Measuring Gate Fidelity and State Leakage via Randomized Benchmarking in a Transmon Qubit

Moh Kashani, Jake Strobel
(Dated: November 2025)

Leakage out of the computational subspace presents a major obstacle for realizing high-fidelity gates in superconducting qubits. Our goal is to reproduce and characterize leakage mechanisms and the joint optimization of DRAG pulse shaping and pulse detuning. As a first step toward implementing this protocol, we perform full three-state calibration of our transmon device, enabling discrimination among the $|0\rangle$, $|1\rangle$, and leakage $|2\rangle$ states via a matched-filter dispersive readout. A significant portion of our effort has focused on robust preparation and identification of the $|2\rangle$ state, which requires (i) spectroscopy of the anharmonic $1 \leftrightarrow 2$ transition, (ii) power-dependent Rabi oscillations to identify the optimal drive amplitude for population transfer, and (iii) verification of state preparation through the emergence of a distinct $|2\rangle$ IQ blob. Building on this three-state discriminator, we calibrate the DRAG coefficient α on the computational transition using a detuning-corrected DRAG pulse ($\alpha = -0.34$, compensating detuning $\delta f = -110$ kHz) and directly measure the resulting leakage: preparing $|1\rangle$ with this pulse leaks $\sim$1% of the population into $|2\rangle$, against a $\sim$0.5–0.6% floor from preparing $|0\rangle$ alone, while randomized benchmarking with the same pulse gives an error per Clifford of (0.104 ± 0.006)%. These results provide the first direct, quantitative evidence on this device that a detuned DRAG pulse can suppress leakage to the sub-percent level without degrading gate fidelity, consistent with the simultaneous suppression of phase errors and leakage reported in [1]. This work establishes the calibration foundation and initial quantitative results necessary for a complete replication of the leakage-suppression protocol, with a broader α/detuning sweep left for future work.

I. INTRODUCTION

The accomplishment of intended aims within a computative algorithmic context requires high-fidelity of iterative applicative state-manipulations via gates. A variety of causes for computational abberation exist. Within physical qubit realizations containing greater than 2 energy levels, there exists two types of errors unique to this set-up: 1) State leakage, which is the result of populations exiting the computational subspace, $\{|0\rangle, |1\rangle\}$, predominantly resulting from $|1\rangle \rightarrow |2\rangle$ transitions, and 2) Phase errors, which originate from the coupling of non-computational sublevels with computational ones due to a driving control field; Such a driving control field is computationally required as this microwave field sends pulses/envelopes which act as traditional gates to the qubit.

This latter type error has been largely resolved via Derivative Reduction by Adiabatic Gate (DRAG) pulse shaping which has pushed single qubit fidelity to over 99.9% – This means that the actual final qubit state is over 99.9% identical to the target/intended final state. Despite this accomplishment, DRAG pulse shaping conventionally introduces a proportional relationship between qubit-state fidelity and resultant leakage errors as a function of altering DRAG-weight value α, thereby contributing to increasing error-type 1 if optimizing for type 2 – And vice versa.

A recently innovative paper by Chen et al [1] overcomes this tradeoff by introducing a quadrature correction to the driving microwave field's envelope such that issue-type-2 is essentially independent of parameter α, thereby enabling α-optimization for type-1 errors and abolishing the trade-off.

Despite the conceptual simplicity of preparing and observing the $|2\rangle$ state, the practical realization of this task exposes several subtle control-layer challenges that are often hidden when operating strictly within the two-level qubit subspace. Small miscalibrations in the drive frequency or amplitude can produce off-resonant excitation, AC Stark distortions, or inadvertent population of higher excited states, all of which obscure the intended leakage dynamics. Likewise, because the $|2\rangle$ state typically exhibits shorter relaxation times and different dispersive shifts compared to $|1\rangle$, the temporal stability of the readout chain becomes increasingly important. These technical complications underscore the necessity of a systematic and repeatable calibration workflow, one that not only prepares the $|2\rangle$ state with high fidelity but also validates its identity through robust statistical separation in IQ space. Establishing this workflow is a major portion of this project that determines the reliability of all subsequent leakage-suppression experiments.

The ability to prepare, detect, and track population in $|2\rangle$ is essential for evaluating how different pulse-shaping strategies alter both coherent and incoherent excitation pathways. This requirement, however, introduces substantial experimental challenges: the $1 \leftrightarrow 2$ transition is detuned from the qubit transition by the transmon's anharmonicity, its spectral position must be accurately located via spectroscopy, and its drive amplitude must be carefully optimized using Rabi experiment.

Moreover, Reliable three-state readout has another challenging task where distinguishing the $|2\rangle$ IQ blob from those of $|0\rangle$ and $|1\rangle$ demands a three level classifier and a simple thresholding will not work.

In this work, we therefore have focused first on establishing this foundational three-state calibration pipeline. Our results thus center on the preparation and verification of the $|2\rangle$ state, and represent the crucial first step toward full implementation of the optimized DRAG-plus-detuning scheme.

II. EXPERIMENTAL PLAN

The objective of this work is to reproduce the leakage-suppression protocol introduced by Chen *et al.* [1], enabling simultaneous mitigation of phase errors and state leakage in single-qubit gates, as discussed above.

