron action is the symmetrised propagation of the same chain, and Lemma 1 bounds the propagated non-stationary component by $(1-\gamma)^{d-1}$. Hence the deviation of $M^{\,d-1}$ from its rank-one Perron part is $\left\|M^{\,d-1} - r(M)^{d-1}\mathbf{v}\boldsymbol{\phi}^\top\right\| \le \kappa(M)\, r(M)^{d-1}(1-\gamma)^{d-1}$, and feeding this into Eq. (B6),

$$Z \;\le\; B'\rho^{\,d} \;+\; A'\,(1-\gamma)^{\,d-1}, \tag{B7}$$

with A' collecting $\boldsymbol{\mu}, \mathbf{1}, \kappa(M)$ and the (bounded) Perron mass. Substituting into Eq. (B4) and setting $A = \frac{1}{2}A'$, $B = \frac{1}{2}B'$ yields Eq. (17). Taking $-\frac{1}{d}\log$ of the dominant term gives the exponential rate $\kappa = -\log\max\{1-\gamma, \rho\} > 0$, and the effective distance $d_{\text{eff}} = \kappa d/\log 2$ (in units where the i.i.d. leader has $\kappa_0 = -\log\rho$) is increasing in γ whenever $1-\gamma > \rho$, i.e. whenever correlation, not weight, is the binding constraint—exactly the regime where a channel-matched decoder helps. □

Remark 1 (Inhomogeneous case). When T_i varies with i, replace $M^{\,d-1}$ by the ordered product $M_0 M_1 \cdots M_{d-2}$; the Perron bound becomes $\prod_i r(M_i)$ and the mixing bound uses the Dobrushin/ergodic coefficient of the product, $\prod_i(1-\gamma_i)$, with γ_i the per-site gap. The conclusion—exponential decay at rate $\min_i$ of the two mechanisms—is unchanged, only the constants track the site-dependent kernels.

3. Consistency and concentration (proof of Theorem 3)

Proof of Theorem 3. Fix a site i and source symbol a. Conditioned on the visits $\{j : e_i^{(j)} = a\}$, the successor symbols $e_{i+1}^{(j)}$ are i.i.d. Bernoulli/categorical draws from $T_i(a,\cdot)$ (the calibration draws are i.i.d. from P_θ, and given $e_i = a$ the next symbol is a fresh draw from the row). Let $N_i(a)$ be the number of such visits; by the η-lower bound and a Chernoff bound, $N_i(a) \ge \eta N/2$ for all i, a simultaneously with probability $\ge 1 - 4d\,e^{-\eta N/8}$, an event we call $\mathcal{G}$ and work on henceforth.

Almost-sure convergence. On the full-measure event that every symbol is visited infinitely often (guaranteed since $\pi_0, T_i \ge \eta > 0$ makes every site-symbol have positive marginal), the strong law of large numbers gives $N_i(a,b)/N_i(a) \to T_i(a,b)$ a.s. The Laplace correction contributes $\hat{T}_i(a,b) - N_i(a,b)/N_i(a) = O(\varepsilon/N_i(a)) \to 0$ since $\varepsilon = o(N)$; hence $\hat{T}_i \to T_i$ and, identically, $\hat{\pi}_0 \to \pi_0$, a.s.

Finite-sample rate. Write the raw frequency $\tilde{T}_i(a,b) = N_i(a,b)/N_i(a) = \frac{1}{N_i(a)}\sum_{j\in V} X_j(b)$ with $X_j(b) = \mathbf{1}\{e_{i+1}^{(j)} = b\}$ i.i.d. Bernoulli of mean $T_i(a,b)$ over the visit set V, $|V| = N_i(a)$. By Hoeffding's inequality, for each entry b and $t > 0$, $\mathbb{P}(\left|\tilde{T}_i(a,b) - T_i(a,b)\right| \ge t \mid \mathcal{G}) \le 2e^{-2N_i(a)t^2}$. Choosing $t = \sqrt{\log(8d/\delta)/(2N_i(a))}$ and union-bounding over the 2 source symbols, 2 target symbols, and $d-1$ sites (at most $4(d-1)$ entries), each entry deviates by at most t with probability $\ge 1-\delta$ on $\mathcal{G}$. Using $N_i(a) \ge \eta N/2$,

$$\max_{i,a,b}\left|\tilde{T}_i(a,b) - T_i(a,b)\right| \le \sqrt{\frac{\log(8d/\delta)}{\eta N}}. \tag{B8}$$

Adding the smoothing bias $\left|\hat{T}_i(a,b) - \tilde{T}_i(a,b)\right| \le 2\varepsilon/N_i(a) \le 4\varepsilon/(\eta N)$ and passing from the entrywise to the max row-sum norm (a row has 2 entries, so $\|\cdot\|_\infty \le 2\,$max-entry) gives Eq. (18) with c a universal constant absorbing the factors of 2. (The same bound follows from matrix Bernstein applied to the sum of the rank-one deviation matrices $\frac{1}{N_i(a)}(X_j - T_i(a,\cdot))\mathbf{e}_a^\top$, whose variance proxy is $\le 1/N_i(a)$ and whose increments are bounded; the scalar Hoeffding route above is tight here because each row is only 2-dimensional.)

Propagation to $\Delta\ell$. For any error e, $\log\hat{P}_\theta(e) - \log P_\theta(e) = \big(\log\hat{\pi}_0(e_0) - \log\pi_0(e_0)\big) + \sum_i\big(\log\hat{T}_i(e_i, e_{i+1}) - \log T_i(e_i, e_{i+1})\big)$. On $[\eta, 1]$ the map

log is $1/\eta$-Lipschitz, so each of the d summands is at most $\frac{1}{\eta}$ times the corresponding entrywise error, whence $\left|\log \hat{P}_\theta(e) - \log P_\theta(e)\right| \;\le\; \frac{d}{\eta} \max_{i,a,b} \left|\hat{T}_i(a,b) - T_i(a,b)\right|$. Since $\Delta\ell$ and $\hat{\Delta}\ell$ are differences of two such log-likelihoods, the same bound (times 2) controls $\left|\hat{\Delta}\ell - \Delta\ell\right|$ uniformly over syndromes, giving the stated $O_{\mathbb{P}}\big(d\sqrt{\log(d/\delta)/(\eta N)}\big)$ rate. $\square$