

Feed-Forward-Free Logical Two-Qubit Randomized Benchmarking with Steane QEC and Residual Machine-Learning Decoders

Jan Balewski

NERSC, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Motivation and protocol. We present an end-to-end, feed-forward-free protocol for logical two-qubit randomized benchmarking (RB) with Steane-code QEC, aimed at neutral-atom platforms such as QuEra that offer broad connectivity but no mid-circuit conditional gates. “Feed-forward-free” here means the QPU executes a fixed circuit with no runtime branching on measurement outcomes; all syndrome information is read out at the end and decoded offline by Pauli-frame tracking (PFT). Two logical qubits are encoded in Steane $[[7, 1, 3]]$ blocks [1, 2] initialized by the Goto fault-tolerant method [3]. Standard two-qubit RB is applied transversally at the logical level, with QEC syndrome-extraction blocks inserted between RB segments. Each QEC round uses the same 14 data qubits plus 12 fresh ancillas; fresh ancillas avoid reuse, so syndrome and data readout can be deferred to the end. We implemented the full workflow—RB-sequence generation, circuit construction, simulator execution, shot packing, LUT-PFT decoding, and optional residual ML—in a single reproducible pipeline targeting QuEra hardware submission next year.

Contributions. This submission contributes: (i) a complete logical RB workflow—circuit generation, shot aggregation, LUT decoding, and PFT—ready for near-term execution on QuEra-class hardware, and (ii) example residual ML decoders (BoostTree and a two-stage NN) that improve on a algorithmic weight-one LUT-PFT baseline rather than replacing it. Noisy simulations illustrate where the learned residual layer adds value.

Logical RB with PFT. For standard two-qubit RB [4], a classical preprocessor draws random Clifford blocks $C_0, \dots, C_{m-1} \in \mathcal{C}_2$ and appends $C_m = (C_{m-1} \cdots C_0)^{-1}$, so the noiseless product is identity. Each Clifford is stored as a symplectic tableau [5], enabling the Pauli frame to be propagated exactly through the RB sequence in post-processing without inserting physical recovery gates. For R QEC rounds, the RB sequence including the inverse block is split into $R + 1$ segments as in Fig. 1. Long RB segments can accumulate more faults than a distance-three code can correct in one round, while very short segments result in too many additional noisy syndrome-extraction circuits; under a fixed noise model the logical error rate therefore tends to be minimized at an intermediate R .

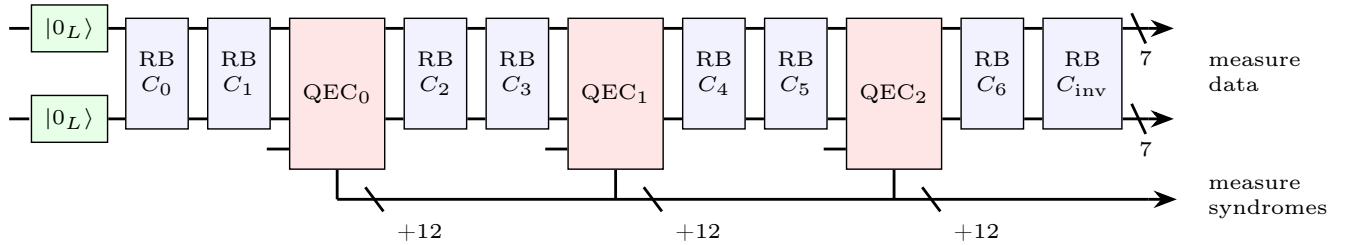


Figure 1: Feed-forward-free logical 2Q RB with two Steane blocks. Logical RB segments $C_0, \dots, C_6, C_{\text{inv}}$ are interleaved with three QEC rounds. Each RB segment acts on two 7-qubit code blocks, while each QEC round uses 12 fresh ancillas whose syndromes are measured at the end. Recovery is deferred to Pauli-frame tracking.

