

Circuit-Aware Gate Calibration via Grouped Channel Direct Fidelity Estimation

Júlia Barberà-Rodríguez¹ and Arthur Strauss²

¹*ICFO - Institut de Ciències Fotoniques, The Barcelona Institute of Science and Technology, 08860 Castelldefels, Barcelona, Spain*

²*Centre for Quantum Technologies, National University of Singapore*

As gate error rates approach the fault-tolerance regime, the imperfections that remain are increasingly hard to characterize. They are often coherent, and tend to depend on the surrounding circuit context, including neighbouring operations, idle periods, and control crosstalk [1]. Characterizing and suppressing these effects calls for a context-preserving protocol, one that measures the fidelity of a process as it appears inside the application circuit, rather than after averaging over an ensemble of artificially modified executions. Such estimators are also becoming important as objective functions for adaptive calibration strategies, including machine-learning-based approaches, where the same circuit must be evaluated repeatedly. Randomized methods [2–4] typically estimate average gate or layer fidelities by twirling the circuit. This averaging is what makes them scalable, but it also discards information about how a specific circuit shapes the noise. Other protocols recover this sensitivity through coherent amplification, but often require implementing the inverse of the circuit under study [5], creating a bootstrapping problem precisely when the gates being characterized are themselves imperfect.

Direct fidelity estimation (DFE) [6, 7] sits in a complementary regime: it estimates the fidelity of a gate against a known target while barely modifying the circuit. DFE importance-samples input-output Pauli pairs from the target, and for each pair it prepares an eigenstate of the input Pauli, applies the channel, and measures the output Pauli. Averaging these outcomes yields an unbiased fidelity estimate using only local preparations and measurements. It thus avoids both the cost of full process tomography and the context-disrupting randomization of benchmarking, which makes it a natural objective for calibration against coherent, circuit-dependent errors. This becomes increasingly relevant as native entangling gates evolve beyond fixed Clifford operations. Current hardware exposes continuously parameterized gates, such as the $RZZ(\theta)$ and $RX(\theta)$ fractional gates on IBM Heron [8] and the $fSim(\theta, \varphi)$ gate on Google processors [9, 10], which are non-Clifford for generic parameters, placing them outside Pauli twirling and much of the benchmarking toolbox. Despite this, DFE remains experimentally expensive. Each sampled Pauli pair generally needs its own preparation and measurement configuration, and the number of pairs grows quickly for non-Clifford targets. Readout error mitigation and better single-qubit gates have reduced DFE’s historical SPAM sensitivity, but the number of distinct configurations, and of channel evaluations, remains the bottleneck in iterative calibration.

We introduce *Grouped channel DFE*, which partitions the support of the target Pauli transfer matrix into groups whose input and output Paulis are simultaneously qubit-wise commuting. This was already proposed for states [11], but the channel case is less trivial, since each coefficient couples an input and an output Pauli. One basis configuration then estimates every Pauli-transfer coefficient in the group. We derive finite-sample guarantees for a robust median-of-means estimator and show that its channel-use complexity is governed by the Rényi- $\frac{1}{2}$ effective support of compatible groups. Grouping therefore offers two complementary advantages: it reduces the number of distinct preparation-measurement configurations required to characterize a process and, when the target has non-uniform weight within compatible groups, it also reduces the channel evaluations needed for a given accuracy. We validate these predictions numerically and study the practical behavior of the grouping. Because it concentrates more shots per configuration, it also improves the statistical effectiveness of standard readout error mitigation. We then benchmark the method on the parameterized $fSim(\theta, \varphi)$ family and use the grouped estimator as the reward in a reinforcement-learning-based calibration loop. *Grouped channel DFE* always reduces the number of distinct preparation and measurement configurations, and when the grouping is favorable it also reduces the number of channel evaluations. This makes it a practical way to measure gate fidelity in context and repeatedly, as needed for machine-learning calibration loops that suppress coherent errors.

