


Characterizing non-Markovian off-resonant errors in quantum gates

Ken Xuan Wei,^{*} Emily Pritchett,[†] David M. Zajac, David C. McKay, and Seth Merkel[‡]
IBM Quantum, IBM T.J. Watson Research Center, Yorktown Heights, New York 10598, USA

 (Received 2 May 2023; revised 12 September 2023; accepted 16 January 2024; published 8 February 2024)

As quantum gates improve, it becomes increasingly difficult to characterize the remaining errors. Here we describe a class of coherent non-Markovian errors—excitations due to an off-resonant drive—that occur naturally in quantum devices that use time-dependent fields to generate gate operations. We show how these errors are mischaracterized using standard quantum computer verification and validation techniques that rely on Markovianity and are therefore often overlooked or assumed to be incoherent. We first demonstrate off-resonant errors within a simple toy model of Z gates created by the ac Stark effect, then show how off-resonant errors manifest in all gates driven on a fixed-frequency transmon architecture, a prominent example being incidental cross-resonance interaction driven during single-qubit gates. Furthermore, the same methodology can access the errors caused by two-level systems, showing evidence of coherent, off-resonant interactions with subsystems that are not intentional qubits. While we explore these results and their impact on gate error for fixed-frequency devices, we note that off-resonant excitations potentially limit any architectures that use frequency selectivity.

DOI: [10.1103/PhysRevApplied.21.024018](https://doi.org/10.1103/PhysRevApplied.21.024018)

I. INTRODUCTION

Quantum processing technologies have matured substantially enabling experiments on full sized logical qubits [1–5] and bringing large distance codes with small errors within reach. However, current capabilities fall short of true fault tolerance, even if approaching that threshold. In reality, the practicalities of overhead and scale necessitate quantum gate errors lower than currently realized, thereby requiring analysis of these errors in previously untested limits. In general, improving performance over generations of quantum devices requires a real-time feedback loop between characterization, design, control, and fabrication in order to determine which (potentially very small) errors dominate systems and remove them. Within this loop, the host of errors present in realistic experiments sort conveniently into two broad categories: coherent (over-rotation, detuning, etc.) and incoherent (amplitude damping, dephasing, etc.). If an error preserves coherence, it may be fixable in the control layer by clever decoupling or pulse engineering [6–9]. Outside of generic techniques, such as shortening gate times and improving circuit compilations (which are almost always already optimized with respect to the rest of a complex control trade space), correcting incoherent errors lies in the realm of error correction or

low-level hardware redesign to mitigate root cause, such as dielectric loss [10,11].

In general, the total gate error, which we would like to partition into coherent and incoherent contributions, is measured using techniques such as randomized benchmarking (RB) [12]. In RB, a random sequence of Clifford gates is applied then inverted at the end of the sequence. When averaged over many sequences, the gate error is simply related to the exponential decay of the ground-state probability as a function of the sequence length [13]. Several methodologies can be used to infer coherent or incoherent contributions to that total gate error. A zeroth-order estimate of the incoherent error—the “coherence limit” [Eq. (A3) in Appendix A]—can be calculated from the gate length and measured noise rates of each qubit, in particular, amplitude damping (T_1) and dephasing (T_2). Since this does not include any dynamic reduction of coherence, the coherence limit typically underestimates the error. A more robust procedure “purity RB” measures the purity of the state after an RB sequence [14,15]; any difference between the RB and purity RB error rates can be ascribed to coherent errors, as discussed in Appendix B. A number of QCVV techniques have been devised to measure coherent errors via amplification, such as gate-set tomography (GST) [16,17], Hamiltonian tomography [18] and Hamiltonian error amplifying tomography (HEAT) sequences [19]. The HEAT sequences we use are tailored to identify small errors to the cross-resonance interaction (see, e.g., Ref. [20]), which is the entangling mechanism utilized by the quantum processors studied in this paper.

^{*}xkwei@ibm.com

[†]emily.pritchett@ibm.com

[‡]seth.merkel@ibm.com

Originally used to amplify only those block-diagonal errors correctable with standard control parameters, here we expand HEAT to include all 15 two-qubit Pauli errors (Appendix C).

We use the aforementioned RB technique to measure a typical set of $2Q$ gate errors (measured individually) on a large quantum device—here the 27 qubit *ibm_peekskill* device—is presented in Fig. 1. While the errors we observe are not well correlated with the coherence limit, they track more closely with the purity error; however, discrepancies remain. Of note, those discrepancies are not well accounted for from coherent errors that are measured from the HEAT sequences, i.e., they are not time-independent Pauli errors. Understanding possible causes of these discrepancies is the topic of this paper, and specifically, we focus on off-resonant errors. We will show that these errors are both coherent and invisible to our HEAT calibration techniques leading, at least in part, to the discrepancy in Fig. 1.

Off-resonant errors are both ubiquitous across many platforms for quantum information and problematic for standard, amplification-based characterization techniques like HEAT and long-sequence GST. These errors result from frequency selectivity, a common control technique where pulses are driven at a frequency resonant with one of the many transitions of the undriven Hamiltonian. However, due to always-on coupling in the Hamiltonian, these pulses additionally drive a number of transitions off resonantly. While a pulse of amplitude Ω detuned from

an unwanted transition (e.g., a higher transmon level or a spectator) by Δ will be suppressed to approximately $(\Omega/\Delta)^2$, there is still residual excitation, which is finite, coherent, and off-resonant with the drive pulse [21]. Ideally Ω/Δ is engineered so that off-resonance error rates fall below the rate of known incoherent processes (e.g., amplitude damping or dephasing) in our devices. Unlike incoherent processes, coherent errors should be amplifiable, allowing us to verify that off-resonant errors are indeed as small as we desire and our system models are complete and accurate. However, as we will show, off-resonant errors are invisible to HEAT, and while not strictly invisible to GST, they produce high amounts of model violation, i.e., GST completely fails to fit an error model (see Appendix D). The model failure of GST hints at the underlying issue detecting off-resonant errors: most methods in the QCVV toolbox rely on a common set of assumptions we shorthand as *Markovianity*. Concisely iterated in Ref. [22], these assumptions ensure that a gate is described by a single quantum process of fixed dimension, independent of how the gate is embedded in a larger quantum circuit. Markovianity is often invoked because it simplifies large-scale simulations and allows for rigorous mathematical statements, particularly in error correction. It is similarly a core assumption for many experimental QCVV protocols, even though most physical processes exhibit some level of non-Markovian behavior. When Markovianity is assumed erroneously, coherent, non-Markovian errors can

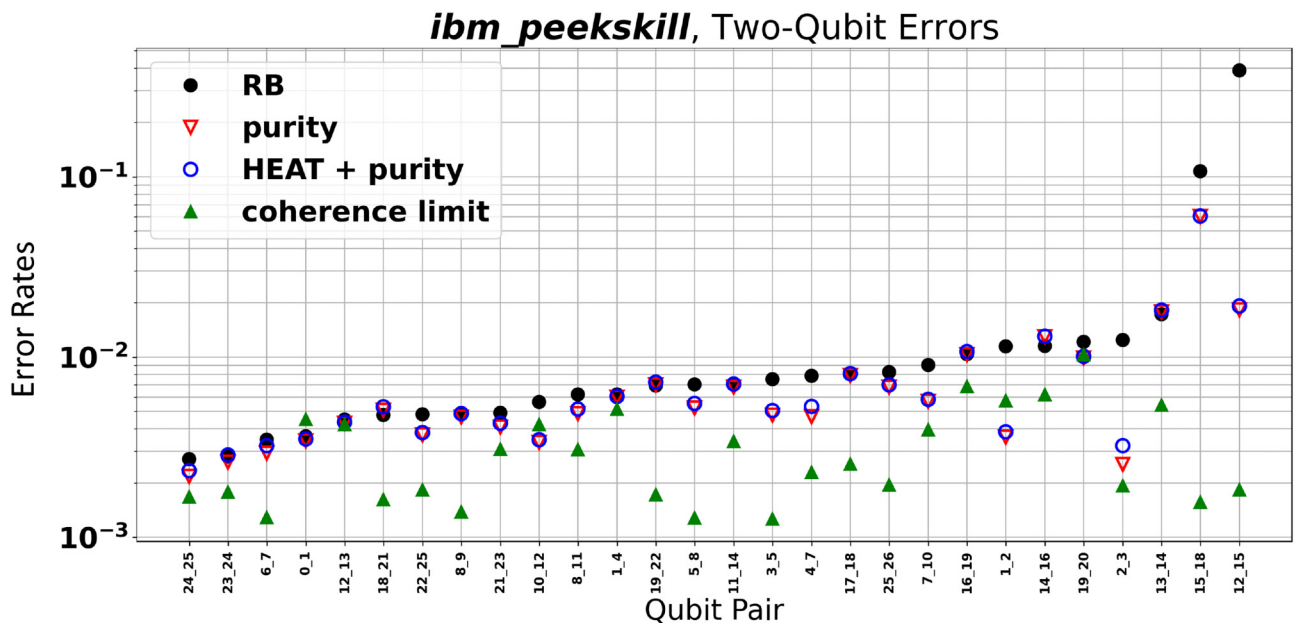


FIG. 1. A comparison of two qubit error rates across 27-qubit device *ibm_peekskill*. Randomized benchmarking (RB) estimates the total error rate per two-qubit gate (black circles) sorted by error rate. The qubit pair for the measured error is listed on the x axis where the control qubit is listed first. The gate errors are compared to the gate error from the coherence limit given by Eq. (A3) (green triangles), the error rate from purity RB (red triangles), i.e., the incoherent gate error contribution, and finally the error rates estimated by adding the purity error to the coherence error estimated by HEAT sequences (Appendix C) (blue circles).

be mischaracterized as incoherent, Markovian errors; this mistake prevents us from pinpointing error mechanisms and engineering solutions. For example, non-Markovianity caused by drifts in device parameters, which may arise due to a number of sources such as magnetic field drifts in neutral atom and ion traps, and drifts of control electronics in most quantum technologies, require different control solutions [9,23] than T_1 decay due to dielectric loss.

Our paper is organized as follows. In Sec. II, we describe how off-resonant errors break the stationary assumption of Markovianity, which states that gate errors must be independent of when the gate is performed. We show that for off-resonant errors this assumption depends on the choice of rotating frame, leading to the situation where for any frame only some, but not all, of the gates in a gate set may have Markovian descriptions. We then propose a practical alternative to GST or HEAT for amplifying and detecting these errors. By carefully interleaving frame changes into amplification sequences, off-resonant errors can be made to add constructively. Next, in Sec. III, we investigate this concept experimentally with a well-controlled version of an off-resonant error, which occurs when using a Stark tone to make a Z gate. Unless otherwise stated, the experiments in this paper were performed on an internal IBM device with specifications similar to *ibm_peekskill*. In Sec. IV we explore off-resonant errors in two-qubit gates that are generated from the cross-resonance interaction. These gates will necessarily have some off-resonant error because the control qubit is driven off-resonantly at the target qubit frequency [21]. We show how derivative removal by adiabatic gate (DRAG) compensation pulses [21,24] can be used to mitigate these errors and achieve an error rate of 1.3×10^{-3} , comparable to the lowest measured in similar CR systems [25,26]. Finally, in Sec. V we explore how spectator qubits are driven off resonantly during single-qubit gates due to always-on coupling. We also observe that these spectator “qubits” are sometimes described by spurious TLS as opposed to the engineered systems in our processors. While a small effect, spectator errors will eventually become bottlenecks as other error sources are improved, revealing the rich physics in our devices.

II. OFF-RESONANT ERRORS AND CONTINUOUS PHASE AMPLIFICATION

In this section we show that the stationary assumption of Markovianity is frame dependent, and we introduce an amplification-based characterization technique sensitive to nonstationary errors. To show how non-Markovian errors manifest we start with a toy model of the Hamiltonian of a single qubit with two drives,

$$H(t) = -\frac{\omega_q}{2}Z + \Omega_0(t) \cos([\omega_q + \Delta_0]t)X + \Omega_1(t) \cos([\omega_q + \Delta_1]t)X, \quad (1)$$

with qubit frequency ω_q , drive detunings $\Delta_{0,1}$, and envelope functions $\Omega_{0,1}(t)$. Even with a single drive, $\Omega_1 = 0$, this Hamiltonian exhibits nonstationary behavior. Imagine a family of envelope functions displaced in time by T_0 , i.e., $\Omega_0(t; 0) = \Omega_0(t + T_0, T_0)$. The resulting Hamiltonian,

$$\begin{aligned} H(t + T_0; T_0) &= -\frac{\omega_q}{2}Z + \Omega_0(t + T_0, T_0) \cos([\omega_q + \Delta_0](t + T_0))X \\ &= -\frac{\omega_q}{2}Z + \Omega_0(t, 0) \cos([\omega_q + \Delta_0](t + T_0))X \\ &\neq H(t; 0), \end{aligned} \quad (2)$$

is not stationary because the carrier does not share the envelope’s symmetry. When we move to the frame rotating at $\omega_q + \Delta_0$ and take the rotating-wave approximation (RWA) dropping terms at $\pm 2(\omega_q + \Delta_0)$, $H(t) \rightarrow \Delta_0/(2)Z + \Omega_0(t)/(2)X$, and the stationary property is restored, i.e., it is only counter-rotating terms that are nonstationary in this frame. If counter-rotating terms were the only source of nonstationary processes in our gates it would be safe to make the stationary assumption; however, it is more complicated when we have multiple drives at different frequencies. While possible to find a rotating frame where any individual term is stationary, it may be impossible to find a single rotating frame where the entire Hamiltonian exhibits stationary behavior. Even if Ω_0 and Ω_1 describe nonoverlapping pulses starting at T_0 and T_1 , respectively, the resulting unitary evolutions integrated over the nonzero domains of the envelopes, $U_0[T_0]$ and $U_1[T_1]$, cannot both be Markovian unless the two drive frequencies are commensurate. If we choose a frame where $U_0[T_0]$ is independent of T_0 , i.e., $U_0[T_0] = U_0$ is stationary, then in that frame $U_1[T_1] = U_{\text{rot}}^\dagger[T_1 + t_g]U_1U_{\text{rot}}[T_1]$, where $U_1 = U_1[T_1]$ in its stationary frame, $U_{\text{rot}}[T] = e^{-i((\Delta_1 - \Delta_0)T)/(2)Z}$ transforms the frame where U_0 is stationary to the frame where U_1 is stationary at time T , and t_g is the duration of U_1 . Clearly this operator depends on T_1 and therefore is nonstationary unless U_1 commutes with U_{rot} , and while sometimes ideal gates commute with frame transformations, their errors may not. Furthermore, Markovianity does not appear to be a property of a given gate, since either gate can be expressed as Markovian in the proper frame, but is instead a holistic property of the gate set.