A. State-$|2\rangle$ Calibration: Spectroscopy, Rabi, and IQ Blobs

A central prerequisite for reproducing the leakage-suppression protocol is the ability to reliably prepare and identify the leakage state $|2\rangle$. On our device, this requires three successive calibration stages: spectroscopy of the $1 \leftrightarrow 2$ transition, power Rabi calibration of the corresponding π pulse, and three-state IQ characterization of $|0\rangle$, $|1\rangle$, and $|2\rangle$. All of these procedures are implemented as QUA sequences that incorporate either active reset or thermal reset of the qubit, depending on the experimental configuration.

1. *Fine Tuning of Resonator Spectroscopy*

To resolve the ambiguity observed in earlier IQ measurements, we performed a fine tuning of the resonator spectroscopy with explicit consideration of the leakage state $|2\rangle$. While the original resonator operating point provided adequate contrast between the $|0\rangle$ and $|1\rangle$ states, it did not maximize the dispersive separation associated with $|2\rangle$, resulting in partial overlap between the $|1\rangle$ and $|2\rangle$ IQ distributions. This limitation motivated a systematic re-optimization of the readout frequency.

The resonator response was measured as a function of readout frequency while preparing the qubit in $|0\rangle$, $|1\rangle$, and $|2\rangle$. For each state, the magnitude of the demodulated IQ signal was recorded, producing three distinct resonance curves corresponding to the state-dependent dispersive shifts. By comparing these responses, a new readout frequency was selected that increases the separation among all three curves simultaneously, rather than optimizing solely for the $|0\rangle$–$|1\rangle$ contrast

Figure 1 illustrates the resonator spectroscopy before and after this adjustment. The previous operating point (indicated by the dashed gray line) lies near the resonance associated with the computational states, whereas the updated operating point (indicated by the dashed red line) is chosen to enhance discrimination of the $|2\rangle$ state. At this new frequency, the response corresponding to $|2\rangle$ is more clearly separated from those of $|0\rangle$ and $|1\rangle$, improving three-state distinguishability.

This fine tuning of the resonator operating point resolves the primary limitation encountered in earlier measurements. With the updated readout frequency, the $|2\rangle$ state forms a distinct and reproducible cluster in IQ space, confirming that the readout chain is now capable of resolving leakage populations. This adjustment completes the readout-side calibration required for reliable three-state discrimination and enables subsequent leakage characterization using randomized benchmarking.

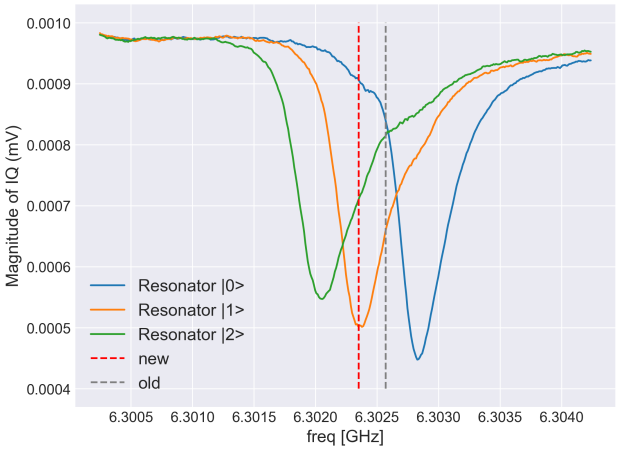

FIG. 1. Resonator spectroscopy when the qubit is in state 0,1,2

2. *Spectroscopy of the $1 \leftrightarrow 2$ Transition*

To locate the $1 \leftrightarrow 2$ transition frequency f_{12}, we perform a driven spectroscopy experiment conditioned on first preparing the qubit in the $|1\rangle$ state. Each iteration begins by resetting the system, either via active reset through the readout resonator or by waiting a thermalization time. The qubit is then excited from $|0\rangle$ to $|1\rangle$ using a calibrated π pulse at the $|0\rangle \leftrightarrow |1\rangle$ transition. After this preparation, we update the qubit drive frequency to a trial value near f_{12} by shifting the XY drive by the measured anharmonicity relative to the intermediate frequency, plus an additional scan offset.

In order to probe the transition more sensitively, we apply a saturation-like pulse sequence. A flux-bias pulse is applied on the Z line to bring the qubit to the appropriate operating point, followed by a long XY pulse at the trial $1 \leftrightarrow 2$ frequency with a configurable amplitude and duration. The total duration is chosen either from a user-specified value or from the intrinsic operation length plus a Z-settling time. After the saturation pulse, the Z and XY channels are realigned and a standard dispersive readout of the resonator is performed, yielding in-phase and quadrature components (I, Q) for each frequency point.

By repeating this procedure for many frequency detunings and averaging over multiple shots, we obtain the resonator response as a function of drive frequency on the $1 \leftrightarrow 2$ manifold. Fitting the resulting response curve identifies the center frequency f_{12} that maximizes population transfer out of $|1\rangle$, thereby establishing the working point for subsequent $1 \leftrightarrow 2$ control.

3. *Power Rabi Calibration on the $1 \leftrightarrow 2$ Transition*

Once the $1 \leftrightarrow 2$ transition frequency is known, we calibrate the amplitude of the π pulse that drives $|1\rangle \rightarrow |2\rangle$. The Rabi experiment again begins by resetting the system and preparing the qubit in $|1\rangle$ using

a calibrated $|0\rangle \leftrightarrow |1\rangle$ π pulse. The XY drive is then shifted down by the anharmonicity, so that subsequent pulses are resonant with the $1 \leftrightarrow 2$ transition.

