TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY

Citation (APA)
Xin, Y., van der Meer, S. L. M., Serra-Peralta, M., Vroomans, T. H. F., Finkel, M., Veen, H. M., Beekman, M. W., & DiCarlo, L. (2026). Improved Error Correction with Leakage Reduction Units Built into Qubit Measurement in a Superconducting Quantum Processor. *PRX Quantum*, *7*(3), Article 033062. https://doi.org/10.1103/y66v-988s

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Improved Error Correction with Leakage Reduction Units Built into Qubit Measurement in a Superconducting Quantum Processor

Yuejie Xin,[1,2] Sean L. M. van der Meer,[1,2] Marc Serra-Peralta,[1,3] Tim H. F. Vroomans,[1,2] Matvey Finkel,[1,2] Hendrik M. Veen,[1,2,†] Marc W. Beekman,[4] and Leonardo DiCarlo[1,2,*]

[1]*QuTech, Delft University of Technology, P.O. Box 5046, 2600 GA Delft, The Netherlands*
[2]*Kavli Institute of Nanoscience, Delft University of Technology, P.O. Box 5046, 2600 GA Delft, The Netherlands*
[3]*Delft Institute of Applied Mathematics, Delft University of Technology, P.O. Box 5046, 2600 GA Delft, The Netherlands*
[4]*Netherlands Organisation for Applied Scientific Research (TNO), P.O. Box 96864, 2509 JG The Hague, The Netherlands*

(Received 2 December 2025; revised 15 May 2026; accepted 7 July 2026; published 21 September 2026)

Leakage to non-computational states is a source of correlated errors in both time and space that limits the effectiveness of quantum error correction (QEC) with superconducting circuits. We present and experimentally demonstrate a high-fidelity, hardware-efficient leakage reduction unit (LRU) operating concurrently with transmon measurement without incurring time overhead for 2-level readout (2RO). Adapted from double-drive reset of population (DDROP), the protocol utilizes simultaneous drives on the transmon and its readout resonator, leveraging the dispersive shift to create a directional process that returns the transmon to the computational subspace. The LRU achieves a 98.4% leakage removal fraction without compromising the 2RO assignment fidelity (99.2%) and can also be configured for 3-level readout. We combine LRU-enhanced measurement and leakage-informed neural-network decoding to successfully suppress logical error rates in both memory and stability QEC experiments without any post-selection. Remarkably, the LRU preserves the logical error-suppression rate in stability experiments as injected leakage increases.

DOI: 10.1103/y66v-988s

I. INTRODUCTION

Quantum error correction (QEC) provides a path toward fault-tolerant quantum computation by encoding information redundantly. Despite remarkable progress [1–6], the performance of QEC is fundamentally limited by physical errors that corrupt quantum information. To achieve fault tolerance, it is crucial to understand and mitigate all significant error channels, especially those that are poorly handled by standard QEC codes. Among the various error channels, leakage of population out of the computational subspace is particularly damaging [7–9]. For instance, in transmons, the most common superconducting qubit, the computational subspace is typically spanned by the ground ($|g\rangle$) and first excited ($|e\rangle$) states, with excitation to the second excited state ($|f\rangle$) being the dominant type of leakage. This type of error poses a fundamental challenge to standard stabilizer codes like the surface code [10]. The power of these codes lies in their ability to discretize arbitrary physical errors that occur within the computational subspace into a set of correctable Pauli errors. However, leakage violates this core assumption because a leakage event cannot be mapped to the Pauli framework. Additionally, leakage can last for several QEC cycles, leading to correlated errors in time. Even worse, leakage can propagate through two-qubit gates, creating not only temporally but also spatially correlated errors that are highly detrimental to the logical performance [7,9,11–13].

Various strategies have been developed to combat leakage. Hardware-level chip design can reduce it [14,15], while leakage-tolerant QEC codes have been proposed at the cost of an increased number of qubits [16,17]. Active methods like leakage reduction units (LRUs), designed to seep leakage to the computational subspace, offer a more direct solution [12,18–24]. However, many LRU protocols require a dedicated time step within the QEC cycle. This can increase the overall physical error rate, potentially negating the benefits of leakage removal. Moreover, some LRUs

*Contact author: l.dicarlo@tudelft.nl
†Present address: SRON Netherlands Institute for Space Research, Niels Bohrweg 4, 2333 CA Leiden, The Netherlands.

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

require additional control electronics or quantum elements on-chip, making their implementation difficult.

In this work, we address these challenges by experimentally demonstrating a high-fidelity, time-efficient, and hardware-efficient LRU for superconducting qubits. Our protocol is inspired by DDROP [25], which we adapt here for leakage removal from $|f\rangle$ to $|e\rangle$. The key innovation is the concurrent operation of the LRU with the measurement process, creating an LRU-enhanced measurement that has no time overhead compared to standard 2-level readout (2RO). The protocol is broadly applicable, as its implementation only requires a dispersively coupled readout resonator, a nearly universal component in current superconducting quantum processors. It also uses the same electronics as for standard readout and single-qubit gates, without the need for extra local oscillators or mixers to reach distant frequency bands [19].

A high-fidelity three-level readout (3RO) can be achieved by the LRU-enhanced measurement, which provides the QEC decoder with significantly richer information than a conventional binary outcome. This has been shown to improve logical performance [26]. Specifically, a readout result corresponding to $|f\rangle$ heralds a leakage event at a known location (measure qubit) and time (QEC round). This transforms a non-Pauli error into a tractable, heralded error for the decoder. A key result of this work is a demonstration that QEC without any post-selection performs best when combining the LRU with 3RO. Notably, the LRU can even stabilize the error-suppression rate in the stability experiment, making it independent of leakage injection rate.

The manuscript is organized as follows. We first detail the calibration and characterization of the LRU, quantifying its leakage removal efficiency and its minimal impact on readout fidelity. We then demonstrate its practical value by integrating it into two QEC benchmarks: a memory experiment of a distance-3 bit-flip repetition code and a 7-qubit stability experiment. In both cases, we show that combining the LRU with 3RO yields the highest logical performance among the four variants (2RO or 3RO with or without LRU) at all levels of controlled leakage injection on controlled-Z (CZ) gates.