Preliminaries. Let $\mathcal{E}$ be an unknown n -qubit quantum channel and $\mathcal{U}$ the target unitary channel, with $d = 2^n$. The Pauli transfer coefficients of a channel $\mathcal{A}$ are $\chi_{\mathcal{A}}(\alpha, \beta) = \text{Tr}(P_{\alpha}\mathcal{A}(P_{\beta}))/d$, where P_{β} is the input Pauli and P_{α} the measured output observable. The entanglement fidelity between the target and the implementation is their overlap,

$$F_e(\mathcal{U}, \mathcal{E}) = \frac{1}{d^2} \sum_{\alpha, \beta} \chi_{\mathcal{U}}(\alpha, \beta) \chi_{\mathcal{E}}(\alpha, \beta). \quad (1)$$

DFE evaluates this overlap by importance sampling. Let $\Omega = \{(\alpha, \beta) : \chi_{\mathcal{U}}(\alpha, \beta) \neq 0\}$ be the support of the target Pauli transfer matrix. Since $\sum_{\alpha, \beta} \chi_{\mathcal{U}}(\alpha, \beta)^2 = d^2$ for a unitary, the distribution $p_{\alpha, \beta} = \chi_{\mathcal{U}}(\alpha, \beta)^2/d^2$ is normalized, and Eq. (1) is the expectation of $X_{\alpha, \beta} = \chi_{\mathcal{E}}(\alpha, \beta)/\chi_{\mathcal{U}}(\alpha, \beta)$ under $p_{\alpha, \beta}$. Each sampled coefficient $\chi_{\mathcal{E}}(\alpha, \beta)$ is estimated

from repeated channel executions, by preparing eigenstates of P_β , applying $\mathcal{E}$, and measuring P_α . Standard DFE samples ℓ pairs independently and estimates each ratio separately, leading to the estimator

$$\tilde{Y} = \frac{1}{\ell} \sum_{t=1}^{\ell} \frac{\tilde{\chi}_{\mathcal{E}}(\alpha_t, \beta_t)}{\chi_{\mathcal{U}}(\alpha_t, \beta_t)}. \quad (2)$$

There are two relevant sampling costs in standard DFE. First, estimating the fidelity to additive error ϵ with failure probability at most δ requires $\ell = \mathcal{O}(1/\epsilon^2\delta)$ independently sampled input-output Pauli pairs. Second, each sampled coefficient must be estimated from repeated channel executions. The total channel-use complexity depends on the support $\Omega = (\alpha, \beta) : \chi_{\mathcal{U}}(\alpha, \beta) \neq 0$ of the target Pauli transfer matrix. Then, the expected number of channel uses scales as

$$N_{\text{ch}} = \mathcal{O}\left(\frac{1}{\epsilon^2\delta} + \frac{|\Omega|}{d^2\epsilon^2} \log\left(\frac{1}{\delta}\right)\right). \quad (3)$$

The cost is thus set by the support size $|\Omega|$: Clifford targets have $|\Omega| = d^2$, whereas a generic unitary reaches $|\Omega| = \mathcal{O}(d^4)$, a worst-case $\mathcal{O}(d^2/\epsilon^2)$ up to logarithmic and failure-probability factors. Crucially, standard DFE treats every sampled pair as a separate experimental configuration; the grouping below removes this restriction by estimating several compatible coefficients from one preparation and measurement setting.

Grouped channel DFE. To reduce both sample complexity and configuration cost, we group commuting input Paulis P_β and output Paulis P_α . Two pairs (α, β) and (α', β') are *jointly compatible* if $[P_\beta, P_{\beta'}] = 0$ and $[P_\alpha, P_{\alpha'}] = 0$. In hardware, we enforce the stronger qubit-wise commuting condition, ensuring that each group can be accessed with local basis rotations. We partition the target support Ω into compatible groups $\mathcal{G} = \{G_m\}_{m=1}^M$ using a norm-weighted greedy clique-cover heuristic [12]. For each group, write $G_m = \{(\alpha_{m,l}, \beta_{m,l})\}_{l=1}^{r_m}$, with $u_{m,l} := \chi_{\mathcal{U}}(\alpha_{m,l}, \beta_{m,l})$, $e_{m,l} := \chi_{\mathcal{E}}(\alpha_{m,l}, \beta_{m,l})$, and collect these coefficients into vectors $\mathbf{u}_m$ and $\mathbf{e}_m$. We sample group G_m with probability $p_m = \|\mathbf{u}_m\|_2^2/d^2$, and estimate its contribution the group contribution $X_m = \mathbf{u}_m \cdot \mathbf{e}_m / \|\mathbf{u}_m\|_2^2$ by an s_m -shot average $\tilde{X}_m$ drawn from the shared eigenbasis. Each replicate $\tilde{X}_{m_t}$ is an unbiased estimate of $F_e(\mathcal{U}, \mathcal{E})$. Averaging $\ell = \mathcal{O}(\epsilon^{-2}\delta^{-1})$ replicates yields the estimator $\tilde{Y} = \frac{1}{\ell} \sum_t \tilde{X}_{m_t}$.