We can now express why nonstationary, off-resonant errors can be tricky to quantify using calibration routines for high-fidelity gates (e.g., HEAT) that amplify coherent errors to fine tune any free parameters in Eq. (1). In the stationary case, repeated coherent errors grow quadratically, and we can design SPAM-free fitting routines. A typical amplification experiment goes as follows: prepare a superposition of eigenstates of the undriven Hamiltonian, repeat

application of a gate a number of times N , then measure some observable in the energy eigenbasis.

However, to amplify errors that anticommute with U , it is necessary to interleave some other interrogation gate, V , constructing sequences of the form $(UV)^N$. While U might have a stationary description, V might not be stationary in the same frame, and in general, *there may be no single choice of rotating frame in which all relevant gates are Markovian simultaneously*. Consider amplifying errors in the gate U_1 from above. We know there is a stationary frame for U_1 , so we can amplify those gate errors that commute with U_1 by repeating it n times in sequence. Since the frame is arbitrary, we can consider this amplification sequence in any rotating frame,

$$\begin{aligned} & U_1[nt_g] \dots U_1[t_g] U_1[0] \\ &= U_{\text{rot}}^\dagger[(n+1)t_g] U_1 U_{\text{rot}}[nt_g] \dots U_{\text{rot}}^\dagger[2t_g] U_1 U_{\text{rot}}[t_g] \\ &\quad \times U_{\text{rot}}^\dagger[t_g] U_1 \\ &= U_{\text{rot}}^\dagger[(n+1)t_g] U_1^n, \end{aligned} \quad (3)$$

and note that there is still coherent amplification. However, if we try and amplify a mixed set of U_0 and U_1 gates, where, for example, we express the sequence in U_0 's stationary frame and define all gates to have duration t_g , we get

$$\begin{aligned} & U_0[(2n-1)t_g] U_1[(2n-2)t_g] \dots U_0[t_g] U_1[0] \\ &= U_0 U_{\text{rot}}^\dagger[(2n-1)t_g] U_1 U_{\text{rot}}[(2n-2)t_g] \dots \\ &\quad \times U_0 U_{\text{rot}}^\dagger[t_g] U_1 \\ &= \prod_{j=0}^{n-1} U_0 U_{\text{rot}}^\dagger[(2j+1)t_g] U_1 U_{\text{rot}}[2jt_g]. \end{aligned} \quad (4)$$

Assuming that U_{rot} does not have a period commensurate with t_g , not only does this sequence not amplify errors in U_1 , it instead suppresses them in an average sense as the phase is essentially randomized from one application to the next.

Fortunately, techniques like HEAT and GST can be recovered for a scenario like Eq. (4) by absorbing the rotating frame into an interleaved gates (i.e., by interleaving a nonstationary gate); however this can be tedious. A simpler approach is to use an interleaved gate that commutes with the rotating frame. Projected onto a qubit, commuting interrogation gates are simply Z rotations, which can be implemented a variety of ways such as tuning the qubit energy, interleaving a delay, decomposing a Z rotation into the standard single-qubit gates, or by performing Z gates in software by updating the phase of subsequent operations [27]. Phase updates $\phi_d \rightarrow \phi_d + \phi$ can be expressed as the transformation $U \rightarrow e^{-i\phi H_{\text{phase}}} U e^{i\phi H_{\text{phase}}}$ for some appropriate choice of H_{phase} . Incrementing the phase ϕ between

successive applications of U results in the unitary

$$\begin{aligned} U_{\text{amp}} &= (e^{-i(n-1)\phi H_{\text{phase}}} U e^{i(n-1)\phi H_{\text{phase}}}) \dots \\ &\quad \times (e^{-i\phi H_{\text{phase}}} U e^{i\phi H_{\text{phase}}}) U, \end{aligned} \quad (5)$$

which can be reordered to make apparent the typical form of an amplification experiment interleaved with a diagonal interrogating gate:

$$U_{\text{amp}} = e^{-in\phi H_{\text{phase}}} (e^{i\phi H_{\text{phase}}} U)^n. \quad (6)$$

We call sweeping the phase ϕ in Eq. (6) *continuous phase amplification*, a technique we find broadly useful for identifying the off-resonant errors that can limit our device's performance. We will explore the use of this technique in the following sections.

III. STARK Z GATES

Now we consider perhaps the simplest example of off-resonant error one can generate on a driven qubit (such as the transmons considered here): the off-diagonal corrections to a $Z_{\pi/2}$ gate generated by driving a two-level system off resonantly, that is, a diagonal gate implemented by a Stark shift. The model Hamiltonian (1) becomes

$$H(t) = -\frac{\omega_q}{2} Z + \Omega(t) \cos(\omega_d t - \phi_d) X \quad (7)$$

when projected onto a single qubit. To make H stationary we choose rotating frame $H_{\text{rf}} = \omega_d/(2)Z$, which, after the RWA, leads to the effective Hamiltonian

$$H'(t) = \frac{\Delta}{2} Z + \frac{\Omega(t)}{2} (\cos \phi_d X + \sin \phi_d Y) \quad (8)$$

with detuning $\Delta \equiv \omega_d - \omega_q$. In the limit where $\Omega(t) \ll \Delta$, H' generates a rotation that is only slightly perturbed from the Z axis. To lowest order and in the square pulse approximation, this perturbation changes the effective Z rotation from $\theta_Z = \Delta t_g \rightarrow \Delta t_g + \Omega^2 t_g / 2\Delta$ over gate duration t_g ; therefore in a frame rotating with the qubit energy, the resonant frame, we see an effective Z rotation of $\theta_{\text{Stark}} = -\Omega^2 t_g / 2\Delta$. In addition to the Z rotation due to Stark shift, there are non-Markovian errors since the rotation axis for the off-resonant drive is slightly tilted from Z . Rotation around the tilted Z axis does not commute with rotation around original Z . In the qubit's resonant frame the tilted Z axis is also rotating, giving rise to time-dependent errors that are difficult to detect in experiments such as Rabi oscillations. In order to measure them directly with Hamiltonian tomography (i.e., process tomography as a function of t_g)

we need to resolve the excitation rate

$$P_{10} \equiv |\langle 1 | U_{\text{eff}}(t_g) | 0 \rangle|^2 = \frac{\Omega^2}{\Omega_r^2} \sin^2 \left(\frac{\Omega_r t_g}{2} \right), \quad (9)$$

where $U_{\text{eff}}(t_g) = e^{-i(\Omega X + \Delta Z)t_g/2}$ and $\Omega_r^2 = \Omega^2 + \Delta^2$. The maximum contrast for this signal is approximately $(\Omega/\Delta)^2$, which has no t_g dependence, requiring that our resolution and shots scale with the expected bare error rate. Repetition does not amplify these excitation errors since they anticommute with the dominant Z rotation. We could try to amplify with interrogating X or Y gates, but these resonantly driven gates are stationary in the *resonant* frame where off-resonant excitations are nonstationary, destructively interfering on average. In Sec. IV, we will show that the off-resonant error in Stark Z gates is a key source of coherent error in CNOT gates using cross-resonance drives. High-fidelity physical Z gates, such as the Stark Z gates described in this section, will be useful in quantum circuits where virtual Z gates cannot be moved across two-qubit interactions such as the $\sqrt{i\text{SWAP}}$ gate.

We have described a simple off-resonant error that is both hard to amplify using standard methods of characterization and calibration and treated as negligible under the Markovian assumption—a dangerous combination especially as error rates approach “the fault-tolerant threshold” requiring unprecedented levels of precision. Fortunately, as is the case for any coherent error, there must be *some* choice of amplification sequence that results in quadratic growth of off-resonant errors. Our solution is to use the continuous-phase amplification technique described in the previous section. We demonstrate this experimentally with the Stark $Z_{\pi/2}$ gate generated by a $t_g = 96$ ns Gaussian square pulse detuned $\Delta = -50$ MHz below the qubit’s resonance (qubit parameters given in Table I). This test case provides a very clean, accessible demonstration of off-resonance physics at gate times and detunings that are comparable to the off-resonant driving of cross-resonance gates discussed in the next sections.

Continuous-phase amplification of the Stark $Z_{\pi/2}$ is performed by the pulse sequence shown in Fig. 2(a). We perform a two-dimensional sweep over the total number of repetitions, N , and the phase incremented in between repetitions, ϕ , as the interrogating frame change (FC).

TABLE I. Summary of parameters describing the Stark $Z_{\pi/2}$ gate.

f_{01} (GHz)	5.165
α (MHz)	-346
f_{readout} (GHz)	7.083
T_1 (μs)	124(6)
$T_{2\text{echo}}$ (μs)	107(8)
$Z_{\pi/2}$ EPG	$4.36 \times 10^{-4} \pm 1.8 \times 10^{-5}$
$Z_{\pi/2\text{DRAG}}$ EPG	$2.52 \times 10^{-4} \pm 9.3 \times 10^{-6}$

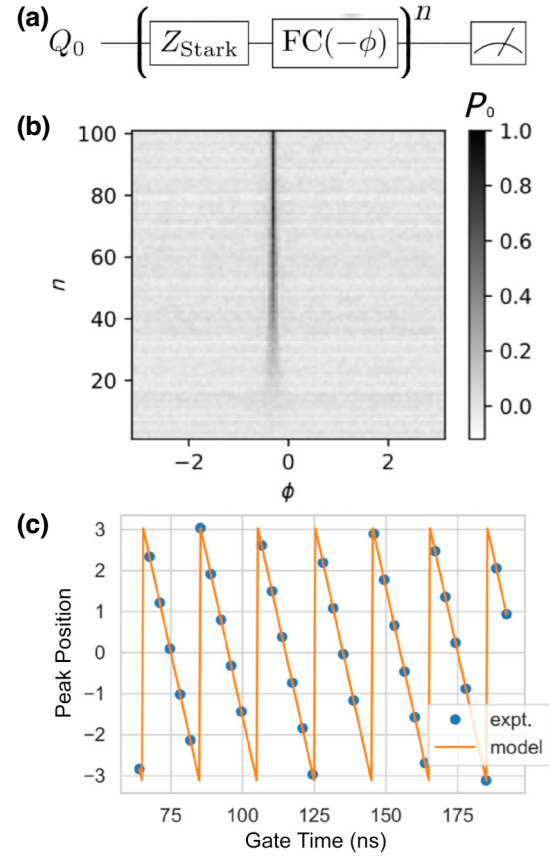


FIG. 2. (a) Continuous phase amplification of Stark $Z_{\pi/2}$ gate errors with an interrogation frame change (FC) of phase increment, ϕ . (b) shows the excitation probability of Q_0 (labeled as P_0) as a function of the phase increment and the number of repetitions with a total population inversion around $n = 100$. (c) The measured phase of maximal error amplification (expt.) has good agreement with the model predictions from Eq. (10).

The measured excited-state probability inverts completely in less than 100 repetitions, as shown in Fig. 2(b). This non-negligible error would be invisible to standard characterization techniques for which $\phi = 0$. The peak location is centered around ϕ_{peak} for which $e^{i\phi_{\text{peak}}H_{\text{phase}}U}$ from Eq. (5) is purely equatorial, that is, the phase update—which acts like a Z rotation in the qubit subspace—completely cancels the Z component of the gate leaving only the small X error. Repetition then amplifies this error giving a large signal (albeit for only a narrow range of ϕ values). Zeroing the Z component of $e^{i\phi_{\text{peak}}H_{\text{phase}}U}$, i.e., setting $\text{Tr}(Ze^{i\phi_{\text{peak}}H_{\text{phase}}U}) \propto \Delta/(\Omega_r) \cos(\phi_{\text{peak}}/2) \sin(\Omega_r t_g/2) - \cos(\Omega_r t_g/2) \sin(\phi_{\text{peak}}/2) = 0$, then expanding with respect to the small parameter Ω/Δ gives

$$\phi_{\text{peak}} \simeq \text{sign}(\Delta)\Omega_r t_g = \Delta t_g + \theta_{\text{Stark}}, \quad (10)$$

where we have used the definition of Stark shift $\text{sign}(\Delta)\omega_{\text{Stark}} = \sqrt{\Omega^2 + \Delta^2} - |\Delta|$ and $\theta_{\text{Stark}} = \omega_{\text{Stark}} t_g$. As

shown in Fig. 2(C), we see excellent agreement between predicted and experimentally extracted peak positions for different t_g demonstrating the robustness of Eq. (10). Note that for a few gate times we did not observe a peak as $\text{mod}(\Omega_r t_g, 2\pi) \simeq 0$ and excitation errors vanish. Our understanding of this particular error makes a 2D sweep to discover ϕ_{peak} unnecessary; however, in more complex examples, it is not practical to predict the phase of one (or more) peaks *a priori*.