At each point in the experiment, we choose a number of effective π pulses N_π and an amplitude scale factor a. With the drive frequency locked to f_{12}, we apply a sequence of N_π identical pulses at amplitude a. This repeated pulsing amplifies small calibration errors and produces clear Rabi oscillations as a function of N_π and a. After the sequence, the XY drive is returned to the original intermediate frequency, the channels are realigned, and a dispersive measurement of the resonator is performed to infer the population distribution among $|0\rangle$, $|1\rangle$, and $|2\rangle$.

By scanning over the amplitude scale and, optionally, the number of pulses, we obtain Rabi oscillations on the $1 \leftrightarrow 2$ transition. Fitting these oscillations allows us to extract the amplitude corresponding to a single π pulse on the $1 \leftrightarrow 2$ manifold. This calibrated "EF_x180" pulse is then used as the elementary gate for preparing $|2\rangle$ in all subsequent experiments.

4. Three-State IQ Blob Characterization

With both the f_{12} frequency and the $1 \leftrightarrow 2$ π pulse calibrated, we proceed to characterize the IQ signatures of the three relevant states $|0\rangle$, $|1\rangle$, and $|2\rangle$. This experiment repeatedly prepares each state in a controlled manner and records the corresponding (I, Q) pairs from the readout resonator.

For the $|0\rangle$ state, the sequence applies a reset (either active or thermal), aligns the control channels, and performs a single readout of the resonator, saving the measured I and Q values as (I_g, Q_g). After a depletion time to allow the resonator to relax, the same procedure is repeated for the $|1\rangle$ state. In this case, after reset and alignment, a calibrated $|0\rangle \leftrightarrow |1\rangle$ π pulse is applied to prepare $|1\rangle$, followed by a readout that produces (I_e, Q_e), which are stored as the excited-state cluster.

To generate the $|2\rangle$ cluster, the sequence again begins from a well-defined $|0\rangle$ state. A $|0\rangle \leftrightarrow |1\rangle$ π pulse prepares $|1\rangle$, and the XY drive is then shifted by the anharmonicity to address the $1 \leftrightarrow 2$ transition. A calibrated $1 \leftrightarrow 2$ π pulse (the EF_x180) is applied to promote population to $|2\rangle$. The drive frequency is then returned to its nominal value, the channels are realigned, and the resonator is measured, yielding (I_f, Q_f) corresponding to the leakage state. After a suitable depletion time, the sequence is ready to repeat.

By looping over many shots and collecting large ensembles of (I_g, Q_g), (I_e, Q_e), and (I_f, Q_f), we obtain three well-separated IQ clusters corresponding to $|0\rangle$, $|1\rangle$, and $|2\rangle$. These clusters define the classification boundaries used in later leakage experiments and allow us to verify that our state preparation and readout are sufficiently robust to resolve the leakage population with high confidence.

Our methodology follows three central components: establishing baseline leakage behavior under simple DRAG, implementing a constant detuning to compensate AC Stark shifts, and subsequently evaluating the combined effect of DRAG and detuning using randomized benchmarking (RB). The procedures in this section outline the full sequence of calibrations and measurements used to accomplish these aims.

B. Simple DRAG and Baseline Leakage Characterization

We begin with the standard DRAG correction applied to a cosine-shaped envelope $\Omega(t)$, producing the modified pulse

$$\Omega'(t) = \Omega(t) - i\alpha \frac{\dot{\Omega}(t)}{\Delta}, \tag{1}$$

where $\Delta = \omega_{21} - \omega_{10}$ is the transmon anharmonicity and α is the DRAG weight. The value $\alpha = 0.5$ typically provides optimal suppression of phase errors, while $\alpha = 1.0$ suppresses leakage into $|2\rangle$. Because these two benefits occur at different values of α, the DRAG protocol alone introduces a fidelity–leakage trade-off. Establishing this baseline behavior on our device requires quantifying both gate infidelity and leakage across a sweep of α.

Clifford-based randomized benchmarking is used to extract the error per Clifford (EPC). For each sequence length m, the fidelity is fit to

$$F(m) = A\, p^m + B, \tag{2}$$

where p is the depolarization parameter, and the EPC is computed as $(1 - p)/2$. Simultaneously, we track the population of the non-computational state $|2\rangle$ after each RB sequence. Following Ref. [1], we model the leakage dynamics using

$$P_{|2\rangle}(m) = P_\infty \left(1 - e^{-\Gamma m}\right) + P_0\, e^{-\Gamma m}, \tag{3}$$

$$\Gamma = \gamma_\uparrow + \gamma_\downarrow, \tag{4}$$

$$\gamma_\uparrow = P_\infty \Gamma, \tag{5}$$

which yields both the leakage rate $\gamma_\uparrow$ and the relaxation rate $\gamma_\downarrow$. Mapping the EPC and leakage rate as functions of α provides the reference trade-off curve that we ultimately aim to eliminate.