II. IMPLEMENTATION OF THE LEAKAGE REDUCTION UNIT

The processor used in our experiments [19] has a standard circuit quantum electrodynamics architecture, with each transmon dispersively coupled to a dedicated readout resonator. The LRU protocol relies on a directional process for unconditional leakage reset. The implementation, shown schematically in Fig. 1, is specifically tailored for leakage removal from $|f\rangle$ to $|e\rangle$. For clarity and simplicity, this schematic omits the dedicated Purcell filter, which is present in our processor to protect each transmon from Purcell decay through the resonator. As we show below (Figs. 2 and 3), the LRU protocol is fully compatible with and effective in the presence of the filter.

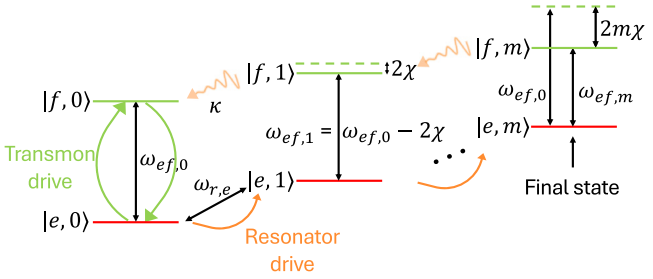

FIG. 1. Simplified schematic of the leakage reduction unit (LRU) protocol. The energy levels of the coupled transmon-resonator system are shown, with states labeled as $|\text{Transmon, Resonator}\rangle$. The protocol utilizes two simultaneous drives: a transmon drive at the frequency of the $|e\rangle \leftrightarrow |f\rangle$ transition when there is no photon in the resonator, $\omega_{ef,0}$, and a resonator drive at the resonator frequency when the transmon is in $|e\rangle$, $\omega_{r,e}$. The process begins with the system in the leakage state $|f,0\rangle$. The transmon drive, which is resonant with the $|f,0\rangle \leftrightarrow |e,0\rangle$ transition, then transfers the population to $|e,0\rangle$. Subsequently, the resonator drive populates the resonator, moving the system to a final steady state, $|e,m\rangle$. The directionality of this process is ensured by the dispersive shift, 2χ. As photons accumulate in the resonator, the transmon transition frequency shifts. The original transmon drive at $\omega_{ef,0}$ becomes off-resonant with the $|e,m\rangle \leftrightarrow |f,m\rangle$ transition by an amount of $2m\chi$, preventing the system from being reexcited to the $|f\rangle$ state manifold. Likewise, the resonator drive is off-resonant when the transmon is in $|f\rangle$, preventing photon accumulation. Once the drives are removed, the system quickly decays to $|e,0\rangle$, effectively removing leakage from $|f\rangle$.

The protocol creates a one-way path from the leakage state $|f\rangle$ to the computational state $|e\rangle$ by applying simultaneous microwave drives to the transmon and the resonator. The core principle is to leverage the state-dependent frequencies of the transmon and resonator. A resonant transmon drive transfers population from $|f,0\rangle$ to $|e,0\rangle$, while a second drive begins to populate the resonator, a process that is efficient only when the transmon is in $|e\rangle$. This induced photon population is the key: it activates the dispersive shift, which detunes the transmon $|e\rangle - |f\rangle$ transition frequency. This detuning effectively locks the transmon in the computational subspace by making the reverse transition to $|f\rangle$ highly off-resonant.

Transitioning from the idealized model above to a realistic system introduces practical limitations that can affect performance due to finite linewidths. In a real system, an off-resonant drive can still induce unwanted transitions, which could reduce the leakage removal efficiency or even re-introduce leakage. This non-ideal effect can be mitigated by designing a system with a larger dispersive shift or decoherence time T_2 (detailed system parameters of our device can be found in Sec. I of Supplemental Material [27]). As will be shown later, tuning the drive parameters to find an optimal operating point also helps to reduce this problem.

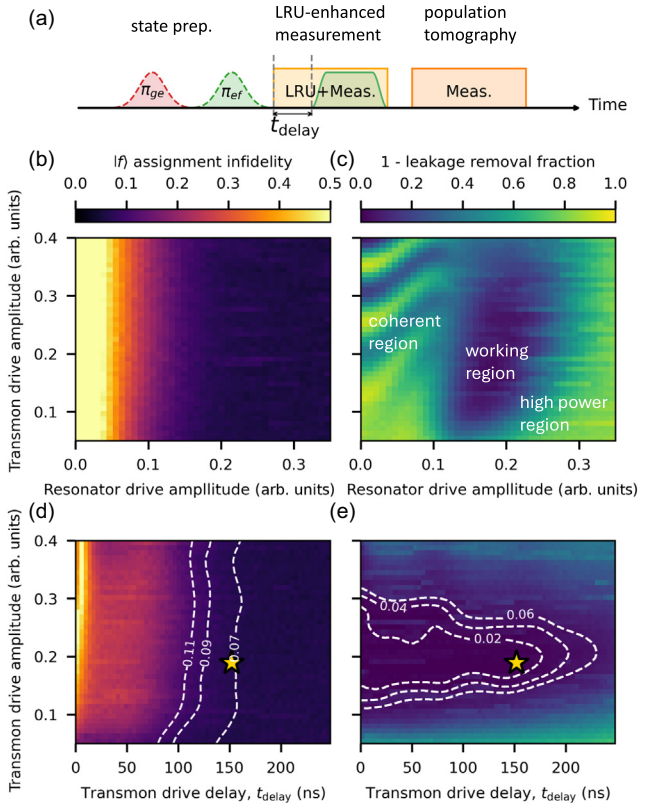

FIG. 2. Calibration of the LRU pulse. (a) The experimental pulse sequence used for calibration. The transmon is first prepared in the $|g\rangle$, $|e\rangle$, or $|f\rangle$, followed by the LRU-enhanced measurement. A delay, t_{delay}, is introduced between the resonator and transmon drives to enable three-level readout. Finally, population tomography is performed using a standard measurement with readout errors corrected. (b)–(e) Performance versus drive parameters. Each point averages over 4000 shots. In (c) three distinct behaviors are observed: a coherent region at low resonator power where Rabi oscillations dominate; a central working region where balanced drives enable the intended LRU process; and a high-power region where the strong resonator drive populates the resonator with photons exceeding the critical photon number of the dispersive regime. In (d),(e) the trade-off between assignment fidelity and leakage removal is explored. In (d) the assignment fidelity increases with t_{delay} before being limited due to transmon relaxation. In (e) the residual $|f\rangle$ population increases as t_{delay} gets longer; it rises slowly at first while photons build up in the resonator, then rises more quickly as the reduced time for the removal process becomes the dominant factor. The white dashed lines indicate contours for infidelity (0.07) and residual population (0.02). The yellow star marks the final choice of parameters.

