

Machine-learned syndrome post-selection for reliable quantum error correction

Tobias Haug^{*1}, Askery Canabarro^{1,2}, and Leandro Aolita¹

¹ Quantum Research Center, Technology Innovation Institute, Abu Dhabi, UAE

² Campus Arapiraca, Federal University of Alagoas, Arapiraca-AL, Brazil

Abstract

Quantum error correction can be enhanced by rejecting runs that are likely to fail logically, but the most informative confidence metrics often require costly decoder-level likelihood information. We introduce a decoder-agnostic post-selection method that learns an abort score directly from syndrome data. A supervised classifier is trained only to distinguish syndromes sampled from low- and high-noise regimes; its output is then used to rank new runs before decoding. No logical-error labels, correction operators, or code-specific likelihood calculations are required. We apply the method to experimental logical magic-state-distillation data from a QuEra neutral-atom processor. The classifier is trained entirely on simulated syndrome data generated from a calibrated noise model and is then transferred directly to experimental data. The learned score improves experimental fidelity beyond syndrome-weight filtering and, when combined with logical-gap post-selection, improves the fidelity-throughput trade-off beyond using the logical gap alone. We have also tested the method in circuit-level simulations of the $[[144, 12, 12]]$ Gross bivariate-bicycle code and code-capacity simulations of the surface code, with complete numerical results reported in an accompanying preprint. These results establish syndrome-only learning as a practical machine-learning layer for reliable quantum error correction.

1. Motivation and approach

Stabilizer quantum error correction (QEC) diagnoses physical noise through syndrome measurements and then uses a classical decoder to infer a correction. Some syndrome patterns remain ambiguous, however, and can lead to logical failure even after decoding. Post-selection trades throughput for reliability by aborting runs whose syndrome record indicates a high risk of failure. This trade-off is especially useful for logical state preparation, magic-state distillation, cultivation, and repeat-until-success primitives, where a smaller accepted ensemble can be worthwhile if its fidelity is substantially higher [1–3].

The strongest post-selection scores typically use soft decoder information. A representative example is the logical gap, namely the difference between the best and second-best logical-class scores [4, 5]. Although highly informative, such quantities may require repeated decoding, explicit likelihood calculations, or hard combinatorial optimization. At the opposite extreme, syndrome weight is cheap and decoder-free but compresses the full syndrome record to a single count, discarding spatial, temporal, and correlation structure [6].

We instead learn *when not to decode*. Let $\mathbf{s} \in \{0, 1\}^m$ denote a syndrome record or a vectorized syndrome history. We construct two datasets, $\mathcal{D}_{\text{low}}$ and $\mathcal{D}_{\text{high}}$, sampled from low- and high-noise regimes, and assign labels $y = 0$ and $y = 1$, respectively. A probabilistic binary classifier produces

$$C(\mathbf{s}) = \Pr(y = 1 \mid \mathbf{s}) \in [0, 1], \quad (1)$$

which is interpreted as a high-noise similarity score. For a cutoff C_0 , a run is accepted when

$$C(\mathbf{s}) < C_0 \quad (2)$$

and rejected otherwise. Varying C_0 controls the acceptance rate R and the conditional output fidelity. The score is used primarily for ranking, so perfect probability calibration is unnecessary.

Crucially, the labels describe only the noise ensemble from which each syndrome was sampled. They do not specify whether decoding succeeds, which correction should be applied, or which logical class is most likely. Training data can therefore be generated from stabilizer simulations, deliberately noise-amplified calibration runs, or mixtures of synthetic and experimental data. The resulting model is modular: it can be placed before any decoder and removed without changing the correction pipeline.

For model selection, we use the Tree-based Pipeline Optimization Tool (TPOT) [7, 8]. TPOT evolves and cross-validates preprocessing-classifier pipelines over several model families. This automated search