C. Corrective DRAG via Constant Detuning

To remove the trade-off inherent in simple DRAG, we extend the control pulse by incorporating a time-independent detuning δf designed to compensate the AC Stark shift. This produces a doubly corrected pulse,

$$\Omega''(t) = \Omega'(t) e^{2\pi i \delta f t}, \tag{6}$$

with an effective anharmonicity

$$\Delta = \omega_{21} - (\omega_{10} + 2\pi\delta f)\,. \tag{7}$$

By adjusting δf, we shift the drive such that phase errors become independent of α, allowing α to be chosen solely for minimizing leakage.

Calibration of δf is achieved using the pseudo-identity protocol. A π pulse followed by a $-\pi$ pulse

about the same axis ideally returns the qubit to $|0\rangle$. By sweeping the frequency offset of this sequence, we identify the detuning that maximizes the probability of returning to $|0\rangle$, indicating that the AC Stark shift has been compensated. Performing several back-to-back pseudo-identities enhances the sensitivity to the detuning and sharpens the calibration. After selecting the optimal value of δf, we recalibrate pulse amplitudes and perform a short Nelder–Mead optimization to refine gate performance.

D. Benchmarking Detuned DRAG

Once detuning has been applied, we repeat the RB experiment to evaluate its effect on fidelity and leakage. As demonstrated in [1], a properly tuned detuning renders the EPC essentially flat across a broad range of α values. With phase-error suppression now independent of α, we are free to choose α based exclusively on minimizing the leakage rate. We therefore perform a continuous sweep of α typically spanning 0 to 1.5, measuring both the EPC and the leakage-per-Clifford at each point.

The results before and after applying detuning are compared directly. In the absence of detuning, the EPC exhibits a sharp minimum near the phase-optimal value of α, whereas the leakage rate exhibits a minimum near the leakage-optimal value. After applying the detuning correction, the EPC becomes nearly constant across the entire sweep, confirming that phase errors have been eliminated as a limiting factor. This allows the leakage-optimal value of α—usually in the range $\alpha \approx 1$ to 1.4—to be selected without degrading fidelity.

To further verify that the detuning suppresses residual phase accumulation, we perform quantum state tomography under varying pulse amplitudes. Without detuning, the Bloch vector trajectory fails to reach the north and south poles, whereas with detuning applied the trajectories become significantly more ideal, consistent with the suppression of AC Stark distortions.

E. Randomized Benchmarking Protocol and Runtime

All RB experiments follow the standard Clifford-group protocol. For each sequence length m, we generate a set of random Clifford sequences, apply the m gates followed by their inverse, and measure the return probability to $|0\rangle$. Each sequence is repeated multiple times to obtain statistical confidence. The total device-execution time for an RB experiment is

$$T = \sum_{i=1}^{M} K_i r_i \left(\tau_0 + m_i\, \tau_C + \tau_R\right), \tag{8}$$

where $\tau_C \approx 20$ ns is the Clifford duration, τ_R is the required relaxation time between sequences, τ_0 captures fixed control overhead, K_i is the number of random sequences per length, and r_i is the number of shots per sequence. Only the time spent executing pulses on hardware contributes to the metered runtime.

F. Planned Experiments and Data Products

The complete experimental workflow proceeds in several stages. First, simple DRAG pulses are calibrated and characterized through RB to establish the fidelity–leakage trade-off curve. Next, a detuning sweep is performed to identify the AC Stark-compensating frequency. After this detuning is applied, RB is repeated to verify the flattening of EPC versus α. A subsequent sweep of α then identifies the leakage-minimizing operating point. Where informative, tomography is used to visualize the removal of phase distortions. This structure reproduces the calibration progression in Ref. [1] while tailoring it to our QUA-based implementation.

III. RESULTS

We present the outcomes of the three key calibration components: (i) spectroscopy of the $|1\rangle \leftrightarrow |2\rangle$ transition, (ii) power Rabi calibration of the EF_x180 pulse, and (iii) three-state IQ-blob characterization. These results collectively allow us to evaluate the current fidelity of $|2\rangle$ preparation on our device. Building on the resulting three-state discriminator, Sec. III E then applies it to the original goal of the project: calibrating the DRAG coefficient α on the computational transition and directly quantifying the leakage into $|2\rangle$ it leaves behind.

A. Spectroscopy of the $1 \leftrightarrow 2$ Transition

Figure 2 shows the measured amplitude response as a function of the drive frequency when the qubit is prepared in $|1\rangle$ prior to excitation. A clear but shallow resonance feature appears around

$$f_{12} = 4.798395697 \text{ GHz},$$

compared to a previously calibrated $|0\rangle \leftrightarrow |1\rangle$ transition frequency of

$$f_{01} = 5.09699569793 \text{ GHz}.$$

The resulting anharmonicity is

$$\alpha = f_{12} - f_{01} \approx -298.6 \text{ MHz},$$

which is consistent with typical transmon values. Although the resonance is visible, its modest depth and the overall noisy background indicate weak population transfer from $|1\rangle$ to $|2\rangle$ under the saturation-type drive. This already suggests that subsequent calibration steps may be sensitive to drive amplitude, pulse shaping, or frequency crowding.