A core advantage of driving the readout resonator is that measurement can be performed simultaneously with the leakage removal. This effectively creates an LRU-enhanced measurement with no time overhead for 2RO. Furthermore, the transmon drive is designed to only induce the transition between $|e\rangle$ and $|f\rangle$, ideally leaving the ground state $|g\rangle$ unaffected. We have experimentally

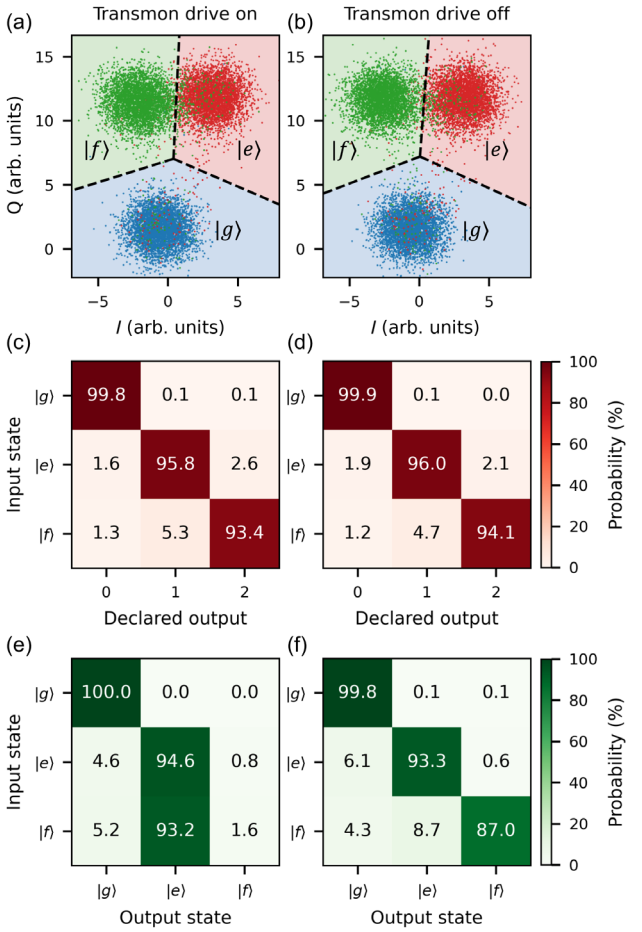

FIG. 3. Performance comparison of LRU-enhanced and standard measurements. The LRU-enhanced measurement (left column) is compared to a standard measurement (right column), which is implemented by turning off the transmon drive while keeping all other parameters identical. (a),(b) Single-shot I-Q data with 4000 shots for each state. (c),(d) Readout assignment matrices, showing the probability of assigning a measurement outcome given a prepared state. The average assignment fidelities are 96.3% (LRU) and 96.7% (standard). (e),(f) Population transfer matrices showing the probability of the transmon being in a final state given its initial state. The matrix in (e) demonstrates a leakage removal fraction of 98.4%.

confirmed that the assignment fidelity between $|g\rangle$ and $|e\rangle$ remains uncompromised during this process (see Fig. 3). By introducing a delay, t_{delay}, between the resonator and transmon drives, a distinct readout signal for $|f\rangle$ can also be obtained (see Fig. 2 for details). This delay allows switching 3RO on and off conveniently without sequencing a different readout.

For calibration, both assignment fidelity and leakage removal efficiency are optimized using the circuit shown in Fig. 2(a). The transmon drive pulse has a sine-squared envelope as in Ref. [42], with 10 ns rise and fall times, while the resonator drive is a square pulse. The total duration for the LRU-enhanced measurement is 500 ns

(380 ns pulses followed by 120 ns for photon depletion), which is typical for a standard measurement in our system. The final population tomography [quantified by a 3×3 matrix as shown in Figs. 3(e) and 3(f)] is corrected for readout errors using an iterative Bayesian unfolding method [43], with a reference matrix for unfolding averaged over 10 000 shots.

Due to the coupled nature of the drive parameters, the covariance matrix adaptation evolution strategy algorithm [44] is used to find a global optimum. We use the geometric mean of assignment fidelity and leakage removal as the cost function, while the parameters to optimize include the frequencies and amplitudes of the two drives, as well as t_{delay}. To illustrate the underlying physics, Figs. 2(b)–2(e) shows 2D sweeps of the key parameters, with all other parameters held at their optimized values.

There are three distinct physical regimes that depend on both drive amplitudes, as highlighted in the landscape of Fig. 2(c). The performance is quantified by the leakage removal fraction, defined as $1 - P(f|f)$ and obtained from the population tomography, where $P(f|f)$ is the probability of both start and end states being $|f\rangle$. The first regime, the coherent region, occurs when the resonator drive is too weak, and an insufficient number of photons populates the resonator. The resulting dispersive shift is not large enough to prevent the transmon drive from causing off-resonant transitions, leading to coherent Rabi oscillations between $|e\rangle$ and $|f\rangle$ rather than a directional reset. This effect can also be quantitatively described by the state jump rate $\gamma_{\mathrm{jump}} = \Omega_R^2/2(\gamma_2 + \Gamma_d)$, where Ω_R is the Rabi drive rate, γ_2 is the qubit dephasing rate and Γ_d is the measurement-induced dephasing rate [45]. Conversely, if the resonator drive is too strong (the high-power region), the reset rate is suppressed, resulting in a lower leakage removal fraction for a fixed duration. The number of photons may also exceed the critical photon number, and the system no longer operates in the dispersive regime, breaking a fundamental assumption of the protocol. Between these two extremes lies a broad working region where the drives are balanced, and the LRU process proceeds with high efficiency. The large area of this working region demonstrates the robustness of the protocol against variations in drive amplitudes.