Now that we have a tool to measure off-resonant errors we can design control sequences to correct them. A small Y rotation will correct this off-resonant error, which has rotated a $Z_{\pi/2}$ gate only slightly in the X - Z plane. Crucially, this Y rotation has to be in phase with the Stark gate. Instead of phase matching a positive Y pulse before the gate and a negative Y pulse after, we take advantage of DRAG [8,24], Fig. 3(a)—by definition a derivative pulse that is in phase with the Stark pulse up to a $\pi/2$ offset. DRAG does not lengthen the gate’s duration and only has one free parameter to calibrate, its relative amplitude. The amplified excitation at ϕ_{peak} shows a clear minima with respect to DRAG amplitude for different numbers of repetition, as shown in Fig. 3(b). Continuous-phase amplification of the Stark gate including an optimized DRAG

pulse shows a dramatic reduction of off-resonant errors [Fig. 3(c)], nearly halving the $Z_{\pi/2}$ error as measured by interleaved randomized benchmarking [Fig. 3(d)]. The gate set used in the reference RB is comprised of $X_{\pm\pi/2}$, $Y_{\pm\pi/2}$, $Z_{\pm\pi/2}$, and $Z_{0,\pi}$; where all Z gates are implemented via software frame changes [27] and X/Y gates are Gaussian pulses with gate time 4σ and $\sigma \approx 7.11$ ns. The interleaved gates are $X_{\pi/2}$, 96-ns Stark $Z_{\pi/2}$ with DRAG correction, and 96-ns Stark $Z_{\pi/2}$ without DRAG correction. The Stark Z gates use flat-topped Gaussian pulses where rise and fall are 2σ long with $\sigma \approx 14.22$ ns.

IV. NON BLOCK-DIAGONAL CNOT ERRORS

Given the results of the previous section, can we apply them to improve two-qubit gates? Fortunately, the Stark example of the previous section relates directly to the CNOT gate implemented by cross resonance (CR), where the control qubit is driven far off resonance at the target qubit’s dressed frequency. The control qubit undergoes a Stark shift due to this drive [20], and while the resulting Z rotation is corrected either by echo or with single-qubit gates, additional off-resonant excitation errors are present and potentially undetected.

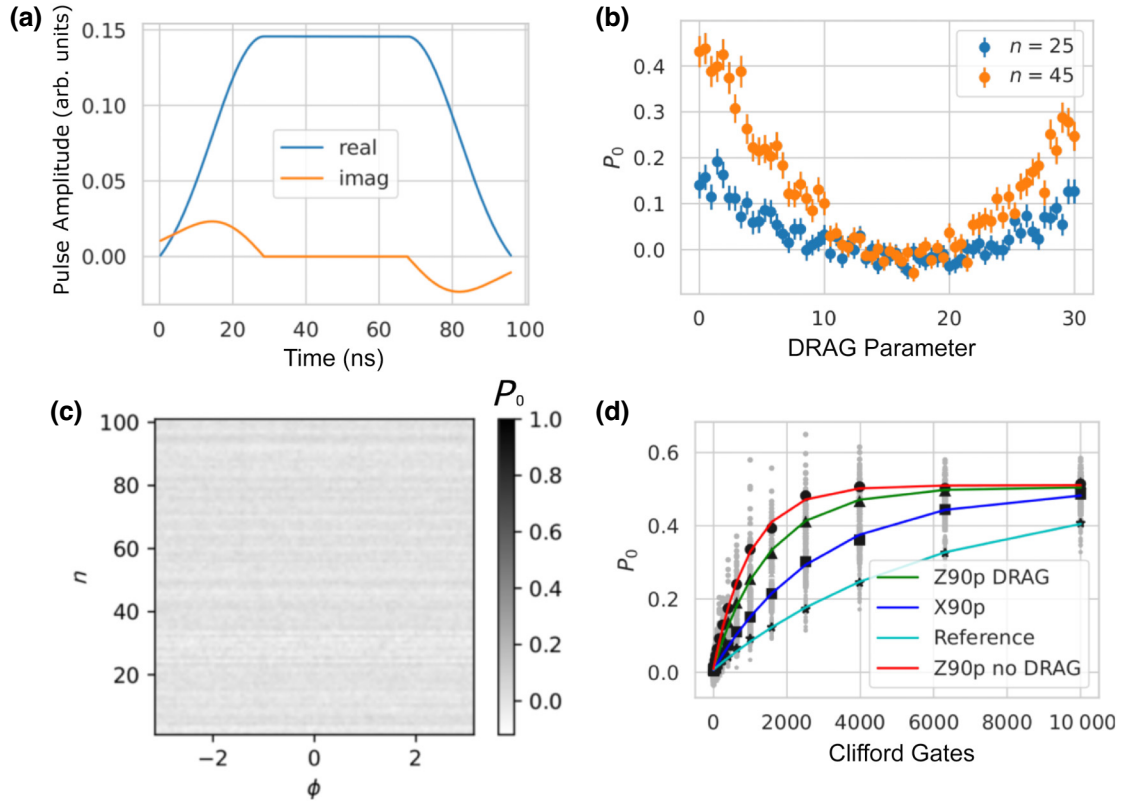


FIG. 3. (a) Drag correction to a Gaussian square pulse. (b) The continuous-phase amplification sequence, at ϕ_{peak} , shows a clear minima for the excitation probability P_0 for both $n = 25$ and $n = 40$ at the optimal DRAG amplitude. (c) For these DRAG parameters the residual off-resonant excitation from Fig. 2(b) vanishes in continuous-phase amplification. (d) The resulting error of the Stark $Z_{\pi/2}$ gate as measured by interleaved randomized benchmarking drops from 4.36×10^{-4} to 2.52×10^{-4} with DRAG (details in main text).

Expressed in the qubits' dressed basis, the full CR Hamiltonian is given by $H_{\text{CR}} = H_0 + \Omega(t) \cos(\omega_{\text{target}}t + \phi)H_{\text{drive}}$, where to leading order $H_0 = \omega_{\text{control}}ZI/2 + \omega_{\text{target}}IZ/2 + \zeta ZZ$ and $H_{\text{drive}} = XI + \mu ZX + \nu IX$. Here μ and ν describe the cross-resonance and crosstalk terms, respectively [20,28]. While ZZ errors are a topic of ongoing research at IBM [18,26], we ignore the ZZ term in the following arguments as it is fairly small and straightforward to detect by conventional techniques. Because the drive Hamiltonian contains both an XI term and a ZX term, the only choice of frame that preserves Markovianity rotates both qubits at the drive (target) frequency: $H_{\text{rf}} = \omega_{\text{target}}(ZI + IZ)/2$. After performing the RWA,

$$H_{\text{CR}}(t) = -\frac{\Delta}{2}ZI + \frac{\Omega(t)}{2} \cos(\phi_d) (XI + \mu ZX + \nu IX) + \frac{\Omega(t)}{2} \sin(\phi_d) (YI + \mu ZY + \nu IY) \quad (11)$$

where $\Delta \equiv \omega_{\text{target}} - \omega_{\text{control}}$. We must transform the control qubit to its resonant frame in order for the resonantly driven single-qubit gates to be stationary. While this frame change commutes with the CNOT gate itself, it will cause all errors that are not block diagonal with respect to the control qubit to become nonstationary (and thus hard to detect with HEAT).

To construct a CNOT from the CR interaction we typically set phase $\phi_d = 0$, choose an envelope that integrates a $ZX_{\pi/2}$, and correct the ZI and IX coefficients with single-qubit gates and/or active cancellation. Gate calibration routines evolve, as thoroughly described in Refs. [19,25, 26,29]; however, usually we neglect the off-resonant XI contribution (similarly to the naive treatment of the Stark gate) due to its small magnitude at desirable detunings. Naively executing a continuous-phase amplification experiment as described in the previous section generates a strong signal of excitation from the XI off-resonant term, but it can be more difficult to interpret. Consider a unitary gate derived from a square pulse CR gate,

$$U_{\text{CR}} = e^{-iH_{\text{CR}}t_g} = U_+ \otimes |+\rangle\langle +| + U_- \otimes |-\rangle\langle -| \\ = e^{-i[\Omega X - (\Delta - \mu\Omega)Z + \Omega\nu I]t_g/2} \otimes |+\rangle\langle +| \\ + e^{-i[\Omega X - (\Delta + \mu\Omega)Z - \Omega\nu I]t_g/2} \otimes |-\rangle\langle -| \quad (12)$$

where t_g is now the CNOT gate time, and U_{\pm} are the control unitaries when the target qubit is in the $|\pm\rangle$ states. The interrogating frame change rotates both the control and the target qubits—which are defined to be in the same frame for CR—rotating off-resonant errors on the control qubit as desired, but also scrambling U_+ and U_- and leading to complicated dynamics. In order to probe off-resonant errors in U_+ and U_- individually we need to undo the phase on the target qubit due to the interrogating frame update. To do so we prepare the target in $|+\rangle/|-\rangle$ states,

which are eigenstates parallel or antiparallel to the very first CR pulse of the amplification sequence. Then, before each repetition of the CR pulse with incremented phase ϕ , we perform an X_{π} pulse with its phase incremented by $\phi/2$ from the previous CR pulse, rotating the eigenstates to those of the next CR pulse. This amplification sequence, shown in Fig. 4(a), keeps the target qubit state parallel or antiparallel with the rotated CNOT gate, allowing the control qubit to independently evolve according to U_+ or U_- and effectively decoupling the two sectors of off-resonant errors. Adapting Eq. (10) for the CR gate we expect peak positions

$$\phi_{\pm\text{peak}} = \text{sign}(\Delta \mp \mu\Omega)\Omega_{\mp r}t_g = \Delta t_g - \theta_{\pm\text{Stark}}, \quad (13)$$

where $\Omega_{\pm r} = \sqrt{\Omega^2 + (\Delta \pm \mu\Omega)^2}$ are the Rabi rates and $\theta_{\pm\text{Stark}}$ are the rotations due to Stark shifts when the target is in the $|\pm\rangle$ states, respectively. Experimentally we observe one peak in the control qubit when the target is in the $|+\rangle$ state and another peak, separated by π , when the target is in the $|-\rangle$ state, as shown in Fig. 4(c) and 4(d). This π separation is a signature of CNOT. We refer to this particular application of continuous-phase amplification as *state-selective frame spectroscopy*.

One subtlety with this type of experiment is that the t_g variable in Eq. (13) should really denote the repetition time of the CNOT plus the interrogating X pulse. In Fig. 4(f), the peak positions extracted at a fixed number of repetitions $n = 45$ vary linearly with X -gate time with slope Δ , as expected. We evaluate $\sin(\phi_{\pm\text{peaks}})$ and fit to $\sin(at_X + \phi_0)$, taking ϕ_0 is the “extrapolated” peak position if the X pulse had no duration. In Fig. 4(e) we compare the extrapolated peak positions with Eq. (13) for different CNOT gate times and observe good agreement. The rotations due to Stark shifts $\theta_{\pm\text{Stark}}$ are measured using Ramsey experiments on the control qubit when the target is prepared in $|\pm\rangle$ state. Like in the previous section, we can optimize the DRAG parameter to suppress the off-resonant excitation in the control qubit for the CNOT gate [21]. As shown in Figs. 5(b) and 5(c), an optimized DRAG pulse eliminates both off-resonant peaks apparent in state-selective frame spectroscopy. The optimized pulse shape is shown in Fig. 5(a), and we observe an improvement in two-qubit error obtained from interleaved randomized benchmarking for the DRAG optimized CNOT, as shown in Figs. 5(d) and 5(e). The estimated error per gate (EPG) for CNOT with DRAG correction is 1.2×10^{-3} , and for CNOT without DRAG is 1.8×10^{-3} . We extract the EPGs from exponential fits to the averages of 18 different RB sequences, shown by the lines in Figs. 5(d) and 5(e). The CNOT gates use 213.33-ns-long flat-topped Gaussian pulse where rise and fall are 2σ with $\sigma \approx 14.22$ ns. We use the CNOT calibration procedure described in Refs. [25,26], where different pulse parameters are simultaneously calibrated. The reference RB sequence uses the following gate