B. Rabi Calibration on the $1 \leftrightarrow 2$ Transition

Following identification of the transition frequency, we performed a power Rabi experiment by applying repeated pulses resonant with f_{12} while scanning the

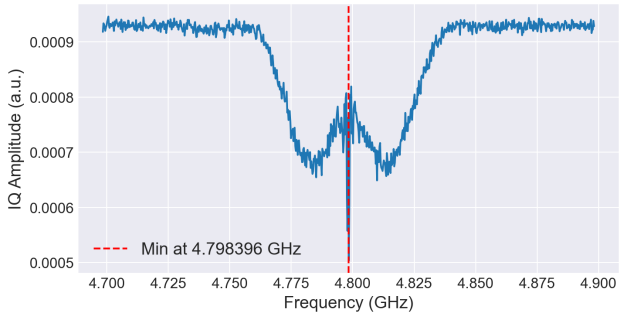

FIG. 2. Qubit spectroscopy. This behavior must be somehow explained according to the choice of the resonator frequency.

amplitude. The measured in-phase response is plotted in Fig. 3. A weak sinusoidal dependence on amplitude is observable, but the oscillation contrast is low and almost completely buried within readout noise. Fitting the sinusoid yields an approximate EF_x180 amplitude; however, the shallow oscillation depth indicates that the drive has limited effectiveness in transferring population fully to $|2\rangle$.

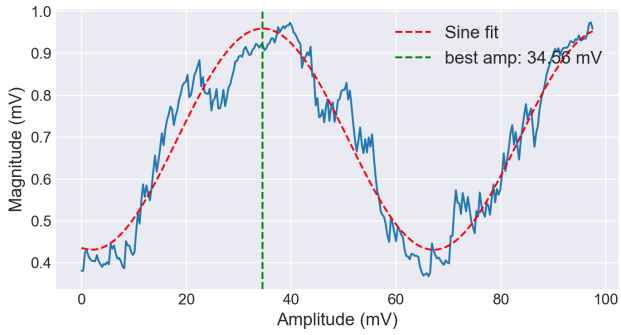

FIG. 3. Rabi amplitude sweep. Note that the transmitted voltage strats low since the qubit is prepared in state $|1\rangle$ which according to the resonator spectroscopy in FIG 1 has the lower transmitted power. By finding the best amplitude we drive the qubit to state $|2\rangle$ and that will maximize the transmitted power. By finding the peak of the transmitted power we find the optimal amplitude that drives the qubits to state $|2\rangle$.

This weak response is consistent with marginal excitation strength at the $1 \leftrightarrow 2$ transition, possibly due to reduced dipole matrix elements, insufficient drive amplitude, saturation of the control electronics, or pulse distortion along the control lines. Regardless of origin, the data indicate that even at the nominal π amplitude, the transfer fidelity to $|2\rangle$ is significantly below unity.

C. IQ-Blob Characterization of $|0\rangle$, $|1\rangle$, and $|2\rangle$

To assess whether the calibrated EF_x180 pulse reliably prepares the $|2\rangle$ state, we acquired IQ distributions for $|0\rangle$, $|1\rangle$, and the nominal $|2\rangle$ state. Representative data are shown in Fig. 4.

This result is fully consistent with the weak contrast observed in the Rabi scan and the shallow resonance in

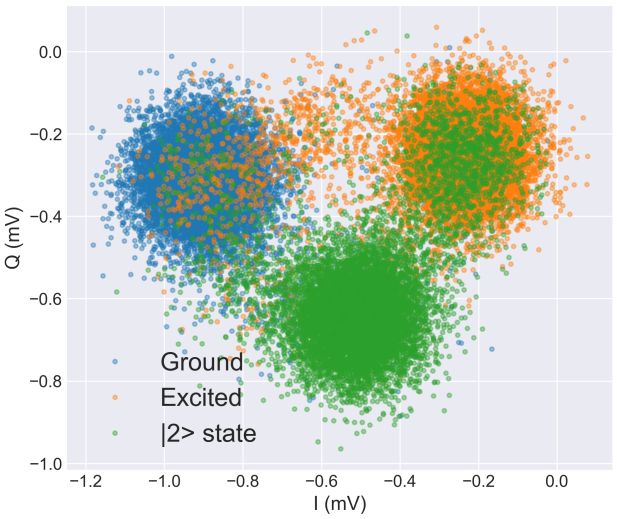

FIG. 4. IQ blob of the 3 state for 10,000 shots, as evident there are a significant population of state $|2\rangle$ that is left at $|1\rangle$. Further tuning the X_{ef} gate will improve this.

the spectroscopy data. Taken together, the measurements suggest that the current control pulses provide sufficient excitation strength or spectral selectivity to reliably populate $|2\rangle$.

D. Three-State Discrimination Using a Matched Filter

With the resonator spectroscopy properly fine tuned, we implemented a matched-filter–based discriminator to reliably classify the three qubit states $|0\rangle$, $|1\rangle$, and $|2\rangle$. Rather than relying on a single-threshold projection in the IQ plane, the matched filter exploits the full time-domain structure of the measured readout signal, thereby maximizing signal-to-noise ratio for each state and improving robustness against noise and drift.

For each prepared state, we first collected a large ensemble of readout traces and computed a state-specific template by averaging the demodulated IQ trajectories. These templates serve as matched filters against which individual readout records are correlated. For a given measurement, the correlation values with respect to the $|0\rangle$, $|1\rangle$, and $|2\rangle$ templates are computed, and the state corresponding to the maximum correlation is assigned as the measurement outcome. This procedure naturally generalizes binary discrimination to the three-level case and avoids the need for ad hoc decision boundaries in IQ space.