One of the key features of the protocol is the ability to enable a distinct readout of $|f\rangle$. As shown in the circuit in Fig. 2(a), this is achieved with finite t_{delay}. This delay creates a pure readout window where the resonator response is sensitive to the transmon state before the leakage removal begins. Note that while the delay necessarily reduces the available time for the subsequent leakage removal, it is not simply a penalty. Building up photons in the resonator before applying the transmon drive helps to reduce leakage induced by the transmon drive itself and can speed up the subsequent reset process. Varying t_{delay} reveals a performance trade-off, as illustrated in Figs. 2(d) and 2(e). We define the $|f\rangle$ assignment infidelity [Figs. 2(b) and 2(d)] as $1 - P(2|f)$, where $P(2|f)$ is the probability of the LRU-enhanced measurement outputting 2 when the transmon is prepared in $|f\rangle$. In Fig. 2(d), increasing t_{delay} initially improves the $|f\rangle$ assignment fidelity by allowing more time for pure readout. This improvement continues until the fidelity becomes limited by the transmon natural relaxation time. At short t_{delay} [Fig. 2(e)], the residual $|f\rangle$ population depends only weakly on t_{delay}. We think that leakage removal in this regime is limited primarily by experimental imperfections. At longer t_{delay}, however, the residual population increases more rapidly with t_{delay} as the reduced pulse duration becomes dominant. This behavior allows us to obtain the benefit of high-fidelity $|f\rangle$-state readout without a significant compromise in leakage-removal performance. The chosen operating point, marked by the yellow star, is where these factors are balanced, achieving a near-saturated assignment fidelity while maintaining a high leakage removal fraction.

With the LRU protocol calibrated, its performance is now directly compared against a standard measurement in Fig. 3. To ensure a fair comparison, the standard measurement is implemented by turning off the transmon drive of the LRU-enhanced measurement while keeping all other parameters identical. The single-shot readout data, plotted on the I-Q plane in Figs. 3(a) and 3(b), shows well-separated state clusters for both measurement types. A more quantitative analysis is provided by the readout assignment matrices in Figs. 3(c) and 3(d). The average 3RO assignment fidelities (calculated as the average of diagonal terms) are nearly identical at 96.3% for the LRU-enhanced measurement and 96.7% for the standard one. To be specific, the assignment fidelities between $|g\rangle - |e\rangle$ and $|g\rangle - |f\rangle$ are unchanged. This confirms that the LRU does not compromise the overall ability to distinguish the computational states. The fidelity between $|e\rangle$ and $|f\rangle$ is slightly lower with the LRU, consistent with the trade-off discussed in Fig. 2. In addition, readout signals are integrated using uniform weights. We expect the assignment fidelity to be further improved with optimal weights [46–48], especially when the transmon drive is on. The LRU performance is revealed in the population transfer matrices [Figs. 3(e) and 3(f)], which are extracted from the final readout-error-corrected population tomography [Fig. 2(a)]. These matrices show the statistics of the final state of the transmon after the measurement. The off-diagonal terms in the lower-left of both matrices, such as the probability $P(g|e)$, are primarily due to relaxation during the measurement window. The crucial result is in the bottom row of Fig. 3(e), which shows that if the transmon starts in $|f\rangle$, the LRU process actively returns it to $|e\rangle$. This corresponds to a leakage removal fraction of 98.4%. In contrast, the matrix for the standard

measurement in Fig. 3(f) shows a high $P(f|f)$, indicating that the leakage state persists as expected. The comparison also indicates that one can easily turn the LRU on or off by simply turning on or off the transmon drive, which highlights the ease of use as a main advantage. In Secs. II and III of Supplemental Material [27], we provide examples of performance under different calibration targets and present a detailed analysis suggesting routes for future improvement.

III. LRU IN QUANTUM ERROR CORRECTION

Having established and calibrated the LRU-enhanced measurement, we explore its benefit in the context of QEC. We perform two experiments that benchmark the basic tasks in fault-tolerant quantum computation: a memory experiment [4–6,49–53], which tests the preservation of a logical observable through time, and a stability experiment [54–56], which tests the movement of a logical observable through space, as in lattice surgery [57]. Leakage is a particularly damaging error source in these two contexts, and especially in stability experiments, as a leaked transmon can induce correlated errors in subsequent rounds. By actively returning leaked transmons to the computational subspace, our LRU protocol should suppress the impact from these error channels and thereby improve the logical performance.

A. Memory experiment

The memory experiment is performed using a distance-3 bit-flip repetition code, which we call Repetition-5. As shown in the device schematic in Fig. 4(a), the logical qubit is encoded in three data qubits (D_7, D_8, D_9), while two measure qubits (Z_3, Z_4) are used to perform parity checks (same device as in Ref. [19], but with X and Z labels swapped). As the experimental circuit in Fig. 4(a) shows, after initializing the data qubits into one of the basis states $|i,j,k\rangle (i,j,k \in \{0,1\})$, the system undergoes R rounds of stabilizer measurements. At the end of each round, X gates are applied to data qubits during measure-qubit readout to make state-dependent errors (mainly relaxation here) more state-independent. These X gates are replaced by two-level readout (2RO) on data qubits in the last round [not shown in Fig. 4(a) for simplicity]. The error information can be extracted from readout outcomes of measure qubits. A parity is defined by the difference between two consecutive measure-qubit outcomes in our case, as we do not reset the measure qubits each round. The parity of the first (last) round is directly computed from data-qubit initialization (measurements). A change in the parity relative to the previous round signals a detection event, indicating that an error may have occurred in the interim. The space-time history of these detection events, together with extra information (leakage flags in our case), is fed to a decoder. We use a neural network decoder as described in Secs. V and VI of Supplemental Material [27].