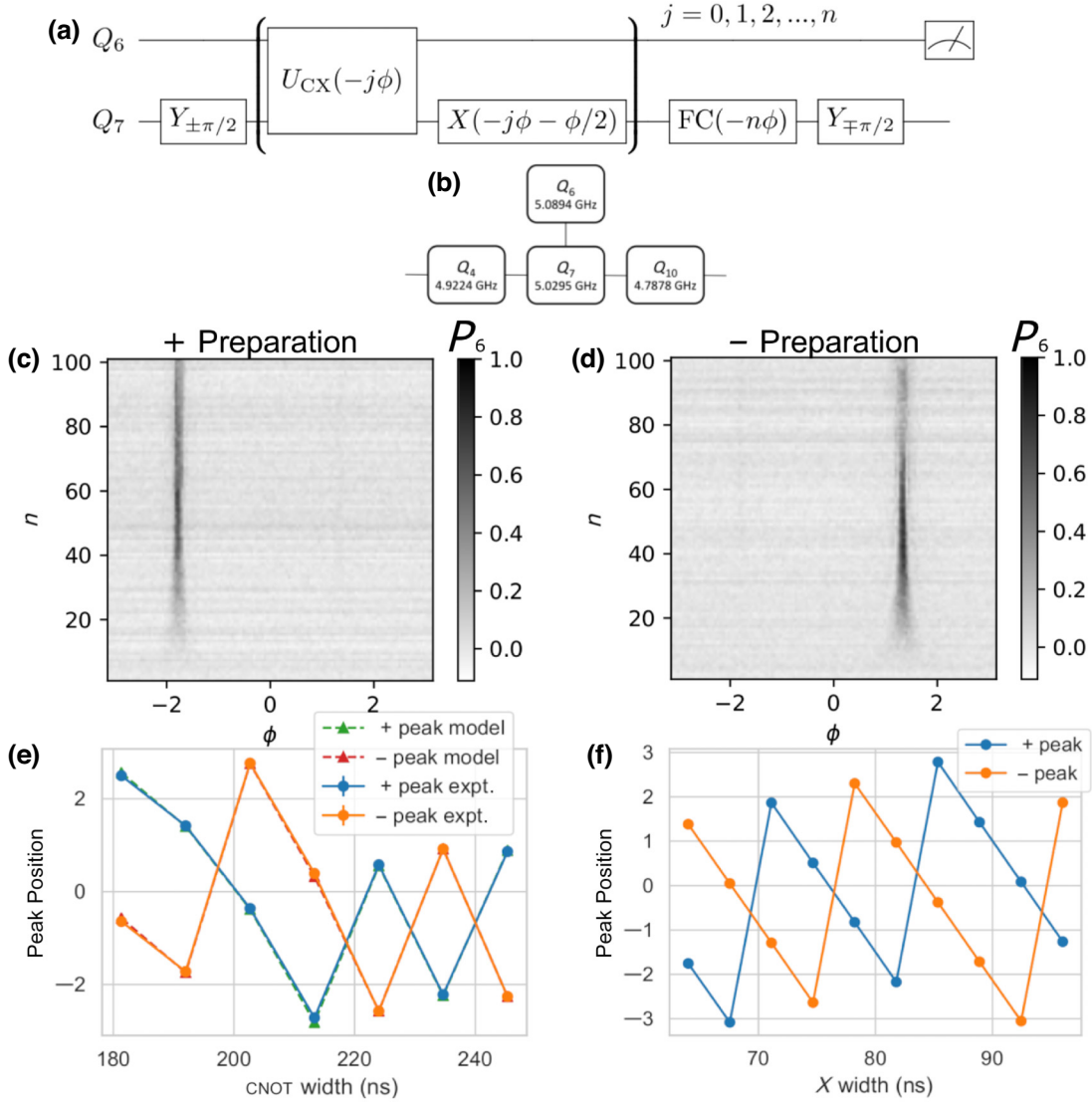


FIG. 4. State-selective frame spectroscopy amplifies CR errors using an interrogating frame change, tracking the eigenstates of the frame-shifted CR pulse with a rotated X pulse (a). We drive CR tones on the control qubit Q_6 resonant with the target qubit Q_7 (b). For the two target eigenstates $|+\rangle$ and $|-\rangle$ we observe peaks in the excitation probability (P_6) of the control qubit, as shown in (c),(d). The observed peak positions (expt.) fit well to the expression in Eq. (13) (model) where there is a dependence on both the length of the CNOT (e) and the length of the X_π pulse following the CNOT (f).

set: $X_{\pm\pi/2}$, $Y_{\pm\pi/2}$, $Z_{\pm\pi/2}$, $Z_{0,\pi}$, and DRAG corrected CNOT. The Z gates are virtual frame changes, and X/Y gates are Gaussian pulses 4σ long with $\sigma \approx 7.11$ ns. Table II summarizes the qubit parameters and gate errors, the detuning for this gate is $|\Delta| \approx 60$ MHz.

V. SINGLE-QUBIT SPECTATOR ERRORS

When neighboring qubits are subject to crosstalk, either classically through the control lines or quantum crosstalk through a fixed coupling, resonant drives can generate off-resonant errors. While these errors will show up in simultaneous RB [30], we would instead like to amplify and detect

them directly. Here we use state-selective frame spectroscopy to detect off-resonant errors in spectator qubits resulting from single-qubit gates. Consider a simplified version of the full Hamiltonian from the previous section where we now drive at the frequency of the control qubit (now labeled qubit 0), and treat the target as a spectator. When we move to the frame where both qubits oscillate at the drive frequency we obtain

$$H_{\text{sq}} = \frac{\Omega}{2}(XI + \mu ZX + \nu IX) - \frac{\Delta}{2}IZ$$

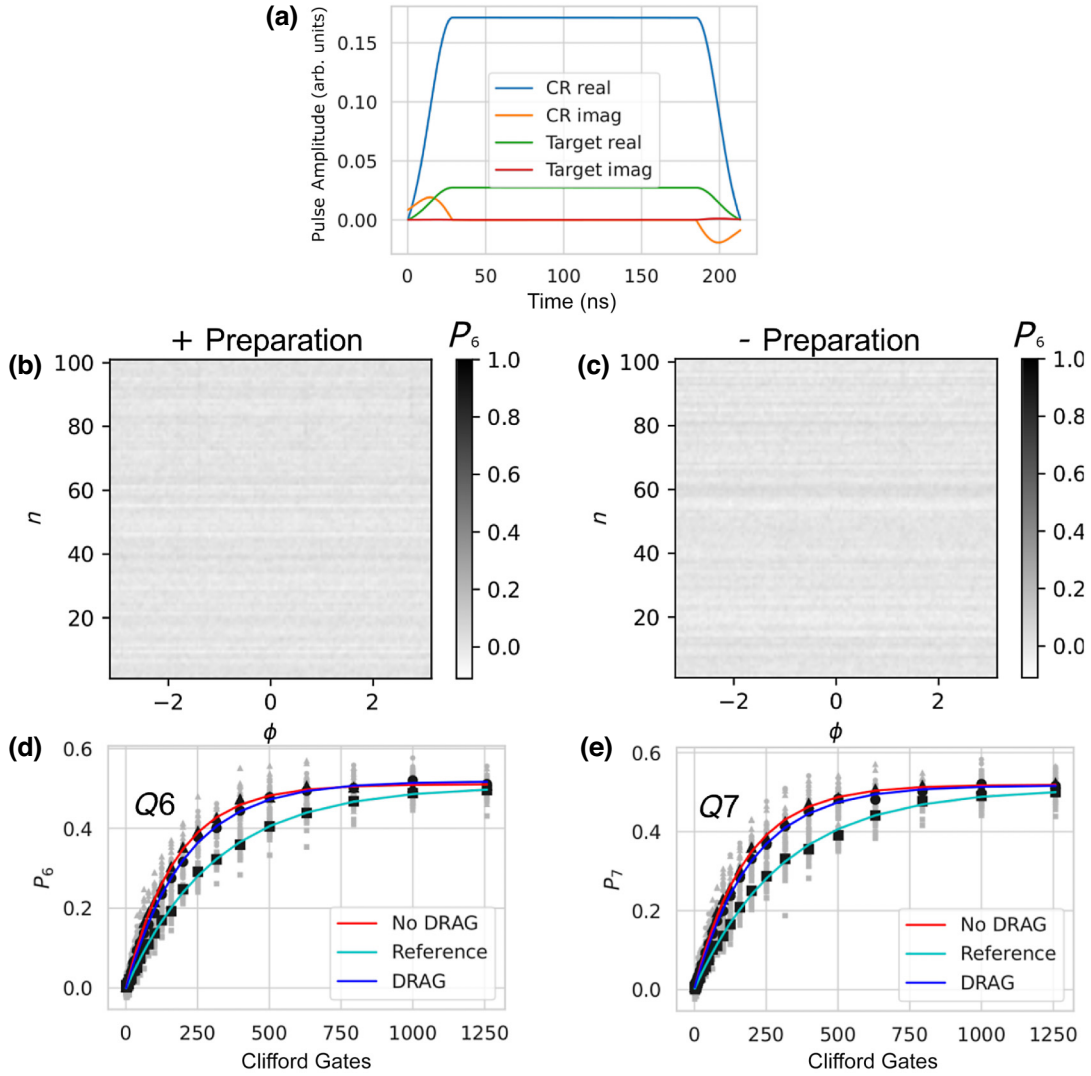


FIG. 5. (a) Optimized control pulse including DRAG and target rotary tone. In (b),(c) we see that once again the off-resonant excitation in the control qubit (Q_6) is suppressed for both target eigenstates $|+\rangle$ and $|-\rangle$. The DRAG correction reduces the error as measured by interleaved RB from 1.8×10^{-3} to 1.2×10^{-3} [(d),(e), details in main text]; P_6 and P_7 are excitation probabilities of Q_6 and Q_7 respectively.

, where $\Delta = \omega_{\text{spectator}} - \omega_q$. In the basis of $|+0\rangle$, $| - 1\rangle$, $| - 0\rangle$, and $| + 1\rangle$

$$H_{\text{sq}} = \frac{1}{2} \begin{bmatrix} -\Delta + \Omega & \mu\Omega & 0 & \nu\Omega \\ \mu\Omega & \Delta - \Omega & \nu\Omega & 0 \\ 0 & \nu\Omega & -\Delta - \Omega & \mu\Omega \\ \nu\Omega & 0 & \mu\Omega & \Delta + \Omega \end{bmatrix} \quad (14)$$

is diagonally dominant so long as $|\Delta - \Omega| \gg \mu\Omega, \nu\Omega$. Entangling and classical crosstalk are generated between the driven qubit and its spectator at rates determined by μ and ν , respectively, and both of these errors will be off resonant with the spectator qubit.

As an intuition building gedanken experiment, consider a single-qubit $X_{\pi/2}$ gate generated by a square pulse where

we further assume $\nu = 0$. Like the CR case, the Hamiltonian is block diagonal, with $|+0\rangle$ only interacting with $| - 1\rangle$, and $| - 0\rangle$ only interacting with $| + 1\rangle$. The pulse

TABLE II. Parameters describing the two qubits used in the off-resonant CNOT experiment as well as resulting error rates.

Parameters	Q_6	Q_7
T_1 (μs)	333(25)	324(22)
$T_{2\text{echo}}$ (μs)	313(45)	271(29)
f_{01} (GHz)	5.089	5.030
f_{readout} (GHz)	7.394	7.142
α (MHz)	-343	-343
CNOT EPG	$0.00191 \pm 9.9 \times 10^{-5}$	$0.00176 \pm 9.8 \times 10^{-5}$
CNOT _{DRAG} EPG	$0.00125 \pm 8.3 \times 10^{-5}$	$0.00124 \pm 9.3 \times 10^{-5}$

sequence is shown in Fig. 6(a) will amplify entanglement errors due to $\mu \neq 0$ as the CR case, modified only in that the rotating X pulse preserves the eigenstates of the primary driven qubit (instead of the target). Thanks to the state-selective nature of our pulse sequence, by preparing the initial state in either $|+\rangle$ or $|-\rangle$ we probe the two off-resonant errors independently. Using the same analysis as the two previous sections, we expect correlated peaks in both the driven and spectator qubits whose positions are given by

$$\phi_{\pm\text{peak}} = \text{sign}(\Delta \mp \Omega)\Omega_{\mp r}t_g \approx (\Delta \mp \Omega)t_g = \Delta t_g \mp \frac{\pi}{2}, \quad (15)$$

where the Rabi rates are $\Omega_{\pm r} = \sqrt{\mu^2\Omega^2 + (\Delta \pm \Omega)^2} \approx |\Delta \pm \Omega|$, and we have neglected terms proportional to μ^2 since μ is small. Experimentally we observe peaks near these predicted value ($\phi = -0.42$) in both the driven

qubit Q_6 and spectator qubit Q_7 when the initial state is in $|Q_6Q_7\rangle = |+\rangle$, as shown in Figs. 6(b) and 6(d), and π away ($\phi = 2.72$) when prepared in $|-\rangle$, as shown in Figs. 6(c) and 6(e).

In addition to peaks found on both the driven and spectator qubits, another peak appears only on the spectator qubit (at $\phi = -1.99$) regardless of the initial state of the driven qubit. This results from $\nu IX \neq 0$, which leaves the Hamiltonian non-block-diagonal. Including this term makes calculating the peak positions analytically more difficult; however, since μ and ν are relatively small, the peak positions can be estimated independently. Since the IX peak is due to an off-resonant error only on the spectator qubit, its peak position is given simply by Δt_g . The positions of the common peaks depend linearly on X -gate duration with slope given by Δ , as shown in Figs. 6(f) and 6(g), as does the position of the peak only appearing in the spectator qubit Q_7 , as shown in Fig. 6(h). Using the same technique as in the CNOT section, we extrapolate