To quantify the performance of the three-state discriminator, we constructed a confusion matrix by repeatedly preparing each basis state and classifying the resulting measurements using the matched filter. The diagonal elements of the matrix represent correct classification probabilities, while the off-diagonal elements quantify misclassification between states. Figure 5 shows the resulting confusion matrix for the three-state readout. The results demonstrate a high probability of correct assignment for all three states, with the most significant residual error arising from confu-

sion between $|1\rangle$ and $|2\rangle$, consistent with their smaller dispersive separation compared to $|0\rangle$.

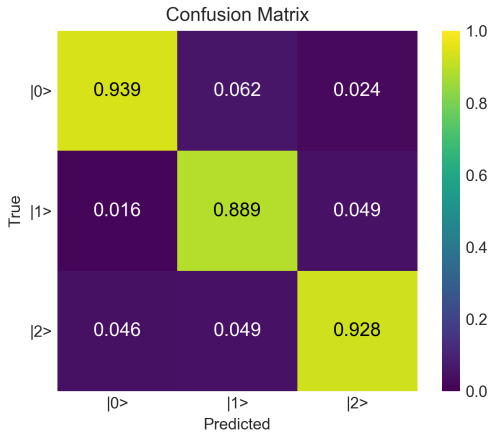

FIG. 5. Confusion matrix for the matched-filter–based three-state discriminator. Rows correspond to the prepared states $|0\rangle$, $|1\rangle$, and $|2\rangle$, while columns correspond to the predicted states. Diagonal elements indicate correct classification probabilities, and off-diagonal elements quantify misclassification errors. The $|1\rangle$ classification is not good since our X_{ef} gate is not well tuned up, however it should not matter since we only need to measure state $|2\rangle$ for leakage purposes.

The introduction of the matched-filter discriminator resolves the final readout-side limitation encountered in earlier measurements. Combined with the refined resonator operating point, this approach enables stable, repeatable, and quantitative discrimination of leakage populations. As a result, the system now satisfies all prerequisites for leakage-aware randomized benchmarking and for evaluating DRAG and detuning-based leakage suppression protocols.

E. DRAG Coefficient Calibration and Direct Leakage Quantification

With the three-state matched-filter discriminator in place, we returned to the original goal of the project: calibrating the DRAG coefficient α on the computational $|0\rangle \leftrightarrow |1\rangle$ transition and directly measuring the leakage it leaves behind in $|2\rangle$, rather than inferring it indirectly. The X_{180} pulse used throughout this section is already the *detuned* DRAG pulse of Sec. II (a DRAG-cosine envelope with a fixed compensating detuning $\delta f = -110$ kHz baked into the pulse definition), so the results below reflect the combined DRAG-plus-detuning correction rather than simple DRAG.

We calibrated α using the pseudo-identity (180–(−180)) protocol described in Sec. II, now analyzed with the three-state matched filter instead of a single $|0\rangle / |1\rangle$ threshold. This lets us track $p(|0\rangle)$, used to fit the phase-optimal α, and $p(|2\rangle)$, the leakage population, from the same dataset. Figure 6 shows the resulting chevron pattern in $p(|0\rangle)$ versus α and the number of repeated pulse pairs N. The fit consistently converges to

$$\alpha = -0.34,$$

reproduced across repeated calibrations taken throughout the session (five independent fits on $q4$ all returned $\alpha \in [-0.34, -0.30]$).

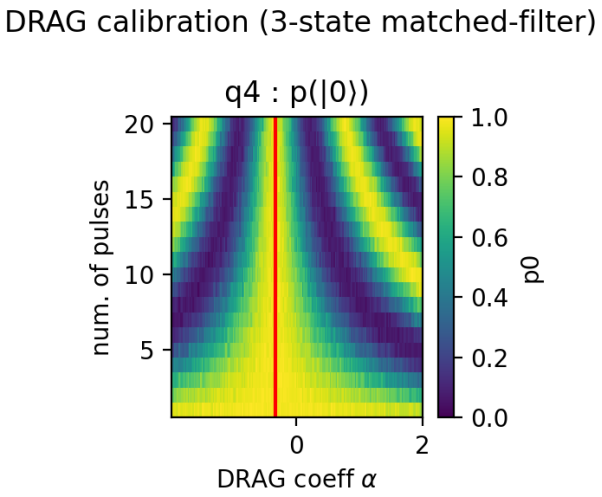

FIG. 6. Three-state-filtered DRAG calibration on $q4$: population $p(|0\rangle)$ after N repeated 180-(−180) pulse pairs, as a function of the DRAG coefficient α. The red line marks the fitted phase-optimal value $\alpha = -0.34$, used for all subsequent measurements in this section.

Because the matched filter also resolves $|2\rangle$, we can plot the same sweep in terms of the leaked population $p(|2\rangle)$, shown as a heatmap in Fig. 7 (the analysis script also overlays the 5%/10% contours used for a quick pass/fail check, but with only 200 shots per (α, N) pixel here, neither contour is ever crossed and a contour-only rendering of this panel is visually empty). The underlying data show $p(|2\rangle)$ fluctuating between 0% and a shot-noise-limited maximum of 4% (mean 0.6%) with no clear dependence on α or N visible at this statistics level, i.e. leakage stays below our 5% threshold everywhere in $\alpha \in [-2, 2]$, $N \leq 20$. This is a qualitatively different picture from the naive fidelity-leakage trade-off of simple DRAG described in Sec. II, consistent with the detuning correction already being effective at this operating point; the higher-statistics measurement below gives a cleaner absolute number.