^{*}tobias.haug@tii.ae

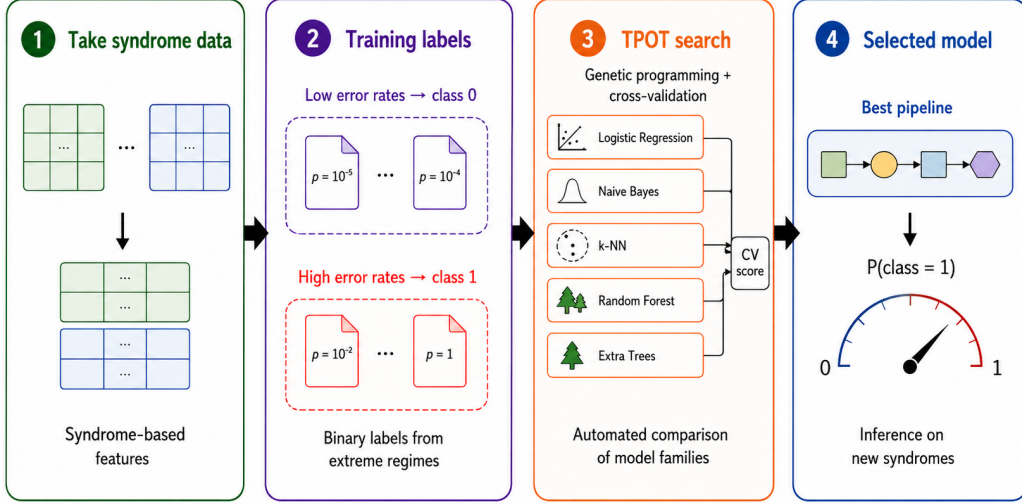


Figure 1: Syndrome-only machine-learning pipeline. Syndrome records from low- and high-noise regimes define a binary classification task. Automated model selection compares candidate pipelines by cross-validation; the selected classifier outputs a continuous high-noise score for new syndromes, which is thresholded before decoding.

avoids committing the method to a specific neural or code-tailored architecture. Relatively simple classifiers were sufficient in our tests, illustrating that useful syndrome structure can be captured without requiring a deep network.

We have additionally benchmarked the method in circuit-level simulations of the $[[144, 12, 12]]$ Gross bivariate-bicycle code and in code-capacity simulations of surface codes with several distances. These studies demonstrate applicability beyond the experimental setting, including qLDPC codes and scalable surface-code post-selection. Full numerical results, training details, and threshold analyses are reported in our accompanying preprint [9].

2. Experimental logical magic-state distillation

We apply the syndrome-only classifier to experimental logical magic-state-distillation data from the QuEra neutral-atom processor [5]. The experiment implements the 5-to-1 distillation protocol using distance-five color codes. Five noisy logical magic states are processed to produce one higher-fidelity output state, conditional on the protocol’s acceptance tests.

The original experiment employed logical-gap post-selection. The logical gap is a highly informative confidence metric because it compares the scores of competing logical explanations of the observed syndrome. However, obtaining it requires solving a costly optimization problem over the decoding graph. Our objective is to construct a lower-cost score directly from syndrome data, without logical labels or decoder likelihoods.

We train on syndrome samples generated using the calibrated Stim model of the experiment. The calibrated gate-error rates are rescaled by factors

$$E \in \{0.01, 0.05, 0.10, 0.15\} \quad (3)$$

for the low-noise class and

$$E \in \{1.75, 2.00, 2.25, 2.50\} \quad (4)$$

for the high-noise class. After class balancing, the simulated dataset contains 809,064 syndrome records, with 404,532 samples in each class. A stratified 10% subset is used for training, and the frozen classifier is then applied directly to the experimental syndromes. This provides a simulation-to-hardware transfer test: no experimental logical outcomes are used to train the model.

We compare four post-selection strategies: machine learning (ML), syndrome weight (SW), logical gap (LG), and a cascade in which ML first filters the data and LG ranks the remaining shots (ML+LG). For ML+LG, ML is first used to post-select 20% of the data, after which the remaining runs are ranked using the logical gap. The output fidelity is evaluated conditionally on the final acceptance rate R using the same Bures-prior treatment as the original experiment.

Figure 2 summarizes the results. ML consistently achieves higher fidelity than SW, showing that the