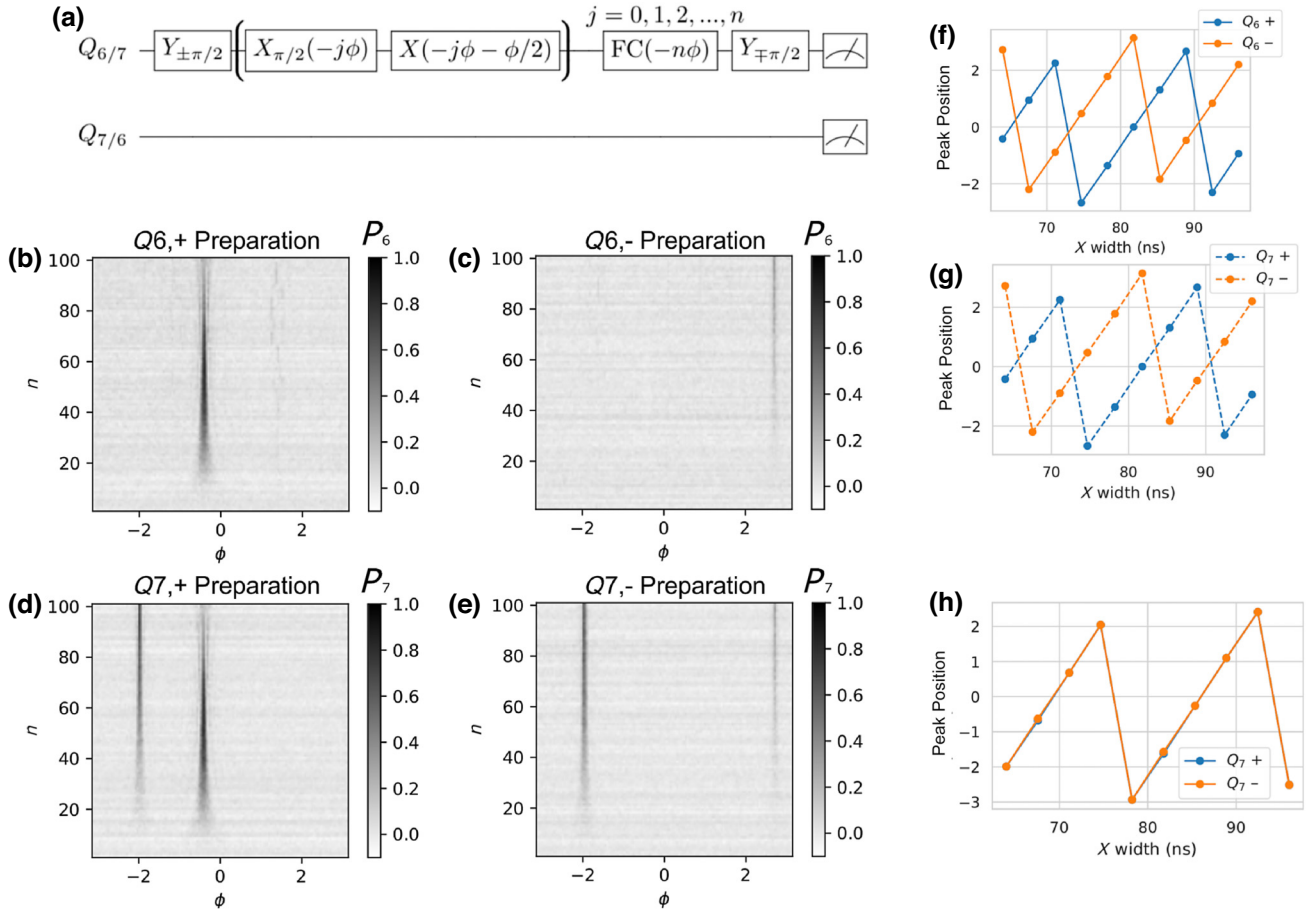


FIG. 6. State-selective frame spectroscopy of an $X_{\pi/2}$ gate (a) on either qubit Q_6 or Q_7 with the other qubit treated as a spectator. In (b),(d), the excitation probabilities for the driven qubit Q_6 (P_6) and for spectator qubit Q_7 (P_7) are measured when Q_6 is in put in the $|+\rangle$ state. In (c),(e), the excitation probabilities are measured when Q_6 is put in the $|-\rangle$ state. In (b),(d), there is a common peak in Q_6 and Q_7 appearing at the same location, which is separated from the common peak in (c),(e) by π , as expected from Eq. (15). The lone peak in (d), (e) is the IX peak. The dependence of the common peak position on the length of the X_{π} pulse are shown in (g),(f), respectively, for Q_6 and Q_7 . The dependence of the IX peak on X_{π} length is shown in (h).

TABLE III. Correlated peaks $\phi_{\pm\text{peak}}$, extrapolated to zero duration X pulse $t_X^* = 0$, on the driven and spectator qubit separated by a comma for the two different preparations: $|+0\rangle, |-0\rangle$. This agrees well with the prediction from Eq. (15) (second column). Similarly, a spectator qubit peak varies only slightly with state preparation, its observed mean (extrapolated) position $\phi_{\pm\text{peak}}^{IX}(t_X^* = 0)$ agrees well with Δt_g .

Qubits	$\phi_{\pm\text{peak}}(t_X^* = 0)$	$\Delta t_g \pm \frac{\pi}{2}$	$\phi_{\pm\text{peak}}^{IX}(t_X^* = 0)$	Δt_g
Q_6 (driven qubit)	0.662, -2.480	0.640, -2.502		
Q_7 (spectator qubit)	0.662, -2.480	0.640, -2.502	-0.969, -0.887	-0.931

the peak positions $\phi_{\pm\text{peak}}$ in the limit of a zero-duration X pulse and compare with our model Eq. (15). The results are summarized in Table III.

Next we repeat the experiments driving Q_7 so that Q_6 is the spectator. For the initial state $|Q_7Q_6\rangle = | + 0 \rangle$, we observe a peak (at $\phi = -2.7$) on both Q_7 and Q_6 and a IX peak (at $\phi = 1.99$) on Q_6 . The 2D sweep for Q_6 is shown in Fig. 7(b). Notice that the peaks in Q_6 and the peaks in Q_7 in Fig. 6(e) are reflections of each other with respect to $\phi = 0$. This makes sense since in our model exchanging the driven

and spectator qubit amounts to changing $\Delta \rightarrow -\Delta$. However, the amplitudes of the peaks are less for $|Q_7Q_6\rangle$ than $|Q_6Q_7\rangle$. For the driven qubit Q_7 , in addition to the peak shared by Q_6 , there is another peak (at $\phi = -2.2$) shared with another spectator qubit Q_4 shown in Fig. 7(c). Indeed Q_7 has three spectator qubits with Q_6 being the closest in detuning (60 MHz), followed by Q_4 (107 MHz) and Q_{10} (241 MHz). We did not observe any spectator excitation on Q_{10} , as shown in Fig. 7(d). The qubit connectivity and frequencies are shown in Fig. 4(b). In addition to obtaining

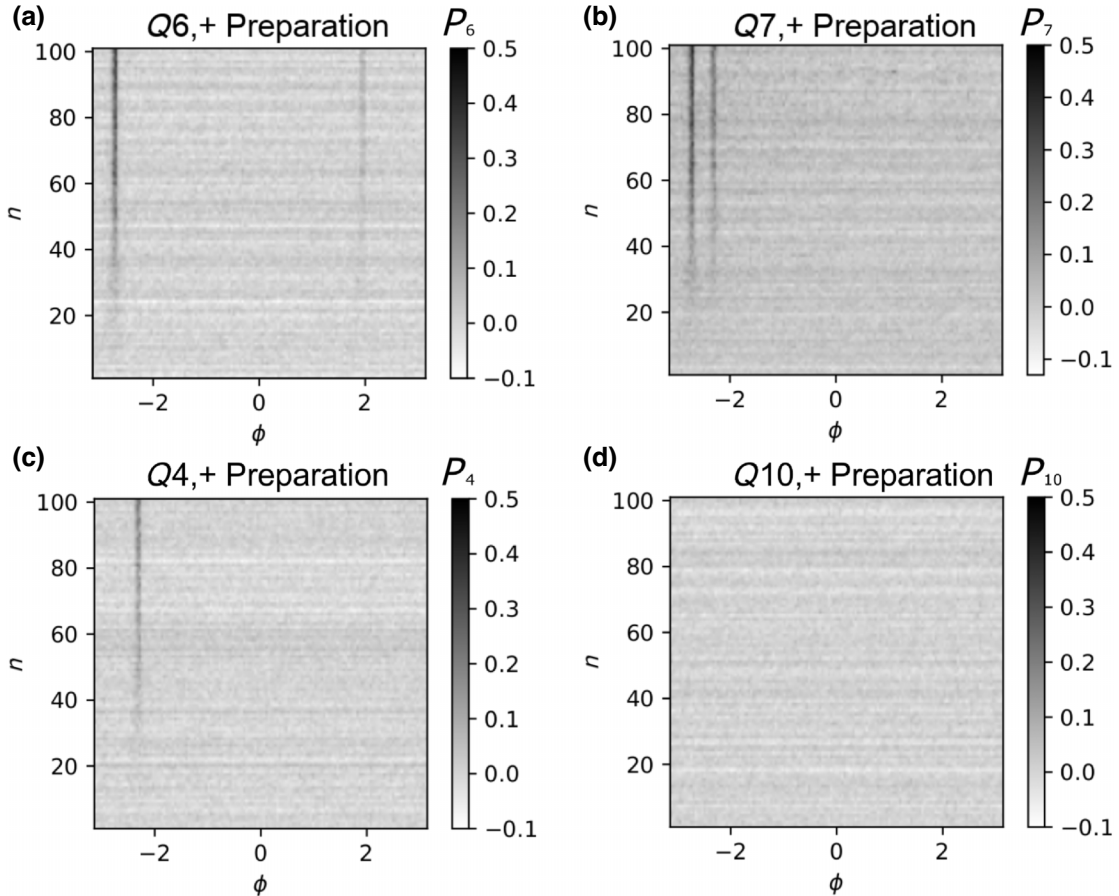


FIG. 7. State-selective frame spectroscopy of an $X_{\pi/2}$ gate performed on the driven qubit Q_7 with Q_4 , Q_6 , and Q_{10} as spectator qubits as seen in Fig. 4(b). In all cases Q_7 is prepared in $|+\rangle$ and we measure the excitation probabilities P_6 (a), P_7 (b), P_4 (c), and P_{10} (d). In (a),(b) there is a common peak in Q_6 and Q_7 appearing at the same location, and a lone IX peak in Q_6 . In (b),(c) there is a different common peak in Q_7 and Q_4 , and no IX peak in Q_4 . Lastly in (d) there are no peaks at all, since Q_{10} is far detuned from the driven qubit Q_7 .

TABLE IV. Correlated peaks $\phi_{\pm\text{peak}}$ on the driven and spectator qubit separated by a comma for the two different preparations: $|+0\rangle$, $| - 0\rangle$. This agrees well with the prediction from Eq. (15) with $t_g \rightarrow t_g + t_X$ (second column). Similarly, a spectator qubit peak varies only slightly with state preparation agreeing well with $\Delta(t_g + t_X)$.

Qubits	$\phi_{\pm\text{peak}}(t_X = 64 \text{ ns})$	$\Delta(t_g + t_X) + \frac{\pi}{2}$	$\phi_{\pm\text{peak}}^{IX}(t_X = 64 \text{ ns})$	$\Delta(t_g + t_X)$
Q_7 (driven qubit)	-2.723, -2.304	-2.732, -2.328		
Q_6 (spectator qubit)	-2.723	-2.732	1.990	1.980
Q_4 (spectator qubit)	-2.356	-2.328		

the peak positions based on extracting the $t_X \rightarrow 0$ limit, one can also keep t_X as it is and modify Eq. (15) as $\phi_{\pm\text{peak}} = \Delta(t_g + t_X) \mp \pi/2$. We use this method to analyze the peak positions observed for the case where Q_7 is the driven qubits. As shown in Table IV, we again see good agreement between experiments and model.

The $X_{\pi/2}$ pulses used in this section are Gaussian pulses with length $t_{X90} = 4\sigma$ and $\sigma = 3.55 \text{ ns}$. We used much

longer pulses, 64 ns, for the X gate in the sequence to minimize spectator off-resonant errors during that pulse since we are focusing on spectator errors due to the $X_{\pi/2}$ gate. Our analysis in this section indicates that one need not drive at an amplitude comparable to detuning to induce appreciable spectator errors; in fact, an excited qubit may exchange or swap excitation with a spectator of small detuning even in the absence of drive [31]. These errors are

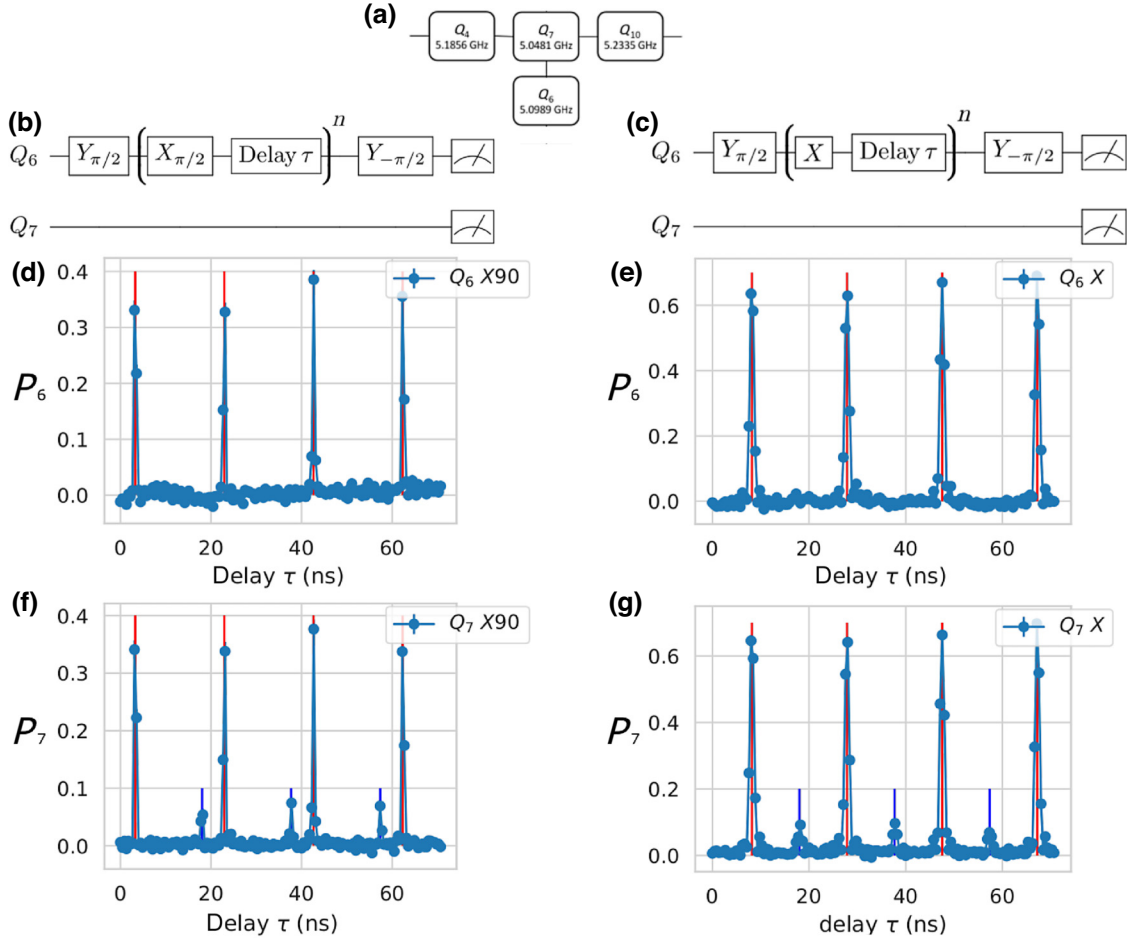


FIG. 8. Amplifying spectator errors with CPMG sequences for $X_{\pi/2}$ (b) and X_{π} (c) gates on Q_6 in (a). We measure excitation probability P_6 in (d),(e) and P_7 in (f),(g) for a $X_{\pi/2}$ gate (d),(f) and X_{π} gate (e),(g). The red vertical lines show the expected positions of the entangling peaks, and the blue vertical lines show the expected positions of spectator (IX) peaks. The entangling peaks appear at the same location in both qubits. These pulse sequences are also used in quantum noise spectroscopy (QNS) and dynamical decoupling magnetometry. The key difference is that we are not using the pulse sequence to probe noise or field during the delay, instead we are using the delay to probe and amplify the spectator error in the pulse.

off-resonant and therefore easy to overlook. Here the $X_{\pi/2}$ pulses had an average drive amplitude of 17 MHz, which is less than 1/3 of the frequency difference between Q_6 and Q_7 . Yet remarkably, the spectator error after 30 $X_{\pi/2}$ pulses on Q_6 is enough to put the driven and spectator qubits into a Bell state. We also point out that the entangling interaction is similar to the FLICFORQ gate described in Ref. [32]. More worryingly our experiments show that single-qubit gates can introduce entangling errors with multiple spectator qubits, even if they are detuned by 100 MHz or more.