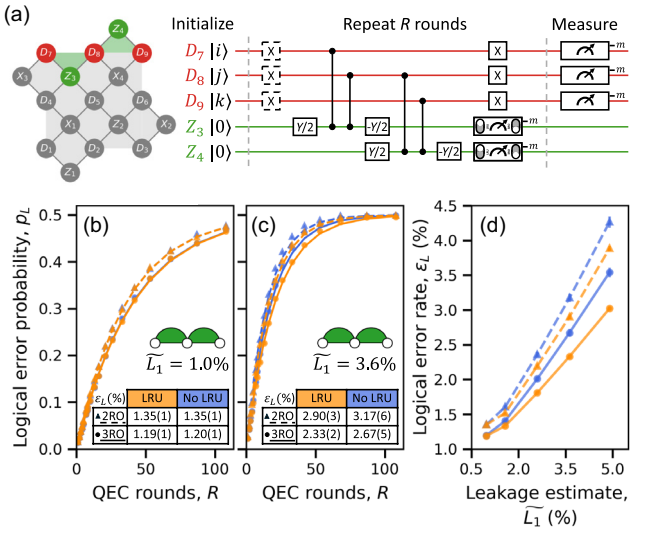

FIG. 4. Memory experiment of a distance-3 bit-flip repetition code. (a) Schematic showing the five qubits used, with data qubits in red and measure qubits in green. Quantum circuit consisting of qubit initialization, R rounds of stabilizer measurements, and final measurement. (b),(c) Logical error probability (p_L) as a function of the number of QEC rounds (R) at baseline leakage ($\widetilde{L_1} = 1.0\%$) and intermediate leakage ($\widetilde{L_1} = 3.6\%$). Performance is compared for protocols with LRU (orange) and without (No LRU, blue), using either a three-level (3RO, solid lines, circles) or two-level (2RO, dashed lines, triangles) readout on the measured qubits. Results are averaged over all basis states $|i,j,k\rangle$. (d) Extracted logical error rate per round (ε_L) as a function of the estimated leakage rate ($\widetilde{L_1}$). The combination of LRU with 3RO consistently provides the lowest error rate, with the advantage growing as leakage increases.

To avoid the neural network decoder learning unwanted information, random logical bit flips (dashed X gates) are applied at the start of each QEC round, as shown in Fig. 4(a).

The goal of the decoder is to use the measured information to determine a logical correction. By repeating the experiment for a fixed number of rounds R, the logical error probability $p_L(R)$ is measured. The key figure of merit is the logical error rate per round, ε_L, which is extracted by fitting $p_L(R) = \frac{1}{2} - \frac{1}{2}(1-2\varepsilon_L)^{R+R_0}$, where R_0 is a constant [58]. To improve the extraction of ε_L, we fix $R_0 = 0$ and fit starting from round 5.

To study the effects of leakage, we controllably sweep the leakage rate of the CZ gates [Figs. 4(b)–4(d)]. The leakage rate is swept by adjusting control parameters along the contour of 180° conditional phase. As a proxy for leakage L_1, we use the upper bound $\widetilde{L_1}$ obtained from conditional oscillation experiments [59] (see Sec. IV of Supplemental Material [27] for details). In all CZ gates, measure qubits are the ones prone to leak as they are the higher-frequency transmons of every pair.

Without intentional leakage injection [Fig. 4(b), at baseline $\widetilde{L_1} = 1.0\%$] using 3RO over 2RO gives the most

important performance improvement, reducing the logical error rate by 11.1%. (Note that 2RO is achieved by actually doing a 3RO and mapping the outcome 2 to 1, see Sec. V of Supplemental Material [27].) Incorporating the LRU shows benefit at all other levels of injected leakage, for both 2RO and 3RO, as shown specifically for $\widetilde{L_1} = 3.6\%$ in Fig. 4(c). The full leakage sweep in Fig. 4(d) shows that combining LRU and 3RO universally gives the best results. Further research is required to see if this holds at larger code distances and for other codes. In addition to investigating leakage during CZ gates in the experiment, we have used simulation to explore the impact of other leakage sources (Sec. VIII of Supplemental Material [27]).

B. Stability experiment

The stability experiment, which we call Stability-7, is performed with a 1D chain of four Z-type stabilizers involving seven qubits [Fig. 5(a)]. The experiment begins by initializing the four measure qubits in one of classical bitstrings $|i, j, k, l\rangle (i, j, k, l \in \{0, 1\})$, and the three data qubits in $|+\rangle$. The circuit is similar to that of Repetition-5; however, there is no need for random flips on data qubits, as the logical observable is now defined by measure qubits instead of data qubits, and data-qubit measurements are not used in the decoding. The logical observable of a stability experiment is defined as the product of the four measure-qubit readout outcomes in a given round; here, we choose the last odd round. A logical error occurs if this measured product, after correction by the decoder, is different from the known fixed product (0 or 1 depending on the initial measure-qubit states). Stability experiments can be viewed as space-time rotated memory experiments [54], where increasing R exponentially reduces p_L, analogous to increasing the code distance in a memory experiment. We fit the data to the model $p_L = Ae^{-\gamma R}$, with error-suppression factor γ and scaling factor A. As the decay of p_L is not purely exponential in the first few rounds (a point we address later in this section), we perform fits starting from round 10. In this experiment, we inject leakage only on the CZ gates between measure and data qubits in which the measure qubit is higher frequency and thus the one susceptible to leakage (all CZ gates except $D_5 - Z_2$ and $D_5 - Z_3$).

In Stability-7, using LRU and 3RO both in isolation and especially in combination gives substantial performance improvements even at the baseline leakage rate [Fig. 5(b)]. Compared to 2RO without LRU, γ increases by 29%, 23%, and 54% for 2RO with LRU, 3RO without LRU, and 3RO with LRU, respectively. We observe significant improvements even at baseline because stability experiments are more susceptible to time-like errors than memory experiments. Leakage is a time-correlated error and can be long lived, making it especially damaging if not dealt with. The improvement is even more significant at increasing levels

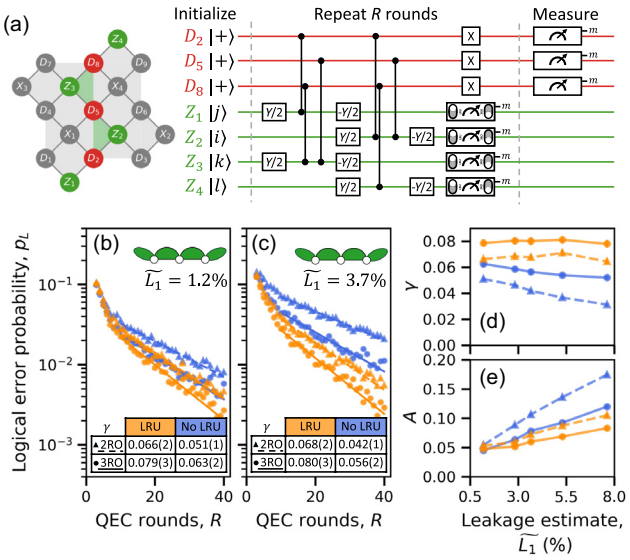

FIG. 5. Stability-7 experiment. (a) Schematic of the seven qubits used and quantum circuit. Device and conventions are the same as in Fig. 4. CZ gates with narrower horizontal spacing between them are executed simultaneously. (b),(c) Logical error probability (p_L) versus QEC rounds (R) at baseline leakage ($\widetilde{L_1} = 1.2\%$) and intermediate leakage ($\widetilde{L_1} = 3.7\%$). The legend is the same as in Fig. 4. The data (scatters) are fit with $p_L = Ae^{-\gamma R}$ starting from round 10 (solid lines). (d,e) Extracted fitting parameters γ (higher is better) and A (lower is better) as a function of the leakage estimate $\widetilde{L_1}$. With the LRU, the error-suppression factor γ remains stable against increasing leakage. The combination of LRU with 3RO consistently provides the best performance for both γ and A.

of injected leakage. For example, at $\widetilde{L_1} = 3.7\%$ [Fig. 5(c)], γ increases by 61%, 33%, and 90%, respectively.