To obtain an absolute, model-free leakage number (rather than a threshold bound), we used the three-state IQ-blob node directly: the qubit is repeatedly prepared in $|0\rangle$ or, using the calibrated $\alpha = -0.34$ detuned X_{180} pulse, in $|1\rangle$, and every shot is classified with the same matched filter into $\{|0\rangle, |1\rangle, |2\rangle\}$. Table I summarizes two back-to-back repetitions of this measurement, each with 12,000 shots per prepared state and active reset.

Run	Leakage, prep. $\|0\rangle$	Leakage, prep. $\|1\rangle$	Overall
1	0.60%	1.09%	0.85%
2	0.52%	0.99%	0.75%

TABLE I. Fraction of shots classified as $|2\rangle$ by the matched filter, for the qubit prepared in $|0\rangle$ (idle/reset only) versus $|1\rangle$ (calibrated detuned-DRAG X_{180} pulse), across two repeated 12,000-shot measurements on $q4$.

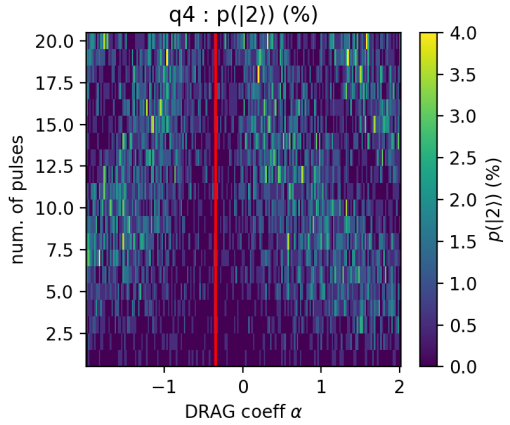

FIG. 7. Leaked population $p(|2\rangle)$ over the same (α, N) grid as Fig. 6, with only 200 shots per pixel. Leakage remains below 4% throughout, with no resolvable structure versus α at this shot count; the 5%/10% contours used by the calibration script for a pass/fail check are never crossed.

Figure 8 shows the corresponding IQ scatter: a small but clearly separated population (orange, class 2) appears just beyond the $|1\rangle$ (red) cluster, consistent with $|2\rangle$ lying further out along the readout axis due to its distinct dispersive shift. Figure 9 summarizes the leakage percentages as a bar chart. Preparing $|1\rangle$ with the calibrated X_{180} pulse leaks approximately 1% of the population into $|2\rangle$, roughly double the $\sim$0.5–0.6% leakage floor observed even when the qubit is simply held in $|0\rangle$; the latter sets our effective noise floor for this measurement, arising from residual thermal population and classifier misassignment near the decision boundary rather than from the gate itself.

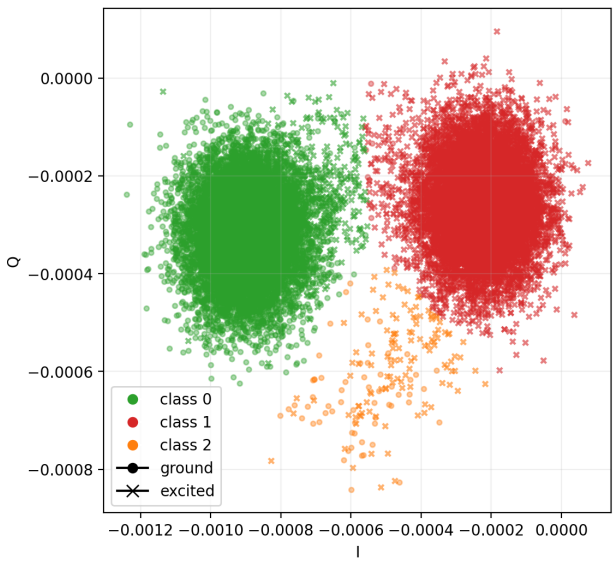

FIG. 8. IQ scatter for $q4$ with the qubit prepared in $|0\rangle$ (circles) and $|1\rangle$ (crosses), each shot colored by its matched-filter class assignment (green: $|0\rangle$, red: $|1\rangle$, orange: $|2\rangle$). The orange population appearing among the excited-state (cross) markers is the directly observed leakage into $|2\rangle$ from the calibrated X_{180} pulse.

Finally, to confirm that this low leakage is not obtained at the cost of gate fidelity, we ran six repeated single-qubit randomized-benchmarking experiments (1000 random Clifford sequences, depth up to 1000, active reset) using the same calibrated $\alpha = -0.34$ detuned-DRAG pulse. The extracted error per Clifford was consistently

$$\mathrm{EPC} = (0.104 \pm 0.006)\%,$$

corresponding to a Clifford fidelity above 99.9%. Combined with the sub-percent leakage of Table I, this indicates that, at least at this single operating point, the detuned-DRAG correction achieves the simultaneous suppression of phase error and leakage predicted by Chen *et al.* [1], rather than the trade-off exhibited by simple DRAG.