We point out the key observation in Secs. IV and V is that by preparing in specific initial states and keeping track of the phase of the interleaved X_{π} pulse [see Figs. 4(a)

and 6(a)], the off-resonant errors in cross resonance and spectator can be reduced to analyzing the dynamics of two independent single-qubit systems, instead of one two-qubit system. This simplification allows us to directly obtain Eqs. (13) and (15) by reading off the Hamiltonian [Eqs. (12) and (14)] using the same reasoning leading to Eq. (10) for the Stark Z gates. We want to emphasize that Secs. III, IV, and V are closely related, and we are essentially analyzing the same problem in slightly different settings.

Unfortunately there is no simple fix to these off-resonant spectator errors using known pulse optimizations such as DRAG. For current devices and fidelity goals, we are able to work in regimes of large enough detuning to

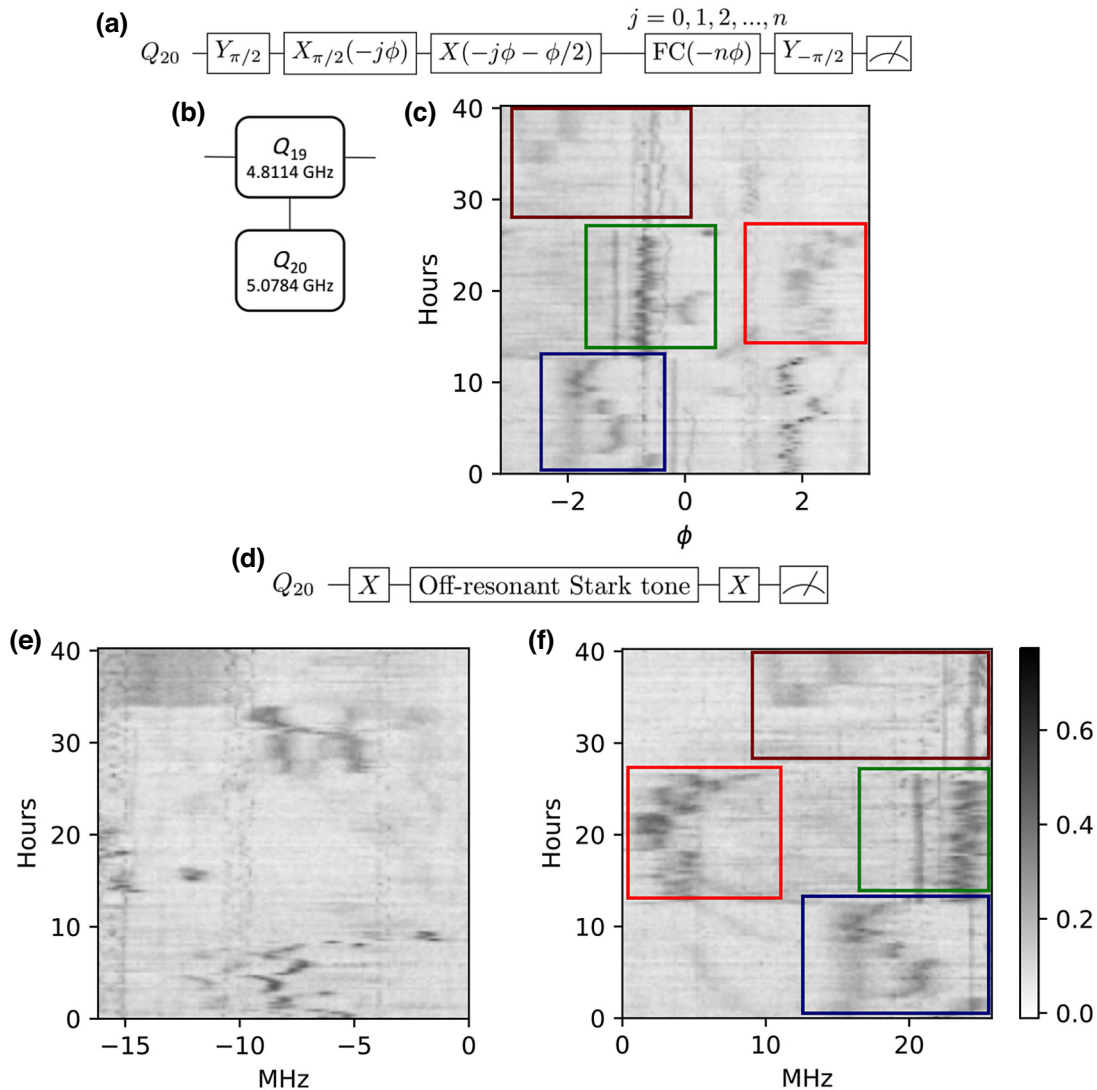


FIG. 9. We compare the features in a continuous-phase sweep (a),(c) to those observed in Stark spectroscopy (d) for an 80-MHz drive tuned above (e) and below (f) the qubits resonance frequency. This data was taken on *ibm_whiplash* and the observed lines in the phase plot are not predicted by spectator interactions as the neighboring qubit (b) is far detuned. Several of the features in the phase sweep are observed in the Stark spectroscopy (highlighted by the colored boxes as a guide to the eye), and so both experiments are probing TLS physics.

mostly ignore these off-resonant errors on single-qubit gates. Designing pulse sequences to correct these errors is out of the scope of this current paper, but could prove a significant future area of research if we find we need to relax constraints on detuning or lower single-qubit gate errors way below the current 10^{-4} levels. For the remainder of this section we will explore two additional examples of off-resonant errors in single-qubit gates.

A. Off-resonant errors and CPMG

Instead of continuous-phase sweeps, it is possible to observe off-resonant errors in the more standard framework of dynamical decoupling (DD) sequences. Consider the pulse sequences shown in Figs. 8(b), 8(c), where the driven qubit is initially prepared in $|+\rangle$ state, followed by a periodic application of either an $X_{\pi/2}$ or X_{π} interleaved by a delay τ . In the case of an X_{π} pulse this is the well-known CPMG sequence. As shown in Fig. 8, we observe excitation in both the driven and spectator qubits at regular intervals separated by $2\pi/|\Delta|$. These peaks are actually the same as those observed before in frame spectroscopy experiments. We can use $\phi_{\text{peak}} = \Delta\tau_{\text{peak}}$ and directly obtain the peak positions as $\tau_{\text{peak}} = -t_g - \theta_g + 2m\pi/|\Delta|$, where m is an integer and $\theta_g = \pi/2$ for $X_{\pi/2}$ pulses and $\theta_g = \pi$ for X_{π} pulses. On the spectator qubit we see another set of peaks. These are the IX peaks, and their positions are given by $\tau_{\text{peak}} = -t_g + 2m\pi/|\Delta|$. Observing off-resonant excitation peaks in DD is limited by the sampling resolution in the delay τ . With a large repetition number n the peak width can be quite small, one can easily lose the peaks if the resolution in τ is not small enough. For this set of data we used a different device *ibmq_cairo* where the electronics allowed us to sweep τ in increments of 0.222 ns (a $16\times$ shorter increment than available on the other two devices used). The qubit parameters are shown in Fig. 8(a). Here both $X_{\pi/2}$ and X gates are Gaussian pulses 4σ long with $\sigma = 5.33$ ns; the repetition numbers are $n = 16$ for X and $n = 32$ for $X_{\pi/2}$. We point out that compared to τ , the minimum phase increment on IBM deployed hardware is almost infinitesimal. Resolving these errors for any but the smallest detunings is much more practical using phase sweeps. We note a recent work describing non-Markovian effects in single-qubit gates in a transmon processor [33].

B. Evidence of TLS in continuous-phase sweeps

So far we have used continuous-phase sweeps to ascertain parameters in otherwise extremely well-understood Hamiltonian models. However, in superconducting qubits loss of coherence is typically described by coupling to other, less well-understood two-level systems (TLSs) typically thought to be due to microscopic irregularities in the device. The dynamics of TLS near the qubit can be

probed via Stark spectroscopy [34], spin-lock spectroscopy [35,36], and careful study of Ramsey sequences [37].

Here we point out that frame spectroscopy developed for observing off-resonant errors can also reveal TLS dynamics. We treat the TLS exactly as we do spectator qubits, although they are typically lower coherence. In Fig. 9(a) we apply the $X_{\pi/2}$ amplification sequence shown in Fig. 6(a) for a large number of repetition number ($n = 1000$), then compare to Stark TLS spectroscopy described in Ref. [34] where excited-state population is measured after 20 μs of off-resonant drive 80 MHz above or below the qubit frequency Fig. 9(f or g). The resulting spectra are monitored every 15 min for over 40 h. The qubit used is far detuned (approximately 260 MHz) from its only neighbor to minimize the effects of spectator errors described in the previous section. As shown in Fig. 9, both Stark TLS and frame spectroscopy display peaks which can move with time. While some peaks from Stark spectroscopy can be associated with peaks in the phase plots, it is clear that the phase plot reveal a much richer structure than one gets from T_1 and T_2 measurements alone.

VI. CONCLUSION

In this paper we demonstrated a broad category of “off-resonant excitation errors” that can show up in common gates utilized on quantum processors and are difficult to quantify with standard QCVV techniques. The main issue is that the assumptions of Markovianity (the stationary assumption, in particular) are violated when the gates in a gate set (and their errors) are stationary only in incommensurate rotating frames. This recasts Markovianity not as an intrinsic property of a single gate, but rather as a statement about an entire quantum processor. We developed an alternative method for characterizing these errors by using interrogation gates (described as phase updates, Z rotations, or frame changes) that are co-Markovian with the gate being measured. This allows us to detect, and in many cases mitigate, off-resonant excitations on multiple gates of a standard fixed-frequency IBM device. In the cases studied here, the magnitude of these errors are just on the cusp of becoming performance bottlenecks [33], which is perhaps not surprising if we consider the long-term efforts of optimizing device parameters with respect to metrics such as randomized benchmarking and quantum volume [29,38]. Without an understanding of the physical mechanism of the errors that affect large-scale metrics they can only be mitigated through the slow process of trial and error in device design and fabrication, or by decreasing drive amplitudes (increasing gate times) until their rates become comparable to those of the background incoherent processes. With alternative protocols, as described here, we hope that we can begin to better understand this wide variety of off-resonant errors and begin the process of engineering corrections so that they are never the limiter

of performance. While DRAG worked for certain scenarios in this paper, the most ubiquitous—spectator errors during single-qubit gates—remains uncorrected and an open question going forward. Also, as the field pushes on techniques to reduce errors algorithmically, such as error mitigation [39,40] and quantum error correction, understanding the strength of these off-resonant errors and how they impact these protocols is of highly significant. Our study should be of immediate interest to quantum noise spectroscopy (QNS) [41,42], quantum signal processing [43], and dynamical decoupling based magnetometry [44], where off-resonant errors, if not accounted for, could lead to spurious results. Furthermore, the experimental techniques developed here could be applied to study many-body resonances in many-body localized systems (MBL) [45,46], potentially shedding light to the stability of MBL systems.

ACKNOWLEDGMENTS

The authors would like to thank Luke Govia, Kentaro Heya, Abhinav Kandala, Isaac Lauer, Moein Malekakhlagh, George Stehlik, Neereja Sundaresan, and

Matthew Ware for insightful discussions and Will Shanks for software support. The devices used in this work were designed and fabricated internally at IBM. This work was supported by IARPA under LogiQ (Contract No. W911NF-16-1-0114) and the Army Research Office under QCISS (W911NF-21-1-0002). All statements of fact, opinion, or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of the U.S. Government.