The non-exponential curvature of p_L at low rounds in Figs. 5(b) and 5(c) warrants further discussion, as it is not expected with LRU. From simulation, we believe that this curvature is not caused by insufficient leakage removal by the LRU, nor by leakage on the central data qubit (D_5) which does not have an LRU. We hypothesize that it may be due to leakage to higher excited states beyond $|f\rangle$. During preparation of this manuscript, we became aware of Ref. [60], which attributes this curvature to a combination of data-qubit leakage and leakage to higher excited states beyond $|f\rangle$. The absence of a clear curvature due to data-qubit leakage in our simulation may be explained by the fact that only one data qubit can leak, which does not capture the leakage mobility [13] in the experiment and therefore may be insufficient to produce a significant effect.

The central finding is that the LRU protocol protects the logical observable against leakage, maintaining a nearly constant γ even as the leakage rate increases [Fig. 5(d)]. Without leakage removal, γ degrades as the leakage rate increases. This demonstrates that the LRU is highly

effective at reducing the long-lived time correlations from leakage events, which damages the logical performance of the stability experiment. Across the studied leakage rates, the combination of the 3RO with LRU consistently provides the best performance. This is especially valuable for lattice surgery and logical qubit mobility, because using LRUs with 3RO can potentially reduce the number of rounds required to achieve a certain p_L.

IV. CONCLUSION AND OUTLOOK

We have presented the mechanism, calibration, and experimental validation of a high-fidelity and hardware-efficient LRU with zero time overhead for 2RO for superconducting qubits. By integrating the reset mechanism directly into the measurement pulse, we achieve a 98.4% leakage removal fraction from $|f\rangle$ and 2RO (3RO) with average assignment fidelity of 99.2%(96.3%). We demonstrated the practical value of this LRU-enhanced measurement by integrating it into two QEC benchmarks. In the Repetition-5 memory experiment, the logical error rate is most reduced by using 3RO, while adding the LRU helps further at finite injected leakage. In the Stability-7 experiment, the LRU successfully mitigates the propagation of leakage-induced correlated errors, evidenced by an increase in the error-suppression factor even at baseline leakage and its stabilization against all levels of injected leakage. Throughout these experiments, we consistently found that leveraging the 3RO capability of the LRU provides a significant, additional benefit to the decoder, showcasing the power of combining active error removal with richer error syndrome information.

Our work establishes a practical and highly effective method for combatting one of the most detrimental error channels in superconducting quantum processors. The zero-time-overhead nature of this technique makes it a valuable tool for improving the performance and scalability of near-term QEC implementations. Future work could focus on deploying this LRU-enhanced measurement across larger-scale QEC codes. Microwave crosstalk may play a role in parallel LRUs, which we have not fully characterized. Besides, carefully designing system parameters to improve the leakage removal efficiency is also necessary for lower overall physical error schemes. Furthermore, it can also be combined with other data-qubit LRUs [12,19–24] during measurements to mitigate leakage on data qubits.

ACKNOWLEDGMENTS

We thank Barbara Terhal, Mackenzie Shaw, and Yuli Nazarov for discussions, Nadia Haider for help in chip design, and Santiago Vallés-Sanclemente, Qblox and Orange Quantum Systems for technical contributions. This research is supported by the European Union Flagship on Quantum Technology (OpenSuperQplus100, no. 101113946), the Dutch National Growth Fund (KAT-1 and DiagnostiQ), Intel Corporation, and the Dutch Ministry of Economic Affairs (TKI).

Y. X. conceptualized the QEC experiments; L. D. C. conceptualized the LRU; Y. X. performed the experiment with contributions from S. L. M. M. and T. H. F. V.; Y. X. performed the data analysis; M. S. P. performed the NN decoding; M. S. P. provided theory support for QEC experiments; M. B. and L. D. C. designed the device; M. F. and H. M. V. fabricated the device; Y. X. wrote the manuscript with contributions from M. S. P. and feedback from S. L. M. M. and L. D. C.; L. D. C. supervised the project.

The authors declare no conflict of interests.

DATA AVAILABILITY

The data for all figures in main text and supplement are available at Ref. [61]. Scripts for training the NNs are available at Ref. [62]. Requests for additional materials should be addressed to LDC.

[1] C. K. Andersen, A. Remm, S. Lazar, S. Krinner, N. Lacroix, G. J. Norris, M. Gabureac, C. Eichler, and A. Wallraff, *Repeated quantum error detection in a surface code*, Nat. Phys. **16**, 875 (2020).

[2] L. Egan, D. M. Debroy, C. Noel, A. Risinger, D. Zhu, D. Biswas, M. Newman, M. Li, K. R. Brown, M. Cetina, and C. Monroe, *Fault-tolerant control of an error-corrected qubit*, Nature (London) **598**, 281 (2021).

[3] C. Ryan-Anderson, J. G. Bohnet, K. Lee, D. Gresh, A. Hankin, J. P. Gaebler, D. Francois, A. Chernoguzov, D. Lucchetti, N. C. Brown, T. M. Gatterman, S. K. Halit, K. Gilmore, J. A. Gerber, B. Neyenhuis, D. Hayes, and R. P. Stutz, *Realization of real-time fault-tolerant quantum error correction*, Phys. Rev. X **11**, 041058 (2021).

[4] Google Quantum AI, *Suppressing quantum errors by scaling a surface code logical qubit*, Nature (London) **614**, 676 (2023).

[5] S. Krinner, N. Lacroix, A. Remm, D. P. Agustin, E. Genois, C. Leroux, C. Hellings, S. Lazar, F. Swiadek, J. Herrmann, G. J. Norris, C. K. Andersen, M. Müller, A. Blais, C. Eichler, and A. Wallraff, *Realizing repeated quantum error correction in a distance-three surface code*, Nature (London) **605**, 669 (2022).