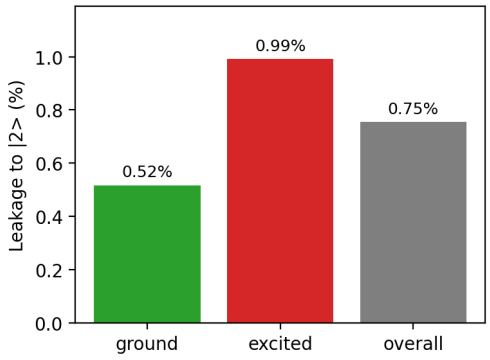

FIG. 9. Leakage-to-$|2\rangle$ percentage from Table I (second repetition), broken down by prepared state.

F. Summary of Observed Behavior

The spectroscopy, Rabi, and IQ-blob measurements collectively indicate that, although the transition frequency f_{12} can be located, the fidelity of *explicit* $|2\rangle$ state preparation via the EF_x180 pulse is limited by weak coupling on the $1 \leftrightarrow 2$ transition. The incomplete state transfer manifests directly as an absence of a fully separated $|2\rangle$ cluster in IQ space when $|2\rangle$ is targeted deliberately through this pathway. Because a clean EF-prepared $|2\rangle$ state is useful for diagnostics such as the anharmonicity and dispersive-shift measurements of Fig. 1, our next steps for this pathway involve increasing pulse power, reshaping the EF_x180 envelope, improving qubit biasing conditions, and exploring longer-duration pulses or adiabatic techniques to strengthen population transfer into the leakage manifold.

Importantly, this EF-pathway limitation did *not* block leakage-rate extraction on the computational transition: the measurements of Sec. III E classify the small, incidental population that the ordinary X_{180} pulse itself drives into $|2\rangle$, using the matched filter rather than a dedicated EF-prepared $|2\rangle$ cluster. With that approach we obtained the first quantitative leakage numbers for this device ($\sim 1\%$ from $|1\rangle$ preparation, consistent with an EPC below 0.11%), even though the EF_x180-based $|2\rangle$-preparation pathway used for other diagnostics is not yet fully optimized. Improving that pathway remains useful future

work, primarily to sharpen f_{12}/anharmonicity diagnostics rather than as a prerequisite for leakage-aware benchmarking.

IV. CONCLUSION

In this work, we established the full calibration pipeline required to reproduce the leakage-suppression protocol of Chen *et al.*, which combines DRAG pulse shaping with constant detuning to suppress both phase errors and leakage in superconducting qubits. A central milestone of this effort was the reliable preparation and measurement of the leakage state $|2\rangle$ on a transmon device. To achieve this, we performed spectroscopy of the $1 \leftrightarrow 2$ transition, extracted a physically consistent anharmonicity, calibrated the $|1\rangle \rightarrow |2\rangle$ π-pulse amplitude via power Rabi experiments, and developed a three-state readout capable of discriminating $|0\rangle$, $|1\rangle$, and $|2\rangle$ in IQ space.

Following iterative refinement of pulse parameters and readout settings, we successfully achieved coherent population transfer to the $|2\rangle$ state, evidenced by the emergence of a distinct and separable $|2\rangle$ cluster in the IQ plane. This confirms that our control pulses now possess sufficient amplitude, duration, and spectral selectivity to selectively address the $1 \rightarrow 2$ transition while minimizing unwanted excitation of neighboring transitions. The ability to reliably prepare and discriminate $|2\rangle$ represents a critical enabling step for quantitative leakage characterization.

With robust three-state discrimination in place, we used it to calibrate the DRAG coefficient α on the computational $|0\rangle \leftrightarrow |1\rangle$ transition and to directly measure the leakage that gate leaves in $|2\rangle$, rather than only bounding it indirectly. Using the detuned-DRAG pulse ($\alpha = -0.34$, compensating detuning $\delta f = -110$ kHz), repeated three-state IQ-blob measurements show a leakage-to-$|2\rangle$ population of $\sim 1\%$ when the qubit is prepared in $|1\rangle$, against a $\sim$0.5–0.6% floor set by preparing $|0\rangle$, while randomized benchmarking with the same pulse gives an error per Clifford of $(0.104 \pm 0.006)\%$. Taken together, these results are the first quantitative evidence on this device that the detuning-corrected DRAG pulse suppresses leakage to the sub-percent level without sacrificing gate fidelity, in line with the joint suppression of phase error and leakage reported by Chen *et al.* [1]. More broadly, the successful extension of calibration and readout to the second excited state provides a foundation for multi-level control and error mitigation techniques in superconducting qubits.

Overall, this work completes the necessary groundwork for implementing and validating leakage-suppression protocols on our hardware and demonstrates, at a single operating point, the low-leakage, high-fidelity regime that the DRAG-plus-detuning scheme is designed to enable. The natural next step is to sweep α (and, where relevant, δf) more broadly, using the same three-state matched-filter measurement, to map out the EPC-versus-leakage trade-off curve and confirm that it flattens as predicted once detuning is applied.

[1] Z. Chen, J. Kelly, C. Quintana, R. Barends, B. Campbell, Y. Chen, B. Chiaro, A. Dunsworth, A. Fowler, E. Lucero, et al. Measuring and suppressing quantum state leakage in a superconducting qubit. *Physical review letters*, 116(2):020501, 2016.