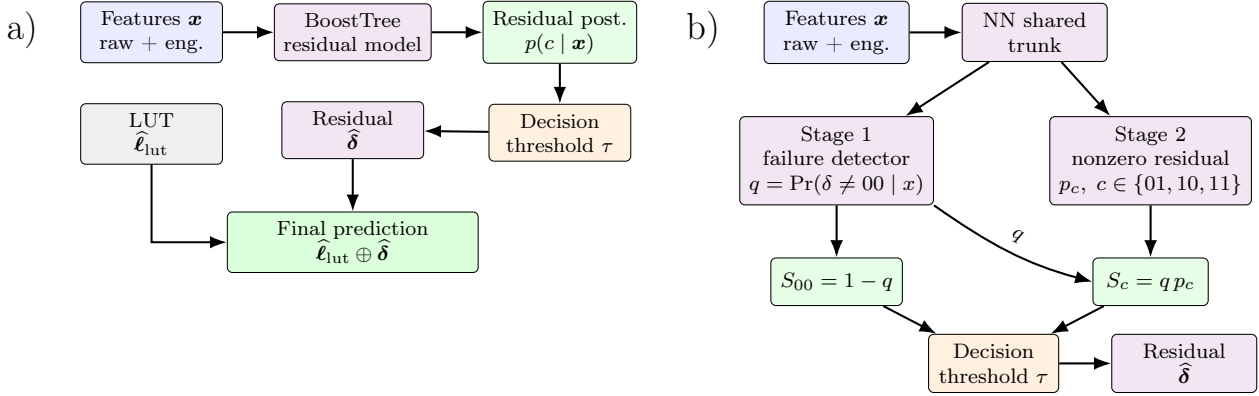


Figure 2: Residual decoder workflows. a) BoostTree residual correction workflow. b) Two-stage NN residual correction workflow.

LUT-PFT baseline. A distance-three Steane block identifies a single-qubit Pauli error from its X - and Z -type syndromes by a weight-one lookup table (LUT). LUT-PFT denotes this analytic decoder coupled to tableau PFT: each syndrome round proposes an incremental Pauli correction that is XORed into the classical frame rather than applied on the QPU. The terminal data bits are interpreted after the accumulated frame, so the same hardware circuit is used for all syndrome outcomes. Shots with nonzero encoding-flag outcomes are discarded before RB analysis, matching standard Goto initializer post-selection.

Residual learning. LUT-PFT is analytic and requires no training data, but it assumes an effectively uncorrelated, weight-one correction picture. Correlated faults from transversal two-qubit gates or device-specific biases can leave structured residual errors; recent learned decoders show that richer models can outperform standard matching baselines [7,8]. We compare two residual classifiers—XGBoost (BoostTree) [6] and a two-stage NN—trained on $\delta_{\text{true}} = \ell_{\text{lut}} \oplus \ell_{\text{in}}$, not on the logical class directly. Features include the 14 final data bits, $12R$ syndrome-history bits, the two logical input bits, and engineered summaries. The dataset includes all four initial logical states $00_L, 01_L, 10_L, 11_L$; train/test splits are at the RB-sequence level, not the shot level, preventing leakage from repeated shots of the same circuit. BoostTree predicts four joint residual classes with a validation-selected intervention threshold. The NN uses a two-stage model: one head detects whether LUT-PFT failed, and a conditional head identifies the nonzero residual among 01, 10, 11. In both cases, $\hat{\ell}_{\text{ML}} = \hat{\ell}_{\text{lut}} \oplus \hat{\delta}_{\text{ML}}$ (Fig. 2). The models are not asked to rediscover Steane decoding from scratch; they learn only when the analytic decoder is predictably biased by the observed record. This residual target is conservative and auditable: if the learned confidence is low, the decoder falls back to LUT-PFT.

Noisy-simulation results. We simulated standard two-qubit logical RB for sequence lengths 60–360 Clifford gates and QEC rounds $R = 0, \dots, 40$ (1000 RB sequences, 10^6 shots per setting). The noise model uses all-to-all connectivity, native Clifford gates (H, S, X, and two-qubit entanglers such as CX or CZ, equivalent under the depolarizing channels used here), one-qubit error 10^{-4} , two-qubit error 4×10^{-4} , and no SPAM noise—the simplest parametrization for a first study; the same pipeline can accept richer device models, including the full Pauli noise specifications published by QuEra. Fig. 3 compares results for two representative RB sequence lengths (120 and 240 Clifford gates). In both panels the logical error rate (LER) versus R has a minimum: too few QEC rounds leave long RB subsequences under-protected, while too many rounds make syndrome noise dominant. The yellow line shows physical two-qubit RB under the same noise model and is included only as an unencoded reference. In the low- R regime, where the hardware outlook below is most relevant, the learned decoders rank $\text{NN} > \text{BoostTree} > \text{LUT-PFT}$; at larger R all three curves converge because the residual models defer to the analytic baseline. For intermediate R the logical curves can fall below the physical reference, indicating that encoded RB has smaller noise than physical two-qubit RB despite using many more entangling gates.