[6] Google Quantum AI and Collaborators, *Quantum error correction below the surface code threshold*, Nature (London) **638**, 920 (2025).

[7] A. G. Fowler, *Coping with qubit leakage in topological codes*, Phys. Rev. A **88**, 042308 (2013).

[8] C. C. Bultink, T. E. O'Brien, R. Vollmer, N. Muthusubramanian, M. W. Beekman, M. A. Rol, X. Fu, B. Tarasinski, V. Ostroukh, B. Varbanov, A. Bruno, and L. DiCarlo, *Protecting quantum entanglement from leakage*

and qubit errors via repetitive parity measurements, Sci. Adv. **6** (2020).

[9] B. M. Varbanov, F. Battistel, B. M. Tarasinski, V. P. Ostroukh, T. E. O'Brien, L. DiCarlo, and B. M. Terhal, *Leakage detection for a transmon-based surface code*, npj Quantum Inf. **6**, 102 (2020).

[10] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, *Surface codes: Towards practical large-scale quantum computation*, Phys. Rev. A **86**, 032324 (2012).

[11] P. Aliferis and B. M. Terhal, *Fault-tolerant quantum computation for local leakage faults*, Quantum Inform. Comput. **7**, 139 (2007).

[12] M. McEwen *et al.*, *Removing leakage-induced correlated errors in superconducting quantum error correction*, Nat. Commun. **12**, 1 (2021).

[13] J. Camps, O. Crawford, G. P. Gehér, A. V. Gramolin, M. P. Stafford, and M. Turner, *Leakage mobility in superconducting qubits as a leakage reduction unit*, Phys. Scr. **101**, 115107 (2026).

[14] T. Thorbeck, A. McDonald, O. Lanes, J. Blair, G. Keefe, A. A. Stabile, B. Royer, L. C. G. Govia, and A. Blais, *High-fidelity gates in a transmon using bath engineering for passive leakage reset*, arXiv:2411.04101.

[15] G. Martín-Vázquez, T. Tolppanen, and M. Silveri, *Passive leakage removal unit based on a disordered transmon array*, Phys. Rev. A **112**, 032408 (2025).

[16] J. Ghosh and A. G. Fowler, *Leakage-resilient approach to fault-tolerant quantum computing with superconducting elements*, Phys. Rev. A **91**, 020302(R) (2015).

[17] Google Quantum AI and Collaborators, *Demonstration of dynamic surface codes*, Nat. Phys. **21**, 1994 (2025).

[18] Google Quantum AI and Collaborators, *Overcoming leakage in quantum error correction*, Nat. Phys. **19**, 1780 (2023).

[19] J. F. Marques, H. Ali, B. M. Varbanov, M. Finkel, H. M. Veen, S. L. M. van der Meer, S. Valles-Sanclemente, N. Muthusubramanian, M. Beekman, N. Haider, B. M. Terhal, and L. DiCarlo, *All-microwave leakage reduction units for quantum error correction with superconducting transmon qubits*, Phys. Rev. Lett. **130**, 250602 (2023).

[20] G. Kim, A. Butler, V. S. Ferreira, X. S. Zhang, A. Hadley, E. Kim, and O. Painter, *Fast unconditional reset and leakage reduction of a tunable superconducting qubit via an engineered dissipative bath*, Phys. Rev. Appl. **24**, 014013 (2025).

[21] L. Chen *et al.*, *Fast unconditional reset and leakage reduction in fixed-frequency transmon qubits*, arXiv:2409.16748.

[22] X. Yang *et al.*, *Coupler-assisted leakage reduction for scalable quantum error correction with superconducting qubits*, Phys. Rev. Lett. **133**, 170601 (2024).

[23] N. Lacroix, L. Hofele, A. Remm, O. Benhayoune-Khadraoui, A. McDonald, R. Shillito, S. Lazar, C. Hellings, F. m. c. Swiadek, D. Colao-Zanuz, A. Flasby, M. B. Panah, M. Kerschbaum, G. J. Norris, A. Blais, A. Wallraff, and S. Krinner, *Fast flux-activated leakage reduction for superconducting quantum circuits*, Phys. Rev. Lett. **134**, 120601 (2025).

[24] Y. Xiao *et al.*, *Flexible readout and unconditional reset for superconducting multi-qubit processors with tunable Purcell filters*, Phys. Rev. Lett. **136**, 070601 (2026).

[25] K. Geerlings, Z. Leghtas, I. M. Pop, S. Shankar, L. Frunzio, R. J. Schoelkopf, M. Mirrahimi, and M. H. Devoret, *Demonstrating a driven reset protocol for a superconducting qubit*, Phys. Rev. Lett. **110**, 120501 (2013).

[26] H. Ali, J. Marques, O. Crawford, J. Majaniemi, M. Serra-Peralta, D. Byfield, B. Varbanov, B. M. Terhal, L. DiCarlo, and E. T. Campbell, *Reducing the error rate of a superconducting logical qubit using analog readout information*, Phys. Rev. Appl. **22**, 044031 (2024).

[27] See Supplemental Material at http://link.aps.org/supplemental/10.1103/y66v-988s for additional information supporting statements and claims made in the main text, which includes Refs. [28–41].

[28] T. Walter, P. Kurpiers, S. Gasparinetti, P. Magnard, A. Potočnik, Y. Salathé, M. Pechal, M. Mondal, M. Oppliger, C. Eichler, and A. Wallraff, *Rapid high-fidelity single-shot dispersive readout of superconducting qubits*, Phys. Rev. Appl. **7**, 054020 (2017).

[29] J. R. Johansson, P. D. Nation, and F. Nori, *Qutip: An open-source python framework for the dynamics of open quantum systems*, Comput. Phys. Commun. **183**, 1760 (2012).

[30] J. A. Nelder and R. Mead, *A simplex method for function minimization*, Comput. J. **7**, 308 (1965).

[31] D. I. Schuster, A. A. Houck, J. A. Schreier, A. Wallraff, J. M. Gambetta, A. Blais, L. Frunzio, J. Majer, B. Johnson, M. H. Devoret, S. M. Girvin, and R. J. Schoelkopf, *Resolving photon number states in a superconducting circuit*, Nature (London) **445**, 515 (2007).

[32] C. J. Wood and J. M. Gambetta, *Quantification and characterization of leakage errors*, Phys. Rev. A **97**, 032306 (2018).