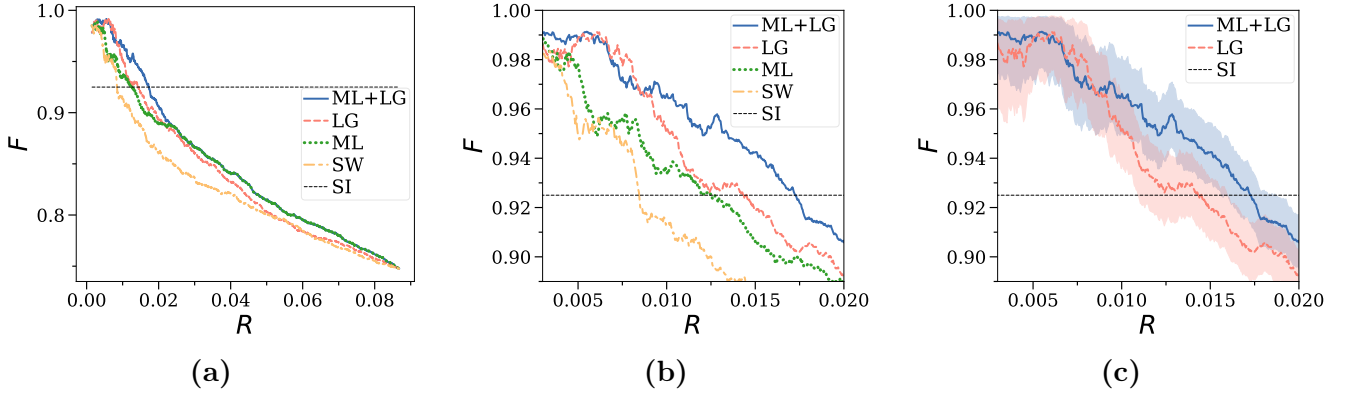


Figure 2: Experimental 5-to-1 logical magic-state distillation using distance-five color codes. **(a)** Output fidelity F versus acceptance rate R for logical gap (LG), machine learning (ML), syndrome weight (SW), and ML followed by logical-gap filtering (ML+LG). ML achieves higher fidelity than SW, while ML+LG provides the strongest fidelity-throughput trade-off. The dashed horizontal line denotes direct state injection without distillation. **(b)** Close-up of the low-acceptance regime. **(c)** ML+LG and LG with 68% credible intervals computed using the Bures prior.

classifier extracts useful structure beyond the total number of detection events. ML alone also exceeds the fidelity of direct logical state injection over part of the operating range, while avoiding the expensive logical-gap computation.

The strongest result is obtained by combining the two scores. ML+LG improves the fidelity-throughput trade-off over LG alone across the relevant acceptance-rate range. Including 68% credible intervals, the combined strategy crosses the direct-injection fidelity at approximately $R \simeq 0.17$, whereas LG alone reaches the same benchmark at approximately $R \simeq 0.14$. Thus, at a fixed target fidelity, the combined method preserves a larger fraction of experimental runs.

3. Discussion

We introduced a decoder-agnostic post-selection method that learns a continuous reliability score directly from syndrome data. Unlike machine-learning decoders, the model is not trained to infer correction operators, logical classes, or logical failures. Instead, it learns to distinguish low- and high-noise syndrome ensembles and uses that distinction to rank new experimental runs.

The experimental magic-state-distillation results demonstrate three practical advantages. First, the classifier transfers from calibrated simulation to hardware data without requiring experimentally labeled logical failures. Second, it outperforms syndrome-weight filtering, showing that relevant spatial and correlation information remains in the full syndrome record. Third, it can be composed with a more expensive decoder-derived metric: using ML as a preliminary filter improves the subsequent logical-gap post-selection.

The method therefore provides a lightweight machine-learning layer that can be evaluated before costly decoding or likelihood estimation. Its requirements are limited to syndrome data collected at distinguishable noise levels, which may be obtained through simulation, calibration, or controlled noise amplification. Additional numerical studies on qLDPC and surface codes, including scalability and threshold behavior, are presented in the accompanying preprint [9].

Future directions include adapting the classifier online as hardware noise drifts, exploiting graph-based models for space-time syndrome records, and jointly optimizing the learned filter with downstream decoders and fault-tolerant protocols.

References

- [1] E. Knill, “Quantum computing with realistically noisy devices,” *Nature* **434**, 39–44 (2005).
- [2] H. Bombín, M. Pant, S. Roberts, and K. I. Seetharam, “Fault-tolerant postselection for low-overhead magic state preparation,” *PRX Quantum* **5**, 010302 (2024).
- [3] C. Gidney, N. Shutty, and C. Jones, “Magic state cultivation: growing T states as cheap as CNOT gates,” arXiv:2409.17595 (2024).
- [4] N. Meister, C. A. Pattison, and J. Preskill, “Efficient soft-output decoders for the surface code,” arXiv:2405.07433 (2024).
- [5] P. Sales Rodriguez *et al.*, “Experimental demonstration of logical magic state distillation,” *Nature* **645**, 620–625 (2025).
- [6] L. H. English, D. J. Williamson, and S. D. Bartlett, “Thresholds for postselected quantum error correction from statistical mechanics,” *Phys. Rev. Lett.* **135**, 120603 (2025).
- [7] R. S. Olson, N. Bartley, J. R. Urbanowicz, and J. H. Moore, “Evaluation of a tree-based pipeline optimization tool for automating data science,” in *Proceedings of GECCO '16*, pp. 485–492 (2016).
- [8] P. Ribeiro *et al.*, “TPOT2: A new graph-based implementation of the tree-based pipeline optimization tool,” in *Genetic Programming Theory and Practice XX* (Springer, 2024).
- [9] T. Haug, A. Canabarro, and L. Aolita, “Machine-learned syndrome post-selection for reliable quantum error correction,” arXiv preprint (2026).