Theorem 1 (Sample complexity). *For any grouping $\mathcal{G}$ of Ω and any $\epsilon, \delta \in (0, 1)$, the grouped estimator satisfies $\Pr[|\tilde{Y} - F_e(\mathcal{U}, \mathcal{E})| \geq \epsilon] \leq \delta$ using*

$$N_{\text{ch}} = \mathcal{O}\left(\frac{\log \delta^{-1}}{d^2\epsilon^2} \sum_{m=1}^M \kappa_m + \frac{1}{\epsilon^2\delta}\right) \quad (4)$$

channel uses and M distinct experimental configurations, where $\kappa_m = \|\mathbf{u}_m\|_1^2 / \|\mathbf{u}_m\|_2^2 = 2^{S_{1/2}(q_m)}$ is the Rényi- $\frac{1}{2}$ effective size of group G_m .

Proof sketch. The estimator error has two independent sources. *Sampling only ℓ groups:* since $\tilde{Y}$ averages ℓ i.i.d. replicates and the physicality bound gives $\mathbb{E}[X_m^2] \leq 1$, this contributes variance at most $1/\ell$, so Chebyshev fixes $\ell = \lceil \epsilon^{-2}\delta^{-1} \rceil$. *Finite shots per group:* each $\tilde{X}_m$ averages s_m bounded single-shot terms of range $\|\mathbf{u}_m\|_1 / \|\mathbf{u}_m\|_2^2$. Applying Hoeffding across all $\sum_t s_{m_t}$ shots, and splitting the budget equally over the ℓ groups, gives the per-group allocation

$$s_m = \left\lceil \frac{2\|\mathbf{u}_m\|_1^2}{\ell \epsilon^2 \|\mathbf{u}_m\|_2^4} \log \frac{2}{\delta} \right\rceil.$$

A union bound over the two sources gives the (ϵ, δ) guarantee. Computing $N_{\text{ch}} = \ell \sum_m p_m s_m$, gives Eq. (4), whose second term is the outer/ceiling cost. Finally $\kappa_m = (\sum_l \sqrt{q_{m,l}})^2 = 2^{S_{1/2}(q_m)}$ with $q_{m,l} := u_{m,l}^2 / \|\mathbf{u}_m\|_2^2$.

The improvement in Eq. (4) is most relevant for structured non-Clifford dynamics. For Clifford channels, the Pauli transfer matrix is a signed permutation and standard DFE is already essentially optimal. The useful regime is a gate with nonzero magic whose Choi–Pauli distribution stays highly non-uniform, with most weight concentrated on a small effective support inside the commuting groups. There, grouping replaces the ungrouped contribution $\sum_m \kappa_m = |\Omega|$ by $\sum_m 2^{S_{1/2}(q_m)}$, which can be much smaller for favorably structured targets, when the coefficients within each group are peaked rather than flat. The largest gains are thus expected for near-Clifford rotations, controlled-phase or phase-gadget circuits, and Clifford circuits doped with few non-Clifford gates, while for generic scrambling unitaries grouping mainly reduces the number of configurations rather than the channel use complexity.