[33] V. Negîrneac, H. Ali, N. Muthusubramanian, F. Battistel, R. Sagastizabal, M. S. Moreira, J. F. Marques, W. J. Vlothuizen, M. Beekman, C. Zachariadis, N. Haider, A. Bruno, and L. DiCarlo, *High-fidelity controlled-z gate with maximal intermediate leakage operating at the speed limit in a superconducting quantum processor*, Phys. Rev. Lett. **126**, 220502 (2021).

[34] B. M. Varbanov, M. Serra-Peralta, D. Byfield, and B. M. Terhal, *Neural network decoder for near-term surface-code experiments*, Phys. Rev. Res. **7**, 013029 (2025).

[35] M. Serra-Peralta, *Data-lru-integrated-with-measurement-for-QEC on GitHub* (2025).

[36] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, *A multi-stream bi-directional recurrent neural network for fine-grained action detection*, in *IEEE CVPR* (2016), pp. 1961–1970.

[37] J. Bausch *et al.*, *Learning high-accuracy error decoding for quantum processors*, Nature (London) **635**, 834 (2024).

[38] B. Krawczyk, *Learning from imbalanced data: Open challenges and future directions*, Progress Artif. Intell. **5**, 221 (2016).

[39] E. Peters, *Sample importance for data-driven decoding*, arXiv:2505.22741.

[40] C. Chamberland and P. Ronagh, *Deep neural decoders for near term fault-tolerant experiments*, Quantum Sci. Technol. **3**, 044002 (2018).

[41] Y.-H. Liu and D. Poulin, *Neural belief-propagation decoders for quantum error-correcting codes*, Phys. Rev. Lett. **122**, 200501 (2019).

[42] F. Battistel, B. M. Varbanov, and B. M. Terhal, *Hardware-efficient leakage-reduction scheme for quantum error correction with superconducting transmon qubits*, PRX Quantum **2**, 030314 (2021).
[43] G. D'Agostini, *Improved iterative bayesian unfolding*, arXiv:1010.0632.
[44] N. Hansen, *The CMA Evolution Strategy: A Comparing Review* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006), pp. 75–102.
[45] J. Gambetta, A. Blais, M. Boissonneault, A. A. Houck, D. I. Schuster, and S. M. Girvin, *Quantum trajectory approach to circuit QED: Quantum jumps and the Zeno effect*, Phys. Rev. A **77**, 012112 (2008).
[46] C. C. Bultink, B. Tarasinski, N. Haandbæk, S. Poletto, N. Haider, D. J. Michalak, A. Bruno, and L. DiCarlo, *General method for extracting the quantum efficiency of dispersive qubit readout in circuit QED*, Appl. Phys. Lett. **112**, 092601 (2018).
[47] C. A. Ryan, B. R. Johnson, J. M. Gambetta, J. M. Chow, M. P. da Silva, O. E. Dial, and T. A. Ohki, *Tomography via correlation of noisy measurement records*, Phys. Rev. A **91**, 022118 (2015).
[48] E. Magesan, J. M. Gambetta, A. D. Córcoles, and J. M. Chow, *Machine learning for discriminating quantum measurement trajectories and improving readout*, Phys. Rev. Lett. **114**, 200501 (2015).
[49] Google Quantum AI, *Exponential suppression of bit or phase errors with cyclic error correction*, Nature (London) **595**, 383 (2021).
[50] Y. Zhao *et al.*, *Realization of an error-correcting surface code with superconducting qubits*, Phys. Rev. Lett. **129**, 030501 (2022).
[51] N. Lacroix, A. Bourassa, F. J. H. Heras, L. M. Zhang, J. Bausch, A. W. Senior, T. Edlich, N. Shutty, V. Sivak, and A. Bengtsson, *Scaling and logic in the colour code on a superconducting quantum processor*, Nature (London) **645**, 614 (2025).
[52] T. He, W. Lin, R. Wang, Y. Li, J. Bei, J. Cai, S. Cao, D. Chen, K. Chen, and X. Chen, *Experimental quantum error correction below the surface code threshold via all-microwave leakage suppression*, Phys. Rev. Lett. **135**, 260601 (2025).
[53] N. Sundaresan, T. J. Yoder, Y. Kim, M. Li, E. H. Chen, G. Harper, T. Thorbeck, A. W. Cross, A. D. Córcoles, and M. Takita, *Demonstrating multi-round subsystem quantum error correction using matching and maximum likelihood decoders*, Nat. Commun. **14**, 2852 (2023).
[54] C. Gidney, *Stability experiments: The overlooked dual of memory experiments*, Quantum **6**, 786 (2022).
[55] L. Caune *et al.*, *Demonstrating real-time and low-latency quantum error correction with superconducting qubits*, Nat. Commun. (2026).
[56] R. Harper, C. Lainé, E. T. Hockings, C. McLauchlan, G. M. Nixon, B. J. Brown, and S. D. Bartlett, *Characterising the failure mechanisms of error-corrected quantum logic gates*, Nat. Commun. **17**, 5039 (2026).
[57] D. Horsman, A. G. Fowler, S. Devitt, and R. V. Meter, *Surface code quantum computing by lattice surgery*, New J. Phys. **14**, 123011 (2012).
[58] T. E. O'Brien, B. Tarasinski, and L. DiCarlo, *Density-matrix simulation of small surface codes under current and projected experimental noise*, npj Quantum Inf. **3**, 39 (2017).
[59] M. A. Rol, F. Battistel, F. K. Malinowski, C. C. Bultink, B. M. Tarasinski, R. Vollmer, N. Haider, N. Muthusubramanian, A. Bruno, B. M. Terhal, and L. DiCarlo, *Fast, high-fidelity conditional-phase gate exploiting leakage interference in weakly anharmonic superconducting qubits*, Phys. Rev. Lett. **123**, 120502 (2019).
[60] A. Eickbusch, M. McEwen, A. Morvan, N. Noll, V. Sivak, J. Atalaya, N. Lacroix, M. Newman, and C. Gidney, *Experimentally Probing Time-Like Errors in Quantum Error Correction with Surface Code Stability Circuits (Part 2)* (APS Global Physics Summit (March Meeting 2026), Denver, Colorado, 2026).
[61] 10.4121/902d7f9e-38bf-48a2-a2f7-52bbf7aeeedf.
[62] https://github.com/MarcSerraPeralta/data-lru-integrated-with-measurement-for-qec.