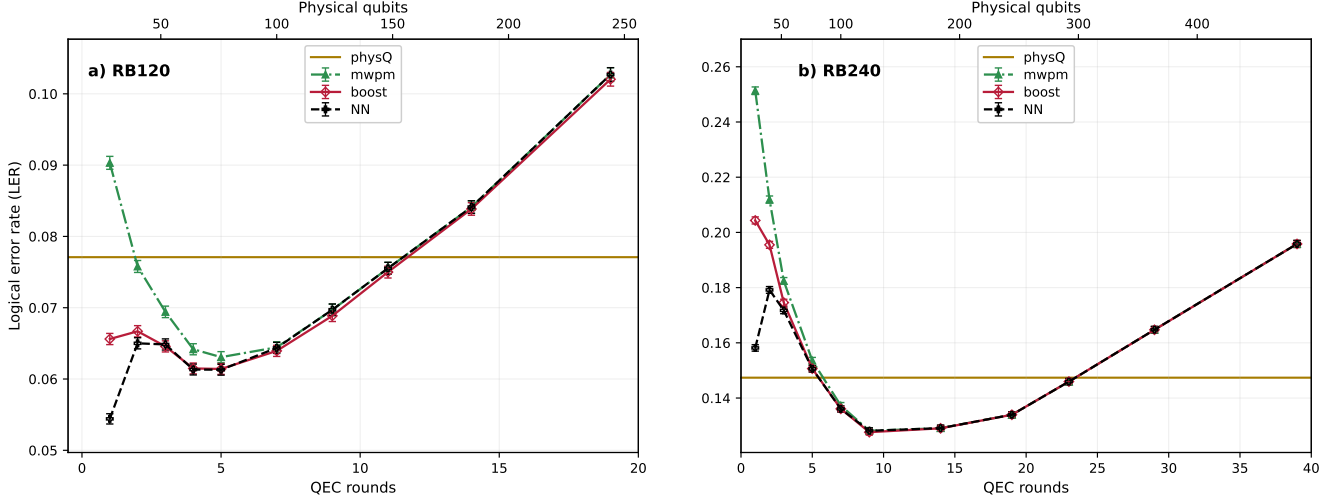


Figure 3: Logical error rate versus QEC rounds for standard 2Q RB simulations with sequence lengths a) RB120 and b) RB240. Top axis: physical qubit count $2(8 + 6R)$. The yellow line marks physical two-qubit RB under the same noise model (reference only).

Discussion and hardware outlook. The primary near-term use case is low- R logical RB on QuEra-class machines: each QEC round adds only 12 ancilla atoms for two logical qubits, so $R \lesssim 10\text{--}15$ requires $\sim 2(8 + 6R)$ atoms (~ 200 at $R = 15$), within the capacity announced for the Gemini generation. That is precisely the regime where residual decoders add the most over LUT-PFT, while keeping the executed circuit feed-forward-free. A practical first experiment would therefore scan modest R at fixed logical RB length, validate LUT-PFT against known logical inputs, and then enable the residual layer only where held-out RB data show structured improvement.

The ML layer does not change code distance; it uses the full syndrome and data record statistically over the device’s noise distribution. LUT-PFT remains a portable physics fallback, while BoostTree or NN add a calibration-aware residual policy when training data support it. Fig. 3 also shows that the LER minimum shifts to larger R for longer RB sequences (RB240 versus RB120), consistent with protecting a similar number of Clifford gates per QEC cycle. Depolarizing noise was chosen as the minimal parametrization compatible with Clifford RB; replacing it by QuEra’s published Pauli channels is straightforward and is the natural next simulation step before hardware execution. Training can be refreshed after calibration updates using the same RB workflow, with LUT-PFT always available for validation. The software stack is backend-agnostic: the same post-processing consumes counts from Stim, Qiskit Aer, or future QuEra job records once shot strings are cast to the common register layout.

References

- [1] A. M. Steane, *Phys. Rev. Lett.* **77**, 793 (1996).
- [2] A. R. Calderbank and P. W. Shor, *Phys. Rev. A* **54**, 1098 (1996).
- [3] H. Goto, *Sci. Rep.* **6**, 19578 (2016).
- [4] E. Knill et al., *Phys. Rev. A* **77**, 012307 (2008); E. Magesan et al., *Phys. Rev. Lett.* **106**, 180504 (2011).
- [5] S. Aaronson and D. Gottesman, *Phys. Rev. A* **70**, 052328 (2004).
- [6] T. Chen and C. Guestrin, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2016), pp. 785–794.
- [7] J. Bausch et al., *Nature* **635**, 834 (2024).
- [8] V. Sivak, M. Newman, and P. Klimov, *Phys. Rev. Lett.* **133**, 150603 (2024).