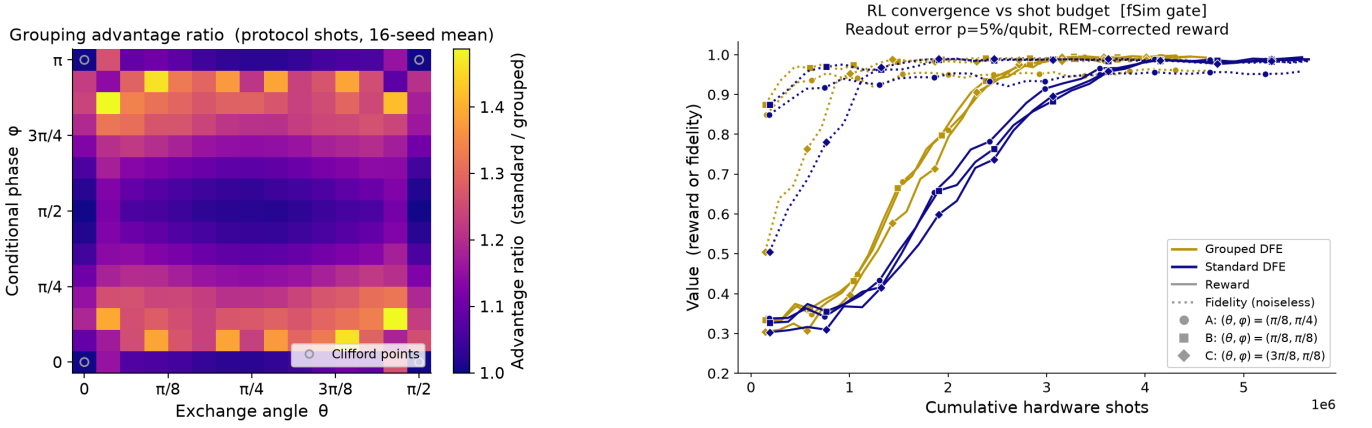


FIG. 1: (Left) Grouping advantage ratio R over (θ, φ) . (Right) Grouping advantage for fSim(θ, φ) calibration.

Readout Error Mitigation under grouping Readout errors are especially relevant for DFE because each Pauli-transfer coefficient is inferred from output parities. If these parities are biased by measurement error, then the resulting fidelity estimate is biased as well. Our protocol is compatible with linear REM [13], which replaces each raw parity $B_j^{(m,l)} \in \{\pm 1\}$ by an unbiased corrected parity $\tilde{B}_j^{(m,l)}$, no longer bounded by ± 1 ; we write its worst-case amplification as $a_{m,l} := \|\tilde{B}^{(m,l)}\|_\infty \geq 1$ ($a_{m,l} = 1$ without mitigation). Then, with REM, the grouped effective size becomes

$$\kappa_m^{\text{REM}} = \frac{(\sum_{l \in G_m} |u_{m,l}| a_{m,l})^2}{\|\mathbf{u}_m\|_2^2}, \quad (5)$$

so the same sample-complexity bound holds after replacing κ_m by κ_m^{REM} . By Cauchy-Schwarz, $\kappa_m^{\text{REM}} \leq \sum_{i \in G_m} a_i^2$, which is precisely the contribution obtained without grouping. Hence, under the same linear REM model, grouping is never worse than ungrouped DFE in the REM-amplified variance bound. If the amplification is approximately constant within a group, $a_{m,l} \simeq a_m$, then $\kappa_m^{\text{REM}} \simeq a_m^2 \kappa_m$, whereas ungrouped DFE contributes $a_m^2 |G_m|$. Hence REM preserves the same grouping advantage r_m / κ_m . Therefore, grouping reduces the REM-amplified sampling overhead whenever the target coefficients are nonuniform inside compatible groups. Moreover, with grouping a fixed experimental or readout-calibration budget can be concentrated on fewer bases. This improves the finite-shot precision of the corresponding readout assignment matrices in practical REM implementations.

Numerics. We benchmark on the fSim(θ, φ) family [14], whose exchange interaction mixes all Pauli sectors, and which interpolates between the native Clifford gates iSWAP ($\theta = \frac{\pi}{2}, \varphi = 0$), $\sqrt{\text{iSWAP}}$ ($\theta = \frac{\pi}{4}, \varphi = 0$) and CZ ($\theta = 0, \varphi = \pi$). Generic (θ, φ) gives the continuously parametrized non-Clifford targets typical of variational and time-evolution circuits. We use grouped DFE as the reward oracle in a closed-loop Proximal Policy Optimization calibration [15] of a five-parameter ansatz (RXX, RYY, CP, local RZ). Figure 1 (Left) maps the grouping advantage ratio $R = N_{\text{ch}}^{\text{std}} / N_{\text{ch}}^{\text{grouped}}$ over the (θ, φ) plane. At the Clifford corners $\{0, \frac{\pi}{2}\} \times \{0, \pi\}$ every Pauli pair forms a group and $R = 1$. Away from them, the peaked Choi–Pauli weight yields large groups with small Rényi- $\frac{1}{2}$ support, and R peaks in the intermediate regime, as predicted by the $\sum_m 2^{S_{1/2}(q_m)}$ scaling of Theorem 1. Figure 1 (Right) tracks fidelity versus cumulative hardware shots at three targets: near-Clifford ($\theta = \frac{\pi}{8}, \varphi = \frac{\pi}{4}$), peak-advantage ($\theta = \frac{\pi}{4}, \varphi = \frac{\pi}{2}$) and sustained non-Clifford ($\theta = \frac{3\pi}{8}, \varphi = \frac{\pi}{2}$), under a 5% per-qubit readout error with readout mitigation. Grouped DFE reaches the high-fidelity regime in fewer shots, most clearly at the peak-advantage point. Both oracles converge to comparable final fidelities, so the gain is in shot efficiency of the optimization trajectory rather than the attainable fidelity.