APPENDIX A: COHERENCE LIMIT

A common technique used to estimate the *minimum* error of a gate is to construct an amplitude and dephasing error channel for the duration of the gate and then calculate the average gate error. This can be done by representing those error channels in the Pauli superoperator form (Pauli transfer matrix, PTM, R) and then applying the formula described in Ref. [30]

$$\epsilon = \frac{d}{d+1} \left(1 - \frac{\text{Tr}[R]}{d^2} \right) \quad (\text{A1})$$

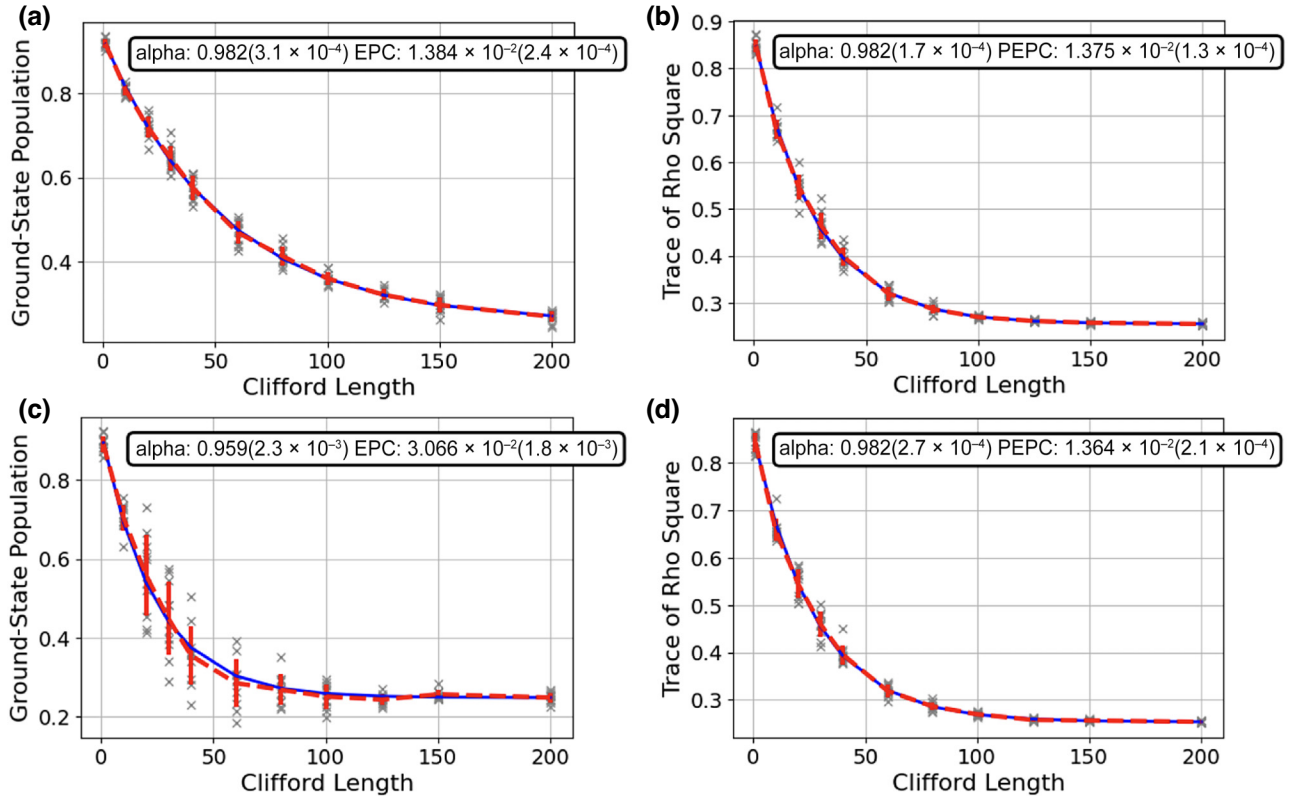


FIG. 10. Simulation of purity RB (b),(d) versus standard RB (a),(c) done in Qiskit [47]. Here we choose $T_1 = T_2 = 40 \mu\text{s}$ and for just the CX gate to be finite width of 300 ns, there is also a 3% measurement error. This has a coherence limit error of 1.35×10^{-2} per Clifford gate. For (a),(b) there is only the T_1 and T_2 decohering errors. For (c),(d) we add a coherent X rotation error after each CX gate which degrades standard RB but does not change purity. Red lines are the average and standard deviation of the RB sequences, and blue lines are fits to the average.

where $d = 2^n$. Applying this formula for the 1 qubit gate we obtain

$$\epsilon_{1Q} = \frac{1}{6} (3 - 2e^{-t_g/T_2} - e^{-t_g/T_1}), \quad (\text{A2})$$

where t_g is the length of the gate and T_1 and T_2 are the amplitude damping decay time and Ramsey decay time, respectively. For the two-qubit gate we take the tensor product of the superoperators for the qubits since the error channels are independent and obtain,

$$\begin{aligned} \epsilon_{2Q} = \frac{1}{20} & \left(15 - \sum_{i=0,1} [2e^{-t_g/T_2, Q_i} + e^{-t_g/T_1, Q_i}] \right. \\ & - e^{-t_g(1/T_1, Q_0 + 1/T_1, Q_1)} - 4e^{-t_g(1/T_2, Q_0 + 1/T_2, Q_1)} \\ & \left. - 2e^{-t_g(1/T_1, Q_0 + 1/T_2, Q_1)} - 2e^{-t_g(1/T_2, Q_0 + 1/T_1, Q_1)} \right). \end{aligned} \quad (\text{A3})$$

APPENDIX B: PURITY RB

The version of purity RB we use here is discussed in Ref. [15] and implemented in Qiskit-Ignis [47]. For each

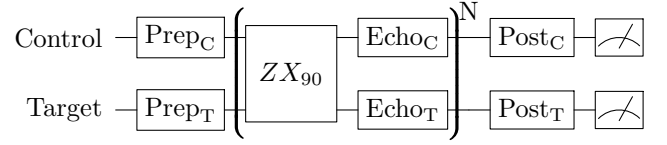


FIG. 11. The k th heat sequence \mathcal{H}_k^N .

Clifford sequences we append 3^n postrotations to measure all 4^n Pauli expectation values required to calculate the purity of the state $\text{Tr}[\rho^2]$. This can be seen from

$$\text{Tr}[\rho^2] = \text{Tr}[\sum_{ij} a_i a_j P_i P_j] = \sum_i d a_i^2, \quad (\text{B1})$$

since $\text{Tr}[P_i P_j]$ is 0 if $i \neq j$ and d otherwise (where P_i is a Pauli operator). Also note,

$$\langle P_i \rangle = \text{Tr}[\rho P_i] = \sum_i d a_i, \quad (\text{B2})$$

TABLE V. Summary table of generalized HEAT sequences to amplify coherent errors.

Sequence	Prep _C	Prep _T	Echo _C	Echo _T	Post _C	Post _T	Measure	α	Paulis
1	I	Y_{90}	X	I	X_{90}	Y_{-90}	C	$\frac{\pi^2}{16}$	$\frac{1}{2}(XI + XX)$
2	I	Y_{-90}	X	I	X_{90}	Y_{90}	C	$\frac{\pi^2}{16}$	$\frac{1}{2}(XX - XI)$
3	I	Y_{90}	I	Z	I	I	T	$\frac{\pi^2}{16}$	$\frac{1}{2}(ZZ + IZ)$
4	X	Y_{-90}	I	Z	X	X_{-90}	T	$\frac{\pi^2}{16}$	$\frac{1}{2}(ZZ - IZ)$
5	I	Y_{90}	I	Y	I	I	T	$\frac{\pi^2}{16}$	$\frac{1}{2}(ZY + IY)$
6	X	Y_{90}	I	Y	X	X_{90}	T	$\frac{\pi^2}{16}$	$\frac{1}{2}(ZY - IY)$
7	I	X_{90}	I	X	I	I	T	1	$\frac{1}{2}(ZX + IX)$
8	X	X_{90}	I	X	X	I	T	1	$\frac{1}{2}(ZX - IX)$
9	I	Y_{90}	Y	I	X_{-90}	Y_{-90}	C	$\frac{\pi^2}{16}$	$\frac{1}{2}(YI + YX)$
10	I	Y_{-90}	Y	I	X_{90}	Y_{90}	C	$\frac{\pi^2}{16}$	$\frac{1}{2}(YI - YX)$
11	I	Y_{90}	Y	Y	Y_{-90}	X_{-90}	C	$\frac{1}{2}$	$\frac{1}{2}(YY - XZ)$
12	I	Y_{-90}	X	Z	Y_{90}	X_{-90}	T	$\frac{1}{2}$	$\frac{1}{2}(YY + XZ)$
13	I	Y_{90}	X	Y	X_{90}	X_{-90}	C	$\frac{1}{2}$	$\frac{1}{2}(XY - XZ)$
14	I	Y_{-90}	Y	Z	X_{-90}	X_{-90}	T	$\frac{1}{2}$	$\frac{1}{2}(XY + XZ)$
15	Y_{90}	I	Z	I	X_{90}	I	C	1	ZI

so

$$\text{Tr}[\rho^2] = \sum_i \frac{\langle P_i \rangle^2}{d}. \quad (\text{B3})$$

If we assume depolarizing error then fitting $\text{Tr}[\rho^2]$ versus Clifford length (n) to $A\gamma^{2n} + B$ we get that the incoherent error per gate is,

$$\epsilon = \frac{3}{4}(1 - \lambda^{1/n_2}), \quad (\text{B4})$$

where n_2 is the number of $2Q$ gates per Clifford gate. We show some simulations of purity RB in Fig. 10 illustrating how coherent errors degrade standard RB, but do not affect purity RB.

APPENDIX C: STANDARD TECHNIQUES TO CHARACTERIZE COHERENT ERRORS

In Ref. [19], sequences were given to amplify ZX , ZY , ZZ , IY , and IZ errors to the cross-resonance gate, as these errors are directly addressed during calibration of the

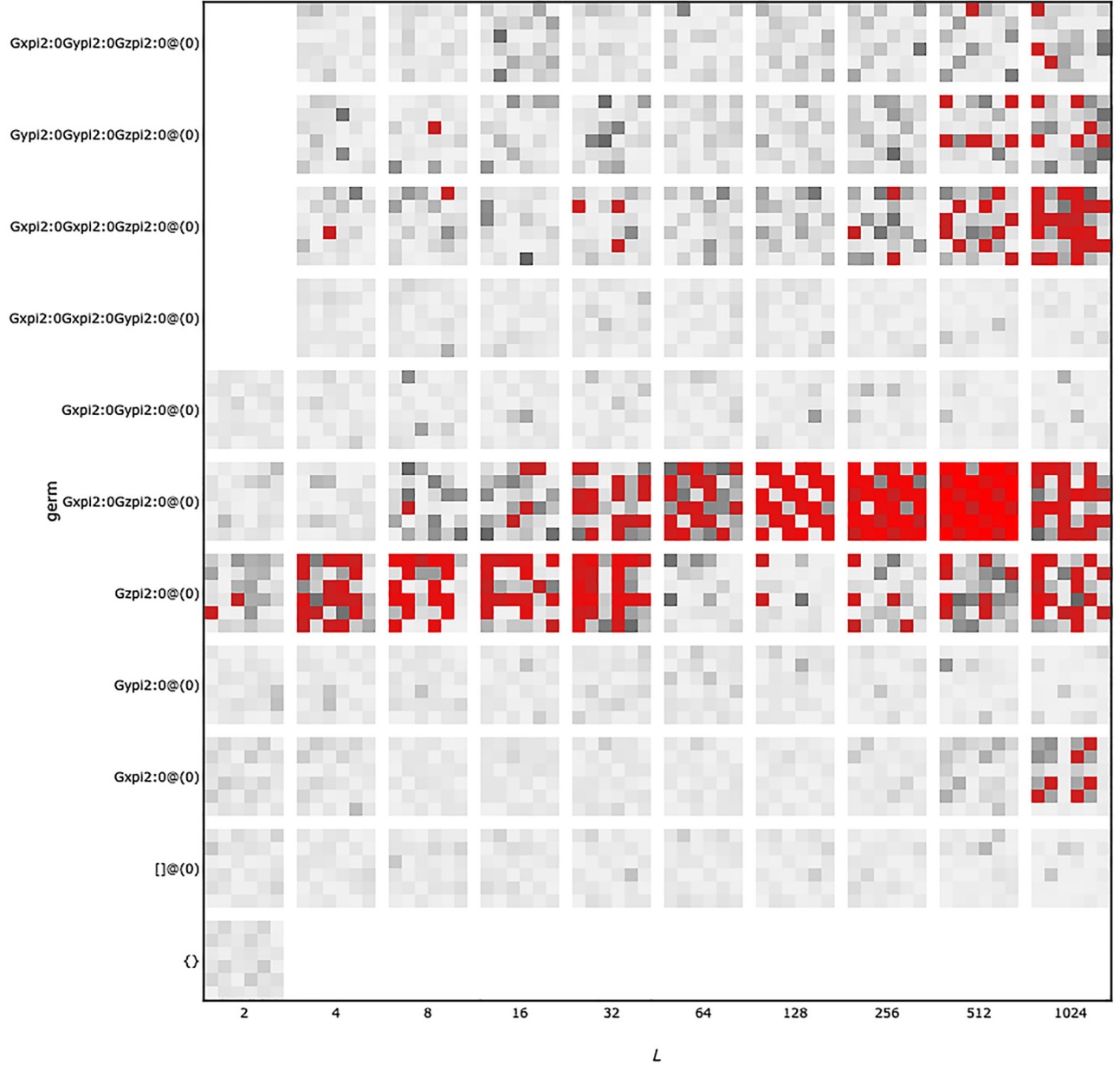


FIG. 12. Model violation in GST as reported by the loglikelihood ratio of the fit (red denotes experiments that are poorly described by the best fit Markovian model). Each row describes a different germ sequence, a small sequence $X_{\pi/2}, Y_{\pi/2}$, and $Z_{\pi/2}$ gates. These are then raised the power L columns. The 36 smaller squares inside each larger block denote different choices of the prep and measure operations.

amplitudes and phases of the cross resonance and target rotary pulse. These sequences assume only that the unitary being amplified is nearly a $ZX_{\pi/2}$ rotation, and all other terms in the amplified unitary's effective Hamiltonian are relatively small. We can extend the sequences given in Ref. [19] to include all other two-qubit Pauli errors by choosing pre, post, and echoing rotations from Table V. The magnitude of Pauli error ϵ_{ij} can be obtained from fitting $\langle Z \rangle_{CT}$ following the appropriate HEAT sequences \mathcal{H}_k^N of the form in Fig. 11 to a line in N of slope α . We choose numbers of repetitions N such that $N = 0 \pmod{4}$. So for example, $\langle \mathcal{H}_1^N \rangle_C \simeq \alpha_1(\epsilon_{xi} + \epsilon_{xx})N/2$, or equivalently, $\epsilon_{xi} \simeq (\langle \mathcal{H}_1^N \rangle_C + \langle \mathcal{H}_2^N \rangle_C)/2\alpha_1N$. We assume that all ϵ_{ij} are small compared to $\pi/2$ so that the resultant HEAT sequences fit to a line for up to $N \simeq 20$. In practice this assumption is nearly always reasonable after the execution of other standard calibration routines.