Outlook *Grouped channel DFE* converts the configuration overhead into a resource: jointly exploiting commuting input preparations and output measurements cuts the number of circuit templates and improves shot economy for non-Clifford targets. Applied to fSim(θ, φ) calibration, it provides faster closed-loop convergence in contexts which are relevant for fine-tuning gates within circuit layers for circuit-noise awareness.

-
- [1] D. C. McKay, I. Hincks, E. J. Pritchett, M. Carroll, L. C. G. Govia, and S. T. Merkel, Benchmarking quantum processor performance at scale (2023), [arXiv:2311.05933 \[quant-ph\]](#).
 - [2] 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, [Physical Review A](#) **77**, 012307 (2008).
 - [3] F. Arute *et al.*, Quantum supremacy using a programmable superconducting processor, [Nature](#) **574**, 505 (2019).
 - [4] A. Erhard, J. J. Wallman, L. Postler, M. Meth, R. Stricker, E. A. Martinez, P. Schindler, T. Monz, J. Emerson, and R. Blatt, enCharacterizing large-scale quantum computers via cycle benchmarking, [Nature Communications](#) **10**, 5347 (2019).
 - [5] D. M. Debroy *et al.*, Context-aware fidelity estimation, [Physical Review Research](#). **5**, 043202 (2023).
 - [6] S. T. Flammia and Y.-K. Liu, Direct fidelity estimation from few pauli measurements, [Physical Review Lett.](#) **106**, 230501 (2011).
 - [7] M. P. da Silva, O. Landon-Cardinal, and D. Poulin, Practical characterization of quantum devices without tomography, [Physical Review Lett.](#) **107**, 210404 (2011).
 - [8] IBM Quantum, Fractional gates, <https://quantum.cloud.ibm.com/docs/en/guides/fractional-gates> (2024), accessed 2026.
 - [9] F. Arute *et al.*, Quantum supremacy using a programmable superconducting processor, [Nature](#) **574**, 505 (2019).
 - [10] R. Acharya, D. A. Abanin, L. Aghababaie-Beni, I. Aleiner, T. I. Andersen, M. Ansmann, F. Arute, K. Arya, and Google Quantum AI and Collaborators, enQuantum error correction below the surface code threshold, [Nature](#) **638**, 920 (2025).
 - [11] J. Barberà-Rodríguez, M. Navarro, and L. Zambrano, Sampling groups of pauli operators to enhance direct fidelity estimation, [Quantum](#) **9**, 1784 (2025).
 - [12] O. Crawford, B. v. Straaten, D. Wang, T. Parks, E. Campbell, and S. Brierley, Efficient quantum measurement of pauli operators in the presence of finite sampling error, [Quantum](#) **5**, 385 (2021).
 - [13] S. Bravyi, S. Sheldon, A. Kandala, D. C. McKay, and J. M. Gambetta, Mitigating measurement errors in multiqubit experiments, [Physical Review A](#) **103**, 042605 (2021).
 - [14] B. Foxen, C. Neill, A. Dunsworth, P. Roushan, B. Chiaro, A. Megrant, J. Kelly, Z. Chen, *et al.*, Demonstrating a continuous set of two-qubit gates for near-term quantum algorithms, [Physical Review Letters](#) **125** (2020).
 - [15] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, Proximal policy optimization algorithms (2017), [arXiv:1707.06347 \[cs.LG\]](#).