APPENDIX D: GATE-SET TOMOGRAPHY FOR A STARK $Z_{\pi/2}$

In this Appendix we use the pyGSTi implementation of long-sequence GST [17,48] to show model violation for a simulation of the Stark $Z_{\pi/2}$ gate described in Sec. III. Starting from Eq. (8), we assume a detuning of $\Delta/2\pi = -50$ MHz, a phase $\phi = 0$, and apply a square pulse of amplitude Ω and duration $t_g = 96$ ns. We then optimize Ω to perform a $Z_{\pi/2}$ gate, which results in $\Omega/2\pi \approx 16.25$ MHz. Integrating this Hamiltonian from $t = 0 \dots t_g$ yields the unitary operator, in the resonant frame, of

$$U \approx 0.717I - i(0.037X + 0.027Y + 0.695Z), \quad (\text{D1})$$

which has an average gate error of 1.5×10^{-3} . We assume the $X_{\pi/2}$ and $Y_{\pi/2}$ gates to be perfect gates with duration 35.55 ns.

We now simulate a GST experiment using this gate set, being careful to track the phase in the resonant frame, and obtain the results in Fig. 12. GST uses a loglikelihood ratio test to determine how plausible the raw data is based on the optimally fitted model. Here the only model constraints are that each gate is described by a physical quantum operation, and that the simulation is Markovian. For this experiment, GST has an extremely hard time finding a Markovian model to fit this data, even for very small numbers of pulses.

-
- [1] R. Acharya *et al.*, Suppressing quantum errors by scaling a surface code logical qubit, *Nature* **614**, 676 (2023).
 [2] 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).

- [3] P. Zhao, P. Xu, D. Lan, J. Chu, X. Tan, H. Yu, and Y. Yu, High-contrast zz interaction using superconducting qubits with opposite-sign anharmonicity, *Phys. Rev. Lett.* **125**, 200503 (2020).
 [4] S. Krinner, P. Kurpiers, B. Royer, P. Magnard, I. Tsitsilin, J.-C. Besse, A. Remm, A. Blais, and A. Wallraff, Demonstration of an all-microwave controlled-phase gate between far-detuned qubits, *Phys. Rev. Appl.* **14**, 044039 (2020).
 [5] J. M. Gambetta, Quantum-centric supercomputing: The next wave of computing (2022).
 [6] L. Viola, E. Knill, and S. Lloyd, Dynamical decoupling of open quantum systems, *Phys. Rev. Lett.* **82**, 2417 (1999).
 [7] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-Herbrüggen, and S. J. Glaser, Optimal control of coupled spin dynamics: design of NMR pulse sequences by gradient ascent algorithms, *J. Magn. Reson.* **172**, 296 (2005).
 [8] F. Motzoi, J. M. Gambetta, P. Rebentrost, and F. K. Wilhelm, Simple pulses for elimination of leakage in weakly nonlinear qubits, *Phys. Rev. Lett.* **103**, 110501 (2009).
 [9] Y. Baum, M. Amico, S. Howell, M. Hush, M. Liuzzi, P. Mundada, T. Merkh, A. R. Carvalho, and M. J. Biercuk, Experimental deep reinforcement learning for error-robust gate-set design on a superconducting quantum computer, *PRX Quantum* **2**, 040324 (2021).
 [10] C. Wang, C. Axline, Y. Y. Gao, T. Brecht, Y. Chu, L. Frunzio, M. H. Devoret, and R. J. Schoelkopf, Surface participation and dielectric loss in superconducting qubits, *Appl. Phys. Lett.* **107**, 162601 (2015).
 [11] J. M. Gambetta, C. E. Murray, Y.-K.-K. Fung, D. T. McClure, O. Dial, W. Shanks, J. W. Sleight, and M. Steffen, Investigating surface loss effects in superconducting transmon qubits, *IEEE Trans. Appl. Supercond.* **27**, 1 (2017).
 [12] E. Magesan, J. M. Gambetta, and J. Emerson, Characterizing quantum gates via randomized benchmarking, *Phys. Rev. A* **85**, 042311 (2012).
 [13] For two-qubit RB this procedure gives the error of the average two-qubit Clifford gate, which is composed of one-qubit and two-qubit gates. Here we quote the two-qubit gate error, which are the two-qubit Clifford gate error divided by 1.5, the average number of two-qubit gates per Clifford gate. Errors from single-qubit gates are assumed to be negligible by comparison.
 [14] J. Wallman, C. Granade, R. Harper, and S. T. Flammia, Estimating the coherence of noise, *New J. Phys.* **17**, 113020 (2015).
 [15] D. C. McKay, S. Filipp, A. Mezzacapo, E. Magesan, J. M. Chow, and J. M. Gambetta, Universal gate for fixed-frequency qubits via a tunable bus, *Phys. Rev. Appl.* **6**, 064007 (2016).
 [16] S. T. Merkel, J. M. Gambetta, J. A. Smolin, S. Poletto, A. D. Córcoles, B. R. Johnson, C. A. Ryan, and M. Steffen, Self-consistent quantum process tomography, *Phys. Rev. A* **87**, 062119 (2013).
 [17] E. Nielsen, J. K. Gamble, K. Rudinger, T. Scholten, K. Young, and R. Blume-Kohout, Gate set tomography, *Quantum* **5**, 557 (2021).
 [18] S. Sheldon, E. Magesan, J. M. Chow, and J. M. Gambetta, Procedure for systematically tuning up cross-talk in the cross-resonance gate, *Phys. Rev. A* **93**, 060302 (2016).

- [19] N. Sundaresan, I. Lauer, E. Pritchett, E. Magesan, P. Jurcevic, and J. M. Gambetta, Reducing unitary and spectator errors in cross resonance with optimized rotary echoes, *PRX Quantum* **1**, 020318 (2020).
- [20] E. Magesan and J. M. Gambetta, Effective Hamiltonian models of the cross-resonance gate, *Phys. Rev. A* **101**, 052308 (2020).
- [21] M. Malekakhlagh and E. Magesan, Mitigating off-resonant error in the cross-resonance gate, *Phys. Rev. A* **105**, 012602 (2022).
- [22] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland, Randomized benchmarking of quantum gates, *Phys. Rev. A* **77**, 012307 (2008).
- [23] T. Proctor, M. Reville, E. Nielsen, K. Rudinger, D. Lobsenz, P. Maunz, R. Blume-Kohout, and K. Young, Detecting and tracking drift in quantum information processors, *Nat. Commun.* **11**, 5396 (2020).
- [24] J. M. Chow, A. D. Córcoles, J. M. Gambetta, C. Rigetti, B. R. Johnson, J. A. Smolin, J. R. Rozen, G. A. Keefe, M. B. Rothwell, M. B. Ketchen, and M. Steffen, Simple all-microwave entangling gate for fixed-frequency superconducting qubits, *Phys. Rev. Lett.* **107**, 080502 (2011).
- [25] A. Kandala, K. X. Wei, S. Srinivasan, E. Magesan, S. Carnevale, G. A. Keefe, D. Klaus, O. Dial, and D. C. McKay, Demonstration of a high-fidelity cnot gate for fixed-frequency transmons with engineered zz suppression, *Phys. Rev. Lett.* **127**, 130501 (2021).
- [26] K. X. Wei, E. Magesan, I. Lauer, S. Srinivasan, D. F. Bogorin, S. Carnevale, G. A. Keefe, Y. Kim, D. Klaus, W. Landers, N. Sundaresan, C. Wang, E. J. Zhang, M. Steffen, O. E. Dial, D. C. McKay, and A. Kandala, Hamiltonian engineering with multicolor drives for fast entangling gates and quantum crosstalk cancellation, *Phys. Rev. Lett.* **129**, 060501 (2022).
- [27] D. C. McKay, C. J. Wood, S. Sheldon, J. M. Chow, and J. M. Gambetta, Efficient z gates for quantum computing, *Phys. Rev. A* **96**, 022330 (2017).
- [28] M. Malekakhlagh, E. Magesan, and D. C. McKay, First-principles analysis of cross-resonance gate operation, *Phys. Rev. A* **102**, 042605 (2020).
- [29] P. Jurcevic *et al.*, Demonstration of quantum volume 64 on a superconducting quantum computing system, *Quantum Sci. Technol.* **6**, 025020 (2021).
- [30] J. M. Gambetta, A. D. Córcoles, S. T. Merkel, B. R. Johnson, J. A. Smolin, J. M. Chow, C. A. Ryan, C. Rigetti, S. Poletto, T. A. Ohki, M. B. Ketchen, and M. Steffen, Characterization of addressability by simultaneous randomized benchmarking, *Phys. Rev. Lett.* **109**, 240504 (2012).
- [31] D. M. Zajac, J. Stehlik, D. L. Underwood, T. Phung, J. Blair, S. Carnevale, D. Klaus, G. A. Keefe, A. Carniol, M. Kumph, M. Steffen, and O. E. Dial, Spectator errors in tunable coupling architectures (2021), doi:10.48550/ARXIV.2108.11221.
- [32] C. Rigetti, A. Blais, and M. Devoret, Protocol for universal gates in optimally biased superconducting qubits, *Phys. Rev. Lett.* **94**, 240502 (2005).
- [33] Z. Li, P. Liu, P. Zhao, Z. Mi, H. Xu, X. Liang, T. Su, W. Sun, G. Xue, J.-N. Zhang, W. Liu, Y. Jin, and H. Yu, Error per single-qubit gate below 10^{-4} in a superconducting qubit, *Npj Quantum Inf.* **9**, 111 (2023).
- [34] M. Carroll, S. Rosenblatt, P. Jurcevic, I. Lauer, and A. Kandala, Dynamics of superconducting qubit relaxation times, *Npj Quantum Inf.* **8**, 132 (2022).
- [35] L. V. Abdurakhimov, I. Mahboob, H. Toida, K. Kakuyanagi, Y. Matsuzaki, and S. Saito, Driven-state relaxation of a coupled qubit-defect system in spin-locking measurements, *Phys. Rev. B* **102**, 100502 (2020).
- [36] L. V. Abdurakhimov, I. Mahboob, H. Toida, K. Kakuyanagi, Y. Matsuzaki, and S. Saito, Identification of different types of high-frequency defects in superconducting qubits, *PRX Quantum* **3**, 040332 (2022).
- [37] B. Gulácsi and G. Burkard, Signatures of non-markovianity of a superconducting qubit, *Phys. Rev. B* **107**, 174511 (2023).
- [38] A. W. Cross, L. S. Bishop, S. Sheldon, P. D. Nation, and J. M. Gambetta, Validating quantum computers using randomized model circuits, *Phys. Rev. A* **100**, 032328 (2019).
- [39] Y. Kim, C. J. Wood, T. J. Yoder, S. T. Merkel, J. M. Gambetta, K. Temme, and A. Kandala, Scalable error mitigation for noisy quantum circuits produces competitive expectation values, *Nat. Phys.* **19**, 752 (2023).
- [40] E. van den Berg, Z. K. Mineev, A. Kandala, and K. Temme, Probabilistic error cancellation with sparse Pauli-Lindblad models on noisy quantum processors, *Nat. Phys.* **19**, 1116 (2023).
- [41] J. Bylander, S. Gustavsson, F. Yan, F. Yoshihara, K. Harrabi, G. Fitch, D. G. Cory, Y. Nakamura, J.-S. Tsai, and W. D. Oliver, Noise spectroscopy through dynamical decoupling with a superconducting flux qubit, *Nat. Phys.* **7**, 565 (2011).
- [42] A. Murphy, J. Epstein, G. Quiroz, K. Schultz, L. Tewala, K. McElroy, C. Trout, B. Tien-Street, J. A. Hoffmann, B. D. Clader, J. Long, D. P. Pappas, and T. M. Sweeney, Universal-dephasing-noise injection via Schrödinger-wave autoregressive moving-average models, *Phys. Rev. Res.* **4**, 013081 (2022).
- [43] G. H. Low and I. L. Chuang, Optimal Hamiltonian simulation by quantum signal processing, *Phys. Rev. Lett.* **118**, 010501 (2017).
- [44] Y.-X. Liu, A. Ajoy, and P. Cappellaro, Nanoscale vector dc magnetometry via ancilla-assisted frequency up-conversion, *Phys. Rev. Lett.* **122**, 100501 (2019).
- [45] C. Berke, E. Varvelis, S. Trebst, A. Altland, and D. P. DiVincenzo, Transmon platform for quantum computing challenged by chaotic fluctuations, *Nat. Commun.* **13**, 2495 (2022).
- [46] A. Morningstar, L. Colmenarez, V. Khemani, D. J. Luitz, and D. A. Huse, Avalanches and many-body resonances in many-body localized systems, *Phys. Rev. B* **105**, 174205 (2022).
- [47] Qiskit Authors, Qiskit: An open-source framework for quantum computing (2021).
- [48] pyGSTi: A python implementation of gate set tomography, Online (2016).